Accès sémantique aux bases de données documentaires. Techniques symboliques de traitement automatique du langage pour l'indexation thématique et l'extraction d'informations temporelles

Laurent Kevers

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
FACULTÉ DE PHILOSOPHIE ET LETTRES

ACCÈS SÉMANTIQUE AUX BASES DE DONNÉES DOCUMENTAIRES

Techniques symboliques de traitement automatique du langage
pour l'indexation thématique et l'extraction d'informations temporelles

Thèse présentée par Laurent KEVERS
en vue de l'obtention du grade de Docteur en Langues et lettres

Sous la direction du Professeur Cédrick FAIRON

Membres du jury : Prof. Heinz BOUILLON (Président)
Prof. Cédrick FAIRON (Directeur, UCL)
Prof. Laurence DANLOS (Université Paris 7)
Prof. Manuel KOLP (UCL)
Prof. Eric LAPORTE (Université Paris-Est Marne-la-Vallée)
Prof. Piet MERTENS (KULeuven)

LOUVAIN-LA-NEUVE
ANNÉE ACADÉMIQUE 2010-2011
CENTRE DE TRAITEMENT AUTOMATIQUE DU LANGAGE

Aux Autres.

Remerciements

Après huit années passées au Cental, je me dois de remercier de nombreuses personnes. Les nommer toutes serait difficile, mais je désire néanmoins les associer au résultat de mon travail.

En particulier, j'exprime ma reconnaissance à Cédrick Fairon, qui depuis cette fameuse « journée découverte entreprise » de 2001, m'a fait confiance et m'a toujours incité, jusqu'à la remise de cette thèse, à aller plus loin. Les années passées au Cental sous sa direction, au sein d'une équipe motivée et enthousiaste, m'ont énormément apporté tant au niveau professionnel que personnel.

Je tiens à remercier les membres du jury – Laurence Danlos, Manuel Kolp, Eric Laporte et Piet Mertens – de s'être rendus disponibles en si peu de temps et d'avoir accepté d'effectuer une lecture critique mais constructive de ce travail. Merci également à Heinz Bouillon d'avoir présidé ce jury.

Si j'ai pu finir dans les temps, c'est un peu grâce à Babette Dehottay et à Hubert Naets, qui m'ont retiré de (très) grosses épines du pied lors de ma période de rédaction. Pour leur lutte contre mon imagination orthographique, je dois remercier Julia Medori et Jacques Kevers. Je leur suis reconnaissant de ne pas m'avoir demandé si ma thèse constituait une proposition pour une réforme de l'orthographe. Merci à Valérie Martin d'avoir répondu à l'avalanche de questions à laquelle je l'ai soumise durant les derniers mois. Enfin, si la défense privée s'est bien déroulée, c'est grâce à l'organisation et l'assistance technique assurée avec brio par Hubert Naets et Julia Medori.

Merci à l'équipe Stratego (Marco Saerens, Hugues Bersini, Pascal Francq, Amin Mantrach, Jérôme Callut, Nicolas van Zeebroeck et Joachim De Beule) pour ces trois ans de collaboration. Tous mes remerciements vont aussi à l'ensemble des membres, passés ou présents, du Cental (Cédrick, Anne, Patrick, Babette, Claude, Michel, Marc, Sébastien, Piet, Hubert, Julia, Noémi, Kévin, Richard, Thomas, Amélie, Sophie, Olivier, Isabelle, Sacha, Jean-Léon, Olga, Nadja, les Stéphanies, Nicolas, Vincent et Julien) pour ces huit années passionantes et enrichissantes. Merci à Patrick Watrin pour ses conseils et ses encouragements, ainsi que d'avoir entretenu mon addiction au café et mon tabagisme passif. Enfin, merci aussi à Bastien Kindt pour, entre autres, ses tentatives de m'apprendre le grec ancien.

Je ne peux évidemment pas terminer cette liste en oubliant ma famille, ainsi que la famille Medori. Merci à vous tous ! Enfin, merci à ma Julia d'être là et d'avoir supporté ces derniers mois... Les prochains seront tout autres !

Cette recherche n'aurait pu être menée sans le soutien de la Région Bruxelloise et de la Région Wallonne, grâce auxquelles les projets B-Ontology (IRSIB) et Stratego (Wist2) ont vu le jour.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

La transmission du savoir et de la connaissance a, au cours de l'histoire, connu diverses formes et emprunté de nombreux canaux. Avec le développement et l'enseignement de l'écriture, la tradition orale a progressivement cédé sa place à une diffusion de l'information prenant la forme d'un document écrit[1]. L'essor de l'imprimerie, et l'alphabétisation croissante de la population, a été une étape importante dans l'adoption de ce format comme vecteur principal de l'information et de la connaissance. Celles-ci ont alors pu se développer sur un média pérenne[2] et diffusable à grande échelle.

La circulation de l'information écrite a cependant connu un tournant radical suite au développement des technologies de l'information et de la communication (TIC), lors de ce qui peut être désigné comme une révolution, à la fin du siècle passé. La numérisation et la dématérialisation des documents, ainsi que la création du réseau Internet, sont à ranger parmi les impacts concrets les plus visibles de cette révolution. Ces facteurs réunis ont eu un effet multiplicateur sur des tendances déjà présentes, à savoir l'augmentation du volume et la rapidité de production et de circulation de l'information.

Au même titre que la révolution industrielle au XIX^{e} siècle, la *révolution numérique* provoque de nombreux et importants impacts sur l'économie, la société, l'environnement, etc. En effet, en une bonne vingtaine d'années, l'informatique encore balbutiante et confinée aux laboratoires scientifiques et militaires[3] s'est transformée en une technologie aujourd'hui utilisée par une large partie de la population à des fins professionnelles ou privées[4]. De nombreux aspects de notre société ont été modifiés et influencés par les TIC. Citons, à titre d'exemple, l'impact sur :

- la vie quotidienne : envoi de courriels à la place de lettres, lecture des sites d'informations plutôt que des *journaux-papier*, achats sur internet, *e-learning*, guichet électronique (administration), *e-health*, travail à domicile, domotique, etc. ;
- l'économie : apparition, disparition ou modification de certains métiers et activités économiques. Par exemple, le domaine du journalisme, et plus précisément la presse écrite, a amorcé une mutation profonde, le format électronique remplaçant progressivement le format papier[5]. Dans ce cadre, les métiers relatifs à la production du support

[1] L'adoption du document écrit ne s'est évidemment pas déroulée de la même manière dans toutes les parties du monde. De même, suivant les milieux et/ou les domaines, la transmission orale du savoir et du savoir-faire subsiste parfois de manière importante, par exemple pour les métiers manuels (David et Foray [2002]).

[2] Mais néanmoins pas exempt de problèmes liés à sa conservation.

[3] Le réseau ARPANET s'est ouvert, sous la forme du réseau Internet, à une utilisation commerciale au début des années 1990 (Zakon [2010]).

[4] Du moins dans les sociétés dites *économiquement développées*.

[5] Dawson [2010] prévoit que les journaux sous leur forme actuelle ne seront plus *pertinents* d'ici 2017 aux États-Unis,

disparaissent progressivement, du moins dans ce secteur d'activité, alors que les métiers en rapport avec l'élaboration du contenu sont modifiés. De nouveaux modes de production apparaissent également, dans la foulée du *crowdsourcing*[6], en impliquant la participation du public dans la création du contenu (Greenslade [2010]) ;

– le domaine social : apparition et accentuation d'inégalités sociales dues à l'incapacité de certains à avoir accès aux technologies de l'information, phénomène qui est parfois nommé *fracture numérique* (La Documentation française [2007], Lacroix [2010]) ;
– la culture : les langues peu représentées sur internet ou peu supportées par les logiciels sont affaiblies et, à terme, menacées de disparition (Diki-Kidiri [2007]). En cela, elles sont accompagnées de la culture qu'elles véhiculent.

En particulier, le fonctionnement des entreprises a été progressivement mais profondément modifié, à tel point que l'informatisation et l'adoption des technologies de l'information ont très souvent eu des conséquences importantes au niveau social et organisationnel[7]. La matière première et le produit de l'activité économique se sont également transformés. Nous sommes en effet passés d'une société industrielle à une société de la connaissance[8]. La richesse et la prospérité ne sont aujourd'hui plus majoritairement créées à partir des matières premières, des usines et des procédés de fabrication qu'elles mettent en œuvre, mais bien à partir de l'information et de la connaissance[9].

La société, et l'économie, de l'information sont cependant parvenus à un paradoxe. L'évolution de l'écrit manuscrit ou imprimé vers le format électronique a d'une part rendu une grande quantité de documents accessibles à un nombre important de personnes, mais a d'autre part eu tendance à noyer ces informations dans une masse documentaire si vaste qu'elle a rendu leur identification difficile. Face à cette (sur)abondance de documents numériques disponibles sur l'Internet, dans les entreprises ou dans les administrations[10] (Boughanem *et al.* [2006]), et étant donné le nouveau statut de matière première de l'information et de la connaissance, le problème de l'accès à celles-ci est devenu un enjeu stratégique.

En pratique, le défi se présente de diverses manières. Parmi les buts poursuivis en ce qui concerne l'accès à l'information, les utilisateurs cherchent à[11] :

2026 en Belgique, 2029 en France, etc.

[6] Processus d'externalisation d'une activité vers une foule, une communauté, qui prend en charge la réalisation, éventuellement collaborative, d'une tâche. Ce principe a notamment été exposé par Howe [2006].

[7] « [...] le développement massif des technologies de l'information et de la communication, ouvre aux entreprises des possibilités considérables de réorganisation de leur production et de recentrage sur les activités à plus forte valeur ajoutée. » (Levy et Jouyet [2006]).

[8] La récente stratégie de développement européenne, *Europe 2020*, présente l'économie de la connaissance comme l'un des piliers de la croissance pour la prochaine décennie (Commission européenne [2010]).

[9] « Durant les Trente Glorieuses, le succès économique reposait essentiellement sur la richesse en matières premières, sur les industries manufacturières et sur le volume de capital matériel dont disposait chaque nation. Cela reste vrai, naturellement. Mais de moins en moins. Aujourd'hui, la véritable richesse n'est pas concrète, elle est abstraite. Elle n'est pas matérielle, elle est immatérielle. C'est désormais la capacité à innover, à créer des concepts et à produire des idées qui est devenue l'avantage compétitif essentiel. Au capital matériel a succédé, dans les critères essentiels de dynamisme économique, le capital immatériel ou, pour le dire autrement, le capital des talents, de la connaissance, du savoir. » (Levy et Jouyet [2006]).

[10] En ce qui concerne les documents disponibles sur le web, une étude d'octobre 2010 fait état de 232.839.963 de sites, pour seulement 182.226.259 deux ans plus tôt (`http://news.netcraft.com/archives/category/web-server-survey/`).

[11] La liste présentée ne se veut pas exhaustive.

- rassembler un ensemble de documents relatifs à un ou plusieurs sujets, soit en sélectionnant des thèmes prédéfinis (ex. : à partir d'une liste de thèmes liés à l'écologie, chercher les documents sur les marées noires), soit en mentionnant des entités d'un type particulier telles que des personnes ou des organisations (ex. : les documents relatifs à « Steve Jobs » ou à la société « Apple ») ou encore en fournissant des mots-clés quelconques (ex. : les documents dans lesquels apparaissent les mots ou expressions « Union Européenne » et « politique agricole commune ») ;
- organiser une masse documentaire en sous-collections cohérentes relativement à leur contenu (ex. : répartir l'ensemble des dépêches de presse d'une journée en plusieurs groupes correspondant à différents sujets) ;
- obtenir une vision synthétique de documents sélectionnés (ex. : résumer un ou plusieurs documents portant sur une mission en Afghanistan) ;
- consulter des informations structurées portant sur un sujet abordé dans une collection de documents (ex. : parcourir des éléments biographiques concernant Herman Van Rompuy) ;
- mener une veille stratégique et être averti de tout nouveau document qui aborde un sujet particulier (ex. : repérer tout article de presse mentionnant des fusions ou des acquisitions d'entreprises) ;
- obtenir une réponse à une question précise (ex. : quels sont les avocats spécialisés en propriété intellectuelle) ;
- filtrer des documents en fonction de certaines de leur caractéristiques (ex. : détecter la langue utilisée, écarter les documents non-pertinents tels que les spams, etc.).

Ces besoins peuvent être comblés, au moins en partie, à l'aide de diverses techniques de traitement automatique du langage (TAL). Parmi celles-ci on citera plus particulièrement celles qui ressortissent :

- à la recherche d'informations dont, entre autres, les technologies sous-jacentes aux moteurs de recherche (tant en ce qui concerne l'indexation que les modes de requête) ;
- à l'extraction d'informations, qu'elle soit à des fins d'indexation ou de création de bases de connaissances ;
- à la classification ;
- au *clustering* ;
- au résumé automatique ;
- aux systèmes de question-réponse.

Les différentes technologies concernées ont atteint des niveaux de performances qui leur permettent d'être utilisées efficacement dans de nombreux cas concrets[12]. Il subsiste cependant de la place pour faire évoluer ces systèmes vers des résultats toujours plus complets et précis. Un axe de recherche important en la matière concerne l'émergence de technologies *sémantiques*, c'est-à-dire qui ne s'arrêtent plus à la *forme* du contenu des documents, mais qui tiennent également compte, d'une manière

[12] Par exemple, les moteurs de recherche tels que Google sont utilisés de manière satisfaisante par de nombreux utilisateurs.

ou d'une autre, du *sens* de celui-ci.

Cette thèse a pour objectif de montrer dans quelle mesure des techniques symboliques de traitement automatique du langage peuvent venir contribuer à l'enrichissement sémantique de la représentation des documents, d'une manière qui serait susceptible d'en améliorer l'accès.

L'information dispose d'une dimension thématique qui définit ce sur quoi elle porte. Nous avons observé qu'elle peut également être accompagnée de dimensions supplémentaires, telles que des informations temporelles (une date ou une période de temps) ou géographiques (un lieu), qui s'apparentent à des *métadonnées*[13]. L'ensemble de ces éléments constitue une information multidimensionnelle. Alors que la dimension thématique est très large, et très variable du point de vue de son expression et de son apparition dans les textes, les dimensions temporelle et géographique sont elles beaucoup plus stables de ce point de vue[14].

Étant donné cette multidimensionnalité de l'information, la recherche proposée dans cette étude se déroule en deux temps. Pour la dimension *thématique*, la partie I présente une méthode semi-automatique de classification de documents qui permet d'améliorer le processus d'indexation lorsque celui-ci est effectué, souvent manuellement, par rapport à un ensemble de catégories déterminées. Ces dernières apportent une sémantique bien définie à chaque document ainsi indexé. Le système proposé repose sur l'exploitation, à l'aide de techniques de traitement automatique du langage, de la définition lexicale de ces catégories[15]. L'approche mise au point est suffisamment générale et adaptable pour pouvoir être appliquée, moyennant l'existence d'une ressource terminologique de départ, à n'importe quel thème.

D'autre part, cette thèse aborde aussi le problème de l'information temporelle (Partie II). Dans un premier temps, un système de reconnaissance et d'interprétation de données temporelles est présenté. L'accent est placé sur l'extraction des expressions temporelles qui peuvent être situées dans un espace du temps usuel, tel qu'un calendrier, mais sans négliger l'aspect imprécis que revêt souvent ce type d'information. Le système d'extraction de données temporelles s'appuie sur une spécification précise des différents types d'expressions, qui constitue un apport méthodologique et pratique important. Dans un second temps, il est montré comment relier la dimension thématique à la dimension temporelle, à partir d'un index thématique et de l'extraction d'informations temporelles, dans le but de fournir une indexation thématique à dimension temporelle – c'est à dires une indexation multidimensionnelle – susceptible d'offrir de nouvelles perspectives aux systèmes de recherche d'informations.

[13] La notion de métadonnée est due en premier lieu à Bagley [1968]. Il s'agit d'une « donnée au sujet de la donnée », ou encore d'une « information au sujet d'une information » (NISO [2004]). Cette définition très ouverte est représentative de la diversité qui se cache derrière le concept de métadonnée. En effet, de nombreuses initiatives ont proposé des cadres de définition de métadonnées spécifiques à certains domaines. Citons par exemple Dublin Core (`http://dublincore.org`), initialement prévu pour la caractérisation de ressources sur le Web, TEI (Text Encoding Initiative, `http://www.tei-c.org/index.xml`), axé sur les textes sous forme digitale, ou encore IPTC (International Press Telecommunications Council, `http://www.iptc.org`), pour les documents de presse.

[14] Ce constat a par exemple été réalisé par Le Parc-Lacayrelle *et al.* [2007] dans le cadre des systèmes de recherche d'informations géographiques.

[15] Celle-ci est contenue dans une ressource terminologique de base, qui prend par exemple la forme d'un thésaurus.

Dans le cadre de cette étude, le champ d'action est limité aux documents écrits. En effet, bien que ces dernières années, les données sonores, les images ou les vidéos aient fait une apparition remarquée dans le monde numérique, le texte reste cependant le média le plus courant[16] et constitue toujours un format privilégié pour la diffusion des informations *importantes*[17]. D'un point de vue pratique, ces documents représentent également le matériau qui se prête actuellement le mieux, à l'aide de moyens informatiques, à une analyse centrée sur la langue. Les différents développements présentés s'appliquent à des textes en français.

Enfin, cette thèse ne s'intéresse pas à l'information déjà structurée, souvent sous la forme de bases de données, qui se prête plus à des analyses telles que le *data mining* ou le *graph mining*. De nombreux liens peuvent cependant être tissés entre ces disciplines et les techniques de TAL, ces dernières pouvant entre autres être exploitées pour créer des données structurées à partir de textes libres (non structurés).

[16] Chaumier et Dejean [2003] ont estimé en 2003 que l'information textuelle atteignait une proportion de 80%.

[17] Il serait cependant caricatural d'attribuer le statut d'information *sérieuse* à tout document écrit, et de *divertissement* aux formats multimédias. Néanmoins, les documents officiels émanant des entreprises ou des administrations restent fidèles au format écrit, ce qui constitue un indicateur important. D'autres secteurs non négligeables en ce qui concerne l'émission d'informations, les organes de presse pour ne citer qu'un seul exemple, ont par contre tendance à produire de plus en plus de documents électroniques sous une forme sonore, photographique, voire multimédia. Ces aspects sortent cependant du cadre de cette thèse.

Première partie

Indexation thématique semi-automatique

Rien ne vaut la recherche lorsqu'on veut trouver quelque chose.
J.R.R. Tolkien – Bilbo le Hobbit

L'information peut tout nous dire. Elle a toutes les réponses.
Mais ce sont des réponses à des questions que nous n'avons pas posées,
et qui ne se posent sans doute même pas.
Jean Baudrillard – Cool Memories - 1980-1985

Le langage est un fameux véhicule et, contrairement aux autres, il ne coûte rien.
Claude Duneton

CHAPITRE 1

INTRODUCTION

1.1 Le problème de l'accès à l'information

Le problème de l'accès à l'information n'est pas neuf. Il a déjà été abordé dans le domaine des sciences documentaires, pour des collections de *documents-papier* dans un premier temps, pour des ensembles de ressources électroniques ensuite. Avec l'avènement du réseau Internet et du Web[1], c'est un nouveau type de collection documentaire qui est apparu. Son importance en ce qui concerne le nombre de documents et d'utilisateurs ainsi que l'accès largement public, au contraire de certaines archives présentes dans les entreprises et autres grandes organisations, ont alors entraîné une concentration importante des innovations dans ce secteur. Si le web reste un cas particulier de collection de documents, les technologies développées pour y accéder sont néanmoins souvent applicables d'une manière générale à tout ensemble documentaire numérique.

Actuellement, l'accès aux collections électroniques de documents est souvent réalisé à l'aide de mots clés. Ce système, s'il rencontre un certain succès, que ce soit sur le Web ou dans le cadre d'autres fonds documentaires, est loin d'être idéal. Le problème de l'ambiguïté lexicale et celui représenté par les multiples possibilités d'expression d'une information sont des obstacles importants au bon fonctionnement des systèmes de recherche. En fait, ces derniers maîtrisent difficilement tout ce qui fait la diversité et la richesse d'une langue naturelle. Une méthode de recherche performante se doit de prendre ces aspects en compte, voire même de les dépasser. Afin de maximiser la couverture et la précision d'une recherche par rapport à une collection de documents, il peut être profitable de passer d'un espace de mots à un espace de concepts. L'accès aux documents devrait donc idéalement se dérouler sur une base sémantique et non lexicale. Si cet objectif est assez ambitieux et encore en grande partie hors de portée des technologies actuelles, il n'en demeure pas moins intéressant de se demander comment, dans un premier temps, apporter des éléments de sens à la représentation et à l'indexation des documents. Ce qui rend cette tâche difficile, c'est le caractère souvent hétérogène des collections de documents qui entraîne de nombreuses difficultés lors de l'inventaire, de la manipulation, du jugement de la qualité et de la pertinence, et finalement de l'indexation même des documents.

[1] World Wide Web, désigne l'ensemble des documents disponibles sur le réseau Internet, reliés par des liens hypertextes et visualisables à l'aide d'un navigateur. Internet est le réseau informatique par lequel sont accessibles ces documents. L'usage courant confond souvent, de manière erronée, les deux termes.

D'abord, les ensembles de documents ne sont pas nécessairement organisés selon un plan précis, que ce soit logiquement ou physiquement. Ensuite, il existe une grande variété de formats de documents (formats de fichiers et organisation du texte dans le document), et leur contenu n'a pas toujours fait l'objet d'une validation. De plus, ces documents sont parfois difficilement accessibles[2]. Ces différents obstacles ne se retrouvent pas dans toutes les collections de documents, mais le Web en concentre une bonne partie. Certaines de ces difficultés étaient déjà connues et présentes avant l'expansion numérique, mais cette dernière a généralement eu un effet amplificateur, et les a rendues plus critiques. Concrètement, pour une ressource documentaire telle que le Web, un certain nombre de difficultés peuvent être mises en évidence :

- Le nombre de documents à traiter est tel qu'en pratique il est très difficile d'atteindre l'exhaustivité.
- La diversité thématique est très élevée, de nombreux domaines étant abordés.
- Le degré d'intérêt[3] des documents est variable. L'information proposée peut être cruciale ou très importante, ou au contraire complètement anecdotique.
- La qualité du contenu peut varier très fortement (la facilité de production et de diffusion permet à tout un chacun de produire des documents, indépendamment de toute contrainte éditoriale).
- L'authenticité des documents n'est pas toujours garantie et est parfois difficile à établir (possibilité de faux, difficulté de distinguer ce qui relève de l'opinion ou des faits, etc.).
- L'existence de redondances complètes suite à la diffusion par différents canaux, ou partielles suite à l'achat ou à la citation de contenu, opérations durant lesquelles le texte peut éventuellement être modifié.
- Les modes de diffusions numériques favorisent la circulation de documents parfois très courts qui ne présentent souvent que des informations partielles (par exemple les flux RSS, le système Twitter, etc.).
- L'information est disséminée en de nombreux endroits.
- L'existence d'une multitude de formats (encodage des caractères, format du document, structuration de l'information à l'intérieur du document, etc.).
- L'information est exprimée au moyen de beaucoup de langues différentes.

Face à ces obstacles, plusieurs domaines de recherche, principalement la *recherche d'informations* et l'*extraction d'informations* ont tenté de proposer des technologies pour améliorer l'accès à l'information selon des approches différentes. Après avoir brièvement introduit ces deux domaines à la section 1.2, nous proposerons un aperçu des différentes solutions qui ont été créées pour l'accès au Web à la section 1.3.

[2] Par exemple, les documents qui font partie de ce qui est appelé le *web invisible* n'ont pas de *pointeurs* permettant d'y accéder facilement.

[3] Cette notion est cependant en partie subjective. Ce qui est souligné ici est le fait que toutes les informations n'ont pas nécessairement le même statut.

1.2 Recherche d'informations et extraction d'informations

Grishman [1997] définit l'extraction d'informations (EI) comme étant :

> « the identication of instances of a particular class of events or relationships in a natural language text, and the extraction of the relevant arguments of the event or relationship. Information extraction therefore involves the creation of a structured representation (such as a data base) of selected information drawn from the text. »

Cette définition se situe dans la droite ligne de l'approche adoptée au cours des conférences MUC[4], *Message Understanding Conference* (Grishman et Sundheim [1996]), qui à partir du début des années 1990, ont contribué à fonder ce courant de recherche. Il peut sembler un peu réducteur de ne mentionner comme objet de l'extraction que les seuls événements et relations, mais ceux-ci peuvent être considérés selon une interprétation large qui se référera à un ensemble beaucoup plus vaste de types d'informations. D'aucuns préféreront cependant une formulation un peu plus générale, comme celle donnée par Moens [2006] :

> « Information extraction is the identification, and consequent or concurrent classification and structuring into semantic classes, of specific information found in unstructured data sources, such as natural language text, making the information more suitable for information processing tasks. » (p. 4)

L'extraction d'informations consiste donc à rechercher des éléments spécifiques, définis par la tâche d'extraction, dans des textes non structurés (en langage naturel) et à les caractériser selon les catégories définies au préalable. Ce processus peut-être vu comme une étape de (pré)traitement destiné à produire un document plus propice au traitement automatique, ou au contraire, si les informations extraites constituent le résultat attendu, comme un aboutissement.

En recherche d'informations (RI), l'approche est différente. Baeza-Yates et Ribeiro-Neto [1999] en exposent le principe général :

> « the primary goal of an IR system is to retrieve all the documents which are relevant to a user query while retrieving as few as non-relevant documents as possible. » (p. 2)

Un aspect important réside dans l'ordre de présentation des résultats :

> « To be effective in its attempt to satisfy the user information need, the IR system must somehow 'interpret' the contents of the information items (documents) in a collection and rank them according to a degree of relevance to the user query.» (Baeza-Yates et Ribeiro-Neto [1999], p. 2)

L'activité de recherche implique une tâche préalable : l'*indexation* des documents. Celle-ci peut être effectuée selon diverses méthodes et produire différents types d'index. La recherche d'informations se déroule donc la plupart du temps en deux phases. Tout d'abord, les documents sont analysés afin d'y relier des clés d'indexation ou de les classer dans des catégories. Ensuite, la recherche consiste

[4] `http://www-nlpir.nist.gov/related_projects/muc/index.html`

à comparer les requêtes formulées par les utilisateurs à cet index afin de retrouver les documents pertinents.

La distinction faite entre extraction d'informations et recherche d'informations n'est, dans la pratique, pas si tranchée. En effet, l'extraction peut faire appel à des techniques de recherche, et inversement. Par exemple, les systèmes de classification mis au points en RI peuvent être utilisés en amont de l'EI afin de séparer les documents en sous-corpus plus homogènes ou, de manière encore plus fine, pour sélectionner des phrases à analyser de manière plus détaillée (Nédellec *et al.* [2001]). De même, l'EI peut, entre autres, réduire un document représenté initialement par son contenu entier à un ensemble particulier de mots ou d'expressions et ainsi diriger l'indexation (Riloff et Lehnert [1994], Fairon et Watrin [2003]). Les deux domaines sont donc complémentaires.

Dans cette thèse, nous nous intéressons principalement à la recherche d'informations, en tant que moyen d'améliorer l'accès aux documents et, par conséquent, à l'information qu'ils contiennent. L'extraction d'informations sera cependant massivement utilisée pour atteindre cet objectif. Plus particulièrement, l'analyse temporelle présentée à la partie II, relève de l'EI mais est finalement mise au service du système de classification et d'indexation présenté au chapitre 8. Bien entendu, ce système ne représente qu'un exemple possible d'utilisation de l'analyse temporelle parmi bien d'autres. Les développements consentis en la matière sont donc exploitables de diverses manières, que ce soit pour des applications en recherche ou en extraction d'informations.

1.3 Les systèmes de recherche d'informations

Avant toute chose, précisons que nous écartons de la recherche d'informations, les systèmes purement encyclopédiques, telles que Universalis[5] ou Wikipedia[6]. Même si ceux-ci satisfont à un certain nombre de critères que nous attendons d'un système de recherche d'informations performant, c'est-à-dire, entre autres, un accès à l'information partiellement basé sur le sens (grâce à des classification par catégories ou par thèmes) ou une certaine qualité de l'information[7], ils doivent avant tout être considérés comme un ensemble de documents parmi d'autres. En effet, la couverture thématique et surtout la diversité des documents proposés est forcément limitée. Nous nous intéressons ici, au contraire, aux méthodes rendant possible l'accès à une collection quelconque de documents (textuels) numériques, potentiellement très vaste, dont l'exemple le plus parlant est le Web[8].

Les obstacles présentés à la section 1.1 expliquent en grande partie pourquoi, encore aujourd'hui, il est parfois ardu de trouver les documents pertinents parmi ce type de ressources, et ce malgré les efforts pour développer des systèmes de recherche performants. Comme nous l'avons déjà mentionné, ces systèmes procèdent généralement en deux temps : l'indexation des documents, qui permet en-

[5] `http://www.universalis.fr`

[6] `http://www.wikipedia.org`

[7] La question de la qualité de l'information fournie par une encyclopédie de type collaboratif, telle que Wikipedia, peut être discutée, mais sort de notre propos.

[8] Dans les paragraphes qui suivent, nous citons à de nombreuses reprises des exemples issus du Web. Celui-ci concernant à la fois un très grand nombre de documents et d'utilisateurs, les technologies de recherche d'informations se sont naturellement développées dans ce milieu. Les principes évoqués restent cependant valables dans le cadre d'un intranet ou d'une collection privée de documents électroniques.

suite une interrogation de l'index au moyen d'une requête. Cette dernière étape se décompose plus précisément en deux parties : d'une part, la formulation de la requête, et d'autre part, la confrontation de celle-ci à l'index.

Au cours du temps, les techniques d'indexation ont bien entendu évolué dans le but de concilier deux objectifs *a priori* opposés, l'efficacité du processus de traitement et l'obtention d'une représentation la plus complète et la plus adéquate possible du document dans l'index. L'indexation peut être réalisée de manière manuelle ou automatique et les clés d'index peuvent se situer dans l'espace des mots, sous la forme de mots clés librement choisis, ou dans un espace plus conceptuel, dont les éléments – des catégories – sont prédéfinis et porteurs d'un sens précis.

En ce qui concerne les requêtes, il existe divers moyen de les exprimer : à l'aide de mots clés, en utilisant le langage naturel, par sélection ou navigation dans un ensemble de catégories prédéfinies ou encore en passant par une architecture de facettes.

Quant au processus de confrontation de la requête à l'index, il peut faire intervenir divers processus, automatiques ou requérant l'intervention de l'utilisateur.

Ces aspects sont exposés et illustrés au fur et à mesure de l'examen des différents systèmes. Après un rapide aperçu des premiers développements (Section 1.3.1), les systèmes actuels sont passés en revue selon qu'ils utilisent un espace fermé ou ouvert de clés d'indexation (Sections 1.3.2 et 1.3.3). Finalement, les dernières évolutions en matière de moteurs sémantiques sont abordés (Section 1.3.4).

1.3.1 Les premiers systèmes

Au début des années 1990, les premiers développements[9] permettant de rechercher des documents sur Internet furent à l'image du réseau auquel ils s'appliquaient, c'est-à-dire limités, surtout en comparaison avec ce qui a vu le jour par la suite. Le premier moteur de recherche, *Archie*[10], se résumait à une simple liste de documents qui permettait des requêtes sur les noms de ceux-ci. Par la suite, un moteur tel que *JumpStation*[11] a permis d'étendre la recherche à une partie limitée du document, les titres en l'occurrence. Avec l'intensification de l'utilisation du réseau Internet et du Web, la quantité de documents devint ensuite de plus en plus importante. Les outils de recherche d'informations s'adaptèrent et se développèrent alors en conséquence, pour finalement aboutir aux systèmes que nous connaissons aujourd'hui. Ceux-ci sont présentés au cours des sections suivantes.

[9] Une présentation de l'histoire des moteurs de recherche peut être consultée sur le site `http://www.searchenginehistory.com`.

[10] Créé à la McGill University, à Montréal, en 1990.

[11] Lancé en 1993 à la University of Stirling, Écosse.

1.3.2 Indexation dans un espace fermé de clés

Les répertoires de liens

Les répertoires de liens construits de manière manuelle sont probablement parmi les systèmes les plus simples technologiquement parlant, mais ils ont malgré tout atteint une certaine popularité, surtout au début du Web. Citons par exemple l'*Open Directory Project* (dmoz)[12] (Figure 1.1), qui continue d'ailleurs encore aujourd'hui de proposer un répertoire entretenu et enrichi manuellement par des éditeurs humains bénévoles. Chaque site web intégré dans la ressource est ajouté à la catégorie qui lui correspond le mieux. L'ensemble des références est organisé selon une hiérarchie de catégories parmi lesquelles l'utilisateur navigue afin de trouver la section qui l'intéresse, pour ensuite explorer les liens qui y sont contenus. Alternativement, une recherche par mots-clés (voir section 1.3.3), limitée aux sites référencés dans le répertoire, est également proposée.

Figure 1.1 : Le répertoire de liens Open Directory Project (dmoz), construit manuellement par des éditeurs bénévoles.

L'utilisation de terminologies

Le principe de recherche par catégories se place dans la droite ligne des répertoires, tout en offrant un peu plus de souplesse et de puissance, en permettant par exemple d'indexer un document à l'aide de plusieurs catégories. Celles-ci peuvent être organisées de manière plus ou moins complexe : liste de termes, taxonomie, thésaurus voir même ontologie. La définition de ces catégories étant une tâche compliquée, et donc longue et coûteuse, leur structuration n'atteint cependant pas toujours les formes

[12] `http://www.dmoz.org`. Fondé en 1998, en réaction à l'attitude jugée trop peu réactive de Yahoo (`http://www.yahoo.com`), qui proposait également un répertoire de liens depuis 1994. Yahoo a ensuite évolué vers un modèle d'indexation automatique et d'interrogation par mots-clés libres, qui est actuellement le plus courant.

les plus complexes. Une fois la terminologie[13] ou la classification disponible, les documents peuvent y être indexés. Généralement, c'est un documentaliste expert du domaine qui attribue manuellement une ou plusieurs catégories à chacun d'entre eux. Si cette méthode se révèle à nouveau assez onéreuse, elle garantit cependant une indexation d'une qualité assez élevée, et qui est effectuée dans l'espace des concepts et non pas celui des mots (Chaumier et Dejean [2003], Da Sylva [2004], Da Sylva [2006]). En raison de son coût, cette méthode est plutôt utilisée par les entreprises ou grandes organisations, mais peut aussi être appliquée pour de plus petites collections de documents.

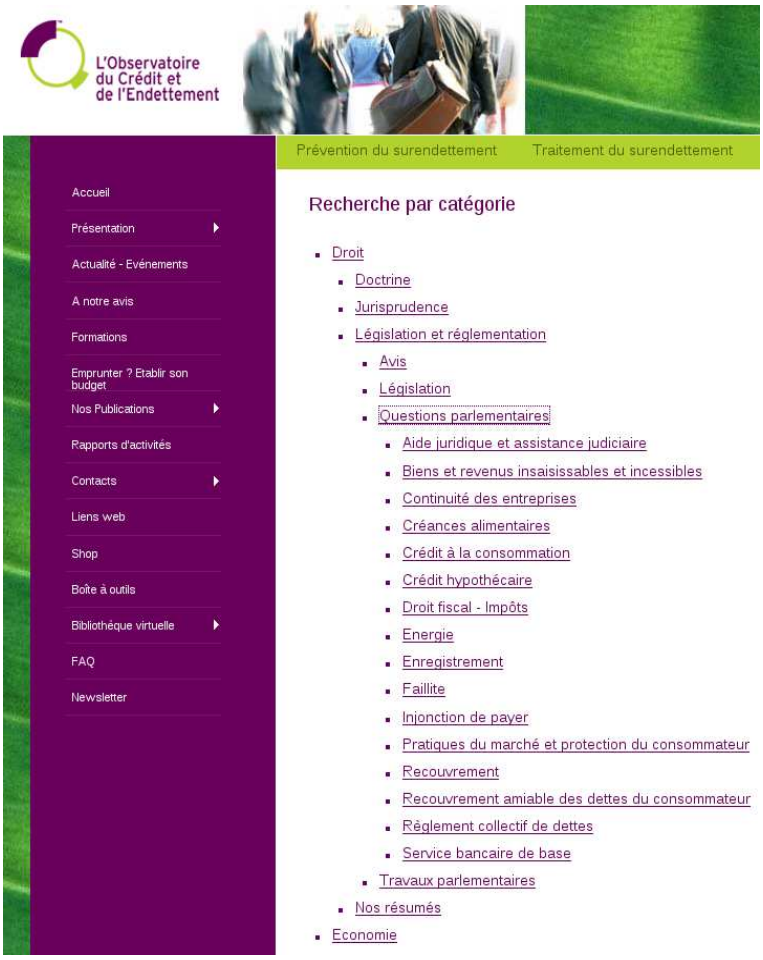


Figure 1.2 : Exemple de recherche par catégories prédéfinies (Site : Observatoire du Crédit et de l'Endettement).

Pour la recherche dans ce type de système (Figure 1.2[14]), l'utilisateur se voit proposer un certain nombre de clés prédéfinies parmi lesquelles il doit faire son choix. Ces catégories peuvent être présentées de manière plate ou hiérarchique. La sélection d'une, ou éventuellement de plusieurs catégories entraîne l'affichage des documents s'étant vus attribuer au moins une de ces catégories lors de la phase d'indexation. Ce mode de recherche peut aussi éventuellement être combiné à une requête par

[13] Le sens donné au mot *terminologie* est ici celui qui désigne une ressource, telle que celles exposées à la section 1.4.1, qui reprennent un ensemble de termes relatifs à un ou plusieurs domaines, activités, etc.

[14] `http://www.observatoire-credit.be` (consulté le 31/07/2010). Voir également le site du Sénat dont la recherche avancée (`http://www.senat.be/www/?MIval=/index_senate&MENUID=12420&LANG=fr`) propose un accès aux documents au travers des catégories du thésaurus Eurovoc (`http://eurovoc.europa.eu`).

mots clés. C'est par exemple le cas de l'interface expérimentale de JSTOR[15], qui permet de chercher des référence bibliographiques par mots-clés (voir section 1.3.3), et d'ensuite réordonner le résultat en pondérant l'importance des différentes catégories (thèmes) concernées par les résultats de la requête (Figure 1.3).

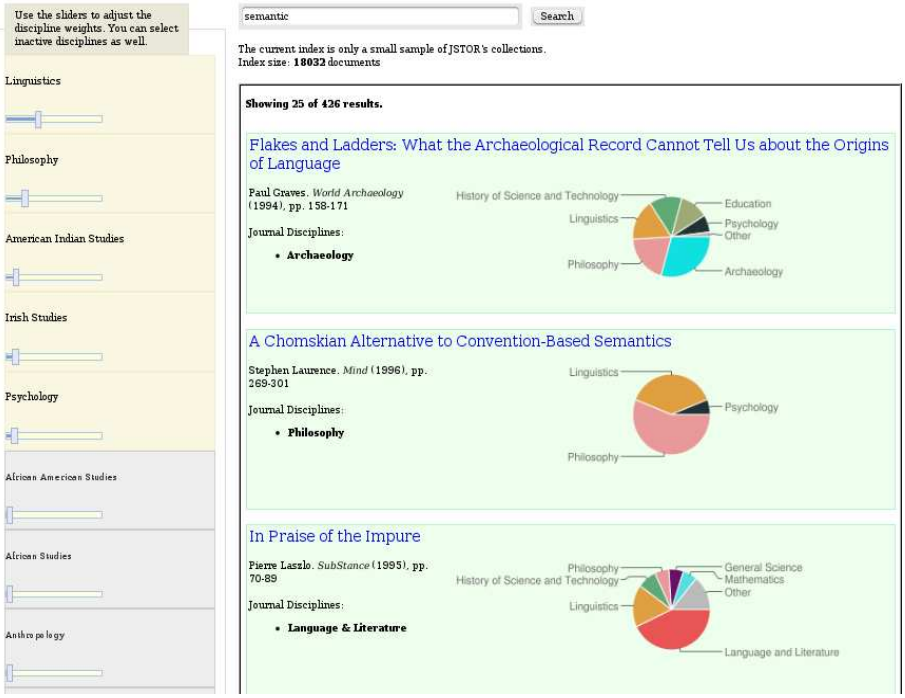


Figure 1.3 : Exemple de recherche par mots-clés et catégories.

La recherche par facettes

Une variation de ce mode d'interrogation est la recherche par facettes. Basé sur le travail initial de Ranganathan [1967], la pertinence de l'application de ce principe dans le cadre des ressources électroniques modernes, telles que le Web, a été démontrée (Zins [2002], Kyung-Sun *et al.* [2006]). Avec la recherche, ou navigation, par facettes, les catégories sont regroupées en plusieurs groupes (les facettes) représentant chacun une caractéristique particulière des documents. L'utilisateur est invité à utiliser plusieurs de celles-ci, de manière simultanée ou successive, afin de raffiner le résultat proposé par le système. Ce type de recherche présente l'avantage de pouvoir plus facilement caractériser des objets complexes, dont la nature peut être définie selon plusieurs axes. Par exemple, le portail *Innovons*[16] (Figure 1.4) a pour vocation d'indexer des documents relatifs à l'*innovation*, sujet qui touche potentiellement à tous les domaines scientifiques et techniques (facette 2), et qui peut s'appliquer à divers secteurs d'activités (facette 3), dans le but de fournir différents types de produits (facette 4), au travers d'un ensemble de métiers et compétences (facette 5). Dans ce cadre, le portail propose d'apporter une réponse à une série de besoins concrets (facette 1).

[15] `http://dbrowser.jstor.org/browser.cgi?q=semantic&btnG=Search`, accédé le 02/12/2010.
[16] `http://www.innovons.be`

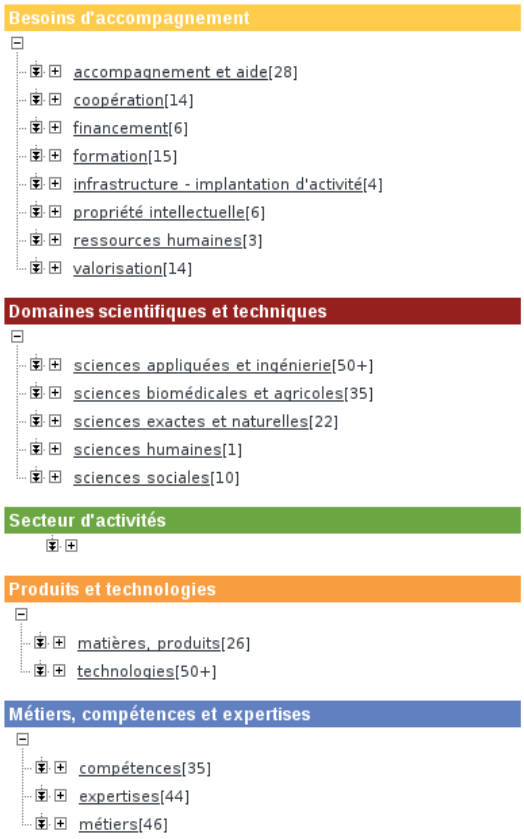


Figure 1.4 : Exemple de catégories organisées par facettes (Site : Innovons en Région wallonne).

1.3.3 Indexation dans un espace ouvert de clés

Recherche par mots clés

Les moteurs de recherche, dont les principaux représentants[17] sont aujourd'hui Google, Yahoo, Baidu et Bing (Microsoft), fonctionnent selon une autre philosophie. Leur principe est d'entretenir, de manière automatique, un index qui exploite principalement le contenu même du document ainsi que certaines métadonnées. Cet index est ensuite confronté aux requêtes formulées sous la forme de mots clés afin de fournir la liste des documents jugés les plus pertinents (Figure 1.5[18]). L'utilisation de ce mode d'interrogation laisse une grande liberté à l'utilisateur : il peut choisir les termes exacts de sa requête et ne doit pas respecter un formalisme spécifique. Il est cependant souvent possible d'utiliser des guillemets pour délimiter des expressions composées, ainsi que certains opérateurs logiques (AND, OR, NOT). En pratique, rares sont les utilisateurs à employer réellement ces possibilités[19].

[17] Selon une statistique établie par Comscore et relayée par le Journal du Net (`http://www.journaldunet.com/cc/03_internetmonde/intermonde_moteurs.shtml`, consulté le 31/07/2010), Google obtiendrait 67,5% des parts de marché au niveau mondial en juillet 2009, contre 7,8% pour Yahoo, 7% pour le moteur chinois Baidu et 2,9% pour Bing.

[18] `http://www.google.be`

[19] L'observation du nombre de mots inclus dans les requêtes montre que celui-ci est assez faible : Assadi et Beaudouin [2002] établissent que trois-quarts des requêtes ont une longueur inférieure ou égale à deux mots (les dernières tendances montrent cependant un allongement progressif des requêtes, selon un rapport Hitwise : `http://image.exct.net/lib/fefc1774726706/d/1/SearchEngines_Jan09.pdf`). Cette longueur peu importante n'invite évidemment pas à l'utilisation d'opérateurs complexes. Jansen et Eastman [2003] rapportent que environ 10% des requêtes utilisent des opérateurs booléens (13% pour Assadi et Beaudouin [2002]), et établissent que leur apport est sou-

Figure 1.5 : Page de résultats fournis par Google suite à une recherche par mots clés.

Dans ce type de système, la qualité du résultat est aussi grandement influencée par l'ordre de présentation des documents. Le moteur se doit, dans la mesure du possible, de proposer en premier lieu le document le plus pertinent, et de continuer ensuite par ordre décroissant d'importance. Par exemple, l'algorithme Page Rank (Brin et Page [1998]) utilisé par Google à cet effet, permet d'exploiter les interconnexions entre les documents afin de faire ressortir les pages les plus *importantes*.

L'avantage déterminant des moteurs de recherche est leur capacité à tenir à jour de manière automatique des index répertoriant un nombre très élevé de documents. Lorsque des liens existent entre ceux-ci, le système est capable de *découvrir* tout seul les nouveaux documents. Au delà de cet aspect, ils rencontrent néanmoins certaines difficultés à appréhender toutes les finesses du langage naturel. En plus des variations grammaticales (singulier/pluriel, forme nominale/forme adjectivale, etc.), qui sont maintenant gérées dans une certaine mesure, la principale difficulté est de pouvoir prendre en compte la nature intrinsèquement variée et ambiguë de la langue. En effet, d'une part, un concept peut souvent être désigné par de nombreux mots ou expressions composées, et d'autre part un mot peut également référer à plusieurs concepts. Cette relation *de plusieurs à plusieurs* entre l'espace des concepts et celui des mots explique pourquoi une requête classique à partir de mots clés ramène rarement l'ensemble des documents pertinents et, dans le même temps, propose souvent des résultats qui ne sont pas en rapport avec la recherche de l'utilisateur. En dépit de cette faiblesse, la technologie de recherche par mots clés est toujours aujourd'hui celle qui est la plus utilisée, aussi bien en entreprise que par le grand public.

Recherches avancées

Les recherches dites *avancées* correspondent à une variation des recherches par mots clés pour lesquelles le système attend de la part de l'utilisateur des clés de recherche qui correspondent à des types d'information définis a priori. On retrouve assez fréquemment ce type d'interrogation sous la forme

vent très faible en ce qui concerne la qualité des résultats, à moins que l'utilisateur ait une connaissance assez pointue du fonctionnement du moteur de recherche.

d'un formulaire proposant des champs spécifiques pour les principales catégories de clés gérées par le système. Pour une bibliothèque, on se verra par exemple proposer un champ pour le nom de l'auteur, un autre pour le titre, un troisième pour la maison d'édition, et ainsi de suite. Cet exemple est parfaitement illustré par le formulaire de recherche proposé par Google Livres[20] (Figure 1.6). Ce type de requête s'applique de préférence à des documents au moins partiellement structurés, ou proposant les métadonnées nécessaires.

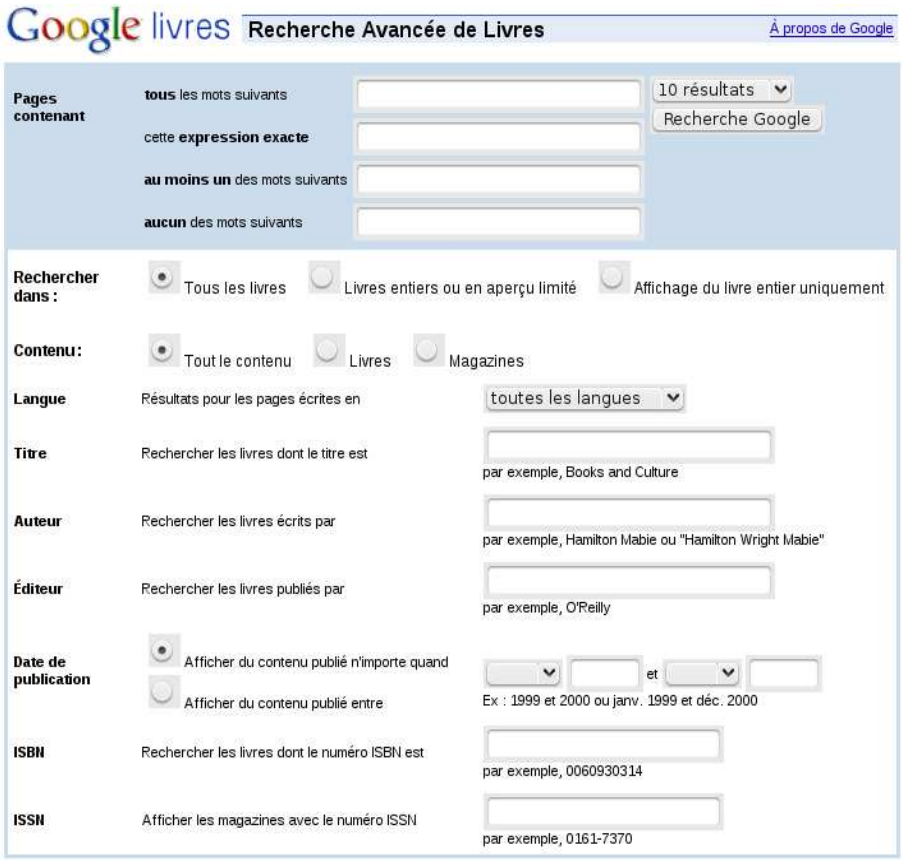


Figure 1.6 : Formulaire de recherche avancée (Google Livres) permettant de contraindre la recherche sur certains types de données.

1.3.4 Les moteurs *sémantiques*

Une nouvelle génération de moteurs, dits *sémantiques*, a fait son apparition depuis quelques années. D'une manière générale, on parle du *Web sémantique* (Berners-Lee et Lassila [2001]), dans le cadre duquel chaque document est accompagné d'un ensemble de données sémantiques qui décrivent son contenu. Cette couche supplémentaire doit permettre à un logiciel d'accéder directement au sens de l'information et non plus à sa matérialisation sous la forme de mots. Cela ouvre évidemment de nombreuses perspectives pour l'amélioration des résultats et l'augmentation de la complexité des recherches. Cependant, la difficulté que représente la production de données correctement annotées et leur exploitation est importante, et cela a pour conséquence que peu d'applications tirent déjà

[20] `http://books.google.be`

complètement parti de tous les aspects du Web sémantique. En pratique, de nombreux systèmes proposent certaines évolutions, qui ne s'inscrivent pas nécessairement dans ce cadre strict, mais qui permettent tout de même d'apporter des éléments de sens aux documents, comblant ainsi en partie le fossé entre l'espace des mots et l'espace des concepts. C'est par exemple le cas des techniques d'extension de requêtes.

Extension de requêtes

La richesse des langues naturelles se traduit par une variété et une ambiguïté du lexique et de son utilisation. En recherche d'informations cela a pour conséquence qu'on observe souvent un écart important entre le contenu lexical des requêtes des utilisateurs et celui des documents (Cui *et al.* [2002]). Les recherches menées sur les possibilités d'extension de requêtes constituent, à cet égard, une démarche intéressante. Le terme *extension* est cependant quelque peu trompeur puisqu'il vise, en réalité, à la fois l'augmentation de la couverture et l'élimination des résultats non pertinents afin d'augmenter la précision. Les techniques utilisées sont assez nombreuses et variées, leur présentation dans le cadre de cette introduction sera donc rapide et forcément incomplète. Ces méthodes nécessitent aussi souvent la résolution de problèmes connexes tels que la désambiguïsation du sens des mots ou la prise en compte de l'aspect multilingue des documents (Gaillard *et al.* [2010]). Ces aspects ne peuvent être considérés comme des tâches triviales et représentent à eux seuls divers défis.

L'idée de l'extension de requête n'est pas neuve puisque Salton et Lesk [1968] montraient déjà que l'usage de synonymes pouvait améliorer les résultats des systèmes de recherche d'informations. Par la suite, des ressources telles que WordNet ont souvent été utilisées afin d'étendre les termes de requêtes avec des termes possédant un sens commun (Voorhees [1994], Moldovan et Mihalcea [2000]).

Des analyses plus complexes peuvent également permettre d'aller plus loin dans l'enrichissement des requêtes. L'exploitation d'informations issues d'ontologies peut par exemple venir compléter la recherche initiale (Bhogal *et al.* [2007], Guelfi *et al.* [2007]). Au delà des synonymes, que nous avons déjà mentionnés, différentes informations peuvent être extraites : hyperonymes, hyponymes, méronymes, ou encore d'autres relations sémantiques (Joho *et al.* [2002]).

L'analyse des logs de requêtes est aussi souvent utilisée pour trouver des termes de recherche associés à la requête initiale. Cui *et al.* [2002] utilisent une méthode consistant en l'extraction, à partir de ces logs, de corrélations probabilistes entre les termes des requêtes contenues dans les logs, et les termes provenant des documents. Les probabilités ainsi obtenues permettent alors d'ajouter les termes qui semblent les plus appropriés à l'extension d'une nouvelle requête.

Une autre idée largement exploitée est la technique de *relevance feedback* (Salton et McGill [1983]) dont le principe est d'utiliser les mots issus des documents les plus pertinents dans la liste de résultats originale. Cette méthode nécessite une participation relativement importante de la part de l'utilisateur. Afin de minimiser celle-ci, il est possible de sélectionner systématiquement les documents les mieux classés dans la liste de départ. Il est évidemment nécessaire que l'algorithme d'ordonnancement des documents donne de bons résultats dès le début, pour que cette approche puisse fonctionner.

Enfin, à l'aide des techniques de *clusering*, il est aussi possible de rassembler les résultats obtenus en plusieurs groupes (*clusters*) représentant diverses interprétations qui ont pu être distinguées (Stefanowski et Weiss [2003]). Cette approche a en particulier été implémentée pour le moteur de recherche Carrot2[21]. La figure 1.7 montre les différents clusters, et les sens correspondants, construits à partir de la recherche « extraction ». On y retrouve entre autres « information extraction », « DNA extraction », ou encore « tooth extraction », qui représentent effectivement des sens assez différents.

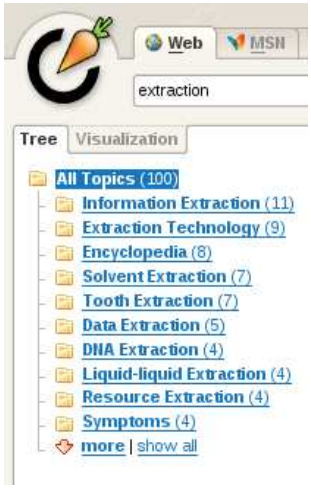


Figure 1.7 : Le moteur de recherche Carrot2 organise les résultats en groupes correspondant aux interprétations qu'il a pu identifier, afin de préciser la recherche initiale (ici, « extraction »).

Toutes ces techniques peuvent être utilisées de différentes manières. La requête initiale peut par exemple être étendue directement afin de présenter dès le départ un résultat *amélioré* à l'utilisateur. Le mécanisme peut aussi s'opérer en deux temps, en faisant intervenir la participation de l'utilisateur suite à l'introduction de sa requête. L'analyse de cette dernière, ou des résultats obtenus par son exécution, est dans ce cas utilisée pour fournir diverses possibilités d'affinement de la requête initiale ou pour orienter l'utilisateur vers de nouvelles recherches, sémantiquement proches.

À titre d'illustration, les propositions émises par le moteur de recherche de Yahoo (Figure 1.8), semblent faire intervenir plusieurs des techniques évoquées. À partir d'une requête « énergie », on remarque dans la première colonne, ce qui ressemble à des mots-clés assez proches (« énergie éolienne », « radio energie », *etc.*), et qui pourraient typiquement être obtenus par analyse de logs de requêtes. Dans les colonnes suivantes, on voit par contre apparaître des expressions plus éloignée au niveau lexical (« électricité ») mais aussi sémantique (« développement durable », « chaleur », « fournisseur de gaz »), ce qui suggère l'intervention de ressources ou procédés sémantiques plus complexes.

[21] `http://search.carrot2.org`

Figure 1.8 : Le moteur de recherche Yahoo fournit des propositions d'affinement de requêtes ainsi que l'orientation vers des recherches sémantiquement proches.

Le web sémantique

La description sémantique du contenu des documents, telle que proposée dans le cadre du Web sémantique (Berners-Lee et Lassila [2001]), a pour objectif de permettre l'élaboration d'applications complexes. Celles-ci exploitent de manière automatique ces données structurées afin de fournir un service particulier, par exemple fournir l'adresse et les heures d'ouvertures d'un restaurant italien, dans une certaine ville, et qui propose un plat bien spécifique. Cela suppose évidemment des possibilités relativement étendues de recherche et d'analyse de l'information. Cette situation semble encore loin d'être atteinte à l'heure actuelle.

L'enrichissement des requêtes en données sémantiques par l'utilisateur lui-même pourrait cependant être considéré comme un premier pas dans cette direction. Cette approche, proposée par Umbrich et Blohm [2008], permet par exemple de préciser, à l'aide d'un balisage XML, que tel mot-clé correspond à un lieu, et que le résultat attendu est une personne. La recherche s'applique à une collection de documents qui a au préalable été indexée à l'aide de Wikipedia[22] et de l'ontologie YAGO[23] (Suchanek *et al.* [2007]). Shah *et al.* [2002] ont une approche relativement similaire, à la différence qu'un mécanisme de traduction transforme automatiquement une requête classique en requête sémantique, exprimée à l'aide du langage DAML+OIL.

La formulation de requêtes en langage naturel, proposée entre autres par Powerset[24], Hakia[25] ou encore Ask[26], suppose également un minimum de prise en compte d'éléments sémantiques. Bien que les résultats présentés soient effectivement différents de ceux obtenus avec un moteur classique (comme à la figure 1.9), il n'apparait pas toujours clairement si le moteur a réellement interprété la question ou s'il l'a juste transformée en un ensemble de mots-clés. D'autre part, il faut aussi souligner que la plupart des moteurs sémantiques se limitent à l'exploitation de ressources particulières, au moins partiellement structurées ou enrichies de données sémantiques, telles que Wikipedia.

Il ne faut pas non plus confondre ce type de moteur avec les systèmes de question-réponse, tels que True Knowledge[27] ou Wolfram[28], qui effectuent de véritables raisonnements et exploitent également des bases de connaissances internes afin de fournir des réponses et non une liste de documents.

[22] `http://www.wikipedia.org`
[23] `http://www.mpi-inf.mpg.de/yago-naga/yago/`
[24] `http://www.powerset.com`
[25] `http://www.hakia.com`
[26] `http://fr.ask.com/`
[27] `http://www.trueknowledge.com`
[28] `http://www.wolframalpha.com`

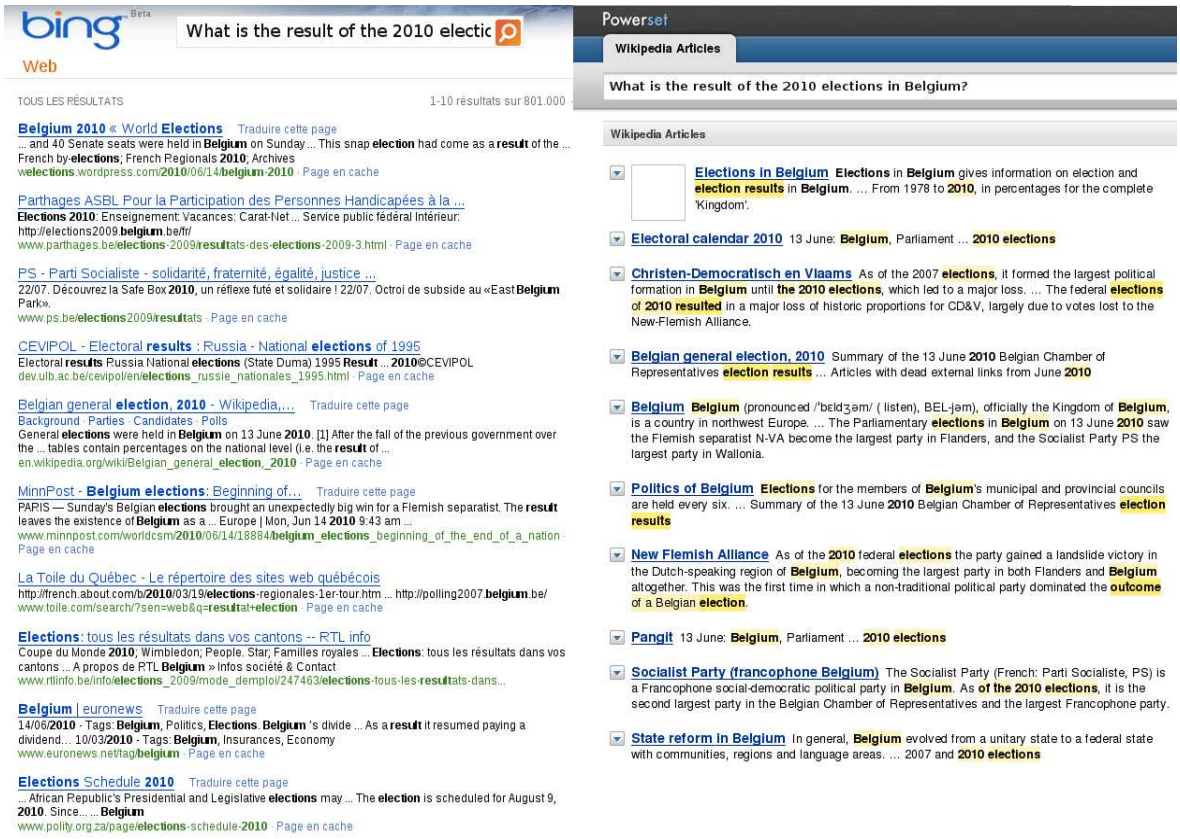


Figure 1.9 : Comparaison des résultats entre le moteur de type mots-clés (bing) et le moteur sémantique (Powerset) de Microsoft.

Au final, si les technologies *sémantiques* poursuivent un objectif légitime, elles semblent avoir quelques difficultés à s'imposer. De nombreux obstacles restent à surmonter en la matière, entre autres la production des documents aux formats et selon les standards définis (ce qui implique un effort important), la question de l'interopérabilité entre ontologies, etc.

1.4 Les systèmes de catégories en tant que couche sémantique pour la recherche d'informations

À l'issue de ce tour d'horizon des technologies de recherche d'informations, il semble clair que le couple indexation *full text* et recherche par *mots-clés*, même s'il reste le système le plus fréquemment utilisé, comporte de nombreuses limitations. Il faut cependant concéder que les espoirs qui ont pu être placés dans les technologies *sémantiques* n'ont apporté pour l'instant que peu de solutions concrètes et à grande échelle.

Comment dès lors dépasser les difficultés liées à la langue naturelle et apporter des éléments de sens lors de l'activité de recherche d'informations ? Un élément de réponse peut être trouvé dans l'utilisation d'un ensemble fermé de clés ou de catégories lors de l'indexation et de la recherche (voir section 1.3.2). Cette approche, si elle n'est pas sans poser certains problèmes (qui seront examinés à la section 1.4.3), permet effectivement d'avoir un certain contrôle sémantique sur l'indexation car celle-ci délaisse l'espace des mots pour s'effectuer au moyen de catégories qui possèdent par défi-

nition un sens particulier. La section 1.4.1 présente un certain nombre de *systèmes terminologiques* qui peuvent servir à définir un ensemble cohérent et organisé de catégories. Quelques exemples de systèmes qui proposent un mode de recherche par catégories sont ensuite passés en revue à la section 1.4.2. Finalement, pour conclure, la section 1.4.3 dresse un bilan des avantages et inconvénients de l'utilisation d'un ensemble fermé de clés, que ce soit pour pour l'indexation ou la recherche, avant de finir sur les perspectives offertes dans ce domaine.

1.4.1 Les systèmes terminologiques

Un ensemble fermé de clés, ou de catégories, nécessaire à l'indexation peut être organisé de diverses façons plus ou moins complexes à l'intérieur d'une terminologie[29].

Vocabulaire contrôlé

L'appellation *vocabulaire contrôlé* constitue la désignation générale de tout ensemble de termes, définis et sélectionnés par un ensemble d'experts. Un tel vocabulaire constitue donc un sous-ensemble du vocabulaire complet d'une langue (de plusieurs langues lorsqu'il s'agit d'une ressource multilingue). Il est généralement mis au point de manière à couvrir et à décrire un ou plusieurs domaines particuliers. Son utilité est de permettre l'organisation des connaissances à des fins de recherche d'informations.

Les termes qui constituent le vocabulaire contrôlé peuvent constituer une simple liste (par exemple le vocabulaire RAMEAU[30]) ou être organisés de diverses manières : taxonomie, thésaurus ou ontologie.

Taxonomie

La taxonomie est une forme assez simple de vocabulaire contrôlé. Elle consiste à organiser les termes à l'aide de relations hiérarchiques. Les taxonomies sont souvent utilisées dans le domaine des sciences de la nature, pour classer les différentes espèces animales et végétales. Par exemple, une taxonomie des virus a été mise au point par l'ICTV[31].

Thésaurus

Un thésaurus est un vocabulaire contrôlé un peu plus complexe. Les termes qu'il regroupe, et qui permettent de définir et de décrire les concepts d'un certain domaine utilisés par un groupe de personnes, sont liés de manière hiérarchique et transversale.

[29] Une rapide introduction peut être obtenue dans l'article de Woody Pidcock, « What are the differences between a vocabulary, a taxonomy, a thesaurus, an ontology, and a meta-model ? » (`http://www.metamodel.com/article.php?story=20030115211223271`, date de dernière consultation : 13/08/2010).

[30] *Répertoire d'autorité-matière encyclopédique et alphabétique unifié*, langage documentaire élaboré et utilisé, entre autres, par la Bibliothèque nationale de France. `http://rameau.bnf.fr`

[31] *International Committee For Taxonomy Of Viruses*, `http://www.ictvonline.org/index.asp?bhcp=1`.

Un concept est représenté par un terme principal appelé *descripteur* qui peut être relié à plusieurs *non descripteurs* ou *synonymes* par une relation *used-for* (UF). Les concepts sont organisés hiérarchiquement à l'aide des relations *broader-than* (BT) et *narrower-than* (NT). La relation *related-term* (RT) permet de définir un lien de similarité entre deux concepts. Les grands thésaurus peuvent être fragmentés en *microthésaurus* qui couvrent chacun un sous-thème particulier. Plusieurs normes internationales dont ISO [1986] et AFNOR [1981] définissent plus précisément les thésaurus.

La portée d'un thésaurus peut être très large, par exemple Eurovoc[32], ou au contraire très spécialisée, tel que Agrovoc[33]. Ces deux thésaurus comptent un grand nombre de niveaux hiérarchiques et de descripteurs[34]. De nombreuses organisations se contentent cependant de vocabulaires de taille plus modeste. Van Slype [1987] préconise l'usage de 500 à 1.500 descripteurs pour des bases de données ayant un accroissement de 10.000 documents par an et de 3.000 à 6.000 descripteurs si la base s'étend jusqu'à 100.000 documents par an.

Ontologie

L'ontologie, concept bien connu dans le domaine du web sémantique, est définie par Gruber [1993] comme « an explicit and formal spécification of a conceptualization ». La construction d'une ontologie revient donc à exprimer de manière formelle la perception que l'on a d'un domaine. Même si elle en reprend de nombreux aspects, l'ontologie déborde largement du champ d'action des *simples* terminologies.

Dans une ontologie, les concepts s'organisent en classes et disposent de propriétés. Celles-ci se rapportent à une classe ou à un type de données particulier. Cela signifie qu'il est possible de définir à peu près n'importe quel type de lien entre les différentes classes, y compris la relation hiérarchique qui est souvent nommée *is a*. Une définition logique accompagne généralement les classes et les relations, de manière à en fournir une spécification formelle, mais aussi un moyen de raisonner sur l'univers ainsi construit. Certains concepts simples ne peuvent pas être définis formellement et constituent les éléments de base de l'ontologie.

En plus des connaissances structurelles, une ontologie peut contenir des connaissances factuelles ou assertionnelles. Ces données actualisent les classes définies dans l'ontologie et sont aussi parfois appelées *instances*. Le peuplement d'instances dans le *schéma* que définit l'ontologie donne naissance à une base de connaissances. En plus des données explicites qui y ont été déposées, il est possible de déduire de la connaissance implicite à l'aide de logiciels de raisonnement. Par exemple, à partir des informations « Pierre est le fils de Jean » et « Paul est le fils de Jean », le *raisonneur* pourra déduire que Pierre et Paul sont frères.

Comme le décrit Guarino [1998], on peut distinguer plusieurs types d'ontologies. La plus générale

[32] Thésaurus du Parlement de la Communauté européenne, couvre une grande diversité de domaines, mais toujours en rapport avec le travail parlementaire : `http://europa.eu/eurovoc/`

[33] Thésaurus de l'Organisation des Nations Unies pour l'alimentation et l'agriculture, se concentre sur l'agriculture : `http://www.fao.org/aims/ag_intro.htm`

[34] Eurovoc : 6.645 pour chaque langue ; Agrovoc : 28.718 en anglais uniquement.

est appelée ontologie de haut niveau (*top-level ontology*) et décrit des concepts très généraux, indépendants d'un quelconque problème particulier. Un exemple est l'ontologie SUMO[35] (Niles et Pease [2001]). L'ontologie de domaine (*domain ontology*) et l'ontologie de tâche (*task ontology*) spécialisent toutes deux les concepts de l'ontologie de haut niveau. La première le fait pour décrire le vocabulaire d'un domaine générique (la musique[36], le domaine médical[37], etc.), alors que la seconde s'attachera à la définition du vocabulaire relié à une tâche ou à une activité générique (la pose de diagnostic, la vente, ou encore l'apprentissage assisté par ordinateur, comme chez Ikeda *et al.* [1997]). Enfin, les ontologies d'application (*application ontology*) sont les plus spécifiques. Elles définissent des concepts qui correspondent à des rôles pris dans le cadre d'une certaine activité par des entités d'un certain domaine. Il s'agit donc d'une spécialisation des types d'ontologies de tâche et d'activité.

1.4.2 Cas concrets d'utilisation de ressources terminologiques pour l'indexation et la recherche de documents

Dans cette section nous allons passer en revue quelques exemples qui montrent dans quels contextes sont utilisés les terminologies à des fins d'indexation et de recherche d'informations. D'une manière générale, les terminologies employées sont plutôt des thésaurus, voir des taxonomies, étant donné d'une part le coût élevé de création de structures plus complexes, et d'autre part la valeur ajoutée relative de celles-ci pour la recherche.

Le thésaurus Eurovoc[38] a été développé par la Commission européenne dans le but d'indexer les documents issus du travail parlementaire européen. Il est également utilisé par d'autres organisations telles que, en Belgique, la Chambre des députés et le Sénat. La figure 1.10, montre, pour le site du Sénat[39] les possibilités de recherche offertes suite à l'utilisation d'Eurovoc pour l'indexation. Une requête permet à la fois de restreindre l'ensemble de documents par rapport au thème à l'aide des descripteurs Eurovoc, mais aussi d'exprimer des contraintes par rapport au type de document, à sa date de publication, à ses auteurs ou encore par rapport aux mots contenus dans les titres et sous-titres.

Un deuxième exemple concerne la recherche de littérature scientifique dans le domaine des sciences de la vie sur le portail PubMed[40]. Celui-ci permet la recherche au moyen de termes issus de MeSH[41], comme illustré à la figure 1.11. La recherche se déroule en plusieurs temps : l'utilisateur introduit un mot-clé pour trouver le descripteur MeSH qu'il recherche, il le sélectionne et lance ensuite la véritable recherche de documents à partir de celui-ci. Il est possible de combiner les descripteurs à l'aide des opérateurs logiques habituels.

[35] Suggested Upper Merged Ontology (`http://www.ontologyportal.org`).
[36] Music Ontology, `http://musicontology.com`.
[37] `http://bioportal.bioontology.org`
[38] `http://eurovoc.europa.eu`
[39] `http://www.senate.be`
[40] `http://www.ncbi.nlm.nih.gov/pubmed/`
[41] Medical Subject Headings (`http://www.nlm.nih.gov/mesh/`).

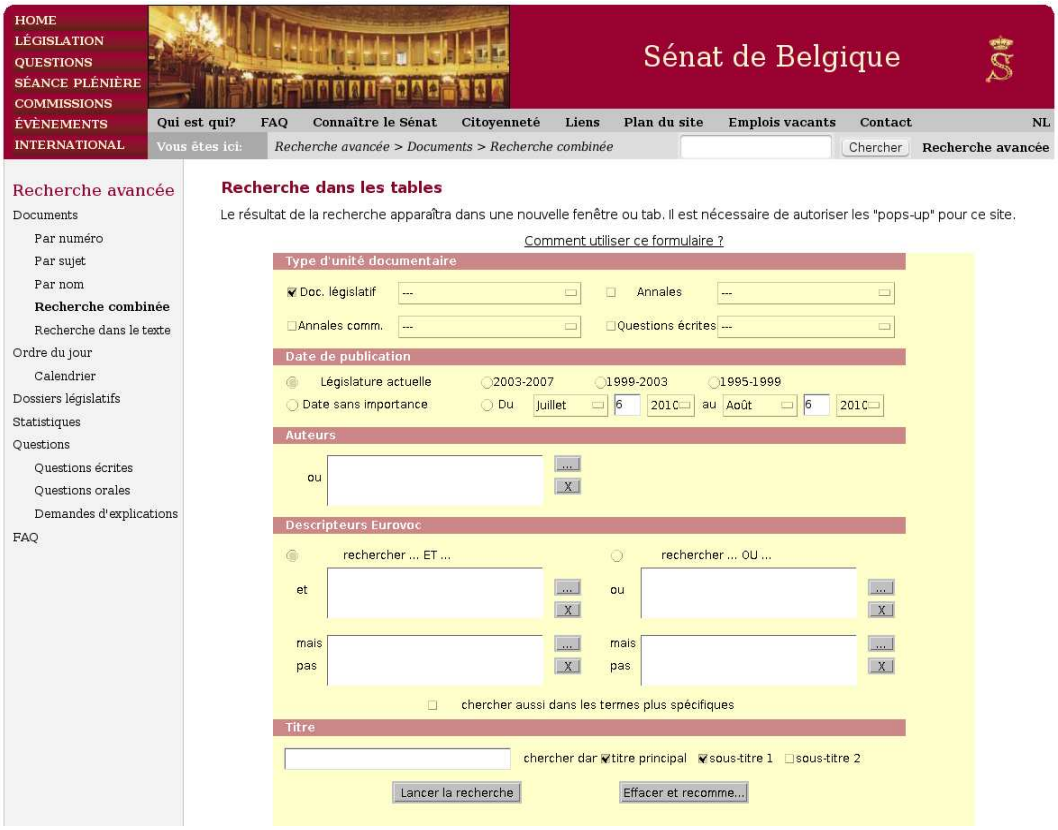


Figure 1.10 : Le formulaire de recherche avancée du Sénat permet l'utilisation des descripteurs d'Eurovoc.

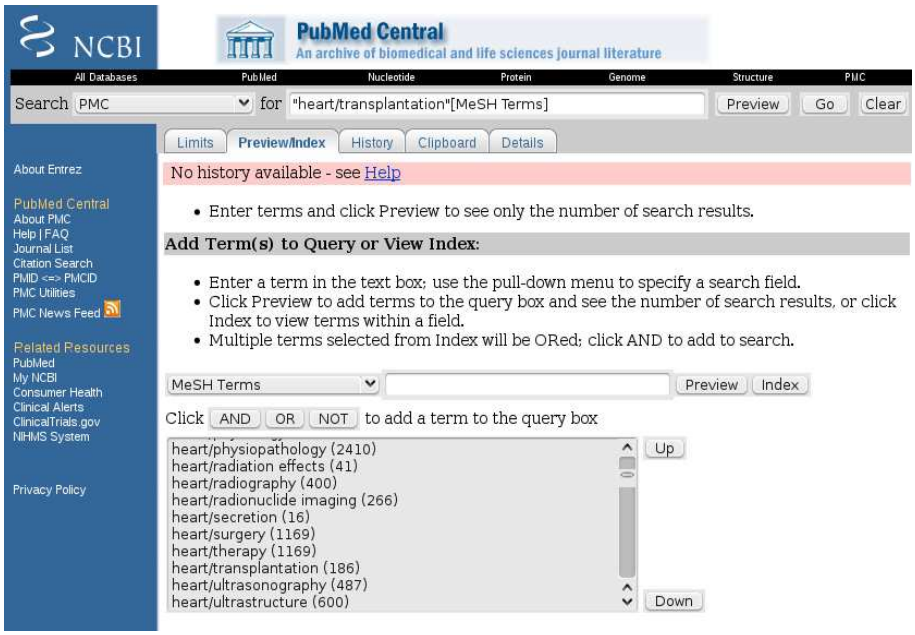


Figure 1.11 : Utilisation de MeSH dans le formulaire de recherche avancée du Pubmed.

1.4.3 Avantages, inconvénients et perspectives

L'utilisation d'un terminologie lors de l'indexation est très souvent mise en œuvre au sein d'un processus manuel. Ce sont alors des documentalistes, généralement experts d'un domaine particulier, qui analysent les textes avant de leur attribuer une ou plusieurs catégories.

Du côté de la recherche, la requête est définie par le choix d'une ou plusieurs catégories, ou par la navigation dans la structure qui les organise. Il est cependant parfois possible de faire appel à une recherche par mots clés qui détermine alors les catégories pertinentes pour cette requête. Le système d'interrogation par mots-clés pourra également être exploité pour effectuer une recherche classique sur le texte complet, tout en utilisant le système de catégories comme *filtre*. Ce dernier restreint alors la collection de documents à un sous-ensemble qui correspond à une certaine catégorie.

Le principal avantage d'une indexation au moyen d'une terminologie est qu'elle permet de contrôler l'espace d'indexation, et par là même d'y apporter un sens précis. La qualité de cette indexation, généralement manuelle, est bien connue dans les milieux documentaires (Chaumier et Dejean [2003], Da Sylva [2004]) et constitue une solution qui est implémentée dans les entreprises et autres organisations.

Cet apport sémantique de qualité est également présent lors de la recherche. En effet, chaque descripteur, ou catégorie, possède un sens et représente un concept bien précis dans le contexte thématique d'une terminologie en particulier. Une recherche sur le descripteur « carotte » n'aura pas le même sens lorsqu'il est question de légumes ou lorsque le sujet est la géologie. Dans un système utilisant une indexation par rapport à une terminologie, la recherche ne sera pas ambiguë car elle ne s'effectue pas sur le mot, mais sur un concept. Le terme est en fait implicitement désambiguïsé par les concepts plus généraux qui y sont hiérarchiquement reliés (par exemple, « légume » ou « géologie ») ou plus simplement par le fait qu'une seule interprétation existe au sein du système terminologique. Dans le cas d'une recherche par mots-clés, l'ambiguïté subsiste à tout moment.

Diverses difficultés viennent cependant tempérer ces avantages, souvent au profit des solutions d'indexation *full text* et des modes de recherche par *mots-clés* (voir la comparaison entre l'indexation humaine et l'indexation automatique, réalisée par Chaumier et Dejean [2003]). En ce qui concerne l'indexation, les coûts engendrés par l'utilisation des terminologies ne sont pas négligeables, tant en ce qui concerne la création même de la ressource que son utilisation. En effet, les documentalistes humains, experts du domaine, qui prennent souvent en charge l'attribution des descripteurs aux documents, représentent une charge financière importante. L'indexation manuelle constitue également un processus relativement lent. De plus, si cette solution apporte effectivement une bonne qualité à l'indexation, elle introduit aussi paradoxalement des problèmes de cohérence, soit au cours du temps, soit entre les différents annotateurs. Van Slype [1987] montre que la cohérence de l'indexation d'un même document par deux documentalistes se situe entre 50% et 80%. De même, Pouliquen *et al.* [2003] rapportent un *accord inter-annotateur* allant de 78% à 87%.

En ce qui concerne la recherche, des restrictions peuvent également être émises. Les interfaces d'interrogation, telles que celles qui ont été présentées à la section 1.4.2, ne sont souvent pas très satisfaisantes pour l'utilisateur. En plus de leur côté peu ergonomique, l'inconvénient principal réside, pour l'utilisateur, en sa connaissance généralement très approximative du système terminologique utilisé pour l'indexation :

> « Il est évident que dans nos dispositifs documentaires actuels, l'utilisateur final qui n'indexe pas et qui n'a pas construit le thésaurus, non professionnel de la documenta-

> tion et parfois non spécialiste du domaine, affronte en réalité une tâche beaucoup plus complexe qu'un documentaliste » (Dalbin [2007], p. 45)

L'effort nécessaire à une bonne connaissance d'une telle ressource est important et de nature à décourager de nombreux utilisateurs :

> « Mais la plupart des utilisateurs souhaitent passer outre cette phase complexe de formulation d'une requête à partir de la sélection de termes dans des vocabulaires contrôlés, préférant porter leur attention sur la fouille du lot de résultat. » (Dalbin [2007], p. 48)

Le problème provient donc du décalage qui existe, dans la maîtrise de la ressource terminologique, entre les documentalistes et les utilisateurs.

Les avantages offerts par une indexation relative à une terminologie se heurtent donc aux obstacles de coût et de lenteur de la tâche en ce qui concerne l'indexation, ainsi qu'à l'inadéquation des modes de recherche lors de l'interrogation des bases documentaires. Il existe cependant des perspectives qui rendent ce choix possible. Pour l'indexation, il s'agit principalement de mettre en place des méthodes automatiques ou semi-automatiques, selon le degré de contrôle que l'on désire conserver. Le chapitre 2 propose à cet égard un processus d'indexation semi-automatique qui utilise le thésaurus comme base d'un processus de classification dans lequel chaque descripteur, identifié par son code et accompagné par ses synonymes, constitue une classe. Pour ce qui concerne la recherche, il est possible d'exploiter l'indexation de manière beaucoup plus souple que par une navigation dans une hiérarchie de catégories. Une solution performante ne peut cependant être atteinte qu'en combinant plusieurs des techniques *sémantiques* présentées à la section 1.3.4. Par exemple, à partir d'une requête par mots-clés :

- extension de requête afin de maximiser la couverture (rappel) ;
- effectuer un regroupement de ces résultats étendus selon les catégories définies par la terminologie ;
- présenter les catégories les plus *pertinentes* afin que l'utilisateur précise sa requête (mécanisme de *relevance feedback*) ;
- le choix exprimé par l'utilisateur, permet de préciser le sens de la requête initiale, et ainsi d'atteindre un résultat sémantiquement plus proche de ses attentes (précision).

La réalisation d'un tel système n'est cependant pas effectué dans le cadre de cette thèse. Les technologies suggérées pour cette solution de recherche ont cependant prouvé leur efficacité. Leur combinaison doit permettre d'atteindre un résultat en rapport avec les exigences des utilisateurs, et encourage par conséquent le développement de systèmes (semi-)automatiques d'indexation guidés par des terminologies.

CHAPITRE 2

INDEXATION SEMI-AUTOMATIQUE, UNE APPROCHE SYMBOLIQUE DE CLASSIFICATION DE TEXTES

2.1 Introduction

Face aux défis rencontrés en recherche d'informations pour apparier au mieux les requêtes des utilisateurs aux documents d'une collection, l'utilisation d'un ensemble fermé de clés constitue une solution permettant de faire un pas de plus, par rapport aux simples mots-clés, sur la route de la *recherche sémantique*. Par ce terme, nous entendons qu'il y a un passage d'un mode de recherche basé sur des éléments lexicaux, mots simples ou expressions composées, à une recherche qui s'effectue dans l'espace des concepts, c'est-à-dire en tenant compte du sens de ces mots et expressions. Les clés définies dans un cadre particulier comportent en effet l'avantage de faire référence à des sens bien précis pour les mots ou expressions qui seraient ambigus autrement[1]. La désambiguïsation du lexique est assurée à la fois par le contexte d'utilisation (par exemple une entreprise minière) et par la définition même de la ressource terminologique[2]. L'organisation des clés peut être réalisée sous différentes formes (Section 1.4.1). Celles-ci font souvent intervenir un lien hiérarchique apte à caractériser et désambiguïser le sens de deux formes homographes.

L'indexation des documents est souvent réalisée par des documentalistes experts du domaine, ce qui garantit un traitement de bonne qualité, mais n'est pas sans poser certains problèmes relatifs au coût et à la cohérence à long terme. Le traitement de volumes de données importants est aussi un point critique pour cette méthode dont le passage à une grande échelle (*scalability*) représente un obstacle. Nous proposons une méthode dont la vocation est d'être utilisée de manière semi-automatique, c'est-à-dire comme une aide au codage. Son but est d'améliorer l'efficacité et l'homogénéité de l'indexation manuelle en suggérant au documentaliste une liste de catégories, ou de mots-clés potentiels, automatiquement construite.

La méthode que nous décrivons adopte une approche dite *symbolique*, par opposition aux méthodes *statistiques*. Ces dernières, basées sur un apprentissage artificiel, nécessitent de grandes quantités de

[1] Citons à nouveau l'exemple du mot « carotte », qui a deux sens différents selon que l'on se situe dans un contexte géologique ou végétal.

[2] Le sens donné au mot *terminologie* est ici celui qui désigne une ressource, telle que celles exposées à la section 1.4.1, qui reprennent un ensemble de termes relatifs à un ou plusieurs domaines, activités, etc.

données annotées. Cela peut représenter un effort important et poser des problèmes lorsque peu de données d'entraînement sont disponibles, par exemple pour certains codes spécifiques et rares. De son côté, si l'approche symbolique donne souvent de bons résultats et est appréciée pour la précision qu'elle apporte, elle présente aussi l'inconvénient de la quantité de travail nécessaire à l'élaboration de ressources *ad-hoc*. La méthode utilisée ici vise à réduire autant que faire se peut cette quantité de travail en automatisant une partie de la construction des ressources.

En ce qui concerne les approches *symboliques*, il en existe une grande variété, qui font intervenir différents types d'analyses linguistiques. Une méthode peut se contenter d'exploiter de larges ressources (dictionnaires, thésaurus, ontologies, etc.) ou au contraire tenter de plonger plus profondément dans la structure du texte, en mettant par exemple en œuvre une analyse syntaxique. Notre système repose à la fois sur l'exploitation d'une ressource de base, en l'occurrence un thésaurus, et sur une analyse locale qui fait intervenir différentes techniques de traitement automatique du langage.

2.1.1 Principes et hypothèses

L'approche adoptée est centrée sur l'utilisation d'une ressource de base qui définit les catégories, par exemple une terminologie telle que celles présentées à la section 1.4.1. Cette ressource constitue le point de départ de la génération automatique d'automates de reconnaissance, ou plus précisément de transducteurs. Ceux-ci permettent d'exprimer des motifs lexicaux qui peuvent faire appel à des dictionnaires électroniques, poser des contraintes sur les catégories grammaticales ou sur la morphologie des mots, et ainsi atteindre une assez grande complexité. L'application des transducteurs à un texte a pour résultat d'extraire un nombre limité d'expressions *pertinentes*. La reconnaissance de chacune de celles-ci provoque la génération d'un code de catégorie associé. L'analyse des termes extraits et de leurs codes permet finalement la classification du document. Nous avons appelé cette méthode *motifs lexicaux étendus* (MLE).

L'intervention humaine dans le processus d'indexation se résume alors à la consultation sommaire du document – le titre, le résumé, éventuellement le premier paragraphe, ou encore tout autre partie du document jugée importante – et à la sélection d'une ou plusieurs catégories dans la liste proposée, et non plus dans la terminologie complète. Afin de préserver le gain de temps obtenu par cette analyse automatique, il est évidemment très important que le documentaliste n'ait à consulter ni le texte en entier, ni la terminologie pour confirmer le choix des catégories. Par conséquent, la liste doit contenir un maximum de mots clés plausibles, quitte à y inclure certaines propositions qui ne seront finalement pas sélectionnées. En ce sens, nous essayons de maximiser le rappel, et ce, dans une certaine mesure, au détriment de la précision.

L'appartenance d'un texte à une catégorie thématique se matérialise souvent, au niveau du document, par l'utilisation d'un certain nombre de mots et d'expressions spécifiques. Cette idée, que l'on retrouve par exemple dans le cas d'un *sous-langage*, avec l'utilisation spécifique de la langue dans un certain contexte (Kittredge et Lehrberger [1982]), est ici transposée aux catégories thématiques, c'est-à-dire à l'échelle des concepts. Dès lors, une ressource terminologique adéquate – c'est-à-dire

qui couvre réellement le champ thématique du type de corpus destiné à être indexé, en tenant compte un maximum de la richesse lexicale[3] – s'apparente à un ensemble de *définitions* relatives à chaque concept. Il est alors possible de trouver une intersection suffisante entre le vocabulaire du document et cette ressource pour décider automatiquement de l'attribution des catégories.

La prise en compte des expressions composées est également un principe important. En effet, on peut constater que celles-ci véhiculent généralement un sens très précis et constituent dès lors de bons candidats en tant que concept descripteur d'un document. Par exemple, en français, le terme « allocations » est parfois utilisé seul, mais il est aussi très souvent rencontré dans des formes composées telles que « allocations de chômage » ou « allocations familiale », qui proposent des sens bien plus précis. Le corollaire de cette constatation est que les expressions composées sont souvent moins polysémiques, comme l'a montré Yarowsky [1993]. L'intégration de ces unités polylexicales permet de dépasser une première limitation induite par l'utilisation de mots-clés simples.

La méthode s'applique sur l'ensemble du texte, mais certaines parties peuvent être privilégiées. C'est le cas du titre qui, pour de nombreux documents, se révèle particulièrement informatif. Ce principe est également souligné par la norme AFNOR Z 47-102 (AFNOR [1993]) et est bien connue dans le milieu du journalisme au travers de la notion de *chapeau*.

Après un état de l'art (Section 2.2), les parties suivantes (Sections 2.3 et 2.4) présentent en détail la méthode semi-automatique développée. Celle-ci propose d'améliorer la rapidité et la cohérence par rapport à l'indexation strictement manuelle grâce à l'analyse automatique. Ce gain est obtenu tout en conservant une précision élevée, garantie par la validation humaine. L'approche, sans apprentissage, ne nécessite pas de données annotées manuellement[4] et est donc fonctionnelle dès le premier document. L'analyse des textes est basée sur des principes simples et performants qui font usage d'une ressource lexicale et sémantique telle qu'un thésaurus.

2.2 État de l'art

L'attribution à un document de clés d'indexation issues d'un vocabulaire contrôlé est comparable au processus de classification de textes. Ce domaine est largement couvert par les techniques d'apprentissage artificiel. Bien que notre méthode n'en fasse pas partie, nous allons brièvement les présenter afin de pouvoir nous situer par rapport à ces travaux.

2.2.1 Apprentissage artificiel

La classification de textes peut être divisée en deux activités distinctes : le clustering et la catégorisation. Le *clustering* regroupe les documents en ensembles de textes similaires sur la seule base des documents. La *catégorisation* s'appuie par contre sur la définition a priori des catégories à attribuer.

[3] Pour un thésaurus, il serait par exemple souhaitable d'avoir, en plus d'un terme descripteur principal, autant de non-descripteurs qu'il existe de synonymes. Il s'agit évidemment là de la situation idéale espérée, et non d'une condition nécessaire à l'application de la méthode.

[4] C'est-à-dire pas de corpus d'apprentissage.

Le clustering et la catégorisation sont aussi qualifiés respectivement de classification non supervisée et supervisée. Notre système s'apparente à la classification supervisée que nous appellerons désormais indifféremment classification ou catégorisation[5]. Dans notre cas, le résultat attendu est une liste de catégories accompagnées de poids, ce qui est généralement qualifié de classification *soft*[6] et *multi labels*[7].

Comme nous l'avons déjà mentionné, la recherche d'informations, et plus particulièrement la classification automatique, est un domaine dans lequel les techniques d'apprentissage artificiel sont souvent appliquées. Les documents traités sont représentés au moyen d'un ensemble de caractéristiques définies au préalable. Par exemple, celles-ci peuvent être constituées de l'ensemble des termes simples du texte. Le document est alors caractérisé au travers de son lexique. Les techniques d'apprentissage artificiel visent à acquérir automatiquement pour chaque classe, sur la base d'une collection de documents annotés, des valeurs pour ces caractéristiques. Différents modèles de représentation des caractéristiques existent, dont les principaux sont les modèles booléens, probabilistes ou encore vectoriels. Ceux-ci sont abordés de manière très brève ci-dessous[8] [9].

Les modèles booléens prennent simplement en compte la présence ou l'absence des caractéristiques. Il s'agit de comparer, pour chaque classe, les valeurs prototypiques de l'ensemble des caractéristiques avec celles obtenues pour tout nouveau document. Pour qu'une classe soit attribuée à un document, la correspondance de leurs représentations doit être exacte.

Soit t_i une caractéristique pouvant prendre une valeur binaire (0 | 1),
$\underline{Classe_1}$: $t_1 = 1 \wedge t_2 = 0 \wedge t_3 = 1 \wedge t_4 = 0$
$\underline{Classe_2}$: $t_1 = 0 \wedge t_2 = 1 \wedge t_3 = 0 \wedge t_4 = 1$
$\underline{Classe_3}$: $t_1 = 1 \wedge t_2 = 1 \wedge t_3 = 0 \wedge t_4 = 0$
$\underline{Document_X}$: $t_1 = 0 \wedge t_2 = 1 \wedge t_3 = 0 \wedge t_4 = 1$
$\underline{Classification}$: $Classe_1 =$ **NON**, $Classe_2 =$ **OUI**, $Classe_3 =$ **NON**

Les modèles probabilistes tentent d'obtenir la probabilité qu'un nouveau texte, représenté au travers des caractéristiques, soit pertinent par rapport à une certaine classe. Cette valeur est calculée au moyen des probabilités disponibles pour chaque caractéristique vis à vis des différentes classes. Ces probabilités ont été produites lors de la phase d'apprentissage. Les modèles *bayésiens naïfs* constituent un exemple très populaire de modèles probabilistes car ils sont simples à mettre en œuvre et donnent généralement d'assez bons résultats. À ce titre, ils sont souvent choisis comme valeurs de référence (*baseline*). Soit $P(t_i|C_j)$ la probabilité d'avoir la caractéristique t_i dans un document appartenant à la classe C_j, et $P(C_Y)$, la probabilité d'attribution d'une classe Y, la probabilité d'attribution de cette classe à un document $P(C_Y|Document_X)$ est obtenue par[10] :

[5] De même, dans le cadre qui nous occupe, les dénominations de *classe* et *catégorie* recouvrent la même notion, qui correspond d'ailleurs aussi à l'appellation *clé d'indexation*.

[6] À l'inverse, la classification *hard* donnerait une valeur binaire et non un poids ou une probabilité.

[7] Dans le cas d'une approche *mono label* par contre, une seule catégorie peut être attribuée.

[8] Ces techniques n'étant pas réellement exploitées dans ce travail, elles ne sont par conséquent pas expliquées en détail.

[9] Les exemples donnés ci-dessous pour illustration n'ont que la vocation de susciter l'intuition des différences entre les modèles, et non d'exposer de manière rigoureuse le fonctionnement de chaque technique.

[10] Cette formule est dérivée du théorème de Bayes, que nous n'exposons pas en détail.

$P(C_Y|Document_X) = P(C_Y) * \prod_{i=1}^{n} P(t_i|C_Y)$.

$\underline{Classe_1}$: $P(t_1|C_1) = 0,5; P(t_2|C_1) = 0,05; P(t_3|C_1) = 0,7; P(t_4|C_1) = 0,05$
$\underline{Classe_2}$: $P(t_1|C_2) = 0,05; P(t_2|C_2) = 0,3; P(t_3|C_2) = 0,05; P(t_4|C_2) = 0,8$
$\underline{Classe_3}$: $P(t_1|C_3) = 0,2; P(t_2|C_3) = 1; P(t_3|C_3) = 0,05; P(t_4|C_3) = 0,05$
$\underline{Document_X}$: les caractéristiques t_1, t_2 et t_4 sont présentes
$\underline{Classification}$: $P(C_Y)$ est ici supposée équivalente pour toutes les classes (0,33),
$P(C_1|Document_X) = 0,33 \times 0,5 \times 0,05 \times 0,05 =$ **0,0004**
$P(C_2|Document_X) = 0,33 \times 0,05 \times 0,3 \times 0,8 =$ **0,0040**
$P(C_3|Document_X) = 0,33 \times 0,2 \times 1 \times 0,05 =$ **0,0033**

Les modèles vectoriels représentent les documents sous la forme d'un vecteur dont les éléments sont des poids associés aux différentes caractéristiques[11]. À chaque classe est lié un vecteur de référence. L'appartenance d'un document à une classe est calculée au moyen de valeurs de similarité. Celles-ci sont produites en comparant les vecteurs de référence de chaque classe à celui calculé pour tout nouveau document. Il s'agit de mesurer l'angle entre ces deux vecteurs, ce qui peut, par exemple, être réalisé à l'aide d'un produit scalaire ou d'un cosinus. Le résultat final est une liste ordonnée des classes selon leur *pertinence*. SVM (*Support Vector Machines*, Cortes et Vapnik [1995]) est un des modèles vectoriels parmi les plus performants et les plus utilisés.

Soit t_i, une caractéristique dont la valeur est un certain poids,
$\underline{Classe_1} = (t_1 = 0,5; t_2 = 0; t_3 = 0,7; t_4 = 0)$
$\underline{Classe_2} = (t_1 = 0; t_2 = 0,3; t_3 = 0; t_4 = 0,8)$
$\underline{Classe_3} = (t_1 = 0,2; t_2 = 1; t_3 = 0; t_4 = 0)$
$\underline{Document_X} = (t_1 = 0,1; t_2 = 0,4; t_3 = 0; t_4 = 0,7)$
$\underline{Classification}$: La mesure de similarité est obtenue par un produit scalaire, $\overrightarrow{D_X} \cdot \overrightarrow{C_Y} = \sum_{i=1}^{n} t_{iD_X} \times t_{iC_Y}$:
$Sim(Document_X, C1) = 0,5 \times 0,1 + 0 \times 0,4 + 0,7 \times 0 + 0 \times 0,7 =$ **0,05**
$Sim(Document_X, C2) = 0 \times 0,1 + 0,3 \times 0,4 + 0 \times 0 + 0,8 \times 0,7 =$ **0,68**
$Sim(Document_X, C3) = 0,2 \times 0,1 + 1 \times 0,4 + 0 \times 0 + 0 \times 0 =$ **0,42**

Une introduction plus complète de ces modèles peut être trouvée en consultant Sebastiani [2002], Baeza-Yates et Ribeiro-Neto [1999], ou encore Grossman et Frieder [2004].

Il n'y a pas de lien direct entre la méthode qui est proposée dans ce chapitre et ces différentes méthodes d'apprentissage artificiel. Notre système ne doit pas apprendre des modèles car il les connaît a priori. Les caractéristiques qui définissent ces modèles sont constituées de termes simples ou composés, et sont déterminées en grande partie par le contenu de la ressource terminologique de départ, ce qui restreint grandement la dimension de cet espace[12]. Les sections suivantes montreront cependant certaines similitudes. La pondération des différentes caractéristiques peut par exemple être rappro-

[11] À titre d'exemple, les caractéristiques qui seraient constituées par les mots simples pourraient obtenir un poids en rapport avec la fréquence de ce mot dans le texte. Des valeurs plus élaborées, telles que le *TF.IDF* (voir section 2.4.3) peuvent évidemment aussi être envisagées.

[12] Par rapport aux techniques d'apprentissage pour lesquelles la dimension du vecteur de caractéristiques peut potentiellement s'élever jusqu'à être équivalente au nombre de mots différents contenus dans le texte.

chée de ce qui se fait pour les approches vectorielles. Le calcul du poids final d'une classe pourrait aussi être comparé à la mesure de similarité par produit scalaire utilisée dans ces approches. La différence provient des poids attribués aux caractéristiques des modèles de référence liés à chaque classe, qui sont dans notre cas booléens (un terme caractérise ou non une classe).

2.2.2 Utilisation de terminologies pour la classification

Il existe certains systèmes qui, par leur démarche orientée vers le traitement linguistique, se rapprochent de la méthode que nous proposons. Avec leur *Open Biomedical Annotator*, Shah *et al.* [2009] s'intéressent comme nous à l'indexation de textes par rapport à une référence terminologique. Il s'agit dans ce cas de textes médicaux indexés au moyen de concepts issus du métathésaurus UMLS. Dans ce cadre, deux méthodes d'extraction de *concepts* sont testées : Mgrep et MetaMap[13]. Mgrep est une approche assez simple et très rapide qui repose sur un vaste lexique dont l'exploitation permet la reconnaissance de concepts particuliers au domaine abordé. MetaMap (Aronson [2001], Aronson et Lang [2010]) constitue un système plus complexe qui met en œuvre une chaîne de traitement lexicale et syntaxique, incluant entre autres, la *tokenisation*, la détection des limites de phrase, l'identification d'acronymes et d'abréviations, la détection des parties de discours (*POS tagging*), l'utilisation de lexiques spécifiques et, finalement, une analyse syntaxique de surface permettant la reconnaissance de groupes polylexicaux et de leurs *têtes*. La liste des groupes retrouvés dans un texte est étendue grâce à une étape de génération de variantes (acronymes, abréviations, synonymes, variantes dérivationnelles, variantes flexionnelles et orthographiques). Ces éléments sont ensuite comparés au contenu d'UMLS de manière à dégager des liens avec certains concepts de celui-ci. Ce processus part donc du texte pour aller vers le thésaurus, ce qui constitue une première différence de taille avec notre approche, qui elle fait le chemin inverse, du thésaurus au texte. Cette dernière approche, qui prend la forme de transducteurs générés automatiquement, offre selon nous de meilleurs perspectives en terme de performance, car elle ne s'attache à traiter que les expressions du texte qui ont une réelle chance d'être reliées au thésaurus. Partir du texte et de l'extraction de tout les concepts potentiels nécessite plus de travail, en partie *inutile* au final, ce qui peut se payer cher lorsque l'analyse déployée est lourde[14]. Une deuxième différence est la dépendance importante de MetaMap par rapport à la langue traitée, c'est-à-dire l'anglais. De notre côté, le système que nous proposons n'est pas non plus complètement indépendant de ce point de vue, mais les éléments particuliers à la langue – le français dans notre cas – se limitent principalement à des outils de prétraitement (*POS tagging*, racinisation, etc.) qui sont disponibles pour de nombreuses langues.

Le français, en tant que langue cible, représente d'ailleurs un élément intéressant, car moins de systèmes lui sont consacrés, par rapport à l'anglais. Il existe cependant quelques initiatives en ce sens. C'est par exemple le cas de Névéol *et al.* [2006] qui se rapproche très fort des principes de notre méthode par l'utilisation d'une série de transducteurs pour l'indexation de documents en français

[13] En fin de compte, les auteurs ont choisi d'utiliser Mgrep, globalement moins complet, mais beaucoup plus rapide, ce qui dans le contexte de l'application *temps réel* visée constituait un critère important.

[14] Si MetaMap peut traiter un texte en moins d'une minute, il est aussi parfois possible que la même analyse prenne plusieurs heures, à cause de la présence de phrases ou d'expressions complexes. Cette augmentation du temps de traitement est imputable à l'analyse des très nombreux liens potentiels générés.

à l'aide de MeSH[15]. Il existe cependant une différence importante puisque ces transducteurs sont construits manuellement, en collaboration avec des experts, alors que nous proposons de les générer automatiquement.

Névéol *et al.* [2005] évaluent deux systèmes hybrides d'indexation de documents médicaux dans MeSH : *Medical Text Indexer* (MTI) pour l'anglais, *MeSH Automatic Indexer for French* (MAIF) pour le français. Ces systèmes combinent à la fois une approche de traitement automatique du langage et une approche plus statistique. Dans le cas de MTI, pour l'anglais, la partie TAL est à nouveau prise en charge par MetaMap, alors que la partie statistique est basée sur une méthode appelée *PubMed Related Citations*. En ce qui concerne le système prévu pour le français, MAIF, l'approche TAL se base sur un dictionnaire dérivé de MeSH, et enrichi avec des variantes, pour l'identification des concepts. Ceux-ci sont ensuite traduits vers les termes MeSH correspondants et reçoivent finalement un score de pertinence. La partie statistique implémente la méthode *k-Nearest Neighbour*. Celle-ci permet de calculer, par rapport à un ensemble de documents déjà annotés, les k plus proches voisins. La comparaison s'effectue sur la base d'un *sac de mots* généré à partir du titre, dont a été enlevé les *stopwords*, et est calculée en comptant le nombre de mots communs. Ces deux systèmes adoptent, comme précédemment pour Shah *et al.* [2009], une approche qui part du texte pour aller vers le thésaurus. L'originalité est ici apportée par la combinaison de deux méthodes, à la fois pour l'anglais et le français.

Toujours dans le domaine médical et parmi les développements réalisés pour le français, Pereira *et al.* [2008] ont proposé, avec F-MTI (*French Multi-Terminology Indexer*), un système pour lequel l'accent a été porté sur l'utilisation de plusieurs terminologies. La méthode d'indexation proprement dite consiste à former des *sacs de mots* à partir des titres. Ceux-ci subissent préalablement quelques traitements : décapitalisation, désaccentuation, élimination des *stopwords* et lemmatisation ou racinisation. L'indexation est réalisée en comparant, sans tenir compte de l'ordre, les *sacs* ainsi constitués avec l'ensemble des termes des quatre terminologies. Ces termes ont été traités selon le même procédé afin de rendre la comparaison possible. Au final, les résultats montrent que l'utilisation de plusieurs terminologies permettent d'améliorer le rappel. Cette amélioration s'accompagne cependant d'une baisse de la précision.

Enfin d'autres travaux, moins proches des nôtres, font aussi usage d'un thésaurus pour améliorer ou guider la classification. Par exemple, KEA++ (Medelyan et Witten [2006]) qui est un système agissant en deux phases : l'extraction de mots clés et leur filtrage à l'aide d'un thésaurus sont suivis par une étape d'apprentissage capable de mettre en œuvre différents types d'algorithmes. Les modèles ainsi appris permettent ensuite de réaliser la classification de tout nouveau document. Par ailleurs, Pouliquen *et al.* [2003] présentent une méthode statistique et associative de classification de documents à l'aide du thésaurus Eurovoc. Cette méthode se base sur différentes mesures de similarité. L'étude a été menée sur diverses langues dont l'anglais, l'espagnol et le français.

[15] Medical Subject Headings : `http://www.nlm.nih.gov/mesh/`

2.3 Adaptation d'une ressource terminologique en ressource d'extraction

2.3.1 Principe général

Le système de classification proposé modifie et transforme une ressource terminologique en une ressource d'extraction sous la forme de transducteurs. Cette *ressource d'extraction* définit un ensemble de motifs lexicaux complexes et possède la caractéristique de pouvoir être directement appliquée aux textes[16] de manière à repérer des séquences textuelles – des occurrences particulières des motifs recherchés – et à leur attribuer un code de catégorie spécifique. L'opération de génération de cette ressource est unique et ne fait pas partie du processus de classification répété pour chaque document. Elle est facilement réalisable, malgré le nombre élevé d'éléments que peut contenir la ressource, car elle est en grande partie automatique.

Plus concrètement, parmi les nombreux types de ressources terminologiques (Section 1.4.1), celui qui semble le mieux convenir est le thésaurus, car il constitue un bon compromis entre la complexité de sa construction et de son utilisation et son potentiel pour l'indexation et la recherche (du point de vue de l'utilisateur). En effet, les thésaurus présentent deux caractéristiques intéressantes : l'organisation hiérarchique des concepts et la définition de synonymes pour une grande partie de ceux-ci. La structure hiérarchique du thésaurus permet, lors de la recherche, la désambiguïsation des termes homographes. Pour l'indexation, la hiérarchie n'est par contre pas réellement utilisée[17], si ce n'est pour réduire le problème de départ à un problème plus général[18]. À titre d'illustration, un extrait du thésaurus utilisé pour l'évaluation de la méthode, à la section 2.5.2, peut être consulté en annexe A.

Il serait donc possible de se baser sur une simple liste de termes, à condition qu'ils soient accompagnés de synonymes. Lors de la recherche, il y aurait cependant une perte au niveau de la désambiguïsation des termes homographes qui ne serait plus assurée par la hiérarchie. Une ontologie pourrait apporter certains éléments supplémentaires pour l'indexation, mais la complexité de sa construction[19] et de son exploitation est un obstacle qui n'est pas conciliable avec l'objectif d'arriver à un traitement symbolique, basé sur un socle linguistique important, mais qui minimise l'effort de production des ressources.

Lors de la génération des transducteurs, il est donc possible de choisir le niveau d'analyse que l'on désire obtenir. Il est par exemple possible de prendre en compte l'ensemble des catégories, quel que soit leur positionnement hiérarchique dans le thésaurus, ou de se limiter à un niveau plus général, comme par exemple celui des microthésaurus. Dans ce cas, la classification est limitée à un seul niveau de profondeur. La totalité de la ressource est tout de même exploitée car l'ensemble des descripteurs et synonymes reliés à un microthésaurus particulier est utilisé pour construire le transducteur. Ce regroupement constitue une généralisation qui permet de réduire le nombre de catégories.

[16] À l'aide des logiciels adéquats, ici Unitex.
[17] Cet aspect est laissé à titre de perspective.
[18] Par réduction du nombre de catégories aux plus générales d'entre elles.
[19] Du moins si le but est d'exploiter toute la richesse proposée par ce type de ressource. Évidemment, si seuls les liens hiérarchiques sont utilisés, on se ramène à un thésaurus.

L'analyse la plus fine est par contre obtenue en conservant toutes les catégories. Pour chacune d'elles, un transducteur est généré. Dans ce cas, de petits transducteurs sont produits en très grand nombre alors que le premier cas mène à des transducteurs plus volumineux mais en nombre plus restreint. Chaque transducteur possède une sortie différente qui indique le code de catégorie auquel il est relié.

Les sections suivantes détaillent les principaux traitements nécessaires au passage des termes et expressions à des listes de motifs lexicaux. Ceux-ci seront ensuite retranscrits sous la forme de transducteurs. L'implémentation réalisée repose sur le logiciel de traitement de corpus Unitex[20] (Paumier [2008], Paumier [2003]) qui est utilisé comme une boîte à outils, et en adopte le format.

2.3.2 Élargissement de la description lexicale des concepts

Pour chaque concept, il existe une forme qui nomme celui-ci. Cette forme, appelée *descripteur* dans un thésaurus, peut être accompagnée de divers synonymes (*non descripteurs*). Il est évident que ces derniers doivent être exploités afin de construire des transducteurs de reconnaissance les plus complets possible. C'est d'autant plus important que la forme sélectionnée pour décrire le concept n'est pas nécessairement la plus fréquente, ce choix pouvant obéir à d'autres contraintes. Dans certains cas, plusieurs terminologies sont liées au sein d'une structure plus générale[21], donnant ainsi la possibilité d'augmenter le nombre de termes synonymes et donc de couvrir de manière plus importante les variations lexicales d'un concept.

La morphologie des termes issus des terminologies est variable. Il peut s'agir de termes simples ou composés. Ces derniers sont souvent utilisés pour exprimer des concepts très précis ou complexes, mais peuvent également être employés pour rassembler plusieurs formulations en une seule. Ce dernier cas se traduit par diverses configurations, dont :

- la coordination (« et », « ou »)
  - « Protection maternelle et infantile »
  - « Rétraction ou atrophie de paupière »
- l'énumération (au moins 3 éléments)
  - « Traitement cruel, inhumain ou dégradant »
  - « Colite, entérite et gastro-entérite infectieuses »
- la précision d'acronymes (deux configurations)
  - « Coronavirus associé au Syndrome Respiratoire Aigu Sévère (SRAS) »
  - « Syndrome NARP (neuropathie, ataxie, rétinite pigmentaire) »

La reconnaissance des formes partielles reliées à ces expressions *complexes* peut poser problème si l'analyse s'en tient à la recherche de la séquence entière de départ. La solution qui s'impose est par conséquent de décomposer ces expressions en termes plus simples. La décomposition des cas que nous avons exposés ci-dessus est réalisée pour la génération des transducteurs.

[20] `http://www-igm.univ-mlv.fr/~unitex/`
[21] C'est par exemple le cas de MeSH et de ICD-9-CM dans UMLS.

2.3.3 Normalisation linguistique : racinisation et lemmatisation

Cette étape de normalisation a pour but d'étendre la couverture aux variations possibles d'une expression, telles que le passage du singulier au pluriel. Par exemple, à partir de l'expression *taux d'intérêt légal* issue d'un thésaurus, nous désirons aussi pouvoir retrouver les formes[22] :

- « taux d'intérêt**s** légal »,
- « taux d'intérêt lég**aux** »,
- « taux d'intérêt**s** lég**aux** ».

Deux techniques permettent d'atteindre ce résultat : la racinisation (*stemming*) et la lemmatisation. La *racinisation* consiste en l'extraction d'un préfixe correspondant à la racine d'un mot. Nous avons utilisé à cet effet l'implémentation Snowball[23] de l'algorithme de Porter [1997]. Cette approche produit des transducteurs principalement composés d'expressions régulières (sous la forme de *filtres morphologiques* dans Unitex). La *lemmatisation* permet de relier une forme fléchie à sa forme canonique (ou lemme). Les patrons générés font alors directement référence aux lemmes et non plus à des formes fléchies, ce qui permet de tirer parti de dictionnaires électroniques. Pour obtenir les lemmes, nous avons utilisé Treetagger[24], un étiqueteur morpho-syntaxique multilingue (Schmid [1994]).

Les résultats obtenus avec les deux méthodes sont globalement similaires, avec un léger avantage pour la lemmatisation. Cette dernière n'est cependant pas nécessairement adaptée à tous les types de textes. Lorsqu'un corpus fait apparaître un nombre important de mots non reconnus par le dictionnaire (néologismes, certains toponymes, termes techniques ou particuliers à un domaine, etc.), et donc pour lesquels aucun lemme ne peut être proposé par Treetagger, la racinisation s'impose. En ce qui concerne la vitesse d'exécution, la lemmatisation est également préférable. Elle peut présenter un temps de traitement un peu plus élevé que la racinisation lors de la phase de création des transducteurs, c'est-à-dire lorsque ces processus sont réellement effectués, mais elle se rattrape largement lors de la phase d'application des transducteurs. En effet, nous avons observé que le temps d'exécution des transducteurs constitués de filtres morphologiques est de loin plus élevé que celui obtenu avec les transducteurs lemmatisés. Il est évidemment possible que cela soit dû à l'implémentation d'Unitex et que d'autres systèmes permettent d'améliorer le temps de traitement pour ce type de transducteurs.

2.3.4 Stopwords et ponctuation

L'utilisation du lexique dans la production des textes est très déséquilibrée. Certains mots sont utilisés de manière très fréquente alors que d'autres ne sont employés qu'occasionnellement. Ces observations ont depuis longtemps été mises en avant par plusieurs lois (loi de Zipf [1949][25], loi de Pa-

[22] Un tel procédé présente un risque de surgénération. Dans un contexte de reconnaissance, cela ne représente pas un véritable problème. Au contraire, cela permet de reconnaître les occurrences mal orthographiées ou s'éloignant de la norme.

[23] `http://snowball.tartarus.org/`

[24] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

[25] Selon cette loi, la fréquence d'utilisation d'un mot est inversement proportionnel à son rang. Plus précisément, la fréquence du mot au X^{me} rang serait égale à la division par X de la fréquence du mot le plus courant.

reto[26]). Plus particulièrement, Baeza-Yates et Ribeiro-Neto [1999] insistent également sur le fait que les mots les plus fréquents ne sont pas toujours très utiles en recherche d'informations car ils perdent tout pouvoir discriminant[27]. Ces mots suremployés se trouvent principalement dans une catégorie particulière du lexique dont la fonction principale dans le langage naturel est de structurer celui-ci. Ces mots permettent d'articuler les phrases, mais ne portent que peu de sens en eux-mêmes. Il s'agit par exemple des articles, des prépositions ou encore des conjonctions. En recherche d'informations, ces unités sont regroupées dans une catégorie appelée *stopwords*. La proportion de *stopwords* dans une collection de documents a été évaluée, pour l'anglais, à 40% par Francis et Kucera [1982] (cité dans Grossman et Frieder [2004]).

Sur la base de cette double caractéristique, absence de pouvoir discriminant et peu d'apport sémantique, les *stopwords* sont généralement éliminés lors des traitements en recherche d'informations. Cela apporte un gain de performance en termes de temps de traitement mais également d'espace de stockage. Ce procédé peut cependant aussi se traduire par un effet néfaste sur le rappel, entre autres en cas de recherche d'une expression précise comprenant un grand nombre de ces *stopwords* (l'exemple « to be or not to be » est donné pour l'anglais par Baeza-Yates et Ribeiro-Neto [1999]).

Lors de la transformation de la ressource terminologique en ressource d'extraction, nous avons bien entendu aussi pris en compte le point concernant les *stopwords*. Concrètement, ces mots particuliers, auxquels nous avons ajouté la ponctuation et les nombres, ont été remplacés par une méta-étiquette (`<TOKEN>`). Celle-ci autorise la reconnaissance de n'importe quel *token*, c'est-à-dire toute suite de caractères non interrompue par un espace ou un signe non-alphabétique. Ce traitement a pour objectif d'améliorer la reconnaissance d'expressions dans lesquelles un *stopword* peut être remplacé par un autre mot. C'est par exemple le cas pour

- « contrôle **de** chômeurs »,
- « contrôle **du** chômeur »,
- et « contrôle **des** chômeurs »[28]

qui seront toutes reconnues par un seul et unique transducteur ne comprenant que le motif :

```
<contrôle><TOKEN><chômeur>.
```

2.3.5 Insertions

De nombreuses variations peuvent survenir, à l'intérieur d'une expression composée, par insertion d'éléments *facultatifs*. Leur prise en compte permet de reconnaître des expressions qui sont sémantiquement très proches, mais qui présentent quelques différences dans leur énonciation. Ces éléments de variation peuvent être constitués par des mots supprimés en passant d'une forme complète à une forme simple. Par exemple :

[26] Cette loi, due à Vilfredo Pareto (1848-1923), stipule que 80% des effets sont générés par seulement 20% des causes.

[27] Ils estiment que la moitié d'un texte peut généralement être couvert par un lexique très restreint de mots, dont le nombre ne dépasse pas quelques centaines. Ces mots apparaissent dans 80% des documents d'une collection.

[28] Ce cas n'est pas couvert par la lemmatisation car le lemme de « des » est « un ».

- « Tribunal Pénal pour l'ex-Yougoslavie » au lieu de « Tribunal Pénal **International** pour l'ex-Yougoslavie » ;
- « Office des Pensions » pour « Office **national** des Pensions » ;
- etc.

Il peut également s'agir de précisions par rapport au concept de base, « produit antibactérien » pourrait par exemple se décliner en :

- « produit **naturel** antibactérien » ;
- « produit **chimique** antibactérien » ;
- etc.

Afin de gérer ce genre de variations, les motifs construits à partir de la forme de base autorisent les insertions de termes facultatifs entre chaque mot. D'une manière générale, ce type d'ajout est souvent constitué par des adjectifs. Pour maximiser la reconnaissance, il est cependant possible d'utiliser à nouveau la méta-étiquette `<TOKEN>`. Cette extension permet donc d'aller plus loin dans la reconnaissance d'expressions similaires, comme c'est le cas pour « agence de protection de l'environnement » qui peut être retrouvée sous la forme de « agence `<TOKEN>` de protection de l'environnement », avec `<TOKEN>` parmi « fédérale », « régionale », « belge », « compétente »[29], etc.

Si la couverture (ou rappel) est favorisée par ce traitement, celui-ci peut également entraîner, dans une certaine mesure, un augmentation du bruit. L'impact négatif est cependant négligeable, la plupart du temps, grâce aux contraintes posées par les éléments lexicalisés qui encadrent les éléments *génériques* (`<TOKEN>`). Le contexte joue en quelque sorte un rôle *protecteur*, qui empêche les méta-étiquettes de reconnaître n'importe quelle unité.

2.3.6 Casse et accentuation

La casse et les accents sont deux éléments qui subissent une nouvelle normalisation. Celle-ci consiste à éliminer toutes les distinctions introduites par les caractères capitalisés ou accentués.

En ce qui concerne la casse, la raison est double.

Premièrement, par rapport aux unités à repérer dans les textes, il y a lieu de pouvoir reconnaître un mot, dont la lettre initiale apparaît en caractère majuscule, lorsque :

– le mot débute une phrase ;
– il y a dans le lexique une déclinaison de ce mot en une version avec et sans lettre capitale initiale (par exemple le nom « Italien » et l'adjectif « italien ») ;
– l'usage des lettres capitales est approximatif (par exemple dans les noms d'organisation).

Deuxièmement, il se peut que, pour diverses raisons de mise en page, le mot soit complètement en

[29] Dans ce dernier cas, l'expression qui pourrait être trouvée serait « agence compétente pour la protection de l'environnement », « pour la » étant reconnu à la place de « de » suite au traitement des *stopwords*.

lettres majuscules.

Par ailleurs, en ce qui concerne la ressource terminologique de base, il arrive que certaines d'entre elles ne proposent que des mots complètement capitalisés, ce qui implique dès lors de ne pas tenir compte du tout de la casse.

Les caractères accentués ne sont quant à eux pas pris en compte pour deux raisons :

- la modification d'accentuation qui peut intervenir entre des formes proches (le nom « rêve » et l'adjectif « rêvé », que l'on pourrait retrouver dans « la voiture rêvée » et « la voiture de rêve » et dont les lemmes non accentués se rejoignent en « reve », ou encore les différentes formes « enlevé », « enlève », « enlever » qui donnent une seule racine non accentuée « enlev ») ;
- d'un point de vue orthographique, la désaccentuation est intéressante pour couvrir les modifications introduites par la nouvelle orthographe (par exemple « évènement » qui devient « événement ») ainsi que les fautes d'orthographe qui pourraient être commises au niveau de l'emploi des accents.

En pratique, la racinisation via Snowball inclut la désaccentuation et la décapitalisation alors que ces étapes doivent être ajoutées après la lemmatisation.

L'objectif de ces traitements est évidemment d'améliorer la couverture, et donc le rappel. Il existe cependant un léger risque de perte de précision, mais celui-ci peut être compensé, la plupart du temps, par l'effet *protecteur* du contexte[30]. Enfin, ces choix ont une conséquence pratique : les dictionnaires électroniques et les textes doivent être traités de la même manière afin que la reconnaissance puisse être menée à bien.

2.3.7 Traitement d'exceptions

Cette dernière étape concerne certaines exceptions qui doivent être prises en compte. Par exemple, l'acronyme « CAS » qui correspond à « Caisse d"allocations sociales » est ambigu avec le nom « cas ». La casse n'est d'aucun secours étant donné le traitement adopté à ce niveau (voir section 2.3.6). Utiliser cet acronyme tel quel conduirait à reconnaître sa présence à chaque occurrence du nom commun « cas », ce qui fausserait fortement l'analyse. Dans ce cas, seul l'acronyme dans sa version avec points (« C.A.S. ») est ajouté au transducteur.

Autre exception : suite au traitement des *stopwords*, certains motifs se résument normalement à une seule méta-étiquette telle que `<TOKEN>`. Par exemple, on peut imaginer un intérêt à retrouver le nom de saison « été », alors que la forme verbale homographe, qui constitue un stopword, n'est d'aucune utilité. Un motif composé d'une seule étiquette *générique* peut conduire à la reconnaissance de toutes les unités du texte, ce qui n'est évidemment pas souhaitable. Dès lors, seule la forme d'origine lemmatisée ou racinisée est conservée et une restriction supplémentaire lui est ajoutée en

[30] Une forme désaccentuée et décapitalisée peut devenir ambiguë si elle est considérée seule, mais ne l'est pas, ou moins, lorsqu'elle est employée dans le contexte d'une expression composée.

imposant la présence d'un déterminant au début du motif.

2.3.8 Génération automatique des transducteurs

Les étapes précédemment exposées ont pour but de fournir, pour un concept nommé par un terme simple ou composé, un ensemble de motifs de reconnaissance apparentés à des expressions régulières. Ces motifs ont la particularité, par rapport au terme de départ, d'augmenter sensiblement la couverture, c'est-à-dire qu'ils sont capables de reconnaître un nombre important d'expressions reliées. Celles-ci constituent des variations plus ou moins importantes de la formulation de départ (voir les exemples exposés dans les sections précédentes).

Les motifs obtenus pour chaque concept, c'est-à-dire chaque catégorie ou classe, doivent pouvoir être appliqués d'une façon efficace aux textes afin de reconnaître les expressions susceptibles d'entrer en compte pour la classification du document. Pour ce faire, comme nous l'avons déjà mentionné à la section 2.3.1, nous avons choisi d'adopter le format des transducteurs défini par le logiciel de traitement de corpus Unitex.

Les transducteurs constituent un format particulier des grammaires locales. Une grammaire locale (Gross [1989, 1997]) permet de représenter des structures lexicales ou syntaxiques plus ou moins complexes. Ces structures sont souvent représentées sous la forme d'un graphe (Figure 2.1). Ce graphe est parcouru depuis l'état initial, représenté par une flèche placée sur la gauche, jusqu'à un état final, le carré contenu dans un cercle disposé à l'extrême droite, et cela en parcourant les transitions représentées par les boîtes. Ces dernières contiennent des éléments qui définissent[31] les séquences de caractères qui peuvent être identifiées dans les textes. Elles peuvent également contenir des appels à des sous-graphes permettant de construire des motifs assez complexes. Les grammaires locales sont aussi appelées des *reconnaisseurs*. Un transducteur est une grammaire locale pourvue à la fois d'un alphabet d'entrée, les séquences qui peuvent être reconnues, et d'un alphabet de sortie, les séquences qui peuvent être produites (« [animal] » dans le cas présenté à la figure 2.1).

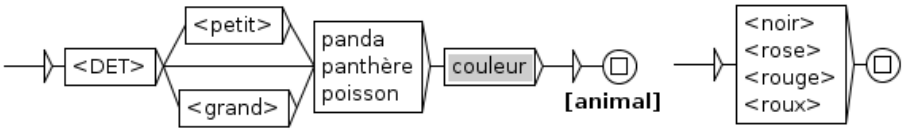


Figure 2.1 : Exemple de graphe Unitex (à gauche), accompagné du sous graphe « couleur » (à droite).

Tout comme pour les tâches exposées dans les sections précédentes, la création des transducteurs à partir des listes de motifs s'effectue de manière automatique. Un transducteur est produit par classe. Chaque motif défini pour cette classe correspond à un *chemin* possible entre l'état initial et l'état final. Le chemin en question est constitué de divers éléments :

– des éléments dits *principaux*, lemmatisés et exprimés au moyen d'une forme cano-

[31] À l'aide de divers moyens : une chaîne de caractères (« porte »), un appel aux dictionnaires via une forme canonique (« <porter> »), un code grammatical (« <V> ») ou sémantique (« <Prenom> »), une méta-étiquette (« <TOKEN> »), une filtre morphologique (« ≪^évén≫ »), etc.

nique (`<forme_canonique>`), ou racinisés et exprimés à l'aide d'un filtre morphologique (≪`^racine`≫)[32] ;
- des méta-étiquettes `<TOKEN>` pour remplacer les *stopwords* ;
- des appels au sous-graphe `insert`, lequel contient une méta étiquette au choix (par exemple `<TOKEN>` ou `<A>`).

Les boîtes sont reliées selon les principes suivants (Figure 2.2) :

(A.) un motif simple composé d'un seul élément (forcément un élément lemmatisé ou racinisé) est directement relié à l'état initial et à l'état final ;

(B.) un motif composé (au moins deux éléments lemmatisés ou racinisés) a son premier élément relié à l'état initial et son dernier à l'état final, chaque élément est de plus relié à son prédécesseur et à son successeur (si ceux-ci existent) ;

(C.) les éléments *principaux* successifs d'un motif composé sont également reliés par l'intermédiaire d'un appel au sous-graphe `insert` ;

(D.) si un *stopword* a été identifié entre deux éléments *principaux*, un élément `<TOKEN>` est inséré entre ceux-ci (si une expression contient plusieurs *stopwords* successifs, ils sont réduit à une seule balise) ;

(E.) un appel au sous-graphe `insert` relie aussi les éléments *principaux* à l'élément `<TOKEN>` qu'ils encadrent ;

(F.) ces deux éléments `insert` peuvent également être empruntés directement, sans passer par l'élément `<TOKEN>` issu du *stopword*, pour relier les deux éléments *principaux*.

La séquence de sortie qui est placée à la fin, avant l'état final, permet de produire la référence à la classe concernée (son numéro ou son code). L'ensemble des transducteurs ainsi produits sont rassemblés dans un transducteur global grâce au mécanisme des sous-graphes.

Le transducteur est généré dans le format GRF correspondant à cette représentation graphique. Un exemple, correspondant au transducteur de la figure 2.2 est détaillé à la figure 2.3. La première partie (jusqu'à la ligne « # ») constitue un en-tête qui permet d'ajuster certains paramètres de mise en page. La ligne suivante indique le nombre de lignes contenues jusqu'à la fin du fichier. Les dernières lignes suivent ensuite toutes le même format :

```
"contenu_de_la_boîte" Position_X Position_Y Nombre_de_liens Liste_des_liens
```

Les deux premières représentent l'état initial et l'état final, alors que les suivantes constituent le reste des transitions du transducteur. À noter que la liste des liens correspond aux numéros de lignes, tels que la ligne 0 correspond à l'état initial. Ce fichier est ensuite compilé à l'aide du programme `Grf2Fst` d'Unitex pour obtenir sa version FST2. Celle-ci est alors exploitable sur un texte au moyen du programme `Locate`. Un exposé plus détaillé concernant les formats de fichiers et les programmes

[32] La forme de départ est également conservée dans les cas où une forme lemmatisée ou racinisée n'aurait pu être obtenue, ainsi que pour pallier partiellement une possible défaillance de ces mécanismes de normaliation linguistique. Dans le cas de la lemmatisation, on ajoute une contrainte grammaticale, sachant que les noms et des adjectifs sont les éléments grammaticaux qui représentent le principal intérêt.

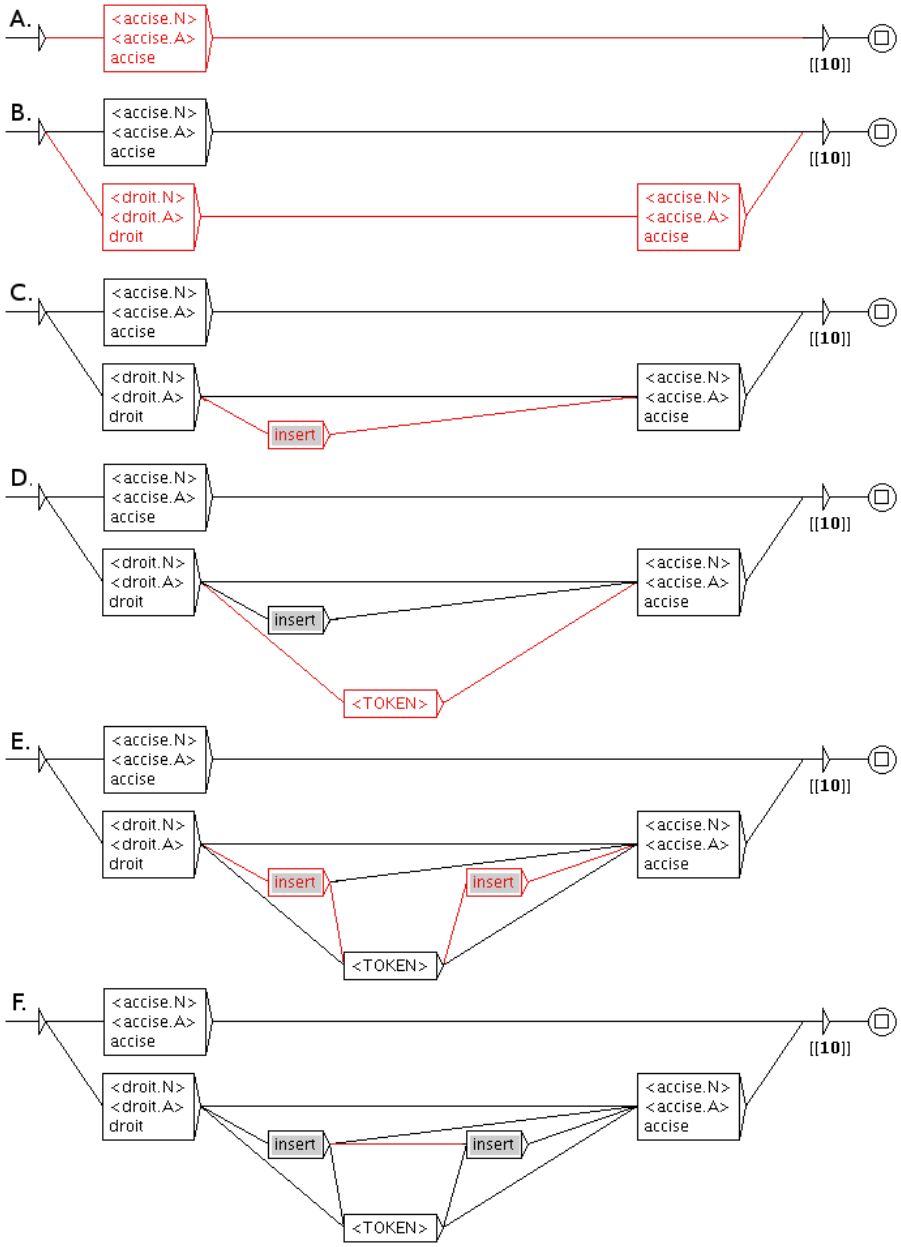


Figure 2.2 : Illustration des différents principes dirigeant la construction du transducteur (ici, en version lemmatisée).

utilisés est disponible dans le manuel d'Unitex (Paumier [2008]).

2.4 Extraction et classification

2.4.1 Prétraitement des textes

Avant l'application aux documents des transducteurs générés à partir du thésaurus, une étape de prétraitement des textes est nécessaire. Lors de la création des transducteurs et du traitement des *stopwords*, les formes élidées telles que « l' » ont été remplacée par une méta-étiquette, par exemple `<TOKEN>`. Or Unitex crée, pour cette forme « l' », deux tokens : « l » et « ' ». Cette forme n'est

```
#Unigraph
SIZE 1188 840
FONT Times New Roman:  9
OFONT Arial Unicode MS:B 9
BCOLOR 16777215
FCOLOR 0
ACOLOR 13487565
SCOLOR 16711680
CCOLOR 255
DBOXES y
DFRAME n
DDATE n
DFILE n
DDIR n
DRIG n
DRST n
FITS 100
PORIENT L
#
9
"<E>" 50 40 2 2 3
"" 1068 40 0
"<accise.N>+<accise.A>+accise" 200 40 1 8
"<droit.N>+<droit.A>+droit" 200 80 3 7 4 5
":insert" 425 120 3 7 6 5
"<TOKEN>" 524 160 2 7 6
":insert" 623 120 1 7
"<accise.N>+<accise.A>+accise" 722 80 1 8
"<E>/[[10]]" 1018 40 1 1
```

Figure 2.3 : Encodage au format GRF d'un transducteur contenant la liste de termes (avec lemmatisation) pour la classe 10.

donc plus reconnaissable par une seule étiquette `<TOKEN>` mais bien par deux. Afin d'éviter cette désynchronisation, un transducteur de prétraitement remplace toutes les formes élidées par une forme complète correspondante, par exemple « le » pour « l' ».

Une procédure de désambiguïsation ciblée peut également être souhaitable afin d'éviter certaines erreurs récurrentes. Cette étape sera évidemment fonction de la ressource terminologique, ou plus précisément des transducteurs générés à partir de celle-ci, ainsi que du type de texte que l'on envisage de traiter. Par exemple, dans le cas de textes juridiques, l'expression « art. 2 » (article 2) peut être interprétée, à tort, comme reliée à une catégorie `ART` (arts plastiques, etc.) issue de la terminologie. L'idéal est de réaliser une étude exhaustive des termes de la terminologie posant un problème d'ambiguïté. Évidemment, cette tâche n'est pas complètement automatisable et est spécifique à une ressource et à une langue en particulier. Afin de minimiser l'effort nécessaire, on peut cependant envisager de mener cette étude lors de la construction même de cette ressource, qui mobilise de toutes façons les compétences de spécialistes. Pour les ressources existantes, il est nécessaire de mettre au point une méthode de détection de la polysémie permettant de repérer les cas problématiques et qui requièrent une intervention.

Finalement, d'autres tâches de prétraitement plus classiques sont réalisées. Comme déjà exposé à la section 2.3.6, le texte doit être désaccentué et décapitalisé afin de pouvoir être confronté aux transducteurs. La suite du processus, *tokenisation* et application des dictionnaires électroniques, est réalisé au moyen d'Unitex (`Tokenize` et `Dico`).

2.4.2 Application des transducteurs au texte

L'application des transducteurs issus du thésaurus aux textes est également effectué à l'aide d'Unitex (`Locate`). Le résultat est récupéré directement dans le fichier *concord.ind*, habituellement utilisé par Unitex pour construire les concordances. Ce fichier se présente sous la forme d'un index de mots ou d'expressions (Figure 2.4), et est par conséquent très commode à analyser automatiquement[33].

```
0 12 @000101024.xml@
14 16 <title>
53 53 aeroport[[MT111]]
57 57 bruxelles[[MT991]]
60 63 </title>
77 77 president[[MT157]]
113 113 ministre[[MT124]]
117 117 transports[[MT111]]
124 124 armee[[MT122]]
124 124 armee[[MT102]]
140 140 aeroport[[MT111]]
144 144 bruxelles[[MT991]]
193 193 batiments[[MT191]]
235 235 controlees[[MT992]]
264 270 personnel de le aeroport[[MT111]]
274 274 bruxelles[[MT991]]
295 295 ministre[[MT124]]
299 299 transports[[MT111]]
348 348 aeroport[[MT111]]
356 356 livre[[MT133]]
360 360 marchandises[[MT192]]
385 385 ministre[[MT124]]
420 420 president[[MT157]]
446 446 deputee[[MT124]]
```

Figure 2.4 : Liste de mots ou d'expressions, retrouvées à l'aide des transducteurs, telle que présentée dans le fichier concord.ind. Le code de catégorie est inclus entre les doubles crochets.

2.4.3 Pondération

Sur la base de la liste construite après application des transducteurs au texte (Figure 2.4), un poids est calculé pour chaque expression et ensuite globalement pour chaque catégorie. Cette pondération est basée sur une mesure de fréquence, mais d'autres critères peuvent également être pris en compte : la longueur d'une expression composée, la présence du terme dans le titre[34], etc. Ces caractéristiques sont implémentés par des multiplicateurs appliqués au poids initial. Ces éléments sont abordés plus en détail dans les paragraphes suivants.

La valeur de base pour la pondération des expressions est constituée de leur fréquence. Elle peut alternativement être donnée par la mesure du TF.IDF (*term frequency-inverse document frequency*). Cette valeur est couramment utilisée pour évaluer le poids d'un terme par rapport à un corpus donné.

[33] Les deux premières colonnes indiquent les numéros des tokens délimitant l'expression

[34] Cela implique bien entendu que le titre soit délimité, ce qui constitue une attente raisonnable. Les textes pour lesquels cette information ne serait pas disponible peuvent faire l'objet d'une détection du titre lors d'un prétraitement.

Les formules appliquées sont :

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \qquad idf_i = log\frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

où n_{ij} est la fréquence d'un terme i dans le document d_j,
$\sum_k n_{kj}$ est la somme des fréquences pour l'ensemble des termes k d'un document d_j,
$|D|$ étant le nombre de documents dans le corpus,
et $|\{d_j : t_i \in d_j\}|$ le nombre de documents dans lesquels le terme i est présent.

La valeur du TF.IDF est obtenue par : $tf.idf_{ij} = tf_{ij} * idf_i$

Le but de cette mesure est de donner plus d'importance aux mots très fréquents dans un document, mais rares à l'échelle du corpus. Chaque expression de la liste obtient donc un poids TF.IDF. Les valeurs IDF sont précalculées[35] sur le corpus en appliquant les transducteurs de reconnaissance issus de la ressource terminologique.

En plus de la fréquence d'un terme, le fait que les informations importantes pour la classification apparaissent souvent au début du document, c'est-à-dire principalement dans le titre et le résumé, ou le paragraphe d'introduction s'il existe, constitue une caractéristique importante. Cette importance est soulignée par la norme AFNOR Z 47-102 (AFNOR [1993]) et est également bien connue dans le milieu du journalisme. En effet, dans un article de presse, c'est au titre et au *chapeau* que reviennent la tâche d'attirer l'attention du lecteur et de l'amener à lire la suite de l'article. Ces parties concentrent par conséquent de nombreuses informations pertinentes et importantes en ce qui concerne le contenu du texte. Nous avons donc introduit un multiplicateur qui est appliqué au score de base (TF.IDF) si l'expression se situe dans le titre.

L'intérêt particulier porté aux expressions composées est également exploité en tant que caractéristique. Bien que cet aspect soit déjà indirectement pris en compte dans la mesure du TF.IDF, nous avons prévu un multiplicateur supplémentaire pour augmenter le score des termes polylexicaux.

La dernière caractéristique prise en compte concerne les entités nommées. Celles-ci peuvent être détectées à l'aide de transducteurs spécifiques lors de la phase de prétraitement du texte. À nouveau, un multiplicateur est employé pour favoriser ce type d'expressions.

Pour chaque expression qui apparaît dans la liste de résultats (voir section 2.4.2 et figure 2.4), une mesure de base est calculée. Cette valeur est ensuite modulée, le cas échéant, par les multiplicateurs abordés ci-dessus. Pour chaque catégorie représentée, les scores obtenus par les diverses expressions qui y sont reliées s'additionnent pour former le poids final. La liste ordonnée des catégories est alors produite. Cette liste, dont les poids peuvent éventuellement être normalisés entre 0 et 1, compte un nombre variable d'éléments suivant les textes analysés.

[35] Cette méthode peut être perçue comme un biais, mais il s'agit d'une approximation raisonnable des scores IDF qui seraient graduellement construits lors du traitement des mêmes documents en situation réelle.

2.4.4 Réduction de la liste de catégories

La liste pondérée obtenue peut, dans certains cas, être assez longue et les différences de poids importantes. Nous désirons donc réduire cette liste afin de ne garder que les candidats les plus probables. Cette sélection est opérée au moyen d'une méthode de seuil. Trois méthodes différentes ont été expérimentées.

La première méthode (*k-first*) est très simple et permet d'obtenir des résultats de référence. Elle consiste à conserver les k premières catégories correspondant aux meilleurs scores.

Les deux autres méthodes s'appuient sur des valeurs *pivot* qui définissent une valeur centrale selon un certain critère. La deuxième méthode de seuil s'appuie sur la moyenne des poids obtenus par les catégories (*averaged weight*), alors que la dernière (*middle weight*) s'organise autour de la valeur centrale de l'intervalle allant de 0 au poids le plus élevé de la liste de catégories. La valeur *pivot* pour la méthode *averaged weight* correspond au poids moyen obtenu à l'aide de :

$pivot_{aw} = \frac{\sum_{i=1}^{n} w_i}{n}$
où w_i est le poids attribué à la catégorie i, et n le nombre total de catégories proposées.

Pour la méthode *middle weight*, la valeur *pivot* s'obtient assez simplement :

$pivot_{mw} = \frac{w_{max}}{2}$
où w_{max} est le poids le plus élevé de la liste de catégories.

À partir de ces pivots et de la valeur maximale, différents niveaux de seuils, plus ou moins stricts, peuvent être obtenus par échantillonnage. Aux cours des expérimentations, plusieurs valeurs ont été testées afin de déterminer dans quelle mesure le seuillage doit être sévère ou non. Les diverses valeurs intermédiaires sont obtenues par sauts de taille fixe à partir du pivot. Pour obtenir x niveaux de seuil (en plus du pivot), $\frac{x}{2}$ points vont être déterminés au dessus et en dessous de la valeur pivot. L'incrément ajouté ou retranché du pivot est calculé, pour les deux méthodes, selon les formules suivantes :

pour les points compris ente $pivot$ et w_{max},

$$increment = \frac{w_{max} - pivot}{x/2}$$

pour les points inférieurs à $pivot$,

$$increment = \frac{pivot}{x/2}$$

La valeur de x à été fixée à 20, ce qui donne 21 valeurs de seuil au total. Notons que la première méthode produit toujours un nombre fixe de propositions par *point* alors que les deux autres en retournent un nombre variable. Le but final est de déterminer quel type de seuil serait le plus approprié dans un environnement applicatif réel.

2.5 Résultats et évaluation

Cette section débute par une rapide introduction des mesures employées pour évaluer les résultats (Section 2.5.1). Les expériences sont ensuite présentées en détail. La première, menée sur le corpus[36] *Parlementaire*, est la plus complète et a permis d'expérimenter de nombreux paramètres (Section 2.5.2). Pour la seconde, qui porte sur le corpus *Médical*, l'évaluation a été moins exploratoire mais a par contre permis de s'assurer de la portabilité de la méthode (Section 2.5.3).

2.5.1 Mesures

Pour évaluer nos résultats, nous avons employé les mesures classiques de précision (P), de rappel (R) et de f-mesure (F_1, noté F).

$$P = \frac{Syst_{OK}}{Syst_{TOT}} \quad R = \frac{Syst_{OK}}{Man_{OK}} \quad F = \frac{2 * P * R}{P + R}$$

où $Syst_{OK}$ est le nombre de catégories correctement proposées par le système,
Man_{OK} est le nombre de catégories attribuées manuellement par l'indexeur humain et
$Syst_{TOT}$ est le nombre total de catégories proposées par le système.
La f-mesure est une combinaison à proportion égale de la précision et du rappel.

Le rappel représente donc la proportion de codes à trouver qui l'ont été réellement, ce qui revient à évaluer l'absence de *silence*. La précision mesure quant à elle la proportion de codes corrects parmi ceux qui ont été proposés par le système, ce qui correspond à la notion complémentaire de celle de *bruit*.

Nous avons choisi de calculer les résultats globaux du système selon une approche macroscopique. La précision, le rappel et la f-mesure sont donc calculés pour chaque document et une moyenne arithmétique de ces valeurs en donne les mesures finales. Ce choix est motivé par la volonté d'évaluer les performances en fonction de l'application visée qui consiste à traiter les documents un à un et non globalement.

Les résultats ont été évalués par rapport à l'indexation manuelle réalisée par un seul documentaliste pour chaque document. Tout les documents n'ont cependant pas nécessairement été indexés par la même personne. Or, il est prouvé qu'il existe, entre les jugements des différents annotateurs, un certain désaccord (voir section 1.4.3). De plus, la manière d'indexer d'un documentaliste peut également varier au cours du temps. On peut donc considérer qu'un système automatique peut difficilement atteindre les 100% lors d'une évaluation s'il est comparé à l'annotation d'une seule personne.

En effet, parmi les *mauvais* descripteurs proposés par le système, il existe une certaine proportion pour laquelle la distance avec le code à trouver n'est pas très importante. Il s'agit par exemple de cas

[36] Les corpus sont présentés ci-dessous, aux sections 2.5.2 et 2.5.3.

dans lesquels le descripteur choisi est le père, le fils ou encore un frère du descripteur de référence. Dans ce genre de situation, il est tout à fait plausible qu'un autre documentaliste ait pris la même décision que le système automatique.

La procédure d'évaluation effectuée par rapport à une indexation de référence est cependant la plus pratique, car elle peut être automatisée. Une évaluation plus fine pourrait être menée en comparant nos résultats avec plusieurs indexations de référence, voire en demandant la validation manuelle de la part de plusieurs documentalistes. Ce type de travail est cependant difficile à mettre en œuvre, la disponibilité de ces personnes étant évidemment très restreinte.

2.5.2 Première expérience : le corpus *Parlementaire*

Contexte

La tâche d'indexation proposée dans cette première expérience est assez classique. Des documentalistes experts analysent les documents et leur attribuent un certain nombre de catégories (descripteurs), issus d'un thésaurus documentaire. Cette indexation permet ensuite une interrogation de la collection de documents à l'aide de ces catégories.

Les documents et le thésaurus utilisés pour cette évaluation proviennent de la base documentaire d'une organisation actuellement en activité. Les textes sont des documents relevant du domaine législatif et parlementaire. Le thésaurus a été, quant à lui, spécialement conçu[37] pour l'indexation de ces documents au sein de cette organisation.

Présentation des données

Le thésaurus[38] contient 2.514 descripteurs et 2.362 synonymes. Les descripteurs sont répartis en 47 microthésaurus. Le nombre de niveaux hiérarchiques monte jusqu'à 6, mais s'établit plus fréquemment entre 2 et 4. Les expressions composées sont bien représentées : 66,59% des descripteurs (1.674 sur 2.514) et 61.85% des synonymes (1.461 sur 2.362).

Le corpus de test compte 12.734 fichiers XML contenant 32.953.724 mots[39]. La taille moyenne d'un document se situe par conséquent à 2.588 mots. Le titre du document est délimité à l'aide de balises particulières. Pour chaque document, on dispose des catégories assignées manuellement par des documentalistes professionnels en situation réelle. Ces informations constituent la référence pour l'évaluation. Le nombre de descripteurs attribués varie entre 1 et 37, la valeur moyenne s'établissant à 1,92. Pour ce corpus, certaines catégories du thésaurus ne sont utilisées pour aucun document alors que d'autres, au contraire, sont employées de manière très soutenue. 669 catégories ne sont jamais

[37] La conception d'un thésaurus est une activité, généralement effectuée manuellement par des documentalistes experts du domaine d'application, qui consiste à définir et organiser de manière cohérente un ensemble de concepts qui couvrent un domaine particulier. La complexité de cette tâche ne nous permet pas de l'exposer en détail, mais diverses normes peuvent être consultées à cet effet (entre autres ISO [1986] et AFNOR [1981]).

[38] Un extrait de ce thésaurus peut être consulté en annexe A.

[39] Cette mesure approximative du texte brut (pas de balises XML) a été obtenue à l'aide de la commande *wc*.

utilisées et le descripteur le plus fréquent est lié à 412 documents. En moyenne, une catégorie est utilisée pour caractériser 9,71 documents.

Objectifs

La méthode proposée a vocation à être utilisée dans un contexte semi-automatique. Le scénario d'utilisation dans lequel elle s'inscrit consiste à proposer au documentaliste, pour chaque document analysé, une liste de catégories. Celui-ci sélectionne celles qu'il juge appropriées et peut éventuellement en ajouter d'autres[40]. Les catégories proposées au documentaliste pour validation sont sélectionnées par le système dans l'ensemble des 2.514 descripteurs mis *à plat*. Malgré l'organisation hiérarchique du thésaurus, le documentaliste peut en effet sélectionner n'importe quel descripteur, quel que soit son niveau hiérarchique (les *feuilles* ne sont pas les seuls choix possibles). Il est également intéressant de comparer ces résultats avec ceux obtenus en se limitant aux 47 microthésaurus qui, s'ils ne constituent pas des catégories attribuables aux documents, permettent cependant d'aiguiller l'indexation du documentaliste vers un ou plusieurs grands thèmes[41].

Paramètres particuliers

Une stratégie particulière d'extension de la ressource de départ a été utilisée dans le cadre de cette expérience. La localisation géographique étant un aspect important, occupant à lui seul tout un microthésaurus, nous avons décidé d'exploiter une base de données toponymique[42]. Celle-ci nous a permis d'ajouter aux noms de pays les formes adjectivales et les gentilés correspondants (« Italie » est étendu par « italien » et « Italien »).

La méthode de normalisation linguistique utilisée dans cette expérience, lors de la phase de création des transducteurs, fait intervenir la lemmatisation (Treetagger). Cette solution a été choisie pour son efficacité et sa rapidité.

La valeur de base pour la pondération est constituée par la mesure du TF.IDF. Pour chacun des multiplicateurs correspondant aux diverses caractéristiques utilisées dans le calcul de la pondération, différentes valeurs ont été testées. Celles-ci ont arbitrairement été fixées successivement à 1, 2, 5, 10, 20, 50 et 100. Toutes les combinaisons ont été testées. Au final, la configuration optimale a fixé le multiplicateur de titre à 100, celui consacré aux expressions composées à 2 et le dernier, en rapport avec les entités nommées, à 2 également.

[40] L'étape de validation des propositions ne fait pas partie de ce travail mais permet de définir un cadre applicatif précis. Dans de futurs développements, il serait néanmoins intéressant de mener une étude d'évaluation de l'aide effective que peut apporter l'aide à l'indexation.

[41] Le documentaliste devra alors explorer le contenu du microthésaurus pour sélectionner les descripteurs pertinents.

[42] Cette base de données toponymique est une ressource non publiée du Cental, mais est comparable, à une plus petite échelle, à celle proposée par Maurel [2008].

Résultats

Pour les différents tests, les trois méthodes de sélection par seuil ont été testées. Elles fournissent chacune 21 points d'évaluation proposant ainsi des listes de résultat de plus en plus longues. Afin d'obtenir une mesure de référence (*baseline*), un test préliminaire a été effectué sur l'ensemble des 2.514 catégories et a porté sur la recherche des expressions d'origine non modifiées et dont la fréquence n'a pas été pondérée. Il a abouti à une f-mesure maximale de 23,83 (rappel=31,65% et précision=19,11%). Le meilleur rappel atteint se situe à 52,80% mais il est accompagné d'une précision très faible (6,91%).

En ce qui concerne les deux expériences principales (portant respectivement sur 47 ou 2.514 catégories), trois mesures ont été mises en avant et sont reprises dans le tableau 2.1 : la maximisation de la f-mesure, la maximisation du rappel en maintenant une précision *acceptable*[43] d'environ 30% et enfin la recherche du rappel maximal. Les résultats se rapportent au niveau de seuil qui produit les meilleurs résultats. Celui-ci étant différent pour chaque méthode, le nombre moyen de catégories suggérées est évidemment différent.

	47 catégories			**2514 catégories**		
	k-first	averaged weight	middle weight	k-first	averaged weight	middle weight
Meilleure f-mesure						
Nbr. de cat.	2	1,8	**1,9**	2	1,9	**2,3**
F-mesure	0,5743	0,6362	**0,6431**	0,4427	0,5066	**0,5117**
Rappel	0,6789	0,6555	**0,6785**	0,4990	0,5009	**0,5296**
Précision	0,4976	0,6180	**0,6113**	0,3978	0,5123	**0,4949**
Meilleur rappel avec précision à +/- 30%						
Nbr. de cat.	4	6,4	**5,6**	3	**10,1**	4
F-mesure	0,4523	0,4516	**0,4714**	0,4004	**0,4141**	0,4799
Rappel	0,8119	0,8630	**0,8587**	0,5610	**0,6291**	0,5876
Précision	0,3135	0,3058	**0,3248**	0,3113	**0,3086**	0,4056
Meilleur rappel						
Nbr. de cat.	21	15,1	**15,1**	21	38,8	**38,8**
F-mesure	0,2424	0,2354	**0.2354**	0,1694	0,1450	**0,1450**
Rappel	0,9077	0,9101	**0,9101**	0,6890	0,7086	**0,7086**
Précision	0,1399	0,1352	**0,1352**	0,0966	0,0807	**0,0807**

Tableau 2.1 : Résultats des test de classification sur le corpus Parlementaire.

On remarque que la méthode *k-first* est significativement moins bonne que les deux autres. L'inconvénient présenté par cette méthode est de proposer un nombre de descripteurs qui varie pour chaque niveau de seuil mais qui reste identique quel que soit le texte analysé : le x^{eme} niveau de seuil proposera toujours x catégories[44]. Les autres méthodes de seuil, basées sur une valeur centrale,

[43] Dans le contexte d'un système semi-automatique, soumis à une validation humaine, il est raisonnable de laisser la précision à un niveau plus faible afin de maximiser les chances de retrouver le plus de *bonnes* catégories possible. Il est cependant souhaitable de ne pas laisser chuter la précision jusqu'à un niveau auquel le nombre de catégories proposées serait trop élevé.

[44] Sauf dans le cas où le nombre total de catégories retrouvées à l'aide des transducteurs est inférieur à x.

adaptent automatiquement le nombre de propositions retournées en fonction du nombre total de descripteurs dans la liste complète, et surtout en fonction de leurs scores. Ces méthodes dynamiques sont bien entendu plus adaptées étant donné le nombre variable de catégories attribuées en réalité par les documentalistes. Comme ces deux méthodes donnent des résultats relativement similaires et sauf indication contraire, nous n'allons détailler les résultats que par rapport à la méthode *middle weight*.

Dans le cas de la classification sur les 47 microthésaurus, les meilleurs résultats en termes de f-mesure sont obtenus avec une valeur de 64,31. Le rappel obtenu est 67,85% pour une précision de 61,13%. Le nombre moyen de catégories proposées est 1,9 (sur 47 catégories possibles). Cela signifie donc qu'en moyenne, en proposant 1,9 microthésaurus par document, le système couvre plus de deux tiers (67,85%) de ceux attribués manuellement. Parmi ces propositions, un peu plus d'un tiers (38,87%) sont incorrectes. Pour l'application visée, il est intéressant de savoir quel rappel il est possible d'obtenir en acceptant une précision moindre. Un rappel de 85,87% est atteint en conservant une précision *acceptable* de 32,48% (f-mesure : 47,14). En moyenne, l'indexeur humain disposerait alors de 5,6 catégories. Enfin, le meilleur taux de rappel obtenu se situe à 91,01% pour une précision de 13.52%, une f-mesure de 23.54 et un nombre moyen de catégories proposées de 15,1.

La classification sur l'ensemble des 2.514 catégories donne des résultats moins élevés. Cela s'explique aisément par le nombre bien plus important de catégories. Par la généralisation qu'elle implique, la classification dans les 47 microthésaurus permet d'éviter une série d'erreurs telles que la classification dans une catégorie sœur ou dans une catégorie mère/fille. La meilleure f-mesure est obtenue à 51,17. Le rappel se situe alors à 52,96% et la précision atteint 49,49%. Le nombre moyen de catégories proposées est 2,3 (sur 2.514 catégories possibles). Pour mémoire, l'attribution de descripteurs dans notre corpus varie entre 1 et 37, la valeur moyenne étant de 1,92. La méthode de sélection par seuil *middle weight* ne nous a pas fourni de valeur de précision s'approchant des 30%. La méthode *averaged weight* indique par contre que pour une précision de 30,86%, le rappel peut atteindre 62,91% (f-mesure : 41,41). Le nombre moyen de catégories proposées au documentaliste est de 10,1. Enfin, le meilleur taux de rappel obtenu se situe à 70,86% pour une précision de 8,07%, une f-mesure de 14,5 et un nombre moyen de catégories proposées de 38,8.

Parmi les multiples causes possibles, nous avons identifié certains éléments qui peuvent expliquer en partie les oublis et erreurs. L'absence de certaines catégories peut s'expliquer par un manque (relatif) de synonymes dans le thésaurus et l'utilisation, dans le texte, de termes trop ou pas assez concrets en regard du thésaurus. La polysémie de certains termes, et l'ambiguïté qu'elle génère lors de l'analyse des documents, peut quant à elle provoquer du bruit.

À titre de comparaison, nous avons testé KEA++ sur le même jeu de données. Les résultats obtenus sont assez décevants, surtout en termes de précision[45]. Ces chiffres sont cependant à prendre avec précaution car ils sont très éloignés de ceux rapportés par les auteurs[46] (Medelyan et Witten [2005]).

[45] Pour un rappel compris entre 0,48 et 0,53, la précision n'a atteint que 0,10.
[46] Rappel=0,53 ; Précision=0.48 ; F-mesure=0,47.

2.5.3 Deuxième expérience : le corpus *Médical*

Cette expérience a en partie été menée sur la base du travail accompli par Julia Medori au sein du projet Capadis (Medori [2008], Medori [2010]), en collaboration avec les Cliniques universitaires Saint-Luc de Bruxelles.

Contexte

Les hôpitaux gardent de nombreuses traces de leurs activités. Parmi celles-ci, l'archivage et l'encodage des *lettres de sortie* est particulièrement important, car ils conditionnent en partie l'obtention de certains financements publics. Les lettres sont rédigées à l'attention du médecin généraliste du patient lorsque celui-ci quitte l'hôpital. Elles résument son séjour en mentionnant, sous la forme de texte libre, les symptômes qu'il présentait, ses antécédents, les analyses effectuées, les actes posés, etc. Les documentalistes spécialisés *résument* le traitement prodigué en un ensemble de codes issus de l'ICD-9-CM[47]. Les codes désignent une variété d'éléments tels que des diagnostics, des procédures ou encore des facteurs aggravants comme les allergies, le tabagisme, mais aussi tout élément dans le passé médical du patient qui pourrait influencer son état de santé actuel. Le codage représente un travail de grande ampleur et est effectué par des documentalistes professionnels, spécialistes de cette tâche. Le coût que représente cette activité est donc assez important. Par conséquent, de nombreux hôpitaux tentent de réduire la charge de travail en essayant d'automatiser, au moins partiellement, le processus.

Présentation des données

La nomenclature ICD-9-CM contient 15.688 codes, composés de 4 ou 5 caractères. Les trois premiers représentent la *catégorie générale* d'un diagnostic, alors que les chiffres restants le caractérisent de manière plus précise (Figure 2.5). ICD-9-CM est divisé en 1.135 catégories générales.

Code	Label
001	**Choleras**
0010	Cholera à Vibrio cholerae
0011	Cholera à Vibrio cholerae el tor
0019	Cholera, sans autre précision

Figure 2.5 : Extrait de la structure hiérarchique d'ICD-9-CM.

Le corpus d'évaluation comporte 19.692 lettres de sortie en français issues du service de Médecine Interne Générale et pour lesquelles les codes ont été manuellement attribués par les documentalistes. Les patients de ce service souffrent de maladies très variées, ce qui implique l'emploi d'un grand nombre de codes différents (dans ce cas, 6.029 codes provenant de 895 catégories). Les documentalistes ont attribué aux lettres de sortie du corpus un total de 150.116 codes, ou 137.336 catégories, ce

[47] International Classification of Diseases – Ninth revision – Clinical Modification (`http://www.cdc.gov/nchs/icd/icd9cm.htm`).

qui représente une moyenne de 7,6 codes (7 catégories) par document. L'emploi de certaines catégories peut être considéré comme *rare*, 27% des catégories (241 sur 895) étant utilisées moins de six fois, alors que la moyenne d'utilisation d'une catégorie est bien plus élevée (153). Les assignations manuelles des codes aux lettres de sorties constituent la référence pour l'évaluation de la désignation automatique des catégories.

Objectifs

Le but de cette expérience est d'évaluer la possibilité d'apporter aux documentalistes une aide au codage en leur fournissant une liste de codes les plus probables pour chaque document à analyser. Dans cette perspective, il a été décidé de classifier par rapport aux catégories générales (trois premiers chiffres du code) et de laisser le soin au documentaliste, expert du domaine, de choisir le ou les codes exacts dans les développements hiérarchiques de celles-ci. En effet, l'automatisation totale du processus semble un objectif difficile à atteindre, sachant que la lettre de sortie contient rarement toute l'information nécessaire aux choix des codes[48].

L'encodage des lettres de sortie dans ICD-9-CM est un processus analogue à la classification, dans lequel les codes issus de cette nomenclature constituent les classes. La méthode symbolique que nous proposons nous semble tout à fait adaptée, entre autres en raison de sa capacité à fonctionner sans apprentissage, et donc sans avoir besoin d'un corpus d'entraînement annoté. Cet aspect est important car l'encodage des lettres de sortie selon ICD-9 va progressivement être abandonné, au profit d'ICD-10[49], classification pour laquelle peu de données d'entraînement seront disponibles au début de son adoption. D'autre part, l'approche symbolique est capable de gérer les codes *rares* de la même manière que les codes plus courants, ce qui constitue également un argument important.

Travaux spécifiquement apparentés

Depuis le début des années 1990, de nombreuses études ont cherché à automatiser le processus d'encodage des documents médicaux (Ananiadou et McNaught [2006], Ceusters *et al.* [1994], Zweigenbaum et Consortium Menelas [1995]). Les deux principales approches, celles basées sur la connaissance (par exemple le système MedLEE, Friedman *et al.* [2004]) et celles par apprentissage artificiel (par exemple Autocoder, Pakhomov *et al.* [2006]), ont donné de bons résultats au *Computational Medicine Challenge* en 2007 (Pestian *et al.* [2007]). Parmi les trois meilleurs systèmes, deux combinent les approches statistique et symbolique, par exemple Farkas et Szarvas [2008], alors que l'autre ne propose qu'une méthode symbolique (Goldstein *et al.* [2007]).

Toutes ces études ont été développées pour l'anglais. En ce qui concerne le français, Pereira *et al.* [2006] proposent une méthode complètement symbolique. Celle-ci est basée sur un système d'indexation par rapport à la version française de MeSH, auquel est couplée une conversion vers ICD-10.

[48] Des informations additionnelles peuvent généralement être trouvées par le documentaliste dans le dossier médical du patient. Cette source de données n'a cependant pas pu être intégrée dans cette étude.

[49] L'OMS (Organisation Mondiale de la Santé) a déjà entrepris la préparation de l'ICD-11 (`http://www.who.int/classifications/icd/ICDRevision/en/index.html`).

La majorité de ces systèmes exploitent donc, d'une manière ou d'une autre, des connaissances linguistiques. Les résultats sont en général assez encourageants – Autocoder atteint par exemple une précision de deux tiers – mais les documents sont souvent déjà partiellement structurés (les diagnostics sont par exemple déjà annotés). Le nombre de codes et la variété des documents utilisés n'est pas toujours très importante non plus.

Paramètres particuliers de l'expérience

Afin d'élargir la couverture de la ressource de base, nous avons utilisé UMLS[50] comme source de variations pour les termes fournis par ICD-9-CM. Le métathésaurus UMLS unifie et intègre en une seule et unique ressource plusieurs nomenclatures ou terminologies en différentes langues. À chaque concept correspond un identifiant unique[51] qui permet d'extraire du métathésaurus différentes variantes lexicales reliées au terme ICD-9-CM d'origine (Figure 2.6).

Classe	Terme	Source
061	Dengue	ICD-9-CM
	Dengues	UMLS
	Fièvre dengue	UMLS
	Infection par le virus de la dengue	UMLS

Figure 2.6 : Définition de la classe « 061 » à l'aide de termes issus de ICD-9-CM et d'UMLS.

Lors de la construction des transducteurs, la méthode de normalisation choisie a été la racinisation (Snowball). Cette technique s'est imposée car elle est mieux adaptée que la lemmatisation (Treetagger) lorsque les textes contiennent un nombre important de termes techniques ou particuliers au domaine médical, pour lesquels aucune proposition de lemme n'est fournie.

En ce qui concerne la pondération des termes, la mesure TF.IDF n'a finalement pas été utilisée. Les tests que nous avons menés n'ont pas montré d'amélioration significative par rapport à la simple mesure de la fréquence. Une cause possible provient de la nature des textes, principalement constitués de termes techniques ou spécifiques au domaine. Or, l'effet discriminant du TF.IDF d'un terme technique important pour le codage peut disparaître si ce terme est utilisé de manière régulière à l'échelle du corpus.

Enfin, nous n'avons pas mené un test aussi complet que pour la première expérience. Les valeurs assignées aux multiplicateurs ont été choisies en fonction des valeurs optimales trouvées lors de la première expérience.

Résultats

L'évaluation a été réalisée sur 19.692 documents dont la codification par les documentalistes experts a servi de référence.

[50] Unified Medical Language System (`http://www.nlm.nih.gov/research/umls/`)
[51] Unique concept identifier (CUI).

Les résultats sont rapportés au tableau 2.2. La réduction de la liste des codes proposés pour un document est réalisée par une méthode de seuillage, dont la sélectivité peut être adaptée, selon que l'on veuille favoriser le rappel ou la précision. Évidemment, ces valeurs évoluent en sens inverse, un rappel élevé étant toujours accompagné d'une précision moindre. C'est la raison pour laquelle le rappel maximal est atteint lorsqu'aucune réduction de la liste de catégories n'est effectuée. Les détails concernant les différentes fonctions de seuil ne sont pas exposés dans le tableau. Seuls les meilleurs résultats, obtenus à l'aide de la fonction *middle weight* (voir Section 2.4.4), y sont repris.

	Rappel (R)	Précision (P)	F-mesure (F)	Nb. classes	Seuil
Meilleur rappel	**52,74**	20,69	27,37	19,6	Non
Meilleure F-mesure	37,97	30,30	**29,43**	9,8	Oui

Tableau 2.2 : Résultats des test de classification sur le corpus Médical.

Par rapport à la première expérience, les résultats obtenus sont inférieurs mais cependant encourageants. Ils montrent que la méthode peut être adaptée à d'autres cas qu'à celui qui a servi à son développement.

Plusieurs explications peuvent être avancées pour expliquer l'écart de performance. Tout d'abord, la présence d'une grande quantité de termes techniques, de jargon ou de néologismes, qui ne sont pas toujours repris dans la terminologie, peut handicaper le rappel.

Ensuite, la complexité des dénominations utilisées dans la ressource servant de base à la construction des transducteurs, ICD-9-CM, s'est avérée être assez élevée. Il ne s'agit en effet pas à proprement parler d'un thésaurus, mais d'une *classification*, qui constitue une sorte de guide pour aider les documentalistes lors de l'encodage. Par conséquent, les dénominations utilisées sont parfois très compliquées (beaucoup plus que les cas prévus initialement, voir section 2.3.2) ou se situent à un niveau conceptuel inadapté (trop général ou trop spécifique).

Enfin, rappelons qu'il a été établi par l'équipe de l'hôpital *Saint-Luc* que les lettres de sortie ne contiennent pas toujours suffisamment d'informations pour permettre leur encodage correct. D'autres documents du dossier médical du patient, dont fait partie la lettre de sortie, sont aussi parfois consultés par les documentalistes afin de formuler les codes adéquats.

2.5.4 Conclusion

La méthode *symbolique* de classification, qui a été proposée pour l'indexation de textes, constitue une approche originale. Celle-ci s'inscrit en porte-à-faux par rapport au courant prédominant constitué par les méthodes *statistiques* qui exploitent des techniques d'apprentissage artificiel. Elle ne se base pas sur une phase d'entraînement, et par conséquent, ne nécessite pas de corpus annoté. De plus, cette méthode ne rencontre pas de problème particulier pour gérer les catégories *rares*, pour lesquelles une grande quantité de données annotées peut difficilement être rassemblée.

Par rapport à d'autres systèmes similaires, c'est-à-dire mettant en œuvre une approche symbolique

basée sur l'utilisation d'une terminologie (voir section 2.2.2), la méthode que nous proposons se distingue de plusieurs manières. Premièrement, le problème de l'indexation de textes a été envisagé selon un angle différent de celui habituellement adopté. Plutôt que d'extraire des concepts potentiels des textes et de les confronter à la ressource terminologique (Névéol *et al.* [2005], Pereira *et al.* [2008], Shah *et al.* [2009], Aronson et Lang [2010]), le principe inverse a été préféré. Celui-ci consiste à partir de la ressource terminologique pour aller vers les textes. Cette démarche n'a pas la vocation d'extraire un maximum de concepts, mais elle permet au contraire de se concentrer sur les termes qui sont réellement utiles à l'indexation, en regard de la terminologie choisie. L'avantage majeur est bien entendu de ne pas perdre de temps à traiter des éléments *sans importance*[52], et permet ainsi de ne pas allonger inutilement le temps d'exécution du système.

L'utilisation de transducteurs pour l'extraction des expressions qui servent de base à l'indexation constitue également une originalité du système. Si ce formalisme a bien été utilisé par d'autres pour l'indexation (Névéol *et al.* [2006]), il n'était alors pas question de génération automatique de cette ressource d'extraction, mais bien d'élaboration manuelle à l'aide d'experts du domaine. Étant donné l'importance du temps généralement nécessaire à la création de ressources linguistiques, inconvénient souvent reproché aux méthodes symboliques, la mise au point d'un processus qui ne requiert que peu d'efforts humains devient dès lors très importante.

Enfin, l'application à des textes en français est un intérêt en soi. Même s'il est vrai que plusieurs systèmes ont déjà été proposés pour traiter cette langue (Névéol *et al.* [2005], Pereira *et al.* [2008]), force est de constater que les travaux portent souvent sur l'anglais. Certains de ceux-ci ne sont d'ailleurs pas adaptables à une autre langue (Aronson et Lang [2010]), ce qui n'est pas le cas de notre approche qui, si elle présente quelques aspects liés à la langue cible, peut tout à fait être étendue à d'autres langues.

Le prérequis à l'utilisation de notre méthode est constitué par l'existence ou la production d'une ressource terminologique adéquate. De telles ressources existent déjà dans divers domaines et de nombreuses entreprises et organisations ont déjà construit leur propre terminologie afin de pratiquer l'indexation manuelle de leurs documents.

Les résultats obtenus démontrent la faisabilité de la démarche dans un contexte semi-automatique. En effet, ce type de mise en œuvre offre l'avantage d'accélérer le processus d'analyse et de catégorisation des documents, d'en augmenter la cohérence tant entre les différents annotateurs qu'au cours du temps, tout en conservant la précieuse validation réalisée par les experts du domaine[53]. L'indexation de documents par rapport à un vocabulaire contrôlé peut donc être rendue plus abordable et plus homogène grâce à ce processus. Dès lors, le gain *sémantique* amené au cours de l'indexation, suite à l'utilisation d'une ressource terminologique particulière dans un contexte particulier, ne doit plus être rejetée au motif du coût représenté par l'analyse manuelle des documents. Cette constatation représente une point important permettant de progresser vers une indexation dans l'espace des concepts, et non plus des mots, ce qui nous semble être un des défis importants pour la recherche d'information,

[52] Du point de vue de l'indexation par rapport à la ressource terminologique en question.

[53] Dans un premier temps, il est imaginable d'indexer de manière complètement automatique afin d'obtenir directement une base de documents à interroger, et de procéder ensuite petit à petit à la validation de l'indexation.

tant pour les collections de documents en entreprise que sur Internet.

2.6 Amélioration des résultats par combinaison avec d'autres méthodes

La méthode (MLE) qui a été mise au point et qui a été exposée dans les sections précédentes a démontré ses qualités, mais présente également quelques inconvénients. Entre autres, l'indexation reste limitée aux éléments définis dans la ressource de départ, même si celle-ci est adaptée de façon à maximiser sa couverture. Dès lors, il est intéressant de déterminer si la combinaison de plusieurs systèmes peut permettre d'améliorer les performances, c'est-à-dire d'examiner le potentiel de complémentarité de ces méthodes.

D'une manière générale, il est possible d'imaginer deux modes de combinaison de notre méthode avec d'autres : en série (mode *collaboratif*) ou en parallèle (mode *concurrent*).

1. Pour le mode collaboratif, la phase d'extraction (section 2.4.2), réalisée à partir de transducteurs qui ont été générés automatiquement, peut servir de *prétraitement* à une autre méthode. Le but est dans ce cas de réduire le texte à un ensemble restreint de termes présentant un intérêt particulier. Cette étape est analogue à la phase de sélection de caractéristiques (*feature selection*) qui est souvent pratiquée en amont des méthodes d'apprentissage artificiel. La classification proprement dite, telle que nous l'avons exposée aux sections 2.4.3 et 2.4.4 n'est donc pas réalisée dans ce cas.

2. Dans le mode concurrent, les différentes méthodes à combiner sont exécutées indépendamment les unes des autres et produisent toutes une liste de résultats pondérés pour chaque document à analyser. Ces listes sont ensuite rassemblées et comparées afin d'en obtenir une meilleure (« plusieurs avis valent mieux qu'un seul »).

Les premiers tests que nous avons réalisés ont été menés avec l'objectif de comparer et combiner les résultats déjà obtenus, avec un algorithme de type SVM[54] (*Support Vector Machine*, Cortes et Vapnik [1995]). Le mode *collaboratif* n'a pas fait l'objet de recherches poussées car il n'a pas montré une réelle amélioration lors des premières expériences[55]. Par contre, le mode de combinaison *concurrent* a été approfondi et a donné de nouveaux résultats sur le corpus *Parlementaire* (Section 2.6.2). La démarche a ensuite été réitérée sur le corpus *Médical* (Section 2.6.3), mais cette fois en combinaison avec une autre méthode symbolique principalement basée sur l'analyse morphologique des termes[56].

[54] Les développements et tests relatifs à SVM ont été réalisés, au sein du projet Stratego, par Amin Mantrach.

[55] Il n'est cependant pas impossible qu'une recherche plus poussée mette en avant le potentiel de ce mode, qui intuitivement semble pouvoir apporter une plus value.

[56] Cette méthode est due à Julia Medori.

2.6.1 Principes de combinaison en mode concurrent

Il est important que les méthodes de classification que l'on combine se différencient par certaines caractéristiques, de manière à fournir des résultats potentiellement complémentaires. Par exemple, les systèmes utilisés peuvent mettre en œuvre des approches différentes (statistiques, symboliques) ou se baser sur des caractéristiques qui se situent à des niveaux différents (termes composés, mots simples, morphèmes, etc.) et qui exploitent des indices qui ne sont pas exactement de même nature.

Plus concrètement, avant l'application d'une fonction de seuil, les différentes méthodes peuvent fournir des listes de résultats de longueur différentes : certaines méthodes fournissent une liste qui compte toujours autant d'éléments qu'il y a de catégories, alors que d'autres renvoient des listes ne contenant qu'un nombre restreint et variable de classes. Dans ces listes, la répartition des poids de catégories peut aussi être fort différente, ce qui représente une opportunité de modifier l'ordre final des catégories dans la liste de résultats[57].

La combinaison des résultats des méthodes de classification, se fait par fusion des listes de classes pondérées et ordonnées. Deux modes de combinaison sont envisagés : le premier correspond à l'union des résultats et le second à leur intersection. En principe, l'intersection favorise plutôt la précision au détriment du rappel, alors que l'union provoque les variations inverses. La combinaison peut être réalisée sur les listes préalablement réduites à l'aide d'une fonction de seuil (voir section 2.4.4) ou sur les listes complètes. Dans ce dernier cas, un seuil peut être appliqué sur la nouvelle liste combinée. La première approche revient donc à fusionner par union (Mix1) ou intersection (Mix2) les listes complètes fournies par les classificateurs avant d'appliquer la fonction de seuil dynamique pour obtenir la sélection finale. La seconde approche consiste à prendre le meilleur résultat obtenu par chaque méthode[58] et à effectuer leur union (Mix3) ou leur intersection (Mix4).

Les poids fournis par les listes de résultats des différentes méthodes ont été normalisés afin de toujours manipuler des valeurs comprises entre 0 et 1. Lors de l'union ou l'intersection des deux listes, pour chaque document et chaque catégorie, les valeurs sont additionnées en appliquant un facteur de pondération qui permet de donner plus ou moins d'importance à chaque méthode : $poids_{combine} = \alpha * poids_{methode1} + (1 - \alpha) * poids_{methode2}$ avec $0 \leq \alpha \leq 1$. Lors des différents tests, nous avons fait varier ces multiplicateurs entre 0,1 et 0,9 (avec des sauts de 0,1) en prenant soin que la somme des deux multiplicateurs soit toujours égale à 1. À noter que, pour Mix3 et Mix4, cette pondération n'a pas d'effet car elle intervient après la réduction des listes par application du seuil.

2.6.2 Expérience 1 : SVM, sur corpus *Parlementaire*

Les listes de classification produites par SVM[59] ont été réutilisées telles quelles. Les détails de son développement ne seront par conséquent pas exposés dans ces pages. Il est cependant intéressant de

[57] L'ordre n'est pas intéressant en tant que tel, mais prend toute son importance lorsque la réorganisation des catégories a lieu avant la réduction de la liste par application d'un seuil.

[58] Par meilleur résultat, nous entendons celui qui maximise la f-mesure (F).

[59] L'algorithme a été mis au point par Amin Mantrach dans le cadre du projet Stratego.

préciser que l'algorithme SVM a été entraîné sur l'ensemble des mots simples, moins les mots vides (*stopwords*). Les termes ont été racinisés (à l'aide de Snowball). La classification multiclasse est en fait constituée par un ensemble de classificateurs *un-contre-un*. Pour N classes, le système apprend N*(N-1)/2 modèles (par exemple pour 4 classes, il y a 6 modèles à exécuter ; pour les 47 classes du cas d'étude, il y en a donc 1081). Le poids attribué au final à chaque classe est le nombre de duels gagnés. Une présentation plus précise est consultable dans Kevers *et al.* [2010].

La combinaison de SVM à MLE est née suite à l'observation des résultats obtenus pour ces deux méthodes. Globalement, les performances sont assez similaires[60], mais celles-ci sont obtenues sur la base de listes de classification partiellement différentes. En plus de leur contenu, elles varient également sur deux caractéristiques : leur longueur et la distribution des poids. En ce qui concerne le nombre d'éléments des listes, SVM renvoie toujours une liste contenant l'ensemble des catégories tandis que MLE livre un résultat de longueur variable. Pour la répartition des poids dans les listes, la méthode SVM génère une pondération qui décroît de manière très progressive, alors que la méthode MLE donne parfois lieu à des sauts brusques. Les premières catégories ont ainsi souvent un poids beaucoup plus élevé que le reste de la liste.

Cette expérience de combinaison a été menée sur le corpus *Parlementaire*. Les catégories utilisées ont été restreintes aux 47 microthésaurus[61].

La synthèse des résultats obtenus pour les méthodes MLE et SVM, ainsi que pour les différentes approches et modes de combinaison de ces deux méthodes, est reprise au tableau 2.3. Nous ne rapportons que le meilleur résultat parmi les combinaisons possibles de multiplicateurs (α, $1 - \alpha$).

On constate que, à l'exception de l'intersection des listes réduites par application d'un seuil (Mix4), les autres combinaisons débouchent sur des résultats supérieurs. La meilleure performance en termes de f-mesure est 66,08 (pour un rappel de 67,70% et une précision de 73,70%) et est atteinte en réalisant l'union des listes complètes fournies par les deux méthodes avant application de la fonction de seuil (Mix1). Cela représente une augmentation de 5,06 par rapport à la méthode MLE et de 6,93 par rapport à SVM.

La détérioration des résultats observés pour Mix4 peut s'expliquer par le faible nombre de catégories fournies par MLE (1,61) et SVM (1,05), qui lors de l'intersection, atteint un niveau très bas (0,67). Certains documents ne reçoivent donc pas de suggestion de catégorie ce qui pénalise fortement le rappel. Le choix d'une unique mauvaise catégorie affecte aussi fortement la précision.

Mix3, avec une augmentation du rappel (à 76,16%) par rapport aux deux méthodes de base et une précision se stabilisant (à 64,96%) légèrement en dessous de MLE, se comporte comme prévu et permet d'atteindre une f-mesure plus élevée (65,13). La forte augmentation du rappel prouve que les catégories correctes comprises dans les deux listes sont en partie différentes.

Avec Mix1 et Mix2, la pondération différente des deux méthodes, modifie les poids initiaux des ca-

[60] La méthode MLE a donné des résultats légèrement plus élevés que ceux obtenus avec SVM (voir tableau 2.3).

[61] La quantité de données à notre disposition n'était pas suffisante pour effectuer un apprentissage satisfaisant sur l'ensemble des 2.514 catégories.

tégories et donne l'opportunité à celles-ci de se réorganiser avant application du seuil. On remarque que les meilleurs résultats sont obtenus à l'aide d'une pondération forte de la méthode SVM. L'ordre et les poids attribués par cette méthode ont donc une grande importance sur le résultat final. L'union (Mix1) et l'intersection (Mix2) atteignent un niveau similaire de f-mesure, supérieur aux méthodes de base. C'est une nouvelle fois l'union qui réalise la meilleure performance (66,08), alors que l'intersection suit de très près (66,01). Les résultats présentent un bon niveau de précision, ce qui est dû à la forte pondération de SVM. Comme prévu, l'union favorise plutôt le rappel (ici, peu élevé en raison du faible nombre de catégories présentées, soit 1,48) et l'intersection, la précision (en partie grâce à l'effet *filtre* de MLE).

	Rapport MLE/SVM	Rappel (R)	Précision (P)	F-mesure (F)	Nbr. de catégories
Méthodes de base					
MLE (Max F)	n/a	64,79	66,05	**61,02**	1,61
SVM (Max F)	n/a	53,93	72,90	**59,15**	1,05
Mix1 : Seuil(MLE ∪ SVM)					
Max F	0,1 / 0,9	67,70	73,70	**66,08**	1,48
F≈0,5 & Max R	0,3 / 0,7	**87,16**	50,06	63,59	3,46
F≈0,3 & Max R	0,6 / 0,4	**91,72**	33,50	49,07	4,77
Mix2 : Seuil(MLE ∩ SVM)					
Max F	0,1 / 0,9	70,20	71,31	66,01	1,62
Mix3 : Seuil(MLE) ∪ Seuil(SVM)					
n/a	n/a	76,16	64,96	65,13	1,99
Mix4 : Seuil(MLE) ∩ Seuil(SVM)					
n/a	n/a	42,57	57,24	46,81	0,67

Tableau 2.3 : Synthèse des résultats obtenus pour les méthodes MLE et SVM, ainsi que pour les différents approches et modes de combinaison.

Dans l'optique de l'indexation semi-automatique, nous pourrions augmenter le nombre de catégories proposées au documentaliste afin d'améliorer le rappel. Avec la meilleure méthode combinée (Mix1), nous pourrions ainsi proposer en moyenne 3,46 catégories pour atteindre un rappel de 87,16% (précision de 50,06%). En acceptant de laisser chuter la précision à 33,50%, et en suggérant en moyenne 4,77 catégories, le rappel pourrait même augmenter jusqu'à 91,72%. La mise en avant du rappel s'accompagne d'un renversement progressif de la pondération vers la méthode MLE (0,3/0,7 dans un premier temps et 0,6/0,4 ensuite), qui démontre donc son apport sur ce point.

Nous avons également évalué, document par document, dans quelle proportion la meilleure méthode combinée (Mix1) offre une f-mesure plus élevée que les deux méthodes de base. Pour MLE, on constate au tableau 2.4 que les résultats restent inchangés pour 56,31% des documents, et que 15,03% subissent une détérioration de la f-mesure (en moyenne -39,21) alors que 28,65% bénéficient d'une meilleure analyse (en moyenne +38,26). Le résultat est donc meilleur ou inchangé pour 84,96% des documents. En ce qui concerne la méthode SVM (voir tableau 2.5), le nombre et la répartition des documents concernés par des variations sont assez semblables. On note cependant une proportion un peu plus grande de documents sans changement (61,78%). Les variations de performances sont un

peu plus importantes, surtout à la hausse (+49,86 dans 24,74% des cas) mais aussi à la baisse (-40,06 dans 13,48% des cas). Au total, les résultats sont meilleurs ou inchangés pour 86,52% des textes.

On constate que par rapport à MLE, l'augmentation des performances apportée par Mix1 vient principalement d'une amélioration de la précision. La même comparaison effectuée par rapport à SVM montre au contraire un gain au point de vue rappel. Ces résultats confirment que les deux méthodes présentent des caractéristiques en partie différentes. Par conséquent, leur combinaison permet d'améliorer la performance finale du système. Nous avons en effet obtenu des gains significatifs en atteignant une f-mesure de 66,08 (+5,06 pour MLE et +6,93 pour SVM).

Variation F-mesure	Amélioration (>)					Egal (=)			Déterioration (<)				
Variation Rappel	>			=	<	>	=	<	>	=	<		
Variation Précision	>	=	<	>	>	<	=	>	<	<	>	=	<
Nb. docs.	1.116	243	47	1.781	10	3	6.257	23	1	541	136	446	553
%	10,00	2,18	0,42	15,96	0,09	0,03	56,08	0,21	0,01	4,85	1,22	4,00	4,96
Total	3.197 (28,65%)					6.283 (56,31%)			1.677 (15,03%)				
Variation moyenne	58,09	31,27	14,15	27,57	10,90	n/a	n/a	n/a	9,53	27,98	15,22	31,00	63,06
f-mesure	38,26					n/a			39,31				

Tableau 2.4 : Analyse des variations de performance (> : amélioration, < : déterioration, = : égal) entre la méthode MLE et l'approche Mix1.

Variation F-mesure	Amélioration (>)					Egal (=)			Déterioration (<)				
Variation Rappel	>			=	<	>	=	<	>	=	<		
Variation Précision	>	=	<	>	>	<	=	>	<	<	>	=	<
Nb. docs.	1.551	880	199	130	0	11	6.882	0	1	1.193	9	36	265
%	13,90	7,89	1,78	1,16	0	0,10	61,68	0	0,01	10,69	0,08	0,32	2,37
Total	2.760 (24,74%)					6.893 (61,78%)			1.504 (13,48%)				
Variation moyenne	65,96	32,47	14,56	29,37	0	n/a	n/a	n/a	5,00	30,81	14,55	31,55	83,85
f-mesure	49,86					n/a			40,06				

Tableau 2.5 : Analyse des variations de performance (> : amélioration, < : déterioration, = : égal) entre la méthode SVM et l'approche Mix1.

2.6.3 Expérience 2 : analyse morphologique, sur corpus *Médical*

Cette expérience combine la méthode MLE à une autre méthode symbolique[62] (MA), décrite dans Kevers et Medori [2010]. Cette méthode se démarque de la notre par sa capacité à prendre en compte, dans le processus de classification, les morphèmes qui composent les mots.

La méthode MA procède en premier lieu à une extraction sur la base de transducteurs construits manuellement pour l'analyse de textes médicaux. Aux éléments extraits sont ensuite ajoutés les morphèmes qui les composent ainsi que les sens de ces derniers. Une mesure de similarité est alors calculée entre cette liste et la représentation de chaque classe, qui a été créée de manière similaire au préalable. L'avantage de cette méthode est de pouvoir rapprocher des termes sémantiquement proches mais morphologiquement éloignés, comme « fibroscopie bronchique » et « bronchoscopie par fibre optique ». Cet exemple est illustré à la figure 2.7, dans laquelle les éléments communs sont présentés en gras et les *stopwords* sont biffés.

Fibroscopie bronchique		Bronchoscopie par fibre optique	
fibroscopie	(mot)	bronchoscopie	(mot)
fibr-	(morphème)	**bronch-**	(morphème)
fibre	(sens)	**bronche**	(sens)
-scopie	(morphème)	**-scopie**	(morphème)
bronchique	(mot)	~~par~~	(mot)
bronch-	(morphème)	**fibre**	(mot)
bronche	(sens)	optique	(mot)
~~-ique~~	(morphème)		

Figure 2.7 : Exemple de décomposition morphologique et sémantique en vue du calcul de similarité.

Les deux méthodes fonctionnent indépendamment l'une de l'autre. Leurs résultats[63] respectifs sont présentés au tableau 2.6. Ici aussi, les différences affichées par les deux méthodes plaident en faveur de leur combinaison. En effet, le mode d'analyse et de représentation des textes est basé, dans les deux cas, sur des principes différents. De plus, comme pour SVM, la liste de résultats fournie par MA contient toujours l'ensemble des classes, au contraire de MLE qui en retourne un nombre variable.

Les résultats obtenus pour la combinaison de ces deux méthodes sont repris au tableau 2.6[64]. Certaines approches combinées améliorent de manière assez claire les résultats obtenus avec les méthodes de base. La meilleure performance est atteinte par Mix1, qui réalise l'union des listes complètes, avant application éventuelle d'un seuil.

Le meilleur rappel est atteint avec Mix1 et s'établit à 60,21%, ce qui représente une amélioration à la fois par rapport à MA (46,13%, +14,08) et à MLE (52,74%, +7,47). Comme pour les deux méthodes de base, ce résultat est obtenu lorsque toutes les catégories de la liste de résultats sont conservées,

[62] Cette méthode a été développée et mise au point par Julia Medori dans le cadre du projet Capadis.

[63] Pour MLE, les résultats sont identiques à ceux présentés à la section 2.5.3. En ce qui concerne la méthode MA, l'évaluation a été effectuée en suivant la même méthodologie que pour MLE (Section 2.5).

[64] L'implémentation de Mix3 et Mix4, c'est-à-dire la fusion des meilleurs résultats obtenus par MLE et SVM, a pour conséquence qu'il n'est pas possible d'obtenir plusieurs points permettant de favoriser le rappel ou la précision.

	Rapport MA/MLE	Rappel (R)	Précision (P)	F-mesure (F)	Nbr. de catégories	Seuil utilisé ?
Méthode MA						
Max R	n/a	**46.13**	14.70	21.10	20	Non
Max F	n/a	34.52	27.34	**28.00**	8.6	Oui
Méthode MLE						
Max R	n/a	**52.74**	20.69	27.37	19.6	Non
Max F	n/a	37.97	30.30	**29.43**	9.8	Oui
Mix1 : Seuil(MA ∪ MLE)						
Max R	Tous	**60,21**	13,20	20,86	30,5	Non
Max F	0,3 / 0,7	37,13	33,12	**31,64**	8,1	Oui
Mix2 : Seuil(MA ∩ MLE)						
Max R	Tous	38,66	29,28	30,52	9,1	Non
Max F	0,3 / 0,7	34,73	34,55	31,50	7	Oui
Mix3 : Seuil(MA) ∪ Seuil(MLE)						
n/a	n/a	43,28	20,59	27,90	14,7	Oui
Mix4 : Seuil(MA) ∩ Seuil(MLE)						
n/a	n/a	24,07	37,95	29,46	4,4	Oui

Tableau 2.6 : Synthèse des résultats obtenus pour les méthodes MA et MLE, ainsi que pour les différents approches et modes de combinaison.

c'est-à-dire 30,5 en moyenne (soit environ 10 de plus qu'avec MA ou MLE). Parmi les catégories fournies par Mix1, 64,21% sont proposés par les deux méthodes de départ et les 35,79% restants uniquement par l'une ou par l'autre. Cette constatation tend à confirmer une certaine complémentarité entre les méthodes MA et MLE.

En s'établissant à 31,64 pour Mix1, l'amélioration la plus importante de la f-mesure n'est pas aussi tranchée mais dépasse cependant les résultats respectifs de MA (28.00, +3.64) et MLE (29.43, +2.21). Ce résultat est principalement dû à l'augmentation de la précision pour les deux méthodes (+5,78 pour MA et +2,82 pour MLE) alors que, dans le même temps, le rappel se maintient à un niveau légèrement inférieur pour MLE (-0.84) et s'améliore pour MA (+2.61). La conséquence de cette précision améliorée s'observe dans une diminution du nombre de catégories proposées qui est réduit à 8,1 alors qu'il atteignait 8,6 pour MA et 9,8 pour MLE.

Les performances proposées par Mix2, qui ne retient que les catégories appartenant à l'intersection des listes des deux méthodes, sont légèrement en retrait par rapport à Mix1. Le rappel subit les conséquences de ce mode de fusion, les catégories correctes ne pouvant plus être issues que de la portion commune aux deux méthodes (64,21% des catégories de départ). Cela explique que Mix2 obtient des valeurs de rappel généralement inférieures à MA et MLE. Cette détérioration est cependant en partie compensée par une amélioration de la précision consécutive à l'effet de *filtre* de la fusion par intersection. La meilleure f-mesure (31,50) atteint ainsi un résultat assez similaire à celui obtenu pour Mix1, mais en proposant un équilibre différent entre le rappel (34,73%) et la précision (34,55%). Cette orientation vers la précision se traduit également dans le nombre moyen de catégories proposées qui chute à 7.

En ce qui concerne les deux dernières tentatives de combinaison, Mix3 s'est avéré être moins performant que MA et MLE, alors que Mix4 n'a apporté que peu d'améliorations par rapport à ces deux

méthodes.

Enfin, une rapide analyse des codes *rares*, c'est-à-dire ceux qui sont utilisés moins de six fois dans tout le corpus de test, montre que 35% de leurs occurrences (212 sur 603) sont reprises dans la liste non réduite obtenue grâce à l'union des deux méthodes (Mix1).

Dans cette seconde expérience, si les méthodes employées sont toutes deux symboliques, elles n'en présentent pas moins des différences importantes, à la fois au niveau des caractéristiques utilisées pour représenter les documents que dans la manière d'attribuer un score aux différentes catégories. Cela confirme donc l'intérêt de mettre en commun les résultats de plusieurs méthodes de classification.

2.6.4 Conclusion

Les deux expériences ont souligné qu'il est tout a fait pertinent de combiner plusieurs méthodes de classification, du moins si celles-ci présentent des caractéristiques différentes, tant au niveau de la nature des éléments exploités (termes simples ou composés, morphèmes, éléments sémantiques) que de la manière de les manipuler (différents modèles : booléens, probabilistes, vectoriels, etc.).

Parmi les différentes modalités de combinaisons testées, Mix1 est la solution qui amène généralement les meilleurs résultats. Celle-ci consiste à rassembler de manière pondérée les listes complètes de résultats, avant application éventuelle d'un seuil, ce qui offre des possibilités d'amélioration à la fois pour le rappel et la précision. En effet, cette manière de procéder permet de tirer avantage d'une méthode qui présenterait un rappel significativement plus élevé tout en laissant la possibilité aux poids des catégories d'évoluer et ainsi de réorganiser leur ordre. Ce dernier élément est important en vue de l'obtention du niveau de précision recherché. Cet objectif est en effet atteint à l'aide de la réduction de la liste par application d'un seuil. D'une manière générale, il est évidemment intéressant de combiner une méthode performante en rappel à une autre plutôt axée sur la précision.

2.7 Perspectives

Diverses pistes d'analyse peuvent être suggérées afin d'obtenir une vision plus précise et nuancée des erreurs commises par le système. Ainsi, le rappel pourrait être analysé classe par classe plutôt que document par document. Il serait alors possible d'examiner les classes pour lesquelles le rappel est élevé, ou au contraire, plutôt faible, ou encore de déterminer s'il y a une corrélation entre la fréquence d'attribution manuelle des classes – par les documentalistes – et automatique – par le système. Ce type de données pourrait aider à révéler les principales difficultés rencontrées par le système et pour lesquelles des améliorations pourraient être apportées.

Les différences de comportement entre MLE et SVM pourraient également être une source intéressante d'informations. En effet, si leurs performances sont relativement comparables, nous avons montré à la section 2.6.2 que les résultats sont partiellement différents. Une comparaison classe par classe pourrait à nouveau amener un éclairage sur les différences qui existent entre les deux systèmes.

Au delà des possibilités de combinaison avec d'autres méthodes de classification, diverses autres perspectives ont été identifiées et viennent compléter les aspects déjà développés pour notre méthode (MLE).

En premier lieu, il est important d'évoquer les possibilités d'enrichissement automatique de la ressource terminologique, qui devraient permettre d'augmenter la couverture et le rappel. Une piste particulièrement intéressante à suivre est celle de la recherche automatique de synonymes. Diverses méthodes peuvent être mises en œuvre et combinées pour atteindre cet objectif. La plus simple et la plus évidente est celle qui consiste à exploiter d'autres ressources terminologiques. Par exemple, le thésaurus de la Communauté européenne, Eurovoc, peut certainement venir combler certaines lacunes du thésaurus qui a été utilisé lors de l'expérience sur le corpus *Parlementaire*. La consultation d'Eurovoc, éventuellement à l'aide de méthodes autorisant les recherches approximatives, permettrait par exemple d'ajouter à l'expression « durée du travail » le synonyme « temps de travail » et l'expression apparentée « semaine de x heures »[65]. Les ressources plus spécialisées, comme la base de données toponymiques utilisée lors de nos expériences, peuvent également venir renforcer certains domaines particuliers du lexique. Enfin, on ne peut s'empêcher de penser que des ressources telles que Wikipedia pourraient également servir de sources de synonymes, entre autres par le biais des redirections de pages[66]

Le deuxième axe de recherche identifié pour l'enrichissement de la ressource de base consiste en l'exploitation des formes verbales. En effet, très souvent le contenu des ressource terminologiques privilégient l'usage de noms ou d'expressions nominales au détriment des verbes et des autres catégories grammaticales. Ce choix est assez logique car le nom est le vecteur par excellence pour exprimer un concept quel qu'il soit. Au contraire, certains verbes peuvent être considérés comme des *stopwords* car ils ne véhiculent aucun sens particulier. Cependant, il est souvent possible qu'un concept, désigné de manière canonique par une forme nominale, puisse être évoqué au moyen d'une forme verbale. Par exemple, le groupe nominal « épuration des eaux usées » peut apparaître sous la forme « épurer les eaux usées ». Ce type de stratégie d'extension de la ressource d'extraction a été testée mais n'a finalement pas été retenue car elle offrait peu ou pas d'amélioration sur le corpus *Parlementaire*. La transformation de groupes nominaux en groupes verbaux a été réalisée à l'aide de Verbaction[67], un lexique de noms d'actions morphologiquement apparentés à des verbes. Le mécanisme mis en place consiste simplement à ajouter à une forme nominale présente dans ce lexique la forme verbale correspondante. Dans le transducteur, cette dernière est exprimée sous la forme d'un élément lemmatisé (l'infinitif), ce qui permet de reconnaître toutes les formes conjuguées du verbe. Cette extension est évidemment largement tributaire de la couverture de la ressource sous-jacente. Son amélioration pourrait donc éventuellement nous amener à en réexaminer l'utilisation.

Enfin une troisième et dernière méthode, à base de corpus, pourrait venir compléter avantageusement la ressource terminologique de départ. L'idée générale part de l'hypothèse distributionnelle de Harris

[65] La recherche « durée du travail » donne deux résultats : « durée du travail » et « durée légale du travail » qui sont respectivement liés aux expressions « temps de travail » et « semaine de x heures ».

[66] La recherche « durée du travail » est en pratique redirigée sur la page intitulée « temps de travail ».

[67] `http://w3.erss.univ-tlse2.fr:8080/index.jsp?perso=hathout&subURL=verbaction/main.html`

[1954] pour qui les expressions qui ont un même sens apparaissent souvent dans un même contexte. Le principe peut s'appliquer aux mots simples et composés. Une simulation manuelle d'une procédure relevant de ce principe a par exemple permis, pour l'expression « droit communautaire », de trouver des similitudes avec d'autres termes composés, tels que « règle communautaire », « règlement communautaire » ou encore « législation communautaire ».

En ce qui concerne la précision offerte par la classification, nous avons souligné que l'ambiguïté de certains termes provoque parfois de nombreuses erreurs et vient donc diminuer les résultats. Il est cependant possible de traiter ces cas, pour autant qu'on en ait conscience. Le problème est classique et a déjà été mis en évidence par des systèmes similaires. Par exemple, Aronson et Lang [2010] propose une méthode de désambiguïsation entre les termes du métathésaurus UMLS. Dans le cadre de notre système, basé sur des transducteurs de reconnaissance, l'ambiguïté se situe plutôt entre les termes du thésaurus et les mots des textes analysés (le concept « ART » ambigu avec l'abréviation « art » de « article », le concept « MAÏS » ambigu avec le mot « mais », etc.). Pour traiter ce type de cas, une procédure de détection de la polysémie capable de pointer les endroits problématiques offrirait la possibilité d'une intervention humaine ciblée, permettant ainsi une résolution semi-automatique de l'ambiguïté.

Ensuite, au delà du travail sur la ressource de départ, un certain nombre de points relatifs à la classification en elle même laissent de la place à des recherches ultérieures plus poussées. C'est par exemple le cas des procédés de pondération et des méthodes de seuil. L'organisation hiérarchique du thésaurus pourrait aussi être prise en compte lors de la classification, et les divers traitements nécessaires à la construction des transducteurs pourraient encore être étendus.

Enfin, plusieurs pistes de combinaison de cette méthode avec d'autres sont envisageables, que ce soit de manière concurrente ou collaborative. Ainsi, l'extraction d'expressions réalisée à l'aide des transducteurs pourrait par exemple servir de base à une méthode d'apprentissage artificiel.

Deuxième partie

Extraction d’informations temporelles et indexation thématique à dimension temporelle

Le fil du temps est couvert de noeuds.
Gaston Bachelard – La dialectique de la durée

« Quelle singulière montre ! » dit-elle.
« Elle marque le quantième du mois, et ne marque pas l'heure qu'il est ! »
« Et pourquoi marquerait-elle l'heure ? » murmura le Chapelier.
« Votre montre marque-t-elle dans quelle année vous êtes ? »
Lewis Carroll – Les Aventures d'Alice au pays des merveilles

Le temps de la réflexion est une économie de temps.
Publius Syrus

CHAPITRE 3

LA NOTION DE TEMPS

3.1 Introduction

Dans cette deuxième partie, c'est le problème du traitement automatique de l'information temporelle qui est abordé. Cette tâche consiste, dans les textes en langage naturel, à repérer les éléments qui véhiculent ce type d'information, à les interpréter afin de leur donner une valeur univoque dans l'espace du temps, et, finalement, à en fournir une représentation normalisée. Le résultat d'une telle analyse est important car il apporte la possibilité d'exploiter plus facilement et de manière plus complète les informations temporelles. Celles-ci apparaissent dans de nombreux types de textes, quel que soit le thème abordé. Bien qu'elles puissent être considérées comme des données à part entière, les informations temporelles constituent souvent une dimension particulière en rapport avec une autre information (statut de *métadonnée* par rapport à celle-ci). Nous considérons que les références temporelles constituent des éléments particulièrement intéressants pour l'indexation de textes, et c'est dans ce but qu'elles sont exploitées au chapitre 8. Cependant, que ce soit dans le cadre de la recherche ou de l'extraction d'informations, et que le but soit l'amélioration de l'accès aux documents ou tout autre objectif, de nombreuses applications peuvent tirer profit de l'analyse temporelle (construction de bases de connaissances à partir de corpus de textes non structurés, résumé automatique, la traduction automatique, etc).

Alors que l'aspect thématique, qui a été présenté à la partie I, représente une dimension très variable de l'information, la dimension temporelle est, elle, beaucoup plus *stable*, que ce soit du point de vue de son expression qu'en ce qui concerne sa distribution dans les différents textes[1]. Dès lors, là où la prise en compte des informations thématiques exigeait une méthode adaptable en fonction du domaine, le traitement de l'information temporelle reste une tâche assez peu variable. Cette caractéristique justifie un investissement beaucoup plus important, entre autres dans des ressources linguistiques *ad-hoc*, car celles-ci sont réutilisables.

Après l'introduction à la notion de temps qui est donnée par le présent chapitre, un tour d'horizon des travaux concernant l'information temporelle dans le langage naturel est réalisé. Celui-ci s'organise en trois chapitres, consacrés à différentes approches de la question : le point de vue linguistique

[1] Les expressions temporelles sont employées dans de très nombreux textes, alors que les informations relatives à un certain thème n'apparaissent que dans un ensemble de documents spécifiques.

(Chapitre 4), la question de la modélisation du temps (Chapitre 5), et finalement l'extraction d'informations (Chapitre 6). Une fois ces fondations posées, le chapitre 7 présente les réalisations concrètes effectuées, dans le cadre de cette thèse, en matière de traitement automatique de l'information temporelle. Finalement, le chapitre 8 réunit les apports de la première partie à ceux obtenus pour les aspects temporels afin de montrer comment ces éléments peuvent conjointement améliorer la représentation sémantique des documents de manière à en améliorer l'accès.

3.2 La notion de temps

De manière générale, tout événement est confronté d'une façon ou d'une autre à des informations temporelles et peut être caractérisé en fonction d'un calendrier ou d'un horaire. Les concepts temporels que nous manipulons, sans même plus y penser, et avec tant de facilité, représentent pourtant, à divers points de vue, une notion fondamentale. Ces concepts sont tellement intégrés dans notre perception du monde qu'il est parfois difficile de se rappeler qu'ils constituent un système de représentation et non une réalité en tant que telle. Il existe par contre, bien évidemment, un lien entre les concepts temporels et le monde réel.

Les notions de temps que nous utilisons proviennent principalement de divers phénomènes naturels cycliques. Ce type de phénomène possède une caractéristique intéressante qui consiste à retrouver son état initial après le passage successif dans divers états intermédiaires. Lorsque le cycle est assez régulier, il possède une propriété remarquable : à partir d'un événement quelconque, il est possible de compter le nombre de fois que le cycle s'opère jusqu'à un autre événement[2]. Il s'agit donc d'un moyen de matérialiser et de mesurer le temps qui passe, en termes de nombre de cycles.

Concrètement, un certain nombre de ces phénomènes ont effectivement servi de base à l'élaboration de concepts temporels. La rotation de la Terre autour de son axe implique une alternance de périodes d'obscurité et de clarté. Ce phénomène a donné naissance au concept de *jour*. Il s'agit du temps nécessaire à une rotation complète de la Terre sur son axe. La succession des saisons provoquées par les révolutions de la Terre autour du Soleil a donné naissance à la notion d'*année*. Enfin, la révolution de la Lune autour de la Terre, et l'alignement Lune-Terre-Soleil, a lui donné lieu à l'observation des différents *quartiers de lune* dont découle le concept de *mois*. Notons que si ces phénomènes sont effectivement cycliques, leur réalisation n'est pas toujours tout à fait constante en raison de l'interaction avec divers autres phénomènes astronomiques.

Au cours de l'histoire de nombreux systèmes de calendriers ont été utilisés. Ceux-ci sont en rapport avec les phénomènes astronomiques que nous venons de citer. Pour les calendriers lunaires, le mois se calque sur la période d'une lunaison alors que pour les calendriers solaires une année doit correspondre à une révolution complète de la Terre autour du Soleil. Le calendrier communément utilisé dans une grande partie du monde est le calendrier grégorien, qui est solaire.

[2] « [...] Galilée a découvert que le pendule était une bonne horloge. La légende dit que, dans l'église de Pise, observant un grand chandelier suspendu osciller lentement, il a compté le nombre de ses battements cardiaques entre chaque oscillation. Comme c'était toujours la même, il en a conclu que le pendule est une bonne façon de mesurer le temps. Depuis, la plupart des horloges utilisent un pendule. » (Le Meur [2010])

De ces concepts de base découlent toute une série d'autres notions, telles que celles de *passé*, *présent* et *futur*. Il faut cependant remarquer que la conceptualisation du temps est culturellement variable. Núnez et Sweetser [2006], cité par Girault [2007], exposent le cas du peuple Aymaras pour qui le temps est *inversé*. Ils considèrent en effet que le passé est *devant*, et le futur *derrière*, ce qui a des répercussions au niveau de leur langue. Une expression telle que « nayra mara », qui signifie « l'année dernière », est traduite littéralement par « l'année devant ».

Un autre exemple peut être trouvé dans l'interprétation de la mort qu'ont les différentes religions. Alors que certains considèrent le décès comme la fin de la période de vie, qui court donc sur un intervalle de temps fini, cet événement est très souvent perçu d'une autre façon. De nombreuses religions ou cultures considèrent, d'une manière ou d'une autre, une certaine continuité de l'être. C'est entre autres le cas de la croyance en la réincarnation, très répandue dans les religions orientales et asiatiques. Cette manière d'envisager la mort amène une vision moins finie du temps.

De même, l'expression des notions de temps dans les différentes langues ne passe pas par les mêmes mécanismes linguistiques, comme cela a été montré par divers travaux en apprentissage des langues, par exemple par Paprocka-Piotrowska et Demagny [2004]. Ceux-ci ont observé que l'acquisition de la maîtrise des aspects temporels d'une langue est très différente selon qu'il s'agit d'un apprenant natif (enfant) ou d'un apprenant d'une langue seconde :

> « l'apprenant d'une langue seconde, qui a acquis les concepts fondamentaux dans l'enfance, n'a pas à effectuer ce travail cognitif important mais doit aussi en quelque sorte déconstruire la grammaire de sa langue maternelle afin d'être au plus proche du discours de la L2 en cours d'acquisition. » (Paprocka-Piotrowska et Demagny [2004], p. 73)

De plus, on peut remarquer que, comparé au français ou à l'anglais qui disposent de nombreux temps verbaux pour véhiculer des informations temporelles et marquer la localisation chronologique, d'autres langues sont moins bien dotées à cet égard. En japonais par exemple, il existe seulement deux temps, le passé et le non-passé (utilisé à la fois pour le présent et le futur). L'expression du temps passe alors par d'autres mécanismes ou registres du langage.

La perception du temps a également évolué au cours des années et des siècles. On peut remarquer qu'il y a un lien fort entre la finesse et la précision des notions de temps utilisées et le niveau de développement technologique. La division du jour en heures a commencé à être utilisée aux environs du moyen-âge car elle a permis d'organiser les activités de prière et de travail au cours de la journée. Dans nos sociétés technologiquement (plus) avancées, nous avons des notions temporelles beaucoup plus fines (Comrie [1985]). Il en découle un impact direct sur le lexique qui va contenir des mots tels que « nanoseconde ». A ce propos, il faut souligner que les unités temporelles modernes, définies par des organismes internationaux (BIPM [2006]) diffèrent quelque peu des unités telles que présentées au début de cette section. En effet, l'horloge astronomique a cédé sa place à l'horloge atomique. L'unité de base actuelle, la seconde, est maintenant définie par rapport à la période d'une certaine radiation, choisie dans le spectre du césium 133[3].

[3] La seconde est la durée de 9.192.631.770 périodes de la radiation correspondant à la transition entre les deux niveaux hyperfins de l'état fondamental de l'atome de césium 133 (BIPM [2006]).

Enfin, notons que le temps a été au centre d'une multitude de recherches et de réflexions. En physique, l'approche classique de Newton, soutient qu'un événement ponctuel est parfaitement déterminé lorsque sa localisation dans l'espace et le temps est connue. Le temps est vu comme « une grande horloge extérieure à l'Univers dont les aiguilles indiquent un même temps absolu pour tout le monde » (Le Meur [2010]). Cette conception est remise en cause par la théorie de la relativité d'Einstein. Celle-ci conteste le statut du temps et de l'espace et conduit à des théories qui les considèrent comme des notions qui ne sont plus indépendantes. En effet, « en relativité générale, l'évolution et la localisation d'un corps ne se repèrent pas par rapport à un espace-temps cadre, mais relativement à l'évolution et à la localisation d'autres corps » (De Saint-Ours [2010]). Les derniers développements de la mécanique quantique tendraient même à éjecter la variable du temps des équations qui la régissent[4].

La définition du temps a également beaucoup occupé les philosophes, sans pouvoir cependant dégager un avis unanime :

> « De fait, au cours... du temps, les philosophes ont convoqué à peu près autant d'arguments pour prétendre que le temps existe que pour prétendre qu'il n'existe pas. » (Klein [2010])

Ce constat poussa Pascal à déclarer que le temps est de ces notions qu'il est impossible et même inutile de définir :

> « Qui pourra le définir ? Et pourquoi l'entreprendre, puisque tous les hommes conçoivent ce qu'on veut dire en parlant du temps, sans qu'on le désigne davantage ? » (Pascal [1838], p. 27).

Dans notre société, il existe de nombreux domaines dans lesquels le temps a pris une importance particulière. Dans le contexte professionnel, de nombreuses personnes sont payées « à l'heure » ou engagées pour prester un certain nombre d'heures par semaines. Un grand nombre d'entreprises importantes ou d'administrations ont instauré une procédure de « pointage » permettant de mesurer le temps de travail de chacun. Toujours dans le monde du travail, on a vu apparaître des expressions telles que « les 35 heures » ou « travail à temps partiel », soulignant bien l'importance de la dimension temporelle dans cet univers.

D'un point de vue juridique, on note également le caractère fondamental du temps, que ce soit pour dater les lois, prononcer des peines (« cinq ans de prison »), établir l'antériorité de tel élément sur un autre (par exemple en propriété intellectuelle). Dans le cadre des archives d'entreprises et de la *gestion de la preuve*, le temps joue également un rôle fondamental (temps de conservation légal des documents, preuve en matière de délit d'inité, etc.). De même, de nombreux documents officiels nécessitent la présence d'une date afin d'être valides et certaines procédures qui exigent l'envoi de courrier font appel au cachet ou à la *date de la poste* (« la date de la poste faisant foi »).

On remarque aussi la présence du temps pour l'organisation de nombreux aspects de la vie quoti-

[4] Pour compléter ces considérations, consulter le numéro de juin 2010 de *La Recherche*, intitulé *Le temps n'existe pas*.

dienne : les agendas pour les rendez-vous, les horaires pour les transports en commun, les services et commerces ou encore les spectacles. Enfin, pour beaucoup d'activités, la notion de rapidité s'est imposée, que ce soit dans l'industrie où il faut toujours produire de plus en plus vite ou dans les compétitions sportives souvent (mais pas toujours) basées sur des critères de vitesse.

3.3 Le temps dans le langage naturel

Étant donné l'importance de la notion de temps, il n'est donc pas vraiment surprenant de voir apparaître, dans de nombreux textes, une grande quantité de marques temporelles. Leur utilité est d'organiser les diverses informations dans l'espace du temps.

L'étude de la temporalité dans le langage naturel a été abordée au travers de nombreux domaines de recherche : la linguistique, les théories du discours, la logique, l'extraction d'information ou encore l'ingénierie des connaissances. Si ces travaux abordent tous le sujet selon un point de vue différent, il existe évidemment de nombreux points communs ou d'interconnexions entre ceux-ci. Il n'est pas pour autant facile d'obtenir une vue d'ensemble tant les différentes approches semblent concentrées sur le point spécifique qui les occupe. Il faut reconnaître que la complexité du problème, ou des problèmes, posé(s) par le traitement du temps en langage naturel ne permet souvent pas de l'aborder sous tous ses aspects.

Dans les chapitres suivants, quelques-unes des théories les plus importantes sont abordées afin de brosser un aperçu de ces différentes approches. Nous allons cependant dès à présent évoquer quelques exemples qui vont permettre de donner une première intuition des obstacles rencontrés lors du traitement du temps dans les textes en langage naturel.

L'information temporelle apparaît souvent de manière directe par l'intermédiaire d'expressions temporelles, souvent adverbiales :

> « le 25 mai 2009 »,
> « jeudi »,
> « ce soir ».

Il existe cependant de nombreux autres moyens d'exprimer la notion de temps. L'emploi des temps morphologiques permet par exemple de placer le propos dans le présent, le futur ou le passé. Certaines constructions syntaxiques apportent également des éléments d'informations. Il existe aussi des connecteurs qui ont une dimension temporelle, comme « et » qui peut marquer la succession de deux actions. De manière générale, l'agencement des phrases et des propositions au sein du discours a son importance. Qu'il s'agisse de propositions simplement juxtaposées, de propositions coordonnées ou subordonnées, il y a souvent une indication sur la temporalité des événements qui peut être déduite :

> « Il traversa la rue prudemment car les voitures roulaient rapidement. Il prit un taxi pour se rendre à l'aéroport »,
> « Il a raté son bus car il s'est levé trop tard »,
> « Il a pris le bus après avoir atteint Bruxelles en train »,

« Il a rattrapé le bus qu'il avait raté ».

Enfin, les événements en eux-mêmes possèdent une dimension temporelle implicite :

« La bombe a explosé »,
« Il a été flashé sur l'autoroute »,

expriment des événements qui semblent être instantanés. Au contraire :

« Il a traversé la Manche à la nage »,
« Il a attrapé un coup de soleil »,

évoquent plutôt des actions qui ont nécessité un certain temps pour s'accomplir.

Les dates telles que « le vendredi 26 juin 2009 » sont des références temporelles, dites *absolues*, qui sont très faciles à interpréter car elles se rapportent directement à un élément déterminé du calendrier ou à une section sur une ligne du temps. Il existe cependant des moyens moins directs pour désigner un moment dans le temps. Certaines expressions temporelles, désignées comme *relatives*, telles que :

« hier »,
« jeudi »,
« à 20 heures »
« il y a une semaine »,

doivent être interprétées dans le contexte temporel de l'énonciation. D'autres encore sont exprimées par rapport au contexte défini par le discours, comme

« la veille »,
« trois mois plus tard ».

L'exemple suivant illustre la différence entre ces deux cas :

« Il y a une semaine, Luc a mangé au restaurant et a eu une intoxication alimentaire. Trois jours après, il en était encore malade ».

Pour localiser le moment du repas, on *recule* d'une semaine par rapport au moment de l'énonciation, alors que pour déterminer le moment auquel Luc est toujours malade, on effectue un déplacement *en avant* par rapport au moment auquel l'intoxication s'est déclarée.

Enfin, la localisation temporelle peut aussi s'effectuer relativement à un point explicitement donné avec l'expression temporelle, mais dont la localisation temporelle précise s'appuie sur une *connaissance du monde*, qui n'est pas nécessairement connue (« trois jours après sa victoire »).

Lors de la production d'un texte, le scripteur aura parfois recours à des expressions temporelles qui lui permettent de rester imprécis quant à la localisation temporelle de son propos. Ces expressions peuvent par exemple être énoncées à l'aide d'une préposition qui permet de rendre imprécise l'identification d'un point :

« vers 15 heures »,
« au début des années soixante ».

L'imprécision peut aussi provenir du déplacement temporel par rapport à un point de référence qui est lui précis :

« environ une heure plus tard »,
« environ une heure après 20 heures ».

Notons également qu'une expression telle que « il y a un an » ne s'interprète pas nécessairement comme étant l'instant situé exactement 365[5] jours plus tôt.

Ce dernier exemple introduit le concept de *granularité*. Nous avons vu qu'au cours de l'histoire différentes unités ont été définies afin de mesurer le temps. Celles-ci sont utilisées en fonction de ce qui est situé dans le temps. Pour évoquer un mandat présidentiel, on ne fait pas référence au temps au moyen des mêmes unités que lorsque le sujet est un 100 mètres. Il serait tout à fait inadapté de dire que la fonction présidentielle s'exerce normalement sur 157.680.000 secondes (5 ans) et que le 100 mètres peut être couru en $6,34e^{-08}$ années (10 secondes). À chaque type d'événement correspond une granularité *naturelle*, à partir de laquelle il est le plus commode de l'évoquer. Évidemment, il arrive souvent que plusieurs granularités puissent convenir. Il est alors possible de passer de l'une à l'autre :

« Il a été nommé au poste de Premier Ministre ce vendredi »,
« Il a été nommé au poste de Premier Ministre en 2010 ».

L'information véhiculée par ces deux phrases est similaire car elle correspond au même fait objectif, mais n'est cependant pas tout à fait identique en ce qui concerne la précision de la localisation temporelle. En langage naturel, un locuteur peut choisir d'utiliser une unité temporelle différente de la granularité *naturelle* et ainsi produire un effet d'imprécision. Ce mécanisme peut être utilisé lorsqu'il juge que, dans le contexte de sa conversation, son propos ne mérite pas un niveau de précision plus élevé

« Son anniversaire est en mars »,

ou tout simplement parce qu'il exprime quelque chose qui est imprécis par nature

« Les navigateurs devraient arriver à destination en décembre ».

L'existence de références temporelles volontairement imprécises induit cependant un phénomène d'ambiguïté. Une proposition telle que

« En septembre, les billets d'avion pour l'Australie seront en promotion »

ne permet pas de conclure sur la période concernée. Il se peut qu'il s'agisse de tout le mois de septembre ou au contraire d'une période précise et plus restreinte de ce mois. D'autres cas ambigus peuvent survenir. La distinction entre une expression générale ou particulière en est un. Avec une

[5] Ou 366 lors des années bissextiles !

affirmation telle que

> « L'hiver est froid »,

parle-t'on de l'hiver en tant que concept de saison, en tant que période approximative correspondant à la partie de l'année durant laquelle il fait froid, ou encore en tant que période très précise délimitée par deux dates fixes ? Cet exemple introduit d'ailleurs un autre type d'ambiguïté, que l'on ne peut résoudre qu'en ayant certaines *connaissances sur le monde* et sur le contexte d'énonciation. En effet, le terme « hiver » ne désignera par exemple pas la même période selon que l'on se situe dans l'hémisphère nord ou sud.

Certaines expressions temporelles peuvent également être employées pour désigner l'événement qui a habituellement lieu à ce moment là ou qui a eu lieu à cette date, par exemple :

> « le 21 juillet » pour la fête nationale belge,
> « le 11 septembre » pour les attentats contre le World Trade Center à New-York en 2001.

Si ces expressions, désignées par Calabrese Steimberg [2008] comme des *héméronymes*, conservent une certaine valeur temporelle, elles désignent avant tout un événement et non la date à laquelle il s'est produit.

Enfin, il arrive aussi très fréquemment que la caractérisation temporelle des faits exposés dans un texte ne soit pas explicite. La proposition

> « Le premier ministre Yves Leterme a déclaré [...] »

exprime implicitement que Yves Leterme a été premier ministre au moment de la rédaction de l'article (ou à un autre moment précisé par le contexte).

3.4 Le texte au travers du triangle de référence

Pour traiter, dans les textes en langage naturel, d'une notion aussi fondamentale que la temporalité, il est important de comprendre ce que ces textes représentent, comment ils ont été produits, et selon quels mécanismes ils pourront être compris par un lecteur. Pour aborder ces différents aspects, nous avons utilisé un outil d'interprétation bien connu : le *triangle de référence* (Ogden et Richards [1969]).

Un texte, quel qu'il soit, n'est qu'une matérialisation sous une forme particulière d'une certaine réalité, objective ou subjective. Les objets (entités, événements, faits, idées, etc.) qui existent en pensée ou en réalité sont donc exprimés selon un *code* qui est le langage naturel. Ogden et Richards [1969] présentent un modèle qui se situe à la croisée des chemins de la linguistique, de la philosophie et de la psychologie : le *triangle de référence* (Figure 3.1), aussi appelé le *triangle sémiotique*. Il permet d'analyser les relations entre les objets du monde, la représentation qui en est faite et les unités lexicales qui permettent de les décrire. Il est donc composé de trois dimensions, une sur chaque

sommet, représentant respectivement le *monde*, le *domaine conceptuel* et le *domaine symbolique*.

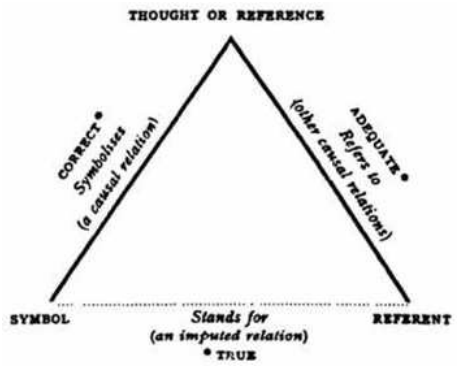


Figure 3.1 : Le triangle de référence (Ogden et Richards [1969]).

Les faits ou événements (les données) ont une existence objective ou subjective dans ce que nous appellerons le *monde* (sommet *referent*). D'autre part, ces objets sont projetés dans le *domaine conceptuel* dès l'instant où une personne les manipule mentalement et s'en fait une représentation (sommet *tought of reference*). Pour un objet du monde réel, il peut bien entendu exister plusieurs conceptualisations différentes. Prenons l'exemple d'une voiture de sport. Un ingénieur en aura une représentation liée à l'aspect technique, le fait qu'elle est composée d'un châssis, d'un moteur et de diverses autres parties. Un pilote y associera plutôt l'idée de vitesse et de performance, alors qu'une personne soucieuse de l'environnement pensera à la pollution qu'elle engendre. Ces représentations subjectives sont souvent, de manière consciente ou inconsciente, incomplètes ou erronées par rapport à la réalité. Cet ensemble de conceptualisations, qui correspondent à un seul et même objet du monde, peut être transposé dans le *domaine symbolique* de diverses manières et selon une multitude de formalismes (sommet *symbol*). Le formalisme le plus évident est bien entendu le langage naturel. Celui-ci se décline sous sa forme orale ou écrite, dans différentes langues et selon différents styles. En l'occurrence, nous nous intéressons principalement à l'écrit. Ce mode d'expression présente une grande variété stylistique. Pour parler de notre voiture de sport, il sera possible de rédiger une fiche technique reprenant ses spécifications détaillées, de la décrire en langue *standard* voire d'écrire un poème ou un roman à son sujet. Notons qu'à côté du langage naturel, il existe d'autres moyens d'expression tels que les langages mathématiques, le langage binaire, la représentation sous la forme d'une base de données relationnelle ou d'une ontologie, etc.

Le temps, en tant que composant du monde, ne fait bien entendu pas exception au principe d'interprétation du *triangle de référence*. Il s'écoule selon des lois précises, est perçu d'une certaine manière, et est exprimé au moyen de mécanismes linguistiques particuliers. L'objectif des chapitres suivants est d'exposer les diverses approches qui s'intéressent aux différentes parties du triangle de référence.

Les théories linguistiques (Chapitre 4) ont pour objectif l'observation et l'explication du fonctionnement de la langue. Il s'agit donc d'analyser le processus de production de textes et d'informations, illustré par le cas du journaliste, au point 1 de la figure 3.2. L'analyse linguistique part donc du domaine symbolique, c'est-à-dire les textes, vers le domaine conceptuel, c'est-à-dire les théories

linguistiques.

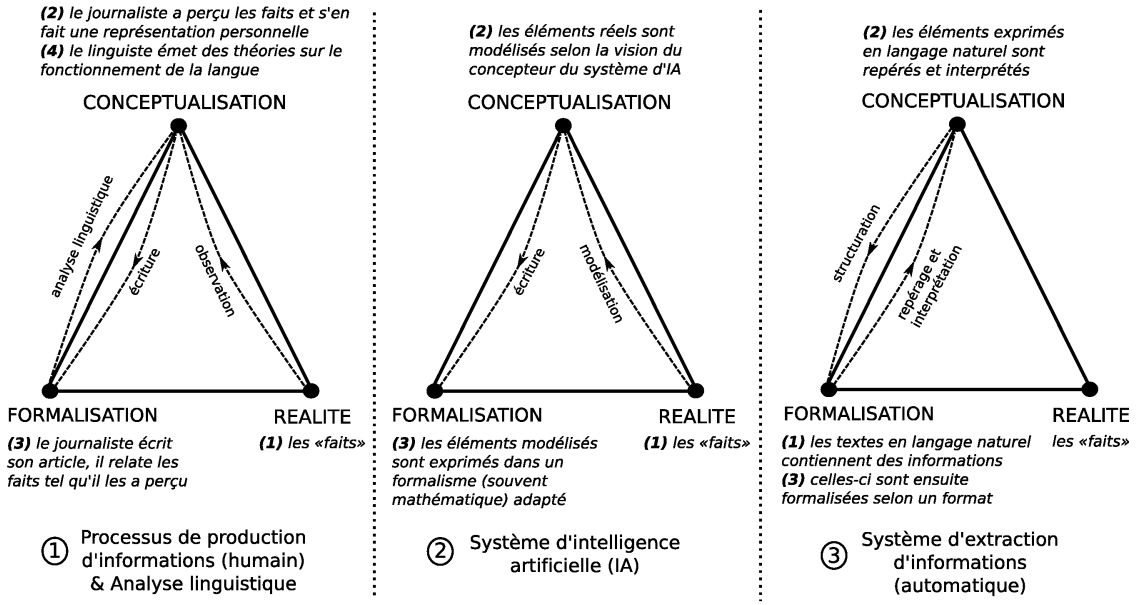


Figure 3.2 : Le triangle de référence appliqué à différents systèmes : production de texte et analyse linguistique, intelligence artificielle, extraction d'information.

Par ailleurs, d'autres théories, principalement issues du domaine de l'intelligence artificielle, vont plutôt s'intéresser aux moyens de modéliser, de représenter et de raisonner directement sur les éléments du monde réel (Chapitre 5). Comme le montre le point 2 de la figure 3.2, ces travaux de modélisation se situent entre les sommets qui représentent la *réalité* (ou le *monde*) et le domaine conceptuel. Évidemment, les modèles doivent être exprimés dans un langage adapté, qui sont du ressort du sommet *formalisation* (domaine symbolique).

Ces deux premiers domaines ne sont bien entendu pas complètement cloisonnés et il arrive que l'un contribue à faire progresser l'autre. C'est par exemple le cas pour certaines théories linguistiques d'analyse du discours qui font appel à une représentation formelle issue de la logique.

Enfin, nous nous intéresserons au domaine de l'extraction d'informations qui, par sa visée plus opérationnelle et applicative, a pour vocation de réunir linguistique et intelligence artificielle (Chapitres 6 et 7). Son objectif est de passer d'une représentation symbolique à une autre (voir le point 3 de la figure 3.2). En l'occurrence, il s'agit d'aller de la langue naturelle vers une représentation plus facilement manipulable par des moyens informatiques, et dans laquelle des attributs ont été attachés à certains éléments d'informations (des données sémantiques par exemple). Cette opération s'effectue nécessairement en effectuant un détour par le domaine conceptuel afin que la nouvelle représentation symbolique réponde à un modèle bien précis.

Comme nous l'avons déjà mentionné, cette thèse s'inscrit dans la problématique de l'accès à l'information, et plus particulièrement de la recherche d'informations (RI). Les différents travaux que nous allons passer en revue (Chapitres 4 à 6) servent de fondations au développement d'une méthode

d'analyse de la temporalité (Chapitre 7), orientée vers l'utilisation dans un cas applicatif concret, l'indexation multidimensionnelle, ou plus précisément thématico-temporelle (Chapitre 8). L'approche proposée n'est pas spécifique à ce cas précis, mais certains aspects ont été mis en avant au détriment d'autres. Cette démarche, courante dans le domaine de l'extraction d'informations, pourrait être vue comme réductrice. Au contraire, elle constitue un intérêt particulier. L'objectif n'est en effet pas d'étudier de manière exhaustive tous les problèmes liés à la temporalité dans le langage naturel, mais bien d'identifier les aspects qui peuvent être utiles dans une perspective de traitement automatique en général, et pour l'amélioration de l'accès à l'information en particulier.

CHAPITRE 4

EXPRESSION DU TEMPS DANS LE LANGAGE NATUREL

4.1 Introduction

Comme nous l'avons déjà évoqué, la démarche menée en linguistique consiste à étudier ce qui dans la langue permet d'exprimer l'information temporelle. L'intuition donnée à la section 3.3 au sujet de la variété des références au temps dans le langage naturel a également été mise en avant par de nombreux auteurs, tel que Bell [1998]. Celui-ci explique que le temps est exprimé à différents niveaux : dans la morphologie et la syntaxe des groupes verbaux, dans les adverbes temporels (lexique ou paraphrase), dans la structure du discours. Gosselin [1996] relève lui aussi le fait qu'il existe un ensemble de marques linguistiques de la temporalité, mais aussi que celles-ci doivent être évaluées conjointement afin de pouvoir en dériver une interprétation correcte :

> « [...] 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. » (Gosselin [1996], p. 23)

Dans les sections suivantes, nous allons examiner successivement les diverses marques linguistiques qui véhiculent la temporalité.

4.2 Les adverbiaux temporels

Un moyen très courant d'exprimer le temps dans les textes est d'utiliser des expressions qui ont, de manière assez explicite et directe, une valeur temporelle. Ces expressions sont regroupées sous le nom d'*adverbiaux temporels*. Le terme *adverbial* est à mettre en relation avec la notion de *complément adverbial* (Gross [1986], cité par Borillo [1998])[1], c'est-à-dire d'un ensemble d'éléments qui peuvent prendre la fonction d'adverbe[2]. La variété rencontrée au niveau des différents adverbiaux est assez

[1] Chez Gross, le nom d'*adverbe généralisé*, d'*adverbe* ou de *complément adverbial* est donné aux compléments qui, dans la terminologie traditionnelle, sont repris sous les catégories d'adverbes, de compléments circonstanciels et de propositions subordonnées circonstancielles.

[2] Pour plus de commodité, nous utiliserons dès maintenant de manière indifférente, sauf indication contraire, les termes adverbes et adverbiaux.

grande, tant en ce qui concerne leur *nature* (la manière dont ils sont construits, voir section 4.2.1), que leur *rôle* (le sens qu'ils véhiculent, voir section 4.2.2) ou encore leur *interprétation* (les mécanismes nécessaires pour passer de l'expression au sens véhiculé, voir section 4.2.3). Les sections suivantes abordent les adverbes temporels selon ces trois axes.

4.2.1 Nature des adverbiaux temporels

En ce qui concerne les adverbes, une caractéristique marquante, en plus de leur caractère facultatif, invariable et mobile, est leur hétérogénéité. Les adverbes de temps n'y échappent pas. En effet, aussi bien morphologiquement que du point de vue de leur comportement syntaxique, les adverbes de temps constituent un groupe varié.

Morphologie et catégories grammaticales

Pinchon [1969] fait remarquer que, dans les adverbes de temps, on rencontre des mots venant du latin (« hier »), des mots formés à l'aide d'un suffixe (« actuellement »), ceux formés par soudure de plusieurs éléments (« aussitôt »), des locutions (« tout de suite »). De plus, certains adverbes sont susceptibles d'être accompagnés de marques d'intensité (« (très) souvent ») alors que d'autres ne les acceptent pas (« aujourd'hui »).

La variété des éléments qui peuvent prendre la fonction d'adverbial temporel est également illustrée par les différentes catégories grammaticales qui sont concernées : *adverbes simples ou composés* tels que « hier », « aujourd'hui », « demain », « plus tard », *noms* ou *groupes nominaux* (« jeudi », « le 25 juin 2009 », « ce matin ») et *groupes prépositionnels* (« dès le lendemain », « après 14h00 », « en 2009 »). Certaines formes peuvent se combiner pour donner naissance à des expressions plus complexes, par exemple un groupe nominal suivi d'un adverbe (« trois ans plus tard »).

Syntaxe

Au niveau syntaxique, et toujours selon Pinchon [1969], il existe des mots dont l'usage peut entraîner une ambiguïté entre adverbe et préposition (« avant ») ou conjonction de coordination (« ensuite »).

Elle note également que si les adverbes de temps peuvent tous porter sur le verbe (« Il convient maintenant de décider [...] »), ils peuvent aussi modifier un adjectif (« [...] deux exigences parfois contradictoires »).

Quelques adverbes peuvent être postposés à des substantifs (« deux jours avant »). Certains adverbes peuvent faire l'objet d'une construction indirecte avec les prépositions *à* (« à tantôt »), *de* (« d'aujourd'hui »), *dès* (« dès potron-minet »), *depuis* (« depuis toujours »), *jusque/jusqu'à* (« jusqu'à maintenant ») ou *pour* (« pour toujours »).

À l'échelle d'une proposition, la place de l'adverbe n'est pas fixe et est souvent déterminée par le

style, même si le nombre de constituants a aussi une influence à cet égard[3].

Enfin, suivant les adverbes, l'insertion entre l'auxiliaire et le participe passé d'un verbe conjugué à un temps composé sera ou non permise.

Certains adverbes de temps peuvent parfois perdre leur valeur temporelle. C'est le cas de « alors », utilisé pour introduire une conclusion, ou de « maintenant » utilisé en tête de phrase comme dans : « Maintenant, il avait peut-être raison ».

Caractérisation sémantique

Plusieurs classes particulières de noms peuvent être distinguées pour exprimer le temps et plus spécialement la durée. Tout d'abord, celle qui reprend les *noms de temps* (*Ntps*) et que Borillo [1986] définit plus précisément comme la catégorie des noms qui désignent des découpages temporels (« jour », « semaine », « saison »).

Une autre catégorie de noms *temporels* est mise en avant par Borillo [1988], les *noms de durée* (*Ndur*) : des noms relatifs à des événements ou des faits qui possèdent un caractère duratif et dont l'emploi permet donc d'exprimer la durée. Ces *Ndur* sont des nominalisations de verbes (« traversée », « voyage ») ou d'adjectifs (« maladie », « jeunesse »). Il peut également s'agir de noms prédicatifs qui, lorsqu'ils sont utilisés avec un verbe support, fournissent un prédicat (« prendre des vacances », « être en congé »). Les *Ndur* peuvent être utilisés en combinaison avec les *Ntps* dans des constructions telles que *Ndur de Quant Ntps* (« une sieste de deux heures »). Borillo [1988] précise que les deux catégories sont liées : les *Ntps* sont des *Ndur* qui présentent la particularité de pouvoir servir d'unité de mesure.

De manière plus générale, Borillo [1998] donne une répartition en sept catégories sémantiques des substantifs utilisés dans les adverbiaux de temps.

- N1 : Noms d'années ou d'époques, dates uniques (« 1989 », « le Moyen-Âge », « le 31 Décembre 1996 »).
- N2 : Noms utilisés comme unités de mesure de temps (« jour », « mois », « année », « minute »). Cette catégorie correspond aux *Ntps* définis précédemment.
- N3 : Noms donnés aux éléments constitutifs de certaines unités de mesure (« lundi », « Pâques »).
- N4 : Noms de sous-parties de l'unité de mesure de 24 heures (« matin », « soir », « après-midi », « nuit »).
- N5 : Désignation d'activités ou d'événements, périodes marquantes de la vie (« repas », « vacances », « naissance », « mort »). Cette catégorie reprend une partie des *Ndur*.
- N6 : Désignation de parties qui se rapportent aux noms N1 à N5 qui se réfèrent à des intervalles (« début », « milieu », « fin »).

[3] On dira « Il est resté longtemps », mais pas « Longtemps il est resté », alors que « Longtemps il est resté sans venir en Suisse » sera correct.

– N7 : Désignation de segments temporels vagues et indéterminés (« pendant ce temps », « à cette occasion »).

Une description très complète de la formation des adverbes de temps a également été menée par Gross [1986] dans le cadre du lexique-grammaire. Nous ne détaillerons pas ici ces importants travaux qui constituent par contre une référence de travail intéressante pour l'élaboration de ressources d'extraction d'informations temporelles.

4.2.2 Rôle de l'adverbe

D'un point de vue conceptuel, Borillo [1986] expose les différents rôles pris par les adverbiaux de temps. Par rapport à une représentation de l'espace temporel sous la forme d'une ligne du temps, ces expressions correspondent à des intervalles ou des points qui peuvent :

– être des *références temporelles* qui servent de repère d'occurrence et qui répondent à la question « quand ? » ;
– être des *durées* qui répondent alors à la question « combien de temps ? » ;
– exprimer une *distribution* (« tous les dimanches ») ;
– exprimer une *fréquence* (« trois fois par jour »).

4.2.3 Interprétation de l'adverbe

En ce qui concerne l'interprétation des adverbes, nous nous intéressons principalement à leurs deux grands rôles, celui de *référence temporelle* et celui de *durée*. Pour chacun de ceux-ci, une catégorisation des adverbes peut être établie de manière à les rassembler en classes homogènes par rapport aux types de traitements à mettre en œuvre. Les deux points consacrés à ces problèmes font principalement référence aux différents travaux de Borillo en la matière (Borillo [1983, 1998, 1988]).

Les adverbiaux de référence temporelle

La PREMIÈRE ANALYSE (1) que nous rapportons provient de Borillo [1983], pour qui les adverbiaux de référence temporelle sont organisés selon une double catégorisation. La *première distinction* (1a) concerne la relation entre l'adverbe et le moment de l'énonciation. Elle débouche sur quatre classes : les adverbes autonomes, déictiques, anaphoriques et enfin polyvalents (ou neutres).

– Les *adverbes autonomes* se caractérisent par leur indépendance à la fois par rapport au moment de l'énonciation et au contexte temporel des phrases précédentes du discours. Une catégorisation plus fine consisterait à les répartir en datation calendaire ou historique (« en 1980 », « au Moyen-Âge »), en datation événementielle (« à sa naissance », « avant la réunion ») et en localisation vague (« une fois », « à l'avenir »).
– Les *adverbes déictiques* sont définis par rapport au moment de l'énonciation (« il y a une semaine », « l'an prochain »). Celui-ci implique une division de l'espace du

temps en trois parties, passé / présent / futur, dans lesquelles viennent s'inscrire ces adverbes.

– Les *adverbes anaphoriques* (« la veille », « dix ans plus tôt ») sont définis par rapport à une référence temporelle déjà établie dans le discours. Une fois cette référence identifiée, l'adverbe anaphorique est interprété selon un rapport d'antériorité, de simultanéité ou de postériorité avec celle-ci.
– Les *adverbes polyvalents* sont définis comme pouvant fonctionner à la fois en tant que déictiques ou en tant qu'anaphoriques. Dans « Il est arrivé dans la soirée », l'interprétation déictique se rapporte à une partie de la journée d'énonciation, alors que son usage anaphorique réfère plutôt à un intervalle contenu dans une autre journée déterminée par le contexte.

La *seconde distinction* (1b) concerne la nature de l'indication temporelle : ponctuelle, inclusive ou durative.

– Les *adverbes ponctuels* identifient sur l'axe du temps un repère équivalent à un point (« à cet instant là », « à huit heures »). Ils peuvent être utilisés en combinaison avec d'autres adverbes qui désignent une section temporelle plus grande (« le matin à huit heures »), mais dont la taille maximale est restreinte à la journée (on ne dira pas « la semaine dernière à midi »).
– Les *adverbes inclusifs* sont assimilables à des intervalles de dimensions variées dans lesquels un événement a lieu (« en 1980 », « dans la semaine », « ce matin »). Cet événement peut se dérouler sur tout ou partie de l'intervalle[4].
– Les *adverbes duratifs* indiquent à la fois une durée et un ancrage temporel (« Depuis un mois, il ne va pas bien »). Les durées seules sont exclues de cette catégorie.

Les deux points de vue exposés ci-dessus constituent deux ensembles de critères que Borillo [1983] croise afin d'obtenir douze catégories illustrées au tableau 4.1.

ADVERBES	Ponctuels	Inclusifs	Duratifs
Autonomes	à sa naissance	en 1980	depuis 1980
Déictiques	à l'instant	cette semaine	depuis hier
Anaphoriques	à ce moment-là	ce matin-là	depuis la veille
Polyvalents	à huit heures	dans la soirée	depuis l'été

Tableau 4.1 : Illustration de la catégorisation des adverbes de référence temporelle (Borillo [1983]).

Une SECONDE ANALYSE (2) de l'organisation des adverbes de référence temporelle a été proposée par Borillo [1998]. À nouveau, deux axes de caractérisation sont utilisés. Le *premier axe* (2a) s'intéresse au type de référence temporelle exprimé par l'adverbe : il peut s'agir d'une localisation directe ou indirecte.

[4] Selon le temps verbal utilisé et/ou le type d'événement. Par exemple, imparfait et passé composé n'impliquent pas la même interprétation (« En 1980, il (était | a été) en prison ») et un état, activité ou un autre type de procès n'a intuitivement pas la même couverture de l'intervalle temporel (« En 1980, il a (bien) gagné (sa vie | un marathon) »).

- La *localisation directe* fournit une localisation sur la ligne du temps à partir de repères établis tels que des dates ou événements historiques (« le 1er janvier 1997 »), de repères liés à des événements supposés connus (« à la naissance de Paul ») ou encore de repères provenant du discours qui précède (« dans les jours qui suivirent »).
- Dans le cas d'une *localisation indirecte*, l'adverbe détermine lui aussi une localisation sur la ligne du temps, mais par l'intermédiaire d'un calcul basé sur des mesures temporelles (« Il partira dans huit jours »).

Le *second axe* (2b) concerne le type de repérage opéré par les adverbes : coïncidence ou inclusion d'un côté, limitation de l'autre. Ce critère porte sur le rapport entre le moment désigné par l'adverbial en entier (*R(Adv)*) et le moment qui correspond au syntagme nominal contenu dans l'adverbial (*R(SN)*)[5].

- Pour les *adverbes de coïncidence ou d'inclusion*, les deux moments coïncident ou se superposent (« lundi dernier », « toute la journée »).
- Dans le cas des *adverbes de limitation*, la localisation temporelle de l'adverbe complet ne correspond pas à celle du syntagme nominal qui y est contenu (« avant la fin de l'année », « depuis Pâques »).

À nouveau, Borillo [1998] croise les deux axes mis en évidence, et les catégories qui y sont définies. Le résultat est une typologie de quatre catégories.

- I. Adverbes de localisation directe :
  peuvent être constitués de tout type de noms (*N1* à *N7*) ;
  - I.A. Adverbes de localisation directe, *R(Adv)* et *R(SN)* coïncident :
    sont de forme *Adv*, *SN* ou *Prep SN* (« demain », « la semaine suivante », « dans la journée ») ;
  - I.B. Adverbes de localisation directe, *R(Adv)* et *R(SN)* ne correspondent pas :
    sont limitatifs à une ou deux bornes et de forme *Prep SN* (« dès lundi », « depuis dimanche ») ;
- II. Adverbes de localisation indirecte :
  ne peuvent faire intervenir que les *Ntps* de mesure (*N2*) et quelques *N7*, les *N* étant toujours précédés d'un numéral ou déterminant quantitatif (« quelques », « plusieurs ») ;
  - II.A. Adverbes de localisation indirecte, *R(Adv)* et *R(SN)* coïncident :
    sont des limitatifs qui impliquent une durée, leur valeur correspond à cette durée à partir d'une certaine borne (« Quand nous sommes arrivés, le train était parti depuis quelques minutes ») ;
  - II.B. Adverbes de localisation indirecte, *R(Adv)* et *R(SN)* ne correspondent pas :
    adverbes ponctuels dont la valeur est obtenue à partir d'un temps d'origine et d'une durée, sont de forme *Prep SN* ou « il y a ... », « cela fait ... » (« Il est revenu au bout de trois mois »).

[5] Si l'adverbial est composé uniquement d'un adverbe ou d'un syntagme nominal seul, il y a *de facto* correspondance de la localisation temporelle puisque Adv=SN.

Dans les deux analyses, les deux principaux axes de catégorisation semblent au premier abord relativement similaires. Ils concernent le *mode de placement* de l'adverbial sur la ligne du temps d'une part et le *type de localisation* sur cette ligne d'autre part. Les approches adoptées pour définir les catégories à l'intérieur de ces axes présentent cependant certaines différences. Pour le premier axe, la première analyse (1a) propose quatre catégories (direct, indirect déictique, indirect anaphorique et polyvalent) et s'appuie sur un critère soulignant les différences en ce qui concerne les mécanismes de localisation temporelle. La seconde analyse (2a) se base plutôt sur un critère purement morphologique scindant les adverbes en deux groupes que l'on pourrait résumer en *adverbes basés sur des unités de mesures temporelles* (localisation indirecte) et *autres adverbes* (localisation directe). En ce qui concerne le second axe, alors que la première analyse (1b) se penche sur la forme sous laquelle la référence temporelle se traduit sur la ligne du temps (un point, un point dans un intervalle, une durée ou un intervalle ancré temporellement), la seconde analyse (2b) s'intéresse au caractère inclusif ou limitatif de la référence sur cette même ligne (la localisation temporelle de l'adverbial correspond, ou pas, à celle du syntagme nominal qui le compose). En réalité, il s'avère que les deux catégorisations sont complémentaires, la deuxième analyse pouvant être utilisée pour raffiner la première. Borillo [1998] en apporte l'illustration en étudiant les adverbes anaphoriques au moyen de la seconde analyse exposée ci-dessus.

Avant d'aller plus loin, revenons sur la première catégorisation qui définit les adverbes ponctuels comme étant un repère assimilable à un point, mais toujours de taille inférieure à la journée. Cette restriction est motivée par la volonté de ne pas couvrir des constructions du type « la semaine dernière à midi », qui sont incorrectes. Selon nous, le caractère ponctuel d'une référence temporelle dépend plutôt du point de vue et de la granularité adopté. Ainsi, une année peut tout autant être représentée comme un point qu'une journée ou une heure bien précise (« 2008 fut une excellente année »).

Les adverbiaux de durée

Tout comme les références temporelles, la durée peut, elle aussi, être exprimée au moyen d'adverbes, qu'il s'agisse d'adverbes simples ou d'expressions adverbiales. Borillo [1988] aborde le cas de ces adverbiaux de durée et les partage en deux catégories.

- Pour l'adverbial *durée-limite*, la durée peut être exprimée au moyen d'un intervalle dont les bornes sont déterminées au moyen d'expressions faisant intervenir des noms de date *Ndate*[6] (« de lundi à samedi », « du 14 juillet au 15 août », « d'ici jusqu'à la fin du mois »). Dans certains cas, il est possible que seule une des deux bornes soit renseignée explicitement (« jusqu'à demain »). La borne manquante doit alors être interprétée comme étant soit le moment de l'énonciation, soit un point de référence donné précédemment dans le discours.
- La durée peut également être exprimée sans avoir à spécifier quelque borne que ce

[6] Un *Ndate* peut être formé à partir de noms de temps (*Ntps*), noms de jours de la semaine, de mois de l'année, de saisons, de jours de fêtes, *etc.* ou au moyen de noms évoquant des événements ponctuels (événements historiques, événements bien connus des participants au discours).

soit, simplement en indiquant la taille de l'intervalle sans l'ancrer temporellement. Ces adverbes de *durée-valeur* incluent des expressions qui sont assez approximatives (« pendant un instant ») ou au contraire plus précises (« quelques mois », « trois jours »). Ils sont obtenus en utilisant un ensemble de noms de temps, *Ntps*, qui doivent être complétés par un quantitatif (un déterminant numéral, un indéfini) afin d'être interprétés en tant que durée. Les adverbes de *durée-valeur* ne sont cependant pas exclusivement composés d'expressions à base de *Ntps* puisque les noms de durée (*Ndur*) peuvent aussi être employés.

Remarque finale sur les adverbes temporels

Il ressort de ces différents travaux qu'il est délicat de définir une catégorisation des types d'adverbiaux temporels. L'adoption d'une nomenclature d'adverbes doit avant tout être dirigée par l'usage envisagé et par l'aspect particulier étudié. Ainsi Gross [1986] explique :

> « Toute la difficulté du problème des temps réside dans l'établissement d'une correspondance cohérente entre des catégories INTUITIVES comme **date absolue** (date du calendrier), **date relative** (date repérée par rapport à un événement), **présent**, **futur**, **début**, **fin**, **date**, **durée** et les formes que sont les mots isolés ou les constructions. La méthode principale consistera à se restreindre à un ensemble de catégories intuitives suffisamment OPÉRATOIRES pour que les interprétations de phrases soient reproductibles entre locuteurs. » (p. 206)

4.3 Les connecteurs temporels

Parmi les nombreux adverbes temporels, il existe une classe particulière qui est désignée sous le nom de *connecteurs temporels*. Il s'agit d'éléments qui possèdent à la fois les caractéristiques d'adverbes temporels et de connecteurs de discours. Charolles *et al.* [2005] donne la définition suivante :

> « [...] among the vast set of adverbial establishing a temporal (or aspectuo-temporal) relation between propositions [...], we consider as connectives only those which entail at the same time a logico-pragmatic relation - *i.e.* which play a role at the level of discourse relations. » (p. 117)

Borillo *et al.* [2004] proposent également une définition de ces adverbes :

> « Within the large category of temporal adverbials, there is a small group of adverbials known as 'relational temporal adverbials'[7] that is to be distinguished on the basis of both their syntactic and their semantic properties. [...] The function of these adverbials is primarily to act as relational elements that establish a transphrastic relation between their host sentence and the preceeding sentence - or a preceeding segment of text larger

[7] Cette apellation est due à Nojgaard [1992].

than a sentence. » (p. 312)

Borillo *et al.* [2004] distinguent en réalité trois catégories d'adverbes qui peuvent prendre le statut de connecteurs :

- *les adverbes relationnels*, qui sont des éléments qui agissent comme des liens, et qui peuvent renseigner le temps entre deux événements (« deux jours après »), ou simplement exprimer un ordre chronologique entre ceux-ci (« ensuite ») ;
- *les adverbes de référence anaphorique*, qui introduisent une référence vers un moment déjà connu qui est différent du moment de l'énonciation (« le lendemain ») ;
- *les adverbes aspectuo-temporels*, qui sont des adverbes qui expriment plutôt une modalité, la rapidité par exemple, mais qui prennent aussi une dimension temporelle (« brusquement »).

Bien qu'ils puissent être placés à l'intérieur même de la structure de la phrase, c'est le plus souvent dans une position détachée, en début de phrase, que ces trois types d'adverbes expriment, de manière plus évidente, leur rôle de connecteur. Différentes relations de discours (*cf.* section 4.8) peuvent être matérialisées par ces connecteurs. Dans Borillo *et al.* [2004], les adverbes « puis » et « un peu plus tard » sont examinés. Même s'ils ont une signification relativement similaire, leur influence sur la structure du discours n'est pas tout à fait identique : la relation de narration introduite par « puis » est en effet beaucoup plus forte que celle impliquée par « un peu plus tard ».

4.4 La notion de procès

L'analyse de la temporalité en langage naturel est indissociable des objets auxquels elle se rapporte. Ces objets sont les faits ou les événements qui, au travers de leur description dans les textes, sont situés dans le temps. Or, comme nous l'avons déjà mentionné à la section 3.3, les événements eux-même possèdent une dimension temporelle implicite. En effet, les actions sont instantanées ou, au contraire, prennent un certain temps pour s'accomplir[8]. Ces actions peuvent aussi être caractérisées par leur nature itérative ou, au contraire, atomique[9].

S'il semble facile d'obtenir une intuition sur cette question, la caractérisation précise de ce que nous avons successivement appelé *fait*, *état*, *événement* ou encore *action*, est moins évidente. Le terme général qui regroupe ces notions est *procès*. La définition et l'organisation des différents types de procès ont été réalisées au moyen de typologies qui ont été, entre autres, proposées par Vendler [1957], Kenny [1963], Mourelatos [1978], Culioli [1983] ou Desclés [1991]. Plusieurs travaux, tels que Fuchs [1991] ou Gosselin et François [1991], offrent une analyse comparative des différentes approches du *procès*.

Le terme de *procès* ou de *type de procès* a dans un premier temps été utilisé pour caractériser la

[8] « Être flashé sur l'autoroute » contre « traverser la Manche à la nage ».

[9] « Se diriger vers la sortie », qui désigne un certain nombre d'actions similaires (des pas), contre « franchir la porte », qui est une action qui se passe *en une fois*.

structure temporelle du verbe. C'est la notion d'*Aktionsart*[10]. Dans ce cas, les types de procès sont principalement définis de manière lexicale (on attribue à un verbe un type de procès particulier).

Plusieurs critiques peuvent être formulées par rapport à cette vision (Fuchs [1991]). D'abord, la construction de la forme verbale, par exemple la présence ou l'absence d'un objet ou d'un circonstant, possède une influence sur le type de procès qui le rend difficilement représentable au niveau du lexique : « manger » est une activité, mais « manger une pomme » est un accomplissement. Ensuite, il faut souligner que les caractéristiques temporelles du procès sont parfois combinées à d'autres caractéristiques, par exemple modales[11] ou actancielles[12]. Enfin, on peut évoquer le rapprochement qui peut être fait entre les procès sous leurs formes verbales et les formes nominales correspondantes[13]. On constate donc que c'est l'ensemble de la construction de la phrase qui est concerné par la détermination du type de procès. On est largement sorti du cadre strict des verbes pour glisser vers ce qu'on pourrait appeler des typologies de prédications ou de situations. Fuchs [1991] résume bien les hésitations sur cette problématique :

> « [...] d'une part les lexèmes sont intrinsèquement chargés de valeurs, ou au moins, de potentialités sémantiques spécifiques, mais d'autre part le contexte d'emploi de ces lexèmes contribue lui aussi à la construction de valeurs ; dès lors, comment se situer dialectiquement entre le « tout dans les mots » et le « tout dans le contexte » ? » (p. 11)

Au cours des années, de nombreuses typologies de procès ont été proposées. Celles-ci ne sont pas détaillées ici, même si certaines seront abordées par la suite (entre autres dans la partie consacrée à l'*aspect*, section 4.6), mais un rapide aperçu est tout de même proposé. D'une manière générale, les différents travaux portent sur des typologies de trois ou quatre grandes classes. Bien entendu, la terminologie utilisée n'est pas toujours identique, mais il est possible de les mettre en parallèle (voir tableau 4.2, construit à partir des constatations de Fuchs [1991]).

RÉFÉRENCES	CLASSES			
Desclés [1991]	état	processus	événement	
Kenny [1963]	état	activité	performance	
Culioli [1983]	compact	dense	discret	
Vendler [1957]	état	activité	accomplissement	achèvement
Mourelatos [1978]	état	processus	développement	occurences ponctuelles

Tableau 4.2 : Différentes typologies de procès.

Les différentes catégories de procès sont généralement définies à l'aide de traits sémantiques, par exemple : dynamique/statif, borné/non-borné, ponctuel/duratif. À partir de cet ensemble de traits, la classe *état* pourrait être définie comme non-dynamique, non-bornée et non-ponctuelle.

[10] Terme d'origine allemande, traduit en français par, entre autres, *mode d'action, modalité d'action*.

[11] Selon Querler [1996], la modalité est « l'expression de l'attitude du locuteur par rapport au contenu propositionnel de son énoncé ».

[12] « Construire une maison » ou « construire des maisons ».

[13] « Il était beau » et « sa beauté était reconnue » constituent deux moyens d'exprimer, sous des formes différentes, verbale et nominale, un même fait.

Si, dans un premier temps, les différentes typologies de procès restent comparables, elles se différentient sur le nombre de traits utilisés, la nature des traits ainsi que leur structuration, donnant finalement naissance à un grand nombre de définitions différentes. Les traits sémantiques qui définissent les typologies sont des variables binaires, ce qui rend impossible la catégorisation d'un procès entre deux catégories existantes, à moins de modifier la typologie.

La catégorisation en types de procès permet d'effectuer un regroupement des verbes et prédicats sur la base de leurs caractéristiques temporelles propres. Le type de procès a également une influence sur l'interprétation temporelle d'autres éléments de la phrase (adverbiaux, etc.)[14]. On comprend dès lors aisément l'intérêt que peut avoir une telle information lors de l'interprétation temporelle d'un texte.

4.5 Le(s) temps

L'utilisation des temps verbaux constitue un élément important pour l'expression de la temporalité. Comrie [1985], cité par Mani *et al.* [2005], définit les temps comme étant « the grammaticalized expression of location in time ».

Sous le terme de *temps*, se cache cependant plusieurs types, ou niveaux, qu'il est nécessaire de distinguer. Benveniste [1974] expose trois niveaux de temps : le temps physique, le temps chronique et le temps grammatical. Le *temps physique* est « un continu uniforme, infini, linéaire, segmentable à volonté. Il a pour corrélat dans l'homme une durée infiniment variable que chaque individu mesure au gré de ses émotions et au rythme de sa vie intérieur. » (p. 70). *Le temps chronique* est « le temps des événements, qui englobe aussi notre propre vie en tant que suite d'événements. » (p. 70). Les calendriers, basés sur la récurrence de phénomènes naturels, permettent de diviser ce temps chronique. Trois conditions sont nécessairement remplies par un calendrier : la condition stative (il a un point zéro), la condition directive (il y a un *avant* et un *après*) et la condition mesurative (il existe des unités de mesure). Quant au *temps linguistique*, il est lié à l'utilisation de la langue dans le but de manifester l'expérience humaine du temps.

Jespersen [1971] distingue le temps notionnel du temps grammatical. Le *temps notionnel* présente trois divisions essentielles : le passé, le présent et le futur. Il correspond au temps chronique de Benveniste. En français, le *temps grammatical* est composé de sept temps : un temps présent, trois temps passés et trois temps futurs. Ils correspondent aux temps que l'on utilise dans la langue. Le temps grammatical est donc à mettre en relation avec le temps linguistique. Dans le cadre du triangle de référence (voir section 3.4), il est possible de rapprocher les types de temps et les sommets du triangle. En particulier, le temps notionnel est lié au domaine conceptuel, alors que, dans le cas des langues naturelles, le temps grammatical est en rapport avec le domaine symbolique.

La difficulté de l'analyse des temps en linguistique provient de la mise en correspondance de deux niveaux de temps, notionnel et grammatical. En effet, Bestougeff et Ligozat [1989] constatent :

[14] Par exemple, « peu après » peut prendre des valeurs fort différentes selon les procès avec lesquelles il est employé (dans « La voiture s'enflamma et peu après elle explosa », on parle probablement de minutes, alors que dans « Il la rencontra durant l'été et peu de temps après il se fiancèrent », il s'agit plus vraisemblablement de semaines ou de mois).

> « Le problème principal qui se pose [...] est lié au fait que temps notionnel et temps grammatical n'ont pas de rapport simple entre eux. » (p. 14)

Gosselin [2005] insiste également sur ce point en s'interrogeant sur l'existence d'un grand nombre de temps morphologiques :

> « L'étude linguistique de la temporalité exprimée par les langues en Europe est traditionnellement centrée sur l'analyse des temps verbaux. Elle cherche à résoudre un paradoxe : si le verbe conjugué exprime le temps et si le temps se décline selon les trois dimensions du présent, du passé et du futur, comment expliquer qu'il puisse exister plus de trois temps morphologiques ? » (p. 31)

Ce propos peut être illustré en comparant les temps censés exprimer le passé, le présent et le futur. Dans les temps de l'indicatif, cinq temps sont destinés au passé (imparfait, passé simple, passé composé, plus-que-parfait et passé antérieur), deux au futur (futur simple et futur antérieur) et un seul au présent. On pourrait dès lors conclure qu'il existe plus d'un temps grammatical pour exprimer chaque temps notionnel. Cette conclusion serait cependant un peu rapide, car d'autre part, on peut constater qu'un temps grammatical peut parfois exprimer plus qu'un seul temps notionel :

> « [...] l'abondance de formes ne correspond pas à une simple profusion de la langue qui permettrait l'utilisation de plusieurs formes équivalentes pour exprimer une même idée. Il se peut au contraire que les choses se passent comme si l'on avait pénurie de formes. » (Bestougeff et Ligozat [1989], p. 19).

Bestougeff et Ligozat illustrent ce propos au moyen des exemples suivants qui, bien que tous exprimés à l'aide du présent, doivent être interprétés comme exprimant trois temps (notionnels) différents.

- « Attention, *le système s'arrête*, il y a une panne de disque. »
- « *Le système s'arrête* ce soir pour la maintenance. »
- « Hier, j'étais en pleine compilation, tout-à-coup *le système s'arrête* et me déconnecte, que puis-je faire maintenant ? »

Par conséquent, il devient évident que les temps ne sont pas les seuls à véhiculer la temporalité dans la langue, et qu'il faut donc prendre en compte d'autres éléments :

> « Pour [...] essayer de décrire la spécificité de chacun des temps morphologiques, grammairiens et linguistes ont progressivement distingué le temps (externe au procès) et l'aspect (structure temporelle interne au procès), le temps absolu (le procès se situe par rapport au moment de l'énonciation) et le temps relatif (le procès est situé relativement à un autre procès) » (Gosselin [2005], p. 31).

4.6 L'aspect

Les sections précédentes ont montré qu'il n'existe pas de relation simple entre les temps grammaticaux (ou linguistiques) et le temps notionnel (ou chronique). Il y a par conséquent au moins une autre notion qui intervient dans l'interprétation des temps grammaticaux : elle est désignée sous le nom d'*aspect*.

La distinction entre temps et aspect peut être trouvée dans la définition de Guillaume [1964], citée par Wilmet [1997] (p. 310-311) :

> « Le verbe est un sémantème qui *implique* et *explique* le temps.
> Le *temps impliqué* est celui que le verbe emporte avec soi, qui lui est inhérent, fait partie intégrante de sa substance et dont la notion est indissolublement liée à celle de verbe. Il suffit de prononcer le nom d'un verbe comme « marcher » pour que s'éveille dans l'esprit, avec l'idée d'un procès, celle du temps destiné à en porter la réalisation.
> Le *temps expliqué* est autre chose. Ce n'est pas le temps que le verbe retient en soi par définition, mais le temps divisible en moments distincts – passé, présent, futur et leurs interprétations – que le discours lui attribue.
> Cette distinction du *temps impliqué* et du *temps expliqué* coïncide exactement avec la distinction de l'*aspect* et du *temps*.
> Est de la nature de l'*aspect* toute différenciation qui a pour lieu le temps impliqué.
> Est de la nature du *temps* toute distinction qui a pour lieu le temps expliqué. »

Bien qu'elle ne soit pas vraiment opératoire, cette définition illustre bien la différence entre le temps notionnel ou expliqué (passé-présent-futur) et l'aspect ou temps impliqué (la dimension temporelle intrinsèque du procès).

L'aspect est une matière sujette à controverses depuis longtemps en linguistique. De nombreuses années de recherches n'ont pas suffi à amener l'unanimité sur le sujet. Il existe d'ailleurs de nombreuses terminologies, souvent différentes et pas toujours compatibles les unes avec les autres. Ce constat est révélateur de la difficulté à appréhender et à caractériser cette notion[15]. Ainsi, la comparaison de quelques définitions données pour l'aspect, révèle rapidement que ce concept n'est pas employé de manière uniforme.

L'aspect tel qu'il est abordé par Guillaume [1964], est à mettre en relation avec les types de procès, tels que nous les avions présentés à la section 4.4. De ce point de vue, une des nomenclatures parmi les plus connues est celle de Vendler [1957] qui propose quatre catégories, que nous illustrons à l'aide d'exemples inspirés de ceux donnés par Bestougeff et Ligozat [1989] (p. 24) :

- *États* : procès qui n'ont ni début, ni fin et qui ne sont pas compatibles avec le trait ponctuel (« Paul connaît le fonctionnement du système ») ;
- *Activités* : procès sans fin déterminée, de nature durative, non conclusifs (« Paul pro-

[15] La présentation qui en est faite ici ne se veut pas exhaustive (se référer à des études telles que Mascherin [2007]), mais a plutôt pour objectif de faire passer l'intuition de la notion d'aspect, et de son importance en ce qui concerne l'interprétation de la temporalité.

gramme en Java ») ;
- *Accomplissements* : par opposition à la classe précédente, ces procès dynamiques sont conclusifs (« Paul programme l'algorithme de Dijkstra ») ;
- *Achèvements* : caractérisés par leur caractère ponctuel (« Paul a perdu son mot de passe »).

Ces catégories sont caractérisées par un ensemble de traits binaires tels que ponctuel / duratif, conclusif / non-conclusif, inchoatif[16] / non-inchoatif, *etc.*

D'autre part, Comrie [1976] propose lui aussi une définition de l'aspect : « aspects are different ways of viewing the internal temporal constituency of a situation ». Il présente trois grands types d'aspect, qu'il spécialise ensuite. Cette catégorisation est reprise ci-dessous et illustrée au moyen d'exemples, à nouveau inspirés par ceux de Bestougeff et Ligozat [1989] (p. 21-22) pour le français :

1. *Perfectif*
   « Le fichier a été sauvegardé à 8h35 »
2. *Imperfectif*
   2.1. *Habituel*
      « Il y a encore dix ans, on *utilisait* des disquettes »
   2.2. *Continu*
   2.2.1. *Non progressif*
      « Autrefois, on *disposait* de deux types de fichiers »
   2.2.2. *Progressif*
      « Je *suis en train d'éditer* mon fichier »
3. *Parfait*
   3.1. *État résultant*
      « Le système *a été rechargé* : tout fonctionne »
   3.1.1. *Expérience*
      « *As-tu* déjà *programmé* en Java ? »
   3.1.1.1. *Situation persistante*[17]
   3.1.1.1.1. *Passé récent*
      « *J'ai copié* ce fichier il y a une heure »

La confrontation des définitions de Vendler [1957] et de Guillaume [1964] avec celle de Comrie [1976] démontre que l'aspect réfère à plusieurs notions différentes. Cet écart est également expliqué par Mascherin [2007] (p. 58) :

> « L'aspect est véhiculé par différents éléments linguistiques incluant des tiroirs, des affixes dérivationnels, des périphrases verbales, des adverbes, des éléments lexicaux. Du fait de ces multiples éléments on distingue différentes catégories aspectuelles (notamment le mode de procès ou aspect lexical et l'aspect flexionnel). L'aspect exprime les

[16] Marque le commencement d'une action ou d'une activité, ou l'entrée dans un état.
[17] D'après Comrie, relayé par Bestougeff et Ligozat [1989], la situation persistante correspond à un emploi propre à l'anglais du present perfect et du pluperfect.

> notions comme la durée, la répétition, etc. des procès (ou des actions ou des événements, selon les théories). Les catégories de la temporalité et de l'aspectualité sont distinctes mais fortement intriquées au sein du discours, c'est pourquoi on parle du champ aspectuo-temporel. »

Parmi les catégories aspectuelles évoquées, on distingue donc principalement d'une part la dimension temporelle intrinsèque du procès et d'autre part la manière dont est exprimé ce procès (est-ce qu'il est en cours de réalisation ou déjà réalisé, etc.). La première notion correspond à ce dont il a déjà été question sous le nom de *type de procès* (voir section 4.4). Les termes *aspect lexical*, *aktionsart* ou encore *mode d'action et de procès* sont également utilisés. La seconde notion est parfois appelée *aspect grammatical* ou *aspect flexionnel*.

Gosselin [2005] distingue lui aussi très clairement aspect lexical et aspect grammatical (voir section 4.7.6). L'*aspect lexical* concerne le procès tel qu'il a été conçu, il détermine le type de procès et est marqué par le verbe. Il est composé de quatre classes de procès (inspirées de Vendler) : les *états* (« être malade », « aimer la confiture »), les *activités* (« marcher », « manger des fruits »), les *accomplissements* (« manger une pomme ») et les *achèvements* (« atteindre un sommet »). L'*aspect grammatical* représente la manière dont est montré le procès. Il peut être *aoristique* (ou *perfectif*) lorsque le procès est montré dans son intégralité (« il traversa le carrefour »), *inaccompli* (ou *imperfectif*) lorsque seule une partie du procès est présentée, *accompli* quand c'est l'état résultant du procès qui est montré (« il a terminé son travail depuis 10 minutes ») ou encore *prospectif* si c'est la phase préparatoire du procès qui est visible (« il allait traverser le carrefour »).

D'après Mascherin [2007], les éléments de la langue qui expriment l'aspect en français sont multiples :

> « Les trois sources qui véhiculent et qui jouent un rôle important dans la détermination de l'aspect sont : le lexème verbal ou le prédicat verbal, la flexion verbale et les périphrases aspectuelles, et enfin les compléments du verbe qui peuvent être des compléments à valeur aspectuelle ou simplement des compléments d'objet. » (p. 120)

Mani et Schiffman [2005] donnent un exemple qui illustre bien le fait que certains verbes puissent appartenir à plusieurs classes aspectuelles, suivant le contexte d'utilisation : dans « le régiment marcha (jusqu'à Saigon) » où le procès est passé d'une activité (« marcher ») à un accomplissement (« marcher jusqu'à Saigon »).

Il apparaît donc que la notion d'aspect est très importante, peut-être même plus que les temps. En effet, Lyons [1980], cité par Mascherin [2007], remarque que :

> « l'aspect est beaucoup plus répandu que le temps dans les langues du monde : il y a de nombreuses langues qui ne possèdent pas de temps grammaticaux mais il y en a fort peu qui ne possèdent pas d'aspect » (p. 325).

En pratique, il existe tout de même de nombreuses langues, dont le français, pour lesquelles la temporalité est exprimée à la fois par le temps et l'aspect. Si le temps (notionnel) concerne la distinction

passé/présent/futur, et l'aspect la manière dont est exprimé le procès (par exemple, est-ce qu'il est en cours de réalisation ou déjà réalisé), ils s'interprètent cependant dans le même espace (la ligne du temps par exemple). L'interprétation d'un temps grammatical donne donc une indication quant au temps notionnel et à la valeur aspectuelle du procès.

Par conséquent, si l'on perçoit bien l'intérêt que représente la détermination de l'appartenance d'un procès à une classe aspectuelle en termes d'interprétation temporelle, il faut cependant constater, comme le font Bestougeff et Ligozat [1989], que du point de vue de l'analyse automatique, il est compliqué de créer des algorithmes pour mener à bien la reconnaissance de l'aspect.

4.7 Modèles des temps verbaux

Le rôle des temps verbaux (ou morphologiques) a déjà été évoqué dans les sections précédentes. Il s'agit, d'un point qui a focalisé de nombreuses études sur l'analyse du temps en langage naturel. Celles-ci ont, depuis longtemps, tenté de décrire, d'expliquer et de représenter le fonctionnement du système des temps verbaux. Pour atteindre cet objectif, les différents modèles manipulent des repères temporels, qui sont disposés sur l'axe du temps afin de définir leur fonctionnement. Ces repères représentent généralement le moment de l'énonciation (S), la zone temporelle relative à l'événement (E), ainsi que, dans certains cas, un point de référence (R). Chaque temps se voit ainsi attribuer une configuration particulière qui, en fonction du placement de ces différents repères, détermine la chronologie entre l'acte de parole[18], l'événement en question, et une éventuelle autre zone temporelle par rapport à laquelle les autres éléments du modèle sont situés.

Les sections suivantes, guidées en partie par Schwer [2009a], reprennent un aperçu des modèles des temps verbaux, principalement de l'indicatif, et de leur évolution au fil des différentes propositions. Le premier modèle, celui de Port-Royal (Arnauld et Lancelot [1810], est basé sur des repères sous la forme de points. Les modèles suivants, avec Girard [1747] (Section 4.7.2) ou Beauzée [1767] (Section 4.7.3), tentent de remédier aux limites de cette représentation en donnant une *épaisseur* à certains points. Reichenbach [1947] (Section 4.7.4), probablement le plus souvent cité de tous, propose un modèle qui, pour les neuf formes *fondamentales*, n'emploie par contre que des points. Plusieurs autres modèles prennent Reichenbach comme base et avancent de nouvelles propositions. C'est le cas avec Vet [2007] (Section 4.7.5), et surtout Gosselin [1996] (Section 4.7.6). Ce dernier, en plus d'avoir l'originalité d'être basé complètement sur des intervalles, propose une approche beaucoup plus large que les autres en intégrant les notions d'aspect et de circonstants temporels. Finalement, après l'exposé de ces théories, la section 4.7.7 examine les perspectives éventuelles concernant leur automatisation et leur utilisation en extraction d'informations.

[18] Ou d'écriture.

4.7.1 Arnauld et Lancelot, la grammaire de Port-Royal

La grammaire de Port-Royal (Arnauld et Lancelot [1810]), dont l'ensemble des temps est repris au tableau 4.3[19], est parue initialement en 1660. Elle distingue trois temps simples : le présent, le passé et le futur. Ces temps, appelés *simples dans le sens*, impliquent deux repères temporels : le temps de l'événement (E) et celui de l'énonciation (S). La relation entre ces deux points peut être celle d'antériorité, de simultanéité ou de postériorité. Afin de distinguer le passé composé (prétérit défini) du passé simple (prétérit indéfini ou aoriste), une règle dite *des 24 heures* est énoncée. Elle stipule que le prétérit indéfini ne peut être utilisé correctement que pour désigner « un temps qui soit au moins éloigné d'un jour de celui auquel nous parlons » (Arnauld et Lancelot [1810], p. 338). Cela implique l'usage d'une seconde relation entre les repères temporels afin de caractériser la proximité ou l'éloignement de ceux-ci.

Temps	Simples dans le sens	Composés dans le sens
Présent	présent *je soupe* E,S	prétérit imparfait *je soupais* E,R - S
Passé	prétérit défini *j'ai soupé* E - S + proximité	plus-que-parfait *j'avais soupé* E - R - S
	prétérit aoriste *je soupai* E - S + éloignement	(non décrit)
Futur	futur *je souperai* S - E	futur parfait *j'aurai soupé* S - E - R

Tableau 4.3 : Les temps verbaux dans la grammaire de Port-Royal.

Il existe également des temps *composés dans le sens* qui utilisent un repère temporel supplémentaire (R). Tout d'abord, le passé avec rapport au présent, le prétérit imparfait, qui « ne marque pas la chose simplement et proprement comme faite, mais comme présente à l'égard d'une chose qui est déjà néanmoins passée » (Arnauld et Lancelot [1810], p. 339). Ensuite, il existe un temps qui marque doublement le passé, le plus-que-parfait. Il doit être interprété comme « passé en soi, mais aussi comme passé à l'égard d'une autre chose qui est aussi passée » (Arnauld et Lancelot [1810], p. 339). Enfin, le troisième temps composé, le futur parfait, marque l'avenir avec rapport au passé. L'action est décrite comme « future en soi, et comme passée au regard d'une autre chose à venir, qui la doit suivre » (Arnauld et Lancelot [1810], p. 339). Afin de conserver la symétrie avec le passé, un quatrième temps aurait encore pu être ajouté : un futur avec rapport au présent. Cependant dans l'usage de la langue, celui-ci se confond au futur simple.

[19] Les conventions de notation suivantes sont utilisées : « , » sépare deux références à des points qui se confondent alors que « - » sépare les références à deux points distincts sur la ligne du temps.

4.7.2 L'abbé Girard

La grammaire de Girard [1747], dont la partie concernant les temps verbaux est exposée dans Schwer [2009a,b], est la première à faire la distinction entre le temps objectif, naturel et le temps linguistique, représenté. Cette distinction permet de ne plus considérer les points temporels comme n'ayant aucune épaisseur, mais de les envisager comme des unités qui possédent une certaine étendue temporelle[20]. La notion d'intervalle est dès lors introduite, sous le nom de *période*. Par conséquent, le nombre de relations possibles entre les repères temporels augmente, avec entre autres l'inclusion, et le sens de la simultanéité qui est élargi afin de dépasser la stricte égalité. Les points deviennent des points épais, ou *extensions temporelles*, dont la particularité est de pouvoir se transformer en intervalles lorsqu'un repère doit y être inséré.

Le système des temps verbaux exploite donc, en plus du repère de l'énonciation (S), trois éléments : l'extension temporelle de l'événement (E)[21], le période (P)[22] qui se substitue à la *règle des 24 heures* de Port-Royal (voir section 4.7.1) et aux notions de proximité et d'éloignement, et enfin l'extension temporelle reliée à un autre événement (R). Le tableau 4.4 reprend les huit temps distingués par l'abbé Girard, dont la répartition est assez similaire à celle de la grammaire de Port-Royal.

Temps	Absolus	Relatifs
Présent	*je soupe* E,S	*je soupais* E,R - S
Prétérit	*j'ai soupé* E - S	*j'avais soupé* E - R - S
Aoriste	*je soupai* E,[P] - S	*j'eus soupé* [P - E - R - P] - S
Futur	*je souperai* S - E	*j'aurai soupé* S - E - R

Tableau 4.4 : Les temps verbaux chez l'abbé Girard.

4.7.3 Le modèle de Beauzée

Le modèle proposé par Beauzée [1767] présente un changement majeur par rapport aux propositions précédentes. Celles-ci exploitaient la relation entre le moment de l'énonciation (S) et le temps de l'événement (E) et s'axaient principalement sur le problème de la distinction entre le passé simple et le passé composé, qui ne peut cependant être résolue à l'aide de cette relation.

Le modèle de Beauzée s'appuie lui sur trois repères temporels : l'époque de l'événement (E), l'époque ou le période de comparaison (R) et l'époque du moment de l'énonciation (S). L'*époque* est définie

[20] « Contrairement au *maintenant* du temps physique, insaisissable et sans épaisseur, le repère du maintenant, ainsi que tous les temps représentés possèdent une étendue stable et permanente » (Schwer [2009a]).

[21] Correspond au temps contenu entre le début et la fin de l'événement.

[22] Nous noterons les bornes supérieure et inférieure de P à l'aide des signes « [P », « P] » ou encore « [P] » lorsque cet intervalle ne contient aucun autre élément particulier.

comme un repère ponctuel alors que *le période* est duratif[23]. L'innovation du modèle est l'utilisation de deux relations, celle de E par rapport à R et la seconde de R par rapport à S, déterminant ainsi les deux principales divisions entre les temps. Notons que la relation entre E et S n'est pas directement exprimée, mais qu'elle peut être dérivée par transitivité à partir des deux autres relations.

Le rapport entre E et R peut être constitué par une relation de simultanéité, d'antériorité ou de postériorité, ce qui définit trois classes de temps : les *Présents*, qui « expriment la simultanéité d'existence à l'égard de l'époque de comparaison » (Beauzée [1767], p. 428) ; les *Prétérits* qui « expriment l'antériorité d'existence à l'égard de l'époque de comparaison » (Beauzée [1767], p. 429) ; et les *Futurs* qui « expriment la postériorité d'existence à l'égard de l'époque de comparaison » (Beauzée [1767], p. 429). Comme on peut le constater, et contrairement aux modèles précédents, la description des temps fait systématiquement appel à l'époque de comparaison (R).

La deuxième division générale des temps est basée sur la manière d'envisager l'époque de comparaison, soit comme *indéterminée* (relative à aucune époque précise ou à une époque dite *indéfinie*), soit comme *déterminée* (relative à une époque précise ou à une époque dite *définie*). Les huit principaux temps de l'indicatif, illustrés au tableau 4.5 rentrent tous dans la catégorie des temps définis[24].

Il existe enfin une troisième division qui consiste à examiner, pour ces temps définis, le positionnement de l'époque de comparaison (R) et le moment de l'énonciation (S). Cette comparaison peut elle aussi déboucher sur une relation de simultanéité, d'antériorité ou de postériorité donnant naissance aux temps *actuels*, *antérieurs* et *postérieurs*. Dans le cas des temps antérieurs, R peut prendre la forme d'un repère simple (R est une époque) ou périodique (R est un période). Les temps futurs ne concernent pas les temps de l'indicatif[25], ils ne sont donc pas repris dans le tableau 4.5 qui récapitule le modèle de Beauzée[26].

4.7.4 Le modèle de Reichenbach

Comme pour le modèle de Beauzée, c'est l'utilisation permanente d'un point de référence qui est une des caractéristiques intéressantes du modèle de Reichenbach [2005] (initialement publié dans Reichenbach [1947]). Pour représenter les *temps fondamentaux*, il fait donc lui aussi intervenir trois points sur la ligne du temps : le moment de l'énonciation (S), le moment de l'événement (E) et le moment de référence (R).

[23] « On donne à ces point fixes de la succession de l'existence ou du temps, le nom d'*époques* ; [...] parce que ce sont des instants dont on arrête, en quelque manière, la rapide mobilité, pour en faire des lieux de repos, d'où l'on observe pour ainsi dire, ce qui coexiste, ce qui précède et ce qui suit. On appelle *période*, une portion de temps dont le commencement et la fin sont déterminés par des époques, [...] une portion de temps bornée de toutes parts, est comme un espace autour duquel on peut tracer un chemin, pour observer ce qui y est enfermé et ce qui l'environne. » (Beauzée [1767], p. 425)

[24] Vetters [1996] explique à ce sujet : « [...] les temps actuels n'ont pas d'expression propre : ils emploient toujours la forme du *temps indéfini* correspondant. Le tiroir appelé couramment *indicatif présent* est un *présent indéfini* car il peut être employé comme *présent actuel* [...], comme *présent antérieur* [...], comme *présent postérieur* [...] ou avec abstraction de toute époque [...]. »

[25] Les formes du futur chez Beauzée sont « je dois souper », « je devais souper » et « je devrai souper ». Portine [1995] en conclut que ces formes verbales ne peuvent être considérées comme des temps verbaux car *devoir* n'est pas un auxiliaire.

[26] Un période est noté à l'aide des symboles « [X » et « X] » ou encore « [X] » lorsqu'aucun élément particulier n'y est contenu. X est un repère temporel quelconque du modèle. Le symbole « ? » désigne l'indétermination qui peut survenir quant à la disposition relative de deux repères temporels et recouvre donc les trois possibilités « X-Y », « Y-X » et « X,Y »

Temps	Présent E,R	Prétérit E - R
Actuel R,S	*je soupe* E,R,S	*j'ai soupé* E - R,S
Antérieur simple (R ponctuel) R - S	*je soupais* [E - R - E] ? S	*j'avais soupé* E - R - S
Antérieur périodique (R périodique) [R]- S	*je soupai* [R - E - R] - S	*j'eus soupé* E - [R] - S
Postérieur S - R	*je souperai* S - E,R	*j'aurai soupé* E ? S - R

Tableau 4.5 : Les temps verbaux chez Bauzée.

Reichenbach précise que, pour certains temps (les *temps étendus*), il est possible de représenter l'étendue temporelle de l'événement. Le point E se mue alors en un intervalle qui peut également représenter la répétition plutôt que la durée. Si en anglais ces *temps étendus* existent[27], ils ne sont pas vraiment utilisés en français. La durée ou la répétition sont plus souvent exprimés à l'aide d'adverbes tels que « toujours » ou « habituellement ». L'imparfait (« Je voyais Jean ») peut cependant être considéré comme un temps *étendu* là où le passé simple (« Je vis Jean ») ne l'est pas.

Tout comme chez Beauzée, les deux relations exploitées sont celles entre R et S, d'une part et entre E et R d'autre part. Notons cependant que l'utilisation de ces relations est inversée[28]. La première division, R par rapport à S, permet de distinguer les temps passés (R-S), présents (R,S) et futurs (S-R). La seconde distinction, entre E et R, exprime le caractère antérieur (E-R), simple (E,R) ou postérieur (R-E) des temps. Le résultat du croisement de ces deux axes d'analyse donne neuf *formes fondamentales*, exposées au tableau 4.6. Les noms des temps ne correspondant pas nécessairement aux noms traditionnels, ceux-ci sont donc indiqués dans ce tableau et illustrés par les exemples de la figure 4.1, traduits de Reichenbach [2005], (p. 290). Dans certains cas, la disposition de R par rapport à S d'une part, et de R par rapport à E d'autre part, ne permet pas de déduire le placement relatif de E et S. C'est le cas du passé postérieur (conditionnel présent) et du futur antérieur. Dans ce cas, l'indétermination est notée par le signe « ? » qui recouvre les trois possibilités « E-S », « S-E » et « S,E ».

Notons que, dans ce tableau 4.6, le plus-que-parfait (E-R-S) a été ajouté ultérieurement par Vet [2007]. On constate que le futur est placé dans deux cases différentes. Reichenbach évoque la possibilité que les deux configurations du futur correspondent, d'une part, à sa forme périphrastique (« je vais voir ») pour R,S-E et, d'autre part, au futur simple (« je verrai ») pour S-E,R. Cependant, Vet [2007] a montré que le futur périphrastique peut en réalité s'accommoder des deux configurations.

On remarque également que le point R permet de différencier clairement le passé simple (E,R-S), le passé composé (E-R,S) et le plus-que-parfait (E-R-S). Par contre, la distinction entre l'imparfait et le passé simple ne peut pas être faite de façon satisfaisante. Reichenbach indique que la distinction entre ces deux temps peut être réalisée en considérant l'aspect, ponctuel pour le passé simple et duratif pour

[27] Par exemple le *simple past* « I saw John » devient « I was seeing John » pour le *simple past extended*.

[28] Beauzée envisageait le rapport entre E et R comme première distinction entre les temps, pour ensuite caractériser ceux-ci à l'aide de la relation R-S.

Temps	Antérieur E - R	Simple E,R	Postérieur R - E
Passé	(Passé antérieur) (Plus-que-parfait)	(Passé simple) (Imparfait)	(Conditionnel présent)
R - S	E - R - S *j'eus soupé / j'avais soupé*	E,R - S *je soupai / je soupais*	R - E ?S *je souperais*
Présent	(Passé composé)	(Présent)	(Futur périphrastique) (Futur simple)
R,S	E - R,S *j'ai soupé*	E,R,S *je soupe*	R,S - E *je vais souper / je souperai*
Futur	(Futur antérieur)	(Futur simple) (Futur périphrastique)	(-)
S - R	E ?S - R *j'aurai soupé*	S - E,R *je souperai / je vais souper*	S - R - E -

Tableau 4.6 : Les temps verbaux chez Reichenbach.

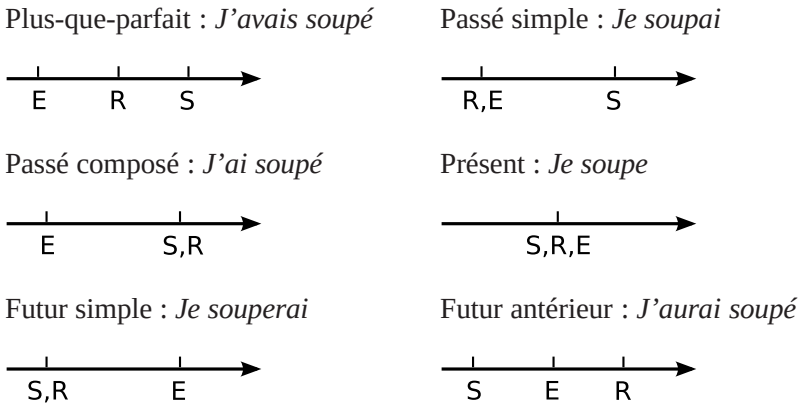


Figure 4.1 : Illustration de la disposition des points temporels chez Reichenbach.

l'imparfait. Cependant, cette caractéristique durative (ou inaccomplie) n'est pas exprimable à l'aide de points. De même, la différence entre le plus-que-parfait et le passé antérieur n'est pas exprimable par ce modèle.

Autre point particulier, la configuration des repères temporels pour le passé postérieur (conditionnel présent) et le futur antérieur débouche sur une indétermination en ce qui concerne le placement de E par rapport à S. Selon Reichenbach, l'indétermination entre les trois placements possibles[29] n'est cependant pas réelle. Vet [2007] fait en effet remarquer que cette situation devrait être le signe de l'ambiguïté de ces temps, ce qui n'est pas le cas. En réalité, sur les trois configurations du futur antérieur, deux n'existent pas dans la langue (seul S-E-R représente correctement le futur antérieur). La phrase « Il aura soupé maintenant » (E,S-R) a une interprétation modale (je crois qu'il a mangé maintenant), et « Il aura mangé hier » ne peut aussi se comprendre que via une interprétation modale.

Après l'affectation des différents temps du français aux cases du tableau, on constate qu'une de celles-ci reste vide. Il s'agit du futur postérieur qui n'est pas attesté dans la langue. Au contraire, certains temps n'ont pas pu être casés. C'est le cas du conditionnel passé (« j'aurais soupé ») qui n'est pas représentable dans le modèle de Reichenbach. Vet [2007] indique à ce propos qu'il serait

[29] Soit les configurations suivantes : E-S, S-E et E,S.

nécessaire de disposer d'un second point de référence (R') afin de représenter ce temps sous la forme de R ?E-R'-S.

Comme le fait remarquer Girault [2007], le caractère absolu et relatif des temps peut être discerné en observant la relation entre R et E. Si R et E sont simultanés, alors l'événement s'exprime par un temps absolu. Au contraire, si R et E sont dissociés, alors l'événement s'exprime par un temps relatif.

4.7.5 Le modèle de Vet

Le modèle de Vet [2007], aussi exposé dans Schwer [2009a], constitue une évolution du modèle de Reichenbach. Les temps verbaux sont toujours organisés à l'aide de repères temporels (S, E et R). Par contre, l'accent est porté sur la zone temporelle qui concerne l'événement. Celle-ci se compose de trois parties : la phase *prospective* (E'), la phase *ipse* (E) et enfin la phase *résultante* (E"). Les trois phases sont évidemment liées : « il ne peut y avoir de phase résultante sans réalisation de la phase *ipse*, la phase prospective ne peut se concevoir que dans l'intention de la réalisation (comme anticipation) de la phase *ipse* » (Schwer [2009a], p. 18). La distinction de trois phases de E est formalisée dans le modèle à l'aide d'un nouveau repère U, qui indique la phase de l'événement qui est utilisée. Le modèle exploite la relation entre R et S afin de délimiter les temps passés (R-S) et les temps présents (R,S). D'autre part, la relation entre U et R permet de distinguer trois cas : les temps antérieurs (U-R), actuels (U,R) et postérieurs (R-U). Évidemment, il faut tenir compte du fait que U peut coïncider aux trois différentes phases de l'événement, ce qui multiplie par trois les possibilités offertes par ces deux relations. C'est donc un ensemble de 18 temps qui est proposé, au tableau 4.7, par le modèle de Vet. Parmi ceux-ci, certains ne sont cependant pas observés dans la langue.

U=E	Antérieur E - R	Actuel E,R	Postérieur R - E
Présent R,S	*je soupai/j'ai soupé* E - R,S	*je soupe* E,R,S	*je souperai/je vais souper* R,S - E
Passé R - S	*j'avais soupé* E - R - S	*je soupais* E,R - S	*je souperais/j'allais souper* R - E ?S
U=E'	Antérieur E' - R	Actuel E',R	Postérieur R - E'
Présent R,S	- E' - R,S	*je vais souper* E',R,S	- R,S - E'
Passé R - S	- E' - R - S	*j'allais souper* E',R - S	- R - E' ?S
U=E"	Antérieur E" - R	Actuel E",R	Postérieur R - E"
Présent R,S	*j'eus soupé/j'ai eu soupé* E" - R,S	*j'ai soupé* E",R,S	*j'aurai soupé* R,S - E"
Passé R - S	*j'avais eu soupé* E" - R - S	*j'avais soupé* E",R - S	*j'aurais soupé* R - E" ?S

Tableau 4.7 : Les temps verbaux chez Co Vet.

4.7.6 Le modèle des intervalles de Gosselin

Le modèle de Gosselin [1996], va au-delà des temps verbaux, sur le rôle joué par diverses autres marques de la temporalité (aspect grammatical, types de procès, circonstanciels, etc.).

Le modèle des temps verbaux de Gosselin entend combler certaines lacunes du modèle de Reichenbach qu'il prend comme point de départ. Il s'attaque entre autres au problème de la distinction du passé simple et de l'imparfait. Les différents points sont remplacés par des intervalles. Ceux-ci permettent d'exprimer certaines différences aspectuelles que les points ne peuvent exprimer[30]. En plus des repères habituels, un nouvel intervalle, circonstanciel, est également introduit. Cela porte donc le nombre d'intervalles manipulés à quatre :

- *l'intervalle d'énonciation* : [01,02], les limites temporelles de l'acte physique d'énonciation ;
- *l'intervalle du procès* : [B1,B2], la portion de l'axe temporel qui est occupé par une « situation » ;
- *l'intervalle de référence* : [I,II], correspond à ce qui est montré sur l'axe temporel, joue un rôle qui est en partie analogue au point R de Reichenbach ;
- *l'intervalle circonstanciel* : [ct1,ct2], marqué par les compléments de localisation temporelle (« mardi dernier ») et les compléments de durée (« pendant 3 heures »).

La construction d'une représentation à l'aide de ces intervalles suit certaines règles. Pour chaque énoncé, il n'est possible d'attribuer qu'un et un seul intervalle d'énonciation [01,01]. Pour chaque proposition, il faut attribuer au moins un intervalle de procès [B1,B2] et au moins un intervalle de référence [I,II]. Et enfin, à chaque complément circonstanciel de temps est raccroché au moins un intervalle circonstanciel [ct1,ct2].

Pour caractériser ces intervalles et les comparer, il est nécessaire de définir les relations possibles entre leurs bornes. Soit deux bornes i et j, il existe trois relations primitives :

- $i = j$: coïncidence ;
- $i \propto j$: i précède j, mais est infiniment proche ;
- $i \{ j$: i précède j, mais pas de façon immédiate.

À partir de ces relations de base, d'autres relations complexes sont ensuite définies :

- $i < j =_{df} (i \propto j) \vee (i \{ j)$;
- $i > j =_{df} j < i$;
- $i \leq j =_{df} (i < j) \vee (i = j)$;
- $i \geq j =_{df} j \leq i$.

Les relations entre intervalles sont ensuite exprimées à partir de ces relations entre bornes :

- antériorité : [i,j] ANT [k,l] $=_{df}$ $j < k$;
- postériorité : [i,j] POST [k,l] $=_{df}$ $l < i$;

[30] Par exemple l'aspect perfectif du passé simple et l'aspect imperfectif de l'imparfait.

- simultanéité : [i,j] SIMUL [k,l] $=_{df}$ $(i \leq l)$ & $(k \leq j)$;
- recouvrement : [i,j] RE [k,l] $=_{df}$ $(i < k)$ & $(j > l)$;
- coïncidence : [i,j] CO [k,l] $=_{df}$ $(i = k)$ & $(j = l)$;
- accessibilité : [i,j] ACCESS [k,l] $=_{df}$ $(i \leq k)$ & $(j \geq l)$;
- succession : [i,j] SUCC [k,l] $=_{df}$ $k < i$;
- précédence : [i,j] PREC [k,l] $=_{df}$ $i < k$.

Les temps verbaux, nous l'avons déjà évoqué, donnent deux types d'indications. En plus de la caractérisation des temps absolus et relatifs, des informations aspectuelles sont également *codées* par la forme verbale. Ces caractéristiques linguistiques peuvent être exprimées au moyen des relations entre intervalles définies jusqu'ici.

En ce qui concerne l'ASPECT, on y distingue deux catégories : l'aspect lexical, représente le procès comme il a été conçu, et l'aspect grammatical, le procès tel qu'il est montré.

L'*aspect lexical* détermine le type de procès et est marqué par le verbe. Trois critères permettent de classer les procès selon leur aspect lexical :

1. le type de bornes
   - intrinsèques, pour les procès perfectifs (téliques, c'est-à-dire qui ne peuvent être que recommencés, pas continués) : [Bi1,Bi2]
   - extrinsèques, pour les procès imperfectifs : [Be1,Be2]
2. les relations entre les bornes
   - si le procès est ponctuel : B1 $\propto$ B2
   - si le procès est non ponctuel : B1 $\{$ B2
3. le fait que l'intervalle de procès contienne :
   - une série de changements
   - une absence de changements
   - un changement atomique

L'aspect lexical est composé de quatre classes de procès (inspirées de Vendler), qui sont définies à partir de ces critères :

1. État [Be1 $\{$ Be2] [absence de changements] (« être malade », « aimer la confiture ») ;
2. Activité [Be1 $\{$ Be2] [série de changements] (« marcher », « manger des fruits ») ;
3. Accomplissement [Bi1 $\{$ Bi2] [série de changements] (« manger une pomme ») ;
4. Achèvement [Bi1 $\propto$ Bi2] [changement atomique] (« atteindre un sommet »).

L'*aspect grammatical* joue sur la relation entre l'intervalle de référence ([I,II]) et l'intervalle du procès (B1,B2). Quatre aspects grammaticaux de base sont distingués :

1. *aoristique (perfectif)* : [I,II] CO [B1,B2], c'est-à-dire I=B1 et II=B2, les deux intervalles coïncident, c'est l'*aspect global*, qui montre le procès dans son intégralité (« il traversa le carrefour ») ;
2. *inaccompli (imperfectif)* : [B1,B2] RE [I,II], c'est-à-dire B1<I et II<B2, l'intervalle de référence est inclus dans celui du procès, seule une partie du procès est présentée (« il traversait le

carrefour ») ;

3. *accompli* : [I,II] POST [B1,B2], c'est-à-dire B2<I, montre l'état résultant du procès (« il a terminé son travail depuis 10 minutes ») ;
4. *prospectif* : [I,II] ANT [B1,B2], c'est-à-dire II<B1, présente la phase préparatoire du procès (« il allait traverser le carrefour »).

En ce qui concerne les TEMPS, et en particulier les *temps absolus*, il faut s'intéresser à la relation entre l'intervalle de référence ([I,II]) et le moment de l'énonciation ([01,02]). Il s'agit d'une différence intéressante par rapport aux modèles traditionnels[31] qui définissent plutôt cette caractéristique relativement aux moments du procès (E) et de l'énonciation (S).

1. *temps passé* : [I,II] ANT [01,02], c'est-à-dire II < 01 ;
2. *temps présent* : [I,II] SIMUL [01,02], les deux intervalles coïncident ou se chevauchent, c'est-à-dire une des configurations suivantes : 01-I-II-02 ou 01-I-02-II ou I-01-II-02 ou I-01-02-II ;
3. *temps futur* : [I,II] POST [01,02], c'est-à-dire 02 < I ;

Les *temps relatifs* sont exprimés par la relation entre deux intervalles de référence, celui de la principale et celui de la subordonnée :

- *antériorité* : [I',II'] ANT [I,II], c'est-à-dire II<I' ;
- *simultanéité* : [I',II'] SIMUL [I,II], c'est à dire une des configurations suivantes : I-I'-II'-II ou I-I'-II-II' ou I'-I-II'-II ou I'-I-II-II' ;
- *postériorité* : [I',II'] POST [I,II], c'est-à-dire II'<I.

Par exemple, dans « Pierre disait que, lundi, Luc aurait terminé son travail depuis longtemps », la contrainte concernant le temps relatif est celle entre les intervalles de référence de la principale et de la subordonnée ([I',II'] POST [I,II])[32]. Dans ce genre de cas, les théories traditionnelles exploitaient plutôt la relation entre les deux procès.

La caractérisation des temps verbaux de l'indicatif, dans une principale ou une indépendante, est reprise au tableau 4.8[33] et est effectuée à l'aide des critères d'aspect (grammatical) et de temps absolu. Le tableau 4.9 adopte une présentation selon un format plus proche de celui utilisé pour présenter les autres modèles. Les relations entre le procès et le moment d'énonciation d'une part, et entre procès différents d'autre part peuvent être obtenues indirectement. Dans certains cas, il est possible que des relations restent indéterminées.

Comme on peut le constater, les deux critères d'aspect et de temps absolu ne permettent pas de discriminer complètement les différents temps. La caractérisation en termes de temps relatif n'a pas été intégrée à ce stade. En effet, pour prendre ce point en compte, il faut que deux intervalles de référence soient comparés, ce qui est le cas lorsque l'on considère le procès conjointement avec un autre procès ou une expression circonstancielle provenant du contexte.

[31] Entre autres Port-Royal et l'abbé Girard.

[32] Dans cet exemple, la relation entre l'intervalle de référence de la subordonnée et l'intervalle d'énonciation n'est pas connue.

[33] Les codes ont été ajoutés.

Temps morphologiques	Temps absolus	Aspect	Exemples
Passé simple (PS)	passé	aoristique	*Luc sortit*
Imparfait (IMP1)	passé	inaccompli	*Luc se promenait*
(IMP2)	passé	aoristique	*Le lendemain, Luc quittait Paris* (imparfait de rupture)
Présent (P1)	présent	inaccompli	*Luc se promène*
(P2)	futur	aoristique	*Je termine demain*
(P3)	passé	aoristique	*Cette année-là Colomb découvre l'Amérique* (présent historique)
Futur (F1)	futur	aoristique	*Je viendrai*
(F2)	futur	inaccompli	*A huit heures, Luc dormira depuis longtemps*
Passé composé (PC1)	passé	aoristique	*Samedi, Luc s'est promené*
(PC2)	présent	accompli	*Luc a terminé depuis longtemps*
Plus-que-parfait (PQP1)	passé	aoristique	*La veille, Luc s'était promené*
(PQP2)	passé	accompli	*Luc avait terminé depuis longtemps*
Futur antérieur (FA1)	futur	accompli	*A huit heures, Luc aura terminé depuis longtemps*
(FA2)	futur	aoristique	*Luc sera fatigué, il va (être en train de) manger*
aller (prst) Vinf (FP1)	futur	aoristique	*Demain, Luc va terminer son travail*
(FP2)	futur	inaccompli	*A huit heures, il va (être en train de) manger*
(FP3)	présent	prospectif	*Luc va se mettre en colère. Je le vois bien.*

Tableau 4.8 : Les temps verbaux chez Gosselin [1996] (p. 29).

Temps \ Aspect	Aoristique [I,II] CO [B1,B2]	Inaccompli [I,II] RE [B1,B2]	Accompli [I,II] POST [B1,B2]	Prospectif [I,II]ANT [B1,B2]
Passé [I,II] ANT [01,02]	PS, IMP2, P3, PC1, PQP1	IMP1	PQP2	-
Présent [I,II] SIMUL [01,02]	-	P1, PC2	-	FP3
Futur [I,II] POST [01,02]	P2, F1, FA2, FP1	F2, FP2	FA1	-

Tableau 4.9 : Les temps verbaux chez Gosselin, présentation alternative.

4.7.7 Un regard final sur les modèles de temps verbaux

Après avoir examiné ces différents modèles de temps verbaux, il apparaît clairement que cette partie de l'analyse temporelle n'est pas une tâche aisée, surtout si on désire l'envisager dans une approche automatique. Les théories exposées ne s'accordent pas toujours entre elles, et présentent certaines divergences sur plusieurs points. Les modèles les plus anciens, basés sur des repères sous la forme de points, ont ensuite vu l'arrivée d'autres modélisations, celles-ci faisant appel à la fois à des points et des intervalles, voire uniquement à cette dernière forme de représentation. Malgré ces différentes propositions de modélisation, certains temps ne sont parfois pas représentés, ou ne peuvent être distingués complètement des autres. Parmi ces modèles, un seul intègre la prise en compte des temps

verbaux dans un cadre plus large. Il s'agit du modèle de Gosselin, qui fait appel aux notions d'aspect et inclut les circonstants temporels.

La mise au point de ces modèles reste cependant encore un défi en linguistique. Par conséquent, leur implémentation dans un processus d'analyse automatique du temps représente une tâche particulièrement ardue. Il existe d'ailleurs peu de systèmes qui intègrent un environnement complet de traitement des temps verbaux. Un de ceux-ci est proposé par Person [2004] et est inspiré du modèle des intervalles de Gosselin. Ce système présente cependant certaines limites, entre autres dues à son objectif, l'analyse de constats d'accidents de la route. Outre l'orientation du système vers l'analyse de ce type particulier de textes, la reconnaissance des circonstants temporels ne constitue pas la priorité du système et n'est donc pas très développée. De plus, l'analyse nécessite parfois, pour certaines phases, l'avis d'un utilisateur (système semi-automatique).

4.8 La structure du discours

L'organisation générale du texte est un élément qui véhicule lui aussi une dimension temporelle. En effet, si l'ordre d'apparition des différentes propositions qui composent ce texte peut correspondre à leur déroulement chronologique, il existe également plusieurs phénomènes qui peuvent venir perturber cette linéarité : flashbacks, projections en avant, recouvrements entre procès, etc. L'ordre d'apparition des propositions conserve cependant une influence importante quant à leur interprétation temporelle. Par exemple :

(1) Jean sortit et prit sa voiture. Ça bouchonnait sur l'autoroute. Il alluma la radio et tomba sur une émission qu'il n'aimait pas.

(2) Jean alluma la radio et tomba sur une émission qu'il n'aimait pas. Il sortit et prit sa voiture. Malheureusement, ça bouchonnait sur l'autoroute.

(3) Jean alluma la radio et tomba sur une émission qu'il n'aimait pas. Il était sorti et avait pris sa voiture. Malheureusement, ça bouchonnait sur l'autoroute.

Les exemples 1, 2 et 3 montrent bien que si l'on modifie l'ordre d'apparition des propositions, l'interprétation est complètement différente. Dans l'exemple 1, on comprend que Jean est sorti de chez lui et qu'il écoute la radio dans sa voiture alors qu'il est pris dans les bouchons. Au contraire, dans l'exemple 2, le texte indique que Jean a écouté la radio chez lui et que, tombant sur une émission qu'il n'apprécie pas, il décide de sortir, de prendre sa voiture et se retrouve dans des embouteillages. Par contre, moyennant quelques modifications des temps des verbes dans l'exemple 3, le sens de l'exemple 1 peut être à nouveau obtenu en conservant l'ordre des propositions de l'exemple 2.

Dans Borillo *et al.* [2004], les principales relations de discours[34] sont identifiées. La première est la *narration*. Il s'agit d'une relation qui, lorsqu'elle est établie entre deux événements, permet de les décrire comme faisant partie *de la même histoire*. Son implication temporelle est que les deux événements sont successifs. Il s'agit généralement de la relation *par défaut*.

[34] En privilégiant celles qui présentent une dimension temporelle.

La relation de *résultat* implique un lien causal entre les deux événements qui, dès lors, entraîne la précédence du premier sur le second. Cette relation peut être provoquée par un marqueur tel que « donc » ou être inféré sur la base de connaissances sur les événements[35].

L'*explication* est l'inverse de la relation de *résultat*. Elle peut être mise en œuvre au moyen d'une séquence au passé composé (« Jean est tombé. Paul l'a poussé ») ou d'une suite composée d'un passé simple ou d'un passé composé suivi d'un imparfait (« Marie (arriva | est arrivée) en retard au cinéma. Elle attendait son mari à la maison avant de partir. »).

Une *élaboration* est le fait de préciser et détailler un premier événement au moyen d'un ou plusieurs autres événements. L'exemple de Kamp et Rohrer, cité par Borillo *et al.* [2004], illustre cette relation : « L'été de cette année-là vit plusieurs changements dans la vie de notre héros. François épousa Adèle, Jean-Louis partit pour le Brésil et Paul s'acheta une maison à la campagne. ». Cette relation peut être repérée à l'aide de connaissances sur les types d'événements et de sous-événements. Sa dimension temporelle implique que tout les événements qui détaillent l'événement principal sont inclus dans celui-ci.

La relation de *background* peut être repérée suite à des discontinuités aspectuelles[36]. Elle implique un chevauchement temporel entre les deux événements.

D'autres relations n'ayant pas d'implication temporelle sont encore identifiées par Borillo *et al.* [2004] : *contraste*, *continuation* ou encore *topic*.

Certains éléments temporels ont également une influence qui dépasse la proposition ou la phrase. C'est par exemple le cas des circonstanciels temporels détachés en début de phrase. L'exemple 4 illustre ce phénomène. On voit bien que le circonstanciel « lundi » ne porte pas uniquement sur la première proposition, mais sur l'ensemble de celles qui suivent, jusqu'à ce qu'un autre circonstanciel soit rencontré (« le lendemain »).

(4) **Lundi**, Jean sortit et prit sa voiture. Malheureusement, ça bouchonnait sur l'autoroute. Il alluma la radio et tomba sur une émission qu'il n'aimait pas. **Le lendemain**, il put par contre écouter son programme préféré.

Ce type de relation est comparable à celle observée dans le cas d'un encadrement du discours (Section 4.9).

4.9 Les cadres de discours

Pour Charolles [1997], le texte ne doit pas être analysé au niveau de la phrase, mais bien au niveau du discours. Certains marqueurs linguistiques, tels que des circonstanciels détachés en début de phrase, jouent un rôle particulier à cet égard. Il s'agit de la notion de *cadres de discours*, définie comme suit :

[35] « tomber » peut être le résultat de « pousser ».
[36] Par exemple, le passage d'un événement à un état.

> « Les cadres de discours intègrent une ou plusieurs propositions en fonction de critères qui sont spécifiés par les expressions les introduisant. Ils contribuent à subdiviser et à répartir les informations apportées par le discours au fur et à mesure de son développement. Les critères servant à la répartition des informations en blocs homogènes peuvent être très divers. » (Charolles [1997], p. 33).

Les cadres sont de longueurs variables, allant de la proposition à un ensemble de paragraphes. Ils correspondent en fait à des univers qui peuvent être de plusieurs types : spatiaux, de connaissances, de représentation, d'énonciation, et bien entendu temporels. C'est évidemment ce dernier type qui nous intéresse plus particulièrement car il souligne le caractère structurant ou organisateur des expressions temporelles au niveau du discours. Un exemple, tiré de Charolles *et al.* [2005], permet d'illustrer le découpage d'un texte en cadres temporels :

> « **En juin 1992**, 747 500 candidats se sont présentés a l'examen, [...] ; près des trois quarts ont été reçus ; mais pour les candidats individuels le taux de réussite a été à peine de 50%. Pour la série collège [...], 76% des candidats des établissements scolaires ont obtenu le brevet [...]. **En 1989**, tant les collégiens du privé que ceux du public ont de meilleurs résultats dans les départements des académies de l'Ouest où les éleves du privé sont nombreux, [...]. Dans le Nord-Ouest, en Ile-de-France et dans l'Est, les taux de réussite des éleves des collèges publics sont généralement inférieurs à la moyenne nationale, [...]. **À la session 1991**, ce sont les académies de Rennes, Grenoble, Dijon, Nantes, Clermont-Ferrand qui obtiennent les meilleurs résultats [...], alors que celles du Midi méditerranéen n'atteignent pas 70%. » (p. 122)

Trois cadres de discours sont présents dans cet exemple. Le premier commence par « En juin 1992 » et se termine lorsque le deuxième commence (« En 1989 »). Enfin le troisième cadre suit, à partir de l'expression « À la session 1991 ». On peut le constater, ces expressions temporelles caractérisent et portent sur tout le cadre qu'elles définissent.

De nombreux travaux ont par ailleurs montré le rôle d'organisation du discours qu'ont les expressions temporelles. Par exemple, Costermans et Bestgen [1991] montrent que leur introduction en début de phrase permettent de marquer des ruptures importantes et d'ainsi signaler la structure hiérarchique du texte. Cette fonction de marqueur de segmentation a en partie été validée de manière expérimentale par Piérard *et al.* [2004].

La faisabilité et l'utilité de l'exploitation des cadres de discours ont déjà été démontrées. Par exemple, Battistelli *et al.* [2006] proposent une application de la gestion de la temporalité au travers des cadres de discours. Le système construit est une aide à la lecture de textes, biographiques en l'occurrence, qui permet de passer d'une lecture linéaire du texte à une lecture chronologique en naviguant d'un cadre de discours à un autre par le biais de leur caractérisation temporelle.

4.10 Conclusion

Dans le langage naturel, et dans le cas présent, en français, de nombreux éléments interviennent pour l'expression de notions temporelles. Ce chapitre a permis d'aborder les principaux : les adverbes et connecteurs temporels, la notion de procès, le(s) temps et l'aspect, y compris les différents modèles linguistiques des temps verbaux, et finalement la structure de discours, et en particulier le mécanisme d'encadrement du discours.

Ces différents vecteurs de l'information temporelle sont tous importants, et devraient idéalement tous pouvoir être intégrés dans un processus de traitement automatique, permettant ainsi leur prise en compte conjointe. Malheureusement, certains sont moins enclins que d'autres à se soumettre à ce type d'analyse. Il est cependant possible de mener des tâches d'extraction d'informations temporelles en s'appuyant sur les éléments exploitables par une procédure automatique.

Les adverbes temporels sont à analyser en priorité, de par leur potentiel informationnel, ainsi qu'en raison des bonnes perspectives qu'offre leur traitement automatique. L'apport de l'analyse des temps des verbes est également incontestable, mais l'implémentation d'un modèle complet est une tâche complexe, surtout au vu de certains points théoriques qui restent ouverts. Le traitement du temps ne peut cependant se passer de l'information temporelle portée par le verbe. L'analyse à ce niveau doit donc tenter d'apporter autant d'informations que possible. Enfin, l'organisation du discours peut aussi apporter des informations intéressantes, mais nécessite des processus d'analyse plus larges, qui agissent au niveau du texte. Certains phénomènes, tel que l'encadrement du discours, n'impliquent pas un traitement trop complexe et sont susceptibles d'améliorer l'analyse temporelle. Une présentation plus détaillée des caractéristiques et des choix réalisés pour l'extraction d'informations temporelles réalisée dans cette thèse, est proposée à la section 7.3.

CHAPITRE 5

MODÉLISATION DU TEMPS

5.1 Introduction

Dans cette thèse, l'accent n'est pas mis de manière aussi importante sur les aspects de modélisation du temps que sur les aspects d'extraction. Par conséquent, les questions relatives à la manière de représenter le temps sont surtout abordées de façon à pouvoir définir clairement les concepts et idées qui sont exploités par la suite. Dans ce chapitre, l'ensemble des modélisations possibles pour un phénomène tel que le temps ne sont donc pas abordées. Plutôt que de fournir une étude qui se voudrait exhaustive, mais qui serait inévitablement incomplète, la présentation s'axe plutôt sur les points importants pour la poursuite de l'exposé. Le problème est abordé en deux temps : quelle modélisation peut on donner à l'espace du temps (Section 5.2) et comment est-il possible de faire référence à cet espace (Section 5.3) ?

5.2 Modélisation de l'espace du temps

La section 3.2, qui a introduit le concept de *temps*, a montré que cette notion fondamentale se manifeste indirectement au travers de la répétition régulière de certains phénomènes astronomiques. Ceux-ci ont été caractérisés à l'aide d'outils de mesure, ce qui a permis de matérialiser le temps. Par conséquent, ces phénomènes constituent généralement la base des représentations et modélisations de l'espace temporel.

5.2.1 Calendriers et autres modélisations

La représentation la plus évidente est matérialisée par les *calendriers*. Ceux-ci offrent la possibilité de représenter le temps sous une forme *structurée* en rapport avec les différents phénomènes naturels et physiques, principalement le déroulement des cycles solaires et lunaires. Leur utilisation et acceptation très large est évidemment un gros avantage. Il faut cependant tempérer ce constat. Tout d'abord, il n'existe pas un seul calendrier[1], qui serait universel, mais bien plusieurs. La définition

[1] Pour simplifier l'exposé, lorsque le terme *calendrier* est employé sans autre précision, il fait référence au calendrier grégorien.

de ceux-ci varie, selon les différentes communautés, en fonction de questions liées à la culture, la religion, etc. Ensuite, la modélisation du temps sous la forme d'un calendrier est assez complexe. Elle intègre différents niveaux, dont les unités correspondent à des périodes de temps plus ou moins longues (heure, jour, semaine, mois, année, etc.). Cette organisation hiérarchique, si elle répond à un besoin pratique (Section 5.2.2) ne constitue pas un avantage pour une manipulation simple du temps et des objets temporels. De plus, l'aspect irrégulier de cette organisation ne fait qu'augmenter la complexité du modèle calendaire. En effet, le nombre de jours dans un mois varie de 28 à 31, selon les mois et les années, ces dernières pouvant être bissextiles. Par conséquent, un jour désigné par une référence *jour/mois* (le « 15 janvier ») ne se voit pas toujours attribuer le même nom de jour (« vendredi » par exemple). De même, si une année débute bien par le 1^{er} janvier, elle ne commence pas toujours par un jour ayant le même nom. Il y a également le changement d'horaire entre l'été et l'hiver, la gestion de fuseaux horaires, etc. En fait, le système de calendrier tente de se synchroniser avec les phénomènes à la base de la notion commune de temps, mais n'y arrive qu'imparfaitement et se voit donc contraint d'introduire des mécanismes de réajustement.

Partant de ce constat de complexité, Ohlbach [2000] mentionne une alternative à ce système :

> « Weeks are not in phase with months and years, but they are in phase with hours and minutes and seconds, which themselves are in phase with months and years. Quite confusing, isn't it ? Things would be much easier if calendar systems would be decimal systems, with one fixed unit of time as basis, and all other units as fractions or multiples of this basis. » (p. 320)

L'idée citée par Ohlbach consiste à représenter le temps comme une suite de périodes d'une certaine taille, par exemple des jours. Ces unités sont alors étiquetées par une suite d'identifiants, généralement monotone et croissante, par exemple des entiers ou des réels. Ce principe est exploité par diverses représentations, dont le système de *jours Julien* (*Julian Day Number*, Meyer [2009]), le système de *Temps Atomique International* (BIPM [2010]), etc. Cette solution est évidemment beaucoup plus facile à manipuler car il n'existe qu'une seule unité temporelle, et qu'un déplacement dans le temps peut être effectué par simple addition ou soustraction. Par contre cette représentation peut devenir peu pratique si les quantités de temps, ou les durées des zones temporelles manipulées, sont très faibles ou, au contraire, très grandes[2]. Ce système de modélisation du temps, s'il reste basé sur une réalité physique via son unité de base, s'éloigne de la représentation habituelle du temps et est par conséquent difficilement directement interprétable par un humain. Plus simple et régulier, il est par contre particulièrement adapté pour un traitement automatique et une manipulation informatique.

Les différences présentées par ces deux représentations en font deux systèmes complémentaires, destinés à des utilisations différentes. Le calendrier correspond aux concepts intelligibles et manipulés habituellement par l'homme, et qui se retrouvent donc dans le langage. Le système de jours Julien est moins naturellement appréhendé par l'homme et est plutôt destiné à une exploitation informatique.

[2] Par rapport à l'unité de base.

5.2.2 La granularité

Le fait que l'espace du temps puisse être envisagé au moyen d'unités de tailles plus ou moins grandes – le jour, le mois, l'année, etc. – est une manifestation de l'intervention des divers phénomènes naturels dans la définition du concept de temps (voir section 3.2). Cette caractéristique est à l'origine du système de *granularités*. Celui-ci a été progressivement élaboré et enrichi sous l'influence des religions[3] ou de l'activité humaine[4] (Bettini *et al.* [2000]).

Fondamentalement, ces diverses granularités dénotent toutes un même objet, le temps. Dès lors, pourquoi est-il nécessaire d'en manipuler plusieurs ? Toujours selon Bettini *et al.* [2000], l'explication la plus directe est qu'il est peu commode de manipuler des grands nombres. D'autre part, ce n'est pas la mesure du temps en elle-même qui constitue l'objectif poursuivi, mais bien l'association d'une certaine période à un événement. Or, chaque type d'événement possède une durée intrinsèque propre (voir la notion de *procès*, à la section 4.4). Par exemple, une explosion ou une traversée de l'Atlantique ne se déroulent intuitivement pas sur la même période de temps. Évidemment, entre différentes réalisations d'un même type d'événement, les durées observées peuvent varier. Mais ce qui nous intéresse ici, c'est l'*ordre de grandeur* de la durée de l'action. Si cette notion peut parfois sembler mal définie, dans la pratique, il est généralement assez intuitif de déterminer cette granularité. Il est par exemple évident que l'ordre de grandeur temporel d'un marathon n'est certainement pas le mois, mais plutôt l'heure.

La notion que nous désignons sous le nom de *granularité*, correspond donc à un ordre de grandeur à partir duquel la dimension temporelle des événements peut être exprimée. Les nombreuses granularités existantes sont organisées en systèmes hiérarchiques. L'utilité de ce type de système est d'ordonner les éléments et de permettre certaines conversions. Un exemple concret est constitué par la norme ISO 8601 (ISO [2004]) qui traite de la représentation des dates, des horaires et des périodes de temps.

Si, comme le font Battistelli *et al.* [2006] (et bien d'autres encore : Alonso *et al.* [2007], Strötgen *et al.* [2010]), on se réfère à cette norme, deux unités de base émergent : le jour et l'an. L'utilisation de ces deux dimensions temporelles permet de désigner un élément quelconque du calendrier dont la granularité est le jour sous la forme d'un couple (année,jour de l'année). D'autres granularités, plus grandes (les siècles) ou plus petites (la seconde), sont évidemment proposées par la norme. Il existe un lien de multiplication ou de division entre ces granularités et le jour. Ce dernier constitue donc la granularité de base et fait office de pivot entre d'une part les granularités plus élevée et d'autre part celles qui sont plus fines.

Jusqu'ici, les termes *unité* et *granularité* ont été employés sans être réellement distingués. Il est cependant important de marquer la différence qui sépare ces deux concepts. Une unité temporelle désigne une certaine quantité de temps, sans référence particulière à un calendrier ou à tout autre modélisation du temps. Une zone temporelle désignée au moyen d'une unité de temps peut donc

[3] Une semaine compte sept jours en référence aux sept dieux des Babyloniens. Ce peuple est également à l'origine de la division d'un jour en 24 heures.

[4] Les notions de week-end ou de jour ouvrable sont à mettre en relation avec l'organisation moderne du travail.

commencer à n'importe quel moment (dans « sa maladie a duré deux ans », la zone temporelle qui dure deux ans ne possède pas d'ancrage temporel particulier, elle peut débuter à n'importe quel moment). La granularité désigne elle un ordre de grandeur temporel d'un événement ou d'un fait. Une instance particulière d'une granularité correspond à un élément de calendrier[5]. Cet élément a pour caractéristique de ne pas débuter à n'importe quel moment. Par exemple, « l'année 2010 » représente une zone temporelle qui dure un an, mais dont le début se situe à un moment précis de l'espace du temps. Parfois, la localisation de la zone temporelle peut être moins précise (« mardi matin »), mais elle conserve la particularité de ne pas pouvoir se dérouler à tout moment. La distinction entre unité et granularité est également visible au travers du fait que les quantités de temps exprimées au moyen d'unités sont convertibles (un jour = 24 heures) alors que le passage d'une granularité à une autre est plus délicat et implique une certaine imprécision, surtout dans le cas d'une diminution d'échelle (« 14/01/2011 » à la granularité du mois donne « 01/2011 », « 14/01/2011 matin » à la granularité de l'heure peut, entre autres, être désigné par « 14/01/2011 à 10h00 »[6].

5.3 Modélisation des références à l'espace du temps

5.3.1 Référence à une zone temporelle

Une fois l'espace du temps défini par une ou plusieurs modélisations, il est nécessaire de disposer de moyens afin d'y accéder. Le choix d'une représentation pour les références à l'espace du temps est un sujet qui fait débat depuis de nombreuses années. Certains estiment que ces références peuvent être réalisées au moyen de points qui désignent une zone temporelle particulière (par exemple Mani *et al.* [2005] citent McCarthy et Hayes [1969] dans le cadre du *Situation Calculus*, Kowalski et Sergot [1986] pour l'*Event Calculus*, ou encore McDermott [1982] en planification). D'autres pensent que cette représentation n'est pas adaptée et préfèrent les représenter sous la forme d'intervalles (entre autres Allen [1984] ou Gosselin [1996]). Ces deux points de vue ne sont pas nécessairement inconciliables, car tout point – par exemple le *mois* – peut être représenté par un intervalle dont les bornes sont exprimées à une granularité inférieure – par exemple le *jour*. L'inverse n'est pas toujours vrai, car il n'existe pas nécessairement une granularité qui correspond exactement à chaque intervalle possible[7].

5.3.2 Manipulation des références temporelles

La formalisation des références à l'espace du temps revêt une importance particulière lorsque la manipulation de celles-ci est abordée. Les différentes opérations sur ces références, entre autres les déplacements et surtout les comparaisons, sont dépendantes de la représentation sous-jacente. Le

[5] Au sens large du terme, il n'est pas ici nécessairement question d'un calendrier journalier.

[6] D'autres valeurs sont cependant possibles. Dans le cadre de l'implémentation d'un système automatique, un choix arbitraire doit être posé pour gérer ce cas.

[7] Tant au niveau de la durée – comment exprimer sous la forme d'un point une période de deux heures et demi ? – qu'en ce qui concerne son emplacement dans l'espace du temps – une période d'un mois à cheval sur un mois particulier du calendrier grégorien sera difficilement désignée au moyen d'un simple point.

fonctionnement de ces opérations est décrit et défini dans un *système logique*. Sans rentrer dans les détails du fonctionnement de celui-ci, il est tout de même intéressant de s'attarder sur certaines opérations parmi les plus importantes, telles que le déplacement temporel ou la comparaison de références.

Dans le cadre d'une modélisation sous la forme de points, les opérations sont relativement simples. Le déplacement temporel consiste, à partir d'un point déterminé, à avancer ou reculer dans le temps d'un certain nombre d'unités. Dans le cas d'une modélisation de l'espace temporel sous la forme d'une suite monotone croissante, cette opération se résume à une addition ou à une soustraction. Si le modèle temporel est un calendrier, le déplacement temporel doit obéir aux règles particulières que celui-ci définit (nombre de jours dans un mois, etc.).

En ce qui concerne la comparaison, la situation est aussi assez claire. Soit deux points temporels t_1 et t_2 situés dans le même espace :

t_1 est antérieur à t_2 si $t_1 < t_2$,
t_1 est postérieur à t_2 si $t_1 > t_2$,
t_1 est simultané à t_2 si $t_1 = t_2$.

Les opérateurs « $<$ », « $>$ » et « $=$ » doivent évidemment être définis en fonction de la modélisation de l'espace temporel. Cette définition correspond par exemple à la relation d'ordre sur les entiers, dans le cas d'une modélisation du temps dans cet espace. Dans le cas d'un calendrier, si la définition formelle est plus complexe[8], elle correspond cependant à une intuition bien ancrée.

Les choses sont cependant un peu plus complexes que cela. En effet, jusqu'à présent il a été implicitement supposé que les opérations de déplacement et de comparaison s'effectuent à une seule et même granularité. Or, ce n'est évidemment pas toujours le cas. Pour un déplacement temporel avec un granularité du point de départ inférieure au *delta*, par exemple $samedi + 1\ semaine$, une opération de conversion de ce dernier peut être réalisée : $samedi + 7\ jours$. Dans certain cas, l'adaptation du *delta* n'est pas toujours possible sans perte de précision. Un mois peut en effet être composé d'un nombre variable de jours. Lorsque le déséquilibre de granularité est inversé, par exemple $novembre\ 2008 + 1\ jour$, le résultat sera nécessairement imprécis[9]. Cette situation est cependant moins naturelle et plus rare.

Le problème de granularité est aussi présent lors de la comparaison entre deux points tels que $novembre\ 2008$ et $samedi$. Une possibilité consiste à *généraliser* la référence qui possède la granularité la plus fine. Cette généralisation donne cependant, à ce nouveau point temporel, un caractère approximatif. Dans ce cas, la représentation sous la forme de points semble être une limite. Il apparaît nécessaire de pouvoir considérer une référence temporelle ponctuelle sous la forme d'un intervalle dont les bornes sont exprimées à un niveau de granularité compatible avec celui de la seconde référence.

[8] Celle-ci sort du cadre de cette thèse et n'est donc pas développée.

[9] La conversion du point de départ en un intervalle dont les bornes sont exprimées à la granularité du *delta*, pourrait éventuellement être envisagée dans cette situation. Le déplacement temporel suit alors les règles spécifiques à la manipulation des intervalles.

En ce qui concerne les intervalles, la situation est moins évidente. En cas de déplacement temporel, l'opération doit être appliquée à chacune des deux bornes. Pour ce qui est de la comparaison, le nombre de situations est beaucoup plus élevé qu'avec des points. Allen [1984] en a dénombré treize, en comptant les relations inverses (à l'exception de l'égalité) et les a définies :

- $During(t_1, t_2)$: l'intervalle t_1 est complètement contenu dans l'intervalle t_2 ;
- $Starts(t_1, t_2)$: l'intervalle t_1 partage le même début que l'intervalle t_2, mais finit avant la fin de t_2 ;
- $Finishes(t_1, t_2)$: l'intervalle t_1 partage la même fin que l'intervalle t_2, mais débute après le début de t_2 ;
- $Before(t_1, t_2)$: l'intervalle t_1 se situe avant l'intervalle t_2 et ceux-ci ne se chevauchent en aucune manière ;
- $Overlap(t_1, t_2)$: l'intervalle t_1 débute avant l'intervalle t_2 et ceux-ci se chevauchent ;
- $Meets(t_1, t_2)$: l'intervalle t_1 se situe avant l'intervalle t_2 et il n'existe pas d'autre intervalle entre ceux-ci ;
- $Equal(t_1, t_2)$: t_1 et t_2 représentent le même intervalle.

Les difficultés relatives aux différences de granularité sont également présentes et se reportent sur les bornes des intervalles.

5.3.3 Systèmes temporels et ontologies

La norme ISO 8601 (ISO [2004]) est un ensemble de définitions de concepts temporels qui permet de représenter les références à l'espace du temps. Celles-ci prennent la forme de points ou d'intervalles qui se placent dans un cadre calendaire. Les différentes granularités et autres phénomènes particuliers aux calendriers sont également abordés et définis.

Les différents concepts de référence à l'espace du temps – les objets temporels tels que des points ou des intervalles – ainsi que leur manipulation, peuvent aussi être rassemblés au sein d'une *ontologie temporelle*. Celle-ci, délimite un système cohérent, comprenant l'ensemble des notions et opérateurs nécessaires à l'expression de données temporelles, mais leur associe également une définition logique. Cette dernière, en plus d'apporter une définition formelle qui profite de la rigueur propre aux langages logiques, permet une opérationnalisation des concepts temporels. En effet, sur la base des définitions logiques, des logiciels souvent nommés *raisonneurs* sont capables d'effectuer des calculs, des vérifications, des *raisonnements*.

De tels systèmes ontologiques sont par exemple définis dans Hobbs *et al.* [2002] et Hobbs et Pan [2004]. Ces deux ontologies expriment les concepts temporels – les instants et les intervalles, les relations temporelles, les liens entre temps et événements, la durée ainsi que le concept de calendrier – au moyen de la logique de premier ordre[10]. Ces définitions sont destinées à être utilisées au sein

[10] Aussi nommée *calcul des prédicats du premier ordre*.

des langages ontologiques DAML[11] ou OWL[12]. D'autres définitions ontologiques ont bien entendu été proposées, par exemple Zhou et Fikes [2002] ou Fikes et Makarios [2004].

Le langage OWL ne propose pas de manière native de mécanisme de raisonnement. Pour effectuer ce type d'opérations sur la base d'une ontologie OWL, le langage SWRL[13] peut être employé. Par exemple, O'Connor *et al.* [2009] et O'Connor et Das [2010] utilisent ce langage pour manipuler des données temporelles dans des ontologies biomédicales.

Enfin, d'autres formalismes, tels que les graphes conceptuels, ont également été utilisés pour représenter et raisonner sur les données temporelles (Amghar *et al.* [2002]).

5.3.4 Imprécision des références temporelles

Dans les sections précédentes, il a, à plusieurs reprises, été question de la possibilité pour une référence à l'espace temporel d'être imprécise. Dans le même temps, il a été montré que ces références peuvent être réalisées à différents niveaux de granularités. Ces deux concepts sont en fait liés. En effet, une référence à l'espace du temps qui s'établit de manière approximative à une granularité très fine, ne sera pas nécessairement considérée comme imprécise si le niveau de granularité est augmenté. Par exemple, la zone temporelle située aux alentours de 15h30, lors d'une certaine journée, ne constitue pas une référence précise à l'espace du temps à l'échelle de la minute. Par contre, si les grains de base sont les journées ou les parties de journées (l'après-midi par exemple), la référence peut être considérée comme précise.

En pratique peu d'ontologies ou de systèmes temporels tiennent compte de cette caractéristique d'imprécision qui constitue cependant un phénomène relativement courant au niveau de l'expression du temps dans la langue.

[11] DARPA Agent Markup Language, `http://www.daml.org`.
[12] Web Ontology Language, `http://www.w3.org/TR/owl-ref/`
[13] Semantic Web Rule Language, `http://www.w3.org/Submission/SWRL/`

CHAPITRE 6

EXTRACTION D'INFORMATIONS TEMPORELLES

6.1 Introduction

Les approches exposées dans les sections précédentes ont montré différents points de vue sur l'information temporelle. Ces théories contribuent à définir et expliquer le phénomène de la temporalité en langage naturel, mais adoptent souvent une démarche plutôt théorique. Le but de ces travaux est d'analyser certains aspects du problème de manière à en expliquer le fonctionnement. La réalisation d'un système automatique d'analyse de la temporalité n'est pas leur objectif. Ils constituent cependant un terreau qui peut permettre à de tels systèmes de germer. C'est dans le domaine de l'*extraction d'informations* que la préoccupation d'opérationnaliser l'analyse (temporelle) sur la base de ces différentes théories est présente. C'est une approche plus pragmatique qui part d'éléments concrets pour atteindre des objectifs ciblés. Le développement d'un système d'extraction d'informations nécessite souvent de prendre en compte les caractéristiques particulières et les contraintes provenant :

- du type de texte à traiter,
- des buts à atteindre,
- des méthodes et outils disponibles.

La confrontation entre les acquis théoriques et ces contraintes permet d'identifier ce qu'un système automatique peut réellement traiter.

L'extraction d'informations temporelles comporte différentes facettes, dont trois principales. La première est la tâche qui consiste à repérer et délimiter à l'aide d'un marquage les expressions temporelles. La deuxième consiste à interpréter ces expressions dans le but d'y associer de manière explicite et univoque la valeur temporelle qu'elles véhiculent, et cela selon un format normalisé. Enfin, le troisième aspect concerne la détection d'*événements*, l'identification des liens qui relie ceux-ci aux expressions temporelles, et finalement leur organisation chronologique. Ces différentes tâches ont été progressivement définies au travers des conférences en extraction d'informations, telles que MUC et ACE. Les formats et directives d'annotation en sont par conséquent souvent issus. Parmi les contributions principales relatives à la définition de la tâche, citons en particulier Timex (Chinchor [1997]), Timex2 (Ferro *et al.* [2005]), la campagne d'évaluation de ACE (NIST [2006]) ou encore les activités autour du format TimeML (Pustejovsky *et al.* [2003a], Boguraev *et al.* [2005] et Saurí *et al.*

[2006]). Notons que de nombreuses autres initiatives sont également centrées sur le sujet, que ce soit dans le domaine de l'extraction d'informations, en linguistique ou sur les aspects de modélisation : les symposiums TIME[1] (International Symposium on Temporal Representation and Reasoning) et les conférences Chronos[2] (International Conference on Tense, Aspect, Mood, and Modality) ou encore la tâche TempEval[3] au sein du workshop SemEval.

Après une rapide définition de la notion d'expression temporelle (Section 6.2, les travaux existants seront examinés selon divers points de vue : leurs objectifs (Section 6.3), le type d'informations et d'expressions prises en compte (Section 6.4), les formats d'annotation (Section 6.5), les techniques de reconnaissance et d'interprétation (Section 6.6), et le support éventuel de différentes langues (Section 6.7). Après ce tour d'horizon, l'implémentation du système d'extraction d'informations temporelles développé dans le cadre de cette thèse, est présenté au chapitre 7.

6.2 Définition de l'information temporelle

Bien que souvent les expressions temporelles ne soient pas définies explicitement, mais plutôt par la tâche d'extraction en elle-même, plusieurs définitions existent. Nous reprenons ici celle que donne Ahn *et al.* [2005] qui définit les expressions temporelles comme étant des « natural language phrases that refer directly to time points or intervals. They not only convey temporal information on their own but also serve as anchors for locating events referred to in text ». Il ne s'agit pas vraiment d'une définition opérationnelle telle qu'on pourrait la concevoir en extraction d'informations, mais elle a le mérite de souligner de manière succincte trois aspects importants de l'information temporelle auxquels le domaine s'attaque :

- l'information temporelle est véhiculée par des groupes de mots, que l'on peut reconnaître et annoter, dans les textes en langage naturel ;
- ces expressions peuvent être interprétées afin de leur donner une valeur relative à un certain espace temporel ;
- les valeurs temporelles structurent le discours en étant reliées aux événements qui le composent, et permettent d'ordonner les événements entre eux.

6.3 Types d'extractions et objectifs poursuivis

Les objectifs principaux concernent le repérage et l'interprétation des expressions temporelles. Les extensions constituées par la reconnaissance d'événements et leur positionnement temporel sortent quelque peu du cadre strict de l'extraction temporelle telle que nous la concevons, mais sont généralement considérés comme une suite logique.

[1] http://time.dico.unimi.it/
[2] http://www.utexas.edu/cola/conferences/chronos-8/main/
[3] http://www.timeml.org/tempeval/

6.3.1 Reconnaissance et interprétation d'expressions temporelles

La reconnaissance et l'annotation des expressions temporelles constituent naturellement la première étape de tout traitement de la temporalité. L'objectif poursuivi est de baliser le plus complètement et précisément possible un ensemble d'expressions temporelles diverses et préalablement définies. Ces définitions peuvent par exemple prendre la forme de directives d'annotation, telles que celles présentées à la section 6.5.

L'attribution d'une valeur aux expressions temporelles est une seconde étape, qui est souvent traitée conjointement avec l'annotation. D'une manière générale, il s'agit de donner un *sens* à la référence temporelle, ce qui est la plupart du temps réalisé en lui attribuant une valeur explicite, relative à une ligne du temps (un système calendaire). Un certain nombre de travaux sont consacrés exclusivement à la question de la reconnaissance et de l'annotation des expressions temporelles (Maurel [1990], Fairon et Senellart [1999], Vazov [2001], Bittar [2008], Weiser [2010]) mais il est très fréquent que le problème connexe de l'attribution d'une valeur temporelle à l'expression soit également abordé dans le même temps (Maurel et Mohri [1994], Mani et Wilson [2000], Filatova et Hovy [2001], Wilson *et al.* [2001], Muller et Tannier [2004] Battistelli *et al.* [2006], Ahn *et al.* [2005, 2007], Vicente-Díez *et al.* [2008], Parent *et al.* [2008], Bittar [2009], Martineau *et al.* [2009]).

6.3.2 Reconnaissance et positionnement temporel d'événements

Un autre aspect du traitement de la temporalité est la reconnaissance des unités considérées comme des *événements*. Cette notion est cependant très large et peut correspondre à de nombreuses choses dans la pratique. Dans le cadre des systèmes d'extraction d'informations temporelles, les événements considérés sont la plupart du temps limités à certains types de syntagmes précis. Concrètement, Mani et Wilson [2000], Filatova et Hovy [2001] et Mani et Schiffman [2005] ne prennent en compte que les verbes. Muller et Tannier [2004] se concentrent également sur les événements introduits par des verbes finis[4], alors que Schilder et Habel [2001] exploitent à la fois les verbes (« increased ») et les groupes nominaux (« the election »). Bittar [2008, 2009] se conforme à la définition donnée par TimeML qui considère l'événement dans un sens large, c'est-à-dire la plupart des verbes, des noms événementiels (« destruction », « guerre »), des adjectifs (« malade ») et des groupes prépositionnels (« à bord ») qui désignent des états. Parent *et al.* [2008] adoptent le même type de définition, mais sous le nom d'*éventualité*, celle-ci dénotant un événement ou un état. Dans Hagège et Tannier [2008], les événements pris en compte sont exprimés par :

- les verbes, qu'ils expriment une action ou un état ;
- les noms déverbaux pour lesquels il existe un lien morphologique clair entre le nom et le verbe (par exemple, en anglais, « interaction » et « interact ») ;
- les autres noms s'ils sont argument de la préposition « during » (« during the war ») ou s'ils sont le sujet des verbes « to last », « to happen », « to occur » lorsque ces

[4] Un verbe ou un auxiliaire qui porte une indication du temps et qui porte des traits de personne et de nombre provenant de l'accord sujet-verbe.

verbes sont modifiés par une expression temporelle explicite (« the siege lasted three days »).

Dans un certain nombre de travaux, la notion d'événement est accompagnée et raccrochée à celle de proposition, qui désigne une portion de phrase qui ne contient qu'un seul événement : Filatova et Hovy [2001], Muller et Tannier [2004], Mani et Schiffman [2005].

Le positionnement temporel d'événements constitue une suite logique à leur reconnaissance. Il existe quelques variations entre les différentes approches, entre autre en ce qui concerne le type de relations à établir – relations d'événement à expression temporelle et/ou d'événement à événement – et la taille de l'ensemble de relations utilisé.

La première approche consiste donc à établir des relations entre les événements et la ligne du temps (Mani et Wilson [2000], Filatova et Hovy [2001], Schilder et Habel [2001]). La seconde s'intéresse plutôt aux relations d'événement à événement. Pour ce type de relation, les divers travaux s'inspirent tous des relations définies par Allen [1984][5]. Cependant, il est apparu que les relations d'Allen constituent un ensemble trop grand et trop précis pour être utilisées lors d'une annotation manuelle (Setzer [2001], Muller et Tannier [2004], Mani et Schiffman [2005]). En effet, un annotateur humain se contentera souvent d'une ou deux relations entre deux événements alors qu'il peut souvent y en avoir plus. Cela pose évidemment un problème dans le cadre d'une évaluation automatique des résultats, lorsque celle-ci est effectuée par rapport à une annotation manuelle. Un ensemble alternatif et plus simple de relations est donc souvent choisi. Schilder et Habel [2001] utilisent sept relations[6], tout comme Hagège et Tannier [2008][7]. Muller et Tannier [2004] n'utilisent eux que six relations[8], alors que Mani et Schiffman [2005] n'en considèrent que trois[9].

6.4 Types d'expressions et d'informations prises en compte

L'extraction d'informations temporelles s'articule généralement en deux phases : le repérage ou l'annotation des expressions et leur interprétation et reformulation en une valeur normalisée. Selon que l'on se situe dans une étape ou dans l'autre, la vision que l'on peut avoir des expressions temporelles n'est pas nécessairement la même. Lors de la reconnaissance, il est intéressant d'analyser et de classer les expressions en fonction des types de constituants qui les composent (section 6.4.1). L'accent est donc mis sur leur morphologie. Par contre, lors de l'interprétation, c'est plutôt le sens, et donc le type de valeur temporelle véhiculée, qui dirige leur catégorisation (section 6.4.2).

Les événements, qui font souvent partie des travaux en extraction d'informations temporelles, ne sont pas à proprement parler des expressions temporelles, même si on peut leur reconnaître une dimension temporelle implicite (voir section 4.4). Ils constituent une catégorie d'informations particulière. En

[5] Allen a défini treize relations différentes qui peuvent s'établir entre deux intervalles : *before*, *meets*, *overlaps*, *starts*, *during*, *finishes*, les relations inverses de ces six premières relations et enfin *equal*.
[6] BEFORE, AFTER, INCL, AT, STARTS, FINISHES, EXCLUSION.
[7] BEFORE, AFTER, DURING, INCLUDES, OVERLAPS, IS_OVERLAPPED et EQUALS.
[8] BEFORE, AFTER, OVERLAPS, IS_OVERLAPED, INCLUDES et IS_INCLUDED.
[9] BEFORE, AFTER, EQUAL.

effet, leur reconnaissance a plutôt pour objectif de les ancrer dans l'espace du temps ou de les ordonner les uns par rapport aux autres, et non pas de les interpréter afin d'en tirer une valeur temporelle[10].

6.4.1 Caractérisation des expressions temporelles

Les travaux en extraction d'informations ne s'attardent pas toujours à décrire en détail la nature des expressions à extraire. D'une part, il s'agit probablement d'une tâche plus traditionnellement réservée aux travaux en linguistique. D'autre part, les systèmes d'extraction d'informations ont une vision très pragmatique et ne s'attachent à décrire que les éléments intéressants pour l'application visée ou ceux pour lesquels un traitement automatique est envisageable.

Certaines descriptions détaillées, mais limitées en étendue, existent cependant. Dans son article sur les déterminants numéraux, Gross [2002] dresse une description des *dates horaires*. Celles-ci peuvent être des heures informelles, constituées d'adverbes construits sur la base de noms tels que :

« matin »,
« après-midi »,
« soir »,
« nuit ».

Il peut également s'agir d'heures numériques, basées sur un cycle de 12 ou 24 heures :

« à onze heures du soir »,
« à 16 heures 24 minutes et 47 secondes ».

Quelques expressions qui ne rentrent pas dans ces deux catégories, sont également considérées comme des dates horaires, par exemple « à l'heure de l'apéritif ».

Vazov [2001] souligne que l'information temporelle est portée par un ensemble de marques linguistiques diverses, principalement grammaticales (morpho-syntaxiques), lexicales (les verbes et adverbes) ou strictement syntaxiques (les anaphores temporelles, la structure temporelle du verbe). Son travail d'extraction porte sur les marques temporelles lexicales non-verbales en français. Celles-ci sont définies comme étant :

- des adverbes (« hier »),
- des groupes nominaux adverbiaux (« trois jours avant le mariage de son plus jeune fils »),
- des expressions adverbiales (« à 10 heures »).

Vicente-Díez *et al.* [2008] établissent une typologie structurelle des expressions temporelles en anglais. Leur approche est un peu différente car elle ne se base pas sur les catégories grammaticales. Deux grands types de constituants sont distingués :

[10] Une exception à cette affirmation est constituée par les héméronymes (Calabrese Steimberg [2008]), c'est-à-dire les événements que l'on peut désigner par l'expression temporelle relative au moment auquel ils ont lieux (« le 21 juillet », « le 11 septembre », « le 20 heure »).

- les unités temporelles,
- les modificateurs.

Les *unités temporelles* sont composées de différents éléments :

- des unités de mesure temporelle (« hour », « minute », « week »),
- des unités déictiques (« today », « yesterday »),
- des unités nommées (« Monday », « 1998 », « 12/10/2007 »).

Les *modificateurs* sont caractérisés par leur position dans l'expression (modificateur PRE pour « last » et modificateur POST pour « after ») et par leur sémantique (modificateur ordinal pour « first » et modificateur de fréquence pour « each »).

D'une manière générale, les divers éléments qui sont exploités dans les systèmes d'extraction d'informations temporelles sont :

- des noms (« lundi », « décembre », « printemps », « jour », « heure », etc.),
- des expressions numériques (« 1984 », « 15 », etc.),
- des adverbes (« tôt », « tard », « actuellement », « désormais », « hier », « demain », « aujourd'hui », « environ », etc.).

Les expressions temporelles peuvent également être introduites par certaines prépositions adverbiales (« vers », « pendant », « depuis », etc.) ou encore suivies par des adjectifs (« dernier », « prochain », « suivant », etc.).

6.4.2 Types de valeurs temporelles

De nombreux travaux abordent l'interprétation des expressions temporelles (Maurel et Mohri [1994], Mani et Wilson [2000], Wilson *et al.* [2001], Filatova et Hovy [2001], Muller et Tannier [2004], Battistelli *et al.* [2006], Ahn *et al.* [2005, 2007], Vicente-Díez *et al.* [2008], Parent *et al.* [2008], Martineau *et al.* [2009]). La définition du type de valeur véhiculée par celles-ci est cependant parfois relativement sommaire. Les nomenclatures les plus abouties sont abordées ci-dessous, alors que les formats d'annotation Timex2 et Timex3 qui proposent de manière indirecte ce type de classification seront abordés à la section 6.5.

Une des classifications parmi les plus complètes et les mieux structurées est donnée par Muller et Tannier [2004] qui définissent 11 catégories :

- les dates non absolues (« le 25 mars », « en juin », etc.),
- les dates absolues (« le 14 juillet 1789 »),
- les dates relatives au moment d'élocution (« il y a 2 ans », « l'année dernière »),
- les dates relatives au focus temporel (« trois jours plus tard »),
- les dates absolues de forme particulière (« au début des années 1980 »),
- les dates relatives de forme particulière (mois, saisons),
- les durées quelconques (« pendant 3 ans »),

- les durées contenant deux dates (« du 11 février au 27 octobre »),
- les durées absolues (« à partir du 14 juillet »),
- les durées relatives au moment de l'élocution (« depuis un an »),
- les durées relatives au focus temporel (« depuis »),
- les atomes de temps (« trois jours », « 4 ans », etc.).

Diverses autres catégorisations ont été proposées par ailleurs, et sont plus ou moins développées. Elles réalisent généralement la distinction entre dates (ou points), durées (ou intervalles) et fréquences (ou ensembles) (Wilson *et al.* [2001], Bittar [2008, 2009] avec la distinction supplémentaire entre une date (DATE) et une heure (TIME), tout comme Parent *et al.* [2008] mais cette fois sans les ensembles).

D'autres travaillent avec deux grandes catégories qui correspondent aux références temporelles absolues (ou explicites), qui peuvent être directement replacées dans le contexte d'un calendrier, et aux références temporelles relatives, qui nécessitent une deuxième référence pour être interprétées (Schilder et Habel [2001], Battistelli *et al.* [2006]).

Vicente-Díez *et al.* [2008] proposent une classification en six catégories :

- les expressions temporelles absolues (« 25/10/2007 »),
- les expressions temporelles relatives (« yesterday »),
- les intervalles (« from May to June »),
- les ensembles temporels (« every day » ou « Mondays »),
- les durées (« during two months ») ,
- les dates nommées (« Christmas Day »).

Finalement, certaines approches sont parfois plus ciblées, comme celle de Weiser [2010] qui se concentre sur les dates et les horaires. Cela l'amène parfois à devoir traiter des expressions assez particulières et souvent complexes, par exemple « Ouvert du lundi au vendredi de 9h à 11h et le dimanche en haute saison ».

6.5 Formats d'annotation

Les formats d'annotation constituent également un moyen de définir l'information temporelle. En effet, en plus de déterminer un ensemble d'éléments uniformisés pour annoter les expressions temporelles, entre autres des balises et des attributs spécifiques, les directives d'utilisation de ces éléments sont exposées. Ce sont ces dernières qui contribuent à définir ce que sont les expressions temporelles. Les différents cas exposés sont accompagnés d'exemples qui illustrent la manière dont ils doivent être annotés. Les trois principaux formats, déjà évoqués précédemment, sont Timex, Timex2 et Timex3 (TimeML).

Le choix d'un langage et d'un formalisme pour annoter les expressions temporelles est important, car il influence à peu près toutes les étapes de l'extraction. C'est bien entendu le cas de la reconnaissance, entre autres en ce qui concerne la création des ressources nécessaires pour cette tâche. Mais

l'importance de ce format, et de son expressivité , est également valable pour l'étape d'interprétation, qui est menée sur la base des informations fournies par l'annotation.

6.5.1 Timex et MUC

L'annotation et l'extraction des expressions temporelles faisaient déjà partie des tâches relatives aux entités nommées définies lors des conférences MUC (Message Understanding Conference), en particulier dans MUC-7 (Chinchor [1997]).

Le format d'annotation prévu pour les entités nommées de type temporel est Timex. Il est conçu pour annoter des expressions absolues ou relatives qui peuvent être de type :

- **DATE**, une période de temps qui correspond à au moins un jour complet,
- **TIME**, une période plus courte qu'un jour complet.

Les expressions temporelles peuvent être ancrées par rapport à la ligne du temps, d'une manière absolue ou relative. Les durées qui ne sont pas explicitement ancrées ne doivent pas être annotées. En plus des références temporelles isolées que constituent les expressions de type DATE et TIME, les expressions d'intervalles sont également prises en compte.

Les **expressions temporelles absolues** sont constituées d'expressions qui indiquent un segment temporel spécifique. Pour le type **TIME**, les expressions de minutes doivent désigner une minute et une heure spécifique alors que les expressions d'heures doivent seulement désigner une heure particulière. Le type **DATE** recouvre les expressions de jours, de saisons, de trimestres ou semestres financiers (ou comptables), d'années, de décennies, de siècles. Les déterminants qui introduisent les expressions ne doivent pas être inclus de même que les mots ou groupes qui les modifient (« around », « about », etc.).

```
<TIMEX TYPE="TIME">twelve o'clock noon</TIMEX>
<TIMEX TYPE="DATE">January 1990</TIMEX>
```

Les **expressions temporelles relatives** sont des expressions qui indiquent une date relative à la date du document (« yesterday », « today », etc.) ou une fraction d'une unité temporelle donnée (« Monday morning »). Les expressions composées qui contiennent un marqueur déictique suivi d'une unité temporelle (« last month », « next year », etc.) sont également concernées.

```
<TIMEX TYPE="TIME">last night</TIMEX>
<TIMEX TYPE="DATE">yesterday</TIMEX> <TIMEX TYPE="TIME">evening</TIMEX>
```

Les **expressions indéfinies ou vagues** telles que les adverbes temporels vagues (« now », « recently », etc.), les expressions de durée indéfinies (« for the past few years »), les expressions relatives à un événement (« the morning after the July 17 disaster ») ne doivent pas être annotées.

Enfin, les **jours spéciaux** tels que les jours de vacances référencés par un nom particulier doivent être annotés.

```
<TIMEX TYPE="DATE"> All Saints' Day </TIMEX>
```

6.5.2 Timex2 et ACE

Le format d'annotation des expressions temporelles dans le cadre de ACE est défini par Timex2 (Ferro *et al.* [2005]). Ce format est plus étendu que Timex. Il prévoit un attribut particulier qui permet d'indiquer une valeur normalisée pour l'expression annotée. Le format de ce champ valeur (VAL) se base sur la norme IS0 8601 (ISO [2004]), en l'étendant pour certains cas.

Pour être annotées, les expressions doivent être composées d'un des déclencheurs lexicaux spécifiques à la temporalité. Il peut s'agir :

- de noms (« minute », « afternoon », « month »),
- d'unités nommées (« Monday », « January »),
- de patrons temporels spécialisés (« 8 :00 », « 1994 », « 1960s »),
- d'adjectifs (« recent », « former », « ago »),
- d'adverbes (« currently », « lately », « hourly »),
- de noms ou adverbes temporels (« now », « today », « yesterday »),
- de nombres (« 3 », « three », « fifth », « Sixties »).

Comme pour Timex, l'annotation des expressions est réalisée au moyen de tags SGML, dont l'étiquette est TIMEX2.

```
<TIMEX2> Halloween </TIMEX2>
```

Plusieurs attributs peuvent être ajoutés à la balise ouvrante afin de spécifier :

- une valeur normalisée de l'expression (VAL) ;
- la présence de modificateurs temporels (MOD) ;
- une valeur normalisée du moment de référence, ou ancrage (ANCHOR_VAL) ;
- la direction de ce point de référence (ANCHOR_DIR) ;
- si l'expression dénote un ensemble temporel (SET) ;
- les commentaires de l'annotateur (COMMENT).

Plusieurs catégories d'expressions temporelles sont distinguées. Celles-ci peuvent être **précises** ou **floues**. Il peut également s'agir de **fréquences** ou d'expressions **non-spécifiques**.

Les expressions temporelles précises

Il s'agit des expressions dont on peut déterminer la date calendaire, l'instant de la journée ou la durée qu'elles dénotent. Les **dates** sont annotées quelle que soit leur granularité.

```
<TIMEX2 VAL="1994"> 1994 </TIMEX2>
<TIMEX2 VAL="1998-11"> November </TIMEX2>
<TIMEX2 VAL="1998-07-14"> yesterday </TIMEX2>
```

Dans le cas d'un intervalle délimité par deux expressions explicites, chaque expression est annotée séparément. Pour les expression ancrées, l'annotation peut être imbriquée.

```
<TIMEX2 VAL="1999-08-03"> two weeks from <TIMEX2 VAL="1999-07-20"> next
    Tuesday </TIMEX2> </TIMEX2>
```

Les expressions qui dénotent des unités plus grandes que l'année reçoivent un attribut VAL dans un format particulier, ce qui constitue une extension par rapport à la norme ISO 8601.

```
<TIMEX2 VAL="196"> the 1960s </TIMEX2>
<TIMEX2 VAL="10"> 11th century </TIMEX2>
```

Les **instants de la journée** incluent la date du jour concerné, s'il est explicite ou s'il peut être déterminé, ainsi que la désignation de la portion de la journée :

```
<TIMEX2 VAL="1984-01-03T12:00"> twelve o'clock January 3, 1984 </TIMEX2>
```

Les **expressions dont l'unité est la semaine** sont encodées de manière spécifique :

```
<TIMEX2 VAL="1999-W29"> next week </TIMEX2>
```

Enfin, les **durées** sont les expressions qui indiquent combien de temps quelque chose dure. Les valeurs (VAL) sont exprimées selon le format ISO 8601. Elles peuvent être ancrées et orientées par rapport à un autre point ou période. Les valeurs de direction d'ancrage sont WITHIN, STARTING, ENDING, AS_OF, BEFORE et AFTER.

```
<TIMEX2 VAL="PT3H" ANCHOR_DIR="WITHIN" ANCHOR_VAL="1999-07-15"> three-hour </TIMEX2>
```

Les expressions temporelles floues

Ces expressions sont celles pour lesquelles les bornes sont imprécises. Dans le cas des **expressions temporellement imprécises**, l'annotation ne portera que sur l'unité (ou granularité) présente dans l'expression. Une date *complète*, qui spécifie un jour précis, ne peut donc pas être mentionnée en tant que valeur.

```
<TIMEX2 VAL="1998"> a year ago </TIMEX2>
```

Pour les **expressions générales relatives au passé, présent ou futur** (« now », « in a couple of days », « a few months ago »), une nouvelle extension à la norme ISO 8601 est introduite. Elle permet d'insérer une valeur alphabétique pour le champ VAL. Il est dès lors possible d'utiliser les valeurs PRESENT_REF, FUTURE_REF et PAST_REF pour indiquer de quelle type de référence il s'agit.

Les **saisons** sont également des unités floues. En effet, certains parleront de l'hiver comme de la période froide de l'année, alors que d'autres feront référence à la période telle que définie par rapport aux solstices et équinoxes. Ces cas sont gérés en ajoutant une expression alphabétique (SP, SU, FA, WI[11]) à la place du mois dans la valeur au format ISO. À noter que l'hiver est un cas problématique car il s'étend sur deux années différentes. Sans indication contraire, l'année par défaut sera celle contenant les mois de janvier à mars.

[11] Spring, Summer, Fall, Winter.

```
<TIMEX2 VAL="1998-FA"> Fall 1998 </TIMEX2>
<TIMEX2 VAL="1999-WI"> an unusually mild winter </TIMEX2>
<TIMEX2 VAL="P1WI" ANCHOR_DIR="STARTING" ANCHOR_VAL="1999"> all winter </TIMEX2>
```

Les **années fiscales** sont également sujettes à interprétation. À nouveau, une valeur alphabétique (FY) sera utilisée, en préfixe de l'année cette fois.

```
<TIMEX2 VAL="FY1998"> fiscal 1998 </TIMEX2>
```

De même, les **trimestres** (Q1, Q2, Q3 et Q4) et les **semestres** (H1 et H2) exploitent des valeurs alphabétiques. Les **week-ends** sont encodés à l'aide du code WE. Les **périodes de la journée** sont aussi des notions variables. Les codes MO (« morning »), MI (« mid-day »), AF (« afternoon »), EV (« evening »), NI (« night »), PM (« PM ») et DT (« day time », « working hours ») sont utilisés.

Dates ou moments de la journée dont des composants sont non-spécifiés

Certaines de ces dates peuvent être prises en compte, en masquant une partie de la valeur temporelle.

```
<TIMEX2 VAL="FY1998"> fiscal 1998 </TIMEX2>
in <TIMEX2 VAL="XX63">'63</TIMEX>
```

Les **durées non-spécifiées** peuvent être codées à l'aide d'une valeur de type PXY, interprétée comme « a period of X years ».

Les **expressions combinant semaines et mois** sont des cas complexes qui ont nécessité une extension de la norme ISO 8601 au niveau du contenu du champ VAL.

```
<TIMEX2 VAL="1998-FA-WXX-5TNI" MOD="START"> early one Friday night in <TIMEX2
   VAL="1998-FA"> fall 1998 </TIMEX2> </TIMEX2>
```

Les **expressions temporelles modifiées** sont annotées en capturant le sens du modificateur. Pour ce faire le champ MOD doit se voir attribuer une valeur parmi BEFORE, AFTER, ON_OR_BEFORE, ON_OR_AFTER uniquement pour les points, LESS_THAN, MORE_THAN, EQUAL_OR_LESS, EQUAL_OR_MORE uniquement pour les durées, ou encore START, MID, END ou APPROX.

```
The trend began in <TIMEX2 VAL="196" MOD="START"> the early 1960s </TIMEX2>
```

Les expressions dont le décalage est approximatif peuvent être codées en combinant les balises VAL et MOD.

```
<TIMEX2 VAL="1994" MOD="AFTER" ANCHOR_DIR="BEFORE" ANCHOR_VAL="1995"> Nearly five
    years ago </TIMEX2>, the plan [...]
```

Les **ensembles d'expressions temporelles** sont particulières, car elles permettent d'exprimer la fréquence d'un événement. La balise SET reçoit la valeur YES et le champ VAL prend une valeur générale permettant de décrire l'ensemble.

```
They watched Millionaire on TV <TIMEX2 SET="YES" VAL="1999-WXX-2"> every Tuesday in
    <TIMEX2 VAL="1999"> 1999 </TIMEX2></TIMEX2>
<TIMEX2 VAL="XXXX-WI" SET="YES"> Some winters </TIMEX2>, he was too sick to go
    to school
```

Les **expressions temporelles non-spécifiques** sont des expressions génériques (« I love December ») ou indéfinies (« The election took place on a Tuesday »). Elles ne sont pas encodées grâce à un attribut spécial mais peuvent être exprimées à l'aide d'une valeur VAL générale pour les expressions calendaires, par l'absence de l'attribut ANCHOR pour les durées ou encore en omettant purement et simplement l'attribut VAL.

Timex2 permet encore d'annoter plusieurs cas tels que les pronoms et les éléments élidés, les expressions temporelles ancrées à un événement, les expressions culturellement variables, les expressions dont la valeur peut changer ainsi que les expressions métonymiques.

6.5.3 Timex3 et TimeML

Timex3 est le format d'annotation temporel faisant partie de TimeML. Ce langage d'annotation a la particularité de s'intéresser au problème du repérage des événements et de leur ancrage temporel (Pustejovsky *et al.* [2003a], Boguraev *et al.* [2005] et Saurí *et al.* [2006]). TimeML est issu du projet TERQAS (Time and Event Recognition for Question Answering Systems, Pustejovsky *et al.* [2002]). TimeML a été principalement élaboré sur la base de Timex2 et du langage d'annotation STAG proposé par Setzer [2001].

Contrairement à Timex2, le langage rend possible l'identification de **signaux** intervenant dans l'interprétation des expressions temporelles, entre autres les prépositions temporelles (« for », « during », « on », « at ») et les connecteurs temporels (« before », « after », « while »).

Une autre particularité par rapport à Timex2 est l'identification de différentes classes d'**événements**. Ils sont définis comme étant des « situations that happen or occur » et peuvent être ponctuels ou être constitués d'une période de temps. Il peut également s'agir d'un état. Les événements sont définis comme étant des verbes conjugués (« was captured », « will resign »), des adjectifs dénotant un état (« sunken », « stalled », « on board ») ou des groupes nominaux (« merger », « Gulf War »).

Enfin, il est également possible de créer des **liens de dépendance entre les événements et les expressions temporelles** : un *ancrage* (« John left on Monday »), un *ordonnancement* (« The party happened after midnight ») ou une *imbrication* (« John said Mary left »).

Quatre structures de données sont donc proposées : EVENT, TIMEX3, SIGNAL et LINK. Un point particulier à souligner est la séparation de la description des événements (balises EVENT) de celle des relations dans lesquelles ils interviennent (balises LINK).

Les expressions **TIMEX3** sont des expressions temporelles explicites, c'est-à-dire :

- des expressions temporelles complètement spécifiées (« June 11 », « Summer », « 2002 »),
- des expressions temporelles sous spécifiées (« Monday », « next month », « two days ago »),
- des durées (« three months », « two years »).

D'une manière générale, TIMEX3, en dehors du contexte de TimeML, n'apporte pas de grandes évolutions à TIMEX2. À noter parmi les changements intervenus, certains éléments de syntaxe qui ont été modifiés, les imbrications qui ne sont plus permises et l'orientation vers une plus grande fragmentation, plutôt que vers le rassemblement des expressions complexes.

Les balises **SIGNAL** sont utilisées pour annoter des sections de textes qui indiquent comment les objets temporels doivent être reliés entre eux. Il peut s'agir :

- de prépositions temporelles (« on », « during »),
- de connecteurs temporels (« when »),
- de conjonctions de subordination (« if »),
- d'indicateurs de polarité (« not », « no », « none »),
- d'indicateurs de quantification temporelle (« twice », « three times »).

Les balises **LINK** se déclinent selon trois types : les liens *temporels* (TLINK), les liens de *subordination* (SLINK) et les liens *aspectuels* (ALINK).

Le lien temporel **TLINK** permet de spécifier une palette assez large de relations entre événements ou entre un événement et une valeur temporelle. Elles correspondent à celles définies par Allen [1984][12].

Les liens de subordination **SLINK** sont utilisés pour relier deux événements ou un événement et un signal. Un lien peut être :

- *modal*, pour les relations introduites par les verbes modaux (« should », « could », « would ») ;
- *factif*, lorsque certains verbes introduisent des présuppositions sur la véracité d'un argument (« Mary regrets that she didn't marry John ») ;
- *contrefactif*, lorsque l'événement introduit une présupposition au sujet de la non-véracité de ses arguments (« Mary was unable to marry John ») ;
- *probant* pour les situations de rapport ou de perception (« John said he bought some wine ») ;
- *non-probant* pour les situations de rapport ou de perception ayant une polarité négative (« John denied he bought only beer ») ;
- *negatif*, lorsque des particules négatives marquées comme SIGNAL (« not », « neither ») sont utilisées.

Les liens aspectuels **ALINK** permettent de relier l'événement *aspectuel* à son événement *argument*. Les différentes relations aspectuelles sont :

- le début d'un événement ;
- l'aboutissement d'un événement ;
- l'arrêt d'un événement ;

[12] Celles-ci expriment le fait que ces entités peuvent être simultanées, identiques, l'une avant l'autre, l'une après l'autre, l'une immédiatement avant l'autre, l'une immédiatement après l'autre, l'une incluant l'autre, l'une incluse dans l'autre, l'une se déroulant durant l'autre, l'une débutant l'autre, l'une étant débutée par l'autre, l'une finissant l'autre, l'une étant finie par l'autre.

– la poursuite d'un événement.

6.5.4 Formats ad-hoc

À côté des formats *standards*, qui constituent des solutions utilisées assez largement, de nombreux travaux choisissent également, pour des raisons diverses, de définir leur propre format d'annotation. Il n'est évidemment pas possible d'en donner un aperçu complet, mais citons entre autres Martineau *et al.* [2009] qui s'insèrent dans un cadre plus large d'extraction d'entités nommées, Weiser [2010] pour les expressions temporelles qui expriment des horaires, ou encore Le Parc-Lacayrelle *et al.* [2007]. Ces derniers estiment que la fragmentation de l'annotation des expressions complexes, et l'attribution de valeurs temporelles à celles-ci[13], n'est pas adaptée à l'utilisation dans le cadre de leur application de recherche d'informations géographiques.

Dans le cadre de notre implémentation d'un système d'extraction d'informations temporelles, qui est présenté au chapitre 7, nous avons également choisi d'utiliser un format intermédaire[14] personnel. En ce qui concerne le format de sortie final, nous en prévoyons plusieurs, dont les formats *standards* ne sont pas nécessairement exclus. Le choix d'un format d'annotation est abordé plus spécifiquement à la section 7.7.1.

6.6 Méthodes et techniques d'extraction

Les méthodes utilisées en extraction d'informations temporelles sont diverses. Comme pour beaucoup de tâches en traitement automatique des langues, on peut classer les approches selon deux principaux types. Premièrement, les méthodes symboliques, souvent basées sur des ressources linguistiques et des systèmes de règles, et dont la caractéristique principale est de pouvoir offrir un très haut niveau de précision. L'obtention d'un rappel acceptable se fait cependant souvent au prix d'un effort conséquent de développement de ressources.

Le deuxième type d'approche est constitué par les techniques axées sur l'apprentissage automatique. Celles-ci sont souvent considérées comme moins dépendantes du domaine d'application (et de la langue) et plus robustes, mais parfois au détriment de la précision. Entre ces deux pôles, il existe également des méthodes hybrides.

Les caractéristiques de ces méthodes les rendent plus propices à certaines tâches particulières de l'extraction d'informations temporelles. Ahn *et al.* [2005] montrent que si les techniques d'apprentissage permettent d'atteindre de bons résultats pour les tâches de reconnaissance des expressions temporelles, l'interprétation de celles-ci reste un problème qu'il est plus aisé de résoudre à l'aide de règles. En effet, le grand nombre de classes[15] que pourrait attribuer un algorithme de classification

[13] Par exemple, « le début de juin 1963 » est annoté à l'aide d'une balise TIMEX3 (avec *value="1963-06"*) pour « juin 1963 », d'une balise SIGNAL pour « début », le tout étant relié par une balise TLINK. Cette dernière n'apporte cependant pas de champ *value* pour attribuer une valeur à l'expression complète.

[14] C'est-à-dire pour les expressions reconnues, avant interprétation et normalisation.

[15] Les classes seraient constituées par les différentes valeurs temporelles.

à une expression temporelle, ainsi que la nécessité d'analyser un contexte parfois fort éloigné de l'expression pour l'interpréter sont des éléments qui rendent les techniques d'apprentissage difficiles à adapter pour cette tâche.

6.6.1 Approches symboliques

D'un manière générale, l'approche symbolique est adoptée par de nombreux travaux, que ce soit pour la phase de reconnaissance ou celle d'interprétation. La description et la reconnaissance des expressions temporelles par automates à états finis[16] a été exposée à plusieurs reprises, entre autres par Maurel [1990] sur les dates et adverbes apparentés et par Gross [2002] au sujet des déterminants numéraux et plus spécifiquement des dates horaires. Dans la même ligne, Fairon et Senellart [1999], Mani et Wilson [2000], Schilder et Habel [2001] ou Baptista et Guitart [2002] adoptent tous une approche par automates à états finis ou transducteurs. Les résultats obtenus se situent généralement à un bon niveau. Par exemple, Schilder et Habel [2001] rapportent une précision de 92,11% et un rappel de 94,09% pour les expressions temporelles simples. Le repérage des expressions complexes (incluant des prépositions) obtient une précision de 87,30% pour un rappel de 90,66%. Toujours dans le monde des méthodes à états finis, Muller et Tannier [2004] présentent un système dans lequel la reconnaissance des expressions temporelles est réalisée au moyen d'une analyse grammaticale partielle. Celle-ci est réalisée à l'aide de l'analyseur Cass (Abney [1996]) et est constituée d'une cascade d'expressions régulières (89 règles réparties en 29 niveaux) permettant :

- d'annoter les dates ;
- de localiser les constituants temporels non essentiels à la phrase (*temporal adjuncts*, groupes prépositionnels) ;
- d'extraire divers marqueurs temporels ;
- de délimiter des propositions ne contenant qu'un seul verbe fini ;
- de rattacher les constituants temporels non essentiels à une des propositions délimitées ;
- de traiter les propositions relatives (de manière similaire à Filatova et Hovy [2001]).

Même si l'approche par automates est parmi les premières à avoir été utilisées, de nombreux travaux récents continuent à l'exploiter. Citons Battistelli *et al.* [2006] pour les expressions calendaires, et Vicente-Díez *et al.* [2008] qui obtiennent un taux de reconnaissance de 82,7% des expressions temporelles à l'aide d'un ensemble de 22 patrons. Bittar [2008, 2009] se base sur la collection Time_French de grammaires locales de Gross [2002] et rapporte une précision de 84,2% et un rappel de 81,8% (f-mesure de 83,0%). Les développements de Parent *et al.* [2008] aboutissent à un rappel de 79% pour une précision de 83% (f-mesure de 81%). Hagège et Tannier [2008] utilisent une approche par automates finis au moyen du logiciel XIP de Xerox, alors que Martineau *et al.* [2009] proposent un système basé sur Unitex et des transducteurs pour l'annotation et la normalisation d'entités nommées, dont certaines expressions temporelles. Enfin, Weiser [2010] se base également sur une série de transducteurs pour repérer des expressions temporelles, principalement des horaires, dans des

[16] Les termes grammaires locales ou transducteurs sont également employés pour désigner ce type de ressources.

pages Web liées au milieu touristique.

Les possibilités d'interprétation à l'aide d'automates semblent moins évidentes à mettre en œuvre à grande échelle, ou du moins celles-ci exigent un travail conséquent pour le développement des ressources nécessaires. Maurel et Mohri [1994] démontrent en tout cas que les automates constituent une solution possible pour un certain nombre d'expressions, telles que des adverbes temporels qui n'impliquent pas de compléments.

Les règles utilisées pour reconnaître ou interpréter les expressions temporelles n'utilisent pas nécessairement des automates à états finis. Elles peuvent être mises en œuvre de différentes façons. Dans Filatova et Hovy [2001], l'interprétation qui suit la reconnaissance et une découpe en propositions se déroule de la manière suivante. Pour les propositions contenant des informations temporelles explicites :

A. Pour un jour de la semaine (nom de jour)~:
  1. si le jour de la semaine de la date interprétée est le même que celui de la date de référence et
    - qu'il n'y a pas d'indice signalant que l'événement s'est déroulé avant ou après la date considérée, alors l'horodatage de référence est attribué;
    - si au contraire de tels indices sont présents, ils sont interprétés conjointement avec la dernière valeur temporelle attribuée;
  2. si le jour de la semaine de la date interprétée n'est pas le même que celui de la date de référence et
    - si des indices (mots-signaux) indiquent que l'événement s'est déroulé avant l'écriture de l'article ou que le temps verbal est le passé, la valeur temporelle est attribuée en accord avec les mots-signaux ou en choisissant le jour le plus récent correspondant au même nom de jour;
    - si des mots-signaux indiquent que l'événement s'est déroulé après que l'article ait été écrit ou que le temps verbal est le futur, la valeur temporelle est attribuée en accord avec les mots-signaux ou en choisissant le jour ultérieur le plus proche qui porte le même nom de jour.

B. Pour un nom de mois : les règles sont identiques, mais la valeur attribuée est un intervalle et non un point.

C. Pour les semaines, jours, mois, années avec modificateurs : un intervalle approximatif est attribué à la place d'une date particulière.

D. « When », « since », « after », « before » dans une phrase sans date : l'utilisateur est invité à insérer une date manuellement, cette date pouvant ensuite être utilisée comme référence pour l'interprétation d'une autre expression temporelle.

Pour les propositions ne contenant pas d'information temporelle explicite :

- les temps verbaux du présent et du past perfect vont induire un intervalle dont la borne de départ est inconnue et dont la borne de fin est la date de référence;
- un temps verbal du futur implique un intervalle démarrant au point de référence et dont la borne de fin est inconnue;
- le présent indéfini provoque l'assignation de la dernière date attribuée;
- le passé indéfini provoque souvent l'assignation de la dernière date attribuée ou de la date de l'article;
- enfin, s'il n'y a pas de verbe, c'est la dernière date attribuée qui est assignée.

Ce système atteint une précision variant entre 77,85% et 82,29%.

Schilder et Habel [2001] utilisent également un système de règles pour l'étape d'interprétation et obtiennent une précision de 84,49%.

Chez Vazov [2001], l'identification des expressions est réalisée au moyen d'une méthode basée sur la *context-scanning strategy* (CSS) de Desclés *et al.* [1997]. Le système met en œuvre à la fois une recherche par expressions régulières et un *chart parsing de gauche à droite* et *de droite à gauche*. Les marqueurs temporels exploités dans les expressions régulières sont au nombre de 121 et sont de deux types. Premièrement, des marqueurs *autonomes* (*stand-alone*), c'est-à-dire :

- des chaînes de caractères constantes (« par la suite », « le lendemain matin ») ;
- des éléments initiaux d'une chaîne considérée comme une expression temporelle (« quand »), accompagnés d'un ensemble d'unités lexicales bornées à droite par une certaine catégorie syntaxique (il s'agit principalement de propositions temporelles subordonnées que l'on peut circonscrire par la détection de la première ponctuation rencontrée après avoir au moins trouvé un verbe) ;

Deuxièmement, des marqueurs *déclencheurs*, c'est-à-dire des unités qui indiquent et font partie d'une expression temporelle plus large :

- des marqueurs qui sont des unités toujours présentes en position la plus à gauche dans l'expression temporelle et qui déclenchent par conséquent une analyse « de gauche à droite » (« il y a », « au cours ») ;
- des marqueurs qui interviennent dans les expressions temporelles, mais jamais dans les positions la plus à gauche ou la plus à droite, et qui impliquent dès lors une analyse à la fois « de gauche à droite » et « de droite à gauche » (« janvier », « minute », « printemps ») ;
- des marqueurs qui peuvent intervenir à n'importe quelle position de l'expression temporelle et qui entraînent également une analyse dans les deux sens ; ils peuvent également être trouvés seuls (« après »).

Le principe général de l'analyse réalisée par le *parser* consiste à vérifier à gauche et/ou à droite si la catégorie du mot correspond à un ensemble déterminé de catégories[17]. Il s'agit donc d'une analyse hors-contexte. Un certain nombres de contraintes, permettant de tenir compte du contexte, sont également observées :

- contrainte d'adjacence pour bloquer prématurément la reconnaissance de l'entité en présence de certaines séquences de catégories ;
- contrainte sémantique qui spécifie que les noms dans le contexte gauche doivent dénoter une période de temps (minutes, secondes, saisons) ;
- contrainte de dépendance symétrique pour gérer la conjonction « et » reliant deux parties d'une expression temporelle et pour exclure les emplois possessifs de « de » et « du »[18] ;

[17] Les catégories autorisées à gauche sont : DET, PRE, ADJ, PRO, NUM et NOC, alors que celles autorisées à droite sont DET, ADJ, NUM, NOC, ADV, INF, PPA, PPR, CLI, COJ, NOP.

[18] Par exemple, écarter « de cet hiver », mais garder « de 1980 à 1985 »

– contrainte sur les propositions relatives (dans le contexte droit) qui permet de différencier les cas tels que « La réunion a commencé et [3 minutes après] son porte-parole qui avait déjà annoncé » (la proposition modifie un sujet grammatical) de « Le ministre est venu [3 minutes après son porte parole] qui avait déjà annoncé » (la proposition modifie un argument du prédicat).

Le système, après une étape de *tokenisation*, exploite d'abord les marqueurs autonomes avant de rechercher des marqueurs déclencheurs. L'évaluation partielle du système a fourni un rappel qui tourne autour de 95% pour une précision d'environ 85%.

Chez Mani et Wilson [2000], l'interprétation des expressions temporelles repérées est réalisée à l'aide d'un ensemble initial de règles construites à la main. Celles-ci sont augmentées par des règles apprises automatiquement, ce qui en fait un système hybride. La précision atteinte est de 83,7% pour un rappel de 82,7%, soit une f-mesure de 83,2%.

6.6.2 Les techniques d'apprentissage

Les techniques d'apprentissage automatique sont évidemment aussi exploitées pour l'extraction d'informations temporelles. Par exemple, Adafre et de Rijke [2005] exposent une méthode qui exploite la technique d'apprentissage *Conditional Random Fields (CRFs)* pour la reconnaissance d'expressions temporelles. Les *features* de base sont constituées du contexte des expressions temporelles ainsi que de certaines de ses caractéristiques internes (capitalisation, utilisation de chiffres). Une autre *feature* importante est obtenue par l'utilisation d'une liste de mots qui apparaissent fréquemment dans les expressions temporelles (noms de jours ou de mois, unités temporelles). Cette liste est également utilisée dans une phase de post-traitement des résultats du classifieur, dans le but d'en augmenter le rappel. Ce système obtient d'assez bons résultats : une précision de 86,1% pour un rappel de 80,4% en reconnaissance exacte et une précision de 97% pour un rappel de 90,6% en reconnaissance partielle.

Hacioglu *et al.* [2005] considèrent la tâche de reconnaissance des expressions temporelles comme un problème de classification dans lequel chaque token peut être considéré comme commençant, finissant, étant à l'intérieur ou à l'extérieur d'une telle expression. Ces quatre classes, ou catégories, sont étendues à dix pour prendre en compte les cas d'imbrications d'expressions temporelles. Le système de classification est composé de 10 classifieurs de type *un contre tous* (un par classe). Ces classifieurs exploitent un certain nombre de *features* :

– lexicales : les tokens, leurs versions décapitalisées, les catégories de discours (POS), des caractéristiques morphologiques telles que la présence d'un trait d'union ou de chiffres, les fréquences (rare, fréquent, inconnu) ;
– syntaxiques : les chunks ;
– sémantiques : les têtes lexicales et les relations entre celles-ci et les autres mots ;
– externes : les indications données par le système d'annotation temporelle distribué à

l'occasion de TERN[19] (basé sur des règles) et celles fournies par le système BBN IdentiFinder[20].

Le système donne des résultats assez intéressants[21] : une précision de 97,8% pour un rappel de 89,4%, soit une f-mesure de 93,5% pour la détection. Ces valeurs sont revues à la baisse si l'annotation complète avec parenthesage est considérée (rappel=91,9%, précision=84,0%, f-mesure=87,8%).

Ahn *et al.* [2007] utilisent une cascade de classifieurs. La tâche de reconnaissance des expressions temporelles est réalisée par un algorithme de type SVM (*Support Vector Machine*). Les performances rapportées s'échelonnent entre 85% et 90,5% pour la précision, et 73,2% et 82,9% pour le rappel. Les composants syntaxiques, restreints à certaines catégories (ADVP, ADJP, NN, NNP, JJ, CD, RB, et PP)[22], sont classifiés en deux classes distinctes : *timex* et *non-timex*. Un deuxième classifieur, lui aussi de type SVM, permet ensuite de déterminer la classe sémantique[23] de l'expression. Enfin, si l'interprétation de l'expression n'est pas directe (pour les expressions temporelles référentielles par exemple), les règles et le classifeur SVM (pour déterminer la direction de la référence) sont combinés pour aboutir à une valeur normalisée. Ce système, même s'il utilise massivement les techniques d'apprentissage, peut être considéré comme hybride car il continue à faire appel à un certain nombre (limité) de règles.

6.7 Langue cible et aspect multilingue

Les travaux en extraction d'informations temporelles sont sensibles à la langue des textes traités. Il est évident que l'extraction en anglais, en français, en coréen ou dans tout autre langue est influencée par les multiples différences (lexicales, grammaticales voire même conceptuelles) présentées par ces langues. Si certaines techniques semblent pouvoir s'affranchir du lien avec la langue cible, elles nécessitent toujours des données de base dans la langue en question.

Parmi tous les travaux menés en extraction d'informations temporelles, nombreux sont ceux à être appliqués à l'anglais. La plupart des grandes initiatives telles que les conférences (MUC, ACE, TERN, TIME, Chronos) ou la mise au point de formats d'annotation (Chinchor [1997] avec Timex, Ferro *et al.* [2005] avec Timex2, et Boguraev *et al.* [2005] pour TimeML et Timex3) et de ressources (TimeBank, Pustejovsky *et al.* [2003b]) proviennent également du milieu anglophone. Cette vitalité s'est traduite par un grand nombre de publications, dont entre autres Mani et Wilson [2000], Filatova et Hovy [2001], Schilder et Habel [2001], Setzer [2001], Mani et Schiffman [2005], Adafre et de Rijke [2005], ou encore Hagège et Tannier [2008].

Ces avancées bénéficient aussi aux travaux entrepris pour d'autres langues. C'est par exemple le cas du français, langue pour laquelle il existe déjà des travaux publiés il y a un certain nombre d'années (Maurel [1990], Maurel et Mohri [1994]), mais aussi plus récemment (Vazov [2001], Muller

[19] `http://fofoca.mitre.org/tern.html`
[20] `http://www.bbn.com/technology/speech/identifinder`
[21] Pour l'anglais. Les résultats sur le chinois, qui exploite un ensemble différent de *features*, étant légèrement inférieurs.
[22] Ces catégories ont été tirées de Ferro *et al.* [2005].
[23] C'est-à-dire son type en tant que valeur temporelle, par opposition au type en tant qu'expression temporelle.

et Tannier [2004], Battistelli *et al.* [2006], Bittar [2008, 2009], Parent *et al.* [2008], Martineau *et al.* [2009] et Weiser [2010]). Il faut également noter l'apparition progressive de ressources linguistiques, telles qu'un *TimeBank* pour le français (Bittar [2010]).

D'autres langues, qu'elles soient indo-européennes (par exemple l'allemand avec Schilder et Habel [2001], l'espagnol avec Vicente-Díez *et al.* [2008], le portugais dans Baptista et Guitart [2002] et l'italien chez Caselli *et al.* [2008]) ou provenant d'autres origines (notons le cas du chinois avec Li *et al.* [2001] et Cheng *et al.* [2007] ou du coréen avec Jang *et al.* [2004]) ont bien entendu bénéficié de l'attention des chercheurs en matière d'extraction d'informations temporelles.

Enfin, divers travaux tendent à mettre en œuvre, du moins pour certaines parties du traitement de la temporalité, des techniques multilingues. C'est entre autres le cas de Wilson *et al.* [2001] ou de Ahn *et al.* [2005]. Dans une optique de traduction automatique, les aspects temporels sont également abordés par Fairon et Senellart [1999] et Lecuit *et al.* [2009].

CHAPITRE 7

IMPLÉMENTATION D'UN SYSTÈME D'EXTRACTION D'INFORMATIONS TEMPORELLES

7.1 Introduction

Dans ce chapitre, nous exposons les développements effectués pour la construction d'un système d'extraction d'informations temporelles. Comme nous l'avons montré au chapitre 4, les expressions temporelles auxquelles nous nous intéressons vont bien plus loin que la simple date « jour-mois-année ». Même si ce type d'expressions, qui désigne une zone temporelle bien identifiée dans l'espace du temps, est effectivement présent dans les textes, il est évident qu'il n'est pas nécessairement le plus fréquent[1]. On trouve aussi très couramment des dates incomplètes, c'est-à-dire que l'on ne peut directement, et de manière univoque, rattacher à un point précis du calendrier. D'autre part, on rencontre également à de nombreuses reprises des expressions imprécises ou approximatives[2]. Celles-ci sont généralement utilisées par le locuteur soit parce que son propos ne nécessite pas de localiser temporellement un fait avec précision, soit parce que cette localisation précise ne lui est pas connue. Le caractère imprécis fait donc partie de la nature même de l'expression du temps en langage naturel[3]. Il n'est donc pas nécessairement toujours pertinent de vouloir interpréter les expressions en une valeur précise. Une analyse temporelle se doit de tenir compte de ces différents aspects afin d'assurer la reconnaissance la plus complète possible des informations temporelles contenues dans un texte.

L'extraction et l'interprétation des éléments qui présentent une valeur temporelle dans un texte, nécessitent un travail de développement important qui comprend :

- la définition des catégories d'expressions temporelles et leur spécification détaillée ;
- la réalisation d'une ressource d'extraction capable de reconnaître ces éléments dans un texte ;
- le rassemblement de diverses autres informations (repérage des verbes et de leurs temps morphologiques, délimitation des propositions, etc.).

[1] La distribution des types d'expressions temporelles varie évidemment en fonction du type de textes. La référence est ici constituée par des textes journalistiques, des dépêches de presse, etc.

[2] La section 7.6.1 expose de manière plus précise les notions d'expression incomplète d'une part, et approximative ou floue d'autre part.

[3] Cette caractéristique est également liée à la notion de granularité. Ce lien est abordé plus en détail à la section 7.3.2.

Tout ces éléments, analysés conjointement, permettent alors d'interpréter temporellement le texte.

7.2 Positionnement et objectifs

7.2.1 Positionnement

Au chapitre 4, les travaux présentés portaient sur le temps tel qu'il est abordé par les théories linguistiques. Sa conceptualisation et sa modélisation ont été ensuite exposées au chapitre 5. Enfin les travaux menés en extraction d'informations ont été présentés, selon divers points de vue, au chapitre 6. Les développements présentés dans le présent chapitre se situent clairement dans ce dernier cadre. Par conséquent, nous adoptons une démarche pratique et concrète, mais n'écartant pas pour autant les apports plus théoriques. En effet, comme nous l'avons souligné à la section 3.4, l'extraction d'informations se nourrit nécessairement de nombreux éléments, provenant à la fois des théories linguistiques et des travaux visant à conceptualiser et modéliser le temps. L'apport de ces différentes théories est exposé dans la partie consacrée au modèle d'interprétation temporel qui a été mis en œuvre (Section 7.3).

D'une manière générale, l'approche adoptée en extraction d'informations a généralement pour finalité la mise au point d'une application concrète. Cette caractéristique a des conséquences importantes. Elle implique souvent de simplifier certains phénomènes linguistiques complexes pour lesquels il n'existe pas de traitement automatique satisfaisant, ou suffisamment rapide et robuste, pour être utilisé dans un but applicatif. Il est aussi fréquemment nécessaire de se contenter d'une couverture partielle du phénomène, du moins dans un premier temps. Dans de nombreux cas, un système imparfait peut déjà apporter une aide satisfaisante[4]. Cette constatation est particulièrement vraie dans un contexte où la masse documentaire est importante et où une analyse manuelle se révélerait longue et ardue.

Comme nous l'avons mentionné au début de cette section, les développements qui sont présentés dans ce chapitre sont à considérer dans cette optique d'extraction d'informations. Ainsi, les différentes étapes nécessaires au traitement automatique du temps dans les textes sont à la fois incomplètes et imparfaites. Elles n'en sont pas moins utiles pour autant. Le système mis en place ne permet certainement pas d'atteindre l'exhaustivité et la finesse d'analyse que pourrait avoir un linguiste face à cette tâche. Mais grâce aux éléments empruntés aux théories issues de ce domaine, l'extraction des informations temporelles peut être effectuée automatiquement, d'une manière rapide, autorisant ainsi le traitement de grandes quantités de données textuelles. L'extraction d'informations doit composer avec diverses contraintes, entre autres la disponibilité et la performance des technologies permettant de mener à bien les analyses de base sur lesquelles repose le système. Au-delà de la réalisation de ce système, l'intérêt réside donc aussi dans la détermination de la meilleure manière d'opérationnaliser les connaissances théoriques, dans le but d'en faire une application performante.

[4] Les performances d'un système d'extraction d'informations peuvent être évalués par rapport à une analyse idéale, telle que pourrait l'effectuer un *expert* humain. Son *utilité* se mesure cependant de manière moins directe.

7.2.2 Objectifs

Dans les nombreux travaux en extraction d'informations temporelles, le traitement du temps a été abordé à plusieurs niveaux, du repérage des expressions temporelles, en passant par leur interprétation et en aboutissant finalement à l'ordonnancement des événements d'un texte. Cette dernière tâche est considérée comme la plus ambitieuse, car elle nécessite la résolution des problèmes posés par les étapes précédentes.

Dans cette thèse, les deux premières tâches évoquées ci-dessus sont couvertes. La troisième n'est pas abordée, et cela pour plusieurs raisons. Tout d'abord, nous estimons qu'elle sort du champ strict de l'extraction d'informations temporelles. Il s'agit plutôt d'une tâche de plus haut niveau[5], qui, à l'instar d'applications telles que l'extraction d'informations biographiques ou l'indexation à dimension temporelle, exploite les résultats des étapes précédentes. D'autre part, s'il est indéniable que les événements possèdent une dimension temporelle, cette dimension est difficilement accessible sans de larges connaissances sur les caractéristiques des différents types d'événements. De plus, alors que les expressions temporelles constituent un ensemble relativement régulier qui peut être formalisé au moyen d'un nombre fini[6] de patrons, les événements ne le sont pas vraiment, du moins de manière large. De fait, le concept d'événement, qui n'est pas défini de manière consensuelle, peut selon la situation, être interprété de nombreuses façons.

La position adoptée dans le cadre de ce travail est qu'il est préférable de concentrer les efforts sur la tâche précise de l'analyse temporelle du texte. L'extraction d'événements, ou de manière plus large l'extraction d'informations, quelles qu'elles soient, est un travail à part entière, qui nécessite souvent un investissement particulier au domaine ou au type de données visé. Il semble donc difficile de traiter ce point en tant que problème annexe à celui de l'extraction temporelle.

Cette limitation du système au cœur même de l'extraction d'informations temporelles permet également d'envisager celui-ci comme un module qui peut être exploité au sein d'autres applications, plus complexes. Il est donc important d'insister sur les possibilités d'interaction et d'intégration avec d'autres tâches, et de penser aux moyens à mettre en œuvre pour rendre celles-ci possibles. Dans cette optique, l'ouverture est une des caractéristiques dont un système d'extraction d'informations temporelles doit être doté. Celui que nous proposons dans ce chapitre est ainsi capable d'intégrer certains processus d'analyse tiers ou, à défaut, de s'intégrer à ceux-ci en leur fournissant des versions sémantiquement annotées (sur les aspects temporels) des textes. Un exemple du premier cas de figure est proposé au chapitre 8 avec une application d'indexation à dimension temporelle. Le second cas de figure est lui illustré à la section 7.10 (Figure 7.14) à l'aide du cas de l'extraction d'informations biographiques[7], dans lequel l'information temporelle tient un rôle particulièrement important.

Avec cette séparation des tâches d'extraction, subsiste cependant un point essentiel qui concerne l'établissement de liens entre les informations temporelles et les autres informations extraites. Ces liens peuvent être établis de nombreuses façons, qui dépendent en partie de l'application visée. Il peut

[5] L'ordonnancement d'événements peut constituer une fin en soi, l'extraction temporelle l'est plus rarement.
[6] Cela ne veut évidemment pas dire que ce nombre est faible.
[7] Voir aussi Kevers [2006]; Kevers et Fairon [2007] sur ce sujet.

par exemple s'agir de relations de co-occurrence, ou de liens syntaxiques plus élaborés. Cet aspect est lui aussi renvoyé au niveau de l'application, qui peut ainsi l'implémenter selon ses exigences. Cependant, dans le cas de l'intégration d'une analyse tierce dans le même processus de traitement que celui consacré aux aspects temporels, la création des liens fait alors partie intégrante de ce processus (voir chapitre 8).

7.3 Modèle pour une interprétation temporelle

Parmi l'ensemble des éléments qui sont porteurs d'une information temporelle dans le langage naturel (Section 4), tous n'ont pas été exploités par la démarche d'extraction d'informations développée. Le *modèle* d'interprétation du temps qui est proposé s'attache donc à déterminer quels sont ceux qui doivent, et qui peuvent, être pris en compte, et de quelle manière.

7.3.1 Éléments d'information pris en compte

La date d'émission du texte est une information fondamentale car elle constitue un point de repère par rapport auquel vont se situer de nombreuses informations temporelles du texte. Dans une série de situations, la date d'émission n'a que peu de rapport avec le contenu informationnel du document. Les romans et les chroniques historiques constituent deux exemples évidents à cet égard. Cependant, nombreux sont les types de textes à être ancrés dans l'actualité de leur moment d'émission. Les textes de presse, et spécialement les dépêches de presse, en constituent un exemple par excellence[8]. Par conséquent, la date d'émission est considérée comme une métadonnée du texte, essentielle et obligatoire à son analyse. Elle constitue un prérequis à l'analyse automatique et doit donc toujours être présente et identifiée en tant que telle dans les textes.

L'analyse temporelle du texte est principalement alimentée par les adverbes et locutions adverbiales (voir section 4.2). Ces éléments constituent véritablement le cœur du système car ils présentent le double avantage d'être un vecteur très fort pour l'information temporelle, tout en pouvant être repérés efficacement. Ils représentent un moyen largement utilisé pour fixer les repères temporels d'un récit, surtout lorsqu'il s'agit d'une référence qui se rapporte de manière assez précise à l'espace du temps modélisé sous la forme d'un calendrier. Bien entendu, de nombreuses expressions adverbiales désignent aussi des zones temporelles de manière imprécise. Cette caractéristique est prise en compte et conservée lors du traitement automatique.

Autre source d'information importante, les temps verbaux donnent des indications utiles lors de l'analyse des adverbes relatifs. Ces derniers sont en effet interprétés à partir d'un point de repère, explicite ou implicite, et dans une certaine direction temporelle. Les temps verbaux contribuent à indiquer si l'interprétation de l'adverbe doit s'effectuer dans le passé, le futur ou le présent.

L'interprétation des temps verbaux actuellement implémentée reste cependant relativement basique,

[8] Rien n'empêche cependant les textes de presse de s'ancrer temporellement ailleurs que dans leur *présent*, mais cela ne constitue pas la règle du genre.

et ne va pas jusqu'à exploiter toutes les finesses de la langue. Ce modèle est par conséquent appelé à évoluer afin d'affiner la mise en relation du temps grammatical (ou linguistique) avec le temps notionnel (ou chronique). Comme nous l'avons vu au chapitre 4, et plus particulièrement à la section 4.7.6, l'interprétation de l'*aspect grammatical*[9], pourrait être une information profitable. L'intégration d'un modèle complet pour les temps verbaux, tel que ceux évoqués à la section 4.7 ne semble cependant pas chose évidente. En effet, l'explication du fonctionnement de ces modèles reste encore un défi en linguistique, et il est dès lors compliqué d'envisager leur implémentation. Il existe d'ailleurs très peu de systèmes qui intègrent un environnement complet de traitement des temps verbaux. Un de ceux qui existent est l'implémentation inspirée du modèle des intervalles de Gosselin proposé par Person [2004]. Ce système présente cependant certaines limites[10].

En pratique, le repérage et l'analyse des groupes verbaux est en grande partie réalisée à partir des informations issues de l'analyse syntaxique. Il s'agit en particulier des éléments annotés par XIP (Aït-Mokhtar *et al.* [2002]) à l'aide des étiquettes relatives aux groupes verbaux finis (« FV »), infinitifs (« IV ») et gérondifs (« GV »). L'annotation qui concerne les formes verbales passives (« AUXIL_PASSIVE ») est également exploitée. Pour les formes verbales simples, les informations fournies par l'analyse en parties du discours (Treetagger, Schmid [1994]) sont également prises en compte. En cas de conflit entre XIP et le Treetagger au sujet d'un code grammatical d'un forme verbale simple, c'est l'analyse du Treetagger qui est privilégiée.

Lors de leur traitement, les adverbes et les temps verbaux sont analysés en fonction de leur contexte. Celui-ci est constitué, au sein de la phrase, par la proposition. La découpe en propositions doit donc faire partie des prétraitements effectués sur le texte. Cette découpe est principalement réalisée sur la base d'informations fournies par l'analyse syntaxique (XIP). En particulier, les séparateurs de propositions peuvent être insérés au niveau des éléments étiquetés « BG », qui marquent le début d'une clause, et « PUN » qui identifient les signes de ponctuation. Dans le cas où une proposition contient plus d'un verbe, celle-ci est scindée, soit sur un signe de ponctuation (par exemple une virgule), soit juste avant un verbe, et cela de manière à n'obtenir qu'un seul verbe par proposition. Le temps du verbe caractérise l'ensemble de la proposition à laquelle il appartient. L'utilité de cette découpe est de lier une expression temporelle avec le verbe qui permet de l'interpréter. Ce lien, lorsqu'il est effectué sur des segments suffisamment fins, peut être réalisé au moyen d'une simple co-occurrence. Si la proposition contient un ou plusieurs adverbes, ceux-ci sont tous interprétés à l'aide du même temps verbal attribué à cette proposition. Le même principe est également employé pour lier indices thématiques et expressions temporelles lors de l'indexation thématico-temporelle (voir chapitre 8).

Enfin, certains phénomènes syntaxiques particuliers apportent de précieux éléments d'information lors de l'interprétation des adverbes temporels. C'est par exemple le cas des expressions dites *cadra-*

[9] Celui-ci détermine la manière dont un événement est montré : dans sa globalité, en tant qu'accomplissement, en cours de réalisation, etc. Cette information est importante pour arriver à construire un ordonnancement temporel correct des données contenues dans le texte.

[10] Entre autres dues à son objectif, l'analyse de constats d'accidents de la route. Outre l'orientation du système vers l'analyse de ce type particulier de textes, la reconnaissance des circonstants temporels ne constitue pas la priorité du système et n'est donc pas très développée. De plus, l'analyse nécessite parfois, pour certaines phases, l'avis d'un utilisateur (système semi automatique).

tives (Charolles [1997], voir section 4.9). Il s'agit d'expressions, qui lorsqu'elles apparaissent dans certaines configurations syntaxiques, ont pour particularité de définir un cadre temporel pour la suite du discours[11]. À ce cadre est attribuée une valeur temporelle qui devient une clé pour l'interprétation des expressions temporelles qui viennent s'y placer. Par exemple, la mention d'une année en début de phrase (« En 2010 ») est cadrative. L'interprétation des expressions qui suivent tient alors compte de cette information : « décembre » est directement interprété comme « décembre 2010 ».

Plusieurs éléments n'interviennent pas dans le modèle temporel actuel, mais devraient probablement y être intégrés dans le futur. Au-delà d'une interprétation plus complète des temps verbaux et de la prise en compte de l'*aspect grammatical*, dont nous avons déjà parlé, la détection et le traitement adéquat du discours rapporté ainsi que celui des propositions relatives nous semblent les plus intéressants. Il reste évidemment encore divers phénomènes, dont certains sont probablement assez ardus à intégrer en pratique et dont il faudrait évaluer l'utilité réelle pour l'extraction d'informations. Citons entre autres la modalité et l'aspect lexical.

Notons que l'on se limite ici aux éléments qui ont une influence potentielle en ce qui concerne l'interprétation des expressions adverbiales temporelles. Comme nous l'avons signalé à la section 7.2.2, nous ne nous intéressons pas directement aux événements et à leur placement dans l'espace du temps. Les éléments qui interviennent dans ce type de processus, comme par exemple l'aspect[12], ne sont pas pris en compte ici, car ils sortent du cadre de ce travail[13].

7.3.2 Caractéristiques importantes de la modélisation temporelle

Le chapitre 5 a abordé la représentation et la modélisation du concept de temps. Dans le cadre de l'implémentation de ce système d'extraction d'informations temporelles, des choix ont été effectués en la matière. Tout d'abord, il faut préciser que le cadre général de cette modélisation temporelle est un calendrier (voir section 5.2.1). En l'occurrence, il s'agit plus précisément du calendrier grégorien. Celui-ci constitue la base de la conceptualisation du temps et a été choisi, d'une part pour son acceptation et son utilisation très large, et d'autre part parce qu'il est naturellement centré sur le niveau de granularité du jour (voir section 5.2.2), qui est adapté aux traitements que l'on envisage.

Plusieurs approches ont été proposées en ce qui concerne la modélisation d'une zone temporelle, principalement sous la forme d'un point, ou d'un intervalle (voir section 5.3). Cependant, ces deux notions sont fortement liées. Entre un point et un intervalle, il n'y a souvent guère plus qu'une question de granularité. Le choix opéré par rapport à ces deux représentations s'est par conséquent plutôt basé sur des critères pratiques. Ainsi, toute expression qui peut être représentée sans perte d'information, sous la forme d'un point à une certaine granularité, adopte effectivement ce format. Par contre les expressions qui font intervenir explicitement deux bornes sont pour leur part représentées à l'aide d'un couple de points. Par exemple « 2010 » est représenté sous la forme d'un point dont la granu-

[11] Ou plus précisément jusqu'à la fin du cadre, par exemple la fin du paragraphe.

[12] L'événement qui concerne le fait que Luc boit de la bière n'est pas temporellement équivalent dans « Luc a bu une bière le 20 janvier » (à un moment précis du 20 janvier) et « Luc a bu de nombreuses bières le 20 janvier » (potentiellement durant toute la journée).

[13] Au contraire de TimeML qui prévoit de prendre ces éléments en compte.

larité est l'année ([2010]) alors qu'il pourrait l'être de manière équivalente à l'aide de l'intervalle [01/01/2010 , 31/12/2010]. De même, l'expression « du 1er décembre 2010 au 31 décembre 2010 » correspondra l'intervalle [01/12/2010 , 31/12/2010] alors que la représentation au moyen d'un point à la granularité du mois ([12/2010]) est tout aussi valable. Notre mode de représentation du temps adopte donc à la fois les points et les intervalles.

Un autre choix réside dans la décision de différencier les expressions temporelles selon qu'elles sont déictiques, relatives au moment de l'énonciation, ou anaphoriques, relatives à un point de référence se situant dans le discours. Cette caractéristique n'a pas vraiment d'influence sur la représentation finale de la zone temporelle, mais bien sur le processus d'interprétation de l'expression qui désigne cette zone.

Vient ensuite la prise en compte du caractère flou, ou imprécis, de certaines expressions temporelles (voir section 5.3.4). Trois valeurs sont possibles pour cette caractéristique nommée *fuzzy*[14] : « 0 » lorsque l'expression ou la zone temporelle est précise, « 1 » pour exprimer une imprécision limitée à la zone définie (imprécision dite *interne*), et enfin « 2 » pour désigner de manière imprécise une zone qui inclus et s'étend autour de la zone temporelle délimitée (imprécision dite *externe*). Une seconde façon de caractériser de manière floue une zone temporelle est de désigner une partie de celle-ci (le début, le milieu ou la fin). L'utilisation de cette caractéristique implique automatiquement celle de l'indicateur d'imprécision.

La notion d'imprécision est importante dans le modèle. D'une part, elle permet de coder et de représenter des expressions naturellement floues, qu'il ne serait pas souhaitable de préciser. Et d'autre part, en utilisant l'indicateur d'imprécision à la manière d'un indice de certitude, cela permet de contrôler les éventuelles approximations de l'analyse automatique. Lorsque le système n'est pas certain de fournir un résultat tout à fait correct, l'accompagner d'une étiquette d'imprécision permet de gérer une certaine marge d'erreur.

Les trois points abordés jusqu'ici – modélisation sous la forme de points ou d'intervalles, caractère absolu ou relatif, et précision ou imprécision – correspondent à trois des quatre caractéristiques importantes qui ont été isolées pour caractériser les expressions temporelles. Cette catégorisation est exposée plus en détail à la section 7.6.1.

Le modèle que nous proposons intègre également une échelle de granularités temporelles importante (Figure 7.1). Une différence doit être faite entre le concept d'unité de mesure temporelle et celui de granularité. Les premières citées servent à mesurer des quantités de temps. Elles proviennent de l'observation de phénomènes naturels (voir section 3.2), font partie de systèmes normalisés de mesures (BIPM [2006]) et présentent des possibilités de conversions, d'une unité plus grande vers une unité plus petite, ou inversement. Les granularités (voir section 5.2.2) concernent plutôt les unités calendaires et servent à exprimer l'*ordre de grandeur* de la zone temporelle qu'elles occupent. Le

[14] L'utilisation du champ fuzzy, et de ses différentes valeurs, pour décrire l'aspect approximatif d'une expression temporelle peut être comparée à un opérateur de logique floue (Zadeh [1965], Hajek [2010]) en ce sens qu'il n'a pas une utilisation booléenne. En effet, ce champ peut prendre trois valeurs différentes qui représentent différents niveaux d'approximation plus ou moins élevés.

choix d'une nomenclature de granularités est en partie arbitraire[15]. Le passage d'une granularité à l'autre est naturellement possible lorsque la transformation s'opère d'un grain fin vers un grain plus important. Le mouvement inverse, s'il n'est pas impossible, est par contre plus délicat à réaliser et implique un certain degré d'imprécision à l'arrivée.

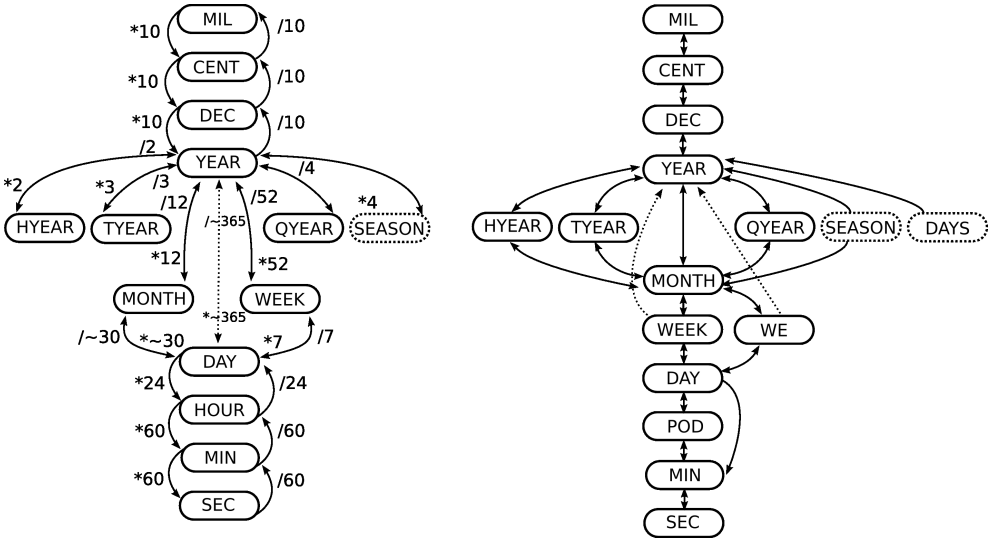


Figure 7.1 : Unités de mesure temporelles et niveaux de granularités.

Parmi la nomenclature de granularités, il en est une qui n'est pas toujours communément employée : la partie de journée (POD, pour « part of day »). Celle-ci constitue un degré supplémentaire entre le jour et l'heure. Elle constitue également une manière de rendre bien définies certaines références qui auraient été considérées comme floues autrement, tout en n'extrapolant pas des valeurs en termes d'heures précises. Cela permet donc de gérer un aspect assez fréquent de l'imprécision naturelle de la langue. L'information est conservée et peut être restituée sans perte à un utilisateur qui aura le loisir de l'interpréter dans le contexte adéquat[16], comme il l'aurait fait à la lecture de l'expression originale. Cela implique cependant une désynchronisation entre les unités de mesure temporelles et les niveaux de granularités. En effet, il ne semble par exemple pas heureux de considérer « après-midi » comme une unité de mesure, dans le sens où l'expression « dans deux après-midi » n'est pas très courante et possède une interprétation particulière qui n'est pas comparable à un usage normal, tel que « dans deux jours ».

La granularité de base choisie est le jour. En plus d'être naturellement adaptée aux systèmes de calendrier, comme nous l'avons expliqué à la section 5.2.2, cette granularité convient particulièrement bien à l'évocation des événements qui sont habituellement relatés dans des textes de la presse quotidienne. Certains de ces événements peuvent bien entendu être référencés au moyen d'autres granularités, plus ou moins fines. Le passage entre celles-ci relève alors simplement d'une question d'échelle. L'ana-

[15] La définition d'une nomenclature de granularités est en relation avec les unités temporelles, mais aussi avec le type d'expressions contenues dans les textes à analyser et avec la nature de la tâche à accomplir. Cela laisse la place à certains choix lors de l'implémentation d'un système qui est amené à manipuler le temps (ajout/suppression d'éléments par rapport aux unités temporelles.).

[16] La notion d'« après-midi » n'est pas nécessairement la même dans tous les pays ou toutes les cultures.

lyse de sources qui décrivent des événements dont la durée est très faible, par exemple de l'ordre de la micro-seconde, ou dont la mesure est très précise devra par contre être envisagée à un niveau beaucoup plus fin et nécessitera dès lors un autre choix de granularité.

Le modèle d'analyse temporelle utilise de manière conjointe les notions de précision/imprécision et de granularité afin de fournir un résultat le plus pertinent possible, tout en ayant soin de minimiser les erreurs d'interprétation. Le mécanisme qui articule l'imprécision et la granularité intervient principalement dans deux situations précises. Premièrement, il peut être nécessaire de faire varier la granularité lorsqu'une différence de granularité est observée entre différents éléments qui entrent en ligne de compte pour l'interprétation temporelle. Cela peut par exemple être le cas entre le point à interpréter (un jour, « la veille ») et le point de référence (une semaine, « la semaine dernière »), ou encore entre un point temporel et un déplacement temporel (« il y a une semaine »). Dans ces situations, l'utilisation de l'étiquette d'imprécision permet de réaliser normalement l'interprétation et, malgré l'approximation qui en résulte, de fournir un résultat cohérent avec le sens du texte.

Le passage vers un grain plus élevé est accompagné d'une imprécision de type 1 (interne), alors que la conversion en une granularité plus fine provoque l'attribution de l'étiquette de type 2 (externe). Par exemple :

> Augmentation de la granularité : avril 2010 –> 2010 + fuzzy=1.
> Diminution de la granularité : avril 2010 –> 15 avril 2010 + fuzzy=2.

Deuxièmement, lorsque l'interprétation temporelle ne mène pas à un résultat complètement certain, le système peut décider d'augmenter la granularité de la réponse et de l'accompagner d'une imprécision interne. Par exemple :

> Résultat incertain : 15 avril 2010 –> Réponse finale du système : avril 2010 + fuzzy=1.

7.3.3 Structure de données temporelles

Une structure de données a été définie afin de pouvoir encoder l'ensemble de ces caractéristiques. Une expression temporelle est enregistrée dans une structure de données qui, en plus de quelques champs relatifs au type et à la morphologie de cette expression, permet de décrire une zone temporelle. Celle-ci correspond à un point ou à un intervalle (Figure 7.2).

La représentation d'un expression temporelle ponctuelle (Figure 7.3) est principalement définie par sa structure de référence au calendrier (CALENDREF). Les expressions complètement spécifiées possèdent au moins une valeur pour chaque champ obligatoire[17]. Au contraire, pour les expressions sous-spécifiées, certains champs obligatoires sont vides. Cette structure est aussi accompagnée, pour les expressions relatives, d'une information au sujet du type de point de référence utilisé (déictique ou anaphorique). Celui-ci est repris dans la structure TEMPSHIFT, qui permet également d'exprimer un déplacement temporel. Quelques exemples :

[17] Pour chaque niveau de granularité utilisé par le système, un certain nombre de champs obligatoires ont été définis.

```
[cat]    => catégorie temporelle générale
[subcat] => sous-catégorie
[sig]    => « signature » (ensemble des étiquettes attribuées aux sous-constituants)
[ssig]   => « signature triée » (les étiquettes apparaissent par ordre alphabétique)

+  Structure expression ponctuelle (
        calendref,
        tempshift,
        [foctemp]
   )
OU Structure intervalle (
        fuzzy,
        partof,
        lower,
        upper
   )
```

Figure 7.2 : Structure de données globale (emballe une expression ponctuelle ou un intervalle).

– « le 24 décembre 2010 » remplit tous les champs obligatoires de CALENDREF, pour la granularité *jour* ;
– « le 24 décembre » remplit seulement certains champs obligatoires de CALENDREF, pour la granularité *jour*, et possède un type de référence codé dans TEMPSHIFT ;
– « le 24 décembre, il y a un an » remplit certains champs de CALENDREF, le type de point de référence ainsi que les autres champs de TEMPSHIFT, dédiés à la description des déplacements temporels ;
– « il y a un an » présente une structure CALENDREF vide, mais une structure TEMPSHIFT remplie des éléments spécifiant le déplacement temporel.

Enfin, lorsque l'expression propose la définition explicite d'un point de référence (« trois jours avant lundi »), celui-ci est repris, selon le même format qu'une référence de calendrier, sous le nom de *focus temporel* (FOCTEMP). Cette dernière structure est donc facultative.

```
[calendref] => description du point temporel (structure CALENDREF)
[tempshift] => description du type de référence et de déplacement temporel
  ( [ref]     => type de référence
    [move]    => direction du mouvement
    [nb]      => amplitude du mouvement
    [grain]   => unité temporelle
    [fuzzy]   => indicateur d'imprécision
    [precise] => indicateur de précision
  )
[foctemp]   => description d'un point de référence explicite (structure CALENDREF)
```

Figure 7.3 : Structure de données pour une expression temporelle ponctuelle.

Pour les intervalles, la structure de description d'un point temporel est dédoublée afin de pouvoir coder les bornes inférieure et supérieure. Quelques champs complémentaires sont également définis au niveau de l'intervalle pour pouvoir rendre compte des imprécisions qui s'appliquent à ce niveau (Figure 7.4).

Enfin, pour le point temporel lui même, il est décrit à l'aide de l'ensemble des champs détaillés à la figure 7.5. Ces champs correspondent à une référence vers un élément au sein d'un calendrier.

```
[partof] =>  partie de l'intervalle (début, milieu ou fin)
[fuzzy]  =>  indicateur d'imprécision de l'intervalle (0, 1 ou 2)
[lower]  =>  borne inférieure
             (structure complète, description d'une expression ponctuelle)
[upper]  =>  borne supérieure
             (structure complète, description d'une expression ponctuelle)
```

Figure 7.4 : Structure de données particulière pour les expression temporelles non ponctuelles (intervalles).

Comme déjà mentionné, pour chaque niveau de granularité utilisé par le système, un certain nombre de champs obligatoires ont été définis. Ces champs sont ceux qui permettent de localiser de manière univoque une zone temporelle dans l'espace du temps (le calendrier)[18].

```
[full]       => le contenu de la structure code une zone temporelle
                « bien/complètement identifiée », ou pas (0 ou 1)
[partof]     => partie d'une zone temporelle (début, milieu ou fin)
[fuzzy]      => étiquette d'imprécision (0, 1 ou 2)
[precise]    => présence d'un marqueur de précision (0 ou 1)
[grain]      => niveau de granularité (minute, ..., siècle)
[named]      => chaîne de caractère qui décrit une zone temporelle nommée
[hour]       => champ « heure » (0 à 23)
[min]        => champ « minute » (0 à 59)
[sec]        => champ « seconde » (0 à 59)
[pod]        => champ « partie de journée » (matin, ..., nuit)
[day]        => champ « jour » (dans un mois, de 1 à 31)
[dayinweek]  => champ « nom de jour » (lundi à dimanche)
[we]         => champ « numéro de week-end » (dans une année, de 1 à 52)
[week]       => champ « numéro de semaine » (dans une année, de 1 à 52)
[month]      => champ « mois » (1 à 12)
[season]     => champ «saison» (printemps, ..., hiver)
[qyear]      => champ « trimestre » (1 à 4)
[tyear]      => champ « quadrimestre » (1, 2 ou 3)
[hyear]      => champ « semestre » (1 ou 2)
[year]       => champ « année » (0 à 2***)
[dec]        => champ « décennie » (0 à 2**)
[cent]       => champ « siècle » (0 à 2*)
[mil]        => champ « millénaire » (0 à 2)
[defaultval] => valeur par défaut (chaîne de caractères)
```

Figure 7.5 : Structure de données pour un point temporel (référence à un calendrier, CALENDREF).

7.3.4 Choix de l'orientation principale pour l'implémentation.

À la section 7.3.1, nous avons signalé l'orientation du système vers les expressions adverbiales. Celles-ci constituent, d'une part, un moyen très fréquent de véhiculer une information temporelle dans un texte, et d'autre part, une information dont l'analyse et l'exploitation se trouvent à la portée des méthodes de traitement automatique du langage.

Une décision importante avant l'implémentation est, étant donné la nature des informations temporelles que nous avons décidé d'utiliser, le choix d'une méthode de repérage et d'extraction. Celui-ci s'est orienté vers une méthode symbolique, mieux à même de décrire le plus complètement et précisément possible ce type d'expression. La méthodologie employée, qui est décrite à la section 7.4, permet de produire ce type de description et de l'encoder dans une ressource d'extraction composée

[18] Pour un jour, les champs *day*, *month* et *year* sont obligatoires. Les autres, par exemple *dayinweek*, sont facultatifs.

de patrons lexico-syntaxiques. Ceux-ci sont créés à la main et implémentés à l'aide du logiciel Unitex (Paumier [2003], Paumier [2008]) sous la forme de grammaires locales et de transducteurs (Gross [1989], Gross [1997], et voir aussi à la section 2.3.8).

Ces transducteurs ont pour objectif d'annoter temporellement le texte. Les expressions temporelles sont donc repérées, délimitées, segmentées en sous-constituants et se voient finalement attribuer un ensemble d'étiquettes qui les caractérisent.

En pratique, d'autres décisions ont dû être posées pour implémenter le modèle tel que défini ci-dessus. Celles-ci sont détaillées aux sections 7.7 et 7.8.

7.4 Méthodologie

Le système d'extraction et d'interprétation des informations temporelles que nous proposons repose avant tout sur la reconnaissance des expressions temporelles. Pour atteindre un résultat valable, il est nécessaire de construire de nombreuses et volumineuses ressources linguistiques. Ce type de développement n'est pas simple à mener car il implique la gestion de ces ressources au fil du temps et d'inévitables mises à jour. Il nous a donc semblé judicieux de définir une procédure pour encadrer le processus d'extraction, en ce compris la création de la ressource linguistique.

La méthodologie proposée, dont les points importants sont représentés à la figure 7.6, touche aussi bien la partie qui s'occupe de définir et décrire les expressions temporelles et leurs annotations, que la partie consacrée à l'interprétation de ces éléments. Elle s'apparente à un cycle de vie qui permet de créer un système et de le faire évoluer au cours du temps.

1. La première approche doit permettre de se faire une idée assez précise des expressions à annoter dans les textes. Pour ce faire, il est possible de consulter des ouvrages de référence (par exemple Gross [1986] dans le cas des adverbes temporels), ou d'examiner un corpus de test. Une première génération de grammaires peut être construite afin de se faire une idée des possibilités de formalisation des expressions recherchées à l'aide de graphes. Ces premiers développements n'ont pas pour but d'atteindre une large couverture, mais bien d'identifier certains critères importants, tant au niveau de la morphologie que du sens, qui pourront servir à organiser les unités à extraire. Suite à cette première analyse, une définition des catégories est réalisée.
2. Une fois les différentes catégories définies, une spécification plus détaillée des types d'expressions contenus dans chaque catégorie peut être effectuée. Celle-ci s'attache à décrire les éléments constitutifs de ces expressions et les accompagne d'exemples.
3. Cette spécification constitue la référence qui doit être utilisée pour construire les grammaires d'extraction. Celles-ci transposeront la spécification vers un format qui pourra être appliqué aux textes. Il est important que la transposition soit la plus fidèle possible, c'est-à-dire que tous les cas prévus par la spécification soient couverts et que, dans le même temps, le bruit[19] éventuellement apporté par les grammaires soit minimisé.

[19] Des séquences imprévues, pertinentes ou non.

4. Les exemples fournis dans la spécification servent ensuite de mini-corpus de test pour vérifier que les grammaires atteignent bien la couverture espérée. Les grammaires sont d'abord validées sur ce corpus *minimal*, avant d'être testées sur des textes réels.
5. Un certain nombre d'itérations des étapes 2 à 4 peut être effectué dans le but d'augmenter la couverture des grammaires. Ce processus de construction de grammaire a été proposé à l'origine par Gross [1999] et est nommé *bootstrap method*.
6. L'application des grammaires sur un corpus peut fournir des statistiques concernant la fréquence et la distribution des différents types d'expression (voir Annexe D). Au-delà de l'intérêt scientifique que peuvent avoir de telles constatations, elles peuvent également permettre d'axer en priorité l'implémentation d'une application (étape 7) vers les cas les plus fréquent.
7. Le résultat de l'extraction doit ensuite être exploité. À ce stade, la spécification sert à nouveau de base afin de déterminer avec précision quels sont les cas qui doivent être pris en charge par le logiciel. Cette étape est également l'occasion de vérifier l'exactitude de l'annotation réalisée par les grammaires. Le programme peut en effet rapporter facilement tout cas sortant de la *norme* définie par la spécification.

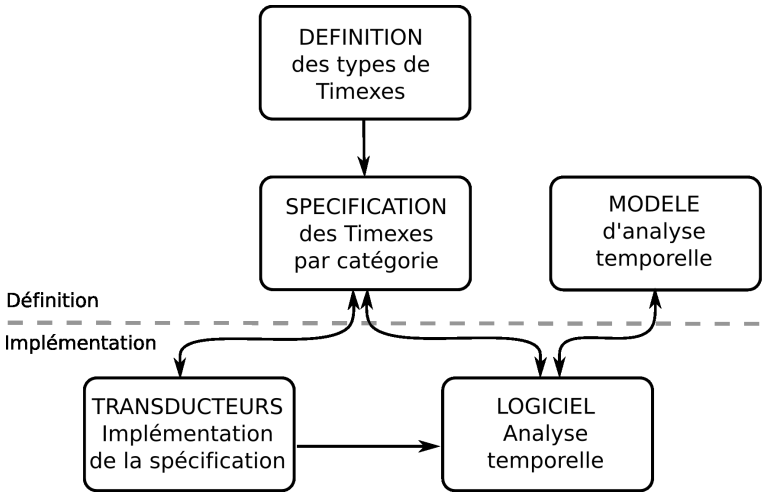


Figure 7.6 : Vue globale des étapes de développement.

La spécification occupe donc une position centrale au cours du développement du système, ainsi que lors de son évolution. Un des éléments qui matérialisent le lien entre la spécification et les différentes autres parties est l'utilisation d'identifiants pour marquer les sous-catégories d'expressions. Ces identifiants, définis dans la spécification, se retrouvent dans les sorties des grammaires et donc également dans les annotations réalisées sur les textes par ces dernières. Cette pratique permet de réaliser des statistiques au niveau des sous-catégories, mais surtout de tracer les éventuels oublis ou erreurs. En effet, toute anomalie détectée lors des tests pourra facilement être localisée au niveau des grammaires, et faire l'objet d'une correction qui sera finalement répercutée dans la spécification.

Le travail de spécification qui a été effectué se trouve à la croisée des chemins entre certains travaux de linguistique, tels que ceux de Borillo (voir section 4.2), et les guides d'annotation tels que proposés en extraction d'informations (voir section 6.5). La spécification ne doit cependant pas être

vue comme une tentative de créer un document normatif. Elle s'impose plutôt comme une nécessité d'accompagner l'écriture des grammaires d'un guide, d'un référentiel, permettant de ne pas se perdre en cours de route. L'utilisation d'un format bien défini, et la spécification des expressions à l'aide de celui-ci, favorise le développement structuré de la ressource ainsi que la production d'une annotation cohérente. Elle contribue également à rendre compte de l'évolution de la couverture de l'extraction, et constitue une information de base pour l'étape d'interprétation[20]. À notre connaissance, en ce qui concerne l'extraction d'expressions temporelles, ce type de démarche ainsi que la spécification qui en résulte, n'ont pas fait l'objet de publications.

En ce qui concerne la création et l'évolution de la ressource linguistique principale – les grammaires – quelques remarques peuvent être émises. Tout d'abord, dès que le nombre de grammaires augmente, il est important d'adopter une certaine discipline en matière de nommage, et d'organisation en répertoires. Cette rigueur facilite l'accès à certains composants déjà développés et qui peuvent être réutilisés dans différentes grammaires. Cela permet aussi une meilleure vision des ressources disponibles lorsqu'il s'agit de modifier ou corriger tel ou tel élément. Pour faciliter leur gestion et leur partage, les grammaires locales peuvent être organisées à l'aide d'outils spécifiques, tels que celui proposé par Constant [2003].

Enfin, quelques outils pourraient être utiles lors de la phase de mise au point des grammaires avec Unitex. Par exemple, il arrive qu'une séquence puisse malencontreusement être reconnue par plusieurs *chemins* de la grammaire, qui lui attribuent ainsi plusieurs annotations potentielles. Lors de l'insertion des séquences de sortie dans le texte[21], une seule annotation sera retenue, sans garantie que ce soit la plus approprié. Le même type de problème se pose lorsque deux séquences reconnues se chevauchent. L'heuristique généralement utilisée, celle du plus long chemin, décidera alors de sacrifier une séquence au profit de l'autre. Ces deux comportements ne constituent pas un problème en eux-mêmes, au contraire ils augmentent la robustesse du processus. Cependant, les cas de recouvrement total ou de chevauchement étant souvent révélateurs d'une erreur de conception des grammaires, ils pourraient être avantageusement détectés et signalés au développeur.

D'une manière générale, lorsque les projets d'élaboration de ressources dépassent une certaine taille, il serait certainement intéressant de disposer d'un véritable environnement de développement, tel qu'on en voit dans le domaine de la programmation de logiciels[22]. Ce type d'outil permet une véritable gestion du projet : regroupement des fichiers concernés, intégration de librairies, gestion des dépendances entre grammaires, renommage à l'échelle du projet (*refactoring*), gestion de la documentation, gestion des versions, etc. Tous ces éléments pourraient être utilement accessibles au travers d'une seule et unique interface.

[20] Permet de savoir quel type d'information se retrouvera dans quel type d'expression, et de quelle manière cela sera annoté.

[21] Ce problème ne se présente pas lors de l'affichage d'une concordance classique.

[22] L'environnement Eclipse (`http://www.eclipse.org`) en est un exemple parmi d'autres.

7.5 Remarque concernant l'exhaustivité

En ce qui concerne la reconnaissance et l'interprétation des expressions temporelles, nous ne prétendons pas viser l'exhaustivité, mais bien atteindre une couverture significative. Le but est de fournir une analyse pouvant servir dans une application de traitement automatique du langage. L'approche est donc pragmatique en ce sens que nous avons pour objectif de couvrir (au moins) les expressions les plus fréquentes, et surtout, celles qui sont exploitables dans le contexte d'une analyse automatique et donc utiles dans un cadre applicatif. Les expressions temporelles visées sont caractérisées et décrites à la section 7.6.

Les grammaires d'extraction constituent une ressource qui est par nature incomplète. Elle est donc appelée à être enrichie graduellement. Même si la spécification de l'extraction permet de décrire de manière systématique le contenu des grammaires, elle ne reprend pas nécessairement la totalité des expressions potentiellement reconnues par celles-ci. En effet, le phénomène de factorisation, qui constitue un avantage offert par la formalisation de motifs lexico-syntaxiques sous la forme de graphes, induit également un certain risque de bruit (reconnaissance de séquences incorrectes ou imprévues). L'objectif de la spécification n'est donc pas nécessairement de coller complètement à la réalité de l'extraction, mais bien de se concentrer avant tout sur les cas qui seront interprétés lors de l'analyse temporelle. Évidemment, il est préférable que l'ensemble des cas soient décrits et que les graphes développés minimisent les risques de bruit, rendant ainsi les deux ressources aussi proches que possible.

7.6 Définition et spécification des expressions temporelles à extraire

7.6.1 Catégorisation des expressions temporelles

Les expressions temporelles sont regroupées dans une classification que nous avons établie dans le cadre de ce travail. Les différentes catégories sont caractérisées par quatre critères binaires :

1. Ponctuel ou Duratif (P ou D) ;
2. Absolu ou Relatif (A ou R) ;
3. Précis ou Flou (P ou F) ;
4. Unique ou Répétitif (U ou R).

Le **critère 1** (Ponctuel ou Duratif) permet de distinguer les expressions dont la représentation temporelle correspond, à une certaine granularité donnée, à un point ou au contraire à un intervalle. L'expression *ponctuelle* désigne toujours une zone temporelle comme un tout indissociable et est exprimée par une référence unique à l'espace du temps. Une expression *durative* est obligatoirement exprimée à l'aide de deux bornes. Dans certains cas, l'une de celle-ci peut cependant ne pas être explicitement décrite.

P : « lundi 20 septembre 2010 »

P : « en 2010 »
P : « au vingtième siècle »
D : « du lundi au jeudi »
D : « jusqu'au mois de septembre »

Le **critère 2** (Absolu ou Relatif) classe les expressions selon que leur localisation sur la ligne du temps est directe, indépendante de tout autre point dans l'espace du temps (localisation *absolue*) ou qu'elle nécessite l'utilisation d'un repère temporel (localisation *relative*). Une expression temporelle est donc absolue si elle contient en son sein tous les éléments nécessaires à son identification univoque dans un calendrier. La nature et le nombre de ces éléments varient en fonction de la granularité de l'expression. Pour être situées sans ambiguïté dans un calendrier, les expressions relatives nécessitent par contre la connaissance d'un point de repère. Celui-ci peut être le moment d'énonciation (référence déictique) ou un autre point temporel contextuel, explicitement renseigné ou non (référence anaphorique).

A : « le 20 septembre 2010 »
A : « 2005 »
R : « lundi »
R : « la veille »

Le **troisième critère** (Précis ou Flou) indique si la zone temporelle décrite par l'expression peut être délimitée de manière précise ou si, au contraire, sa délimitation manque de précision. Une expression est qualifiée de *précise* si on considère, quelle que soit sa granularité, qu'elle couvre l'entièreté de la zone qu'elle désigne, à l'exclusion de tout le reste.

P : « le jeudi 23 septembre 2010 » ;
P : « à 14h00 » ;
P : « le 20ème siècle ».

Une expression *floue* étend la couverture, de manière proche, mais indéterminée, autour de la zone désignée ou, au contraire, la restreint à une fraction de celle-ci. Le caractère flou peut donc être obtenu de plusieurs façons[23] :

F : à l'intérieur d'une plage temporelle de granularité élevée, « durant le mois de décembre ») ;
F : aux alentours d'un repère temporel précis, « vers le 22 décembre 2009 » ;
F : spécification relative imprécise, « dans environ trois semaines ».

Le caractère flou d'une expression est également en partie lié à la notion de granularité (voir section 5.2.2). Le choix de la granularité de base, le jour en ce qui nous concerne (voir section 7.3.2), a une influence sur ce qui est considéré comme précis ou flou. Ainsi, toute expression dont la granularité est inférieure ou égale au jour est interprétée comme étant précise, à moins qu'elle soit accompagnée d'une marque d'imprécision. À l'inverse, toute expression de granularité supérieure au

[23] Notons que l'utilisation d'une granularité trop élevée pour décrire temporellement un événement, crée également une imprécision. Ce phénomène est cependant beaucoup plus difficile à repérer et à traiter automatiquement.

jour est considérée comme floue (« au 20e siècle », « en décembre »), à moins qu'une marque de précision soit explicitement présente (« le 20e siècle », « tout le mois de décembre »).

Les expressions floues ne doivent pas être confondues avec les expressions précises mais sous-spécifiées[24], telle que « jeudi ». L'évaluation du critère de précision s'opère au niveau du sens de l'expression, et non de la morphologie de celle-ci.

Enfin, le **dernier critère** sépare les expressions donnant lieu à une représentation unique (U) ou répétée (R) sur la ligne du temps.

U : « mardi matin » ;
R : « chaque dimanche » ;
R : « durant trois vendredis ».

Ces quatre critères binaires peuvent être croisés, ce qui produit une nomenclature de seize catégories. Dans cette thèse, nous avons décidé de laisser de côté les entités de type répétitives. Celles-ci nous ont paru, dans une certaine mesure, moins intéressantes par rapport à notre approche car elles désignent assez souvent des horaires plutôt que des points d'ancrage précis dans l'espace du temps. Ce type d'entités a d'ailleurs été traité en tant qu'horaires, et à l'aide de techniques similaires, par Weiser [2010], dans le cadre de l'extraction d'informations touristiques. L'intégration de ces expressions reste malgré tout un point important pour de futurs développements. Les différentes catégories actuellement traitées sont donc au nombre de huit.

Les codes qui sont attribués à ces catégories sont composés de quatre caractères. Le premier renseigne la valeur relative au premier critère, c'est à dire l'aspect ponctuel ou duratif, et peut donc être constitué par la lettre « P » ou la lettre « D ». De même, le deuxième caractère code le critère 2 (la localisation absolue ou relative, « A » ou « R »), le troisième le critère 3 (l'aspect précis ou flou, « P » ou « F ») et enfin le dernier le critère 4 (l'unicité ou la répétition, « U » ou « R »). Les huit catégories sont donc désignées par les codes suivants : PAPU, PAFU, PRPU, PRFU, DAPU, DAFU, DRPU et DRFU.

Nous y ajoutons quatre autres catégories un peu particulières, qui ne sont pas des expressions qui doivent être interprétées, mais dont la reconnaissance permet d'éviter les problèmes d'ambiguïtés :

- les durées ;
- les âges ;
- les durées imprécises ;
- les âges imprécis.

Cette catégorisation se rapproche, dans l'esprit, de celles proposées par Borillo [1983, 1988]. Elle apporte cependant quelques différences et extensions qui permettent d'enrichir ou d'affiner certains aspects. Une correspondance peut être partiellement établie entre ces catégories et celles qui ont été définies dans cette section. Le tableau 7.1 en reprend une synthèse.

[24] C'est-à-dire spécifiées de manière incomplète, sans tous les éléments nécessaires à leur localisation directe et absolue dans le calendrier. Cette caractéristique dénote plutôt, comme expliqué au critère 2, une expression relative.

Ainsi, pour la première distinction opérée par Borillo [1983], les adverbes *autonomes* sont à mettre en relation avec les expressions *absolues* (*A**) alors que les adverbes *déictiques* et *anaphoriques* se rapportent aux expressions *relatives*, avec référence au moment de l'énonciation ou à un autre moment déterminé (*R**, $ref = (now|focus)$). Quant aux adverbes *polyvalents*, ils rentrent également dans cette catégorie, mais sont identifiés, selon l'adverbe en question, comme étant soit déictiques, soit anaphoriques[25].

La seconde distinction réalisée par Borillo [1983] identifiait des adverbes *ponctuels*. Ceux-ci correspondent aux expressions du même nom (P***). Les adverbes *inclusifs* rentrent dans la catégorie des expressions *floues* (avec une imprécision interne, de type 1). Et enfin, les adverbes *duratifs* sont liés aux expressions du même nom, mais sans inclure les intervalles à deux bornes. Ceux-ci sont cependant abordés sous le nom de *durée-limite* dans Borillo [1988].

ADVERBES	Ponctuels	Inclusifs	Duratifs
Autonomes	PAPU	PAFU$^{fuzzy=1}$	DAPU
Déictiques	PRPU$_{ref=now}$	PRFU$^{fuzzy=1}_{ref=now}$	DRPU$_{ref=now}$
Anaphoriques	PRPU$_{ref=focus}$	PRFU$^{fuzzy=1}_{ref=focus}$	DRPU$_{ref=focus}$
Polyvalents	PRPU$_{ref=*}$	PRFU$^{fuzzy=1}_{ref=*}$	DRPU$_{ref=*}$

Tableau 7.1 : Comparaison de la catégorisation des expressions temporelles avec Borillo [1983, 1988].

La catégorisation implémentée dans cette section apporte donc deux catégories additionnelles qui ne sont pas reprises par Borillo : DAFU et DRFU, qui correspondent aux expressions qui sont à la fois duratives et imprécises (duratif et inclusif selon la terminologie de Borillo). En ce qui concerne l'imprécision d'une manière générale, le système de catégories proposé ici permet une plus grande finesse. Il est en effet possible de préciser si le caractère imprécis est interne ou externe à la zone temporelle, alors que seule la première possibilité est présente chez Borillo. Enfin, même si elles ne sont pas implémentées dans cette thèse, une place est prévue pour les expressions répétitives (***R : fréquences, horaires, etc.) alors qu'il n'en est pas fait état dans la catégorisation à laquelle on se réfère.

D'autre part, une autre remarque peut être faite concernant les adverbes ponctuels. À leur propos, Borillo [1983] dit :

> « Ils fixent sur l'axe temporel un repère assimilable à un point [...] Sur la base d'une relation d'inclusion ils peuvent être combinés avec des adverbes désignant un intervalle supérieur, l'intervalle maximum restant la journée. » (p. 112)

[25] Dans un contexte d'extraction d'informations, il est nécessaire d'attribuer une catégorie non ambiguë afin de pouvoir mener la suite du traitement. Une procédure de désambiguïsation dédiée à ce problème livrerait évidemment une analyse plus nuancée et probablement plus correcte, mais celle-ci n'a pu être abordée dans le cadre de cette thèse.

Cela semble signifier que les adverbes ponctuels se situent obligatoirement à un niveau inférieur à la journée. Or, par l'intermédiaire du jeu des granularités, il est permis de considérer une zone temporelle comme un point, quel que soit le niveau auquel on se place[26]. Quant aux inclusions, si elles semblent plus naturelles pour les granularités inférieures au jour, il n'est pas impossible d'en retrouver à des niveaux plus élevés (« lundi de la semaine dernière »). L'option prise avec la catégorisation proposée ici est de rester le plus souple et général possible, et par conséquent de ne pas introduire ce genre de restrictions.

7.6.2 Spécification des expressions temporelles

Introduction

La catégorisation présentée à la section 7.6.1 n'est que la première étape dans le processus de définition de l'extraction des expressions temporelles (voir section 7.4). Son prolongement consiste à détailler plus profondément ce que chaque catégorie recouvre. Cet objectif prend la forme d'une spécification détaillée des expressions temporelles. Celle-ci, organisée en catégories et sous-catégories, donne pour chaque type d'expression la caractérisation des sous-constituants qui la composent. Ainsi, à toute expression est attachée une séquence d'étiquettes qui correspond à son annotation interne, c'est-à-dire la suite des étiquettes qui ont été attribuées à ses sous-constituants, et qui est nommée *signature*[27].

La spécification des expressions temporelles, dont le rôle a déjà été abordé à la section 7.4, se présente sous une forme assez technique. Il s'agit avant tout d'une référence qui définit de manière détaillée l'étiquetage qui doit être attribué aux expressions, et à ce titre, elle est utilisée à la fois lors du développement des grammaires d'extraction et lors de l'implémentation du logiciel d'interprétation. Malgré le côté un peu moins lisible inhérent à ce genre de ressource, celle-ci constitue néanmoins la meilleure source pour décrire précisément ce que recouvre chaque catégorie d'expressions temporelles. Cette section ne livre pas la spécification en tant que telle (voir l'annexe B pour consulter celle-ci), mais propose plutôt une description commentée, plus *accessible*. La présentation veille cependant à être suffisamment détaillée afin de définir de manière précise les différents types d'expressions temporelles concernées par l'extraction.

D'une manière générale, les expressions les plus *simples* sont présentées avant les plus *complexes*. Ces dernières intègrent souvent un ou plusieurs éléments dont l'exposé complet a déjà été effectué dans un premier temps. Dès lors, ces éléments ne seront pas nécessairement redéveloppés dans le cadre de ces expressions afin de ne pas compliquer outre mesure leur présentation[28].

[26] Dans le sens où un siècle peut tout aussi bien être représenté par un point qu'une journée.

[27] Par exemple, la signature d'une date suivie d'une heure pourrait être, de manière simplifiée, DATE-HEURE.

[28] Cela correspond d'ailleurs à la réalité de l'implémentation puisque, par exemple, dans le cas des expressions duratives (D***), les bornes peuvent être interprétées comme des expressions simples (P***).

Conventions de notation

Une série de notations sont utilisées de manière à rendre l'exposé le plus lisible et compréhensible possible. Pour une définition plus complète et systématique des annotations utilisées dans les grammaires, il est possible de consulter l'annexe B.

Chaque élément porteur d'information temporelle est annoté et possède par conséquent une étiquette particulière. Il existe des *éléments de base* – par exemple un nom de jour, une année ou une heure – et des *groupes* plus complexes qui peuvent accueillir un ou plusieurs de ceux-ci. Ces groupes permettent, entre autres, de définir une date ou une heure. Ces derniers sont notés :

$$DATE\ (\ )\quad HEURE\ (\ )$$

Leur contenu n'est jamais détaillé dans la *signature* qui résume chaque cas, mais il l'est dans la spécification détaillée (Annexe B). Des exemples viennent cependant illustrer les différentes configurations possibles. Un élément fait exception au masquage de la composition interne des groupes. Il s'agit de celui qui renseigne sur la granularité du groupe et qui est noté :

> $G_{(*)}$, avec $*$ qui représente n'importe quel niveau de granularité (ceux-ci sont repris à la section 7.3.2).

L'élément de granularité peut aussi parfois apparaître seul, en dehors de tout groupe. Une notation spéciale est prévue pour $G_{(pod\_*)}$, qui équivaut à l'ensemble des niveaux de granularité commençant par « pod », c'est-à-dire une partie de journée (*part of day*).

D'autres éléments peuvent intervenir en dehors des groupes $DATE$ et $HEURE$. Il s'agit des marques d'imprécision et de partition qui sont notées :

$$Fuzzy_{(*)}\quad PartOf_{(*)}$$

À nouveau, $*$ désigne toute valeur qui peut être prise par cet élément. En ce qui concerne $Fuzzy_{(*)}$, les valeurs possibles sont « 1 » et « 2 » pour signaler les imprécisions internes et externes, alors que pour $PartOf_{(*)}$ il s'agit de « BEG », « MID » ou « END » pour désigner le début, le milieu ou la fin d'une zone temporelle (voir section 7.3.2).

Enfin, trois groupes particuliers interviennent dans la spécification des expressions relatives. Ils permettent respectivement la caractérisation de la cible, c'est-à-dire la zone temporelle que l'expression désigne, d'un point de référence temporel (focus temporel) ou encore d'un déplacement temporel :

$$TARGET_{(*)}\ (\ )\quad FOCTEMP_{(*)}\ (\ )\quad TEMPSHIFT\ (\ )$$

Les caractères $*$ attribués en indices à $TARGET$ et à $FOCTEMP$ sont soit $part$, soit $full$, c'est à dire que le groupe désigne de manière complète ou partielle une zone temporelle[29]. Ces deux groupes peuvent donc contenir les différents éléments déjà exposés. Cependant, pour des raisons de lisibilité des *signatures*, seule la granularité apparaîtra dans ces groupes.

[29] $part$ peut être assimilé à la catégorie PRPU, alors que $full$ est à mettre en relation avec PAPU.

Le contenu de $TEMPSHIFT$, qui code un déplacement temporel, est lui un peu plus particulier et sera détaillé. L'élément indispensable est $Ref_{(*)}$ qui indique quel est le point de référence à considérer : le moment de l'énonciation ($Ref_{(now)}$) ou un autre point donné par le contexte ($Ref_{(focus)}$). Souvent, cet élément sera complété par l'indication de direction vers l'avant ($Move_{(fwd)}$) ou vers l'arrière ($Move_{(rew)}$), une granularité $G_{(*)}$, ainsi que par une amplitude ($Amp_{(*)}$) qui provient d'un élément lexical et a été *traduit* en une valeur numérique. Ces deux derniers éléments peuvent provenir d'un adverbe qui les code implicitement (« la veille » équivaut à un déplacement temporel de un jour dans le passé), mais sont parfois également exprimés par une véritable expression numérique comprenant une quantité accompagnée d'une unité de mesure temporelle (« il y a 10 ans »). Ce cas est noté $NbTempUnit_{(*,*)}$. La présence d'un double indice symbolise les éléments d'amplitude et de granularité sous-jacents.

Enfin, le dernier groupe particulier intervient dans le cadre des expressions duratives et sert à regrouper les éléments qui décrivent une borne :

$$BOUND_{(*)}\ (\ )$$

Son indice peut prendre la valeur $lower$ ou $upper$ selon qu'il s'agit de la borne inférieure ou supérieure de l'intervalle. Étant donné que ces bornes sont apparentées à des expressions temporelles bien identifiées ou à des expressions relatives, ces éléments seront utilisés pour décrire leur formation. Ils seront simplement notés selon leur catégorie, soit respectivement :

$$PAPU\ (\ ) \quad PRPU\ (\ )$$

Quelques éléments de notation interviennent encore. Tout d'abord, comme pour des expressions régulières, les cardinalités permettent de préciser le nombre d'occurrences minimal et maximal que peut atteindre l'élément visé. La notation adopte le format (min, max) placé en exposant (par exemple $G^{(0,1)}_{(pod\_*)}$). Ensuite, les parenthèses et le signe « | » permettent de formuler un *ou* logique entre différents éléments. Dans les listes d'exemples, les crochets (« [ ] ») pourront être employés pour exprimer l'aspect facultatif de ce qu'ils encadrent. Enfin, les différents cas à l'intérieur d'une catégorie sont identifiés par un numéro d'*ID*. Celui-ci constitue le lien avec la spécification complète, ainsi qu'avec les grammaires, dans lesquelles cette référence apparaît également.

PAPU : Référence Ponctuelle, Absolue, Précise et Unique

P - Utilise le point, envisagé à une certaine granularité, comme mode de représentation
A - Peut être située directement sans faire appel à un repère temporel
P - Localisation et délimitation précise et bien déterminée de la plage temporelle concernée
U - N'est composée que d'une seule plage temporelle

Les expressions temporelles de type PAPU sont composées à l'aide des différents éléments qui permettent de définir une zone temporelle bien identifiée, et cela à un certain niveau de granularité. En plus des différentes formes de dates, les granularités inférieures au jour – partie de journée (G_{pod*}) et heure précise ($HEURE$) – sont aussi prises en compte.

Spécification générale :

$$DATE \left(G_{(*)} \right) G^{(0,1)}_{(pod\_*)} HEURE^{(0,1)} \left(G_{(min)} \right)$$

Les motifs reconnus par $DATE$, lorsque la granularité est supérieure ou égale à l'année sont :

- une année (ID=1) : « 2005 » ;
- une décennie (ID=2) : « les années 30 » ;
- un siècle (ID=3) : « le 20e siècle » ;

Lorsque $DATE$ est de granularité inférieure à l'année (ID=4 à 8) :

- un mois et une année : « mars 2007 » ;
- une semaine ou un week-end : « le week-end des 12 et 13 juin 1999 » ;
- un jour, un mois et une année : « le 12 juillet 2007 ».
- Un motif qui désigne un jour peut aussi être accompagné d'une indication de granularité plus fine, relative à une partie de la journée et/ou d'une heure :
  a. $HEURE$: « le 12 juillet 2007 à 13h30 » ;
  b. $G_{(pod\_*)}$: « le 12 juillet 2007 au matin » ;
  c. $G_{(pod\_*)}\ HEURE$: « le 12 juillet 2007 après midi, à 16h00 ».
- Une date (complète) peut être associée à un déplacement temporel. L'utilité de ce dernier est d'insister sur le temps qui s'est écoulé (ou qui va s'écouler) entre cette date et l'instant de l'énonciation. Comme la date est bien identifiée, il ne sera pas nécessaire d'interpréter cet élément supplémentaire : « il y a deux jours, le 12 juillet 2007 ».

Enfin, rentrent également dans la catégorie PAPU, les zones temporelles désignées à l'aide d'une *période nommée* (nom de fête, de vacances, etc.), d'un nom de saison (météorologique) ou d'une autre période de l'année (ID=9) :

« Noël 2005 »,
« les grandes vacances 2006 »,
« l'été 2007 »,
« le carnaval 2008 »,
« le premier trimestre 2009 ».

PAFU : Référence Ponctuelle, Absolue, Floue et Unique

P - Utilise le point, envisagé à une certaine granularité, comme mode de représentation
A - Peut être située directement sans faire appel à un repère temporel
F - Localisation et/ou délimitation floue ou imprécise de la plage temporelle concernée
U - N'est composée que d'une seule plage temporelle

D'une manière générale, les expressions de type PAFU constituent le pendant imprécis de la catégorie PAPU. L'imprécision peut être obtenue à l'aide d'un marqueur d'approximation ($Fuzzy_{(*)}$)

ou de partition ($PartOf_{(*)}$), ou encore à partir d'une combinaison de ces éléments. Ceux-ci sont associés à une expression temporelle qui désignent une zone temporelle bien identifiée, quelle que soit sa granularité. Le cas où l'approximation est réalisée sur la base d'une granularité inférieure au *jour* (spécification générale A) est distingué de celui faisant intervenir tout autre point temporel de granularité supérieure (spécification générale B).

Spécification générale A : granularité inférieure au *jour*

Deux *signatures* sont reliées à ce premier cas.

(1) L'imprécision se situe au niveau d'une heure ($G_{(min)}$) ou d'une partie de la journée ($G_{(pod\_*)}$).

$$DATE\ \left(G_{(day)}\right)\ G^{(0,1)}_{(pod\_*)}\ Fuzzy_{(*)}\ HEURE\ \left(G_{(min)}\right)$$

Une heure imprécise, mais située dans une zone temporelle bien identifiée peut être obtenue à l'aide de la mention d'une heure approximative en complément :

- d'une indication de journée (ID=1) : « 22/12/2009 vers 18h24 » ;
- de partie de journée (ID=2) : « 22 décembre 2009 matin vers 10h24 ».

(2) L'imprécision se situe au niveau d'une partie de la journée ($G_{(pod\_*)}$).

$$DATE\ \left(G_{(day)}\right)\ (\ Fuzzy_{(*)}\ |\ PartOf_{(*)}\ )^{(1,N)}\ G_{(pod\_*)}\ (\ Fuzzy_{(*)}\ HEURE\ \left(G_{(min)}\right)\ )^{(0,1)}$$

L'imprécision peut cependant également s'établir à la granularité intermédiaire de la partie de journée ($G_{(pod\_*)}$). Les marqueurs d'imprécision et de partition font alors leur apparition, seuls ou en combinaison :

- $Fuzzy_{(*)}$ (ID=3) : « mardi 22 décembre 2009 durant l'après-midi » ;
- $PartOf_{(*)}$ (ID=4) : « 22 décembre 2009, fin de matinée » ;
- $Fuzzy_{(*)}\ PartOf_{(*)}$ (ID=5) : « vingt-deux décembre 2009 en fin d'après-midi ».

Dans ces trois derniers cas, la mention supplémentaire d'une heure approximative est possible, mais pas obligatoire (ID=6, 7 ou 8).

Spécification générale B : granularité supérieure ou égale au *jour*

$$(\ Fuzzy_{(*)}\ |\ PartOf_{(*)}\ )^{(1,N)}\ DATE\ \left(G_{(*>=day)}\right)$$

Dans le cas d'une approximation sur une date d'une granularité supérieure ou égale au jour (mois, année, décennie ou siècle), les marqueurs d'imprécision et de partition interviennent à nouveau :

- $Fuzzy_{(*)}$ (ID=9) : « aux environs de l'an 2000 » ;
- $PartOf_{(*)}$ (ID=10) : « au début des années 30 » ;
- $Fuzzy_{(*)}\ PartOf_{(*)}$ (ID=11) : « vers la fin de l'été deux-mille-quatre ».

PRPU : Référence Ponctuelle, Relative, Précise et Unique

P - Utilise le point, envisagé à une certaine granularité, comme mode de représentation
R - Un repère temporel est nécessaire à l'interprétation de l'expression temporelle
P - Localisation et délimitation précise et bien déterminée de la plage temporelle concernée
U - N'est composée que d'une seule plage temporelle

Les expressions qui rentrent dans la catégorie PRPU possèdent la particularité d'avoir besoin d'un point de référence afin de pouvoir être identifiées correctement. Cette caractéristique se matérialise sous la forme de dates sous-spécifiées ou de déplacements temporels.

Spécification générale A : date sous-spécifiée

Les expressions temporelles de type PRPU peuvent être exprimées au moyen d'une date sous, ou incomplètement, spécifiée ($TARGET_{(part)}$). Cela signifie que l'expression ne contient pas suffisamment d'informations pour permettre d'identifier correctement le point du calendrier qu'elle désigne.

$$TARGET^{(1,2)}_{(part)} \left(G_{(*)} \right) TEMPSHIFT \left(Ref_{(*)} \left(Move_{(*)} \; Amp_{(*)} \right)^{(0,1)} \right)$$

Les informations proposées par une date sous-spécifiée peuvent être une date d'une granularité quelconque[30] ($DATE$), mais doit toujours rester sans mention de l'année[31]. La date peut éventuellement être accompagnée d'une $HEURE$, d'une partie de journée ($G_{(pod\_*)}$) ou d'une modification de la granularité vers la semaine ou le week-end $G_{(we|week)}$ (ID=1) :

« décembre » ;
« jeudi » ;
« le jour de la Toussaint » ;
« [le matin,] [à 10 heures,] [le] jeudi quatorze décembre » ;
« le week-end du samedi seize décembre » ;
« la semaine du quatorze décembre ».

D'autre part, les mentions d'une heure ($HEURE$), ou d'une partie de journée $G_{(pod\_*)}$, peuvent aussi apparaître, seules ou ensembles, mais sans mention de $DATE$ (ID=2) :

« à 13 heures » ;
« le soir » ;
« à l'aube, à 5h45 ».

Pour relier une date incomplète – ou sous-spécifiée – au calendrier, il est nécessaire d'explorer l'espace temporel autour d'un point de repère. Celui-ci est constitué par le moment de l'énonciation ($Ref_{(now)}$) ou un autre point de référence ($Ref_{(focus)}$). Une indication quant au sens ($Move$) et à l'amplitude (Amp) de cette recherche peut être mentionnée. Ces éléments apparaissent dans le groupe $TEMPSHIFT$.

« jeudi en huit » ;

[30] Pour que l'expression soit sous-spécifiée, la granularité doit rester inférieure à l'année.
[31] Dans le cas contraire, il s'agirait d'une date bien identifiée.

« la Toussaint précédente ».

Spécification générale B : un déplacement temporel explicite

L'exploration de l'espace temporel peut aussi être donné par une unité temporelle ainsi qu'une amplitude et un sens de déplacement ($NbTempUnit_{(*,*)}$). Ces éléments définissent un déplacement temporel.

$$TARGET^{(0,1)}_{(part)} \left(G_{(*)} \right)$$
$$TEMPSHIFT \left(Ref_{(*)}\ Move^{(0,1)}_{(*)}\ Precise^{(0,1)}\ NbTempUnit_{(*,*)} \right)$$
$$FOCTEMP^{(0,1)}_{(*)} \left(G_{(*)} \right)$$

Le déplacement temporel s'effectue à partir d'un point de départ qui peut être constitué par :

- l'instant de l'énonciation ($Ref_{(now)}$, ID=3) ;
  « [à 13h,] il y a [exactement] une demi heure [, à 13h] » ;
  « [lundi,] il y a exactement un an [, lundi] ».
- un autre repère temporel, implicite ou explicite ($Ref_{(focus)}$, ID=4).
  « [à 12h,] [exactement] une demi heure avant [12h30] [, à 12h] » ;
  « [lundi,] exactement un an avant [le 11 septembre 2001] [, lundi] ».

Comme ces exemples le montrent, des éléments facultatifs sont présents. Il s'agit, dans le premier cas, de la caractérisation de la cible ($TARGET_{(part)}$), et dans le second, du point de repère ($FOCTEMP_{(*)}$).

Les déplacements temporels de granularité supérieure au jour ne sont pas acceptés (considérés comme flous et versés dans la catégorie PAFU), à moins qu'un marqueur de précision ($Precise$) ou qu'une cible ($TARGET_{(part)}$) soit explicitement spécifié. Lorsque cette dernière est présente il est possible que le déplacement soit spécifié de manière imprécise. Dans ce cas, il n'est pas tenu compte de cette marque d'imprécision.

Spécification générale C : déplacement temporel implicite (adverbe)

Le déplacement temporel ($TEMPSHIFT$) peut encore être indiqué au moyen d'un adverbe temporel qui *code* les informations de sens ($Move$) et d'amplitude (Amp) de recherche.

$$TARGET^{(0,1)}_{(part)} \left(G_{(*)} \right)$$
$$TEMPSHIFT \left(Ref_{(*)}\ G_{(day\ |\ undef)}\ Move_{(*)}\ Amp_{(*)} \right)$$
$$FOCTEMP^{(0,1)}_{(*)} \left(G_{(*)} \right)$$

Le point de référence peut à nouveau être :

- le moment de l'énonciation ($Ref_{(now)}$, ID=5) : « avant-hier [, le samedi 14] »
- un autre repère temporel ($Ref_{(focus)}$, ID=6) : « [le jeudi 14,] la veille [du 15/03/2010] ».

Dans ces cas également, la caractérisation de la cible ($TARGET_{(part)}$), pour le premier cas, ou du point de référence ($FOCTEMP_{(*)}$), pour le second cas, peut être mentionnée.

Spécification générale D : déplacement temporel implicite (grain et localisateur)

(1) Enfin, un déplacement temporel précis peut aussi être *codé* par la désignation d'une granularité temporelle accompagnée d'un élément de positionnement relatif (ID=7).

$$TEMPSHIFT\ \left(Ref_{(*)}\ G_{(*)}\ Move_{(*)}\ Amp_{(*)}\right)$$

Le point de référence peut à nouveau être le moment de l'énonciation ($Ref_{(now)}$) ou un point de repère contextuel ($Ref_{(focus)}$) :

« la semaine dernière »,
« l'année prochaine »,
« l'heure d'après ».

(2) La formulation précédente possède un cas particulier (ID=8) qui survient lorsque le déplacement temporel est nul. Dans ce cas, la granularité se rapporte directement à la cible. Un élément $TARGET_{(part)}$ apparaît alors en complément du déplacement temporel ($TEMPSHIFT$).

$$TARGET_{(part)}\ \left(G_{(*)}\right)\ TEMPSHIFT\ \left(Ref_{(*)}\right)$$

« ce soir »,
« cette année là »

PRFU : Référence Ponctuelle, Relative, Floue et Unique

P - Utilise le point, envisagé à une certaine granularité, comme mode de représentation
R - Un repère temporel est nécessaire à l'interprétation de l'expression temporelle
F - Localisation et/ou délimitation floue ou imprécise de la plage temporelle concernée
U - N'est composée que d'une seule plage temporelle

Les expressions temporelles de type PRFU combinent à la fois un élément d'approximation et les caractéristiques en rapport avec leur caractère relatif. Il existe bien entendu différentes sources d'approximation pour ces expressions temporelles relatives. D'une manière générale, on peut différencier les expressions qui présentent un déplacement temporel imprécis de celles qui localisent de manière imprécise une zone temporelle.

Approximation par déplacement temporel imprécis

Cette catégorie d'expressions se caractérise par l'utilisation des unités de mesure temporelle dans le but de spécifier un déplacement dans l'espace du temps. L'amplitude du déplacement est donnée par une expression à valeur numérique.

Spécification générale A :

Comme pour PRPU, le caractère relatif de l'expression peut provenir d'un phénomène de déplacement temporel à partir d'un point de référence. Cela se traduit principalement sous la forme d'une quantité associée à une unité de mesure temporelle ($NbTempUnit_{(*,*)}$), ainsi que d'un sens de dé-

placement ($Move_{(*)}$). Dans ce cas, l'imprécision provient de l'approximation de la valeur numérique (présence de $Fuzzy_{(*)}$).

$$TEMPSHIFT\ \Big(\ Ref_{(*)}\ Move_{(*)}\ Fuzzy^{(0,1)}_{(*)}\ (\ G_{(*)}\ |\ NbTempUnit_{(*,*)}\ )\ \Big)$$
$$FOCTEMP^{(0,1)}_{(*)}\ \Big(\ G_{(*)}\ \Big)$$

Le point de référence qui sert de point de départ peut être :

- l'instant de l'énonciation, $Ref_{(now)}$, (ID=2) : « il y a [environ] un an » ;
- un autre repère temporel, $Ref_{(focus)}$, (ID=4) : « [à peu près] deux mois plus tôt ».

La marque d'approximation $Fuzzy_{(*)}$ est obligatoire pour les déplacements de granularité inférieure ou égale au jour, mais facultative pour les granularités plus élevées qui sont considérées de toutes façons comme imprécises.

L'imprécision peut cependant encore être plus importante lorsque la spécification numérique du déplacement temporel ($NbTempUnit_{(*,*)}$) est remplacée par un adjectif indéfini, tel que « quelques » ou « plusieurs », accompagné d'une unité de mesure temporelle ($Fuzzy_{(*)}\ G_{(*)}$). À nouveau, deux types de points de référence sont possibles :

- l'instant de l'énonciation, $Ref_{(now)}$, (ID=1) : « il y a quelques jours » ;
- un autre repère temporel, $Ref_{(focus)}$, (ID=3) : « plusieurs semaines après ».

De manière facultative, on peut voir apparaître une caractérisation du point de repère ($FOCTEMP_{(*)}$) pour (ID=3) et (ID=4) :

« quelques jours après jeudi » ;
« [environ] un an avant le 22 décembre 2009 » ;
« quasiment 4 jours après la Toussaint ».

Ces expressions décrivent généralement une zone temporelle dont la granularité est inférieure ou égale au jour, même si l'unité employée pour le déplacement temporel est d'un grain plus élevé.

Approximation par localisation imprécise d'une zone temporelle

La localisation imprécise d'une zone temporelle consiste en une localisation telle que déjà exposée pour la catégorie PRPU, à laquelle vient se greffer une marque d'imprécision. La zone temporelle peut être désignée par :

LOC1. un déplacement temporel précis *codé* par la désignation d'une granularité temporelle («semaine») accompagné d'un élément de positionnement relatif (« cette », « prochaine »).

LOC2. une date sous-spécifiée (« le 22 décembre »). L'aspect relatif provient de la nécessité de considérer un point de repère temporel lors de la recherche d'un point qui satisfait aux contraintes de la date sous-spécifiée.

LOC3. une expression adverbiale (« hier », « la veille ») qui localise de manière relative à un point de référence une zone temporelle. Celle-ci est déterminée à l'aide des

instructions que *code* l'adverbe.

Ces localisations temporelles peuvent être réalisées à des granularités diverses.

Comme pour les expressions de type PAFU, l'imprécision peut être obtenue à l'aide d'un marqueur d'approximation ($Fuzzy_{(*)}$), de partition ($PartOf_{(*)}$), ou d'une combinaison de ceux-ci ($Fuzzy_{(*)}\ PartOf_{(*)}$).

Spécification générale B : LOC1, quel que soit la granularité

(1) La zone temporelle qui subit l'approximation est déterminée suite à un déplacement temporel précis, mais qui n'est pas exprimé par une valeur numérique (ID=5). Dans ce cas, ce sont des adjectifs qui présentent une valeur chronologique qui permettent la localisation dans le temps ($Move$ et Amp), par rapport à un point de référence.

$$(\ Fuzzy_{(*)} \mid PartOf_{(*)}\ )^{(1,N)}\ TEMPSHIFT\ \Big(Ref_{(*)}\ G_{(*)}\ Move_{(*)}\ Amp_{(*)}\Big)$$

Le point de référence peut être :

- le moment de l'énonciation ($Ref_{(now)}$) : « aux environs de la semaine prochaine » ;
- un point de repère contextuel ($Ref_{(focus)}$) : « au cours du jour d'après »

(2) Un cas particulier de ce cas est rencontré lorsque le déplacement temporel est nul (ID=6). La localisation de la zone temporelle est alors identique à celle du point de référence, en tenant cependant compte de la granularité visée pour la cible. L'expression « cette année », envisagée avec un point de référence dont la granularité est le jour, prend par exemple la valeur de ce point mais à la granularité de l'année.

$$(\ Fuzzy_{(*)} \mid PartOf_{(*)}\ )^{(1,N)}\ TARGET_{(part)}\ \Big(G_{(*)}\Big)\ TEMPSHIFT\ \Big(Ref_{(*)}\Big)$$

Spécification générale C : LOC2 et LOC3, à une granularité supérieure ou égale au jour.

(1) L'approximation peut s'appliquer à une date cible (relative, $TARGET_{(part)}$) d'une granularité supérieure ou égale au jour (ID=7).

$$(\ Fuzzy_{(*)} \mid PartOf_{(*)}\ )^{(1,N)}\ TARGET_{(part)}\ \Big(G_{(*>=day)}\Big)$$
$$TEMPSHIFT\ \Big(Ref_{(*)}\ (Move_{(*)}\ Amp_{(*)}\ G_{(*)})^{(0,1)}\Big)$$

« (vers | le début du | vers le début du) 22 décembre [prochain] »,
« (durant le | la fin du | durant la fin du) mois de janvier [[de l'année] dernièr[e]] »,
« (lors | le début | lors du début) de la [prochaine] semaine pascale ».

Plusieurs combinaisons de marqueurs sont possibles : $Fuzzy_{(*)}$, $PartOf_{(*)}$ ou encore $Fuzzy_{(*)}$ $PartOf_{(*)}$. Une indication quant au sens ($Move$) et à l'amplitude (Amp) de cette recherche peut être mentionnée.

(2) La cible peut, dans quelques cas, être remplacée par un adverbe qui « code » un déplacement temporel et permet donc de déterminer une valeur pour $Move$ et Amp (ID=8) :

$$(Fuzzy_{(*)} \mid PartOf_{(*)})^{(1,N)}\ TEMPSHIFT \left(Ref_{(*)}\ Move_{(*)}\ Amp_{(*)}\ G_{(*>=day)} \right)$$

« aux environs d'hier »,
« vers le lendemain ».

Spécification générale D : LOC2 et LOC3, à une granularité égale à une partie de journée.

Le niveau de granularité qui désigne une partie de journée peut également faire l'objet d'une localisation imprécise. Comme pour la spécification C, les marqueurs $Fuzzy_{(*)}$, $PartOf_{(*)}$ ou encore $Fuzzy_{(*)}\ PartOf_{(*)}$ sont utilisés. Dans tous les cas, la mention d'une heure approximative est possible, mais pas obligatoire.

(1) La partie de journée peut être déterminée par une date (sous-spécifiée) et un grain $G_{(pod\_*)}$ (ID=9) :

$$TARGET_{(part)} \left(G_{(day)} \right)$$
$$(Fuzzy_{(*)} \mid PartOf_{(*)})^{(1,N)}\ TARGET_{(part)} \left(G_{(pod)} \right)\ (Fuzzy_{(*)}\ TARGET_{(part)} \left(G_{(min)} \right))^{0,1}$$
$$TEMPSHIFT \left(Ref_{(now)} \right)$$

« mardi peu de temps avant l'aube [, vers 5 heures] »,
« le jour de Noël, le début de la soirée [, aux environs de 20h45] »,
« mardi, en début de matinée [, vers 9 heures] ».

(2) De même, la date peut être remplacée par un adverbe (ID=10) qui *code* un déplacement temporel ($Move$ et Amp) par rapport au moment de l'énonciation ($Ref_{(now)}$) ou un point de repère contextuel ($Ref_{(focus)}$) :

$$TEMPSHIFT \left(Ref_{(*)}\ Move_{(*)}\ Amp_{(*)}\ G_{(day)} \right)$$
$$(Fuzzy_{(*)} \mid PartOf_{(*)})^{(1,N)}\ TARGET_{(part)} \left(G_{(pod)} \right)\ (Fuzzy_{(*)}\ TARGET_{(part)} \left(G_{(min)} \right))^{0,1}$$

« hier, peu de temps après midi [, vers 12 heures 30] »,
« aujourd'hui début d'après-midi [, vers 15 heures] »,
« la veille, en fin de matinée [, vers 10 heures] ».

Spécification générale E : LOC2 et LOC3, à une granularité égale à l'heure.

Enfin, on retrouve le même phénomène rencontré aux spécifications C et D pour la granularité correspondant aux heures.

(1) La mention de l'heure peut apparaître en complément d'une indication de journée (date sous-spécifiée), ou de partie de journée (ID=11) :

$$TARGET_{(part)} \left(G_{(day)} \right)$$
$$TARGET^{(0,1)}_{(part)} \left(G_{(pod)} \right)\ Fuzzy_{(*)}\ TARGET_{(part)} \left(G_{(min)} \right)$$
$$TEMPSHIFT \left(Ref_{(now)} \right)$$

« mardi, aux environs de 13h45 »,
« le jour de Noël au soir, aux environs de 20h ».

(2) Comme précédemment, il est également possible d'utiliser un adverbe pour désigner le jour concerné (ID=12). Celui-ci pourra éventuellement être omis[32]. À nouveau, l'adverbe *code* le déplacement temporel ($Move$ et Amp) par rapport au moment de l'énonciation ($Ref_{(now)}$) ou à un point de repère contextuel ($Ref_{(focus)}$) :

$$TEMPSHIFT\ \left(Ref_{(*)}\ Move_{(*)}\ Amp_{(*)}\ G_{(day)}\right)$$
$$TARGET^{(0,1)}_{(part)}\ \left(G_{(pod)}\right)\ Fuzzy_{(*)}\ TARGET_{(part)}\ \left(G_{(min)}\right)$$

« aujourd'hui, vers 13 heures 30 »,
« aux environs de 13h45 »,
« la veille, vers 13 heures 40 ».

DAPU : Référence Durative, Absolue, Précise et Unique

D - La représentation, à un certain niveau de granularité, fait intervenir deux points (un intervalle), dont un au moins est exprimé
A - Peut être située directement sans faire appel à un repère temporel
P - Localisation et délimitation précise et bien déterminée de la plage temporelle concernée
U - N'est composée que d'une seule plage temporelle

La catégorie DAPU regroupe des zones temporelles exprimées sous la forme d'intervalles, et dont la ou les bornes sont exprimées sous la forme de points temporels bien identifiés. Cela signifie qu'il est possible de déterminer directement, et avec précision (en fonction d'une certaine granularité), l'emplacement dénoté par l'expression dans l'espace du temps.

Spécification générale A : Jour et intervalle de parties de journées ou d'heures.

Lorsqu'un jour est spécifié, l'intervalle peut s'établir à un niveau de granularité moindre, soit les parties de journée, soit les heures. Dans ce cas, la valeur attribuée au jour se distribue aux bornes. Les deux configurations, un borne (ID=1) ou deux (ID=2), sont possibles. Lorsque les deux bornes sont explicitement spécifiées, la première constitue le début de l'intervalle et la seconde la fin. D'une manière naturelle, les deux bornes s'établiront au même niveau de granularité[33].

$$PAPU\ \left(G_{(day)}\right)\ BOUND^{(1,2)}_{(*)}\ \left(G_{(pod\ |\ min)}\right)$$

« 10/03/2005, entre 10h45 et 12h30 »,
« 10 mars 2005, depuis le matin »,
« Noël 2005, depuis 10h45 ».

Spécification générale B : Intervalle de dates.

Une formule plus classique d'intervalles de type DAPU consiste à faire intervenir, pour chacune des

[32] Le jour sera alors supposé être le moment de l'énonciation ($Ref_{(now)}$). Cette décision est cependant arbitraire car le point de référence pourrait tout aussi bien être contextuel.

[33] En pratique, rien n'empêche qu'il y ait une différence à cet égard. L'intervalle « à partir de 14h00 jusqu'au soir » est tout à fait valable. La construction des grammaires, qui a donné la priorité aux cas les plus évidents et fréquents, n'a cependant pas encore prévu cette possibilité à ce stade. À l'avenir, ce genre de cas pourra évidemment être ajouté.

bornes exprimées, une expression de type PAPU, quelle que soit sa granularité. À nouveau, les deux bornes peuvent être explicites (ID=3) ou, au contraire, seule la borne inférieure (ID=5) ou supérieure (ID=6) peut apparaître. Lorsque deux bornes sont spécifiées, et que certains éléments, tels que la désignation du mois et de l'année, sont communs, une factorisation de ceux-ci est possible.

$$BOUND^{(1,2)}_{(*)} \left(PAPU \left(G_{(*)} \right) \right)$$

« les années 2009 à 2010 »,
« du 8 mars 2005 au 10 mars 2005 »,
« du 8 au 10 mars 2005 » (mois et année factorisés),
« depuis décembre 2003 »,
« jusqu'au 10 mars 2005 ».

Un cas particulier doit cependant encore être inclus. Il s'agit de la forme particulière qui fait intervenir, avant la désignation des bornes, une indication d'une partie de journée (ID=4). Il s'agit plus précisément des expressions telles que « la nuit du 9 au 10 mars 2005 ». Dans ce cas, la signature est identique à celle donnée ci-dessus (dans sa version incluant deux bornes), mais il faut lui adjoindre l'élément introducteur $G_{(pod\_night)}$.

DAFU : Référence Durative, Absolue, Floue et Unique

D - La représentation, à un certain niveau de granularité, fait intervenir deux points (un intervalle), dont un au moins est exprimé
A - Peut être située directement sans faire appel à un repère temporel
F - Localisation et/ou délimitation floue ou imprécise de la plage temporelle concernée
U - N'est composée que d'une seule plage temporelle

La catégorie DAFU correspond à une expression de type DAPU pour laquelle il existe une approximation de localisation. Les bornes sont donc constituées d'expressions répondant aux exigences de la catégorie PAPU. Cependant, au moins l'une d'elles reçoit une marque d'approximation ($Fuzzy_{(*)}$), de partition ($PartOf_{(*)}$) ou une combinaison des deux ($Fuzzy_{(*)}\ PartOf_{(*)}$).

Spécification générale A : Jour et intervalle de parties de journées ou d'heures.

Dans le cas de la désignation, à l'intérieur d'une date spécifiée, d'un intervalle de parties de journées ou d'heures, l'approximation se porte sur ces éléments (ID=1 et 2).

$$PAPU \left(G_{(day)} \right) BOUND^{(1,2)}_{(*)} \left((\ Fuzzy_{(*)} \mid PartOf_{(*)}\ )^{1,N}\ G_{(pod \mid min)} \right)$$

« 10 mars 2005, depuis le début de matinée »,
« 10 mars 2005, depuis 10h45 jusqu'aux alentours de 12h30 ».

Spécification générale B : Intervalle de dates.

À nouveau, la configuration de la seconde signature est sensiblement identique à celle proposée pour la catégorie DAPU, à la différence des marques d'imprécision et de partition : $Fuzzy_{(*)}$ et

$PartOf_{(*)}$ (ID=3 à 5).

$$BOUND^{(1,2)}_{(*)} \left(\left(Fuzzy_{(*)} \mid PartOf_{(*)} \right)^{1,N} PAPU \left(G_{(*)} \right) \right)$$

« depuis le 8 mars 2005 jusqu'aux environs du 10 mars 2005 »,
« à partir de début décembre 2009 »,
« jusqu'à fin 2010 ».

Enfin, pour le cas particulier faisant intervenir la partie de journée « nuit », l'approximation se porte sur cet élément :

« durant la nuit du 9 au 10 mars 2005,
« au début de la nuit du 9 au 10 mars 2005 ».

DRPU : Référence Durative, Relative, Précise et Unique

D - La représentation, à un certain niveau de granularité, fait intervenir deux points (un intervalle), dont un au moins est exprimé
R - Un repère temporel est nécessaire à l'interprétation de l'expression temporelle
P - Localisation et délimitation précise et bien déterminée de la plage temporelle concernée
U - N'est composée que d'une seule plage temporelle

D'une manière générale, les expressions de type DRPU sont, du moins en surface assez similaires à celles rencontrées pour la catégorie DAPU. La différence entre ces deux catégories intervient dans le fait que les expressions qui constituent les bornes s'apparentent à la catégorie PRPU et non plus PAPU.

Spécification générale A : Jour et intervalle de parties de journées ou d'heures.

Les intervalles de parties de journées ou d'heures se rapportent ici à une date de type PRPU (ID=1 et 2). Cette date est cependant facultative. Dans le cas où ce point d'ancrage n'est pas fourni, une information supplémentaire est présente au niveau de la borne. Il s'agit de l'élément $Ref_{(focus)}$ qui est abrité par $TEMPSHIFT$ et dont l'utilité est de renseigner sur le type de point de référence auquel doit être rattaché l'intervalle.

$$PRPU^{(0,1)} \left(G_{(day)} \right) BOUND^{(1,2)}_{(*)} \left(G_{(pod \mid min)} \, TEMPSHIFT \left(Ref_{(*)} \right) \right)$$

« [jeudi,] entre 10h45 et 12h30 »,
« [le 10 mars,] depuis 10h45 »,
« [le jeudi 10 mars,] jusqu'à soir ».

Spécification générale B : Intervalle de dates.

Les expressions visées par ce point (ID=3 à 6) correspondent à des intervalles mono- ou bi-borne qui se positionnent dans l'espace du temps au moyen d'éléments relatifs. Comme dans le cas PRPU, certains éléments, tels que le mois et l'année, peuvent être factorisés lorsqu'ils sont identiques pour les deux bornes.

$BOUND^{(1,2)}_{(*)}\ \left(\ PRPU\ \left(\ G_{(*)}\ \right)\ \right)$

« du 8 au 10 mars »,
« jusqu'au 10 mars »,
« dès le 18 mai prochain »,
« jusqu'au 31 décembre de l'an dernier ».

À nouveau, le cas particulier de la nuit située à cheval sur deux dates est présent : « la nuit de mercredi à jeudi ».

Une deuxième sorte d'expressions répond à la même spécification générale, mais en affichant cependant certaines différences. Il s'agit d'expressions mono-bornes dont la définition fait appel à une expression PRPU *codant* un déplacement temporel (ID=7 et 8).

« dans les trois jours »,
« depuis deux mois ».

DRFU : Référence Durative, Relative, Floue et Unique

D - La représentation, à un certain niveau de granularité, fait intervenir deux points (un intervalle), dont un au moins est exprimé
R - Un repère temporel est nécessaire à l'interprétation de l'expression temporelle
F - Localisation et/ou délimitation floue ou imprécise de la plage temporelle concernée
U - N'est composée que d'une seule plage temporelle

Enfin, la catégorie DRFU est le pendant approximatif des expressions de type DRPU. Les différents cas identifiés pour cette catégorie sont par conséquent transposables à celle-ci, en prenant soin d'y ajouter les marques d'imprécision ($Fuzzy_{(*)}$) et de partition ($PartOf_{(*)}$) nécessaires au caractère approximatif de DRFU.

Spécification générale A : Jour et intervalle de parties de journées ou d'heures.

Les intervalles de parties de journées ou d'heures se rapportant à une date (facultative), reçoivent obligatoirement les marques d'approximations au niveau de la ou des bornes (ID=1).

$PRPU^{(0,1)}\ \left(\ G_{(day)}\ \right)$
$BOUND^{(1,2)}_{(*)}\ \left(\ (\ Fuzzy_{(*)}\ |\ PartOf_{(*)}\ )^{1,N}\ G_{(pod\ |\ min)}\ TEMPSHIFT\ \left(\ Ref_{(*)}\ \right)\ \right)$

« [le 10 mars,] depuis environ 10h45 »,
« [le jeudi 10 mars,] jusqu'en début de soirée »,
« [jeudi,] jusqu'aux alentours de 12h30 ».

Spécification générale B : Intervalle de dates.

Les intervalles de dates (ID=2, 4 et 5) subissent également les effets de l'approximation au niveau des bornes. Par conséquent, l'expression de type PRPU est accompagnée par une ou plusieurs marques d'imprécision ($Fuzzy_{(*)}$) et de partition ($PartOf_{(*)}$).

$$BOUND_{(*)}^{(1,2)} \left(\left(Fuzzy_{(*)} \mid PartOf_{(*)} \right)^{1,N} PRPU \left(G_{(*)} \right) \right)$$

« entre début janvier et fin février »,
« depuis début décembre »,
« jusqu'aux environs du 10 mars ».

Le cas particulier concernant la période de la nuit entre deux jours (ID=3) est évidemment toujours présent : « durant la nuit de mercredi à jeudi ».

Enfin, les cas d'intervalles désignés par une seule borne constituée d'une spécification de déplacement temporel peuvent également être approximés (ID=6 et 7).

« d'ici plus ou moins trois jours »,
« depuis à peu près deux mois »,
« d'ici deux à trois semaines ».

7.7 Création des grammaires locales d'extraction

Les grammaires locales ont été créées manuellement, à l'aide de l'éditeur de graphes d'Unitex (Paumier [2003, 2008]), et selon la méthode de *bootstrap*, telle qu'exposée à la section 7.4. Ces grammaires sont plus précisément des transducteurs, puisqu'elles cumulent le rôle de reconnaissance et d'annotation. Leur application aux textes permet un balisage précis et détaillé des expressions temporelles. Les graphes qui représentent les principaux transducteurs sont repris en annexe C.

7.7.1 Choix d'un format d'annotation

De nombreux formats ont été définis au cours du temps pour l'annotation d'expressions temporelles. Les principaux, dont il a déjà été question à la section 6.5, se nomment Timex, Timex2 et Timex3 (TimeML). Leurs capacités descriptives et leurs objectifs ne sont bien entendu pas identiques. D'une manière générale, Timex est le plus limité, car il ne concerne que la délimitation des expressions dans le texte. Avec Timex2, la couverture des expressions à annoter et la précision du balisage augmente. Mais c'est surtout l'attribution d'une valeur interprétée et normalisée qui constitue l'apport le plus important. Timex3 n'amène pas de grands bouleversements, mais est lié au cadre plus large de TimeML. Le champ d'action de ce langage s'étend aux *événements* (verbes conjugués, adjectifs d'état, groupes nominaux) et aux *signaux* (prépositions temporelles, connecteurs, etc.).

Pour différentes raisons, ces formats ne nous ont pas paru totalement appropriés par rapport à nos objectifs. En ce qui concerne Timex, outre les multiples aspects non couverts – les intervalles et les expressions imprécises entre autres – c'est aussi la destination exclusivement orientée vers le repérage des expressions qui rend ce langage trop limité. Au contraire, Timex2 constitue un bon candidat, au vu de ses capacités à attribuer une valeur aux expressions. Quant à Timex3, pris en dehors de TimeML[34], sa valeur ajoutée par rapport à Timex2 ne constitue pas une avancée significative. Les

[34] L'objectif poursuivi est plutôt axé sur les événements et sur la manière de les relier aux expressions temporelles, ce

remarques qui suivent se rapportent par conséquent au format Timex2, qui se rapproche le plus de nos besoins.

La couverture de Timex2 s'étend des expressions précises aux expressions floues, en passant par les fréquences et les expressions non spécifiques. Globalement, Timex2 correspond assez bien à l'annotation désirée. Un certain nombre d'éléments requis, tels que les expressions relatives, les modificateurs (début, fin, etc.) ou les instants de la journée (matin, midi, soir, etc.), sont supportés par ce format. Par contre, d'autres critères ne sont pas rencontrés. Tout d'abord, la prise en compte des expressions temporelles imprécises n'est pas assez poussée, tant au niveau de la variété des expressions couvertes par le format, que dans la manière de signaler l'imprécision. Ainsi, « vers le 30 septembre » n'est pas référencé comme expression à annoter, et « il y a un an » se voit attribuer une valeur d'une granularité égale à l'année, sans autre explication quant au type de l'imprécision. D'autre part, un autre point important est l'absence d'annotation interne, c'est-à-dire au niveau des sous-constituants de l'expression. Timex2 autorise bien l'imbrication de balises, mais cela reste insuffisant pour atteindre une annotation fine, et en profondeur, des expressions temporelles. Or ces éléments d'informations sont capitaux. En effet, chaque sous-constituant est susceptible d'apporter une instruction particulière, nécessaire à l'étape d'interprétation pour permettre l'attribution d'une valeur à cette expression.

Par conséquent, Timex2 semble mieux adapté pour une annotation manuelle (humaine) ou en tant que format de sortie, plutôt que comme format intermédiaire, permettant de baliser le résultat de la reconnaissance automatique avant son interprétation. Un format ad-hoc, conforme aux besoins réels de notre système, a finalement été adopté[35]. Sa définition est reprise dans la partie consacrée à la spécification complète des expressions temporelles, en annexe B.

La définition d'un format ad-hoc offre une grande liberté, qui permet de concentrer un maximum d'analyses au niveau des grammaires locales. Cela évite de devoir les mener dans un deuxième temps, que ce soit pour rassembler les sous-constituants s'ils avaient été annotés séparément, ou pour déterminer leur rôle dans l'interprétation temporelle de l'expression. Cette démarche est réalisée dans un souci de simplification et de performance du processus d'analyse. Elle possède également l'avantage d'inclure les éléments nécessaires à l'interprétation dans la dynamique de spécification, telle qu'exposée à la section 7.4.

Le choix d'une annotation non-standard représente cependant certains inconvénients. En ce qui concerne la réutilisabilité et l'utilisation des résultats par d'autres systèmes, un format ad-hoc constitue une difficulté, car un module spécifique doit alors être développé pour importer les données. D'autre part, l'évaluation des performances par rapport à d'autres systèmes ou au moyen de corpus annotés, par exemple le TimeBank français (Bittar [2010]), est plus compliquée, voire impossible. Au final, cela peut avoir pour conséquence de diminuer la visibilité et l'éventuelle diffusion du système dans la communauté scientifique. Il est cependant possible de remédier à ce défaut en prévoyant, à partir de la représentation interne des données, une sortie qui soit conforme à un format standard, par

qui sort du cadre que nous avons défini pour cette thèse.

[35] Dans la suite du texte, il nous arrivera, par commodité, de désigner les expressions temporelles extraites sous le nom de *timex* (pour *time expression*), sans pour autant faire référence à un format d'annotation précis.

exemple Timex3 (TimeML).

7.7.2 Remarques particulières

Il existe un graphe *principal* par catégorie d'expressions temporelles. À ceux-ci on ajoute les graphes qui prennent en charge les durées et les âges, qu'ils soient exprimés de manière précise ou imprécise. Enfin, un dernier graphe reprend les exclusions : des termes qui sont ambigus avec des expressions porteuses d'informations temporelles, mais dont le sens ne se rapporte pas au temps (« la gare de Bruxelles Midi », « le journal Le Soir », « une minute de gloire », etc.). L'ensemble de ces transducteurs, au total 85 graphes[36], sont rassemblés en un graphe principal, qui contient aussi les patrons nécessaires à la reconnaissance des balises de texte (« <texte ...> », voir section 7.8.2).

Les références à la spécification apparaissent sous la forme d'un identifiant présent dans le balisage. Son utilité est d'indiquer, pour chaque expression reconnue, à quelle sous-catégorie celle-ci appartient. Cette information, parfois complétée par des commentaires ou des exemples, est utile pour le processus d'entretien, de mise à jour et d'extension des grammaires (voir section 7.4). Cette annotation particulière est également exploitée pour la réalisation de statistiques précises sur la distribution des expressions temporelles dans les textes analysés (voir Annexe D).

Lors du processus de création des grammaires locales, certaines questions ont surgi. En particulier en ce qui concerne les choix d'organisation et de *découpage*. Le problème se pose au sujet du regroupement des patrons selon leur sens ou selon leurs caractéristiques morphologiques. Dans le premier cas, l'avantage est de regrouper les séquences qui ont les mêmes sorties, et qui ne doivent donc pas être répétées (toutes les expressions décrites en un endroit ont le même type d'annotation). De plus l'organisation interne des grammaires respecte une certaine logique, ce qui les rend plus compréhensibles et plus lisibles. Par contre, l'inconvénient provient du non-regroupement de certains patrons relativement similaires, qui restent dissociés en raison de leurs sorties différentes. Dans le second cas, un regroupement des séquences similaires est possible[37], mais au prix d'une plus forte répétition des séquences de sortie[38]. Dans ce cas, les patrons relatifs aux différentes catégories se retrouvent mélangés, ce qui diminue également la clarté de la ressource.

L'option choisie a été de privilégier un regroupement selon le sens, et donc de conserver une structure la plus logique possible afin de faciliter les futures mises à jour. Pour diminuer les répétitions de séquences similaires, la *factorisation* et la réutilisation de portions de graphes ont par contre été exploitées à de nombreuses reprises. La création d'un sous-graphe et son utilisation dans divers contextes offrent une facilité lors de la mise à jour ou de la correction. Cependant, la mise en pratique de cette méthode est parfois difficile à effectuer. En effet, dans le cas de l'ajout ou du retrait d'un élément dans le sous-graphe, ou encore d'une modification des éventuelles séquences de sortie, il peut être nécessaire de le dédoubler. Dès lors, la multiplication de sous-graphes quasiment identiques peut

[36] Auxquels il faut cependant encore ajouter les graphes issus de bibliothèques partagées (éléments généraux, réutilisables).

[37] Par exemple, lorsqu'il y a des possibilités d'insertion de marqueurs d'imprécision ou de partition.

[38] Ce qui s'avère moins pratique lorsqu'une modification doit y être apportée.

devenir un effet pervers de la méthode.

7.8 Implémentation de l'analyse temporelle

7.8.1 Architecture du système

Le système a été conçu de manière à privilégier deux caractéristiques : la modularité et l'ouverture. En ce qui concerne l'aspect modulaire, il consiste en une séparation claire des différentes étapes d'analyses. Cette découpe en sous-tâches autorise la modification ou le remplacement du module responsable d'une étape par un autre, sans pour autant affecter le processus général[39].

Le système se veut également ouvert. Cela signifie que, sous certaines conditions, il est possible d'intégrer dans la chaîne de traitement dédiée à l'analyse temporelle, d'autres traitements automatiques du langage. Alternativement, le résultat de l'analyse temporelle des textes peut faire l'objet de sorties sous différentes formes, facilement adaptables. Cela doit permettre une intégration aisée de l'analyse temporelle dans une chaîne de traitement tierce.

Le processus complet est illustré à la figure 7.7. Les étapes qui font l'objet d'une présentation détaillée dans le cadre de cette thèse (extraction des expressions temporelles, analyse temporelle, et analyse spécifique à l'indexation pour la recherche d'informations) sont identifiées par les encadrés dont le fond est coloré en gris. Les autres étapes sont assurées par des ressources et des logiciels existants. Le traitement se déroule en plusieurs phases successives, à partir de textes conformes au format d'entrée (Section 7.8.2).

– Phase 1 : *Analyses de base, menées en parallèle*

1. Annotation des expressions temporelles

   Les grammaires de reconnaissance des expressions temporelles (transducteurs), qui ont été construites (Section 7.7) à l'aide d'Unitex (Paumier [2003, 2008]), sont appliquées aux textes afin d'insérer les annotations (balisage). Cette étape est également réalisée à l'aide d'Unitex.

2. Analyse syntaxique

   Cette étape a pour objectif la production d'une analyse permettant la détection des groupes verbaux complexes, ainsi que la découpe en propositions. Le logiciel utilisé pour mener cette analyse est XIP (Aït-Mokhtar *et al.* [2002]). L'analyse syntaxique est utilisée comme une boîte noire. On se contente donc de récupérer l'annotation qu'elle produit et d'exploiter les éléments relatifs aux propositions et groupes verbaux (voir section 7.3.1).

3. Analyse spécifique pour les besoins d'une application tierce (facultatif)

   Le texte de départ est également envoyé au module de traitement dédié à l'application tierce. Dans le cadre de cette thèse, il s'agit d'une classification effectuée dans un but d'indexation thématique du document (voir chapitre 2). L'intégration au processus d'analyse temporelle est réalisée dans le but de relier le résultat des deux analyses (voir chapitre 8).

[39] Pour autant que les résultats intermédiaires soient toujours envoyés selon le même format.

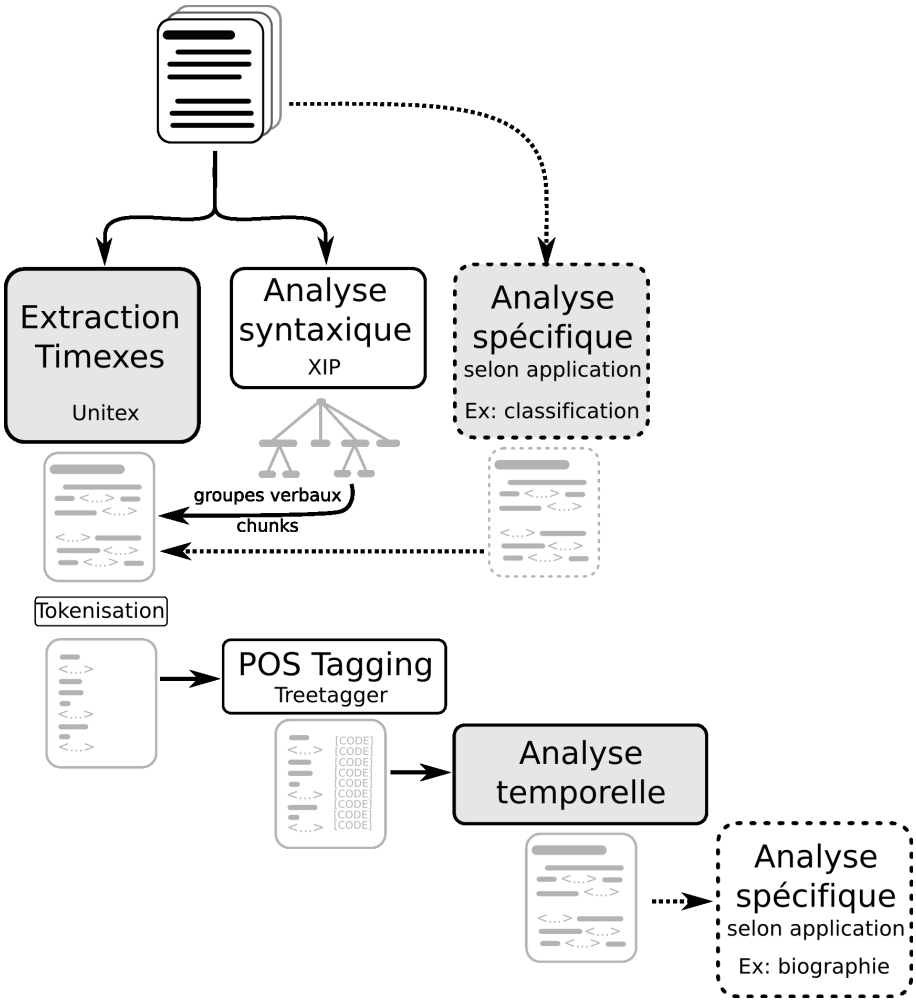


Figure 7.7 : Aperçu des grandes étapes de traitement

– Phase 2 : *Enrichissement de l'annotation du texte*

  Cette étape consiste à ajouter au texte, dans lequel sont déjà annotées les expressions temporelles, les informations issues de l'analyse syntaxique. Il s'agit donc principalement des groupes verbaux complexes ainsi que de la découpe en propositions (via la reconnaissance de chunks). Dans le cas où une application tierce est intégrée au processus de traitement, et qu'elle produit des annotations, celles-ci doivent également venir enrichir le texte temporellement annoté[40]. Le résultat de cette étape est un texte, enrichi de plusieurs types d'annotations, et présenté selon un format intermédiaire (Section 7.8.2) qui reste cependant proche du format de départ.

– Phase 3 : *Analyse en parties du discours (POS tagging)*

  L'analyse en parties du discours (catégories grammaticales) a pour objectif d'obtenir les informations concernant les temps morphologiques des formes verbales. Les unités, simples ou composées, qui font l'objet d'une annotation spécifique issue des phases précédentes (expressions temporelles, expressions qui constituent des indices thématiques pour la classification, etc.) ne sont pas analysées, et sont donc conservées telles quelles. Le POS tagging, assuré par

[40] Il est techniquement plus facile de concilier des analyses qui exploitent les mêmes outils et proposent donc le même type de sortie. L'utilisation d'Unitex pour l'extraction temporelle et pour la classification simplifie grandement la mise en commun des annotations.

Treetagger (Schmid [1994]), est précédé d'une étape de tokenisation du texte.

– Phase 4 : *Analyse temporelle*

  Le texte complètement annoté et tokenisé est ensuite chargé dans une structure de données qui facilite son parcours séquentiel (Section 7.8.3). Cette structure reprend, pour chaque élément, qu'il soit un token simple ou une expression composée, toutes les informations disponibles qui ont été récoltées lors des analyses de base réalisées au préalable[41]. L'analyse temporelle utilise cette structure pour parcourir le texte afin d'interpréter (Sections 7.8.4 à 7.8.9), et de normaliser vers un format défini, les expressions temporelles.

– Phase 5 : *Sortie des résultats*

  Le résultat de l'analyse temporelle peut être réalisé selon différents formats, en fonction de l'utilisation envisagée. Si des traitements particuliers ont été réalisés (Phase 1.3), leurs résultats peuvent évidemment être ajoutés à la sortie.

– Phase 6 : *Récupération des résultats pour exploitation dans une application tierce (facultatif)*

  L'annotation temporelle réalisée peut servir de matière première à une application tierce désirant exploiter, entre autres choses, les informations temporelles contenues dans le texte.

L'évolution de ce processus de traitement est évidemment possible. Certains points susceptibles d'être modifiés sont d'ores et déjà identifiés. C'est par exemple le cas de l'analyse syntaxique, dont l'exploitation très partielle ne justifie pas nécessairement la complexité de cette étape. Il serait, par exemple, possible de trouver des méthodes de remplacement pour la reconnaissance des groupes verbaux. À ce sujet, les travaux de Gross [2000] sur l'anglais, de Constant *et al.* [2002] et Nakamura [2006] pour le français, constituent de bonnes pistes. L'implémentation sous la forme de grammaires locales représente évidemment un avantage qui permettrait une plus grande uniformisation et cohérence des traitements tout au long du processus. De même, la découpe en propositions pourrait être confiée à un module spécialisé. Enfin, il est également possible de s'interroger sur la nécessité de mener une analyse morpho-syntaxique à partir du moment où les temps verbaux sont reconnus de manière satisfaisante par un système d'analyse complet des formes verbales.

L'architecture en pipeline réalisée correspond à une organisation souvent adoptée en traitement automatique du langage. Elle se compose d'une succession d'étapes de complexités diverses, allant du prétraitement jusqu'à la sortie des résultats, en passant par des étapes plus complexes et spécifiques à la tâche réalisée. De nombreuses plate-formes proposent ce genre d'architecture, par exemple GATE[42] (Cunningham *et al.* [2002]), LinguaStream[43] (Wildöcher et Bilhaut [2007]) ou encore UIMA[44].

Bien que l'objectif de ce travail n'est pas d'aboutir à un système allant au delà du stade de prototype ou de la preuve de concept, il est intéressant de s'interroger sur l'architecture qui serait la plus appropriée pour une application finale.

[41] Il est possible de conserver les codes grammaticaux et lemmes des mots quelconques (ni expression temporelle, ni verbe, etc.), même s'ils ne sont pas nécessaires pour l'analyse temporelle en tant que telle.

[42] `http://gate.ac.uk/`

[43] `http://www.linguastream.org/home.html`

[44] `http://uima.apache.org/`. A été mis au point par IBM avant de passer en open-source.

D'une manière générale, l'intégration et la communication entre les différents composants du système doivent être améliorées, de manière à avoir un flux d'information plus simple et cohérent. Par exemple, en ce qui concerne la tokenisation, celle-ci est actuellement réalisée à deux endroits différents de la chaîne de traitement, par deux systèmes distincts (le programme Tokenize d'Unitex et le script de tokenisation du Treetagger).

En ce qui concerne les modules utilisés pour des tâches particulières (par exemple la chaîne de traitement d'Unitex), il serait intéressant qu'ils soient disponibles en tant que service permanent et indépendant, par exemple sous la forme d'un serveur. Du point de vue des performances, le maintient en mémoire des ressources nécessaires à l'analyse (dictionnaires, etc.) représente un gain de temps intéressant.

Au niveau de l'organisation globale, il semble donc intéressant de disposer de modules dont la tâche est de rendre un service bien précis. Ces modules doivent adopter un format de communication commun (par exemple basé sur XML). La décomposition de la tâche en services autonomes offre également l'avantage de pouvoir distribuer plus facilement l'application sur plusieurs ordinateurs. À ce stade du développement, il est cependant un peu prématuré pour se prononcer sur un schéma d'implémentation plus précis.

L'importance du choix d'un paradigme de programmation et d'un langage est aussi encore assez relatif, car une grande partie du travail se situe au niveau des grammaires. Le code informatique sert principalement à lancer certains programmes (Unitex, Treetagger, XIP), à récupérer leurs résultats, à organiser les interactions entre les différents modules, et à effectuer des opérations de base (par exemple transformer un nombre exprimé en lettres en une valeur en chiffres, charger la structure de données, etc.). La partie dédiée aux tâches de haut niveau ne représente pas la majorité du code.

En pratique, pour des raisons de simplicité d'implémentation et de récupération de certaines parties déjà codées, c'est un langage de script qui a été utilisé (langage impératif, procédural, en l'occurrence PHP). Pour une implémentation future du système, le choix d'un langage orienté objet pourrait s'avérer intéressant pour modéliser les objets temporels selon les critères déterminés à la section 7.6.1. Enfin, les langages fonctionnels ou logiques semblent moins adaptés à l'implémentation d'un système complet de traitement des expressions temporelles. Certains modules, en fonction de leurs objectifs, pourraient cependant adopter des choix d'implémentation particuliers. Par exemple, un module de raisonnement sur le temps pourrait être plus naturellement codé à l'aide d'un langage logique.

7.8.2 Formats de départ et intermédiaire des textes analysés

Les textes fournis en entrée du système doivent se conformer à un format déterminé. Celui-ci est cependant assez léger et peu contraignant. Il est prévu pour conserver une certaine flexibilité, notamment pour permettre l'intégration d'applications tierces au processus d'analyse temporel. Le format de fichier, dont une illustration est donnée à la figure 7.8, exploite un formalisme proche d'XML, sans qu'une validation stricte de ce format soit obligatoire.

```
<text id='F040703A_0103.txt' date='20040703-18:04'>

    <header>
        INT032 3 GEN 0122 F BELGA-0188 109 L
    </header>

    <keywords>
        DIVERS/LOISIRS/DEFENSE/AGENDA/
    </keywords>

    <title>
        Quelque 15.000 visiteurs aux portes ouvertes de la base de Coxyde.
    </title>

        §
         Paragraphe 1.
        §
         Paragraphe 2.
        §

</text>
```

Figure 7.8 : Format pour les fichiers d'entrée (exemple d'une dépêche BELGA).

Les éléments obligatoires sont tout d'abord constitués par les balises qui marquent les limites du texte (<text> et </text>). La balise ouvrante contient deux arguments : l'un pour attribuer un identifiant unique (*id*) sous la forme d'une chaîne de caractères quelconques[45], et l'autre (*date*) pour renseigner sur le moment d'émission du texte, au format « AAAAMMJJ-hh :mm ».

La deuxième information importante qui est demandée par le système est la délimitation des paragraphes. Celle-ci est réalisée en plaçant le symbole « § » de manière à encadrer les différents paragraphes. La détection automatique des endroits où il est nécessaire d'insérer cette marque n'est pas toujours aisée et est donc parfois assez approximative. Il a donc été jugé préférable que ces éléments soient inclus au document de départ. Cela ne représente généralement pas une grosse difficulté car cette information est souvent fournie lorsque les documents originaux sont au moins partiellement structurés.

Enfin d'autres balises peuvent encore être introduites dans le texte, pour autant qu'elles ne cassent pas sa structure syntaxique et qu'elles ne rentrent pas en compétition avec les annotations utilisées par l'analyse temporelle. Elles peuvent donc encadrer des phrases ou apparaître indépendamment de tout élément du texte. Les balises sont reconnues en tant que telles, et non comme du texte, ce qui offre une grande flexibilité (l'utilisation des balises peut facilement être adaptée). Dans l'exemple présenté à la figure 7.8, les balises <header> encadrent une information d'en-tête, qui n'est pas réellement utile mais qui est conservée par cohérence au document de départ. Les balises <keywords> et <title> ont été introduites en prévision du traitement de classification et d'indexation qui est appliqué parallèlement à l'analyse temporelle (voir chapitre 8). D'autres balises peuvent être ajoutées selon les besoins.

En ce qui concerne le format intermédiaire du texte, c'est-à-dire celui obtenu après mise en commun des informations issues de l'analyse temporelle, de l'analyse syntaxique et, le cas échéant, d'une ana-

[45] Pour des raisons pratiques, l'identifiant a systématiquement été le nom du fichier, ce qui permet de faire le lien avec le document d'origine.

lyse tierce, il reste assez similaire au format de départ. Certaines balises peuvent cependant apparaître dans le corps du texte (Figure 7.9). C'est évidemment le cas pour les expressions temporelles (<TIMEX>), pour les formes verbales (<VERB>), ainsi que pour les limites de propositions (<PROP>). Mais d'autres balises, propres à une application particulière, intégrée au processus général, peuvent également venir se glisser dans le texte (par exemple <CLASSIF> pour la classification).

```
<TIMEX=[[[#GRAIN=DAY] [samedi#J_N]#DATE]#TARGET=PART][[#REF=NOW]#TEMPSHIFT]
    [#ID=1],.Time+PRPU> samedi </TIMEX>

<VERB sentence="25" form="sont aussi présentés" verb="sont présentés"
    passive="1"> sont aussi présentés </VERB>

<PROP type="PUN">

<CLASSIF code='4822' weight='2'> base aérienne </CLASSIF>
```

Figure 7.9 : Quelques exemples d'insertions de balises dans le format intermédiaire du texte.

7.8.3 Représentation interne du texte

Le texte, enrichi par l'ensemble des analyses récoltées lors des traitements de base, est chargé dans une structure de données qui facilite son parcours ainsi que la consultation des différentes informations. D'une manière générale, la structure s'apparente à un tableau dont les éléments sont des expressions temporelles, des formes verbales ou groupes verbaux complexes, des formes ou groupes qui font l'objet d'une annotation particulière (en fonction des besoins), des indicateurs de saut de proposition ou encore, plus simplement des *tokens* quelconques. Parmi les champs disponibles pour caractériser les éléments de cette structure (Figure 7.10), un seul apparaît toujours (CODE), alors que les autres sont fonction du type d'élément concerné. Cette représentation est donc tout à fait adaptable et très flexible.

```
[code]      => type d'élément : forme (FORM), saut de proposition (PROP, BREAK?),
               forme particulière (CLASSIF, par exemple), etc.
[form]      => forme complète (avec annotation « interne » le cas échéant)
[verb]      => forme verbale nettoyée de ses insertions
[notagform] => forme sans annotation
[gram]      => codes grammaticaux et sémantiques
[passive]   => indicateur de forme verbale passive (0 ou 1)
[struct]    => structure de données détaillée (pour les expressions temporelles)
[class]     => pour la classification, identifiant de la classe
[weight]    => pour la classification, poids attribué
```

Figure 7.10 : Structure de données pour la représentation interne du texte (définition des champs d'un élément).

Une expression temporelle sera par exemple représentée par les champs CODE, FORM, NOTAGFORM, GRAM et STRUCT. Ce dernier contient une structure propre à la définition des aspects temporels, tel que décrit à la section 7.3.3. Une forme verbale est par contre décrite par les champs CODE, FORM, VERB, GRAM et enfin PASSIVE. Les formes ou groupes annotés spécifiquement pour les besoins d'une application particulière reçoivent la caractérisation qui leur convient : CODE

(qui prend la valeur de la balise attribuée, CLASSIF par exemple), FORM, CLASS et WEIGHT.

7.8.4 Représentation interne du temps

Le choix effectué pour la représentation interne du temps dans le système doit évidemment respecter les orientations définies pour le modèle temporel (Section 7.3.2). Cela signifie, en particulier, que cette représentation doit pouvoir manipuler à la fois des points, à différents niveaux de granularité, ainsi que des intervalles. Ces derniers étant, à peu de choses près, considérés comme un couple de points, la représentation interne du temps peut sans problème être limitée à cette unité de base. Étant donné la définition du modèle temporel, nous avons également été attentif à avoir une représentation centrée sur la granularité du jour.

Le format choisi remplit parfaitement ces exigences : il s'agit du système de jours Julien[46] (Meyer [2009]). Celui-ci fonctionne à l'aide d'entiers qui représentent les jours. Le numéro de jour Julien est défini comme le nombre de jours écoulés depuis le 1^{er} janvier 4713 avant J.C.[47] L'avantage de cette représentation est qu'elle est très commode pour effectuer des calculs temporels. Ceux-ci peuvent en effet être réalisés au moyen d'une simple addition ou soustraction, au contraire d'un système de calendrier, tel que le calendrier grégorien. En effet, avec une représentation calendaire, il est nécessaire de gérer des unités irrégulières (mois, années bissextiles, etc.). Cependant, au final, c'est bien la valeur exprimée dans le système grégorien qui nous intéresse. Ce besoin est rencontré par le système de jours Julien pour lequel il existe des algorithmes de conversion vers le système de calendrier grégorien.

7.8.5 Localisation temporelle des expressions de type PA*U

Les expressions bien identifiées

Les expressions de type PAPU ou PAFU sont les plus simples à traiter. En effet, celles-ci présentent la caractéristique de disposer de toutes les informations nécessaires à leur localisation précise dans l'espace du temps (un calendrier), et cela à un certain niveau de granularité donné. C'est la raison pour laquelle il y sera aussi référé comme des expressions désignant une zone temporelle *bien identifiée*. Par conséquent, la structure de données est déjà correctement remplie grâce aux éléments issus de l'annotation. La seule action à entreprendre consiste à compléter la structure en calculant la valeur des champs facultatifs[48]. Ces champs sont définis en fonction du niveau de granularité.

[46] Le système de jours Julien (*Julian Day number* en anglais) a été inventé par l'astronome John F. Herschel (1738-1822).

[47] Cette date est due à Joseph Justus Scaliger (16e siècle) et résulte de la combinaison de trois cycles calendaires. Le cycle solaire (28 ans) au bout duquel les dates se répètent avec le même nom de jour dans le calendrier Julien. Le cycle des *Golden Numbers* (19 ans) représente la période après laquelle les phases de la Lune se répètent aux même dates calendrier. Et enfin, le troisième, l'*indiction cycle* (15 ans), est un cycle d'origine romaine. Une année peut être caractérisée par un triplet qui possède une valeur pour chacun des cycles. Ainsi l'année de la naissance du Christ est représentée par (9,1,3). L'objectif de leur combinaison est de déterminer un point de départ pour un système calendaire qui soit antérieur à tout enregistrement historique, afin de ne pas devoir gérer de dates négatives. Le point départ a été fixé en l'année (1,1,1) du cycle contenant la naissance du christ, soit l'an 4713 B.C.(Meyer [2009])

[48] Ces valeurs sont directement dérivables à partir des informations obligatoires qui sont elles déjà disponibles.

Prise en compte du caractère flou

En ce qui concerne la gestion de l'imprécision, la situation est également assez simple. Les informations sont véhiculées au moyen des champs $fuzzy$ et $partof$. Le premier peut prendre la valeur « 1 », qui signifie que l'imprécision se situe à l'intérieur de la zone désignée, ou « 2 » qui élargit l'approximation à un certain périmètre[49] autour de cette zone. Ces valeurs sont directement renseignées par l'annotation produite par les grammaires. Le champ $partof$ donne une indication sur la partie d'une zone temporelle qui nous intéresse plus particulièrement. Cette information entraîne automatiquement l'assignation de la valeur « 1 » au champ $fuzzy$, ce qui signifie qu'il existe une imprécision de localisation à l'intérieur de la partie désignée.

7.8.6 Localisation temporelle des expressions de type PR*U

La particularité des expressions de type PRPU et PRFU réside dans le fait qu'elles nécessitent un point de référence pour localiser la zone temporelle désignée. Ce point de référence peut correspondre à la date d'émission du texte (référence déictique) ou à une date issue du contexte (référence anaphorique). Le caractère relatif de l'expression peut apparaître sous diverses formes :

- une expression *sous-spécifiée* ;
- un déplacement temporel ;
- un déplacement temporel avec spécification partielle de la cible ;
- un déplacement temporel avec spécification du point de référence ;
- un déplacement temporel indéterminé.

Dans chacun de ces cas, une procédure d'interprétation particulière est prévue.

Expression sous-spécifiée

La *sous-spécification* d'une expression temporelle est la situation opposée à celle qui désigne une zone temporelle *bien identifiée*, tel qu'exposé pour les expressions de type PA*U à la section 7.8.5. Il s'agit donc d'une expression qui, pour le niveau de granularité auquel elle se trouve, ne fournit pas tous les éléments nécessaires à la localisation de la zone temporelle qu'elle désigne dans l'espace du temps. Celle-ci n'est pas indéterminée pour autant, car dans de nombreux cas le contexte donne les informations nécessaires à son interprétation.

Un point de référence, déictique ou anaphorique doit donc être utilisé. Le processus d'interprétation consiste alors, à partir de ce point, à explorer l'espace du temps à la recherche de la zone temporelle la plus proche qui respecte les *contraintes* exprimées par l'expression sous-spécifiée (la *cible*). La

[49] La taille de ce périmètre n'est pas explicitement fixée et peut donc être adaptée lors de l'exploitation de la donnée, en fonction du degré de précision que l'on désire atteindre ainsi que de la tolérance aux erreurs. Dans tous les cas, l'ordre de grandeur de cette zone pourra être réglé en fonction de la granularité. Il pourrait par exemple être décidé de fixer les limites en se référant à la taille de la granularité supérieure. Une imprécision de type 2 se rapportant à une zone temporelle de granularité *jour* pourrait alors être fixée à 15 jours avant et après celle-ci (un mois comptant environ 30 jours).

direction de recherche est déterminée soit par l'expression elle même, soit, à défaut, par l'indication fournie par le temps verbal attribué à la proposition dans laquelle se situe l'expression. Une fois qu'une solution acceptable est trouvée, l'expression est dite *résolue*. Dans le cas où une solution satisfaisante ne serait pas trouvée dans un voisinage raisonnable, le système décide de garder la valeur de la date d'émission du texte, d'augmenter d'un niveau la granularité de celle-ci et de lui attribuer l'indicateur d'imprécision externe ($fuzzy = 2$).

Une difficulté supplémentaire vient cependant s'ajouter à cette procédure. Il s'agit de la situation dans laquelle le point de référence ne se situe pas au même niveau de granularité que la cible (voir exemple à la figure 7.11). Dans ce cas, il est nécessaire de définir une granularité de recherche commune afin d'effectuer la résolution de l'expression sous-spécifiée. Cette granularité est la plus petite commune granularité obtenue en augmentant la granularité la plus faible[50] (étape de *conciliation* entre niveaux de granularité différents). Il est possible que, lors de cette augmentation, la cible perde toute information discriminante[51] : « le soir » envisagé au niveau de granularité du jour ne propose par exemple aucune information intéressante. Dans ce cas, le point de référence sera directement validé comme solution. Sinon, l'opération de résolution peut se dérouler normalement. Enfin, dans un troisième et dernier temps, la solution trouvée est *réconciliée* avec la granularité employée à l'origine pour la cible. Pour ce faire, les précisions nécessaires sont récupérées à partir de l'expression de départ. Il s'agit par exemple de réattribuer la valeur *soir* à une solution trouvée au niveau de granularité *jour*.

Le mécanisme d'interprétation d'une expression sous-spécifiée, avec adaptation de la granularité, que nous avons nommé « Conciliation / Résolution / Réconciliation » (CRR), est illustré à la figure 7.11.

Déplacement temporel

Un déplacement temporel désigne une opération quasi mathématique. Il s'agit, à partir d'un point de référence, d'ajouter ou de retrancher un nombre donné d'une certaine unité de mesure temporelle.

Comme dans le cas précédent, il est cependant possible que la granularité du point de référence ne soit pas directement compatible avec la granularité du déplacement temporel. Dans ce cas, un mécanisme d'adaptation (Figure 7.12), désigné sous le nom de « Conciliation / Déplacement / Réconciliation » (CDR), doit à nouveau être mis en place. Lors de celui-ci, c'est toujours le niveau de granularité du point de référence qui est adapté – ou *concilié* – à la granularité impliquée par le déplacement temporel. Cela signifie donc que la variation peut autant être effectuée vers une granularité plus fine que plus élevée. Si l'augmentation du niveau de granularité ne pose pas de problème particulier, il n'en va pas de même pour la diminution. En effet, quelle valeur choisir lors de l'adaptation de la granularité de « mardi 8/3/2005 matin » vers la granularité *heure* ? La solution retenue permet de gérer cette situation en attribuant une valeur moyenne, selon la définition du système («mardi 8/3/2005 à 9h00» pour l'exemple cité).

[50] La granularité utilisée à l'étape de résolution ne peut cependant pas être inférieure au jour.

[51] Le point de référence est lui toujours *bien identifié* et ne présente pas de problème particulier lors d'une éventuelle augmentation de sa granularité.

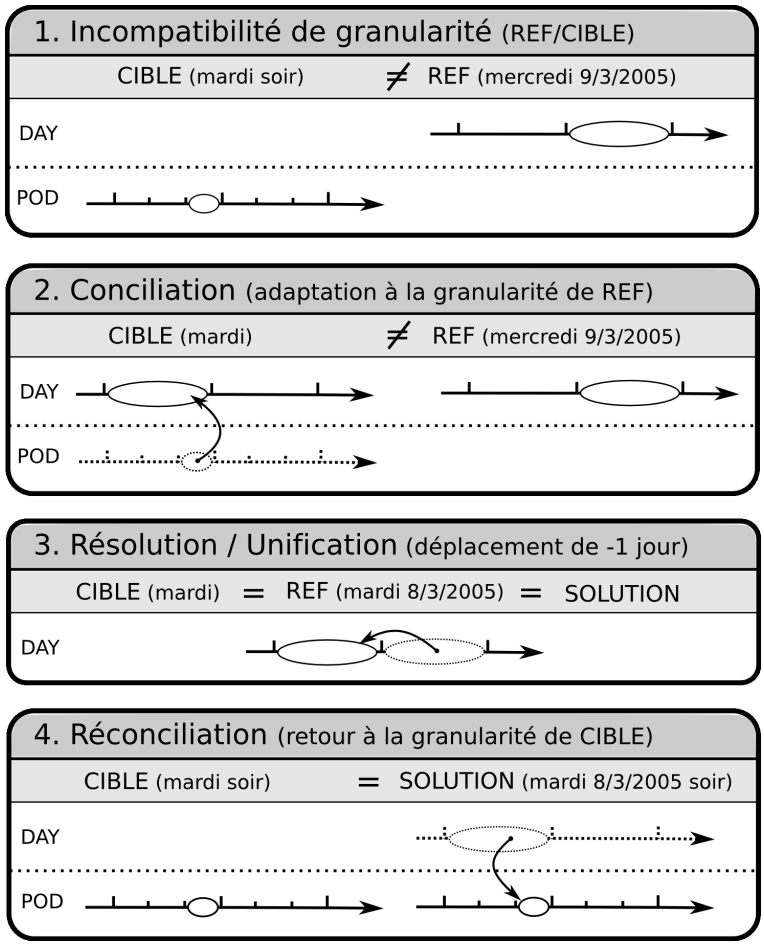


Figure 7.11 : Mécanisme « Conciliation / Résolution / Réconciliation » (CRR)

Une fois le déplacement calculé, la phase de *réconciliation* permet, lorsqu'une diminution de granularité a été effectuée, de revenir à la granularité de départ, mais en l'accompagnant d'un indicateur d'imprécision de type 2 (externe) à cette solution. Si l'adaptation de granularité est une augmentation, la réconciliation n'est pas effectuée et le niveau de granularité reste inchangé. Un indicateur d'imprécision de type 1 (interne) est cependant attribué à la solution.

Déplacement temporel avec spécification partielle de la cible

Le cas précédent peut parfois apparaître sous une forme quelque peu différente. En effet, un déplacement temporel peut être accompagné d'une information complémentaire sur la cible à atteindre. Lorsque c'est le cas, une version un peu particulière du mécanisme « Conciliation / Déplacement / Réconciliation » peut être mise en œuvre, comme cela est illustré à la figure 7.13 (CDR_cible).

Les premières étapes sont identiques au mécanisme CDR déjà exposé : dans le cas d'une incompatibilité de granularité entre le point de référence et l'unité de mesure temporelle, une *conciliation* est effectuée par l'adaptation de la granularité du point de référence vers un niveau plus ou moins élevé. Le déplacement est alors réalisé[52].

[52] Ces étapes ne sont pas détaillées dans la figure 7.13, mais sont similaires à celles présentées à la figure 7.12.

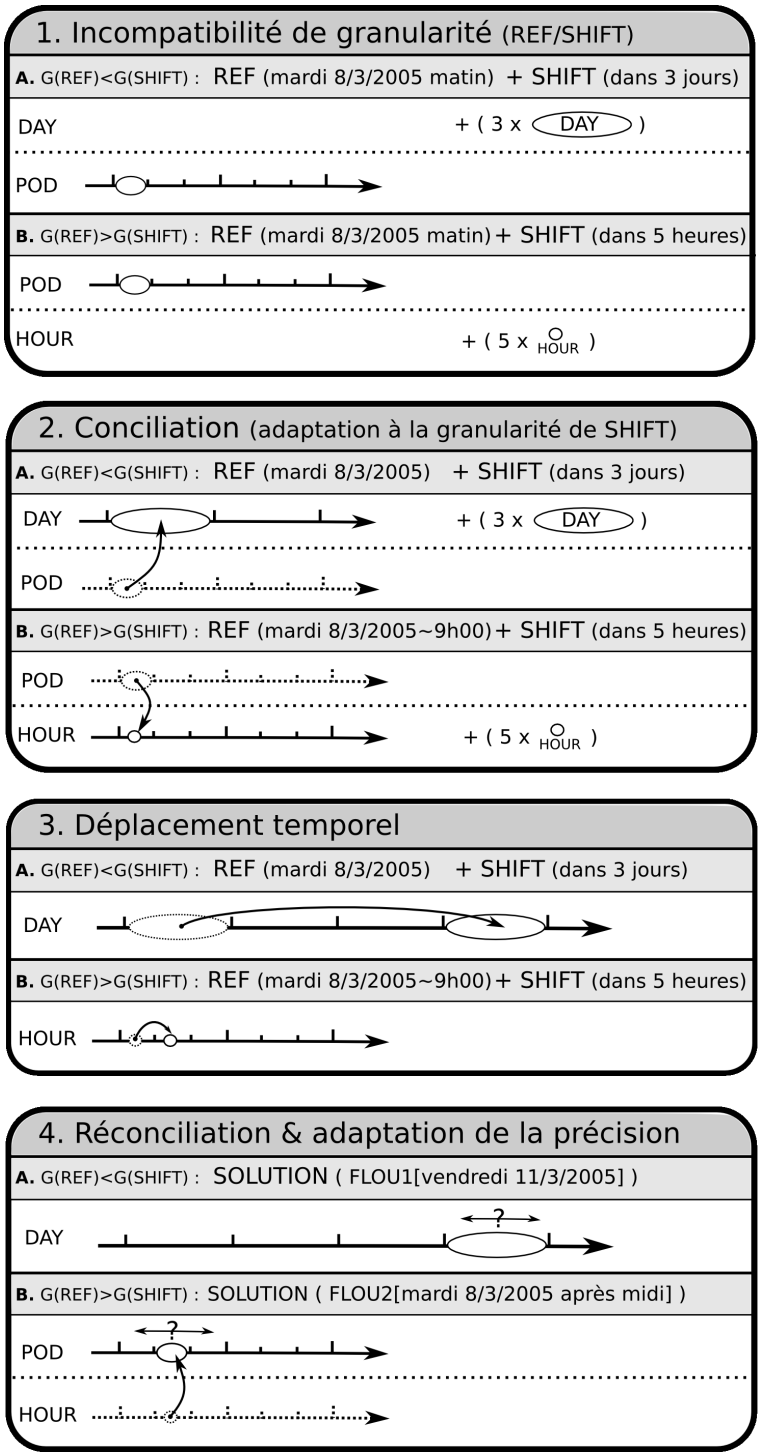


Figure 7.12 : Mécanisme « Conciliation / Déplacement / Réconciliation » (CDR)

L'originalité provient de l'étape de *réconciliation* avec la granularité d'origine de l'expression. Celle-ci comporte deux parties possibles (4a et 4b). La première consiste à compléter la solution intermédiaire, trouvée suite au déplacement temporel, avec les informations de l'expression de départ qui caractérisent la cible. Le résultat est évalué de manière à vérifier s'il remplit bien les critères d'une zone temporelle *bien identifiée*, et s'il ne constitue pas une aberration (par exemple, un « 30 février », ou une date *jours/mois/année* qui ne correspond pas au nom du jour). Si la solution est satisfaisante

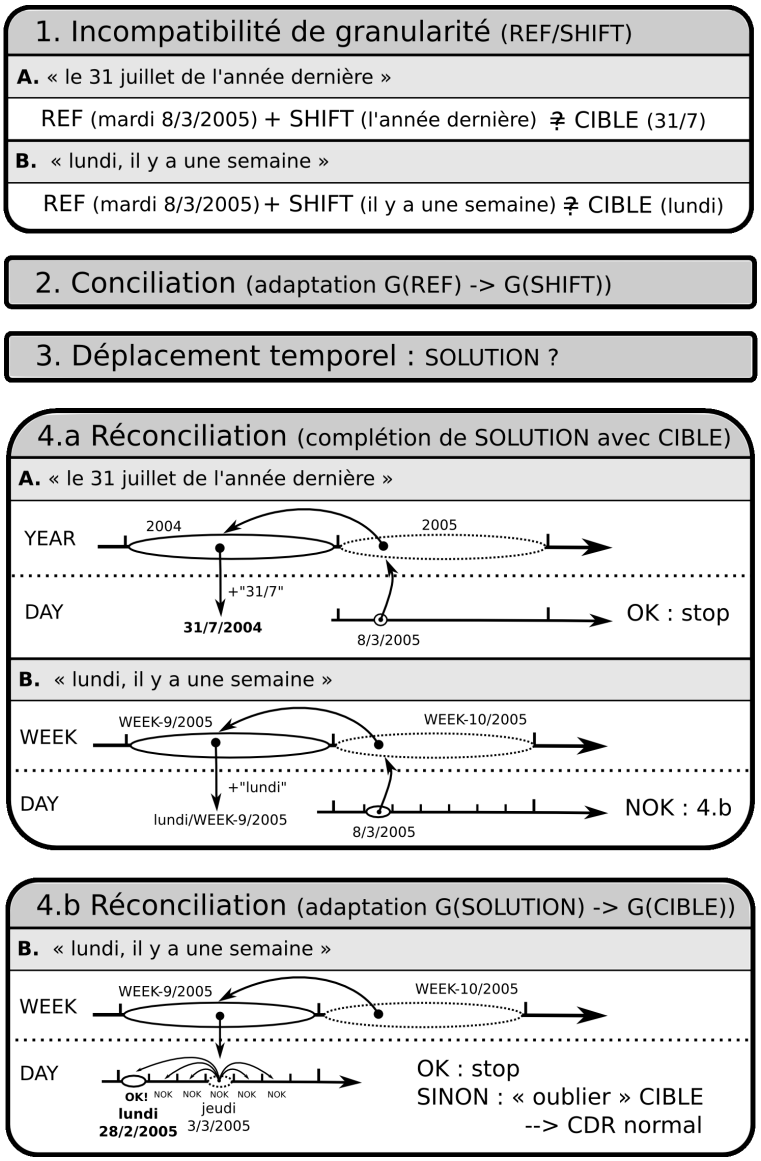


Figure 7.13 : Mécanisme « Conciliation / Déplacement / Réconciliation » avec spécification partielle de la cible (CDR_cible)

elle est conservée telle quelle. Dans le cas contraire, la deuxième partie de la réconciliation est exécutée. Celle-ci consiste en une adaptation de la granularité selon les mêmes principes que pour l'étape de conciliation. À partir du point temporel obtenu, à la granularité souhaitée, une vérification des contraintes posées par la spécification de la cible dans l'expression de départ est effectuée. Si ces contraintes ne sont pas rencontrées, un balayage du contexte direct, passé et futur, est enclenché, jusqu'à obtention d'un point satisfaisant. Une limite sur l'amplitude de ce balayage est cependant fixée, dans ce cas, à 3 unités temporelles (dans les deux sens). Si aucune solution ne devait être trouvée, il est alors décidé de reprendre l'interprétation à zéro en procédant à un déplacement temporel classique, qui accompagnera le résultat d'un indicateur d'imprécision adéquat.

Déplacement temporel avec spécification du point de référence

Une seconde variation du déplacement temporel peut encore être rencontrée, lorsqu'en plus de celui-ci, est spécifié de manière explicite le point de référence (par exemple, « trois jours après lundi »). L'interprétation de ce cas ne varie pas énormément de celle exposée pour le cas classique (CDR). La seule différence réside dans le fait qu'il est nécessaire, en tout premier lieu, d'effectuer la résolution du point de référence en question avant de réaliser le déplacement temporel sur cette base. Cette première étape sera, selon la nature exacte du point de référence, semblable à celle menée pour une expression de type PA*U (section 7.8.5) ou pour une expression temporelle *sous-spécifiée*.

Déplacement temporel indéterminé

Enfin, le cas particulier d'un déplacement temporel indéterminé (« il y a quelques semaines ») ne rentre pas dans les mécanismes de traitement exposés jusqu'ici. En réalité, il n'est pas vraiment possible d'interpréter de manière très précise ce genre d'expression. Dès lors, la valeur temporelle renvoyée par le système sera exprimée au moyen du point de référence accompagné de l'unité temporelle et du sens relatifs au déplacement selon le format *Ref*(+|-)*Unit*S[53]. Cette manière de procéder permet à la fois de respecter l'imprécision naturelle de l'expression, tout en fournissant une valeur normalisée plus facilement exploitable par la suite. Évidemment, cette valeur est accompagnée de l'indicateur d'imprécision « 2 » (externe).

Prise en compte du caractère flou

La prise en compte du caractère flou de l'expression est à nouveau repris par les champs $fuzzy$ et $partof$. Hormis l'attribution d'une valeur à $fuzzy$ lors des processus d'interprétation exposés précédemment, l'imprécision peut également provenir d'une formulation explicite dans l'expression de départ. Lorsque celle-ci est *sous-spécifiée*, il peut s'agir de formulations telles que « vers » ou « au début ». D'autre part, l'imprécision peut aussi porter sur l'amplitude d'un déplacement temporel (« dans environ 5 jours »). Enfin, les déplacements temporels qui utilisent des unités de mesure d'une forte granularité (plus grandes que le *jour*) sont automatiquement considérées comme imprécises, à moins qu'une marque de précision explicite ne vienne démentir cette règle.

7.8.7 Traitement des expressions D**U

Les expressions qui prennent la forme d'un intervalle ne sont pas traitées de manière très différente des autres expressions rencontrées jusqu'ici. Elles possèdent évidemment la particularité d'être composées par deux bornes, dont une peut éventuellement ne pas être exprimée. Dans ce cas, celle-ci sera laissée vide. Pour chaque borne explicite, une interprétation est menée, selon le mécanisme approprié à son type (expression de type PA*U ou PR*U).

[53] Avec *Ref*, la valeur temporelle du point de référence ; + ou -, le sens du déplacement temporel ; et *Unit*S, l'unité de mesure temporelle utilisée (complétée par un « S »). Par exemple : 29/12/2009-WEEKS.

Les intervalles temporels présentent cependant tout de même quelques particularités. En ce qui concerne le caractère imprécis, celui-ci peut se porter sur les bornes (« depuis environ jeudi »), mais peut également être présent au niveau de l'intervalle (« (durant | à la fin de) la période 2001-2003 »). Les champs $fuzzy$ et $partof$ sont par conséquent présents à la fois au niveau de l'intervalle et des bornes. Ces différentes informations sont fournies par l'annotation réalisée par les grammaires. Au niveau des bornes, les champs d'imprécision peuvent éventuellement être mis à jour en fonction du résultat de l'interprétation.

Enfin, la présence d'intervalles dans le système complique quelque peu l'évaluation des expressions relatives anaphoriques. Le problème se pose plus précisément lorsque le point de référence utilisé pour le mécanisme de résolution se trouve être un intervalle. Ce point est abordé plus complètement à la section suivante (7.8.8).

7.8.8 Le problème de la gestion du point de référence contextuel

La description des mécanismes d'interprétation des expressions temporelles de type PR*U met en avant un problème important et complexe à gérer. Il s'agit du choix du point de référence à partir duquel s'opère la recherche de la solution. Il arrive que celui-ci soit explicitement mentionné, mais ce n'est pas toujours le cas. Les annotations effectuées à l'aide des grammaires fournissent alors une indication qui permet de déterminer si le point à sélectionner est la date d'émission du texte (référence déictique) ou un point de repère issu du contexte (référence anaphorique). La référence déictique est gérée de manière assez aisée dès l'instant où la date d'émission du texte est fournie en tant que métadonnée. Par contre, le point de référence anaphorique est moins facilement identifié.

La première question qui se pose, dans le cadre d'une référence anaphorique, est de déterminer quel point sélectionner pour servir de départ à l'interprétation. Pour notre approche de ce problème, une solution simple a été adoptée. Elle consiste à choisir la dernière zone temporelle *bien identifiée* qui a été rencontrée par le système dans l'ordre normal du texte. Une seconde question se pose alors. Elle concerne la portée de la validité de cette référence. L'hésitation concerne le fait de savoir s'il faut limiter d'une manière ou d'une autre le contexte dans lequel le point de référence peut être choisi. La limite peut être déterminée au moyen d'une fenêtre d'un certain nombre de mots, en respectant la limite de la phrase, celle du paragraphe ou encore en élargissant celle-ci à tout le texte. Pour cette implémentation, la limite a été laissée au niveau du texte entier. Cette solution est également celle qui a été appliquée par Mani et Wilson [2000] dans le cadre de leur système de caractérisation temporelle des événements. Leur analyse a cependant montré qu'une erreur pouvait alors être facilement propagée aux événements voisins. Pour cette raison, Filatova et Hovy [2001] ont plutôt choisi de réinitialiser le contexte temporel à chaque paragraphe. Cette approche pourrait également être adoptée dans notre système, le point de référence potentiel serait alors réinitialisé à l'aide de la date d'émission du texte à chaque nouveau paragraphe. Cette valeur est ensuite remplacée par celles attribuées aux différentes expressions temporelles rencontrées dans la suite du paragraphe. Cette solution n'a cependant pas été choisie car, sans mécanisme supplémentaire, elle empêche les références inter-paragraphes. L'effet négatif de la conservation du point de référence devra cependant

être surveillé lors de l'évaluation du système.

Dans ce système, le cas des intervalles est un peu particulier. Il y a deux aspects à examiner. Tout d'abord, lorsque un intervalle constitue le point de référence anaphorique pour l'interprétation d'une expression temporelle relative. Dans cette situation, une des deux bornes est sélectionnée pour remplacer l'intervalle. Si la direction de recherche est négative (vers le passé), c'est la borne supérieure qui est choisie. Dans le cas contraire d'une direction de recherche positive (vers le futur), la borne inférieure l'emportera. Dans l'hypothèse où la borne désirée ne serait pas *bien identifiée*, l'autre borne la remplace alors. La seconde situation particulière concerne l'interprétation des expressions relatives lorsqu'elles se situent dans le cadre d'un intervalle (bornes). Ce cas de figure est géré de manière tout à fait habituelle, en fournissant aux bornes qui en ont besoin le (même) point de référence qui a été déterminé au préalable.

7.8.9 Prise en compte des expressions cadratives

L'exploitation des expressions cadratives correspond à une mise en pratique de la théorie des cadres de discours initiée par Charolles [1997], dont il a déjà été question aux sections 4.9 et 7.3. En pratique, pour être considérée comme cadrative, l'expression temporelle doit être située en début de paragraphe et être suivie par un signe de ponctuation. Elle doit aussi pouvoir être considérée comme un point de référence valide, c'est-à-dire qui est *bien identifié* dans l'espace temporel. Le cadre de discours est toujours interrompu par la fin du paragraphe. Ces critères peuvent bien entendu être affinés si nécessaire, mais cela demande des moyens d'analyse supplémentaires.

L'utilisation des informations en provenance de l'expression cadrative peut servir à interpréter les expressions *sous-spécifiées* sans même avoir recours aux mécanismes habituels. En effet, lorsque les circonstances le permettent, la cible dont la spécification est partielle est complétée à l'aide de l'expression cadrative. Pour cela, il faut que le niveau de granularité de celle-ci soit plus élevé que celui de la cible à interpréter. Pour être approuvée comme solution, la représentation complétée de l'expression subit une validation qui permet de vérifier qu'elle désigne bien une zone temporelle *bien identifiée*.

L'exemple ci-dessous illustre le type de situation dans lequel les cadres de discours sont efficaces. Le cadre est constitué par la mention de l'année, les différentes expressions qui désignent des mois sont interprétées comme se situant dans cette année-là.

> « En **2003**, le Costa-Rica avait déjà subi de nombreuses catastrophes naturelles. Un tremblement de terre avait eu lieu en mars, des inondations en juin et une tempête durant le mois de septembre. »

7.8.10 Comparaison par rapport à l'existant

Si on se réfère aux systèmes ayant un but similaire qui ont été cités à la section 6.6, plusieurs présentent des points communs avec les mécanismes que nous proposons.

Tout d'abord, en ce qui concerne le résultat de la phase d'interprétation et de normalisation, deux tendances se dégagent. La première consiste à fournir une représentation sous une forme *mathématique*, qui code les opérations nécessaires pour passer du point de référence au point exprimé par l'expression temporelle interprétée. Cette approche est par exemple adoptée par Maurel et Mohri [1994] ou Battistelli *et al.* [2006]. La seconde approche, plus fréquente, avec entre autres Mani et Wilson [2000], Filatova et Hovy [2001] ou Ahn *et al.* [2007], a pour objectif de fournir une valeur temporelle classique. Nous nous situons également dans cette catégorie.

Pour atteindre leur objectif, la majorité des systèmes exploitent, comme nous, les informations fournies par les temps verbaux.

Parmi ces différents travaux, celui qui expose le plus complètement son mécanisme d'interprétation, et qui semble être le plus proche du nôtre est Filatova et Hovy [2001] (leur processus est décrit succinctement à la section 6.6). Cependant, le mécanisme d'adaptation des granularités et de gestion de l'imprécision tel que nous le proposons n'y apparaît pas. Le jeu des granularités est par contre présent dans le système d'interprétation temporelle de Battistelli *et al.* [2006], ainsi que chez Han et Lavie [2004].

La détection et l'exploitation des cadres de discours temporels ne sont pas très souvent citées non plus. Ce principe est cependant utilisé par Battistelli *et al.* [2006] dans le cadre d'un système d'aide à la lecture de textes biographiques.

Nous n'avons pas trouvé de système reprenant simultanément les caractéristiques proposées par notre approche, c'est-à-dire principalement une large couverture des différents types d'expressions temporelles[54], l'interprétation des temps verbaux, la prise en compte des niveaux de granularités et de l'imprécision, ainsi que des cadres de discours.

7.9 Évaluation

L'évaluation a été conduite en deux grandes phases. Premièrement, une estimation de la performance, du repérage et de la reconnaissance des expressions et de leur catégorie a été menée. Dans un second temps, c'est l'exactitude de l'interprétation et de la normalisation des expressions retrouvées qui a été évaluée. Entre ces deux grandes phases, une mise à jour des grammaires a été effectuée afin que les oublis ou erreurs de repérage aient le moins d'impact possible sur la deuxième partie de l'évaluation.

7.9.1 Première partie : le repérage des expressions et la reconnaissance de leurs catégories

Cette première étape de l'évaluation concerne donc la délimitation à l'intérieur des textes des expressions reprises dans la spécification (Section 7.6.2). La validité du code de catégorie (par exemple PRPU) est également vérifié. L'évaluation a été réalisée par un évaluateur extérieur, dont le travail a

[54] À l'exclusion toutefois des horaires, fréquences et expressions d'ensembles en général.

ensuite été vérifié.

Constitution des corpus d'évaluation

L'analyse des résultats a été conduite sur deux corpus différents. Tout d'abord, sur un corpus *News* composé de dépêches de presse en français provenant de l'agence Belga. Le corpus a été constitué en choisissant au hasard 365 dépêches de l'année 2005, de façon à en avoir une par jour. Les dépêches de l'agence Belga ayant en partie servi comme corpus d'*apprentissage* pour la construction manuelle des transducteurs, il est important de signaler que le corpus de *test* est constitué de textes qui n'ont jamais été exploités auparavant (l'année 2005 n'était pas concernée par le corpus d'*apprentissage*).

Afin de s'affranchir complètement de la possible influence des dépêches de presse, en tant que type de texte, sur l'élaboration des transducteurs, un second corpus a été utilisé. Celui-ci est constitué de 365 textes pris au hasard dans le corpus *Parlementaire* présenté à la section 2.5.2, dans le cadre de l'indexation thématique. Il s'agit d'un ensemble de textes relevant du domaine législatif et parlementaire. Ceux-ci présentent donc un style complètement différent du premier corpus.

Procédure

L'évaluateur a reçu l'ensemble des textes annotés au format défini par la spécification des expressions temporelles (Section 7.6.2). L'annotation a été réalisée automatiquement au moyen du système présenté dans ce chapitre (voir section 7.8 en particulier). D'autre part, un formulaire a été fourni afin qu'il puisse reporter ses observations. Ce formulaire, une feuille de calcul de type Excel, dispose d'une ligne par document à évaluer. Elle reprend un certain nombre de colonnes afin de caractériser la reconnaissance des expressions temporelles. Il est donc demandé d'inscrire, pour chaque texte, le nombre total d'expressions qui correspondent aux cas représentés par les différentes colonnes[55]. En ce qui concerne le repérage proprement dit, six colonnes sont prévues pour les différents niveaux de reconnaissance.

1. *OK* : les expressions correctement reconnues ;
2. *Frac* : les expressions complètement reconnues, mais en plusieurs parties ;
3. *Trop* : les expressions complètement reconnues, mais qui sont accompagnées d'éléments supplémentaires, qui ne sont pas pertinents ;
4. *Partiel* : les expressions partiellement reconnues, dont un ou plusieurs éléments manquent ;
5. *A tort* : les expressions reconnues et qui n'auraient pas dû l'être ;
6. *Manqué* : les expressions non reconnues, mais qui auraient dû l'être.

Pour le typage des expressions reconnues, c'est-à-dire celles qui rentrent dans les catégories 1 à 4 ci-dessus, c'est la proximité avec le code correct qui est vérifiée. Pour rappel, un code est com-

[55] Lorsque le total pour une colonne est 0, celle-ci peut évidemment être laissée vide pour faciliter la tâche de l'évaluateur.

posé de quatre caractères, qui représentent chacun une caractéristique d'égale importance[56] (voir section 7.6.1). Par conséquent, un code attribué à une expression possède entre quatre et zéro[57] composantes correctes. Cinq colonnes ont donc été définies à cet effet, allant du code correct au code incorrect : *C4ok*, *C3ok*, *C2ok*, *C1ok*, et enfin *Cnok*. Ces catégories ne distinguent pas les erreurs en fonction de la partie incorrecte du code (*C3ok* reprend indifféremment les écarts de un caractère imputables aux caractéristiques 1, 2 ou 3).

Les durées et les âges sont évalués séparément. Un ensemble de colonnes sont donc prévues spécialement pour ces deux types d'expressions. Pour la reconnaissance, elles sont identiques à celles définies pour les autres expressions temporelles (six colonnes, de *OK* à *Manqué*). En ce qui concerne le typage, trois colonnes sont proposées.

1. *OK* : pour les expressions correctement typées ;

2. *Nok(Dur/Age)* : lorsque le typage est incorrect, l'erreur provenant d'une confusion entre une durée et un âge (ou inversement) ;

3. *Nok(ExpTemp)* : lorsque le typage est incorrect, l'erreur provenant d'une confusion entre une durée ou un un âge et une autre expression temporelle.

La séparation de l'évaluation entre les durées et âges et le reste des expressions est motivée par le fait que seules ces dernières seront sujettes à interprétation, et donneront donc un ancrage dans l'espace du temps (voir section 7.6.1). Dans un souci de clarté dans le commentaire des résultats, ce dernier groupe sera désigné sous le nom d'*expressions temporelles HDA*[58].

Finalement, une ultime colonne est prévue pour mentionner la *validité* du texte. Comme les textes sont choisis au hasard, et que les collections de documents *News* et *Parlementaire* contiennent des textes qui peuvent être *non standards*, il est prévu que l'évaluateur puisse rejeter le document en question lorsque le cas se présente[59]. Plus précisément, un texte peut être écarté lorsqu'il :

- ne contient pas d'expression temporelle, qu'elle soit annotée[60] ou non, à l'exclusion de la date d'émission du document qui est toujours présente ;
- n'est pas écrit en français ou s'il est multilingue ;
- s'agit de texte *structuré*, non standard, comme des résultats sportifs, ou des cours de bourse, sous la forme de listes ou de tableaux ;
- a manifestement subi des erreurs lors de l'extraction de la base documentaire.

[56] Il n'y a pas de notion d'ordre entre les différentes caractéristiques. Il n'y a donc pas de caractéristique principale ou secondaire.

[57] Ce cas correspond à la confusion d'une expression de type P*** ou D*** avec une durée ou un âge.

[58] HDA : Hors Durées/Âges.

[59] Une exclusion préalable aurait pu être effectuée afin que le corpus réellement évalué soit maximal. En pratique, la masse de documents à évaluer s'est révélée assez, voire trop, conséquente par rapport à l'investissement humain disponible dans le cadre de cette évaluation. La situation de pénurie ne s'est donc jamais présentée.

[60] À tort ou à raison.

Mesures

Les mesures de *rappel* (R), *précision* (P) et *f-mesure* (F_1, noté F) sont calculés selon les formules suivantes :

$$R = \frac{OK}{(OK + Fract + Trop + Partiel + Manque)}$$

$$P = \frac{OK}{(OK + Fract + Trop + Partiel + Atort)}$$

$$F = \frac{2 * R * P}{R + P}$$

Résultats

Pour les deux corpus, les résultats concernant la reconnaissance et le typage des expressions sont détaillés.

Corpus *News*

Pour le corpus *News*, sur les 365 textes proposés, 70 ont été écartés, l'évaluation s'effectuant donc finalement sur 295 documents. Les résultats obtenus pour la reconnaissance et le typage des expressions temporelles HDA – c'est à dire à l'exclusion des durées et des âges – sont repris aux tableaux 7.2 et 7.3. Pour l'étape de reconnaissance, le rappel s'établit à 90,49%, pour une précision de 97,74%, soit une f-mesure de 93,98. Le typage des expressions reconnues est correct à 98,69%.

	OK	Frac	Trop	Partiel	A tort	TOTAL	Manqué
Nombre d'expressions	1.646	2	4	27	5	1.684	140

Tableau 7.2 : Évaluation de la reconnaissance des expressions temporelles (HDA) sur le corpus News.

	C4ok	C3ok	C2ok	C1ok	Cnok	TOTAL
Nombre d'expressions	1.656	17	1	3	1	1.678
Importance relative (%)	98,69	1,01	0,06	0,18	0,06	100

Tableau 7.3 : Évaluation du typage des expressions temporelles (HDA) sur le corpus News.

Toujours pour le corpus *News*, en ce qui concerne les durées et les âges, les résultats sont repris aux tableaux 7.4 et 7.5. Pour la reconnaissance, le rappel atteint 92,45% alors que la précision est de 96,08%, la f-mesure s'établit donc à 94,23. La précision du typage est de 92,16%.

	OK	Frac	Trop	Partiel	A tort	TOTAL	Manqué
Nombre d'expressions	294	0	2	10	0	306	12

Tableau 7.4 : Évaluation de la reconnaissance des durées et âges sur le corpus News.

Si l'ensemble des expressions temporelles (durées et âges compris) est évalué globalement, le rappel obtenu est de 90,78%. Il est accompagné d'une précision de 97,49%. La f-mesure atteint elle 94,02.

	OK	Nok(Age/Dur)	Nok(ExpTemp)	TOTAL
Nombre d'expressions	282	17	7	306
Importance relative (%)	92,16	5,67	2,29	100

Tableau 7.5 : Évaluation du typage des durées et âges temporelles sur le corpus News.

L'analyse de ces résultats fait ressortir très clairement la principale faiblesse du système. Celle-ci réside dans la non-reconnaissance de certaines formes d'expressions temporelles HDA (140, soit un handicap de 8,34% pour le rappel HDA), et dans une moindre mesure d'expressions de durées ou d'âges (12, soit 3,92% de perdu pour la mesure du rappel relatif aux durées/âges). Une analyse des oublis a par conséquent été menée. Il est apparu que sur les 140 expressions HDA manquées, 66 (soit 47%) incluaient la mention d'une année. Les différentes formes manquées apparaissent :

- seules ou dans une énumération (31, soit 22%) ;
- accompagnées d'un événement (25, soit 18%), soit de très grande notoriété (« Jeux Olympiques 2008 »), soit exprimé sous une forme assez régulière (« Journée internationale de *** 2008 ») ;
- ou encore dans le cadre d'un intervalle de deux années (10, soit 7%) (« 1963-64 »).

D'autres points faibles ont également été identifiés. Les intervalles imprécis (« depuis le début de l'année », « depuis plusieurs mois ») représentaient 18 oublis (soit 13%), et diverses énumérations (autres que des années) impliquaient 10 expressions (soit 7%). Ces quelques cas regroupent à eux seuls deux tiers des expressions HDA manquantes.

Au total, toutes expressions temporelles confondues (dont les durées et les âges), et tous types d'erreurs rassemblés (oublis, reconnaissances partielles, trop larges ou fractionnées), une rapide révision des transducteurs a permis de corriger 139 erreurs sur 197. Ces corrections pourraient potentiellement permettre d'améliorer le rappel jusqu'à environ 97%. La vérification de l'impact réel de ces améliorations n'a cependant pas pu être précisément mesurée sur un nouveau corpus. Les performances réelles devraient dès lors probablement se situer entre les deux valeurs de rappel citées (90,78% - 97%).

Corpus *Parlementaire*

La même procédure d'évaluation, à l'aide des versions originales des transducteurs, a également été conduite sur le second corpus, composé de textes *Parlementaires*. Ces textes étant parfois très longs, l'ensemble des 365 documents initialement choisis au hasard n'a pas pu être vérifié. L'évaluation a porté sur les 101 premiers textes, parmi lesquels 16 ont été écartés, fixant ainsi finalement le nombre de documents évalués à 85. Les résultats obtenus pour la reconnaissance et le typage des expressions temporelles HDA sont présentés aux tableaux 7.6 et 7.7. Pour la reconnaissance, le rappel de 78,70% est plus faible que celui observé lors de l'évaluation du corpus *News*. La précision, en s'établissant à 97,03%, conserve cependant un très bon niveau, ce qui permet d'obtenir une f-mesure de 87,14. La précision du typage pour les expressions HDA correctement reconnues reste lui aussi assez élevé (97,61%).

En ce qui concerne les durées et les âges qui apparaissent dans le corpus *Parlementaire*, les résul-

	OK	Frac	Trop	Partiel	A tort	TOTAL	Manqué
Nombre d'expressions	1.308	0	0	32	8	1.348	314

Tableau 7.6 : Évaluation de la reconnaissance des expressions temporelles HDA sur le corpus Parlementaire.

	C4ok	C3ok	C2ok	C1ok	Cnok	TOTAL
Nombre d'expressions	1.308	29	2	0	1	1.340
Importance relative (%)	97,61	2,16	0,15	0	0,07	100

Tableau 7.7 : Évaluation du typage des expressions temporelles HDA sur le corpus Parlementaire.

tats sont repris dans les tableaux 7.8 et 7.9. Pour la reconnaissance, les résultats obtenus sont du même ordre que pour le corpus *News*, le rappel s'établissant à 98,33%, et la précision à 98,66%. La combinaison de ces deux mesures a permis à la f-mesure d'atteindre 98,50.

	OK	Frac	Trop	Partiel	A tort	TOTAL	Manqué
Nombre d'expressions	295	0	0	2	2	299	3

Tableau 7.8 : Évaluation de la reconnaissance des durées et âges sur le corpus Parlementaire.

	OK	Nok(Age/Dur)	Nok(ExpTemp)	TOTAL
Nombre d'expressions	278	21	0	299
Importance relative (%)	92,98	7,02	0	100

Tableau 7.9 : Évaluation du typage des durées et âges sur le corpus Parlementaire.

Les performances globales, tous types d'expressions temporelles confondues s'établissent à 82,04% pour le rappel, à 97,33% pour la précision et à 89,03 en ce qui concerne la f-mesure.

Le fait important qui est ressorti de l'évaluation menée sur ce deuxième corpus est la chute du rappel pour les expressions temporelles HDA. Une analyse détaillée a donc été réalisée pour trouver l'origine de celle-ci. À nouveau, la reconnaissance des années, seules ou sous forme d'intervalles, a été un élément important (157 expressions concernées). Cependant, l'impact le plus important a été provoqué par un format de date *jour-mois-année* particulier, du type « jj.mm.aaaa ». L'absence de cette forme *pointée* est à elle seule responsable de la non-reconnaissance de 149 expressions. Après correction des transducteurs, 279 expressions temporelles HDA additionnelles ont pu être correctement reconnues. L'effet potentiel de ces améliorations sur le rappel de ces expressions permettrait d'augmenter celui-ci à environ 96%[61], ce qui serait plus en rapport avec les performances précédemment obtenues sur le corpus *News*. Cette amélioration n'a cependant pas pu être vérifiée sur de nouveaux textes.

La comparaison de ces résultats avec ceux obtenus par d'autres systèmes est délicate. De nombreux paramètres peuvent venir influencer les mesures dans un sens ou dans un autre, entre autres : la couverture du domaine assurée par le système[62], la définition des catégories d'évaluation (expressions reconnues de manière stricte, partielle, etc.), la définition des mesures, etc. La comparaison peut néanmoins être réalisée en ne perdant pas de vue ces restrictions. Les différentes évaluations

[61] De même pour le rappel global, toutes expressions temporelles confondues.

[62] En comparaison avec un système qui tend à maximiser la reconnaissance des différentes formes, une extraction qui se concentre sur quelques cas particuliers aura plus de facilité à atteindre des valeurs de rappel et de précision très élevées.

citées à la section 6.6, entre autres Mani et Wilson [2000], Schilder et Habel [2001], Vazov [2001], Vicente-Díez *et al.* [2008], Parent *et al.* [2008] et Bittar [2009], font état de mesures de rappel qui s'échelonnent de 79% à 95%, et de valeurs de précision allant de 83% à 87,30%. Le rappel le plus élevé est dû à Vazov [2001] (95%) et est accompagné d'une précision de 85%. La précision la plus élevée a été relevée chez Schilder et Habel [2001] (87,30%) et est assortie d'un rappel de 90,66%.

7.9.2 Deuxième partie : l'interprétation des expressions temporelles

Cette deuxième partie de l'évaluation concerne l'étape d'interprétation et de normalisation des expressions temporelles reconnues. Afin de minimiser l'impact des expressions non reconnues sur l'évaluation de l'interprétation temporelle des textes, la version améliorée des transducteurs a été utilisée.

Corpus d'évaluation

L'évaluation n'a été menée que sur le corpus *News*. La raison est multiple. Il s'agit tout d'abord d'une question de faisabilité, le temps nécessaire à une vérification rigoureuse, systématique et méthodique étant important. D'autre part, le corpus *Parlementaire* a tendance à s'éloigner plus rapidement d'un texte standard (abréviations, mise en page particulière, etc.), par exemple dans le cas de textes de loi. De plus, des éléments gênants, tels que des phrases incomplètes, des erreurs d'OCR ou des textes parfois multilingues, ont été observés. D'une manière générale les caractéristiques de ces textes se prêtaient un peu moins bien que les dépêches au type d'analyse que nous mettons en place et qu'il fallait évaluer. Lors de la première étape de l'évaluation (sur la reconnaissance des expressions), l'utilisation d'un deuxième corpus était principalement motivée par la nécessité de déceler une éventuelle influence du corpus ayant servi à construire les grammaires – des dépêches de presse similaires à celles présentes dans le corpus *News* – sur les performances. Ce type de problème ne se pose pas à cette étape.

Procédure

À nouveau, l'évaluateur a reçu le corpus *News* annoté, cette fois dans un format utilisant des balises de type SGML pour délimiter les expressions et pour en spécifier la valeur interprétée et normalisée. L'évaluation proprement dite est, comme pour la première phase, consignée dans une feuille de calcul. Celle-ci propose une expression temporelle par ligne. Les expressions apparaissent selon leur ordre d'apparition dans le corpus. Dans les différentes colonnes proposées pour caractériser la manière dont l'expression a été interprétée, une seule doit être cochée (remplie avec la valeur « 1 »). Ces colonnes correspondent à une évaluation correcte (*OK*), incorrecte (*NOK*), ou intermédiaire (*+/-*)[63] et cela pour les huit types d'expressions temporelles HDA. À ces 24 colonnes vient à nouveau

[63] Pour cette catégorie, la zone temporelle désignée est adéquate, mais un problème, par exemple de granularité ou de marque d'imprécision, est présent. Il peut également s'agir d'expressions temporelles pour lesquelles le point de référence est un événement dont la localisation temporelle n'est pas explicitement donnée (« huit jours avant les élections »). Ce type

s'ajouter celle qui permet de rejeter le texte[64].

Résultats

Les jugements de l'évaluateur, matérialisés par le chiffre « 1 » qui a été disposé dans une des 24 colonnes possibles, a fait l'objet d'une somme qui permet d'évaluer, d'abord par type d'expressions temporelles, puis globalement, la précision de la tâche d'interprétation et de normalisation.

	OK		+/-		NOK		Total	
	Nb.	%	Nb.	%	Nb.	%	Nb.	%
PAPU	147	100	0	0	0	0	147	8,24
PAFU	151	100	0	0	0	0	151	8,46
PRPU	995	89,24	11	0,99	109	9,78	1.115	62,46
PRFU	119	87,50	7	5,15	10	7,35	136	7,62
DAPU	74	100	0	0	0	0	74	4,15
DAFU	12	100	0	0	0	0	12	0,67
DRPU	43	64,18	3	4,48	21	31,34	67	3,75
DRFU	79	95,18	0	0	4	4,82	83	4,65
Total	1.620	**90,76**	21	**1,18**	144	**8,07**	1.785	100

Tableau 7.10 : Évaluation de la précision de l'interprétation et de la normalisation des expressions temporelles (HDA) du corpus News.

Une première constatation concerne la distribution des expressions sur ce corpus. La majorité des expressions appartiennent à la catégorie PRPU[65] (62,46%), suivie de loin par PAFU[66] (8,46%), PAPU[67] (8,24%) et PRFU[68] (7,62%). Il est vrai que la catégorie PRPU cache de nombreux cas différents – sous-spécification, déplacement temporel, adverbes déictiques ou anaphoriques, etc. – et qu'il s'agit d'expressions assez fréquentes. En effet, une fois le contexte temporel bien établi, il est plus simple pour le scripteur d'utiliser ce genre d'expressions, plus courtes et plus variées qu'une date complète.

La précision atteinte dépasse le seuil des 90%. À nouveau, il est très délicat de comparer ce résultat avec ceux obtenus par d'autres systèmes, tant leurs caractéristiques peuvent être différentes. Cependant, si on se rapporte aux résultats relevés à la section 6.6, on constate que, pour la tâche d'interprétation des expressions temporelles, Filatova et Hovy [2001] obtiennent une précision qui varie entre 77,85% et 82,29%. Schilder et Habel [2001] mentionnent pour leur part une précision de

d'expression n'est pas pris en compte par le système, qui repère néanmoins la partie « huit jours avant ». Il peut également s'agir d'expressions repérées mais dont l'interprétation n'a pas encore été complètement implémentée.

[64] Les critères sont les mêmes que pour la phase précédente de l'évaluation.

[65] Expressions Ponctuelles Relatives Précises Uniques (voir section 7.6.1). Par exemple : une date sous-spécifiée (« jeudi »), un déplacement temporel explicite (« il y a un an ») ou implicite (« la veille », « la semaine dernière »). Voir aussi à la section 7.6.2.

[66] Expressions Ponctuelles Absolues Floues Uniques (voir section 7.6.1). Par exemple : « 22/12/2009 vers 18h24 », « aux environs de l'an 2000 » (voir section 7.6.2).

[67] Expressions Ponctuelles Absolues Précises Uniques (voir section 7.6.1). Par exemple : « 2005 », « le 12 juillet 2007 », « Noël 2010 ». (voir section 7.6.2)

[68] Expressions Ponctuelles Relatives Floues Uniques (voir section 7.6.1). Par exemple : « il y a environ un an », « plusieurs semaines après », « aux environs de la semaine prochaine », « vers le 22 décembre », « mardi, peu de temps avant l'aube » (voir section 7.6.2).

84,49%.

Les erreurs commises se concentrent dans les catégories *relatives* (*R**). Une analyse a donc été menée pour dégager de manière plus précise les facteurs qui entraînent une mauvaise interprétation. Plusieurs cas ont pu être dégagés, dont les principaux sont : les erreurs dues à l'absence ou à une mauvaise information temporelle en provenance d'une forme ou d'un groupe verbal (102 erreurs constatées, soit environ 62%) et la prise en compte d'un point de référence temporel qui n'est pas adéquat (23 erreurs constatées, soit environ 14%). Ces différents cas sont analysés de manière plus détaillée dans les paragraphes suivants. D'autres erreurs, moins fréquentes, sont encore à signaler : les expressions dont la référence temporelle s'établit par rapport à un événement[69] (12 erreurs, 7%), ainsi que diverses autres causes (28 erreurs, 17%), dont des fautes de reconnaissance ou d'annotation (8 sur 28) et la non-interprétation de date *nommées*[70] (9 sur 28).

(I) La première source d'erreurs concerne donc l'interprétation de temps verbaux. En particulier, l'interprétation d'expressions issues de phrases sans verbe a généré 45 mauvais résultats. L'interprétation des temps (grammaticaux) des formes ou des groupes verbaux, lorsqu'elle ne fournit pas le bon temps (notionnel), a aussi parfois provoqué des erreurs (57 cas).

(I.1) Une mauvaise interprétation du temps notionnel (passé/présent/futur) a pour conséquence d'aiguiller le processus d'interprétation de l'expression temporelle dans la mauvaise direction de l'espace temporel. La tentative de satisfaire les contraintes définies par l'expression temporelle (généralement relative dans ce cas) n'a dès lors aucune chance d'aboutir à la bonne solution. Toutes les erreurs constatées – 57 cas – portaient sur des expressions qui désignaient une zone située dans le futur[71], par rapport au point de référence, mais qui ont été interprétées dans le passé. Différents cas particuliers présentent ce type d'erreur.

(I.1.a) En premier lieu, viennent les erreurs relatives à l'interprétation de la valeur temporelle du temps verbal. Dans le corpus de test, le problème a été mis en évidence avec des verbes conjugués selon différents temps, dont les cas les plus fréquents sont les verbes au présent, les participes passés seuls, et les infinitifs :

> « L'ultime étape mène les coureurs <timex> dimanche </timex> sur 174,4 km de Kulmbach à Neumarkt/Oberpfalz. ».
>
> « Aucun des deux ne semble en mesure d'obtenir la majorité absolue ce qui fait craindre des contestations avant même l'annonce, attendue <timex> lundi </timex>, des premiers résultats officiels. »
>
> « Il devançait de plus d'une minute le Norvégien Petter Solberg (Subaru Impreza), le

[69] Par exemple « moins d'une semaine avant le 60ème anniversaire de la libération d'Auschwitz ». Cas non traité pas notre système, dont le traitement nécessite des processus complexes, tels que des mécanismes de résolution d'anaphore, ou la prise en compte de connaissances externes *sur le monde*. Le début de l'expression, « moins d'une semaine avant » est cependant reconnue par les grammaires, mais son interprétation est la plupart du temps erronée. Idéalement, l'événement pourrait être détecté, et ce genre de séquence ne devrait alors pas être interprété.

[70] Par exemple « Noël ».

[71] Le comportement par défaut du système étant de se tourner vers le passé, il est normal que ce type d'erreur soit majoritaire.

> Finlandais Marcus Gronholm (Peugeot 307) et l'Espagnol Carlos Sainz (Citroën Xsara), remplaçant de notre compatriote François Duval, avant les deux dernières spéciales à courir <timex> dimanche matin </timex>.

D'autres cas ont encore été relevés, avec le participe présent, le présent à la forme passive, etc. L'interprétation de ces verbes aurait dû fournir une valeur de futur. Le modèle d'analyse des temps verbaux actuellement implémenté ne permet cependant pas de couvrir ces cas.

(I.1.b) Ensuite, les erreurs dues à l'interprétation des temps morphologiques sont cependant souvent précédées et masquées par une mauvaise reconnaissance de certains groupes verbaux, en particulier ceux qui font intervenir un infinitif. Différents cas où celui-ci se combine avec le présent, le futur, ou encore le conditionnel ont été notés :

> « En appel, le TPIR doit rendre son verdict <timex> le 20 mai </timex> dans l'affaire de l'ex-maire de Bicumbi (centre), Laurent Semanza. »

> « La Russe Yelena Isinbayeva, qui a amélioré son record du monde du saut à la perche en franchissant 4,93 m <timex> le 5 juillet </timex> à Lausanne, tentera d'aller un centimètre plus haut lors de la réunion de Madrid <timex> samedi </timex>. »

> « Le travail devrait normalement reprendre <timex> ce vendredi </timex> sur le site de Fleurus. »

Dans ce type de situations, on aurait pu s'attendre à ce que les groupes verbaux composés d'infinitifs soient repérés par l'analyse syntaxique, ce qui n'est pas le cas. Étant donné le nombre d'erreurs, ce problème pourrait cependant être géré en aval, lors de l'interprétation des temps.

(I.1.c) Enfin, quelques erreurs de détection des formes verbales sont aussi imputables au Treetagger.

(I.2) Une deuxième situation assez fréquente, en ce qui concerne les mauvaises interprétations dues aux verbes, survient lorsque il n'y a pas de verbe à proximité de l'expression à interpréter (pour rappel, 45 cas constatés). Dans ce cas, une mauvaise association peut être effectuée avec un verbe qui n'a normalement aucun rapport avec l'expression, mais qui est présent dans la même proposition. L'absence totale de verbe dans la proposition constitue également un problème pour l'interprétation, car il n'y a alors pas d'indice pour décider de l'orientation temporelle à donner à l'interprétation. Cette dernière situation est le plus souvent rencontrée dans les titres d'articles.

> Les Américains sont bien armés pour reprendre leur domination sur le 400 m haies où la victoire leur échappe <timex> depuis 1995 </timex>, lors de la finale de l'épreuve des Mondiaux <timex> 2005 </timex> d'athlétisme, <timex> mardi soir (20 h 25) </timex> à Helsinki.

> « Dominique Bruyneel au départ <timex> jeudi </timex> du championnat européen des rallyes »

(II) Après la prise en compte des temps verbaux, la deuxième plus importante source d'erreurs – 23 cas – est la gestion du point de référence nécessaire à l'interprétation des expressions relatives.

L'analyse des erreurs a également permis de dégager différentes situations.

(II.1) Tout d'abord, dans certains cas, la valeur donnée au type de référence (REF, déictique ou anaphorique) lors de l'annotation ne convient pas. Ainsi, plusieurs expressions marquées comme déictiques avaient en réalité un point de référence situé dans un contexte proche ou éloigné (6 erreurs). Ces expressions pourraient être correctement traitées si une exception était générée par rapport à leur annotation en tant que déictique, lorsqu'un élément fort est détecté dans le contexte. Par exemple, pour les erreurs constatées, nous avions entre autres les expressions suivantes à *résoudre* :

- « le 13 mars », résolu en 13-3-2004 au lieu de 13-3-2001, avec dans le contexte la date « 14 mars 2001 » ;
- « après juin », résolu en 6-2005, au lieu de 6-2004, avec dans le contexte l'intervalle « entre février et juin 2004 » ;
- « avant la fin de l'année », résolu en END-2005 au lieu de END-2004, avec dans le contexte la date « en 2004 » ;

(II.2) D'autre part, le cas inverse d'une expression marquée comme anaphorique, mais dont le point de référence est équivalent à la date d'émission de l'article, est également présent (4 erreurs). Ces cas, observés principalement pour des granularités inférieures au jour (« à 8h15 », « l'aube »), ne semblent cependant pas constituer de véritables erreurs, mais échouent néanmoins par manque de contexte explicitement exprimé.

(II.3) D'autres erreurs surviennent encore, malgré le choix correct d'un mode de référence de type anaphorique. L'interprétation de ces expressions relatives peut cependant ne pas réussir, soit parce que la référence à utiliser est située bien avant dans le texte (référence lointaine) et que d'autres expressions temporelles viennent interférer entre les deux (6 erreurs), soit parce que l'expression de référence est située dans la suite du texte et non dans la partie déjà analysée (2 erreurs).

La question de la portée de la référence temporelle est, d'une manière générale, également un point important. Le choix de la taille de l'unité textuelle – phrase, paragraphe, chapitre, etc. – à l'intérieur de laquelle sont recherchés les points de références a évidemment une influence sur l'analyse. Dans le cas des dépêches cela semble être moins le cas, en raison du caractère généralement bref des textes. La portée de la référence temporelle s'étend donc potentiellement sur l'entièreté du texte.

(II.4) Enfin, les erreurs restantes, qui concernent également des expressions relatives anaphoriques, sont à attribuer à un effet de *réaction en chaîne*, dû à une mauvaise interprétation de l'expression qui sert de point de référence (5 erreurs).

7.10 Perspectives et conclusion

Le premier résultat à mettre en avant est la réalisation concrète d'un système d'extraction et d'interprétation d'informations temporelles pour le français, proposant à la fois finesse de description et couverture importante. Les caractéristiques principales du système résident dans la prise en compte de l'imprécision, ainsi que l'exploitation d'un système de granularités étendu. La combinaison de ces

éléments a rendu possible la mise au point de mécanismes d'interprétation capables de faire face à la richesse et à l'imprécision naturelle de la langue.

L'évaluation des résultats a montré que ceux-ci correspondent aux critères attendus en extraction d'informations. La première phase de reconnaissance a en effet permis d'atteindre des valeurs de rappels allant jusqu'à 90,78%, rapidement étendus jusqu'à un niveau potentiel d'environ 97%[72]. La précision, en atteignant son maximum à 97,74% est sans conteste le point fort de la méthode. Ce constat est très positif, le rappel pouvant toujours être amélioré au fil du temps, comme cela a été le cas lors des corrections réalisées après l'évaluation. Le système peut ainsi petit à petit atteindre ses performances optimales. Lors de la seconde phase de traitement, qui consiste en l'interprétation et la normalisation des expressions temporelles reconnues, et dont le but est de transformer celles-ci en *valeurs* temporelles, une précision de 90,76% a été obtenue.

Comme souvent en extraction d'informations, les résultats laissent de la place à certaines améliorations. Tout d'abord, la reconnaissance et l'annotation des expressions temporelles peuvent bien entendu encore être améliorées : certaines séquences non prises en compte, meilleur reconnaissance des intervalles, expressions avec point de référence constitué par un événement nommé, etc. Du côté de l'interprétation des expressions temporelles reconnues, les différents mécanismes mis en place ont rencontré un certain succès. L'amélioration des performances passe cependant par quelques corrections mais également par une sophistication de plusieurs points, dont l'interprétation des temps verbaux est le principal. De nombreuses erreurs ont été commises lors de l'interprétation des expressions temporelles relatives suite à une mauvaise *traduction* du temps du verbal. La gestion du point de référence constitue le second point délicat qui pourrait être amélioré. La question de la portée de la référence ainsi que celle de la résolution des références lointaines ou vers l'avant constituent des problèmes dont la résolution demande une analyse beaucoup plus globale du texte. Diverses autres améliorations, comme la prise en compte de l'aspect lexical et grammatical, pourraient venir compléter le modèle d'interprétation temporel. Enfin, une fois les limites de l'analyse locale atteintes, les derniers éléments ne pourront probablement pas être grappillés sans exploiter véritablement une analyse syntaxique.

L'analyse automatique mise sur pied prouve également que, même sans atteindre une analyse linguistique parfaite de l'ensemble des phénomènes reliés au temps dans les textes, il est possible de manière pragmatique, en simplifiant certains aspects et en ignorant même d'autres, d'extraire une information suffisamment complète et précise pour espérer pouvoir l'exploiter dans une application concrète (voir chapitre 8). De plus, les perspectives d'évolutions subsistent à divers points de vue et permettent d'encore envisager une amélioration future des performances.

En outre, ce travail propose une méthodologie. Celle-ci est centrée autour d'un document de référence, la spécification des expressions temporelles, autour duquel s'organise de manière cohérente l'ensemble des étapes de développement du système. L'intérêt est à la fois théorique – pour la description des expressions temporelles – et pratique – définition d'une marche à suivre pour le déve-

[72] Cette seconde mesure est une estimation du rappel maximal qui pourrait être atteint suite à certaines corrections effectuées. Une seconde évaluation, sur un nouveau corpus de test n'a cependant pas pu être menée.

loppement des ressources linguistiques et du système d'analyse.

Enfin, l'ouverture du système a été démontrée par l'intégration du processus de traitement avec un autre système de traitement automatique du langage (voir chapitre 8). Cela a permis au final la réalisation d'une application de recherche d'informations (RI) à dimension temporelle. De plus, l'analyse temporelle peut aussi jouer le rôle de module pour de nombreuses autres applications, telles que l'extraction d'informations, le résumé automatique, les systèmes de question-réponse, la traduction automatique, les activités de veille stratégique, etc.

De manière plus concrète, l'analyse temporelle mise sur pied va être intégrée aux développements en extraction d'informations biographiques actuellement en cours dans le cadre de la *Spin-off* Knowbel. L'objectif est d'améliorer le traitement du temps actuellement réalisé. Celui-ci se limite à la reconnaissance des expressions temporelles, sans interprétation ni normalisation. Or, dans de nombreux cas, l'information biographique peut tirer parti de ce type de traitement. À la figure 7.14, on constate par exemple que deux valeurs temporelles sont disponibles pour l'événement *mort* (« lundi après-midi » et « lundi à 14 heures 07 »). Elles sont toutes deux sous-spécifiées, ce qui implique que l'on ne connaît pas réellement la date en question. Cette information est susceptible d'être fournie par l'analyse temporelle. De plus, les deux références temporelles indiquent des parties de la journée différentes, mais compatibles («après-midi» et « 14 heures 07 »). Ces deux informations pourraient être croisées afin de les vérifier et de les valider[73], ce qui permettrait alors d'éviter la redondance actuellement obtenue. Suite à l'interprétation du jour du décès en une date précise, et connaissant la date de naissance, il est alors possible d'inférer l'âge de la personne au moment de sa mort. Enfin, d'une manière générale, l'ensemble des faits biographiques pourraient être accompagnés de la période temporelle à laquelle ils se rapportent. Ces quelques exemples illustrent bien les nombreux points qui peuvent être enrichis grâce à une analyse temporelle précise.

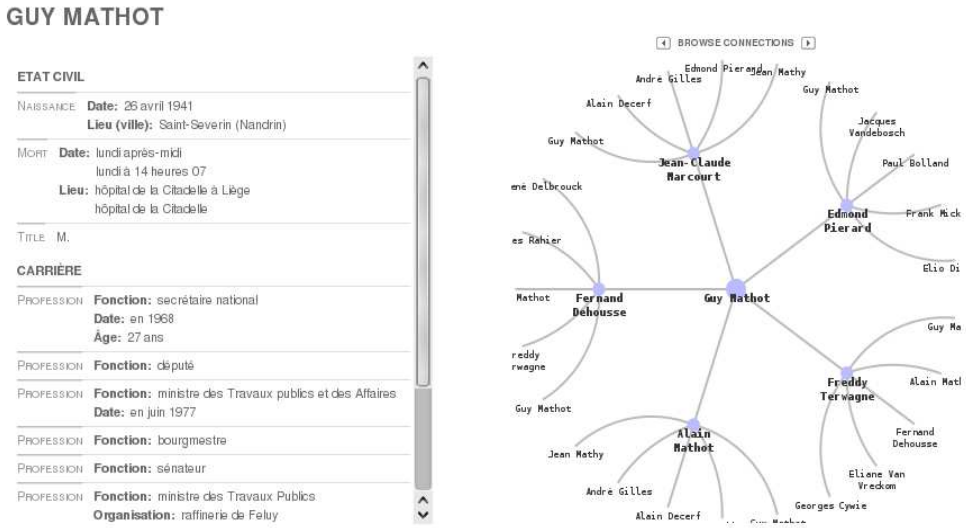


Figure 7.14 : Interface de consultation des données biographiques extraites par Exabyte (Spin-off Knowbel).

[73] Ce type de processus de validation peut également permettre de confronter des informations temporelles imprécises et des dates bien identifiées.

Notre démarche présente ainsi le traitement de l'information temporelle de manière large, allant du stade fondamental de la définition et de la catégorisation des expressions, jusqu'à leur utilisation au sein d'applications concrètes.

CHAPITRE 8

INDEXATION THÉMATICO-TEMPORELLE DE DOCUMENTS TEXTUELS

8.1 Introduction

8.1.1 Notion de recherche d'informations à dimension temporelle

La recherche d'informations est, comme cela a déjà été exposé (voir l'introduction à l'indexation, la recherche d'informations et les moteurs de recherche, au chapitre 1), l'activité qui permet à un utilisateur de rechercher dans une base documentaire un ensemble de documents pertinents au regard d'une requête qu'il a exprimée. Cette recherche se base sur un ou plusieurs index qui ont été préalablement construits lors de l'insertion du document dans la collection. Sans entrer dans des détails qui dépassent le cadre de cette thèse, on peut constater que les moteurs de recherche sont souvent basés sur les *tokens*, éventuellement sur les termes composés, contenus dans les textes (voir aussi les sections 1.2 et 1.3). Certains traitements, tels que la *racinisation* ou l'extension de requêtes, sont parfois entrepris afin d'améliorer la couverture, mais il n'existe que peu d'interventions au niveau du sens du contenu indexé[1], même si cela fait l'objet de nombreuses recherches (voir section 1.3.4).

Si l'on attribue le même statut à tous les *tokens*, de nombreuses informations utiles pour la caractérisation du contenu du document ne sont pas considérées à leur juste valeur. C'est particulièrement le cas pour l'information temporelle. En effet, bon nombre de documents relatent des événements ou des faits qui, s'ils peuvent être reliés à un ou plusieurs thèmes, possèdent également une ou plusieurs dimensions supplémentaires, qui sont souvent de nature temporelle ou spatiale. Cela semble évident si l'on considère spécifiquement des documents d'actualités (articles de journaux, dépêches, etc.). Dès lors, il est pertinent de prendre en compte ces dimensions lors de l'indexation, afin qu'elles puissent être exploitées lors de la recherche de documents. Comme le constatent Alonso *et al.* [2007], Nunes *et al.* [2008] ou Vicente-Diez et Martinez [2009], peu de systèmes de recherche d'informations incluant un support réel des aspects temporels ont vu le jour ces dernières années. Dans le cadre de tels systèmes, un utilisateur peut affiner sa requête en ajoutant des critères spécifiques sur la dimension spatiale ou temporelle. Cette possibilité représente une valeur ajoutée réelle pour la recherche d'informations qui, sans traitement spécifique au temps, ne prendra pas nécessairement en compte correctement cette dimension. Une expression telle que « jeudi » ne peut être mise en rapport

[1] Les développements relatifs au web sémantique sont une tentative d'aller dans cette direction. Les résultats obtenus jusqu'ici ne correspondent cependant pas toujours aux attentes en la matière.

avec une requête qui porterait sur un jour exprimé au moyen d'une valeur *jour/mois/année*. De même, comme le suggère Palacio *et al.* [2010], une recherche sur l'année 1984 ne permet pas de retrouver les documents dans lesquels apparaissent l'expression « les années 1980 ». La situation inverse pose le même type de problème : la requête qui contient des valeurs temporelles sous-spécifiées, imprécises ou d'une granularité élevée peut difficilement être mise en relation avec des expressions précises, bien identifiées ou d'une granularité différente. Enfin, lorsqu'une valeur temporelle est prise en compte, il s'agit souvent de la date de création du document, qui n'est pas nécessairement en rapport avec son contenu (Alonso *et al.* [2007], Alonso *et al.* [2009]).

8.1.2 Utilisation concrète dans les systèmes de recherche d'informations actuels

Afin d'illustrer en pratique l'absence de support de la dimension temporelle, quelques systèmes de recherche d'informations actuels ont été passés en revue.

La consultation du portail de l'Union européenne[2] est un bon exemple. Le formulaire de recherche avancée (Figure 8.1) ne propose pas de champ spécifique pour l'entrée d'une valeur temporelle. Le résultat de la requête « tremblement de terre 2009 » propose un certain nombre de textes en rapport avec le séisme de L'Aquila (Italie) en 2009, ce qui est tout à fait correct. Mais la liste reprend également, dans les dix premières propositions, trois documents non pertinents, c'est à dire qui parlent de tremblements de terre d'une part, d'événements relatifs à l'année 2009 d'autre part, mais en aucun cas d'éléments reliés à ces deux critères en même temps.

La figure 8.2 illustre le même type de situation, dans le cas d'une base documentaire consacrée aux désastres[3]. Pour la requête « earthquake 2009 », le moteur de recherche renvoie un seul résultat dont le thème est bien relatif aux tremblements de terre, mais dont le positionnement temporel concerne l'année 2004. L'erreur d'interprétation provient ici de la prise en compte de la date d'émission du texte et non des valeurs temporelles relatives au contenu du document.

Enfin, citons encore le cas de la base documentaire du Sénat belge, déjà évoqué à la section 1.4.2 (Figure 1.10) : son formulaire de recherche propose une interrogation à dimension temporelle, mais limitée aux dates de publication des documents.

8.1.3 Premier bilan

Ces quelques exemples, issus de systèmes d'information d'organisations importantes, permettent de mettre le doigt sur certaines lacunes en matière de recherche de documents, et plus spécifiquement en ce qui concerne l'exploitation des données temporelles. En effet, de nombreux moteurs de recherche ne proposent, en guise de support de ces informations temporelles, qu'une recherche par mots-clés

[2] `http://europa.eu/index_fr.htm`, consultation le 26/10/2010

[3] Cette base a été créée et est maintenue par le *Centre for Research on the Epidemiology of Disasters* (`http://www.cred.be`, consultation le 27/07/2010).

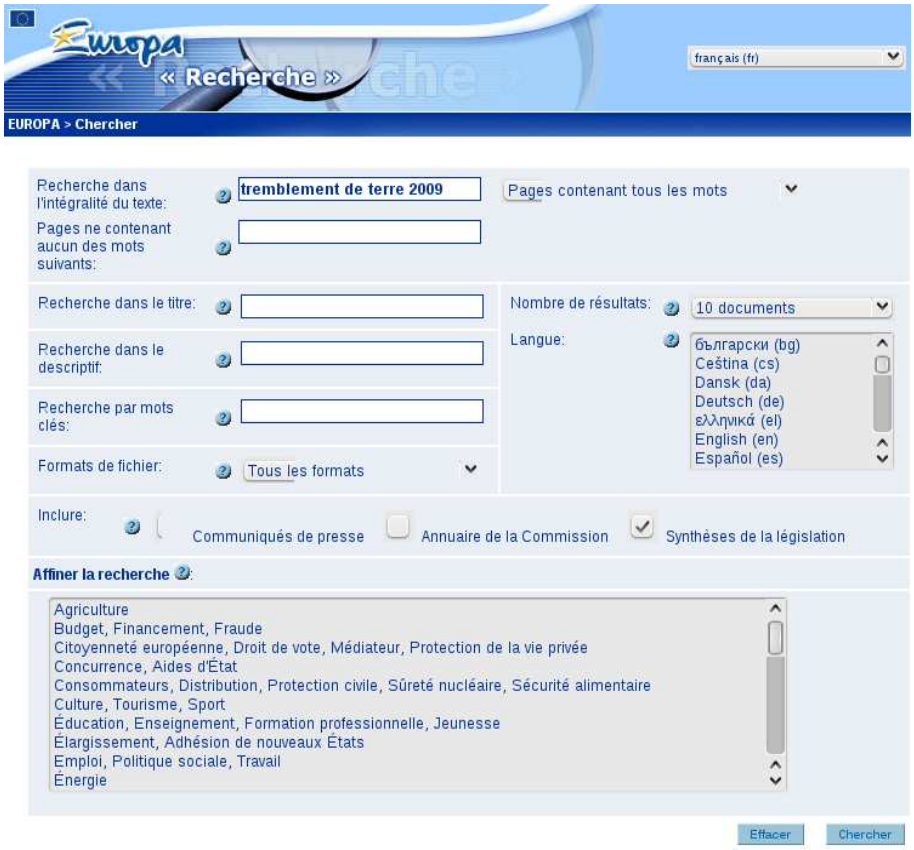


Figure 8.1 : Formulaire de recherche avancée du portail de l'Union européenne.

Figure 8.2 : Résultat insatisfaisant d'une recherche de document à l'aide d'une requête incluant une dimension temporelle.

classique ou une fonction de *filtre* sur la date d'émission du document[4].

Afin de dépasser cette limitation, l'indexation des documents doit évoluer d'une approche strictement

[4] Cette fonction apparaît par exemple dans le formulaire de recherche avancée de Google News (`http://news.google.com/news/advanced_news_search`, consulté le 02/12/2010).

thématique vers une approche multidimensionnelle. Ainsi, la liste traditionnelle de catégories peut céder sa place à une liste de *tuples* dont un des composants est une information temporelle. D'autres dimensions peuvent également venir enrichir les éléments de l'index. D'une manière assez intuitive et naturelle, le temps est par exemple assez souvent lié à une valeur spatiale. Chaque *tuple* peut dans ce cas être représenté par une catégorie accompagnée facultativement par sa dimension spatio-temporelle. Par exemple, l'indexation d'un document par les catégories thématiques

[attaque à main armée ; banque]

peut être remplacée avantageusement par la liste

[(attaque à main armée, 17/03/2005, Bruxelles) ; (banque,_,_)].

Ce type de *tuple* permet à la fois de conserver une indexation classique, mais également d'y adjoindre des dimensions supplémentaires lorsqu'il a été possible de les détecter.

8.2 Travaux apparentés à l'indexation à dimension temporelle

La prise en compte de la dimension temporelle dans les systèmes de recherche d'informations a suscité un intérêt croissant ces dernières années. Cette évolution est due principalement aux développements relatifs aux moteurs de recherche, ainsi qu'au domaine de la recherche d'informations géographiques.

8.2.1 Les moteurs de recherche

D'une manière générale, de nombreuses études (Alonso *et al.* [2007], Nunes *et al.* [2008] ou Vicente-Diez et Martinez [2009]) s'accordent pour souligner que la dimension temporelle de l'information est sous exploitée, voire ignorée, par la majorité des systèmes d'indexation et de recherche d'informations. C'est bien entendu le cas, en tant qu'application-phare pour les moteurs de recherche. Cependant, le principe suscite un intérêt croissant, et certaines expériences en cours tentent de mieux exploiter l'information temporelle. C'est par exemple le cas pour Google News Timeline[5].

De manière un peu surprenante, l'analyse des logs de requêtes de moteurs de recherches commerciaux réalisée par Nunes *et al.* [2008] montre cependant que, de manière globale, la proportion de requêtes incluant une valeur temporelle n'est pas si élevée qu'on pourrait le croire (environ 1,5%). Ce premier constat doit être nuancé. Tout d'abord, en chiffres absolus, à l'échelle du Web, cela représente toujours un nombre important de requêtes. Ensuite, une analyse plus fine à montré que certains domaines étaient plus favorables à l'apparition d'une dimension temporelle (par exemple *Auto*, 7,8%, *Sport*, 5,2% ou *News & Society*, 3,9%) ; c'est également l'avis de Vicente-Diez et Martinez [2009]. En dehors du Web, diverses collections particulières de documents se distinguent par l'importance des informations temporelles, comme par exemple les bases documentaires géographiques (voir section 8.2.2), les bases de *lettres de sortie* en milieu hospitalier, ou encore les collections de documents

[5] `http://newstimeline.googlelabs.com/`

qui présentent de nombreuses informations biographiques. Enfin, l'hypothèse concernant l'emploi par l'utilisateur d'interfaces tierces lorsqu'il est confronté à un besoin de recherche plus poussé, entre autres au niveau temporel, est avancée par Nunes *et al.* [2008]. Ce point de vue est conforté, selon nous, par le fait que les outils à disposition du public ne sont pour la plupart pas prévus pour prendre en charge ce genre de requête. Les utilisateurs ne sont dès lors pas encouragés à formuler de tels critères de recherche.

Les avantages présentés par un traitement particulier de dimensions de recherches telles que le temps sont assez clairs. Alonso *et al.* [2007] cite quelques possibilités offertes par le traitement temporel : le *clustering temporel*, les modes de présentation des résultats, et évidemment, l'extension de la recherche classique. Lors du déroulement de cette dernière, la prise en compte des difficultés déjà citées à la section 8.1.1, en ce qui concerne la mise en correspondance de formes différentes, mais présentant des valeurs temporelles compatibles, est très importante. De plus, l'idée que les valeurs temporelles puissent intervenir dans l'ordonnancement des résultats en tant qu'extension du mécanisme de réputation ou de popularité actuellement en cours (par exemple Page Rank, Brin et Page [1998]), a également été formulée (Alonso *et al.* [2009]). En ce qui concerne la possibilité de *clustering temporel* des résultats d'une recherche classique, par exemple pour une requête sur la guerre en Irak[6], il est facilement imaginable que les résultats puissent être organisés en deux groupes correspondant aux deux guerres différentes qui s'y sont déroulées. Le clustering temporel a également été proposé par Alonso *et al.* [2009]. Enfin, toujours selon Alonso *et al.* [2007], les possibilités de présentation des résultats selon un angle d'attaque temporel sont très appréciées des utilisateurs. De plus, de nombreuses propositions alternatives à la classique ligne du temps voient le jour et pourraient à l'avenir améliorer l'accès aux documents résultant des requêtes adressées aux moteurs de recherche.

En ce qui concerne la phase d'indexation, base du système de recherche d'informations, les développements présentés par Pasca [2008] ou par Vicente-Diez et Martinez [2009] permettent de se rendre mieux compte des apports nécessaires à la prise en charge de la dimension temporelle. D'une manière générale, le processus se déroule en trois phases : la reconnaissance et l'interprétation des expressions temporelles, l'indexation classique, et finalement, lors d'une requête, le recoupement de ces deux index. L'index final est donc constitué par l'ajout de l'index temporel à l'index classique. Les liens entre les expressions temporelles et les termes de cet index classique sont par conséquent réalisés au niveau de l'unité indexée. Pour Vicente-Diez et Martinez [2009], il s'agit du texte en entier, alors que Pasca [2008] descend au niveau des parties de phrases, nommées *nuggets*, qui sont définies au moyens de certains patrons particuliers, tels que « *Date*, when *Nugget* » (par exemple, « By 1910, when Korea was annexed to Japan »). Cette plus grande précision dans le lien entre valeur temporelle et élément thématique doit cependant être tempérée par la limite en termes de reconnaissance que ce type de patron peut provoquer. De même, le système proposé par Pasca [2008] est handicapé par la couverture temporelle limitée des dates bien identifiées par rapport au calendrier (années, décennies, couples mois et année, ou encore triplets mois, jour dans le mois et année). En fin de compte, les deux systèmes donnent des résultats positifs pour la recherche d'informations, allant même jus-

[6] Cet exemple est proposé par Alonso *et al.* [2007] pour illustrer l'ambiguïté temporelle de certaines requêtes.

qu'à un amélioration très importante des performances[7] chez Vicente-Diez et Martinez [2009], en particulier en ce qui concerne la précision.

8.2.2 Les systèmes de recherche d'informations géographiques

Le domaine de la recherche d'informations géographiques constitue également un moteur important pour les travaux en indexation à dimension temporelle. En effet, comme déjà mentionné à la section 8.1.1, il est assez courant que des éléments thématiques soient accompagnés d'autres dimensions telles que le temps, mais aussi l'espace. C'est bien évidemment le cas lorsqu'il s'agit de données géolocalisées. Ces dernières années ont vu l'apparition de plusieurs initiatives allant dans ce sens, entre autres Bilhaut *et al.* [2007] et le projet GéoSem, Le Parc-Lacayrelle *et al.* [2007] et Palacio *et al.* [2010] au sein du projet PIV, Manguinhas *et al.* [2009] et DIGIMAP, ou encore Strötgen *et al.* [2010]. Ce dernier expose très clairement la justification de tels développements :

> « Aspects related to temporal and geographic information embedded in documents that go beyond just the timestamp and location of publication of a document have been of particular interest in many practically relevant document search and exploration tasks. Almost all types of documents contain a variety of temporal and geographic information that describes events, the location of such events, typically in combination with other named entities such as persons or organizations. » (p. 1)

L'objectif de ces systèmes est donc de repérer, d'extraire et d'interpréter les informations temporelles et géographiques, et de les stocker de manière à pouvoir y accéder à des fins de recherche. Ces informations spatio-temporelles constituent un intérêt en soi, mais elles apparaissent également souvent en tant que dimension d'une autre information, dénommée *phénomène*, dont la nature varie fortement suivant les textes analysés et le contexte du système de recherche d'information. Par conséquent, la majorité de ces systèmes manipulent de objets multidimensionnels, représentés sous la forme de triplets *(phénomène, espace, temps)*.

La majorité de ces systèmes partagent la caractéristique de traiter les différentes dimensions de manière séparée, c'est-à-dire d'en effectuer une indexation indépendante. Très souvent, l'unité indexée se situe à un niveau inférieur à celui du texte, par exemple le paragraphe ou la phrase. Dans les sections suivantes, les différentes dimensions sont successivement examinées pour ensuite voir comment elles peuvent être liées.

Reconnaissance des phénomènes

D'une manière générale, la reconnaissance des éléments thématiques, ou *phénomènes*, est réalisée à l'aide de techniques assez classiques dans le domaine de l'indexation.

Par exemple, chez Bilhaut *et al.* [2007], la reconnaissance de ces phénomènes est réalisée principa-

[7] La mesure utilisée, MRR (Mean Reciprocal Rank), passe de 0,40 (baseline) à 0,81 (avec prise en compte du temps).

lement à l'aide d'une mesure de type TF.IDF, calculée au niveau intra-documentaire. Cette mesure permet de trouver les termes d'un passage qui sont statistiquement plus importants que les autres, et cela en comparaison avec le reste du document. Les termes recherchés sont de préférence des expressions composées plutôt que des mots simples. Ces derniers sont en effet jugés moins pertinents et discriminants dans le contexte de documents en rapport avec un domaine de spécialité (l'information géographique).

Dans le même ordre d'idées, Le Parc-Lacayrelle *et al.* [2007] utilise également une approche basée sur une mesure TF.IDF, modifiée pour atténuer les effets négatifs provoqués par les longs documents (Robertson *et al.* [1998]).

Dimension spatiale

Pour le moteur GéoSem, Bilhaut *et al.* [2007] reconnaissent les expressions spatiales grâce à la présence de toponymes. Deux types d'opérateurs peuvent venir s'ajouter à un toponyme simple. Les opérateurs spatiaux tels que « le nord de X », « le triangle X Y Z » ou encore « de X à Y » constituent un premier groupe. Le second est formé par des opérateurs de sélection d'entités au sein d'un zone donnée (« quelques villes maritimes de Normandie »). Ces expressions spatiales sont ensuite transformées en structures de traits capables de représenter la diversité induite par l'usage des différents opérateurs. Une phase d'interprétation permet ensuite d'obtenir une entité géographique (un type, un identifiant et un qualificatif) et son placement dans un espace géographique (géolocalisation à l'aide des latitudes et longitudes des zones).

Le processus présenté par Strötgen *et al.* [2010] est assez comparable. Le repérage d'entités géographiques, qui repose en grande partie sur des ressources lexicales, est suivi d'une phase de normalisation. L'étape ultime consiste alors à attribuer une valeur géographique au lieu identifié.

Dimension temporelle

Le traitement des expressions temporelles pris en charge par GéoSem (Bilhaut *et al.* [2007]) n'est malheureusement que peu détaillé. Les unités temporelles reconnues sont des dates éventuellement modifiées à l'aide d'opérateurs d'intervalles (« de X à Y », « entre X et Y », « les années X ») ou d'approximation (« le début de X », « aux alentours de X »). L'interprétation de ce type d'expression fournit une valeur approximative de la période de temps concernée sous la forme d'un intervalle délimité par deux dates.

Dans Strötgen *et al.* [2010], le traitement des expressions temporelles n'est pas exposé de manière beaucoup plus explicite. Les éléments temporels reconnus sont ceux définis par Schilder et Habel [2001], c'est-à-dire des expressions temporelles explicites (expression calendaire complète), implicite (dates nommées avec mention de l'année), ou encore relatives (qui sont interprétées par rapport à la date de création du document).

Enfin, le support de la dimension temporelle qui apparaît comme le plus complet est dû à Le Parc-

Lacayrelle *et al.* [2007]. Les valeurs temporelles exploitées sont celles à connotation calendaire, c'est à dire les jours, mois, saisons, années ou siècles. Les entités relatives aux heures, aux durées, ou celles rattachées à des événements historiques ne sont pas prises en compte. Une distinction est faite entre les expressions *complètes*, qui peuvent être situées sur une échelle de temps absolue, et *incomplètes*. Ces dernières font l'objet d'un traitement syntaxico-sémantique, dont la nature exacte n'est pas exposée, mais qui semble destinée à les ancrer dans l'espace du temps.

Lien entre les différentes dimensions de l'indexation

Chaque dimension étant indexée indépendamment, le lien entre ces dimensions s'effectue de manière assez naturelle au niveau de la partie de texte qui constitue une unité indexable minimale. La découpe du texte en passages auxquels les termes d'indexation sont attribués peut être réalisée selon diverses méthodes et à différents niveaux. Tout d'abord, il est possible de fractionner le document en exploitant des marques typographiques tels que les sauts de paragraphe ou les signes de ponctuation qui délimitent les phrases (Le Parc-Lacayrelle *et al.* [2007], Strötgen *et al.* [2010]). Une autre possibilité, évoqué par Bilhaut *et al.* [2007], est d'exploiter les techniques de segmentation thématique par cohésion lexicale (Ferret *et al.* [2001]). Cependant, la méthode finalement mise en avant pour GéoSem pour la définition de *passages* est celle mettant en œuvre les cadres de discours. Pour rappel, l'hypothèse de l'encadrement du discours proposée par Charolles [1997], dont il a déjà été question aux sections 4.9 ou 7.3, montre que plusieurs propositions successives du texte peuvent avoir un lien identique avec un élément particulier et dès lors être regroupées en un bloc, dénommé *cadre*. Bilhaut [2007] démontre que divers types de cadres de discours peuvent être utilisés afin de mener une analyse thématique du texte. Ceux exploités pour GéoSem se réfèrent bien entendu à des éléments spatiaux ou temporels. En ce qui concerne le point spécifique de la détection automatique des cadres de discours spatiaux et temporels, un exposé plus complet a été réalisé par Bilhaut *et al.* [2003] et Ferrari *et al.* [2005].

8.2.3 Extraction d'informations

Une certaine similitude peut être établie entre l'indexation thématico-temporelle et la tâche d'extraction d'informations qui consiste à repérer les *événements* d'un texte et à les relier à des valeurs temporelles. Cette tâche, que nous nommerons ici extraction d'événements, a été brièvement présentée à la section 6.3.2 et est souvent reliée au langage TimeML (voir section 6.5.3).

Le but poursuivi n'est cependant pas tout à fait le même. Dans notre cas, il s'agit de construire une représentation thématico-temporelle d'un document, alors que dans l'autre, il s'agit d'ordonner les événements dans l'espace temporel.

Les événements visés par la tâche d'extraction sont également différents. Pour l'extraction d'événements, ceux-ci sont principalement des verbes, même si le repérage inclut aussi parfois des groupes nominaux (« les élections législatives »), des noms événementiels (« guerre »), des adjectifs (« malade ») ou des groupes prépositionnels (« à bord »). Notre approche de la reconnaissance d'éléments

thématiques (voir chapitre 2) a par contre plutôt tendance à repérer des groupes nominaux, bien que les verbes puissent également être exploités[8]. De plus, là où avec notre approche la reconnaissance est sémantiquement guidée, les systèmes d'extraction d'événements ont une démarche plus ouverte, voire générique.

Malgré ces différences, ces deux activités pourraient probablement être rapprochées, et éventuellement se renforcer mutuellement.

8.3 Implémentation d'un système d'indexation à dimension temporelle

Le système proposé ici se place dans la même tendance que ceux présentés à la section 8.2, c'est-à-dire l'indexation multidimensionnelle. En l'occurrence, notre objectif est d'adjoindre aux index thématiques habituels une dimension temporelle, afin de proposer de nouvelles perspectives aux systèmes de recherche d'informations.

L'implémentation exposée dans cette section s'appuie largement sur les développements présentés, pour les aspects thématiques et temporels, aux chapitres 2 et 7. Une présentation complète des ces aspects n'est donc pas répétée. Le processus général de traitement est identique à celui présenté pour l'extraction d'informations temporelles, illustré à la figure 7.7, dans lequel l'insertion d'analyses tierces avait déjà été prévu. Le processus de classification, permettant l'indexation thématique des documents, est d'ailleurs clairement identifié dans ce schéma.

La chaîne de traitement inclut deux analyses indépendantes, dédiées à la reconnaissances des expressions temporelles d'une part, et au repérage et à la pondération des indices thématiques d'autre part. Le résultat de chaque module est un texte annoté au moyen de balises SGML. Il sont ensuite rassemblés en une seule version du texte, à laquelle ont également été ajoutées certaines informations nécessaires à la découpe en proposition et à la reconnaissances des groupes verbaux complexes[9]. La représentation du texte obtenue contient les expressions temporelles, les limites de propositions, les verbes et leurs temps morphologiques, ainsi que les indices thématiques. Une étape d'interprétation des expressions temporelles est alors conduite afin de déterminer, dans l'espace du temps, une valeur pour chaque expression temporelle. Finalement, l'indexation thématico-temporelle est réalisée en associant à chaque indice thématique, et par conséquent à la catégorie qu'il dénote, la ou les valeurs temporelles[10] situées dans la même proposition. Le lien entre les deux dimensions est donc une simple co-occurrence au niveau de la proposition.

D'autres méthodes peuvent être adoptées en ce qui concerne la manière d'effectuer les liens. Il est par exemple possible de se baser sur les liens de dépendance établis par une analyse syntaxique. Ceux-ci pourraient fournir des informations plus fines pour rapprocher indices thématiques et informations temporelles. Cela ne nous semble pas indispensable dans le cas de l'indexation multidimensionnelle

[8] Voir l'essai d'exploitation de Verbaction à la section 2.7.
[9] Ces informations sont issues de l'analyse syntaxique.
[10] Lorsqu'elles existent.

telle que nous l'envisageons, mais cela pourrait très probablement l'être pour des tâches d'extraction d'informations plus précises. Une autre possibilité consiste à faire varier la taille de la fenêtre à l'intérieur de laquelle les liens de co-occurrence peuvent intervenir. Au lieu de s'appuyer sur la proposition, on peut par exemple adopter la phrase ou le paragraphe comme unité de base, ce qui devrait avoir un effet bénéfique sur le rappel, mais au risque de diminuer la précision.

Afin d'illustrer le processus complet, les figures 8.3 à 8.5 montrent la progression de l'analyse d'un texte. Les résultats sont également présentés de manière synthétique au tableau 8.1.

```
<text id='F051230A_0098.txt' date='20051230-15:59'>
<header>SPF016 3 GEN 0092 F BELGA-0176 </header>
<title>VTT - F. Meirhaeghe engagé pour 3 ans chez Landbouwkrediet (1LEAD) . </title>

§   ANDERLECHT 30/12 (BELGA) = Le coureur belge de mountainbike
Filip Meirhaeghe a signé jeudi un contrat de trois ans avec la
formation cycliste Landbouwkrediet-Colnago, a communiqué à la
presse, vendredi, la direction de l'équipe dirigée par Gérard
Bulens.
§   Meirhaeghe, suspendu depuis l'été 2004 jusqu'au 14 janvier 2005
pour avoir utilisé de l'EPO, a été engagé pour cette période de 3
ans afin de pouvoir se préparer au mieux pour les Jeux OLympiques de
Pékin en 2008. Le biker flandrien sera alors âgé de 37 ans.
./.TED
§ </text>
```

Figure 8.3 : Exemple d'un texte de départ.

Indice thématique	Expression temporelle	Valeur temporelle interprétée
belge	30/12 jeudi	30-12-2005 29-12-2005
contrat	30/12 jeudi	30-12-2005 29-12-2005
communiqué à la presse	vendredi	30-12-2005
Jeux OLympiques	en 2008	2008 (fuzzy=1)

Tableau 8.1 : Synthèse du résultat de l'indexation thématico-temporelle.

L'indexation de textes selon plusieurs dimensions, entre autres temporelle, constitue l'idée de base de l'ensemble des systèmes présentés à la section 8.2, et est également celle qui a été exploitée ici. Notre implémentation présente évidemment des point communs avec les différents systèmes passés en revue, mais divers aspects permettent de les différencier. Tout d'abord, comme dans Vicente-Diez et Martinez [2009], notre méthode a pour but l'indexation de documents entiers et non de passages comme c'est par exemple le cas chez Bilhaut *et al.* [2007] ou Pasca [2008]. Ensuite, la recherche des clés d'indexation principales, les *phénomènes* dans les systèmes de recherche géographiques, est réalisée selon des techniques très différentes (voir chapitre 2). Si les divers travaux relatifs aux moteurs de recherche intègrent bien la dimension temporelle, la majorité des systèmes de recherche d'informations géographiques semble souvent mettre l'accent sur les informations spatiales. Notre système est quant à lui axé principalement sur la dimension temporelle des clés d'indexation, et à ce titre propose une prise en compte de l'information temporelle qui est plus étendue que ce qui est

```
<text id='F051230A_0098.txt' date='20051230-15:59'>
<header> SPF016 3 GEN 0092 F BELGA-0176 </header>
<title> VTT -  F. Meirhaeghe engagé pour <timex cat="Time+Duree">3 ans</timex>
      chez Landbouwkrediet (1LEAD).
<PROP>
</title>
<PARAG>
   ANDERLECHT
   <timex grain="DAY" val="30-12-2005" partof="0" fuzzy="0">30/12</timex>
   ( BELGA )  =  Le coureur <CLASSIF code='4839' weight='1'> belge </CLASSIF>
   de mountainbike Filip Meirhaeghe a signé
   <timex grain="DAY" val="29-12-2005" partof="0" fuzzy="0">jeudi</timex>
   un <CLASSIF code='164' weight='1'> contrat </CLASSIF> de trois ans
   avec la formation cycliste Landbouwkrediet-Colnago,
<PROP>
   a <CLASSIF code='31' weight='2'> communiqué à la presse</CLASSIF>,
   <timex grain="DAY" val="30-12-2005" partof="0" fuzzy="0">vendredi</timex>,
<PROP>
   la direction de l'équipe dirigée par Gérard Bulens.
<PROP>
<PARAG>
   Meirhaeghe, suspendu depuis l'
   <timex grain="SEASON" val="S2-2004" partof="0" fuzzy="0">été 2004</timex>
   <timex grain="ITVL" partof="0" fuzzy="0">
     <bound type="lower" grain="unknown" val="unknown"  partof="0" fuzzy="0" />
     <bound type="upper" grain="DAY" val="14-1-2005" partof="0" fuzzy="0"  />
     jusqu' au 14 janvier 2005
   </timex>
<PROP>
   pour avoir utilisé de l'EPO,
<PROP>
   a été engagé pour cette période de 3 ans afin
<PROP>
   de pouvoir se préparer au mieux pour les
   <CLASSIF code='1530' weight='2'> Jeux OLympiques</CLASSIF> de Pékin
   <timex grain="YEAR" val="2008" partof="0" fuzzy="1">en 2008</timex>.
<PROP>
   Le biker flandrien sera alors âgé de 37 ans.
<PROP>
   .
<PROP>
   /.
<PROP>
   TED
<PARAG>
</text>
```

Figure 8.4 : Exemple d'un texte annoté thématiquement et temporellement.

généralement présenté dans le contexte des systèmes de recherche d'informations. Enfin, la méthode utilisée pour lier les repères temporels aux indices thématiques permet de travailler au niveau de la proposition alors que les différents travaux opèrent une échelle allant du document (Vicente-Diez et Martinez [2009]) jusqu'au paragraphe ou à la phrase (Pasca [2008], Le Parc-Lacayrelle *et al.* [2007], Strötgen *et al.* [2010]).

8.4 Évaluation

8.4.1 Corpus d'évaluation

Notre évaluation a été réalisée sur le corpus de 365 dépêches de presse Belga déjà utilisé pour l'évaluation de l'extraction et de l'interprétation temporelle (Section 7.9). La ressource choisie comme

```
<text id='F051230A_0098.txt' date='20051230-15:59'>

<classif id='4839' label='Belgique' weight='1'>
  <timex grain="DAY" val="30-12-2005" partof="0" fuzzy="0"> </timex>
  <timex grain="DAY" val="29-12-2005" partof="0" fuzzy="0"> </timex>
</classif>

<classif id='164' label='droit des contrats _ contrat#signature de contrat
                                            _ conclusion de contrat' weight='1'>
  <timex grain="DAY" val="30-12-2005" partof="0" fuzzy="0"> </timex>
  <timex grain="DAY" val="29-12-2005" partof="0" fuzzy="0"> </timex>
</classif>

<classif id='31' label='communiqué de presse _ communication à la presse
                                             _ déclaration à la presse' weight='2'>
  <timex grain="DAY" val="30-12-2005" partof="0" fuzzy="0"> </timex>
</classif>

<classif id='1530' label='jeux Olympiques' weight='2'>
  <timex grain="YEAR" val="2008" partof="0" fuzzy="1"> </timex>
</classif>

</text>
```

Figure 8.5 : Exemple d'un fichier de résultat.

base du processus de classification est le thésaurus Eurovoc de la communauté européenne. Celui-ci n'est pas tout à fait adapté à l'indexation de dépêches de presse, ce qui entraîne une absence d'indexation sur certains thèmes absents ou peu couverts dans le thésaurus. De plus, aucun traitement particulier, par exemple des cas ambigus, n'a été entrepris ici.

Pour l'évaluation de cette partie, l'attention s'est portée sur les liens effectués entre les catégories thématiques et les valeurs temporelles, afin de mettre en évidence le gain d'information apporté aux systèmes de recherche d'informations. Dans cette optique, on considère que le repérage des indices thématiques est *parfait*, les erreurs et oublis présents à ce niveau sont donc ignorés. De même pour la reconnaissance et l'interprétation des expressions temporelles. Ces erreurs ont déjà été évaluées par ailleurs (Sections 2.5.2 et 7.9) et leur conservation nuirait à l'évaluation du principe de lien thématico-temporel.

8.4.2 Procédure

La pertinence des indices thématiques, ainsi que l'interprétation des expressions temporelles reconnues, ne sont donc pas prises en compte ici. Plus précisément, afin d'isoler au maximum l'évaluation du lien thématico-temporel, certains éléments sont écartés s'ils représentent des fautes manifestes. C'est le cas des erreurs liées à l'ambiguïté, non gérée ici, et qui peut rapidement créer de nombreuses interférences. Par exemple, le thésaurus Eurovoc contient un catégorie « maïs » qui, étant donné les particularités des traitements relatifs à la classification, se retrouve attribuée à chaque mot « mais ». Les erreurs de reconnaissance d'expressions temporelles sont également écartées. Celles-ci sont cependant peu nombreuses. Une mauvaise interprétation d'une expression temporelle ne constitue par contre pas un obstacle à son rapprochement avec une catégorie thématique.

En pratique, l'évaluateur reçoit une liste telle que celle illustrée à la figure 8.5, qui contient pour

chaque document l'ensemble des catégories thématiques attribuées au document. Si certaines d'entre elles ont pu être liées à une ou plusieurs expressions temporelles, la mention de celles-ci est ajoutée. Seules les catégories qui ont reçu au moins une valeur temporelle sont vérifiées. Un code couleur est attribué suivant le cas :

- surlignement ROUGE, l'association n'est pas pertinente ;
- surlignement BLEU, l'association est pertinente, et la valeur temporelle correspond à la date d'émission de l'article ;
- surlignement VERT, l'association est pertinente, et la valeur temporelle est différente de la date d'émission de l'article ;
- surlignement ORANGE, l'association est pertinente, mais la valeur temporelle est imprécise ;
- caractères ORANGES, erreur de reconnaissance d'un indice thématique, due à l'ambiguïté.

8.4.3 Résultats

Une première approche des résultats obtenus montre que sur les 365 textes de départ, 341 (soit 93,42%), possèdent au moins une catégorie accompagnée d'une valeur temporelle. Au total, ce sont 1.798 catégories qui ont été dotées d'au moins un élément temporel. Par rapport au total de 5.406 catégories thématiques attribuées à l'ensemble des textes, cela représente un tiers de celles-ci (33,26%). Enfin, tous documents et catégories confondus, 2.420 liens thématico-temporels ont été effectués. Lorsqu'une catégorie thématique reçoit une dimension temporelle, celle-ci est donc en moyenne constituée de 1,35 valeurs temporelles.

Une analyse plus minutieuse est cependant nécessaire afin de souligner l'apport informationnel possible au moyen de l'analyse thématico-temporelle proposée. Pour ce faire, les liens établis entre les indices et catégories thématiques et les informations temporelles ont été vérifiés et jugés sur l'apport d'information que représente la dimension temporelle. Nous estimons qu'il y a plus-value informationnelle, lorsqu'un index thématique a été augmenté d'une dimension temporelle dont la valeur est différente de la date de l'article. À noter que certains textes ont été écartés, toujours pour les mêmes raisons que lors des évaluations précédentes, et que les liens entre une information temporelle et une catégorie thématique *victime* de l'ambiguïté lexicale ont également été rejetés. Les résultats sont repris aux tableaux 8.2 et 8.3.

Les chiffres repris au tableau 8.2 montrent une très bonne précision du lien entre catégorie thématique et valeur temporelle. En effet, seuls 21 liens (0,87%) ont été jugés incorrects, c'est-à-dire que la valeur temporelle ne se rapportait pas à l'élément d'information décrit par la catégorie thématique. Si les textes et liens écartés pour leur non-pertinence ne sont pas pris en compte, on dénombre donc 2.015 cas corrects sur 2.036, soit une précision de 98,97%[12]. Par conséquent, l'analyse de co-occurrence au niveau de la proposition étant suffisamment précise, il ne semble pas obligatoire d'avoir recours

[12] En comptant les liens écartés (textes inadéquats et erreurs dues à l'indexation thématique ou à l'extraction temporelles) comme des cas négatifs, ce qui est assez sévère, la précision s'établit encore à 83,26%.

	Nombre de liens	Proportion (%)
Liens incorrects	21	0,87
Liens corrects, avec date de l'article	720	29,75
Liens corrects, avec autre info. temporelle	1.260	52,07
Liens corrects, avec info. temporelle vague[11]	35	1,45
Liens écartés (ambiguïté thématique)	348	14,38
Liens écartés (texte écarté)	36	1,49
Total	2.420	100

Tableau 8.2 : Évaluation des liens thématico-temporels lors de l'indexation multidimensionnelle.

à une analyse syntaxique pour effectuer ce type de lien. Elle pourrait par contre être utile dans les situations où il serait nécessaire d'obtenir un lien plus précis (par exemple dans le cadre d'une tâche d'extraction d'informations sur des éléments bien déterminés).

En ce qui concerne la nature de l'information temporelle qui intervient dans les liens corrects, on constate, toujours au tableau 8.2, que pour plus de la moitié des liens trouvés (1.260 liens, soit 52,07%), il y a une réelle valeur ajoutée, c'est-à dire que la date est différente de la date d'émission de l'article. Cette dernière est elle explicitement attribuée à des catégories thématiques à 720 reprises (29,75% des liens effectués).

Le rappel est quant à lui difficile à chiffrer. Il imposerait une évaluation plus poussée et systématique, principalement au niveau des catégories thématiques, démarche que nous ne pouvons malheureusement pas entreprendre dans cette thèse. Rappelons également que les indices thématiques sont retrouvés à l'aide d'une ressource d'extraction dérivée d'Eurovoc. Ce thésaurus ne constitue probablement pas la ressource la plus adaptée à la catégorisation des dépêches de presse. Le rappel et la précision de l'aspect thématique de l'indexation pourrait dès lors très probablement être améliorés, ce qui devrait également avoir un impact sur le résultat final de l'indexation multidimensionnelle. L'utilisation d'Eurovoc constitue cependant une solution qui offre une couverture suffisante pour évaluer l'intérêt et la faisabilité d'une indexation thématico-temporelle.

Lors de la mise en relation entre catégories et informations temporelles, il est évidemment possible que plusieurs liens soient effectués sur une seule et même catégorie thématique, au sein du même texte. C'est l'ensemble fusionné de ces liens qui constitue *in fine* la dimension temporelle de la catégorie. Une analyse des résultats sur la base de ces catégories et de leur éventuelle dimension temporelle (fusionnée), et non plus sur l'ensemble des liens thématico-temporels, est probablement plus à même d'offrir une évaluation de l'apport réel d'information généré par l'ajout de la dimension temporelle. Ces résultats sont repris au tableau 8.3.

	Nombre de catégories	Proportion (%)
Avec dimension temporelle (dates non-vagues, date article exclue)	1.022	18,90
Avec dimension temporelle (toutes dates non-vagues)	1.477	27,32
Toutes	5.406	100

Tableau 8.3 : Évaluation de l'apport d'information temporelle aux catégories thématiques.

Sur les 5.406 catégories thématiques de départ, 1.022 (18,90%) ont pu être étendues avec une dimension temporelle qui apporte une réelle plus-value informationnelle. Ce ratio devrait cependant évoluer à la hausse avec l'élimination au moins partielle des catégories thématiques sélectionnées erronément à cause de l'ambiguïté. En effet, si celles-ci ne sont pas reprises dans le compte des indexations multidimensionnelles (1.022), elles sont cependant incluses dans le nombre total de catégories (5.406). Si tous les liens thématico-temporels corrects sont comptés, y compris ceux qui renseignent la date d'émission du texte, ce sont 1.477 catégories (27,32%) qui deviennent multidimensionnelles.

8.5 Perspectives et conclusion

Dans cet ultime chapitre, les analyses thématiques et temporelles ont été rassemblées afin d'expérimenter une méthode d'indexation peu répandue : l'indexation multidimensionnelle, et plus précisément thématico-temporelle. Celle-ci propose pourtant des perspectives intéressantes pour l'accès à l'information, que ce soit en termes d'expression des critères de recherche, d'ordonnancement des résultats ou encore du mode de présentation de ceux-ci.

Les techniques utilisées, qui ont été présentées en détail dans les chapitres 2 et 7, reposent sur des principes similaires – l'analyse par grammaires locales et transducteurs – et sont par conséquent facilement combinables. C'est également de ce type de techniques que résulte la finesse de l'analyse, tant en ce qui concerne les informations temporelles que les liens thématico-temporels. En ce qui concerne plus particulièrement ces derniers, si le système présenté permet de lier les indices thématiques à des valeurs temporelles au niveau subphrastique de la proposition, c'est grâce à l'indexation thématique qui conserve, en contexte dans le document d'origine, l'ensemble des indices (expressions) ayant contribué à la catégorisation. Cette traçabilité de l'indexation et de la catégorisation thématique permet de relier précisément les expressions temporelles et leurs valeurs aux indices, et par extension aux catégories, thématiques.

S'il est difficile d'obtenir une évaluation chiffrée très pointue sans faire appel à des experts humains, il est tout de même possible de dégager certains éléments probants à partir des résultats obtenus lors de l'expérimentation. Tout d'abord, la précision du lien entre indice thématique et valeur temporelle est assez élevée (98,97%[13]), et dans plus de la moitié des cas (52,07%), l'information temporelle apporte une plus-value informationnelle. Ensuite, la répartition des liens est assez uniforme puisque 93,42% des 365 documents en disposent au moins d'un. Au final, le caractère multidimensionnel de l'indexation a pu être effectivement créé pour 27,32% des catégories. L'information temporelle s'est avérée être différente de la date d'émission du texte pour 18,90% des catégories. Ces ratios sont à mettre en perspective avec les éléments suivants :

- toutes les catégories thématiques ne sont pas appelées à recevoir une dimension temporelle ;
- l'indexation thématique n'a pas fait l'objet d'un traitement poussé pour cette expérience (pas de traitement de l'ambiguïté, et thésaurus non-spécifique au domaine) ;

[13] Lorsqu'on écarte les textes inadéquats ainsi que les erreurs dues à l'indexation thématique ou à l'extraction temporelle.

– le calcul des ratios est tiré vers le bas suite à l'élimination, lors du comptage de l'indexation thematico-temporelle mais pas lors de celui de l'indexation thématique seule, des catégories *victimes* de l'ambiguïté[14].

Dès lors, ces résultats sont assez positifs et nous semblent de nature à encourager l'élaboration de moteurs de recherche capables d'exploiter une indexation multidimensionnelle, et thématico-temporelle en particulier.

La réalisation d'un tel moteur va bien au-delà de l'objectif de cette thèse. Cela ne nous a cependant pas empêché d'imaginer ce à quoi pourrait ressembler son interface d'interrogation. Une simple évolution des pages proposées actuellement est déjà à même d'apporter diverses améliorations à l'utilisateur. Au niveau des *interfaces homme-machine*, les possibilités d'exploitation sont assez nombreuses et variées et mériteraient une approche approfondie du problème.

L'ébauche imaginée (Figure 8.6[15]) n'a pas la prétention d'une telle étude, mais illustre, dans un contexte *plausible*, les possibilités offertes par l'indexation multidimensionnelle telle que nous la proposons. Ainsi, comme nous le proposions déjà en partie à la section 1.4.3, la démarche de recherche pourrait se dérouler de la manière suivante :

1. L'utilisation directe de catégories est un moyen de recherche peu attrayant pour la majorité des utilisateurs (Dalbin [2007], voir section 1.4.3). Il semble donc préférable que ceux-ci commencent par entrer leur requête sous la forme de mots-clés libres, dans le champ texte prévu à cet effet.
2. Des techniques d'extension de requêtes peuvent alors intervenir de manière à enrichir la recherche initiale. Cela doit permettre d'augmenter le rappel, éventuellement au détriment de la précision.
3. Les documents retrouvés sur la base de cette requête (enrichie) sont regroupés en fonction des catégories thématiques qui leur avaient été attribuées[16].
4. L'affichage classique d'une liste de résultats est alors accompagnée, dans la colonne de gauche, par une hiérarchie de catégories. Celle-ci permet d'indiquer à l'utilisateur quels sont les concepts ou sujets qui semblent les plus pertinents par rapport à sa recherche. De cette manière, la rapidité d'une recherche fructueuse par mots-clé est conservée, tout en fournissant un outil sémantique pour clarifier le sens d'une requête complexe ou ambiguë.
5. Dans ce dernier cas, la sélection d'une ou plusieurs catégories par l'utilisateur (mécanisme de *relevance feedback*) adapte la liste de résultats proposée. Cette opération doit permettre d'améliorer la précision et ainsi réduire, voire dépasser, l'éventuel effet négatif engendré à cet égard par l'extension de requête.
6. Enfin, l'aspect temporel peut être activé à tout moment au moyen d'un calendrier et

[14] Car les catégories qui n'ont pas reçu de dimension temporelle n'ont pas été vérifiées.

[15] Cette interface, purement fictive, s'inspire de celle proposée par le moteur de recherche Google. Les catégories reprises dans la colonne de gauche sont issues du thésaurus Eurovoc.

[16] Un document peut généralement recevoir plusieurs catégories, il peut donc aussi apparaître dans plusieurs groupes.

d'une ligne du temps (disposés au-dessus de la liste de résultats). Ces deux représentations de l'espace temporel devraient permettre la sélection d'un point précis ou d'une zone temporelle plus étendue. Le calendrier permet de se situer à l'échelle du jour, alors que la ligne du temps permet d'utiliser des granularités plus élevées ou plus faibles si elle est combinée au calendrier. D'autre part, la ligne du temps et le calendrier peuvent également servir à montrer la répartition des informations thématiques dans la zone temporelle sélectionnée[17], en plus de l'expression de la composante temporelle de la recherche.

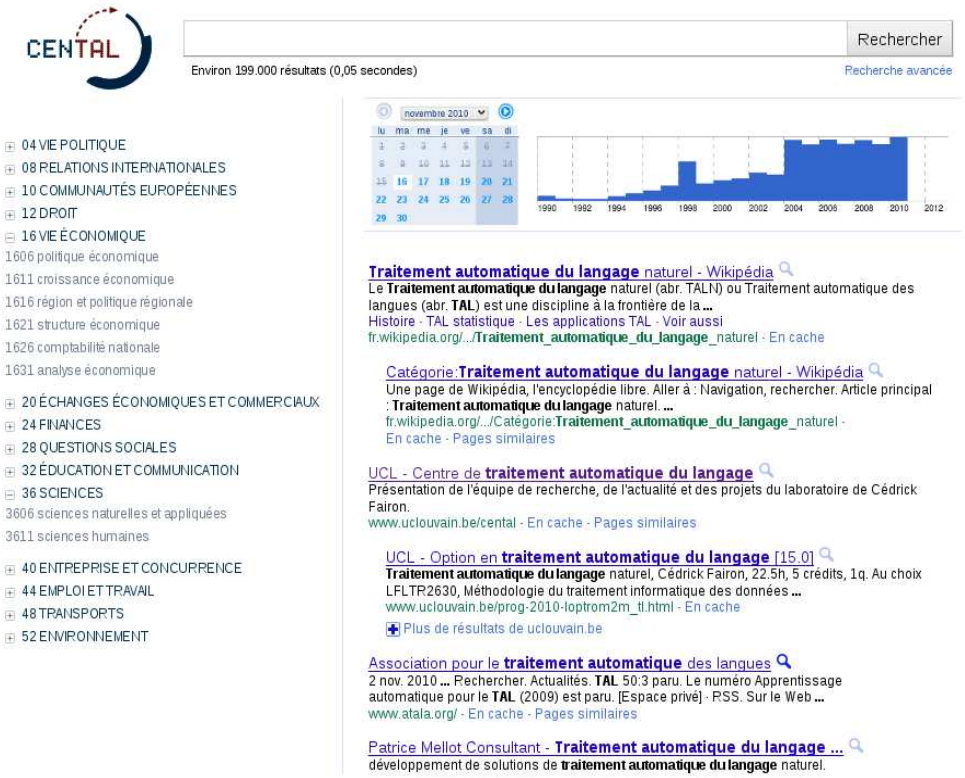


Figure 8.6 : Interface fictive possible pour un moteur de recherche thématico-temporel.

Évidemment de nombreuses autres perspectives et moyens d'accéder à l'information pourront être imaginés. Par exemple, un outil qui pourrait s'avérer très pratique proposerait une vue synthétique (extraits de textes, événements, sujets ou catégories à la une, etc.) des informations disponibles pour une période ou une date déterminée. Une présentation sous la forme d'un nuage de mots ou de catégories peut également être proposée. Ce type d'interface permettrait, entre autres, à un journaliste qui s'intéresse à des événements qui ont eu lieu à un certaine période (par exemple, les élections de juin 2010), d'obtenir une vue sur le contexte dans lequel ceux-ci s'insèrent.

Avec ce type d'interfaces, et grâce à l'indexation multidimensionnelle et à l'utilisation de catégories, l'utilisateur aurait donc la possibilité de formuler des recherches dont les termes seraient porteurs d'un sens précis et non-ambigu, que ce soit au niveau thématique ou temporel.

[17] Cette option est d'ailleurs proposée par Google dans ses *Search Tools*.

Perspectives et conclusion générales

Face au défi que représente l'accès à l'information, et dans le contexte d'une masse documentaire électronique toujours plus importante, nous avons constaté que la majorité des outils de recherche fonctionnent au niveau de l'espace des mots. Afin de maximiser la couverture et la précision d'une recherche par rapport à une collection de documents, il peut être profitable de passer d'un espace de mots à un espace de concepts, ce qui revient à se déplacer du sommet *symbol* au sommet *tought of reference* du triangle de référence (voir section 3.4). Ce principe peut être mis en œuvre par un enrichissement sémantique de la représentation du document lors de son indexation. Nous avons cherché à voir dans quelle mesure certaines techniques de traitement automatique du langage pouvaient être exploitées dans ce sens, pour contribuer à une meilleure mise en relation de la demande et de l'offre documentaire

L'information dispose d'une dimension thématique qui définit ce sur quoi elle porte. Nous avons observé qu'elle peut également être accompagnée de dimensions supplémentaires, telles que des informations temporelles (une date ou une période de temps) ou géographiques (un lieu), qui s'apparentent à des *métadonnées* relatives à l'information thématique. L'ensemble de ces éléments constitue une information multidimensionnelle. Alors que la dimension thématique est très large, et très variable du point de vue de son expression et de son apparition dans les textes, les dimensions temporelle et géographique sont elles beaucoup plus stables de ce point de vue.

En fonction de cette stabilité, nous avons proposé une approche générale et adaptable pour traiter les données thématiques, alors que nous avons consenti un investissement important dans un traitement très spécifique en ce qui concerne la dimension temporelle. Ces deux approches ont en commun le souci de dépasser l'espace des mots pour atteindre celui des concepts. Le texte sémantiquement enrichi par ces traitements, offre alors des perspectives pour un accès à l'information selon le sens et non plus selon son expression.

Du point de vue thématique, l'approche que nous avons proposé consiste en une indexation thématique originale, basée sur des techniques symboliques. Son objectif est d'apporter un sens à l'indexation par l'attribution de catégories définies (classification). Cette tâche, traditionnellement effectuée manuellement par des documentalistes experts, présente l'inconvénient de nécessiter un investissement humain et financier conséquent. Elle s'accompagne par ailleurs d'un certain manque de cohérence à long terme, ce qui réduit quelque peu l'apport d'une indexation humaine. Notre solution propose un processus (semi-)automatique, conçu comme une aide à l'indexation. Une utilisation complètement automatique est également envisageable si la quantité de documents à traiter est im-

portante, la validation pouvant ensuite être menée au fur et à mesure. Cette démarche conserve ainsi la validation humaine tout en rendant la tâche plus rapide, et donc financièrement plus abordable. Le système présente également un comportement plus systématique et cohérent dans le choix des catégories.

Concrètement, la méthode mise en œuvre s'appuie sur une ressource terminologique telle qu'un thésaurus ou une ontologie. Celle-ci définit les catégories thématiques qui peuvent être assignées aux documents. L'attribution de ces catégories est réalisée sur la base d'indices thématiques – des mots simples ou des expressions composées – retrouvés dans les textes. Alors que la plupart des systèmes similaires commencent par extraire des *concepts* pour ensuite les comparer avec les catégories disponibles, nous adoptons la démarche inverse, dans laquelle l'extraction de termes est guidée par la ressource terminologique dès le départ. Par rapport aux nombreux travaux basés sur un apprentissage artificiel, notre approche se distingue aussi par l'absence de phase d'entraînement, ce qui la rend fonctionnelle et indépendante de l'existence d'un corpus annoté dès le premier document. De plus, elle évite la difficulté à traiter des catégories jugées rares par l'apprentissage en raison de leur faible distribution dans l'ensemble d'entraînement. L'originalité de notre approche tient également à la génération automatique de la ressource d'extraction à partir de la ressource terminologique, sous la forme de transducteurs. Cette étape peut donc être répétée quel que soit le domaine, pourvu qu'il y ait une ressource adéquate disponible. De cette manière, l'effort nécessaire à la production des ressources linguistiques d'extraction, point faible souvent mis en avant pour les méthodes symboliques, est minimisé et la méthode s'avère donc reproductible. Notre préoccupation de départ pour le traitement de la dimension thématique de l'information – l'adaptabilité – est donc bien rencontrée.

Pour la dimension temporelle, notre objectif était l'enrichissement sémantique de ce type d'information *via* l'affectation d'une valeur temporelle univoque et normalisée. Les textes en langage naturel abondent d'indications relatives au temps, mais celles-ci ne sont pas toujours explicites. Il est dès lors difficile de les relier directement à une valeur précise dans un espace temporel normalisé, par exemple un calendrier. Il en résulte que de nombreuses applications, qui pourraient tirer parti d'informations temporelles, sous-exploitent cet aspect.

L'expression du temps en langage naturel est assez variée, mais présente cependant une régularité suffisante pour pouvoir être décrite manuellement. L'investissement dans une telle description manuelle est motivé par la richesse d'analyse qui peut alors être obtenue et se justifie par le fait que les ressources ainsi produites peuvent être exploitées dans de nombreux contextes. Au delà des divers systèmes de TAL (recherche d'informations, extraction d'informations, traduction automatique, systèmes de question-réponse, etc.), d'autres applications plus éloignées pourraient tirer parti d'un module d'analyse temporel. L'analyse de processus, le planning, l'aide à la décision ou encore la gestion de projets sont autant de domaines pour lesquels l'analyse automatique de documents peut fournir de précieuses informations, entre autres temporelles[18]. Nous avons donc créé manuellement une série de transducteurs. Cette approche classique permet de reconnaître et d'annoter les expressions temporelles, principalement adverbiales, et cela de manière fine et détaillée, tout en assurant de

[18] Par exemple, la génération de réseaux PERT (Program – or Project – Evaluation and Review Technique) pour l'ordonnancement des tâches d'un projet, comme suggéré par Allen [1991].

manière progressive une couverture la plus importante possible. Cette ressource est exploitée dans un système d'extraction et d'interprétation des expressions temporelles, développé pour la langue française.

Les caractéristiques principales de ce système résident dans la prise en compte de l'imprécision, ainsi que l'exploitation d'un riche système de granularités. La combinaison de ces éléments a permis de mettre au point des mécanismes d'interprétation capables de faire face à la richesse et à l'imprécision naturelle de la langue. L'analyse automatique mise sur pied prouve également que, même sans atteindre une analyse linguistique parfaite de l'ensemble des phénomènes reliés au temps, il est possible d'extraire de manière pragmatique une information suffisamment complète et précise pour être exploitée dans une application concrète.

Le travail important de création de la ressource d'extraction temporelle a été organisé selon une méthodologie visant la maîtrise du processus de développement. Cette méthode est centrée autour d'un document de référence – la spécification des expressions temporelles – autour duquel s'organise de manière cohérente l'ensemble des étapes de développement du système. L'intérêt est à la fois théorique – pour la descriptions des expressions temporelles – et pratique – par la définition d'une marche à suivre pour le développement des ressources linguistiques et du système d'analyse. À notre connaissance, peu de publications abordent une telle démarche, ainsi que la spécification des expressions temporelles qui en résulte.

Finalement, le traitement des dimensions thématique et temporelle de l'information ont été réunis. Si l'apport de sens pour chacune de ces dimensions constitue déjà un pas important, leur prise en compte conjointe apporte une plus-value informationnelle supplémentaire. Le lien entre les dimensions est réalisé selon une approche simple de co-occurrence, mais qui permet d'atteindre une très bonne précision car elle est réalisée au niveau de la proposition. Cette caractéristique constitue une originalité par rapport à la majorité des systèmes similaires (voir section 8.2), qui travaillent plutôt sur des unités plus larges : phrase, paragraphe ou même texte entier. L'adoption de l'indexation dans un espace de concepts plutôt que de mots, ainsi que l'évolution de cette indexation d'un mode monodimensionnel, uniquement thématique, à un mode multidimensionnel, thématique et temporel, est susceptible d'améliorer l'accès à l'information. Pour ce faire, il existe diverses techniques qui permettent de tirer parti de cet apport sémantique. Comme nous l'avons suggéré, en proposant les étapes de traitement à mettre en œuvre pour intégrer cette évolution aux moteurs de recherche, le défi consistera à combiner plusieurs technologies afin d'offrir des outils sémantiques plus performants à l'utilisateur.

Annexes

ANNEXE A

EXTRAIT DE THÉSAURUS DOCUMENTAIRE

Cette annexe reprend un extrait du thésaurus utilisé pour les expériences de classification de la section 2.5.2. Il s'agit plus précisément d'une partie du microthésaurus consacré à l'agriculture.

```
MT171    agriculture

  166      agriculture

    121      moyen de production agricole
      477      engrais
      541      exploitation agricole
        1493     materiel agricole
        2422     batiment agricole
      967      semence
    50       politique agricole
      1130     travail agricole
        2134     ouvrier agricole
        3881     centrale paysanne
        49       agriculteur
        97       association agricole
      114      recherche agronomique
        1171     orientation agricole
        3101     developpement agricole
        732      modernisation agricole
      1445     administration des services techniques de l'agriculture
        1496     service agricole
        3248     station de chimie agricole
        971      service d'economie rurale
        986      service veterinaire
          1108     veterinaire
          723      medicament veterinaire
      2583     economie rurale
        152       bien rural
        929       remembrement
        931       rendement agricole
        945       revenu de reference
      5035      campagne agricole
      5182      quota-laitier
      5370      politique agricole commune
```

```
860      production agricole
  131      elevage
    1120     volaille
    150      betail
      1135     amelioration de la race
      1516     police sanitaire du betail
      4386     animal de boucherie
    2017     apiculture
    3712     maladie animale
      1143     fievre aphteuse
      1149     rage canine
      1160     brucellose bovine
      1903     epizootie
      5046     leucose
      5340     encephalopathie spongiforme des bovins
    5052     saillie
    64       alimentation animale
  3682     production animale
    2        abattoir
      2886     equarrissage
```

ANNEXE B

SPÉCIFICATION DÉTAILLÉE DES EXPRESSIONS TEMPORELLES

B.1 Introduction

La présente spécification est un point de repère central à l'ensemble du processus de développement du système d'analyse des expressions temporelles (voir section 7.4). Elle sert de registre répertoriant les cas pris en compte, de base à la construction des grammaires locales, de mini corpus de test, de référence pour implémenter un logiciel capable de parser et ensuite interpréter les expressions extraites. Elle s'inscrit aussi dans une dynamique d'enrichissement, entre autres suite aux résultats observés lors du repérage (à l'aide des grammaires) et de l'interprétation (par le logiciel) des expressions.

Les aspects importants consistent donc à définir quelles expressions sont reconnues, de quelles catégories elles resortissent et enfin de quelle annotation elles vont bénéficier.

B.1.1 Formats d'annotation

L'**extraction** des expressions temporelles consiste en une annotation à l'aide des grammaires Unitex. Le jeu d'étiquettes exploitées est exposé à la section B.3. Cette annotation, adopte les principes suivants :

- l'expression temporelle complète est encadrée par des accolades : {...}.
- la catégorie attribuée à l'expression, au moyen de plusieurs codes combinés par le signe « + », suit directement celle-ci et en est séparée par les signes « ,. ».
- les différents sous-constituants sont encadrées par des crochets : [...].
- la caractérisation d'un élément encadré par des crochets est donnée par un code en majuscule précédé du signe « # ». Celui-ci est situé à la fin de l'encadrement ([...#CODE]).
- l'imbrication d'encadrements *à crochets* est possible,
- un encadrement *à crochets* peut apparaître vide (à l'exception du code).

Pour des raisons pratiques de lisibilité et de clarté, la syntaxe de la spécification n'est pas totalement identique à celle de l'annotation. L'objectif poursuivi consiste surtout à alléger certaines notations et à autoriser l'usage d'éléments ayant une portée plus générale. Le lien entre les deux formats reste donc très clair et ne nécessite pas une longue discussion. La correspondance entre les deux notations est

reprise à la section B.3. L'exemple ci-dessous illustre quelques unes des différences. À l'expression « à 18 heures le 22 décembre 2009 » correspond :

l'annotation

à [[18#H_C] [heures#GRAIN=HOUR] #HEURE] le [[22#J_C] [décembre#M_L] [2009#A_C] #DATE]

la spécification

[à | le] HEURE[$\text{Heure}_{(1)}$] DATE[$\text{Jour}_{(1)}\text{Mois}_{(N)}\text{An}_{(1)}$]

B.2 Conventions de notation

B.2.1 Cardinalités

- En l'absence de toute indication, la cardinalité par défaut est $^{(1,1)}$, c'est-à-dire *un et un seul*.
- Une autre cardinalité peut être assignée avec la notation $^{(min,max)}$.
- Dans la spécification des *signatures*, les crochets « [ ] » sont utilisés comme délimiteurs de groupe. Ils n'ont donc pas de signification en terme de cardinalité. Par contre, dans les exemples, les crochets sont uniquement utilisés pour exprimer le caractère facultatif d'un élément.
- L'opérateur « OU » est exprimé à l'aide du signe « | », par exemple : (un | deux | trois). Dans le contexte de cette notation, il est possible d'employer le signe <E>, qui désigne un élément vide.

B.2.2 Signes particuliers

- Le symbole « * », s'il ne se trouve pas dans le cadre d'une cardinalité, signifie n'importe quel élément parmi les valeurs possibles (*wildcard*, dépend du contexte).

B.3 Définition des éléments d'annotation des expressions temporelles

Cette section définit et explique les différents éléments qui peuvent intervenir dans une annotation. Une étiquette (ou tag) peut être :
– **fixe** (ex. : HEURE, DATE, PRECISE, etc.),
– **variable** (ex. : GRAIN=*, où * représente un ensemble de valeurs possibles).
D'autre part, les étiquettes peuvent également être :
– **simples** (ex. : PRECISE),
– **composées** (ex. : pour DATE, qui sera alors noté DATE[ ], et dont la structure interne doit être caractérisée par d'autres étiquettes).
Pour les étiquettes composées, dans un souci de rendre la spécification la plus lisible possible, certaines notations utilisées pour l'annotation automatique ont été remplacées par un équivalent plus clair et explicite (voir l'exemple de comparaison à la section B.1.1). Dans les tableaux suivants, les deux notations sont mentionnées en parallèle.

D'une manière générale « $_{(1)}$ » désigne une entité composées de chiffres, « $_{(I)}$ » une entité en chiffres romains, et « $_{(a)}$ » une entité désignant un nombre, mais écrit en toutes lettres. « $_{(N)}$ » est utilisé pour un élément « nommé » (un nom de jour, de mois, etc.) et « $_{(US)}$ » est un cas particulier (mois en format US).

B.3.1 Étiquettes variables

GRAIN=*	Granularité (unité)	
$G_{(*)}$	Notation équivalente pour la spécification	
	Valeurs	Signification
	MIL	Millénaire
	CENT	Siècle
	DEC	Décénie
	YEAR	Année
	SEASON	Saison
	HYEAR	Semestre (6 mois)

GRAIN=*	Granularité (unité)	
$G_{(*)}$	Notation équivalente pour la spécification	
	Valeurs	Signification
	TYEAR	Quadrimestre (4 mois)
	QYEAR	Trimestre (3 mois)
	MONTH	Mois
	DAYS	Période de plusieurs jours ne correspondant pas à une division bien établie du calendrier
	WEEK	Semaine
	WE	Week-end
	DAY	Jour
	HOUR	Heure
	MIN	Minute
	SEC	Seconde
	POD_DA	Aube
	POD_MO	Matin
	POD_MI	Midi
	POD_DT	Journée
	POD_AF	Après-midi
	POD_EV	Soir
	POD_DU	Crépuscule
	POD_NI	Nuit
	UNDEF	Non défini (code temporaire, à transformer en valeur porteuse de sens !)

GRAIN peut être trouvé comme élément racine d'une annotation ou imbriqué dans les balises DATE et HEURE. Au moins une indication GRAIN doit obligatoirement être fournie dans toute annotation.

Lorsque plusieurs granularités aparaissent dans l'annotation des expressions temporelles, par défaut, c'est la granularité la plus fine qui l'emporte. Certaines granularités sont cependant prioritaires sur d'autres (WE et WEEK sur DAY).

$G_{(pod\_*)}$ équivaut à ($G_{(pod\_da)}$ | $G_{(pod\_mo)}$ | $G_{(pod\_mi)}$ | $G_{(pod\_dt)}$ | $G_{(pod\_af)}$ | $G_{(pod\_ev)}$ | $G_{(pod\_du)}$ | $G_{(pod\_ni)}$)

PARTOF=*	Indicateur de partition (partie d'une unité temporelle)	
$PartOf_{(*)}$	Notation équivalente pour la spécification	
	Valeurs	Signification
	BEG	Interprétation : Début
	MID	Interprétation : Millieu
	END	Interprétation : Fin

FUZZY=*	Indicateur d'imprécision.	
$Fuzzy_{(*)}$	Notation équivalente pour la spécification	
	Valeurs	Signification
	1	Imprécision interne à la zone temporelle désignée
	2	Imprécision interne et externe à la zone temporelle désignée

REF=*	Point de référence	
$Ref_{(*)}$	Notation équivalente pour la spécification	
	Valeurs	Signification
	NOW	Référence déictique, le point de référence est le moment de la parole
	FOCUS	Référence relative, anaphorique, le point de référence est un point à déterminer dans le contexte

MOVE=*	Direction du déplacement	
$Move_{(*)}$	Notation équivalente pour la spécification	
	Valeurs	Signification
	REW	Explorer dans le passé
	FWD	Explorer vers le futur

NB=*	Spécification d'un nombre d'unités, utilisé pour un déplacement temporel (amplitude)	
$Amp_{(*)}$	Notation équivalente pour la spécification	
	Valeurs	Signification
	[0,N]	Valeur numérique (nombre d'unités)

TARGET=*	Caractérisation de la cible (zone temporelle) désignée par l'expression temporelle. La cible doit respecter les contraites exprimées par l'expression contenue dans TARGET	
$TARGET_{(*)}$	Notation équivalente pour la spécification	
	Valeurs	Signification
	FULL	Désigne de manière complète une zone temporelle (équivalent PAPU)
	PART	Désigne de manière partielle une zone temporelle (équivalent PRPU)

FOCTEMP=*	Caractérisation du point de repère (focus) temporel à partir duquel un déplacement temporel devra être effectué afin d'atteindre la cible finale. Le focus temporel, qui doit lui même être localisé dans l'espace temporel, doit respecter les contraintes exprimées par l'expression contenue dans FOCTEMP.	
$FOCTEMP_{(*)}$	Notation équivalente pour la spécification	
	Valeurs	Signification
	FULL	Désigne de manière complète une zone temporelle (équivalent PAPU)
	PART	Désigne de manière partielle une zone temporelle (équivalent PRPU)

B.3.2 Étiquettes fixes simples

Valeurs numériques			
Annotation	Spécification	Signification	Commentaire
NB_C	$\text{Nb}_{(1)}$	Nombre entier exprimé en chiffres	Valeur directe
NB_CR	$\text{Nb}_{(I)}$	Nombre entier exprimé en chiffres romains	Valeur à traduire en chiffres
NB_L	$\text{Nb}_{(a)}$	Nombre entier exprimé en lettres	Valeur à traduire en chiffres
NBF_L	$\text{Nbf}_{(a)}$	Nombre (fraction) exprimé en lettres	Valeur à traduire en chiffres

Autres			
Annotation	Spécification	Signification	Commentaire
PRECISE	Precise	Indicateur de référence précise	Présent à titre indicatif, ne doit pas être interprété

B.3.3 Étiquettes fixes composées

DATE[ ]	Date			
Annotation	Spécification	Signification	Valeurs	Commentaire
GRAIN=	$\text{G}_{(*)}$	Granularité	voir supra	Valeur (définie par nomenclature)
S_C	$\text{Siecle}_{(1)}$	Siècle exprimé en chiffres	[1,...]	Valeur directe
S_L	$\text{Siecle}_{(a)}$	Siècle exprimé en lettres	[premier,deuxième,...]	Valeur à traduir en chiffres
S_CR	$\text{Siecle}_{(CR)}$	Siècle exprimé en chiffres romains	[I,II,...]	Valeur à traduir en chiffres
J_C	$\text{Jour}_{(1)}$	Jour exprimé en chiffres	[1,31]	Valeur directe (calendrier)
J_L	$\text{Jour}_{(a)}$	Jour exprimé en lettres	[un,trente et un]	Valeur à traduir en chiffres
J_N	$\text{Jour}_{(N)}$	Nom de jour	[lundi,dimanche]	Valeur à convertir en valeur calendrier
M_C	$\text{Mois}_{(1)}$	Mois exprimé en chiffres	[1,12]	Valeur directe (calendrier)
M_N	$\text{Mois}_{(N)}$	Nom de mois	[janvier,décembre]	Valeur à convertir en valeur calendrier

DATE[ ]	Date			
Annotation	Spécification	Signification	Valeurs	Commentaire
M_US	$\text{Mois}_{(US)}$	Mois (format US)	[JAN,DEC]	Valeur à convertir en valeur calendrier
A_C	$\text{An}_{(1)}$	Année exprimée en chiffres	[1984,2010,...]	Valeur directe (calendrier)
A_L	$\text{An}_{(a)}$	Année exprimée en lettres	[deux mille,...]	Valeur à convertir en valeur calendrier
DEC_C	$\text{Décénie}_{(1)}$	Décénie exprimée en chiffres	[1910,2090,...]	Valeur directe (calendrier)
			[10,90]	Valeur à interpréter (19..)
DEC_L	$\text{Décénie}_{(a)}$	Décénie exprimée en lettres	[dix,nonante]	Valeur à convertir et/ou interpréter
NAMED	Named	Période nommée	[Noël,Pâques,...]	Valeur à convertir en valeur calendrier
SEASON	Season	Saison	[Printemps,hiver]	Valeur à convertir en valeur calendrier
ORD_C	$\text{Ord}_{(1)}$	Ordinal exprimé en chiffres	[1er,4eme]	Valeur à interpréter conjointement avec GRAIN
ORD_L	$\text{Ord}_{(a)}$	Ordinal exprimé en lettres	[premier,quatrième]	Valeur à convertir et interpréter conjointement avec GRAIN
POY	Poy	Partie de l'année (part-of-year)	[tri-,se-mestre]	Ne doit pas être interprété -> pris en charge par grain

HEURE[ ]	Heure			
Annotation	Spécification	Signification	Valeurs	Commentaire
GRAIN=	$\text{G}_{(*)}$	Granularité	voir supra	Valeur (définie par nomenclature)
H_C	$\text{Heure}_{(1)}$	Heure exprimée en chiffres	[1,24]	Valeur directe (calendrier)
H_L	$\text{Heure}_{(a)}$	Heure exprimée en lettres	[un,vingt-quatre]	Valeur à traduir en chiffres
H_M_C	$\text{Min}_{(1)}$	Minute exprimée en chiffres	[1,60]	Valeur directe (calendrier)
H_M_L	$\text{Min}_{(a)}$	Minute exprimée en lettres	[un,soixante]	Valeur à traduir en chiffres
H_S_C	$\text{Sec}_{(1)}$	Seconde exprimée en chiffres	[1,60]	Valeur directe (calendrier)
H_S_L	$\text{Sec}_{(a)}$	Seconde exprimée en lettres	[un,soixante]	Valeur à traduir en chiffres
H_PJ=	PJour	Partie de la journée	AM	Avant-midi
			PM	Après-midi

B.3.4 Étiquettes variables composées

BOUND=*	Borne d'un intervalle	
$BOUND_{(*)}$	Notation équivalente pour la spécification	
	Valeurs	Signification
	UPPER	Interpréter comme une expression indépendante. Utiliser ensuite en tant que borne supérieure
	LOWER	Interpréter comme une expression indépendante. Utiliser ensuite comme borne inférieure
Elements	Signification	Commentaire
GRAIN=	information de granularité	
DATE	date	peut être complète (équivalent PAPU) ou partielle (équivalent PRPU)
HEURE	heure	
TEMPSHIFT	déplacement temporel	

B.4 Catégories d'expressions temporelles, description et spécification des principaux cas

B.4.1 Remarques préliminaires

D'une manière générale, l'ordre des éléments de la signature respecte l'ordre des éléments dans l'expression. Cependant, dans les cas où certaines parties peuvent être inversées, une seule spécification est donnée, sachant que certaines inversions sont possibles.

Pour des raisons de lisibilité, l'étiquette $G_{(*)}$ (GRAIN=) ne sera généralement pas reprise dans les exemples (mais est donnée dans la spécification du cas général correspondant). Comme nous l'avons déjà mentionné (voir section B.3), pour une expression temporelle donnée, c'est la granularité la plus fine qui l'emporte (sauf exception : WE et WEEK). En ce qui concerne la spécification des heures, les cas où l'expression contient l'information d'heure et de minute ne sont pas distingués de ceux dans lesquels seule l'heure est mentionnée (15h équivaut à 15h00). Par conséquent, la granularité correspond donc à la minute.

Pour chaque catégorie, un certain nombre de cas généraux sont distingués, spécifiés et illustrés par leurs différentes réalisations. Toutes les possibilités ne sont cependant pas toujours explicitement illustrées. La quantité d'exemples présentés est néanmoins ajustée de manière à ce qu'elle soit suffisante pour appréhender la variation induite par la catégorie générale.

Les cas généraux sont numérotés de (1) à (N). Ces identifiants sont identiques à ceux qui se trouvent dans les graphes ainsi qu'à la section 7.6.2. Une subdivision (chiffres romains en caractères minuscules) est parfois utilisée. La spécification des cas généraux est présentée en caractères gras et en couleur : le **bleu** est utilisé pour des éléments fixes, qui ne varient pas pour le cas en question, alors que le **vert** est employé pour désigner les éléments qui permettent de différencier les différentes réalisations particulières de ce cas général. Dans la spécification du cas particulier, les éléments bleus sont remplacés par «. . .». Pour plus de lisibilité, d'autres éléments sont imprimés en **gris**, afin de ne faire ressortir que les éléments caractéristiques du cas. L'exemple ci-dessous illustre ces conventions :

(7) **DATE[** $G_{(day)}$ **]** **[Fuzzy]** $G_{(pod\_*)}$ **HEURE[** $G_{(min)}$ **]**
DATE[$Jour_{(1)}Mois_{(1)}An_{(1)}$] … HEURE[$Heure_{(1)}Min_{(1)}$] 22/12/2009, (à | vers) l'aube, à 5 heures et 24 minutes $G_{(pod\_da)}$

Dans certains cas, des expressions peu vraissemblables ou impossibles sont mentionnées lorsqu'elles constituent des exceptions aux règles générales de composition des expressions temporelles. Elles sont alors présentées complètement en **gris** et précédées du signe « N/A ».

B.4.2 PAPU : Référence Ponctuelle, Absolue, Précise et Unique

Spécification générale : $DATE\ \left(G_{(*)}\right)\ G^{(0,1)}_{(pod\_*)}\ HEURE^{(0,1)}\ \left(G_{(min)}\right)$

Spécification	Exemple	Commentaires
(1) DATE[$G_{(year)}$]		
DATE[$An_{(1)}$]	2004	
l' an | année DATE[$An_{(1)}$]	l'an 2004	
	l'année 2004	
la saison DATE[$An_{(1)}$]	la saison 2004	
l' an | année DATE[$An_{(a)}$]	l'an deux mille	
	l'année deux mille quatre	
la saison DATE[$An_{(a)}$]	la saison deux mille quatre	
l'exercice DATE[$An_{(1)}$]	l'exercice 2004	
l'exercice DATE[$An_{(a)}$]	l'exercice deux mille quatre	
(2) DATE[$G_{(dec)}$]		
les années DATE[$Décénie_{(1)}$]	les années 30	
les années DATE[$Décénie_{(a)}$]	les années trente	
(3) DATE[$G_{(cent)}$]		
les DATE[$Siècle_{(1)}$]	le 20e siècle	
les DATE[$Siècle_{(a)}$]	le vingtième siècle	
les DATE[$Siècle_{(I)}$]	le XXe siècle	
(4) DATE[$G_{(we\|week)}{}^{(0,1)}\ G_{(*<=year)}$]		
[le] DATE[[We | Semaine] $Jour_{(1)}Mois_{(1)}An_{(1)}$]	(le | la) [(semaine | week-end) du] 22/12/2009	
[le] DATE[[We | Semaine] $Jour_{(1)}Mois_{(N)}An_{(1)}$]	(le | la) [(semaine | week-end) du] 22 décembre 2009	
[le] DATE[[We | Semaine] $Jour_{(a)}Mois_{(N)}An_{(1)}$]	(le | la) [(semaine | week-end) du] vingt-deux décembre 2009	
[le] DATE[[We | Semaine] $Jour_{(a)}Mois_{(N)}An_{(a)}$]	(le | la) [(semaine | week-end) du] vingt-deux décembre deux-mille-neuf	
[le] DATE[[We | Semaine] $Jour_{(N)}Jour_{(1)}Mois_{(1)}An_{(1)}$]	(le | la) [(semaine | week-end) du] mardi 22/12/2009	
[le] DATE[[We | Semaine] $Jour_{(N)}Jour_{(1)}Mois_{(N)}An_{(1)}$]	(le | la) [(semaine | week-end) du] mardi 22 décembre 2009	
[le] DATE[[We | Semaine] $Jour_{(N)}Jour_{(a)}Mois_{(N)}An_{(1)}$]	(le | la) [(semaine | week-end) du] mardi vingt-deux décembre 2009	
[le] DATE[[We | Semaine] $Jour_{(N)}Jour_{(a)}Mois_{(N)}An_{(a)}$]	(le | la) [(semaine | week-end) du] mardi vingt-deux décembre deux-mille-neuf	
[le] DATE[[We | Semaine] $Jour_{(1)}Mois_{(US)}An_{(1)}$]	(le | la) [(semaine | week-end) du] 22 DEC 2009	
DATE[$Mois_{(N)}An_{(1)}$]	décembre 2009	
DATE[$Mois_{(N)}An_{(a)}$]	décembre deux-mille-neuf	
(5) DATE[$G_{(day)}$] HEURE[$G_{(min)}$]		
REM : L'ordre de DATE et HEURE peut être inversé		
[à | le] HEURE[$Heure_{(1)}Min_{(1)}$] DATE[$Jour_{(1)}Mois_{(1)}An_{(1)}$]	à 18h24 le 22/12/2009	

Spécification	Exemple	Commentaires
[à | le] HEURE[Heure$_{(1)}$Min$_{(1)}$] DATE[Jour$_{(1)}$Mois$_{(N)}$An$_{(1)}$]	à 18h24 le 22 décembre 2009	
[à | le] HEURE[Heure$_{(1)}$Min$_{(1)}$] DATE[Jour$_{(a)}$Mois$_{(N)}$An$_{(1)}$]	à 18h24 le vingt-deux décembre 2009	
[à | le] HEURE[Heure$_{(1)}$Min$_{(1)}$] DATE[Jour$_{(a)}$Mois$_{(N)}$An$_{(a)}$]	à 18h24 le vingt-deux décembre deux-mille-neuf	
[à | le] HEURE[Heure$_{(1)}$Min$_{(1)}$] DATE[Jour$_{(N)}$Jour$_{(1)}$Mois$_{(1)}$An$_{(1)}$]	à 18h24 le mardi 22/12/2009	
[à | le] HEURE[Heure$_{(1)}$Min$_{(1)}$] DATE[Jour$_{(N)}$Jour$_{(1)}$Mois$_{(N)}$An$_{(1)}$]	à 18h24 le mardi 22 décembre 2009	
[à | le] HEURE[Heure$_{(1)}$Min$_{(1)}$] DATE[Jour$_{(N)}$Jour$_{(a)}$Mois$_{(N)}$An$_{(1)}$]	à 18h24 le mardi vingt-deux décembre 2009	
[à | le] HEURE[Heure$_{(1)}$Min$_{(1)}$] DATE[Jour$_{(N)}$Jour$_{(a)}$Mois$_{(N)}$An$_{(a)}$]	à 18h24 le mardi vingt-deux décembre deux-mille-neuf	
(6) DATE[G$_{(day)}$] G$_{(pod_*)}$		
REM : L'ordre de DATE et G$_{(pod_*)}$ peut être inversé		
DATE[Jour$_{(1)}$Mois$_{(1)}$An$_{(1)}$] …	22/12/2009 à l'aube	G$_{(pod_da)}$
DATE[Jour$_{(1)}$Mois$_{(N)}$An$_{(1)}$] …	la journée du 22 décembre 2009 aux aurores	G$_{(pod_dt)}$
DATE[Jour$_{(a)}$Mois$_{(N)}$An$_{(1)}$] …	vingt-deux décembre 2009 matin	G$_{(pod_mo)}$
DATE[Jour$_{(a)}$Mois$_{(N)}$An$_{(a)}$] …	vingt-deux décembre deux-mille-neuf midi	G$_{(pod_mi)}$
DATE[Jour$_{(N)}$Jour$_{(1)}$Mois$_{(1)}$An$_{(1)}$] …	mardi 22/12/2009 après-midi	G$_{(pod_af)}$
DATE[Jour$_{(N)}$Jour$_{(1)}$Mois$_{(N)}$An$_{(1)}$] …	mardi 22 décembre 2009 au crépuscule	G$_{(pod_du)}$
DATE[Jour$_{(N)}$Jour$_{(a)}$Mois$_{(N)}$An$_{(1)}$] …	mardi vingt-deux décembre 2009 à la nuit tombée	G$_{(pod_ni)}$
DATE[Jour$_{(N)}$Jour$_{(a)}$Mois$_{(N)}$An$_{(a)}$] …	mardi vingt-deux décembre deux-mille-neuf au soir	G$_{(pod_ev)}$
(7) DATE[G$_{(day)}$] [Fuzzy] G$_{(pod_*)}$ HEURE[G$_{(min)}$]		
REM : L'ordre de DATE et du groupe « G$_{(pod_*)}$ HEURE » peut être inversé / pas de G$_{(pod_dt)}$ possible		
DATE[Jour$_{(1)}$Mois$_{(1)}$An$_{(1)}$] … HEURE[Heure$_{(1)}$Min$_{(1)}$]	22/12/2009, (à | vers) l'aube, à 5 heures et 24 minutes	G$_{(pod_da)}$
DATE[Jour$_{(1)}$Mois$_{(N)}$An$_{(1)}$] … HEURE[Heure$_{(1)}$Min$_{(1)}$]	22 décembre 2009, aux [alentours des] aurores, à 5h24	G$_{(pod_da)}$
DATE[Jour$_{(a)}$Mois$_{(N)}$An$_{(1)}$] … HEURE[Heure$_{(1)}$Min$_{(1)}$]	vingt-deux décembre 2009 [vers le] matin, à 9h32	G$_{(pod_mo)}$
DATE[Jour$_{(a)}$Mois$_{(N)}$An$_{(a)}$] … HEURE[Heure$_{(1)}$Min$_{(1)}$]	vingt-deux décembre deux-mille-neuf [vers] midi, à 12h15	G$_{(pod_mi)}$
DATE[Jour$_{(N)}$Jour$_{(1)}$Mois$_{(1)}$An$_{(1)}$] … HEURE[Heure$_{(1)}$Min$_{(1)}$]	mardi 22/12/2009 [dans l'] après-midi, à 15h20	G$_{(pod_af)}$
DATE[Jour$_{(N)}$Jour$_{(1)}$Mois$_{(N)}$An$_{(1)}$] … HEURE[Heure$_{(1)}$Min$_{(1)}$]	mardi 22 décembre 2009 (au | aux environs du) crépuscule, à 19h42	G$_{(pod_du)}$
DATE[Jour$_{(N)}$Jour$_{(a)}$Mois$_{(N)}$An$_{(1)}$] … HEURE[Heure$_{(1)}$Min$_{(1)}$]	mardi vingt-deux décembre 2009, (à | dans) la nuit [tombée], à 20h30	G$_{(pod_ni)}$
DATE[Jour$_{(N)}$Jour$_{(a)}$Mois$_{(N)}$An$_{(a)}$] … HEURE[Heure$_{(1)}$Min$_{(1)}$]	mardi vingt-deux décembre deux-mille-neuf (au soir | dans la soirée), à 21h10	G$_{(pod_ev)}$
(8) DATE[G$_{(*)}$] Ref$_{(now)}$ Move$_{(*)}$ Precise$^{(0,1)}$ (Nb$_{(a)}$ | Nb$_{(1)}$ | Nbf$_{(a)}$) G$_{(*)}$ [Nbf$_{(a)}$]		
REM : L'ordre de « DATE Ref » et de « Move … » est inversable.		
REM : L'interprétation est différente selon que la granularité de la date est plus petite ou égale à celle du déplacement (on se déplace pour atteindre la date), ou inversément (on se déplace à l'« intérieur » de la date).		
[le] DATE[Jour$_{(1)}$Mois$_{(1)}$An$_{(1)}$] … Move$_{(rew)}$ … Nb$_{(a)}$ G$_{(year)}$	le 22/12/2008, il y a [exactement] un an	
[le] DATE[Jour$_{(1)}$Mois$_{(N)}$An$_{(1)}$] … Move$_{(fwd)}$ … Nb$_{(1)}$ G$_{(month)}$	le 22 décembre 2008, dans [exactement] 1 mois	
[le] DATE[Jour$_{(a)}$Mois$_{(N)}$An$_{(1)}$] … Move$_{(rew)}$ … Nbf$_{(a)}$ G$_{(week)}$	le vingt-deux décembre 2008, il y a [précisément] une demi semaine	
[le] DATE[Jour$_{(a)}$Mois$_{(N)}$An$_{(a)}$] … Move$_{(fwd)}$ … Nb$_{(a)}$ G$_{(day)}$ Nbf$_{(a)}$	le vingt-deux décembre deux-mille-huit, dans [précisément] un jour et demi	
[le] DATE[Jour$_{(N)}$Jour$_{(1)}$Mois$_{(1)}$An$_{(1)}$] … Move$_{(rew)}$ … Nb$_{(1)}$ G$_{(hour)}$ Nbf$_{(a)}$	le mardi 22/12/2008, il y a [exactement] 1 heure et demi	
[le] DATE[Jour$_{(N)}$Jour$_{(1)}$Mois$_{(N)}$An$_{(1)}$] … Move$_{(fwd)}$ … Nb$_{(a)}$ G$_{(min)}$	le mardi 22 décembre 2008, dans [exactement] dix minutes	

Spécification	Exemple	Commentaires
[le] DATE[Jour$_{(N)}$Jour$_{(a)}$Mois$_{(N)}$An$_{(1)}$] … Move$_{(rew)}$ … Nb$_{(1)}$ G$_{(day)}$	le mardi vingt-deux décembre 2008, il y a [précisément] 3 jours	
[le] DATE[Jour$_{(N)}$Jour$_{(a)}$Mois$_{(N)}$An$_{(a)}$] … Move$_{(fwd)}$ … Nbf$_{(a)}$ G$_{(week)}$	le mardi vingt-deux décembre deux-mille-huit, dans [précisément] une demi semaine	
N/A : [le] DATE[Jour$_{(1)}$Mois$_{(US)}$An$_{(1)}$] …	N/A : 22 DEC 2008	
DATE[Mois$_{(N)}$An$_{(1)}$] … Move$_{(rew)}$ … Nb$_{(a)}$ G$_{(month)}$ Nbf$_{(a)}$	décembre 2008, il y a [exactement] un mois et demi	
DATE[Mois$_{(N)}$An$_{(a)}$] … Move$_{(fwd)}$ … Nb$_{(1)}$ G$_{(year)}$ Nbf$_{(a)}$	décembre deux-mille-huit, dans [exactement] 2 ans et demi	
DATE[Siècle$_{(1)}$] … Move$_{(fwd)}$ … Nb$_{(a)}$ G$_{(year)}$ Nbf$_{(a)}$	dans [précisément] deux ans, le 21e siècle	
(9) DATE[(Named | Season | Poy) G$_{(*)}$]		
(i) Named G$_{(days\|week\|day)}$		
[la] DATE[Named G$_{(week)}$ An$_{(1)}$]	la semaine pascale 2004	
[la] DATE[Named G$_{(week)}$ An$_{(a)}$]	la semaine sainte deux-mille-quatre	
[le] [jour de] DATE[Named G$_{(day)}$ An$_{(1)}$]	le jour de Noël 2004	
[le] [jour de] DATE[Named G$_{(day)}$ An$_{(a)}$]	la Toussaint deux-mille-quatre	
[le] DATE[Jour$_{(N)}$ Named G$_{(day)}$ An$_{(1)}$]	le dimanche de Pâques 2004	
[le] DATE[Jour$_{(N)}$ Named G$_{(day)}$ An$_{(a)}$]	le mardi de carnaval deux-mille-quatre	
[les] DATE[Named G$_{(days)}$ An$_{(1)}$]	les grandes vacances 2004	
[les] DATE[Named G$_{(days)}$ An$_{(a)}$]	les vacances de Pâques deux-mille-quatre	
(ii) Season G$_{(season)}$		
[le | l'] DATE[Season G$_{(season)}$ An$_{(1)}$]	le printemps 2004	
[le | l'] DATE[Season G$_{(season)}$ An$_{(a)}$]	l'été deux-mille-quatre	
(iii) Poy G$_{(qyear\|tyear\|hyear)}$		
[le] DATE[Ord$_{(a)}$ Poy G$_{(qyear)}$ An$_{(1)}$]	le premier trimestre 2004	
[le] DATE[Ord$_{(1)}$ Poy G$_{(hyear)}$ An$_{(1)}$]	le 1er semestre 2004	
[le] DATE[Ord$_{(a)}$ Poy G$_{(tyear)}$ An$_{(1)}$]	le deuxième quadrimestre 2004	
[le] DATE[Ord$_{(a)}$ Poy G$_{(qyear)}$ An$_{(a)}$]	le premier trimestre deux-mille-quatre	
[le] DATE[Ord$_{(1)}$ Poy G$_{(hyear)}$ An$_{(a)}$]	le 1er semestre deux-mille-quatre	
[le] DATE[Ord$_{(a)}$ Poy G$_{(tyear)}$ An$_{(a)}$]	le deuxième quadrimestre deux-mille-quatre	

B.4.3 PAFU : Référence Ponctuelle, Absolue, Floue et Unique

Spécification générale A :

$$DATE\ \left(G_{(day)}\right)\ G^{(0,1)}_{(pod\_*)}\ Fuzzy\ HEURE\ \left(G_{(min)}\right)$$

Spécification	Exemple	Commentaires
(1) DATE[$G_{(day)}$] Fuzzy HEURE[$G_{(min)}$]		
REM : L'ordre de DATE et du groupe « Fuzzy HEURE » peut être inversé		
DATE[$Jour_{(1)}Mois_{(1)}An_{(1)}$] … HEURE[$Heure_{(1)}Min_{(1)}$]	22/12/2009 vers 18h24	
DATE[$Jour_{(1)}Mois_{(N)}An_{(1)}$] … HEURE[$Heure_{(1)}Min_{(1)}$]	22 décembre 2009 vers 18h24	
DATE[$Jour_{(a)}Mois_{(N)}An_{(1)}$] … HEURE[$Heure_{(1)}Min_{(1)}$]	vingt-deux décembre 2009 vers 18h24	
DATE[$Jour_{(a)}Mois_{(N)}An_{(a)}$] … HEURE[$Heure_{(1)}Min_{(1)}$]	vingt-deux décembre deux-mille-neuf vers 18h24	
DATE[$Jour_{(N)}$ $Jour_{(1)}Mois_{(1)}An_{(1)}$] … HEURE[$Heure_{(1)}Min_{(1)}$]	mardi 22/12/2009 vers 18h24	
DATE[$Jour_{(N)}$ $Jour_{(1)}Mois_{(N)}An_{(1)}$] … HEURE[$Heure_{(1)}Min_{(1)}$]	mardi 22 décembre 2009 vers 18h24	
DATE[$Jour_{(N)}$ $Jour_{(a)}Mois_{(N)}An_{(1)}$] … HEURE[$Heure_{(1)}Min_{(1)}$]	mardi vingt-deux décembre 2009 vers 18h24	
DATE[$Jour_{(N)}$ $Jour_{(a)}Mois_{(N)}An_{(a)}$] … HEURE[$Heure_{(1)}Min_{(1)}$]	mardi vingt-deux décembre deux-mille-neuf vers 18h24	
DATE[$Jour_{(1)}Mois_{(US)}An_{(1)}$] … HEURE[$Heure_{(1)}Min_{(1)}$]	22 DEC 2009 vers 18 :24	
(2) DATE[$G_{(day)}$] $G_{(pod\_*)}$ Fuzzy HEURE[$G_{(min)}$]		
REM : L'ordre de DATE et du groupe « $G_{(pod\_*)}$ Fuzzy HEURE » peut être inversé / pas de $G_{(pod\_dt)}$ possible		
DATE[$Jour_{(1)}Mois_{(1)}An_{(1)}$] … HEURE[$Heure_{(1)}Min_{(1)}$]	22/12/2009 à l'aube vers 6h24	$G_{(pod\_da)}$
DATE[$Jour_{(1)}Mois_{(N)}An_{(1)}$] … HEURE[$Heure_{(1)}Min_{(1)}$]	22 décembre 2009 matin vers 10h24	$G_{(pod\_mo)}$
DATE[$Jour_{(a)}Mois_{(N)}An_{(1)}$] … HEURE[$Heure_{(1)}Min_{(1)}$]	vingt-deux décembre 2009 à midi vers 12h10	$G_{(pod\_mi)}$
DATE[$Jour_{(a)}Mois_{(N)}An_{(a)}$] … HEURE[$Heure_{(1)}Min_{(1)}$]	vingt-deux décembre deux-mille-neuf aux aurores vers 6h24	$G_{(pod\_da)}$
DATE[$Jour_{(N)}Jour_{(1)}Mois_{(1)}An_{(1)}$] … HEURE[$Heure_{(1)}Min_{(1)}$]	mardi 22/12/2009 après-midi vers 16h24	$G_{(pod\_af)}$
DATE[$Jour_{(N)}Jour_{(1)}Mois_{(N)}An_{(1)}$] … HEURE[$Heure_{(1)}Min_{(1)}$]	mardi 22 décembre 2009 au crépuscule vers 20h24	$G_{(pod\_du)}$
DATE[$Jour_{(N)}Jour_{(a)}Mois_{(N)}An_{(1)}$] … HEURE[$Heure_{(1)}Min_{(1)}$]	mardi vingt-deux décembre 2009 à la nuit tombée vers 21h24	$G_{(pod\_ni)}$
DATE[$Jour_{(N)}Jour_{(a)}Mois_{(N)}An_{(a)}$] … HEURE[$Heure_{(1)}Min_{(1)}$]	mardi vingt-deux décembre deux-mille-neuf soir vers 20h24	$G_{(pod\_ev)}$

$$DATE\ \left(G_{(day)}\right)\ (\ Fuzzy\ |\ PartOf_{(*)}\ )^{(1,N)}\ G_{(pod\_*)}\ (\ Fuzzy\ HEURE\ \left(G_{(min)}\right)\ )^{(0,1)}$$

Spécification	Exemple	Commentaires
(3) et (6) DATE[$G_{(day)}$] Fuzzy $G_{(pod\_*)}$ (Fuzzy HEURE[$G_{(min)}$])$^{(0,1)}$		
REM : L'ordre de DATE et du groupe « Fuzzy $G_{(pod\_*)}$ [Fuzzy HEURE] » peut être inversé		
DATE[$Jour_{(1)}Mois_{(1)}An_{(1)}$] … [HEURE[$Heure_{(1)}Min_{(1)}$]]	22/12/2009 peu de temps avant l'aube [, vers 6h24]	$G_{(pod\_da)}$
DATE[$Jour_{(1)}Mois_{(N)}An_{(1)}$] … [HEURE[$Heure_{(1)}Min_{(1)}$]]	22 décembre 2009 en matinée [, vers 10h24]	$G_{(pod\_mo)}$
DATE[$Jour_{(a)}Mois_{(N)}An_{(1)}$] … [HEURE[$Heure_{(1)}Min_{(1)}$]]	vingt-deux décembre 2009 vers midi [, vers 12h10]	$G_{(pod\_mi)}$
DATE[$Jour_{(a)}Mois_{(N)}An_{(a)}$] … [HEURE[$Heure_{(1)}Min_{(1)}$]]	vingt-deux décembre deux-mille-neuf en journée [, vers 14h20]	$G_{(pod\_dt)}$
DATE[$Jour_{(N)}Jour_{(1)}Mois_{(1)}An_{(1)}$] … [HEURE[$Heure_{(1)}Min_{(1)}$]]	mardi 22/12/2009 durant l'après-midi [, vers 16h24]	$G_{(pod\_af)}$
DATE[$Jour_{(N)}Jour_{(1)}Mois_{(N)}An_{(1)}$] … [HEURE[$Heure_{(1)}Min_{(1)}$]]	mardi 22 décembre 2009 aux environs du crépuscule [, vers 20h24]	$G_{(pod\_du)}$
DATE[$Jour_{(N)}Jour_{(a)}Mois_{(N)}An_{(1)}$] … [HEURE[$Heure_{(1)}Min_{(1)}$]]	mardi vingt-deux décembre 2009 peu après la nuit tombée [, vers 21h24]	$G_{(pod\_ni)}$
DATE[$Jour_{(N)}Jour_{(a)}Mois_{(N)}An_{(a)}$] … [HEURE[$Heure_{(1)}Min_{(1)}$]]	mardi vingt-deux décembre deux-mille-neuf en cours de soirée [, vers 20h24]	$G_{(pod\_ev)}$
(4) et (7) DATE[$G_{(day)}$] PartOf$_{(*)}$ $G_{(pod\_*)}$ (Fuzzy HEURE[$G_{(min)}$])$^{(0,1)}$		
REM : L'ordre de DATE et du groupe « PartOf$_{(*)}$ $G_{(pod\_*)}$ [Fuzzy HEURE] » peut être inversé		
DATE[$Jour_{(1)}Mois_{(1)}An_{(1)}$] … [HEURE[$Heure_{(1)}Min_{(1)}$]]	22/12/2009, début de matinée [, vers 8h30]	$G_{(pod\_mo)}$, PartOf$_{(beg)}$
DATE[$Jour_{(1)}Mois_{(N)}An_{(1)}$] … [HEURE[$Heure_{(1)}Min_{(1)}$]]	22 décembre 2009 à la mi-journée [, vers 13h10]	$G_{(pod\_dt)}$, PartOf$_{(mid)}$
DATE[$Jour_{(a)}Mois_{(N)}An_{(1)}$] … [HEURE[$Heure_{(1)}Min_{(1)}$]]	vingt-deux décembre 2009, fin d'après-midi [, vers 17h30]	$G_{(pod\_af)}$, PartOf$_{(end)}$
DATE[$Jour_{(a)}Mois_{(N)}An_{(a)}$] … [HEURE[$Heure_{(1)}Min_{(1)}$]]	vingt-deux décembre deux-mille-neuf, fin de nuit [,vers 4h20]	$G_{(pod\_ni)}$, PartOf$_{(end)}$
DATE[$Jour_{(N)}Jour_{(1)}Mois_{(1)}An_{(1)}$] … [HEURE[$Heure_{(1)}Min_{(1)}$]]	mardi 22/12/2009, première partie de soirée [, vers 20h30]	$G_{(pod\_ev)}$, PartOf$_{(beg)}$
DATE[$Jour_{(N)}Jour_{(1)}Mois_{(N)}An_{(1)}$] … [HEURE[$Heure_{(1)}Min_{(1)}$]]	mardi 22 décembre 2009, deuxième partie de journée [, vers 16h00]	$G_{(pod\_dt)}$, PartOf$_{(end)}$
DATE[$Jour_{(N)}Jour_{(a)}Mois_{(N)}An_{(1)}$] … [HEURE[$Heure_{(1)}Min_{(1)}$]]	mardi vingt-deux décembre 2009 fin de matinée [, vers 11h30]	$G_{(pod\_mo)}$, PartOf$_{(end)}$
DATE[$Jour_{(N)}Jour_{(a)}Mois_{(N)}An_{(a)}$] … [HEURE[$Heure_{(1)}Min_{(1)}$]]	mardi vingt-deux décembre deux-mille-neuf milieu d'après-midi [, vers 16h30]	$G_{(pod\_ev)}$, PartOf$_{(mid)}$
(5) et (8) DATE[$G_{(day)}$] Fuzzy PartOf$_{(*)}$ $G_{(pod\_*)}$ (Fuzzy HEURE[$G_{(min)}$])$^{(0,1)}$		
REM : L'ordre de DATE et du groupe « Fuzzy PartOf$_{(*)}$ $G_{(pod\_*)}$ [Fuzzy HEURE] » peut être inversé		
DATE[$Jour_{(1)}Mois_{(1)}An_{(1)}$] … [HEURE[$Heure_{(1)}Min_{(1)}$]]	22/12/2009 en début de matinée [, vers 8h30]	$G_{(pod\_mo)}$, PartOf$_{(beg)}$
DATE[$Jour_{(1)}Mois_{(N)}An_{(1)}$] … [HEURE[$Heure_{(1)}Min_{(1)}$]]	22 décembre 2009 aux environs de la mi-journée [, vers 13h10]	$G_{(pod\_dt)}$, PartOf$_{(mid)}$
DATE[$Jour_{(a)}Mois_{(N)}An_{(1)}$] … [HEURE[$Heure_{(1)}Min_{(1)}$]]	vingt-deux décembre 2009 en fin d'après-midi [, vers 17h30]	$G_{(pod\_af)}$, PartOf$_{(end)}$
DATE[$Jour_{(a)}Mois_{(N)}An_{(a)}$] … [HEURE[$Heure_{(1)}Min_{(1)}$]]	vingt-deux décembre deux-mille-neuf en fin de nuit [, vers 4h20]	$G_{(pod\_ni)}$, PartOf$_{(end)}$
DATE[$Jour_{(N)}Jour_{(1)}Mois_{(1)}An_{(1)}$] … [HEURE[$Heure_{(1)}Min_{(1)}$]]	mardi 22/12/2009 en première partie de soirée [, vers 20h30]	$G_{(pod\_ev)}$, PartOf$_{(beg)}$
DATE[$Jour_{(N)}Jour_{(1)}Mois_{(N)}An_{(1)}$] … [HEURE[$Heure_{(1)}Min_{(1)}$]]	mardi 22 décembre 2009 en deuxième partie de journée [, vers 16h00]	$G_{(pod\_dt)}$, PartOf$_{(end)}$
DATE[$Jour_{(N)}Jour_{(a)}Mois_{(N)}An_{(1)}$] … [HEURE[$Heure_{(1)}Min_{(1)}$]]	mardi vingt-deux décembre 2009 au cours de la fin de la matinée [, vers 11h30]	$G_{(pod\_mo)}$, PartOf$_{(end)}$
DATE[$Jour_{(N)}Jour_{(a)}Mois_{(N)}An_{(a)}$] … [HEURE[$Heure_{(1)}Min_{(1)}$]]	mardi vingt-deux décembre deux-mille-neuf vers le milieu de l'après-midi [, vers 16h30]	$G_{(pod\_ev)}$, PartOf$_{(mid)}$

Spécification générale B : $(\ Fuzzy \mid PartOf_{(*)} \ )^{(1,N)} \ DATE \ \left(G_{(*>=day)} \right)$

Spécification	Exemple	Commentaires
(9) Fuzzy DATE[$G_{(*)}$]		
(i) $G_{(cent)}$		
… DATE[Siècle$_{(1)}$]	aux environs du 19e siècle	
… DATE[Siècle$_{(a)}$]	au vingtième siècle	
… DATE[Siècle$_{(I)}$]	vers le XVIIe siècle	
(ii) $G_{(dec)}$		
… DATE[Décénie$_{(1)}$]	dans les années 30	
… DATE[Décénie$_{(a)}$]	vers les années trente	
(iii) $G_{(year)}$		
… l’ an | année DATE[An$_{(1)}$]	vers l’an 2000	
	aux environs de l’année 2004	
… la saison DATE[An$_{(1)}$]	durant la saison 2004	
… l’ an | année DATE[An$_{(a)}$]	vers l’an deux mille	
	au cours de l’année deux mille quatre	
… la saison DATE[An$_{(a)}$]	peu avant la saison deux mille quatre	
… l’exercice DATE[An$_{(1)}$]	peu de temps avant l’exercice 2004	
… l’exercice DATE[An$_{(a)}$]	peu de temps après l’exercice deux mille quatre	
(iv) $G_{(month)}$		
… DATE[Mois$_{(N)}$ An$_{(1)}$]	durant septembre 2004	
… DATE[Mois$_{(N)}$ An$_{(a)}$]	en décembre deux mille quatre	
(v) Named $G_{(days\|week\|day)}$		
… DATE[Named G$_{(week)}$ An$_{(1)}$]	lors de la semaine pascale 2004	
… DATE[Named G$_{(week)}$ An$_{(a)}$]	au cours de la semaine sainte deux-mille-quatre	
… [jour de] DATE[Named G$_{(day)}$ An$_{(1)}$]	aux environs du jour de Noël 2004	
… [jour de] DATE[Named G$_{(day)}$ An$_{(a)}$]	pendant la Toussaint deux-mille-quatre	
… DATE[Jour$_{(N)}$ Named G$_{(day)}$ An$_{(1)}$]	vers le dimanche de Pâques 2004	
… DATE[Jour$_{(N)}$ Named G$_{(day)}$ An$_{(a)}$]	peu avant le mardi de carnaval deux-mille-quatre	
… DATE[Named G$_{(days)}$ An$_{(1)}$]	peu de temps après les grandes vacances 2004	
… DATE[Named G$_{(days)}$ An$_{(a)}$]	durant les vacances de Pâques deux-mille-quatre	
(vi) Season $G_{(season)}$		
… DATE[Season An$_{(1)}$]	courant du printemps 2004	
… DATE[Season An$_{(a)}$]	vers l’été deux-mille-quatre	
(vii) Poy $G_{(qyear\|tyear\|hyear)}$		
… DATE[Poy G$_{(qyear)}$ An$_{(1)}$]	lors du premier trimestre 2004	
… DATE[Poy G$_{(hyear)}$ An$_{(1)}$]	pendant le 1er semestre 2004	

Spécification	Exemple	Commentaires
… DATE[Poy $G_{(tyear)}$ $An_{(1)}$]	vers le deuxième quadrimestre 2004	
… DATE[Poy $G_{(qyear)}$ $An_{(a)}$]	aux environs du premier trimestre deux-mille-quatre	
… DATE[Poy $G_{(hyear)}$ $An_{(a)}$]	dans le courant du 1er semestre deux-mille-quatre	
… DATE[Poy $G_{(tyear)}$ $An_{(a)}$]	vers le deuxième quadrimestre deux-mille-quatre	
(10) PartOf$_{(*)}$ DATE[G$_{(*)}$]		
(i) G$_{(cent)}$		
… DATE[Siècle$_{(1)}$]	au début du 19e siècle	
… DATE[Siècle$_{(a)}$]	à la fin du vingtième siècle	
… DATE[Siècle$_{(I)}$]	milieu XVIIe siècle	
(ii) G$_{(dec)}$		
… DATE[Décénie$_{(1)}$]	au début des années 30	
… DATE[Décénie$_{(a)}$]	à la fin des années trente	
(iii) G$_{(year)}$		
… l' an | année DATE[$An_{(1)}$]	début de l'an 2000	
	première partie de l'année 2004	
… la saison DATE[$An_{(1)}$]	fin de la saison 2004	
… l' an | année DATE[$An_{(a)}$]	début de l'an deux mille	
	première partie de l'année deux mille quatre	
… la saison DATE[$An_{(a)}$]	fin de la saison deux mille quatre	
… l'exercice DATE[$An_{(1)}$]	début de l'exercice 2004	
… l'exercice DATE[$An_{(a)}$]	fin de l'exercice deux mille quatre	
(iv) G$_{(month)}$		
… DATE[Mois$_{(N)}$ $An_{(1)}$]	début septembre 2004	
… DATE[Mois$_{(N)}$ $An_{(a)}$]	fin décembre deux mille quatre	
(v) Named G$_{(days\|week\|day)}$		
… DATE[Named $G_{(week)}$ $An_{(1)}$]	milieu de semaine pascale 2004	
… DATE[Named $G_{(week)}$ $An_{(a)}$]	au début de la semaine sainte deux-mille-quatre	
… [jour de] DATE[Named $G_{(day)}$ $An_{(1)}$]	tôt le jour de Noël 2004	
… [jour de] DATE[Named $G_{(day)}$ $An_{(a)}$]	tard la Toussaint deux-mille-quatre	
… DATE[Jour$_{(N)}$ Named $G_{(day)}$ $An_{(1)}$]	le début du dimanche de Pâques 2004	
… DATE[Jour$_{(N)}$ Named $G_{(day)}$ $An_{(a)}$]	la fin du mardi de carnaval deux-mille-quatre	
… DATE[Named $G_{(days)}$ $An_{(1)}$]	première partie des grandes vacances 2004	
… DATE[Named $G_{(days)}$ $An_{(a)}$]	deuxième partie des vacances de Pâques deux-mille-quatre	
(vi) Season G$_{(season)}$		
… DATE[Season $An_{(1)}$]	début du printemps 2004	
… DATE[Season $An_{(a)}$]	fin de l'été deux-mille-quatre	
(vii) Poy G$_{(qyear\|tyear\|hyear)}$		

Spécification	Exemple	Commentaires
… DATE[Poy $G_{(qyear)}$ $An_{(1)}$]	milieu du premier trimestre 2004	
… DATE[Poy $G_{(hyear)}$ $An_{(1)}$]	début du 1er semestre 2004	
… DATE[Poy $G_{(tyear)}$ $An_{(1)}$]	fin du deuxième quadrimestre 2004	
… DATE[Poy $G_{(qyear)}$ $An_{(a)}$]	milieu du premier trimestre deux-mille-quatre	
… DATE[Poy $G_{(hyear)}$ $An_{(a)}$]	début du 1er semestre deux-mille-quatre	
… DATE[Poy $G_{(tyear)}$ $An_{(a)}$]	fin du deuxième quadrimestre deux-mille-quatre	
(11) Fuzzy PartOf$_{(*)}$ DATE[G$_{(*)}$]		
(i) G$_{(cent)}$		
… DATE[Siècle$_{(1)}$]	durant le début du 19e siècle	
… DATE[Siècle$_{(a)}$]	vers la fin du vingtième siècle	
… DATE[Siècle$_{(I)}$]	aux environs du milieu du XVIIe siècle	
(ii) G$_{(dec)}$		
… DATE[Décénie$_{(1)}$]	dans le courant du début des années 30	
… DATE[Décénie$_{(a)}$]	peu de temps avant la fin des années trente	
(iii) G$_{(year)}$		
… l’ an | année DATE[$An_{(1)}$]	vers le début de l’an 2000	
	aux environs de la première partie de l’année 2004	
… la saison DATE[$An_{(1)}$]	durant la fin de la saison 2004	
… l’ an | année DATE[$An_{(a)}$]	lors du début de l’an deux mille	
	au cours de la deuxième partie de l’année deux mille quatre	
… la saison DATE[$An_{(a)}$]	peu avant la fin de la saison deux mille quatre	
… l’exercice DATE[$An_{(1)}$]	peu de temps avant le début de l’exercice 2004	
… l’exercice DATE[$An_{(a)}$]	peu de temps après la fin de l’exercice deux mille quatre	
(iv) G$_{(month)}$		
… DATE[Mois$_{(N)}$ $An_{(1)}$]	durant le début septembre 2004	
… DATE[Mois$_{(N)}$ $An_{(a)}$]	dans le courant de la mi-décembre deux mille quatre	
(v) Named G$_{(days\|week\|day)}$		
… DATE[Named $G_{(week)}$ $An_{(1)}$]	lors du milieu de la semaine pascale 2004	
… DATE[Named $G_{(week)}$ $An_{(a)}$]	au cours du début de la semaine sainte deux-mille-quatre	
… [jour de] DATE[Named $G_{(day)}$ $An_{(1)}$]	vers le début du jour de Noël 2004	
… [jour de] DATE[Named $G_{(day)}$ $An_{(a)}$]	pendant la fin de la Toussaint deux-mille-quatre	
… DATE[Jour$_{(N)}$ Named $G_{(day)}$ $An_{(1)}$]	vers la fin du dimanche de Pâques 2004	
… DATE[Jour$_{(N)}$ Named $G_{(day)}$ $An_{(a)}$]	peu avant le début du mardi de carnaval deux-mille-quatre	
… DATE[Named $G_{(days)}$ $An_{(1)}$]	peu de temps après la première partie des grandes vacances 2004	
… DATE[Named $G_{(days)}$ $An_{(a)}$]	durant la dernière partie des vacances de Pâques deux-mille-quatre	
(vi) Season G$_{(season)}$		
… DATE[Season $An_{(1)}$]	courant du début du printemps 2004	

Spécification	Exemple	Commentaires
... DATE[Season An$_{(a)}$]	vers la fin de l'été deux-mille-quatre	
(vii) Poy G$_{(qyear\|tyear\|hyear)}$		
... DATE[Poy G$_{(qyear)}$ An$_{(1)}$]	lors du milieu du premier trimestre 2004	
... DATE[Poy G$_{(hyear)}$ An$_{(1)}$]	pendant le début du 1er semestre 2004	
... DATE[Poy G$_{(tyear)}$ An$_{(1)}$]	vers la fin du deuxième quadrimestre 2004	
... DATE[Poy G$_{(qyear)}$ An$_{(a)}$]	aux environs du milieu du premier trimestre deux-mille-quatre	
... DATE[Poy G$_{(hyear)}$ An$_{(a)}$]	dans le courant du début du 1er semestre deux-mille-quatre	
... DATE[Poy G$_{(tyear)}$ An$_{(a)}$]	vers la fin du deuxième quadrimestre deux-mille-quatre	

B.4.4 PRPU : Référence Ponctuelle, Relative, Précise et Unique

Spécification générale A : date sous spécifiée.

$$TARGET^{(1,2)}_{(part)}\ \left(G_{(*)}\right)\ TEMPSHIFT\ \left(Ref_{(*)}\ (\ Move_{(*)}\ Amp_{(*)}\ )^{(0,1)}\right)$$

Spécification	Exemple	Commentaires
(1) TARGET$_{(part)}$[(Fuzzy$^{(0,1)}$ G$_{(pod_*)}$)$^{(0,1)}$ HEURE[G$_{(min)}$]$^{(0,1)}$ G$_{(we|week)}$$^{(0,1)}$ DATE[G$_{(*)}$]] TEMPSHIFT[Ref$_{(*)}$ (Move$_{(*)}$ Amp$_{(*)}$)$^{(0,1)}$]		
REM : G$_{(pod_*)}$ et HEURE[G$_{(min)}$] ne sont compatibles qu'avec une date de granularité « day »		
REM : Fuzzy ne peut apparaître que devant G$_{(pod_*)}$ et seulement si l'élément HEURE[G$_{(min)}$] est présent.		
(i) TARGET$_{(part)}$[(Fuzzy$^{(0,1)}$ G$_{(pod_*)}$)$^{(0,1)}$ HEURE[G$_{(min)}$]$^{(0,1)}$ DATE[G$_{(*)}$]] TEMPSHIFT[Ref$_{(now)}$]		
REM : L'ordre de DATE, G$_{(pod_*)}$ et HEURE peut varier		
… Mois$_{(N)}$ …	décembre	
G$_{(pod_*)}$$^{(0,1)}$ … HEURE[G$_{(min)}$]$^{(0,1)}$ … Jour$_{(N)}$ …	[à l'aube,] [à 5 heures,] [le] jeudi	
G$_{(pod_*)}$$^{(0,1)}$ … HEURE[G$_{(min)}$]$^{(0,1)}$ … Jour$_{(N)}$Jour$_{(1)}$ …	[le matin,] [à 9h30,] [le] jeudi 14	
G$_{(pod_*)}$$^{(0,1)}$ … HEURE[G$_{(min)}$]$^{(0,1)}$ … Jour$_{(N)}$Jour$_{(a)}$ …	[le soir,] [à 18h,] [le] jeudi quatorze	
G$_{(pod_*)}$$^{(0,1)}$ … HEURE[G$_{(min)}$]$^{(0,1)}$ … Jour$_{(N)}$Jour$_{(1)}$Mois$_{(1)}$ …	[la nuit,] [à 22h,] [le] jeudi 14 / 12	
G$_{(pod_*)}$$^{(0,1)}$ … HEURE[G$_{(min)}$]$^{(0,1)}$ … Jour$_{(N)}$Jour$_{(1)}$Mois$_{(N)}$ …	[à la nuit tombée,] [à 20h15,] [le] jeudi 14 décembre	
N/A : TARGET$_{(part)}$[DATE[Jour$_{(N)}$Jour$_{(a)}$Mois$_{(1)}$]] …	N/A : jeudi quatorze / 12	
G$_{(pod_*)}$$^{(0,1)}$ … HEURE[G$_{(min)}$]$^{(0,1)}$ … Jour$_{(N)}$Jour$_{(a)}$Mois$_{(N)}$ …	[à l'aube,] [à 5 heures,] [le] jeudi quatorze décembre	
G$_{(pod_*)}$$^{(0,1)}$ … HEURE[G$_{(min)}$]$^{(0,1)}$ … Jour$_{(1)}$Mois$_{(1)}$ …	[le matin,] [à 9h30,] [le] 14/12	
G$_{(pod_*)}$$^{(0,1)}$ … HEURE[G$_{(min)}$]$^{(0,1)}$ … Jour$_{(1)}$Mois$_{(N)}$ …	[le soir,] [à 18h,] [le] 14 décembre	
N/A : TARGET$_{(part)}$[DATE[Jour$_{(a)}$Mois$_{(1)}$]] …	N/A : quatorze / 12	
G$_{(pod_*)}$$^{(0,1)}$ … HEURE[G$_{(min)}$]$^{(0,1)}$ … Jour$_{(a)}$Mois$_{(N)}$ …	[la nuit,] [à 22h,] [le] quatorze décembre	
G$_{(pod_*)}$$^{(0,1)}$ … HEURE[G$_{(min)}$]$^{(0,1)}$ … Named G$_{(day)}$ …	[à la nuit tombée,] [à 20h15,] le jour de la Toussaint	
… Named G$_{(week)}$ …	la semaine Pascale	
… Named G$_{(days)}$ …	les vacances de Pâques	
(ii) TARGET$_{(part)}$[G$_{(we\|week)}$ DATE[G$_{(day)}$]] TEMPSHIFT[Ref$_{(now)}$]		
N/A : TARGET$_{(part)}$[G$_{(we\|week)}$,DATE[Jour$_{(N)}$]] …	N/A : la semaine du jeudi	
… Jour$_{(N)}$Jour$_{(1)}$ …	la semaine du jeudi 14	
… Jour$_{(N)}$Jour$_{(a)}$ …	le week-end du samedi seize	
… Jour$_{(N)}$Jour$_{(1)}$Mois$_{(1)}$ …	la semaine du jeudi 14/12	
… Jour$_{(N)}$Jour$_{(1)}$Mois$_{(N)}$ …	la semaine du jeudi 14 décembre	
N/A : TARGET$_{(part)}$[G$_{(we\|week)}$,DATE[Jour$_{(N)}$Jour$_{(a)}$Mois$_{(1)}$]] …	N/A :la semaine du jeudi quatorze / 12	
… Jour$_{(N)}$Jour$_{(a)}$Mois$_{(N)}$ …	le week-end du samedi seize décembre	
… Jour$_{(1)}$Mois$_{(1)}$ …	la semaine du 14/12	
… Jour$_{(1)}$Mois$_{(N)}$ …	le week-end du 16 décembre	

Spécification	Exemple	Commentaires
N/A : TARGET$_{(part)}$[G$_{(we\|week)}$,DATE[Jour$_{(a)}$Mois$_{(1)}$]] …	N/A : la semaine du quatorze /12	
… Jour$_{(a)}$Mois$_{(N)}$ …	la semaine du quatorze décembre	
… Jour$_{(1)}$ …	le week-end du 16	
… Jour$_{(a)}$ …	la semaine du quatorze	
… Named G$_{(day)}$ …	la semaine de Pâques	
(iii) TARGET$_{(part)}$[(Fuzzy$^{(0,1)}$ G$_{(pod_*)}$)$^{(0,1)}$ HEURE[G$_{(min)}$]$^{(0,1)}$ DATE[G$_{(*)}$]] TEMPSHIFT[Ref$_{(now)}$ Move$_{(*)}$ Amp$_{(*)}$]		
REM : A part DATE[Jour$_{(N)}$] ou DATE[Named G$_{(day)}$], il n'y a pas d'autre spécification du jour possible ici.		
REM : L'ordre de DATE, G$_{(pod_*)}$ et HEURE peut varier ; l'ordre de DATE et de « Ref Move NB » peut être inversé		
… Mois$_{(N)}$ … Move$_{(rew)}$ Amp$_{(1)}$	décembre dernier	pas de G$_{(pod_*)}$ / HEURE[G$_{(min)}$] possible
… (Fuzzy$^{(0,1)}$ G$_{(pod_*)}$)$^{(0,1)}$ HEURE[G$_{(min)}$]$^{(0,1)}$ … Jour$_{(N)}$ … Move$_{(fwd)}$ Amp$_{(2)}$	[à l'aube,] [à 5 heures,] jeudi en huit	
… (Fuzzy$^{(0,1)}$ G$_{(pod_*)}$)$^{(0,1)}$ HEURE[G$_{(min)}$]$^{(0,1)}$ … Named G$_{(day)}$ … Move$_{(rew)}$ Amp$_{(1)}$	[à la nuit tombée,] [à 20h15,] [au] dernier Noël	
… Named G$_{(week)}$ … Move$_{(fwd)}$ Amp$_{(1)}$	la semaine Pascale prochaine	pas de G$_{(pod_*)}$ / HEURE[G$_{(min)}$] possible
… Named G$_{(days)}$ … Move$_{(fwd)}$ Amp$_{(1)}$	les prochaines vacances de Pâques	pas de G$_{(pod_*)}$ / HEURE[G$_{(min)}$] possible
(iv) TARGET$_{(part)}$[(Fuzzy$^{(0,1)}$ G$_{(pod_*)}$)$^{(0,1)}$ HEURE[G$_{(min)}$]$^{(0,1)}$ DATE[G$_{(*)}$]] TEMPSHIFT[Ref$_{(focus)}$ Move$_{(*)}$ Amp$_{(*)}$]		
REM : A part DATE[Jour$_{(N)}$] ou DATE[Named G$_{(day)}$, il n'y a pas d'autre spécification du jour possible ici.		
REM : L'ordre de DATE, G$_{(pod_*)}$ et HEURE peut varier ; l'ordre de DATE et de « Ref Move NB » peut être inversé		
… Mois$_{(N)}$ … Move$_{(rew)}$ Amp$_{(1)}$	le mois de décembre précédent	pas de G$_{(pod_*)}$ / HEURE[G$_{(min)}$] possible
… (Fuzzy$^{(0,1)}$ G$_{(pod_*)}$)$^{(0,1)}$ HEURE[G$_{(min)}$]$^{(0,1)}$ … Jour$_{(N)}$ … Move$_{(rew)}$ Amp$_{(1)}$	[le soir,] [à 18h,] [le] jeudi d'avant	
… (Fuzzy$^{(0,1)}$ G$_{(pod_*)}$)$^{(0,1)}$ HEURE[G$_{(min)}$]$^{(0,1)}$ … Named G$_{(day)}$ … Move$_{(rew)}$ Amp$_{(1)}$	[la nuit,] [à 22h,] [à] la Toussaint précédente	
… Named G$_{(week)}$ … Move$_{(fwd)}$ Amp$_{(1)}$	la semaine Pascale d'après	pas de G$_{(pod_*)}$ / HEURE[G$_{(min)}$] possible
… Named G$_{(days)}$ … Move$_{(rew)}$ Amp$_{(1)}$	les vacances de Pâques d'avant	pas de G$_{(pod_*)}$ / HEURE[G$_{(min)}$] possible
(2) TARGET$_{(part)}$[(G$_{(pod_*)}$$^{(0,1)}$ HEURE[G$_{(min)}$]$^{(0,1)}$)$^{(1,2)}$] TEMPSHIFT[Ref$_{(focus)}$]		
… HEURE[Heure$_{(1)}$] …	à 13 heures	
… G$_{(pod_*)}$ …	le soir	
… G$_{(pod_*)}$ HEURE[Heure$_{(1)}$Min$_{(1)}$ …	à l'aube, à 5h45	

Spécification générale B : un déplacement temporel explicite

$$TARGET_{(part)}^{(0,1)}\ \left(G_{(*)}\right)\ TEMPSHIFT\ \left(Ref_{(*)}\ Move_{(*)}^{(0,1)}\ Precise^{(0,1)}\ NbTempUnit_{(*,*)}\right)\ FOCTEMP_{(*)}^{(0,1)}\ \left(G_{(*)}\right)$$

Spécification	Exemple	Commentaires
(a) TEMPSHIFT[Ref$_{(now)}$]		
(3) TARGET$_{(part)}^{(0,1)}$ TEMPSHIFT[Ref$_{(now)}$ Move$_{(*)}$ Precise$^{(0,1)}$ NbTempUnit$_{(*,*)}$] FOCTEMP$_{(*)}^{(0,1)}$		
REM : TARGET peut apparaitre en début ou fin d'expression ; un seul TARGET par expression.		
REM : Precise est facultatif lorsque TARGET est présent (car contient un PRPU) ou dans le cas d'un G$_{(*<=day)}$, sinon il est obligatoire.		
REM : TARGET ne peut pas contenir d'expression de type PAPU, l'expression complète serait alors de ce type.		
REM : Certaines combinaisons entre granularités de TARGET et de G$_{(*)}$ n'apparaîtront jamais (ex. : lundi, il y a deux minutes).		
REM : Lorsque TARGET est présent, Fuzzy peut éventuellement apparaître avant NbTmpUnit$_{(*,*)}$, mais ne sera pas annoté.		
… Move$_{(rew)}$ … Nb$_{(a)}$ G$_{(*)}$ …	[lundi,] il y a exactement un an [, lundi]	G$_{(year)}$, Precise obligatoire, idem pour G$_{(*>year)}$ mais peu courant
… Move$_{(rew)}$ … Nb$_{(a)}$ G$_{(*)}$ …	[lundi,] il y a exactement un trimestre [, lundi]	G$_{(qyear)}$, Precise obligatoire, idem pour G$_{(tyear)}$ et G$_{(hyear)}$
… Move$_{(fwd)}$ … Nb$_{(a)}$ G$_{(*)}$ …	[à pâques,] dans précisément deux mois [, à pâques]	G$_{(month)}$, Precise obligatoire
… Move$_{(rew)}$ … Nb$_{(1)}$ G$_{(*)}$ …	[le mercredi 14,] voici exactement 3 semaines [, le mercredi 14]	G$_{(week)}$, Precise obligatoire
… Move$_{(fwd)}$ … Nb$_{(1)}$ G$_{(*)}$ …	[le 14 décembre,] dans [précisément] 4 jours [, le 14 décembre]	G$_{(day)}$
… Move$_{(rew)}$ … Nbf$_{(a)}$ G$_{(*)}$ …	[à 13h,] il y a [exactement] une demi heure [, à 13h]	G$_{(hour)}$
… Move$_{(fwd)}$ … Nbf$_{(a)}$ G$_{(*)}$ …	[à 22h,] dans [précisément] trois-quarts de seconde [, à 22h]	G$_{(dec)}$
… Move$_{(rew)}$ … Nb$_{(a)}$ G$_{(*)}$ Nbf$_{(a)}$ …	[à 10h15] voici [exactement] une heure et quart [, à 10h15]	G$_{(hour)}$
… Move$_{(fwd)}$ … Nb$_{(a)}$ G$_{(*)}$ Nbf$_{(a)}$ …	[à 14h17] dans [précisément] une minute et demi [,à 14h17]	G$_{(min)}$
(b) TEMPSHIFT[Ref$_{(focus)}$]		
(4) TARGET$_{(part)}^{(0,1)}$ TEMPSHIFT[Precise$^{(0,1)}$ NbTempUnit$_{(*,*)}$ Ref$_{(focus)}$ Move$_{(*)}$] FOCTEMP$_{(*)}^{(0,1)}$		
REM : TARGET peut apparaitre en début ou fin d'expression ; un seul TARGET par expression.		
REM : Pour que TARGET apparaisse en fin d'expression, il doit être précédé de FOCTEMP.		
REM : Precise est facultatif lorsque TARGET est présent (car contient un PRPU) ou dans le cas d'un G$_{(*<=day)}$, sinon il est obligatoire.		
REM : TARGET ne peut pas contenir d'expression de type PAPU, l'expression complète serait alors de ce type. Pas de contrainte de ce type sur FOCTEMP.		
REM : Lorsque TARGET est présent, Fuzzy peut éventuellement apparaître avant NbTmpUnit$_{(*,*)}$, mais ne sera pas annoté.		
… Nb$_{(a)}$ G$_{(*)}$ … Move$_{(rew)}$ …	[lundi,] exactement un an avant [le 11 septembre 2001] [, lundi]	G$_{(year)}$, Precise obligatoire, idem pour G$_{(*>year)}$ mais peu courant
… Nb$_{(a)}$ G$_{(*)}$ … Move$_{(rew)}$ …	[lundi,] exactement un semestre avant [le 11 septembre 2001] [, lundi]	G$_{(hyear)}$, Precise obligatoire, idem pour G$_{(tyear)}$ ou G$_{(qyear)}$
… Nb$_{(a)}$ G$_{(*)}$ … Move$_{(fwd)}$ …	[à pâques,] [précisément] deux mois après [le 12 juin] [, à pâques]	G$_{(month)}$, Precise obligatoire
… Nb$_{(1)}$ G$_{(*)}$ … Move$_{(rew)}$ …	[le mercredi 14,] [exactement] 3 semaines plus tôt [-] [, le mercredi 14]	G$_{(week)}$, Precise obligatoire
… Nb$_{(1)}$ G$_{(*)}$ … Move$_{(fwd)}$ …	[le 14 décembre,] [précisément] 4 jours plus tard [-] [, le 14 décembre]	G$_{(day)}$
… Nbf$_{(a)}$ G$_{(*)}$ … Move$_{(rew)}$ …	[à 12h,] [exactement] une demi heure avant [12h30] [, à 12h]	G$_{(hour)}$
… Nbf$_{(a)}$ G$_{(*)}$ … Move$_{(fwd)}$ …	[-] [précisément] trois-quarts de seconde après [22h] [-]	G$_{(sec)}$
… Nb$_{(a)}$ G$_{(*)}$ Nbf$_{(a)}$ … Move$_{(rew)}$ …	[à 10h15] [exactement] une heure et quart plus tôt [-] [, à 10h15]	G$_{(hour)}$
… Nb$_{(a)}$ G$_{(*)}$ Nbf$_{(a)}$ … Move$_{(fwd)}$ …	[à 14h17] [précisément] une minute et demi plus tard [-] [,à 14h17]	G$_{(min)}$

Spécification générale C : déplacement temporel implicite (adverbe)

$$TARGET^{(0,1)}_{(part)}\ \left(G_{(*)}\right)\ TEMPSHIFT\ \left(Ref_{(*)}\ G_{(day\ |\ undef)}\ Move_{(*)}\ Amp_{(*)}\right)\ FOCTEMP^{(0,1)}_{(*)}\ \left(G_{(*)}\right)$$

Spécification	Exemple	Commentaires
(a) TEMPSHIFT[Ref$_{(now)}$]		
(5) TARGET$^{(0,1)}_{(part)}$ TEMPSHIFT[Ref$_{(now)}$ G$_{(day\|undef)}$ Move$_{(*)}$ Amp$_{(*)}$]		
REM : TARGET peut apparaitre en début ou fin d'expression ; un seul TARGET par expression.		
REM : TARGET ne peut pas contenir d'expression de type PAPU, l'expression complète serait alors de ce type.		
… G$_{(day)}$ Move$_{(no)}$ Amp$_{(0)}$	aujourd'hui [vendredi]	
… G$_{(day)}$ Move$_{(rew)}$ Amp$_{(2)}$	avant-hier [, le samedi 14]	
… G$_{(day)}$ Move$_{(rew)}$ Amp$_{(1)}$	hier [, à Pâques]	
… G$_{(day)}$ Move$_{(fwd)}$ Amp$_{(1)}$	demain [, le 14 décembre]	
… G$_{(day)}$ Move$_{(fwd)}$ Amp$_{(2)}$	après-demain [à 17h30]	
… G$_{(undef)}$ Move$_{(no)}$ Amp$_{(0)}$	actuellement [, mercredi [soir]]	
(b) TEMPSHIFT[Ref$_{(focus)}$]		
(6) TARGET$^{(0,1)}_{(part)}$ TEMPSHIFT[Ref$_{(focus)}$ G$_{(day)}$ Move$_{(*)}$ Amp$_{(*)}$] FOCTEMP$_{(*)}$		
REM : TARGET peut apparaitre en début ou fin d'expression ; un seul TARGET par expression.		
REM : TARGET ne peut pas contenir d'expression de type PAPU, l'expression complète serait alors de ce type. Pas de contrainte de ce type sur FOCTEMP.		
… Move$_{(rew)}$ Amp$_{(1)}$ …	[le jeudi 14,] la veille [du 15/03/2010]	
… Move$_{(rew)}$ Amp$_{(2)}$ …	[vendredi [midi],] l'avant-veille [du dimanche 17 juin]	
… Move$_{(fwd)}$ Amp$_{(1)}$ …	[le 26 décembre,] le lendemain [du jour de Noël]	
… Move$_{(fwd)}$ Amp$_{(2)}$ …	[à Pâques,] le surlendemain [du vendredi]	

Spécification générale D : déplacement temporel implicite (grain et localisateur)

$$TEMPSHIFT\ \left(Ref_{(*)}\ G_{(*)}\ Move_{(*)}\ Amp_{(*)}\right)$$

Spécification	Exemple	Commentaires
(7)		
Ref$_{(now)}$ G$_{(week)}$ Move$_{(rew)}$ Amp$_{(1)}$	la semaine dernière	
Ref$_{(now)}$ G$_{(year)}$ Move$_{(fwd)}$ Amp$_{(1)}$	l'année prochaine	
Ref$_{(focus)}$ G$_{(hour)}$ Move$_{(fwd)}$ Amp$_{(1)}$	l'heure d'après	

$$TARGET_{(part)}\ \left(G_{(*)}\right)\ TEMPSHIFT\ \left(Ref_{(*)}\right)$$

Spécification	Exemple	Commentaires
(8)		
TARGET[G$_{(podnight)}$] TEMPSHIFT[Ref$_{(now)}$]	ce soir	
TARGET[G$_{(year)}$] TEMPSHIFT[Ref$_{(focus)}$]	cette année là	

B.4.5 PRFU : Référence Ponctuelle, Relative, Floue et Unique

Spécification générale A :

$$TARGET^{(0,1)}_{(part)}\ \left(G_{(*)}\right)\ TEMPSHIFT\ \left(Ref_{(*)}\ Move_{(*)}\ Fuzzy^{(0,1)}\ (\ G_{(*)}\ |\ NbTempUnit_{(*,*)}\ )\right)\ FOCTEMP^{(0,1)}_{(*)}\ \left(G_{(*)}\right)$$

Spécification	Exemple	Commentaires
(1) TARGET$^{(0,1)}_{(part)}$ TEMPSHIFT[Ref$_{(now)}$ Move$_{(*)}$ Fuzzy G$_{(*)}$]		
REM : Poy $G_{(hyear\|tyear\|qyear)}$ peut aussi être utilisé comme unité pour le déplacement temporel.		
… Move$_{(rew)}$ … G$_{(*)}$	il y a quelques jours	
… Move$_{(fwd)}$ … G$_{(*)}$	dans quelques semaines	
(2) TARGET$^{(0,1)}_{(part)}$ TEMPSHIFT[Ref$_{(now)}$ Move$_{(*)}$ Fuzzy$^{(0,1)}$ NbTempUnit$_{(*,*)}$]		
REM : Fuzzy est obligatoire lorsque $G_{(*)}$ est plus petit ou égal à « day ».		
REM : Poy $G_{(hyear\|tyear\|qyear)}$ peut aussi être utilisé comme unité pour le déplacement temporel.		
… Move$_{(rew)}$ … Nb$_{(a)}$ G$_{(*)}$	il y a [environ] un an	
… Move$_{(fwd)}$ … Nb$_{(a)}$ G$_{(*)}$	dans [à peu près] deux mois	
… Move$_{(rew)}$ … Nb$_{(1)}$ G$_{(*)}$	voici [un peu plus de] 3 semaines	
… Move$_{(fwd)}$ … Nb$_{(1)}$ G$_{(*)}$	dans pas moins de 4 jours	
… Move$_{(rew)}$ … Nbf$_{(a)}$ G$_{(*)}$	il y a au plus une demi heure	
… Move$_{(fwd)}$ … Nbf$_{(a)}$ G$_{(*)}$	dans moins de trois-quarts de seconde	
… Move$_{(rew)}$ … Nb$_{(a)}$ G$_{(*)}$ Nbf$_{(a)}$	voici près d'une heure et quart	
… Move$_{(fwd)}$ … Nb$_{(a)}$ G$_{(*)}$ Nbf$_{(a)}$	dans plus ou moins une minute et demi	
(3) TEMPSHIFT[Fuzzy G$_{(*)}$ Move$_{(*)}$ Ref$_{(focus)}$] FOCTEMP$^{(0,1)}_{(*)}$		
REM : $G_{(*)}$ inférieur ou égale à « day » est autorisé ici car il est toujours précédé d'une marque d'imprécision		
REM : Lorsque $FOCTEMP_{(*)}$ est donné, $G_{(*)}$ ne peut être inférieur à la granularité de celui-ci		
REM : Poy $G_{(hyear\|tyear\|qyear)}$ peut aussi être utilisé comme unité pour le déplacement temporel.		
(i) pas de FOCTEMP$_{(*)}$		
… G$_{(*)}$ Move$_{(rew)}$ …	plusieurs semaines après	
… G$_{(*)}$ Move$_{(fwd)}$ …	quelques jours plus tard	
(ii) FOCTEMP$_{(full)}$		
… G$_{(*)}$ Move$_{(rew)}$ … FOCTEMP$_{(full)}$[DATE[G$_{(*)}$]]	plusieurs semaines avant le mardi 22/12/2009	
… G$_{(*)}$ Move$_{(fwd)}$ … FOCTEMP$_{(full)}$[DATE[G$_{(*)}$]]	quelques jours après le 22 décembre 2009	
(iii) FOCTEMP$_{(part)}$		
… G$_{(*)}$ Move$_{(rew)}$ … FOCTEMP$_{(part)}$[DATE[G$_{(*)}$]]	plusieurs semaines avant jeudi	
… G$_{(*)}$ Move$_{(fwd)}$ … FOCTEMP$_{(part)}$[DATE[G$_{(*)}$]]	quelques jours après jeudi	
… G$_{(*)}$ Move$_{(rew)}$ … FOCTEMP$_{(part)}$[HEURE[G$_{(min)}$]]	plusieurs heures avant 16 heures	

Spécification	Exemple	Commentaires
… $G_{(*)}$ Move$_{(fwd)}$ … FOCTEMP$_{(part)}$[HEURE[$G_{(min)}$]]	quelques secondes après 16h43	
(4) TEMPSHIFT[Fuzzy$^{(0,1)}$ NbTempUnit$_{(*,*)}$ Move$_{(*)}$ Ref$_{(focus)}$] FOCTEMP$^{(0,1)}_{(*)}$		
REM : Fuzzy est obligatoire pour tout $G_{(*)}$ inférieur ou égal à « day » (sinon voir PRPU (4))		
REM : Lorsque FOCTEMP$_{(*)}$ est donné, $G_{(*)}$ ne peut être inférieur à la granularité de celui-ci		
REM : Poy $G_{(hyear\|tyear\|qyear)}$ peut aussi être utilisé comme unité pour le déplacement temporel.		
(i) pas de FOCTEMP$_{(*)}$		
… Nb$_{(a)}$ G$_{(*)}$ Move$_{(rew)}$ …	[environ] un an avant	
… Nb$_{(a)}$ G$_{(*)}$ Move$_{(fwd)}$ …	[à peu près] deux mois plus tôt	
… Nb$_{(1)}$ G$_{(*)}$ Move$_{(rew)}$ …	[un peu plus de] 3 semaines après	
… Nb$_{(1)}$ G$_{(*)}$ Move$_{(fwd)}$ …	quasiment 4 jours plus tard	Fuzzy obligatoire
… Nbf$_{(a)}$ G$_{(*)}$ Move$_{(rew)}$ …	au plus une demi heure plus tôt	Fuzzy obligatoire
… Nbf$_{(a)}$ G$_{(*)}$ Move$_{(fwd)}$ …	moins de trois-quarts de seconde après	Fuzzy obligatoire
… Nb$_{(a)}$ G$_{(*)}$ Nbf$_{(a)}$ Move$_{(rew)}$ …	près d'une heure et quart avant	Fuzzy obligatoire
… Nb$_{(a)}$ G$_{(*)}$ Nbf$_{(a)}$ Move$_{(fwd)}$ …	plus ou moins une minute et demi plus tard	Fuzzy obligatoire
(ii) FOCTEMP$_{(full)}$		
REM : FOCTEMP$_{(full)}$ n'est pas détaillé de façon exhaustive, se référer à la spécification de PAPU		
… Nb$_{(a)}$ G$_{(*)}$ Move$_{(rew)}$ … FOCTEMP$_{(full)}$[DATE[G(*)]]	[environ] un an avant le 22 décembre 2009	
… Nb$_{(a)}$ G$_{(*)}$ Move$_{(fwd)}$ … FOCTEMP$_{(full)}$[DATE[G(*)]]	[à peu près] deux mois après décembre 2009	
… Nb$_{(1)}$ G$_{(*)}$ Move$_{(rew)}$ … FOCTEMP$_{(full)}$[DATE[G(*)]]	[un peu plus de] 3 semaines avant le mardi 22/12/2009	
… Nb$_{(1)}$ G$_{(*)}$ Move$_{(fwd)}$ … FOCTEMP$_{(full)}$[DATE[G(*)]]	quasiment 4 jours après le 22 décembre 2009	Fuzzy obligatoire
… Nbf$_{(a)}$ G$_{(*)}$ Move$_{(rew)}$ … FOCTEMP$_{(full)}$[DATE[G(*)]]	[au plus] un demi mois avant le mardi 22 décembre 2009	
… Nbf$_{(a)}$ G$_{(*)}$ Move$_{(fwd)}$ … FOCTEMP$_{(full)}$[DATE[G(*)]]	[moins d'une] demi semaine après le 22 décembre 2009	
… Nb$_{(a)}$ G$_{(*)}$ Nbf$_{(a)}$ Move$_{(rew)}$ … FOCTEMP$_{(full)}$[DATE[G(*)]]	[près d'] un mois et demi avant le mardi 22 décembre 2009	
… Nb$_{(a)}$ G$_{(*)}$ Nbf$_{(a)}$ Move$_{(fwd)}$ … FOCTEMP$_{(full)}$[DATE[G(*)]]	[plus ou moins] un an et demi après le 22 décembre 2009	
(iii) FOCTEMP$_{(part)}$		
REM : FOCTEMP$_{(part)}$ n'est pas détaillé de façon exhaustive, se référer à la spécification de PRPU		
… Nb$_{(a)}$ G$_{(*)}$ Move$_{(rew)}$ … FOCTEMP$_{(part)}$[DATE[G$_{(*)}$]]	[environ] une semaine avant décembre	
… Nb$_{(a)}$ G$_{(*)}$ Move$_{(fwd)}$ … FOCTEMP$_{(part)}$[DATE[G$_{(*)}$]]	[à peu près] deux mois après décembre	
… Nb$_{(1)}$ G$_{(*)}$ Move$_{(rew)}$ … FOCTEMP$_{(part)}$[DATE[G$_{(*)}$]]	[un peu plus de] 3 semaines avant jeudi	
… Nb$_{(1)}$ G$_{(*)}$ Move$_{(fwd)}$ … FOCTEMP$_{(part)}$[DATE[G$_{(*)}$]]	quasiment 4 jours après jeudi	Fuzzy obligatoire
… Nbf$_{(a)}$ G$_{(*)}$ Move$_{(rew)}$ … FOCTEMP$_{(part)}$[HEURE[$G_{(min)}$]]	au plus une demi heure avant 16 heures	Fuzzy obligatoire
… Nbf$_{(a)}$ G$_{(*)}$ Move$_{(fwd)}$ … FOCTEMP$_{(part)}$[HEURE[$G_{(min)}$]]	moins de trois-quarts de seconde après 16h43	Fuzzy obligatoire
… Nb$_{(a)}$ G$_{(*)}$ Nbf$_{(a)}$ Move$_{(rew)}$ … FOCTEMP$_{(part)}$[DATE[G$_{(*)}$]]	[près d'] un mois et demi avant jeudi	
… Nb$_{(a)}$ G$_{(*)}$ Nbf$_{(a)}$ Move$_{(fwd)}$ … FOCTEMP$_{(part)}$[DATE[G$_{(*)}$]]	[plus ou moins] une semaine et demi après jeudi	
(iv) FOCTEMP$_{(part)}$ avec DATE de type Named		
… Nb$_{(a)}$ G$_{(*)}$ Move$_{(rew)}$ … FOCTEMP$_{(part)}$[DATE[G$_{(*)}$ Named]]	[environ] un an avant Noël	
… Nb$_{(a)}$ G$_{(*)}$ Move$_{(fwd)}$ … FOCTEMP$_{(part)}$[DATE[G$_{(*)}$ Named]]	[à peu près] deux mois après le jour de Pâques	

Spécification	Exemple	Commentaires
… Nb$_{(1)}$ G$_{(*)}$ Move$_{(rew)}$ … FOCTEMP$_{(part)}$[DATE[G$_{(*)}$ Named]]	[un peu plus de] 3 semaines avant les vacances de Pâques	
… Nb$_{(1)}$ G$_{(*)}$ Move$_{(fwd)}$ … FOCTEMP$_{(part)}$[DATE[G$_{(*)}$ Named]]	quasiment 4 jours après la Toussaint	Fuzzy obligatoire
… Nbf$_{(a)}$ G$_{(*)}$ Move$_{(rew)}$ … FOCTEMP$_{(part)}$[DATE[G$_{(*)}$ Named]]	au plus une demi heure avant Noël	Fuzzy obligatoire
… Nbf$_{(a)}$ G$_{(*)}$ Move$_{(fwd)}$ … FOCTEMP$_{(part)}$[DATE[G$_{(*)}$ Named]]	moins de trois-quarts d'heure après Noël	Fuzzy obligatoire
… Nb$_{(a)}$ G$_{(*)}$ Nbf$_{(a)}$ Move$_{(rew)}$ … FOCTEMP$_{(part)}$[DATE[G$_{(*)}$ Named]]	[près d'] un mois et demi avant Noël	
… Nb$_{(a)}$ G$_{(*)}$ Nbf$_{(a)}$ Move$_{(fwd)}$ … FOCTEMP$_{(part)}$[DATE[G$_{(*)}$ Named]]	[plus ou moins] une semaine et demi après Noël	

Spécification générale B :

$$(Fuzzy \mid PartOf_{(*)})^{(1,N)}\ TEMPSHIFT\ \left(Ref_{(*)}\ G_{(*)}\ Move_{(*)}\ Amp_{(*)} \right)$$

Spécification	Exemple	Commentaires
(6) (Fuzzy | PartOf$_{(*)}$ | Fuzzy PartOf$_{(*)}$) TEMPSHIFT[G$_{(*)}$ Ref$_{(*)}$ Move$_{(*)}$ Amp$_{(*)}$]		
REM : Le groupe « Ref Move Amp » peut apparaître après ou, éventuellement, avant G$_{(*)}$		
REM : Poy G$_{(hyear\|tyear\|qyear)}$ peut aussi être utilisé comme unité pour le déplacement temporel.		
… Ref$_{(now)}$ Move$_{(fwd)}$ …	(vers | le début de) la semaine prochaine (vers le début de)	[PartOf$_{(beg)}$], G$_{(week)}$, Amp$_{(1)}$
… Ref$_{(now)}$ Move$_{(rew)}$ …	(aux environs du | la fin du) mois dernier (aux environs de la fin du)	[PartOf$_{(end)}$], G$_{(month)}$, Amp$_{(1)}$
… Ref$_{(now)}$ Move$_{(rew)}$ …	(aux environs du | la fin du) mois dernier (aux environs de la fin du)	[PartOf$_{(end)}$], G$_{(month)}$, Amp$_{(1)}$
… Ref$_{(focus)}$ Move$_{(fwd)}$ …	(durant le | la première partie du) jour d'après (durant la première partie du)	[PartOf$_{(beg)}$], G$_{(day)}$, Amp$_{(1)}$
… Ref$_{(focus)}$ Move$_{(rew)}$ …	(dans le courant de | la fin de) l'heure précédente (dans le courant de la fin de)	[PartOf$_{(end)}$], G$_{(hour)}$, Amp$_{(1)}$

$$(Fuzzy \mid PartOf_{(*)})^{(1,N)}\ TARGET_{(part)}\ \left(G_{(*)} \right)\ TEMPSHIFT\ \left(Ref_{(*)} \right)$$

Spécification	Exemple	Commentaires
(5) (Fuzzy | PartOf$_{(*)}$ | Fuzzy PartOf$_{(*)}$) TARGET$_{(part)}$[G$_{(*)}$] TEMPSHIFT[Ref$_{(*)}$]		
REM : G$_{(sec)}$ (seconde) n'est pas repris car ambigu avec « deuxième ».		
REM : Poy G$_{(hyear\|tyear\|qyear)}$ peut aussi être utilisé comme unité pour le déplacement temporel.		
REM : le cas Ref$_{(now)}$ G$_{(*)}$ hors contexte est assez ambigu et pourrait très bien ne pas être Fuzzy (cette (année | après midi), j'irai chez le dentiste / cette (année | après midi) a été excellente pour les affaires)		
Fuzzy TARGET$_{(part)}$[G$_{(*)}$] TEMPSHIFT[Ref$_{(now)}$]	dans le courant de (cet |l') après-midi	G$_{(pod_af)}$
Fuzzy TARGET$_{(part)}$[G$_{(*)}$] TEMPSHIFT[Ref$_{(focus)}$]	dans le courant de cet après-midi-là	G$_{(pod_af)}$
PartOf$_{(*)}$ TARGET$_{(part)}$[G$_{(*)}$] TEMPSHIFT[Ref$_{(now)}$]	(cette | la) fin d'année	PartOf$_{(end)}$, G$_{(year)}$
PartOf$_{(*)}$ TARGET$_{(part)}$[G$_{(*)}$] TEMPSHIFT[Ref$_{(focus)}$]	la fin de ce mois-là	PartOf$_{(end)}$, G$_{(month)}$
Fuzzy PartOf$_{(*)}$ TARGET$_{(part)}$[G$_{(*)}$] TEMPSHIFT[Ref$_{(now)}$]	durant le début de (cette | la) soirée	PartOf$_{(beg)}$, G$_{(pod_ev)}$
Fuzzy PartOf$_{(*)}$ TARGET$_{(part)}$[G]$_{(*)}$ TEMPSHIFT[Ref$_{(focus)}$]	vers le début de cet après-midi-là	PartOf$_{(beg)}$, G$_{(pod_af)}$

Spécification générale C :

$$(Fuzzy \mid PartOf_{(*)})^{(1,N)} \; TARGET_{(part)} \; \left(G_{(*>=day)} \right) \; TEMPSHIFT \; \left(Ref_{(*)} \; (Move_{(*)} \; Amp_{(*)} \; G_{(*)})^{(0,1)} \right)$$

Spécification	Exemple	Commentaires
(7) (Fuzzy | PartOf$_{(*)}$ | Fuzzy PartOf$_{(*)}$) TARGET$_{(part)}$[G$_{(*<=day)}$] TEMPSHIFT[Ref$_{(*)}$ (Move$_{(*)}$ Amp$_{(*)}$ G$_{(*)}$)$^{(0,1)}$]		
REM : Les groupes TEMPSHIFT et TARGET peuvent être inversés.		
(i) G$_{(day\|week\|we)}$		
… G$_{(day)}$ … Ref$_{(now)}$ Move$_{(fwd)}$ Amp$_{(1)}$ …	(vers | le début du) 22 décembre [prochain] (vers le début du)	PartOf$_{(beg)}$
… G$_{(week)}$ … Ref$_{(focus)}$ Move$_{(fwd)}$ Amp$_{(1)}$ …	(vers | la première partie de) la semaine du 22 décembre [suivant] (vers la première partie de)	PartOf$_{(beg)}$
(ii) G$_{(month)}$		
… G$_{(month)}$ … Ref$_{(now)}$ Move$_{(rew)}$ Amp$_{(1)}$ G$_{(*)}$ …	(durant le | la fin du) mois de janvier [[de l'année] dernièr[e]] (durant la fin du)	PartOf$_{(end)}$
… G$_{(month)}$ …	(au cours de | le début de) septembre [-] (au cours du début de)	PartOf$_{(beg)}$
(iii) Named G$_{(days\|week\|day)}$		
… Named G$_{(days)}$ … Ref$_{(focus)}$ Move$_{(rew)}$ Amp$_{(1)}$ …	(dans le courant des | la fin des) [précédentes] vacances de Pâques (dans le courant de la fin des)	PartOf$_{(end)}$
… Named G$_{(week)}$ … Ref$_{(now)}$ Move$_{(fwd)}$ Amp$_{(1)}$ …	(lors de | le début) la semaine pascale [prochaine] (lors du début de)	PartOf$_{(beg)}$
… [jour de] Named G$_{(day)}$ … Ref$_{(now)}$ Move$_{(rew)}$ Amp$_{(1)}$ …	(aux environs | le milieu) du jour de Noël [passé] (aux environs du milieu)	PartOf$_{(mid)}$
(iv) Season G$_{(season)}$		
… Season G$_{(season)}$ … Ref$_{(now)}$ Move$_{(fwd)}$ Amp$_{(1)}$ …	(courant | fin) du printemps [prochain] (courant fin)	PartOf$_{(end)}$
(v) Poy G$_{(qyear\|tyear\|hyear)}$		
… Poy G$_{(qyear)}$ …	(lors du | debut du) premier trimestre [-] (lors du début du)	PartOf$_{(beg)}$
… Poy G$_{(hyear)}$ …	(pendant le | la fin du) 1er semestre [-] (pendant la fin du)	PartOf$_{(end)}$
… Poy G$_{(tyear)}$ …	(vers le | le milieu du) deuxième quadrimestre[-] (vers le milieu du)	PartOf$_{(mid)}$

$(Fuzzy \mid PartOf_{(*)})^{(1,N)}\ TEMPSHIFT\ \left(Ref_{(*)}\ Move_{(*)}\ Amp_{(*)}\ G_{(*>=day)} \right)$

Spécification	Exemple	Commentaires
(8) (Fuzzy | PartOf$_{(*)}$ | Fuzzy PartOf$_{(*)}$) TEMPSHIFT[Ref$_{(*)}$ Move$_{(*)}$ Amp$_{(*)}$ G$_{(day)}$]		
… Ref$_{(now)}$ Move$_{(rew)}$ Amp$_{(1)}$ …	aux environs d'hier	
… Ref$_{(focus)}$ Move$_{(fwd)}$ Amp$_{(1)}$ …	vers le lendemain	

Spécification générale D :

$TARGET_{(part)}\ \left(G_{(day)} \right)\ (Fuzzy \mid PartOf_{(*)})^{(1,N)}\ TARGET_{(part)}\ \left(G_{(pod)} \right)\ (Fuzzy\, TARGET_{(part)}\ \left(G_{(min)} \right))^{0,1}\ TEMPSHIFT\ \left(Ref_{(now)} \right)$

Spécification	Exemple	Commentaires
(9) TARGET$_{(part)}$[G$_{(day)}$] (Fuzzy | PartOf$_{(*)}$ | Fuzzy PartOf$_{(*)}$) TARGET$_{(part)}$[G$_{(pod_*)}$] (Fuzzy TARGET$_{(part)}$[G$_{(min)}$])$^{(0,1)}$ TEMPSHIFT[Ref$_{(now)}$]		
REM : L'ordre du groupe TARGET$_{(part)}$[G$_{(pod_*)}$] peut varier. C'est aussi le cas pour la mention facultative de l'heure.		
… G$_{(day)}$ … Fuzzy … G$_{(pod_*)}$ …	mardi peu de temps avant l'aube [, vers 5 heures]	G$_{(pod_da)}$
… Named G$_{(day)}$ … Partof$_{(*)}$ … G$_{(pod_*)}$ …	le jour de Noël, le début de la soirée [, aux environs de 20h45]	G$_{(pod_ev)}$
… G$_{(day)}$ … Fuzzy Partof$_{(*)}$ … G$_{(pod_*)}$ …	mardi, en début de matinée [, vers 9 heures]	G$_{(pod_da)}$
… Named G$_{(day)}$ … Fuzzy … G$_{(pod_*)}$ …	le jour de Noël, en soirée [, aux environs de 23h45]	G$_{(pod_ev)}$
… G$_{(day)}$ … Partof$_{(*)}$ … G$_{(pod_*)}$ …	mardi, fin de matinée [, vers 9 heures]	G$_{(pod_da)}$
… Named G$_{(day)}$ … Fuzzy Partof$_{(*)}$ … G$_{(pod_*)}$ …	le jour de Noël, lors de la fin de soirée [, aux environs de 23h45]	G$_{(pod_ev)}$

$TEMPSHIFT\ \left(Ref_{(*)}\ Move_{(*)}\ Amp_{(*)}\ G_{(day)} \right)\ (Fuzzy \mid PartOf_{(*)})^{(1,N)}\ TARGET_{(part)}\ \left(G_{(pod)} \right)\ (Fuzzy\, TARGET_{(part)}\ \left(G_{(min)} \right))^{0,1}$

Spécification	Exemple	Commentaires
(10) TEMPSHIFT[Ref$_{(*)}$ G$_{(day)}$ Move$_{(*)}$ Amp$_{(*)}$] (Fuzzy | PartOf$_{(*)}$ | Fuzzy PartOf$_{(*)}$) TARGET$_{(part)}$[G$_{(pod_*)}$] (Fuzzy TARGET$_{(part)}$[G$_{(min)}$])$^{(0,1)}$		
REM : L'ordre du groupe TARGET$_{(part)}$[G$_{(pod_*)}$] peut varier. C'est aussi le cas pour la mention facultative de l'heure.		
… Ref$_{(now)}$ G$_{(day)}$ Move$_{(no)}$ Amp$_{(0)}$ … Fuzzy … G$_{(pod_*)}$ …	aujourd'hui en cours d'après-midi [, vers 15 heures]	G$_{(pod_af)}$
… Ref$_{(now)}$ G$_{(day)}$ Move$_{(*)}$ Amp$_{(*)}$ … Fuzzy … G$_{(pod_*)}$ …	hier, peu de temps après midi [, vers 12 heures 30]	G$_{(pod_mi)}$
… Ref$_{(focus)}$ G$_{(day)}$ Move$_{(*)}$ Amp$_{(*)}$ … Fuzzy … G$_{(pod_*)}$ …	la veille en matinée [, vers 10 heures]	G$_{(pod_mo)}$
… Ref$_{(now)}$ G$_{(day)}$ Move$_{(no)}$ Amp$_{(0)}$ … Partof$_{(*)}$ … G$_{(pod_*)}$ …	aujourd'hui début d'après-midi [, vers 15 heures]	G$_{(pod_af)}$
… Ref$_{(now)}$ G$_{(day)}$ Move$_{(*)}$ Amp$_{(*)}$ … Partof$_{(*)}$ … G$_{(pod_*)}$ …	hier à la mi-journée [, vers 12 heures 30]	G$_{(pod_mi)}$
… Ref$_{(focus)}$ G$_{(day)}$ Move$_{(*)}$ Amp$_{(*)}$ … Partof$_{(*)}$ … G$_{(pod_*)}$ …	la veille, fin de matinée [, vers 10 heures]	G$_{(pod_mo)}$

Spécification	Exemple	Commentaires
… Ref$_{(now)}$ G$_{(day)}$ Move$_{(no)}$ Amp$_{(0)}$ … Fuzzy Partof$_{(*)}$ … G$_{(pod_*)}$ …	aujourd'hui en début d'après-midi [, vers 15 heures]	G$_{(pod_af)}$
… Ref$_{(now)}$ G$_{(day)}$ Move$_{(*)}$ Amp$_{(*)}$ … Fuzzy Partof$_{(*)}$ … G$_{(pod_*)}$ …	hier au cours de la mi-journée [, vers 12 heures 30]	G$_{(pod_mi)}$
… Ref$_{(focus)}$ G$_{(day)}$ Move$_{(*)}$ Amp$_{(*)}$ … Fuzzy Partof$_{(*)}$ … G$_{(pod_*)}$ …	la veille, en fin de matinée [, vers 10 heures]	G$_{(pod_mo)}$

Spécification générale E :

$$TARGET_{(part)}\ \left(G_{(day)}\right)\ TARGET^{(0,1)}_{(part)}\ \left(G_{(pod)}\right)\ Fuzzy\,TARGET_{(part)}\ \left(G_{(min)}\right)\ TEMPSHIFT\ \left(Ref_{(now)}\right)$$

Spécification	Exemple	Commentaires
(11) TARGET$_{(part)}$[G$_{(day)}$] TARGET$^{(0,1)}_{(part)}$[G$_{(pod_*)}$] Fuzzy TARGET$_{(part)}$[G$_{(min)}$] TEMPSHIFT[Ref$_{(now)}$]		
REM : L'ordre du groupe Fuzzy TARGET$_{(part)}$[G$_{(min)}$] et de TARGET$_{(part)}$[G$_{(day)}$ peut être inversé		
REM : L'ordre du groupe TARGET$_{(part)}$[G$_{(pod_*)}$] peut varier		
… G$_{(day)}$ … Heure$_{(1)}$Min$_{(1)}$	mardi, aux environs de 13h45	
… Named G$_{(day)}$ … Heure$_{(1)}$Min$_{(1)}$	le jour de Noël, aux environs de 13h45	
… G$_{(day)}$ … G$_{(pod_*)}$ … Heure$_{(1)}$	mardi à l'aube, aux environs de 5 heures	G$_{(pod_da)}$
… Named G$_{(day)}$ … G$_{(pod_*)}$ … Heure$_{(1)}$	le jour de Noël au soir, aux environs de 20h	G$_{(pod_ev)}$

$$TEMPSHIFT\ \left(Ref_{(*)}\ Move_{(*)}\ Amp_{(*)}\ G_{(day)}\right)\ TARGET^{(0,1)}_{(part)}\ \left(G_{(pod)}\right)\ Fuzzy\,TARGET_{(part)}\ \left(G_{(min)}\right)$$

Spécification	Exemple	Commentaires
(12) TEMPSHIFT[Ref$_{(*)}$ G$_{(day)}$ Move$_{(*)}$ Amp$_{(*)}$] TARGET$^{(0,1)}_{(part)}$[G$_{(pod_*)}$] Fuzzy TARGET$_{(part)}$[G$_{(min)}$]		
REM : L'ordre du groupe Fuzzy TARGET$_{(part)}$[G$_{(min)}$] et de TEMPSHIFT peut être inversé		
REM : L'ordre du groupe TARGET$_{(part)}$[G$_{(pod_*)}$] peut varier		
… Ref$_{(now)}$ G$_{(day)}$ Move$_{(*)}$ Amp$_{(*)}$ … Heure$_{(1)}$Min$_{(1)}$	aujourd'hui, vers 13 heures 30	
… Ref$_{(now)}$ G$_{(day)}$ Move$_{(no)}$ Amp$_{(0)}$ … Heure$_{(1)}$Min$_{(1)}$	aux environs de 13h45	G$_{(min)}$
… Ref$_{(now)}$ G$_{(day)}$ Move$_{(*)}$ Amp$_{(*)}$ … Heure$_{(1)}$Min$_{(1)}$	hier, vers 13 heures 15	
… Ref$_{(focus)}$ G$_{(day)}$ Move$_{(*)}$ Amp$_{(*)}$ … Heure$_{(1)}$Min$_{(1)}$	la veille, vers 13 heures 40	
… Ref$_{(now)}$ G$_{(day)}$ Move$_{(*)}$ Amp$_{(*)}$ … G$_{(pod_*)}$ … Heure$_{(1)}$	aujourd'hui après-midi, vers 15h	G$_{(pod_af)}$
… Ref$_{(now)}$ G$_{(day)}$ Move$_{(*)}$ Amp$_{(*)}$ … G$_{(pod_*)}$ … Heure$_{(1)}$	hier midi, vers 12 heures	G$_{(pod_mi)}$
… Ref$_{(focus)}$ G$_{(day)}$ Move$_{(*)}$ Amp$_{(*)}$ … G$_{(pod_*)}$ … Heure$_{(1)}$	la veille au matin, vers 10 heures	G$_{(pod_mo)}$

B.4.6 DAPU : Référence Durative, Absolue, Précise et Unique

Spécification générale A : Jour et intervalle de parties de journées ou d'heures.

$$PAPU\ \left(G_{(day)}\right)\ BOUND_{(*)}^{(1,2)}\ \left(G_{(pod\ |\ min)}\right)$$

Exemple	Commentaires
(1) DATE[G$_{(*)}$] BOUND.*[(HEURE[G$_{(min)}$] | G$_{(pod)}$)]	
(2) DATE[G$_{(*)}$] BOUND.Lower[(HEURE[G$_{(min)}$] | G$_{(pod)}$)] BOUND.Upper[(HEURE[G$_{(min)}$] | G$_{(pod)}$)]	
REM : L'ordre de DATE et du groupe BOUND peut être inversé	
10/03/2005, entre 10h45 et 12h30	
10 mars 2005, depuis le matin	
Noël 2005, depuis 10h45	

Spécification générale B : Intervalle de dates.

$$BOUND_{(*)}^{(1,2)}\ \left(PAPU\ \left(G_{(*)}\right)\right)$$

Exemple	Commentaires
(3) BOUND.Lower[HEURE[G$_{(min)}$]$^{(0,1)}$ G$_{(pod)}{}^{(0,1)}$ DATE[G$_{(*)}$]] BOUND.Upper[HEURE[G$_{(min)}$]$^{(0,1)}$ G$_{(pod)}{}^{(0,1)}$ DATE[G$_{(*)}$]]	
(5) BOUND.Lower[HEURE[G$_{(min)}$]$^{(0,1)}$ G$_{(pod)}{}^{(0,1)}$ DATE[G$_{(*)}$]]	
(6) BOUND.Upper[HEURE[G$_{(min)}$]$^{(0,1)}$ G$_{(pod)}{}^{(0,1)}$ DATE[G$_{(*)}$]]	
les années 2009 à 2010	
du 8 mars 2005 au 10 mars 2005	
du 8 au 10 mars 2005	(mois et année factorisés)
depuis décembre 2003	
jusqu'au 10 mars 2005	
(4) G$_{(pod)}$ BOUND.Lower[HEURE[G$_{(min)}$]$^{(0,1)}$ G$_{(pod)}{}^{(0,1)}$ DATE[G$_{(*)}$]] BOUND.Upper[HEURE[G$_{(min)}$]$^{(0,1)}$ G$_{(pod)}{}^{(0,1)}$ DATE[G$_{(*)}$]]	
la nuit du 9 au 10 mars 2005	

B.4.7 DAFU : Référence Durative, Absolue, Floue et Unique

Spécification générale A : Jour et intervalle de parties de journées ou d'heures.

$$PAPU\ \left(G_{(day)}\right)\ BOUND_{(*)}^{(1,2)}\ \left(\left(\ Fuzzy \mid PartOf_{(*)}\ \right)^{1,N}\ G_{(pod\ \mid\ min)}\right)$$

Exemple	Commentaires
(1) DATE[$G_{(*)}$] BOUND.*[FuzzyPartOf (HEURE[$G_{(min)}$] | $G_{(pod)}$)]	
(2) BOUND.Lower[FuzzyPartOf$^{(0,1)}$ HEURE[$G_{(min)}$]$^{(0,1)}$ $G_{(pod)}{}^{(0,1)}$ DATE[$G_{(*)}$]] BOUND.Upper[FuzzyPartOf$^{(0,1)}$ HEURE[$G_{(min)}$]$^{(0,1)}$ $G_{(pod)}{}^{(0,1)}$ DATE[$G_{(*)}$]]	
REM : Au moins une des deux bornes doit intégrer une marque d'imprécision Fuzzy et/ou PartOf.	
10 mars 2005, depuis le début de matinée	
10 mars 2005, depuis 10h45 jusqu'aux alentours de 12h30	

Spécification générale B : Intervalle de dates.

$$BOUND_{(*)}^{(1,2)}\ \left(\left(\ Fuzzy \mid PartOf_{(*)}\ \right)^{1,N}\ PAPU\ \left(G_{(*)}\right)\right)$$

Exemple	Commentaires
(3) BOUND.Lower[HEURE[$G_{(min)}$]$^{(0,1)}$ $G_{(pod)}{}^{(0,1)}$ DATE[$G_{(*)}$]] BOUND.Upper[HEURE[$G_{(min)}$]$^{(0,1)}$ $G_{(pod)}{}^{(0,1)}$ DATE[$G_{(*)}$]]	
(5) BOUND.Lower[FuzzyPartOf$^{(0,1)}$ HEURE[$G_{(min)}$]$^{(0,1)}$ $G_{(pod)}{}^{(0,1)}$ DATE[$G_{(*)}$]]	
(6) BOUND.Upper[FuzzyPartOf$^{(0,1)}$ HEURE[$G_{(min)}$]$^{(0,1)}$ $G_{(pod)}{}^{(0,1)}$ DATE[$G_{(*)}$]]	
depuis le 8 mars 2005 jusqu'aux environs du 10 mars 2005	
à partir de début décembre 2009	
jusqu'à fin 2010	
(4) FuzzyPartOf $G_{(pod)}$ BOUND.Lower[HEURE[$G_{(min)}$]$^{(0,1)}$ $G_{(pod)}{}^{(0,1)}$ DATE[$G_{(*)}$]] BOUND.Upper[HEURE[$G_{(min)}$]$^{(0,1)}$ $G_{(pod)}{}^{(0,1)}$ DATE[$G_{(*)}$]]	
durant la nuit du 9 au 10 mars 2005	
au début de la nuit du 9 au 10 mars 2005	

B.4.8 DRPU : Référence Durative, Relative, Précise et Unique

Spécification générale A : Jour et intervalle de parties de journées ou d'heures.

$$PRPU^{(0,1)}\left(G_{(day)}\right)\ BOUND^{(1,2)}_{(*)}\left(G_{(pod\ |\ min)}\ TEMPSHIFT\left(Ref_{(*)}\right)\right)$$

Exemple	Commentaires
(1) DATE[G$_{(*)}$]$^{(0,1)}$ BOUND.*[(HEURE[G$_{(min)}$] | G$_{(pod)}$) TEMPSHIFT]	
(2) DATE[G$_{(*)}$]$^{(0,1)}$ BOUND.Lower[(HEURE[G$_{(min)}$] | G$_{(pod)}$) TEMPSHIFT] BOUND.Upper[(HEURE[G$_{(min)}$] | G$_{(pod)}$) TEMPSHIFT]	
REM : L'ordre de DATE et du groupe BOUND peut être inversé	
[jeudi,] entre 10h45 et 12h30	
[le 10 mars,] depuis 10h45	
[le jeudi 10 mars,] jusqu'à soir	

Spécification générale B : Intervalle de dates.

$$BOUND^{(1,2)}_{(*)}\left(PRPU\left(G_{(*)}\right)\right)$$

Exemple	Commentaires
(3) BOUND.Lower[HEURE[G$_{(min)}$]$^{(0,1)}$ G$_{(pod)}{}^{(0,1)}$ DATE[G$_{(*)}$] TEMPSHIFT] BOUND.Upper[HEURE[G$_{(min)}$]$^{(0,1)}$ G$_{(pod)}{}^{(0,1)}$ DATE[G$_{(*)}$] TEMPSHIFT]	
(5) BOUND.Lower[HEURE[G$_{(min)}$]$^{(0,1)}$ G$_{(pod)}{}^{(0,1)}$ DATE[G$_{(*)}$] TEMPSHIFT]	
(6) BOUND.Upper[HEURE[G$_{(min)}$]$^{(0,1)}$ G$_{(pod)}{}^{(0,1)}$ DATE[G$_{(*)}$] TEMPSHIFT]	
du 8 au 10 mars	
jusqu'au 10 mars	
dès le 18 mai prochain	
jusqu'au 31 décembre de l'an dernier	
(4) G$_{(pod)}$ BOUND.Lower[HEURE[G$_{(min)}$]$^{(0,1)}$ G$_{(pod)}{}^{(0,1)}$ DATE[G$_{(*)}$] TEMPSHIFT] BOUND.Upper[HEURE[G$_{(min)}$]$^{(0,1)}$ G$_{(pod)}{}^{(0,1)}$ DATE[G$_{(*)}$] TEMPSHIFT]	
la nuit de mercredi à jeudi	
(7) BOUND.Lower[TEMPSHIFT]	
(8) BOUND.Upper[TEMPSHIFT]	
dans les trois jours	
depuis deux mois	

B.4.9 DRFU : Référence Durative, Relative, Floue et Unique

Spécification générale A : Jour et intervalle de parties de journées ou d'heures.

$$PRPU^{(0,1)}\ \left(G_{(day)}\right)\ BOUND^{(1,2)}_{(*)}\ \left(\ (\ Fuzzy \mid PartOf_{(*)}\ )^{1,N}\ G_{(pod\ \mid\ min)}\ TEMPSHIFT\ \left(Ref_{(*)}\right)\right)$$

Exemple	Commentaires
(1) DATE[G$_{(*)}$]$^{(0,1)}$ BOUND.*[FuzzyPartOf (HEURE[G$_{(min)}$] | G$_{(pod)}$) TEMPSHIFT]	
REM : L'ordre de DATE et BOUND peut être inversé	
[le 10 mars,] depuis environ 10h45	
[le jeudi 10 mars,] jusqu'en début de soirée	
[jeudi,] jusqu'aux alentours de 12h30	

Spécification générale B : Intervalle de dates.

$$BOUND^{(1,2)}_{(*)}\ \left(\ (\ Fuzzy \mid PartOf_{(*)}\ )^{1,N}\ PRPU\ \left(G_{(*)}\right)\right)$$

Exemple	Commentaires
(2) BOUND.Lower[FuzzyPartOf HEURE[G$_{(min)}$]$^{(0,1)}$ G$_{(pod)}$$^{(0,1)}$ DATE[G$_{(*)}$] TEMPSHIFT] BOUND.Upper[FuzzyPartOf HEURE[G$_{(min)}$]$^{(0,1)}$ G$_{(pod)}$$^{(0,1)}$ DATE[G$_{(*)}$] TEMPSHIFT]	
(4) BOUND.Lower[FuzzyPartOf HEURE[G$_{(min)}$]$^{(0,1)}$ G$_{(pod)}$$^{(0,1)}$ DATE[G$_{(*)}$] TEMPSHIFT]	
(5) BOUND.Upper[FuzzyPartOf HEURE[G$_{(min)}$]$^{(0,1)}$ G$_{(pod)}$$^{(0,1)}$ DATE[G$_{(*)}$] TEMPSHIFT]	
entre début janvier et fin février	
depuis début décembre	
jusqu'aux environs du 10 mars	
(3) FuzzyPartOf G$_{(pod)}$ BOUND.Lower[HEURE[G$_{(min)}$]$^{(0,1)}$ G$_{(pod)}$$^{(0,1)}$ DATE[G$_{(*)}$] TEMPSHIFT] BOUND.Upper[HEURE[G$_{(min)}$]$^{(0,1)}$ G$_{(pod)}$$^{(0,1)}$ DATE[G$_{(*)}$] TEMPSHIFT]	
durant la nuit de mercredi à jeudi	
(6) BOUND.Lower[FuzzyPartOf TEMPSHIFT]	
(7) BOUND.Upper[FuzzyPartOf TEMPSHIFT]	
d'ici plus ou moins trois jours	
depuis à peu près deux mois	
d'ici deux à trois semaines	

B.5 Autres catégories

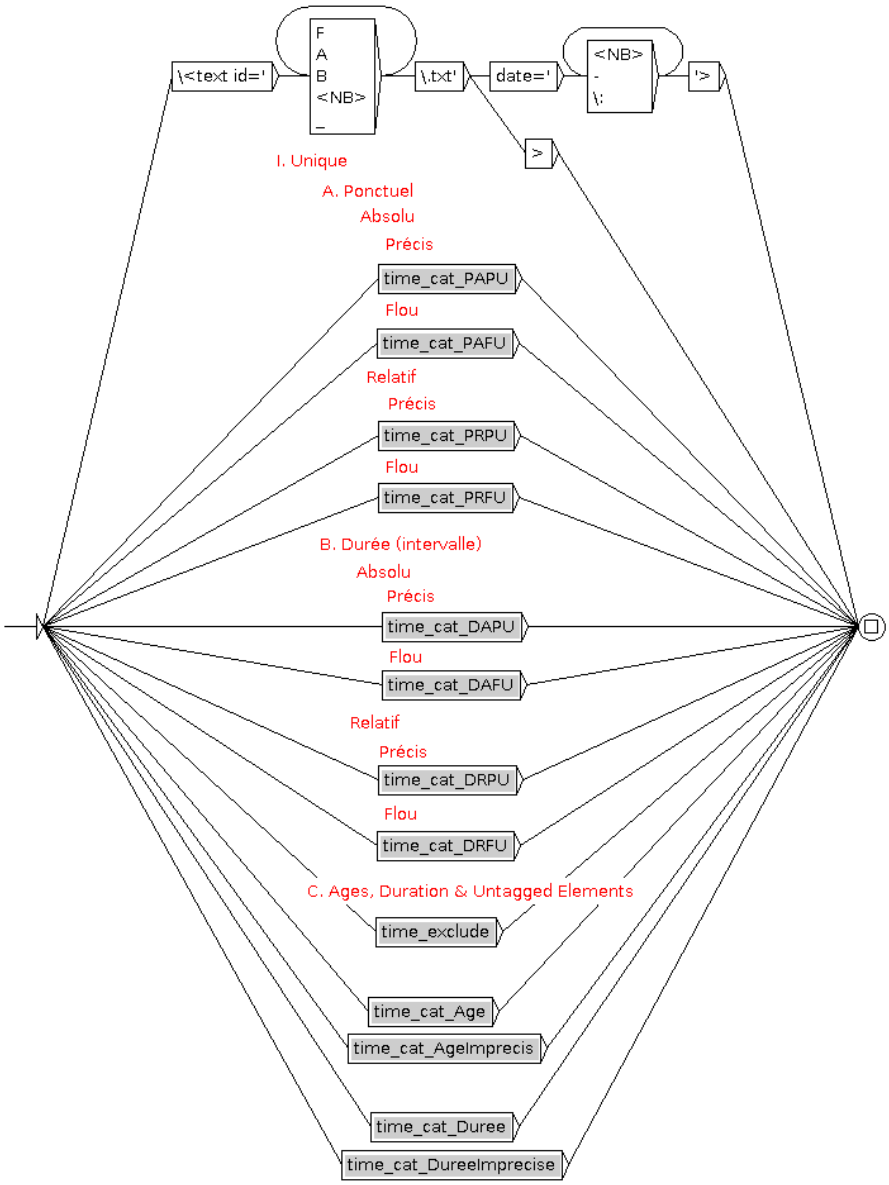
Les catégories présentes dans cette section recouvrent les durées, les âges ainsi que les expressions (souvent figées) utilisant le lexique des expressions temporelles sans en faire véritablement partie. Ces catégories ne sont donc pas exploitées en tant qu'expressions temporelles à part entière. Leur lien avec celles-ci est cependant évident et le bon fonctionnement de l'extraction exige qu'elles soient reconnues.

- Exclude : Cas particuliers à exclure des expressions temporelles
- Dur : Durée
- DurI : Durée imprécise
- Age : Âge
- AgeI : Âge imprécis

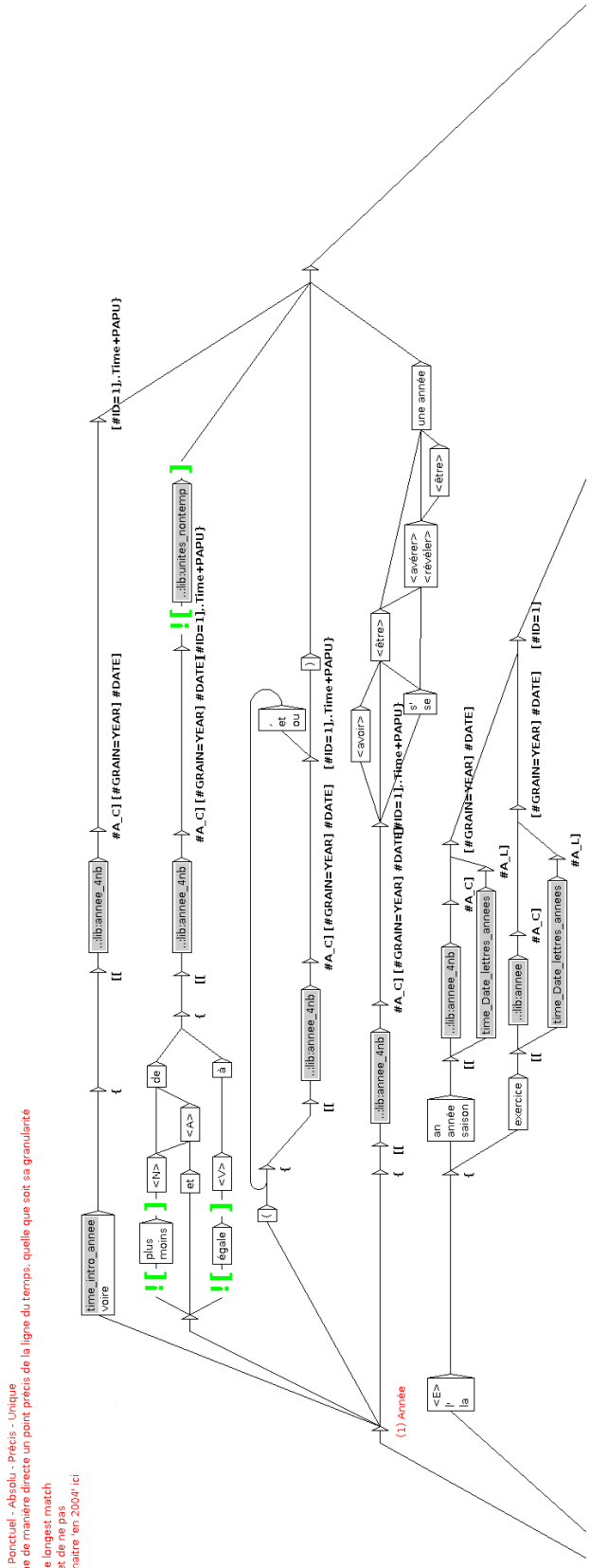
ANNEXE C

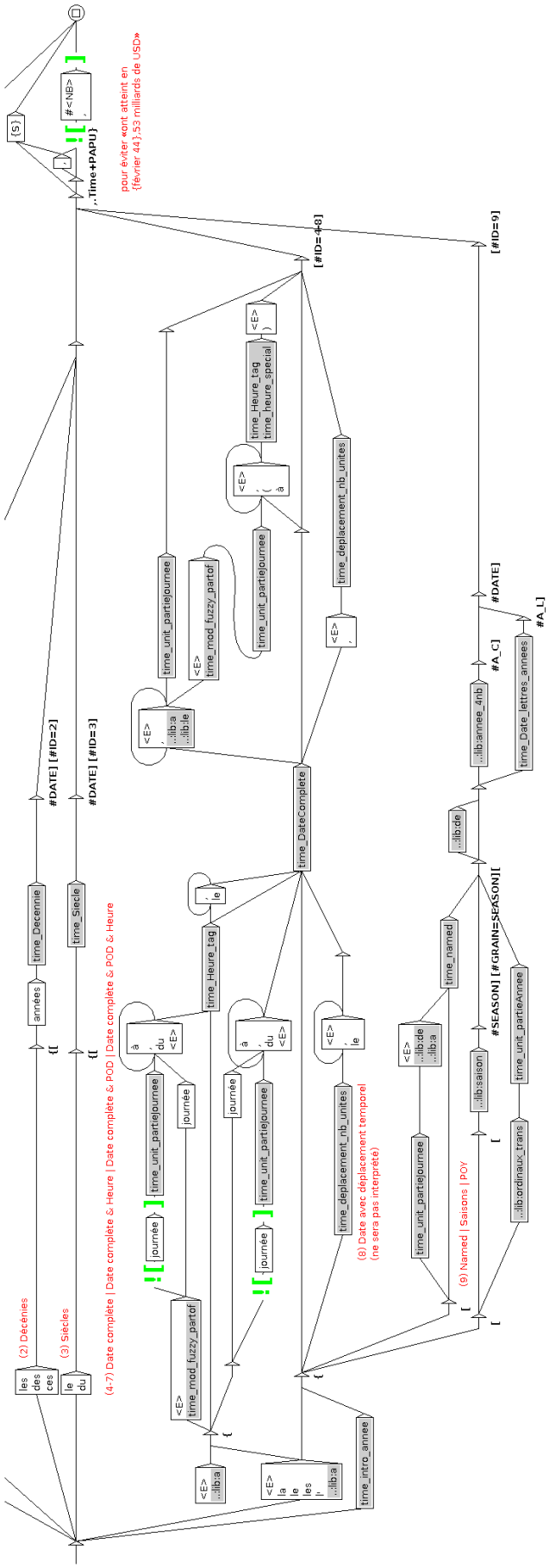
GRAPHES D'EXTRACTION

C.1 Graphe principal



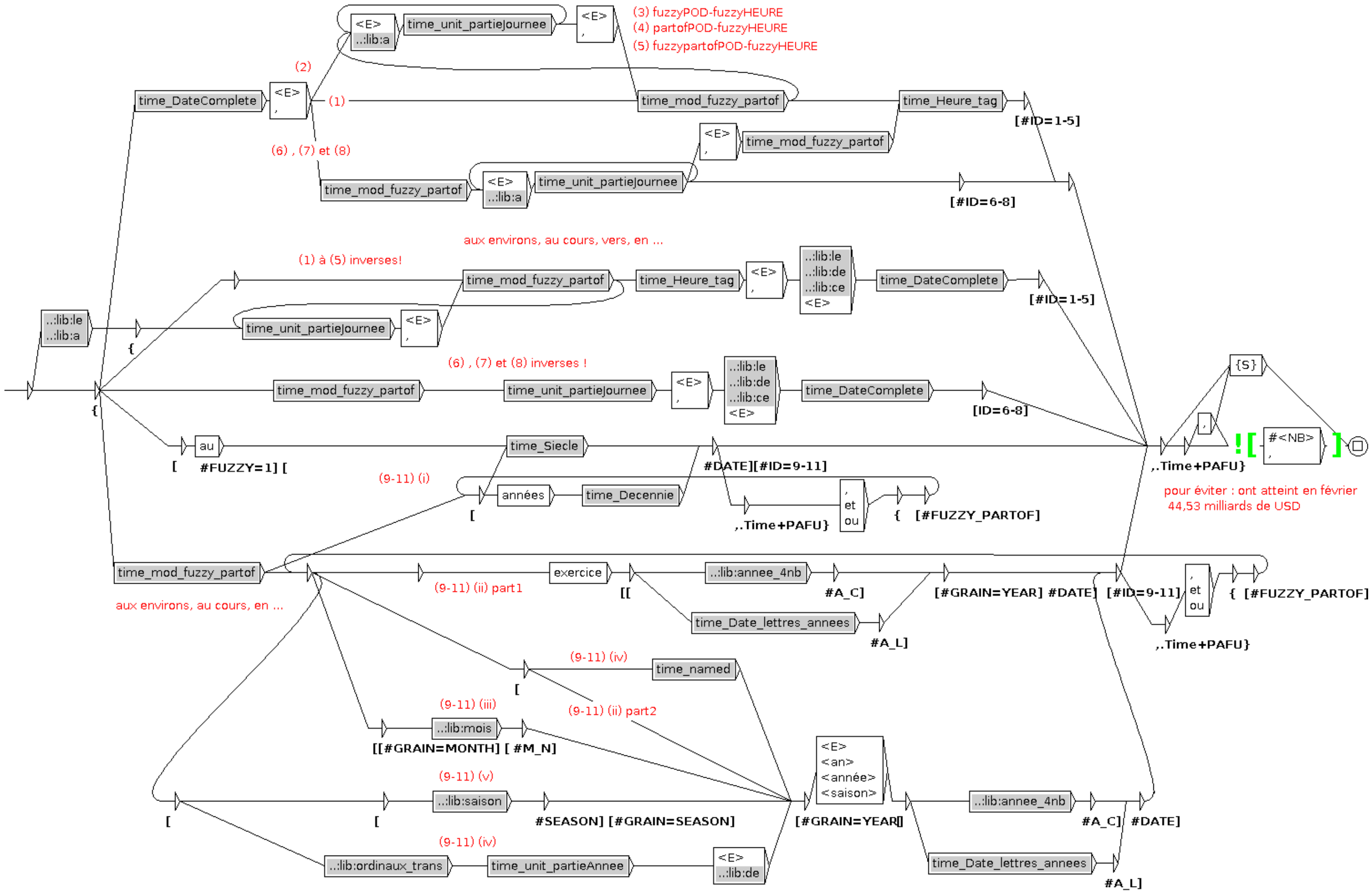
C.2 PAPU


PAPU : Ponctuel - Absolu - Précis : Unique
Désigne de manière directe un point précis de la ligne du temps, quelle que soit sa granularité
REM : le longest match permet de ne pas reconnaitre 'en 2004' ici
time_intro_annee
voire
plus
moins
égale
<N>
<A>
<V>
et
de
à
..lib:annee_4nb
..lib:unites_nontemp
#A_C] [#GRAIN=YEAR] #DATE]
[#ID=1]..Time+PAPU}
,
et
ou
)
(
une année
<être>
<avoir>
<avérer>
<révéler>
s'
se
an
année
saison
(1) Année
<E>
à
la
exercice
time_Date_lettres_annees
#A_C]
#A_L]
[#GRAIN=YEAR] #DATE]
[#ID=1]

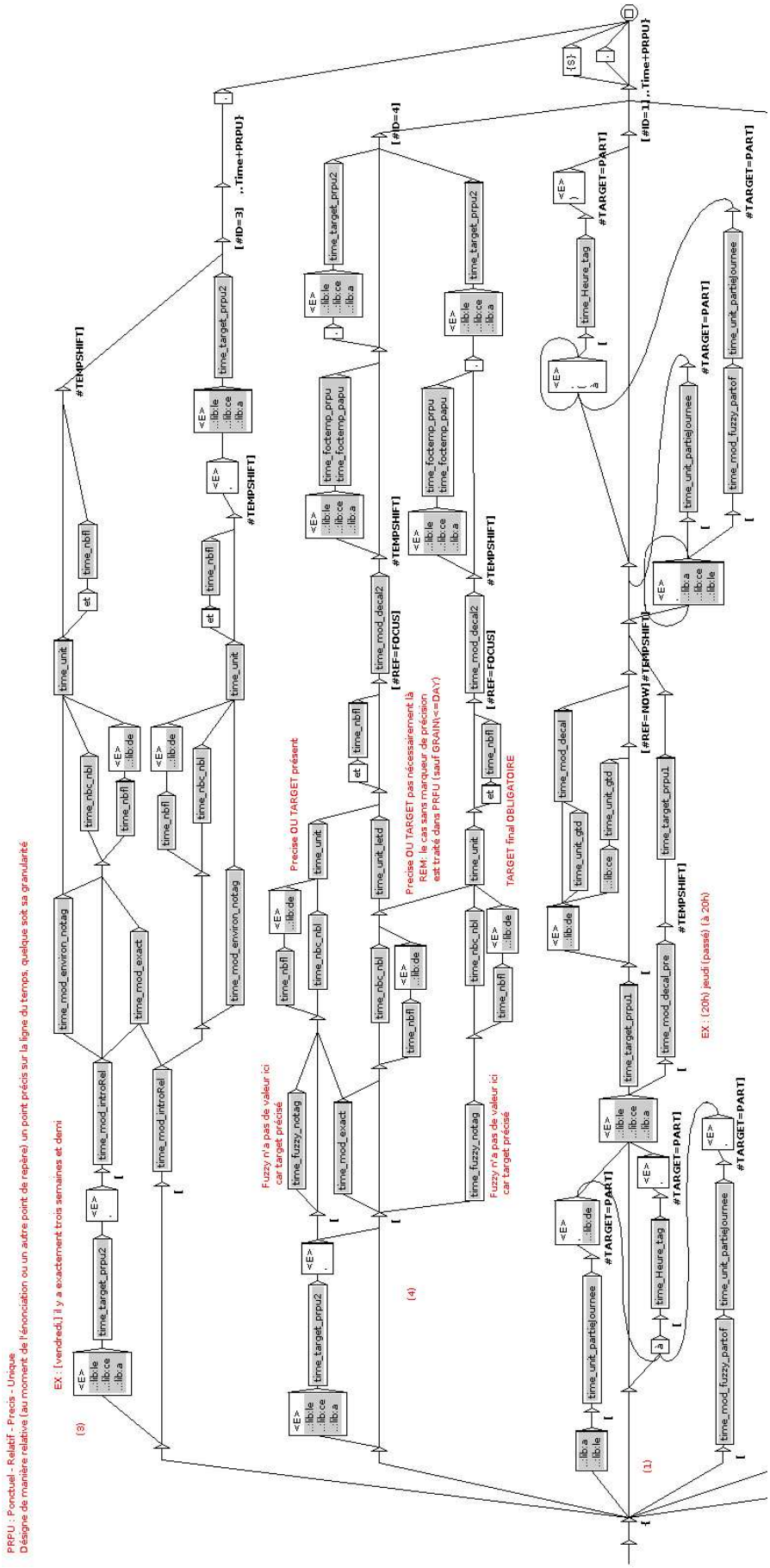

{S}
#<NB>
,
..Time+PAPU}
pour éviter «ont atteint en
{février 44},53 milliards de USD»
[#ID=4-8]
[#ID=9]
#DATE] [#ID=2]
#DATE] [#ID=3]
time_Decennie
années
time_Siecle
les
des
ces
le
du
(2) Décénies
(3) Siècles
(4-7) Date complète | Date complète & Heure | Date complète & POD | Date complète & POD & Heure
time_Date:complete
time_Heure_tag
time_unit_partie|journee
time_mod_fuzzy_partof
time_Heure_tag
time_heure_special
time_deplacement_nb_unites
time_deplacement_nb_unites
(8) Date avec déplacement temporel
(ne sera pas interprété)
time_named
time_unit_partie|journee
(9) Named | Saisons | POY
#SEASON] [#GRAIN=SEASON]
#DATE]
#A_C]
#A_L]
time_intro_annee
time_unit_partieAnnee
time_Date_lettres_annees
..lib:annee_4nb
..lib:saison
..lib:ordinaux_trans
..lib:de
..lib:a
..lib:le

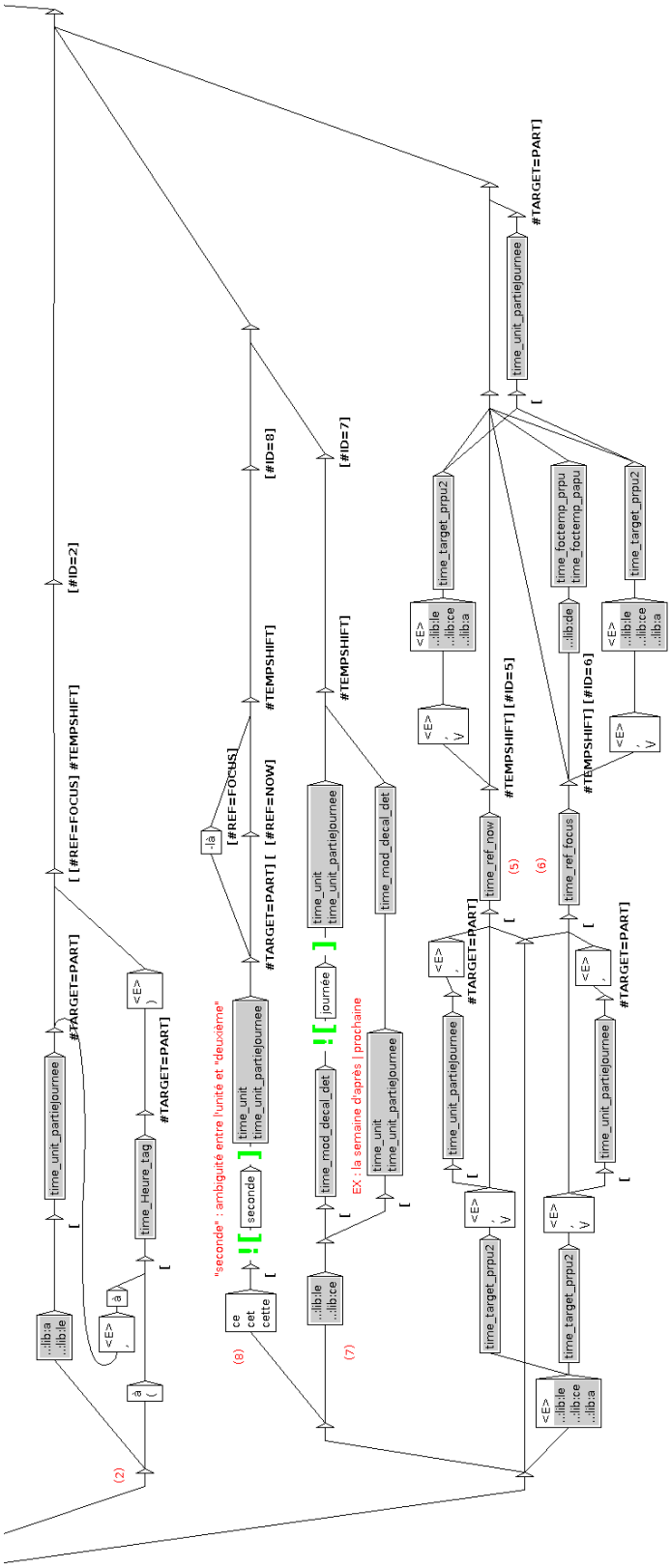
C.3 PAFU

PAFU : Ponctuel - Absolu - Flou - Unique
Désigne de manière directe un point imprécis de la ligne du temps, quelle que soit sa granularité

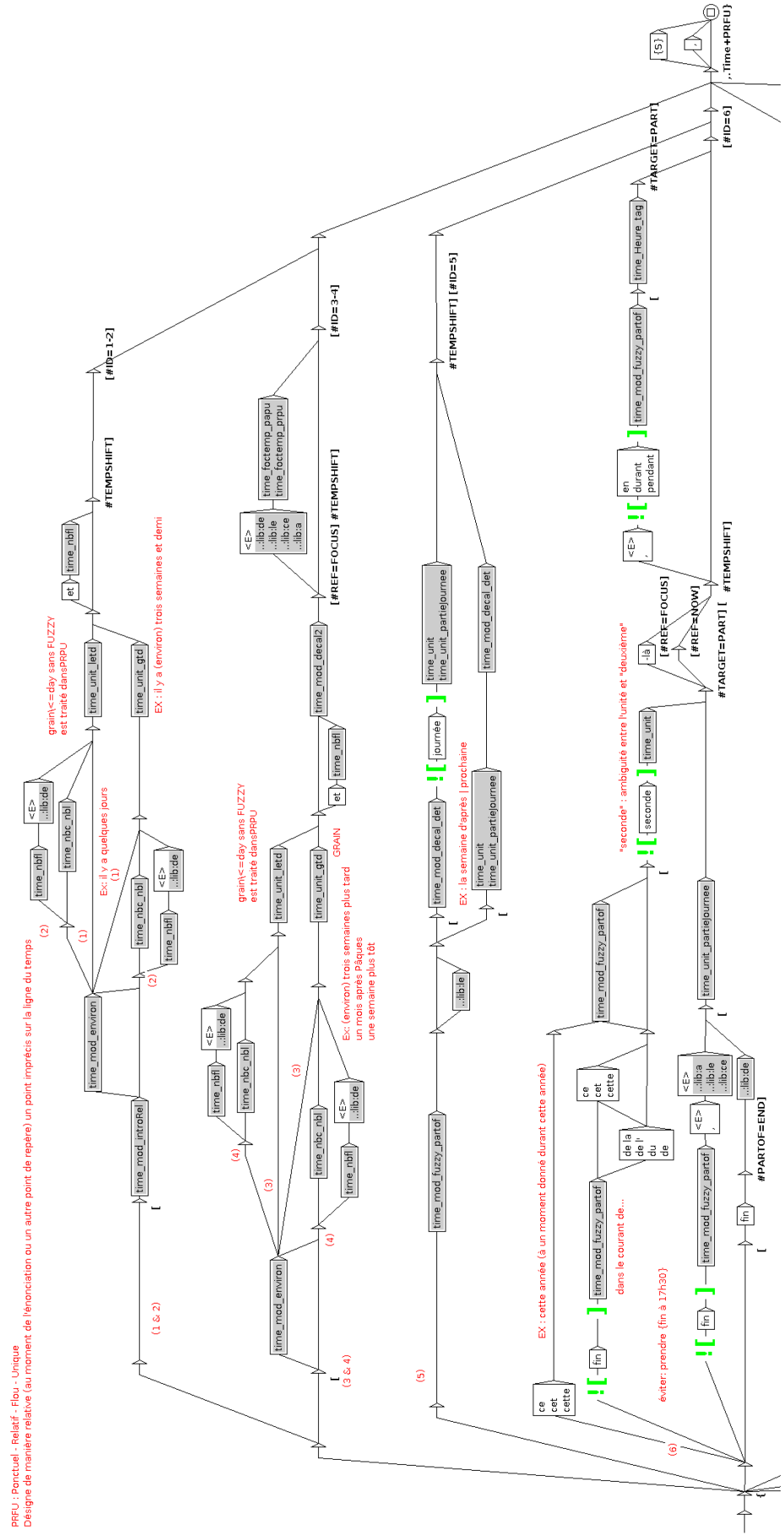


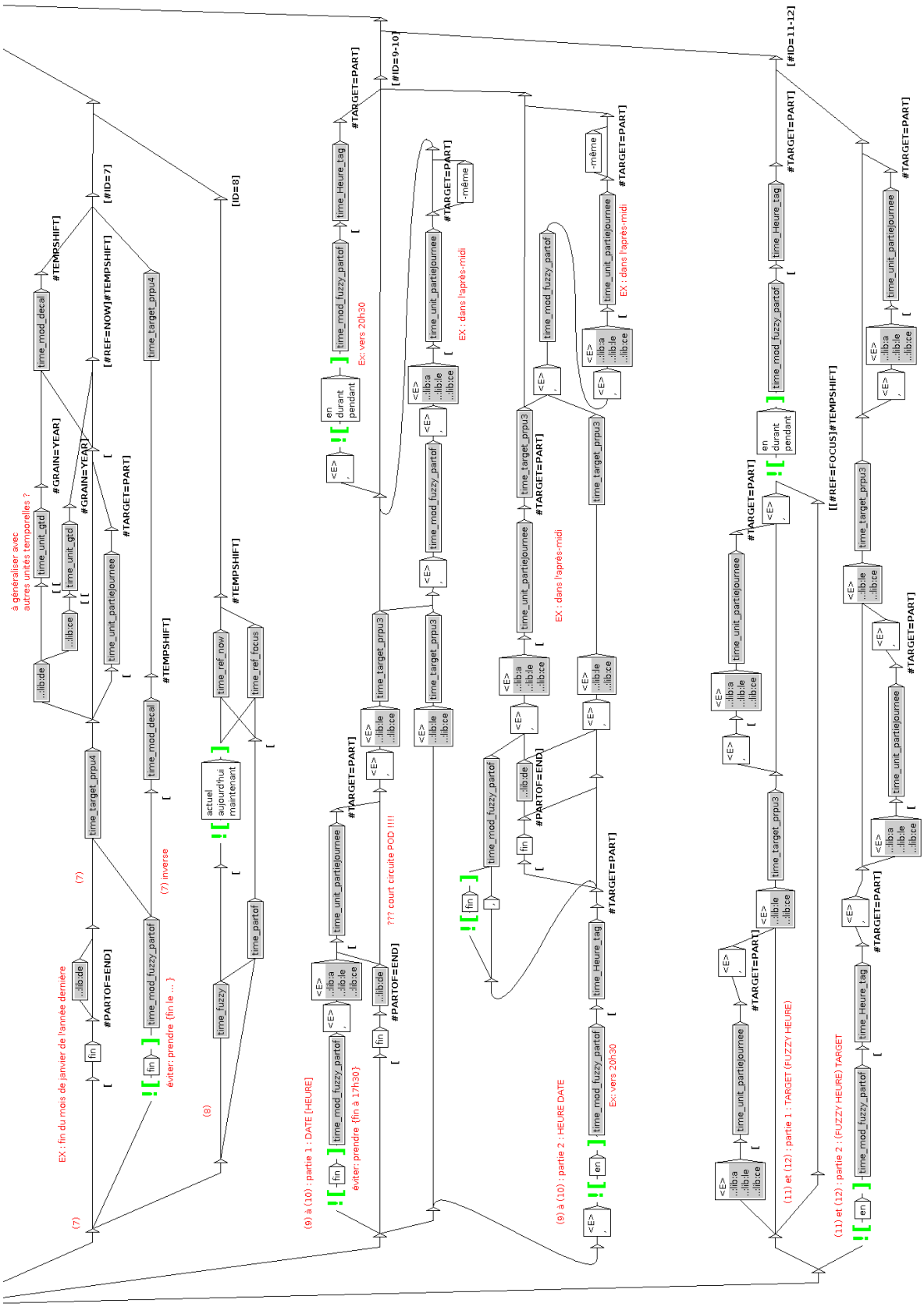
C.4 PRPU


PRPU : Ponctuel - Relatif - Precis - Unique
Désigne de manière relative (au moment de l'énonciation ou un autre point de repère) un point précis sur la ligne du temps, quelque soit sa granularité
EX : [vendredi] Il y a exactement trois semaines et demi
(3)
#TEMPSHIFT
[#ID=3] ..Time+PRPU}
Precise OU TARGET présent
Fuzzy n'a pas de valeur ici car target précisé
(4)
[#ID=4]
[#REF=FOCUS]
Precise OU TARGET pas nécessairement là
REM: le cas sans marqueur de précision est traité dans PRPU (sauf GRAIN\<=DAY)
TARGET final OBLIGATOIRE
Fuzzy n'a pas de valeur ici car target précisé
[#ID=1] ..Time+PRPU}
[#REF=NOW] #TEMPSHIFT]
#TARGET=PART]
EX : (20h) jeudi (passé) (à 20h)
(1)

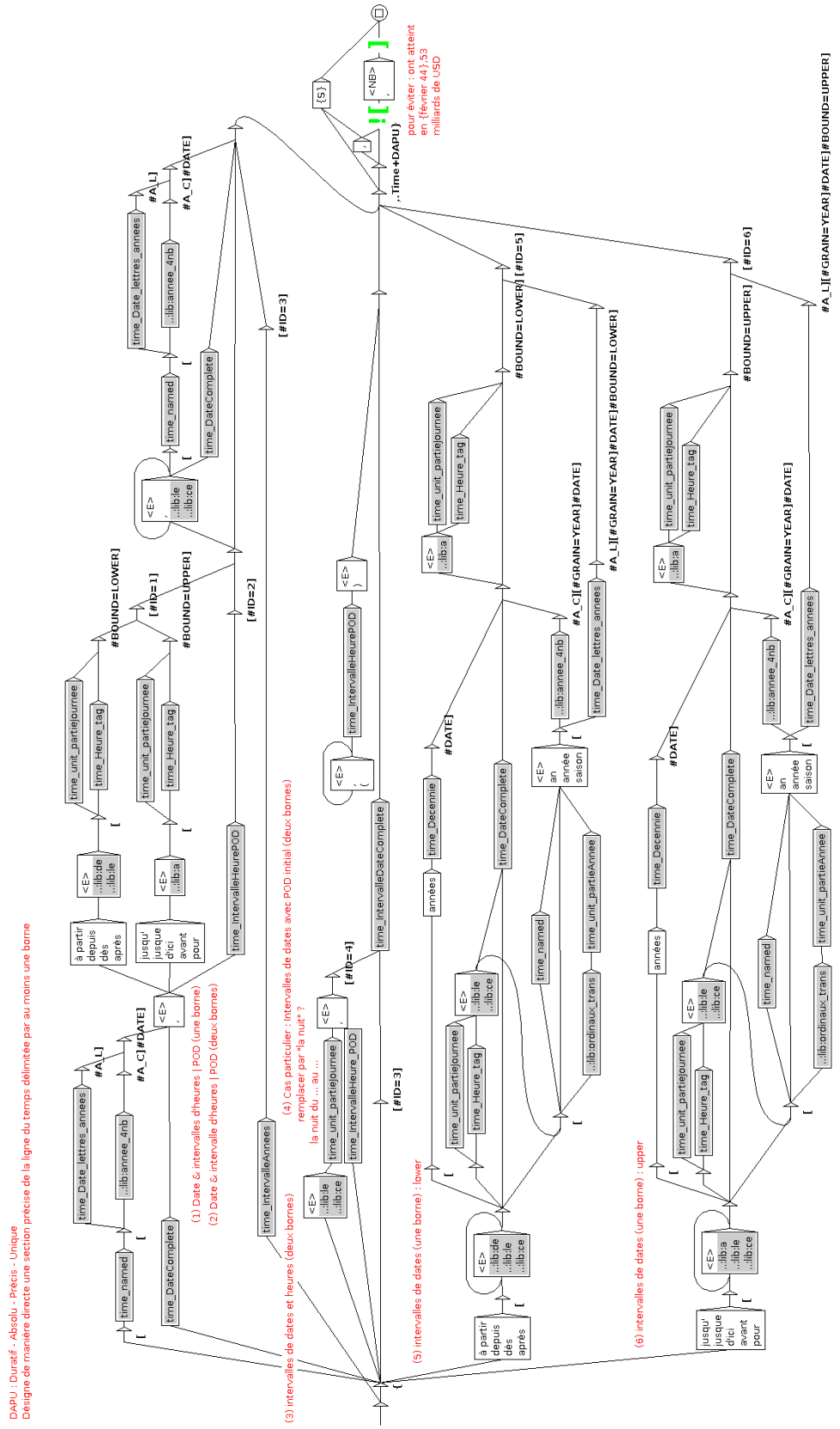

"seconde" : ambiguité entre l'unité et "deuxième"
EX : la semaine d'après | prochaine
#TARGET=PART
#TEMPSHIFT
#REF=FOCUS
#REF=NOW
#ID=2
#ID=5
#ID=6
#ID=7
#ID=8
time_unit_partiejournee
time_Heure_tag
time_unit
time_mod_decal_det
time_ref_now
time_ref_focus
time_target_prpu2
time_focterm_prpu
time_focterm_prpu2
journée
seconde
(2)
(5)
(6)
(7)
(8)

C.5 PRFU

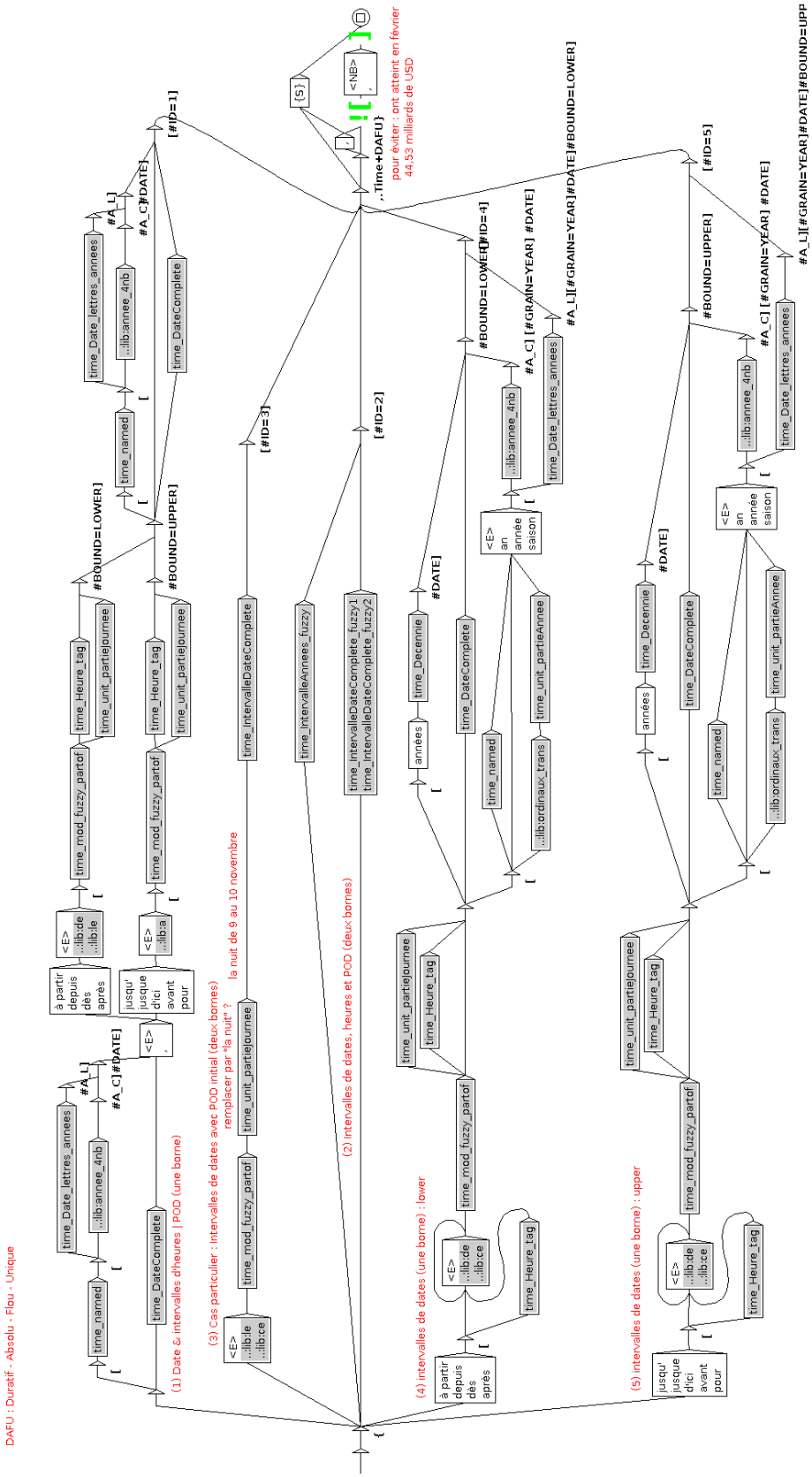

PRFU : Ponctuel - Relatif - Flou - Unique
Désigne de manière relative (au moment de l'énonciation ou un autre point de repère) un point imprécis sur la ligne du temps
#TEMPSHIFT
#REF=FOCUS
#REF=NOW
#TARGET=PART
#PARTOF=END
#ID=1->2
#ID=3-4
#ID=5
#ID=6
grain\<= day sans FUZZY est traité dansPRPU
EX : il y a quelques jours
EX : il y a (environ) trois semaines et demi
Ex: (environ) trois semaines plus tard
un mois après Pâques
une semaine plus tôt
GRAIN
EX : la semaine d'après | prochaine
"seconde" : ambiguité entre l'unité et "deuxième"
EX : cette année (à un moment donné durant cette année)
dans le courant de...
éviter: prendre {fin à 17h30}
time_mod_environ
time_mod_introrel
time_mod_fuzzy_partof
time_mod_decal2
time_mod_decal_det
time_unit_partiejournee
time_Heure_tag
time_unit
time_nbfl
time_nbc_nbl
time_unit_letd
time_unit_gtd
time_focternp_papu
time_focternp_prpu
journée
seconde
ce cet cette
de la de l' du de
en durant pendant
à
et
fin
, .Time+PRFU}


[#ID=7]
[#REF=NOW][#TEMPSHIFT]
[ID=8]
[#ID=9-10]
[#ID=11-12]
[[#REF=FOCUS][#TEMPSHIFT]
#TEMPSHIFT
#GRAIN=YEAR
#TARGET=PART
#PARTOF=END
#PARTOF=BEGIN
time_mod_decal
time_target_prpu4
time_unit_gtd
time_unit_partiejournee
time_mod_fuzzy_partof
time_target_prpu3
time_Heure_tag
time_ref_now
time_ref_focus
time_partof
time_fuzzy
time_unit_lib:de
time_unit_lib:ce
time_unit_lib:a
time_unit_lib:le
time_unit_lib:ce
fin
en
durant
pendant
actuel
aujourd'hui
maintenant
même
à généraliser avec autres unités temporelles ?
EX : fin du mois de janvier de l'année dernière
éviter: prendre {fin le ... }
(7) inverse
éviter: prendre {fin à 17h30}
??? court circuite POD !!!!
Ex: vers 20h30
EX : dans l'après-midi
(7)
(8)
(9) à (10) : partie 1 : DATE [HEURE]
(9) à (10) : partie 2 : HEURE DATE
(11) et (12) : partie 1 : TARGET (FUZZY HEURE)
(11) et (12) : partie 2 : (FUZZY HEURE) TARGET

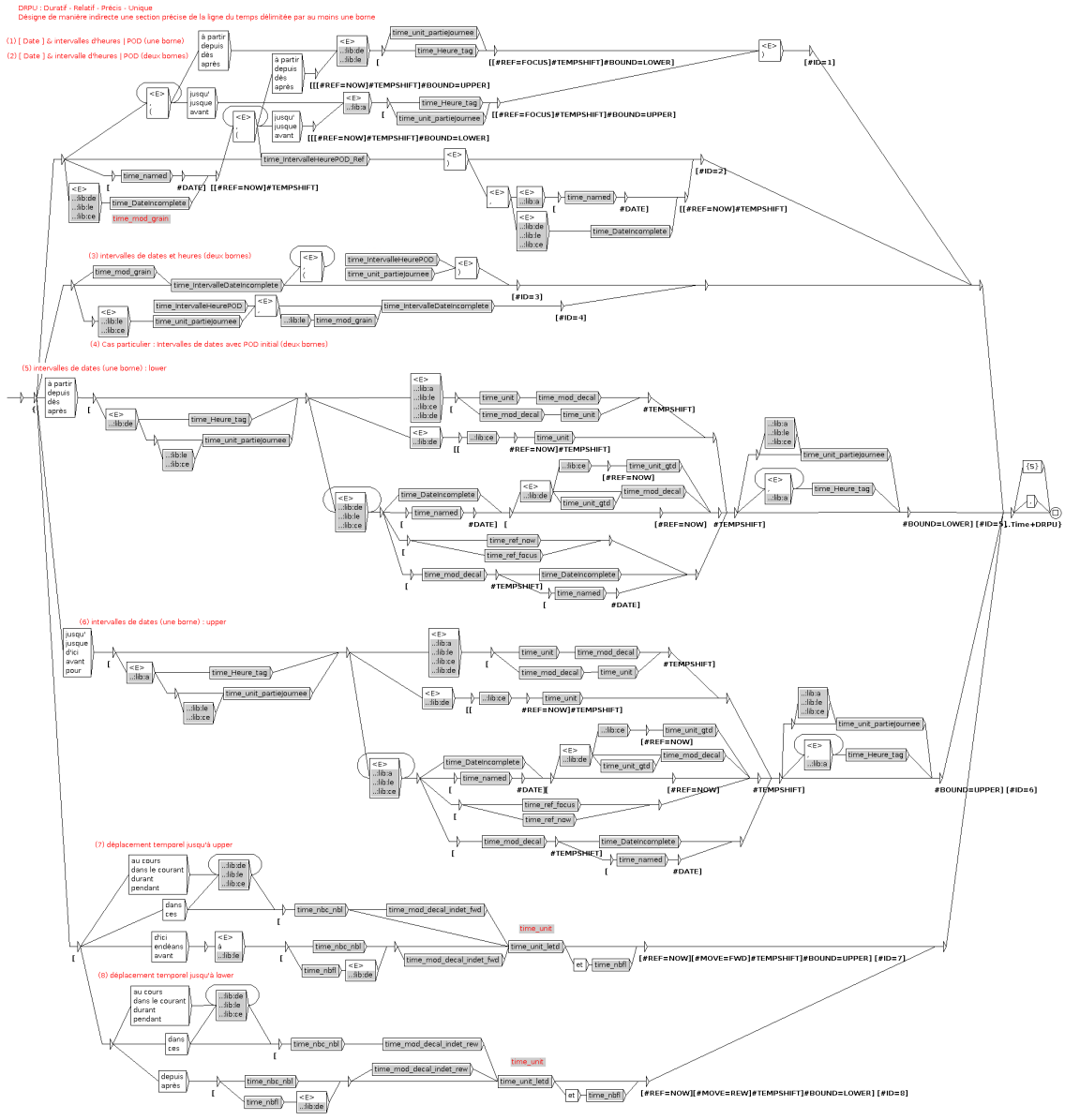
C.6 DAPU


DAPU : Duratif - Absolu - Précis - Unique
Désigne de manière directe une section precise de la ligne du temps délimitée par au moins une borne
(1) Date & intervalles d'heures | POD (une borne)
(2) Date & intervalle d'heures | POD (deux bornes)
(3) intervalles de dates et heures (deux bornes)
(4) Cas particulier : intervalles de dates avec POD initial (deux bornes)
remplacer par "la nuit" ?
la nuit du ... au ...
(5) intervalles de dates (une borne) : lower
(6) intervalles de dates (une borne) : upper
..Time+DAPU
pour éviter : ont atteint
en :février 44}:53
milliards de USD
#A_L
#A_C[#DATE]
#BOUND=LOWER
#BOUND=UPPER
#A_C[#GRAIN=YEAR]#DATE]
#A_L][#GRAIN=YEAR]#DATE][#BOUND=LOWER]
#A_L][#GRAIN=YEAR]#DATE][#BOUND=UPPER]
[#ID=1]
[#ID=2]
[#ID=3]
[#ID=4]
[#ID=5]
[#ID=6]
time_Date_lettres_annees
time_named
time_annee_4nb
time_DateComplete
time_intervalleAnnees
time_intervalleHeurePOD
time_intervalleHeure_POD
time_intervalleDateComplete
time_unit_partieJournee
time_Heure_tag
time_Decennie
time_unit_partieAnnee
time_ordinaux_trans
à partir depuis dès après
jusqu' jusque d'ici avant pour
an année saison
années

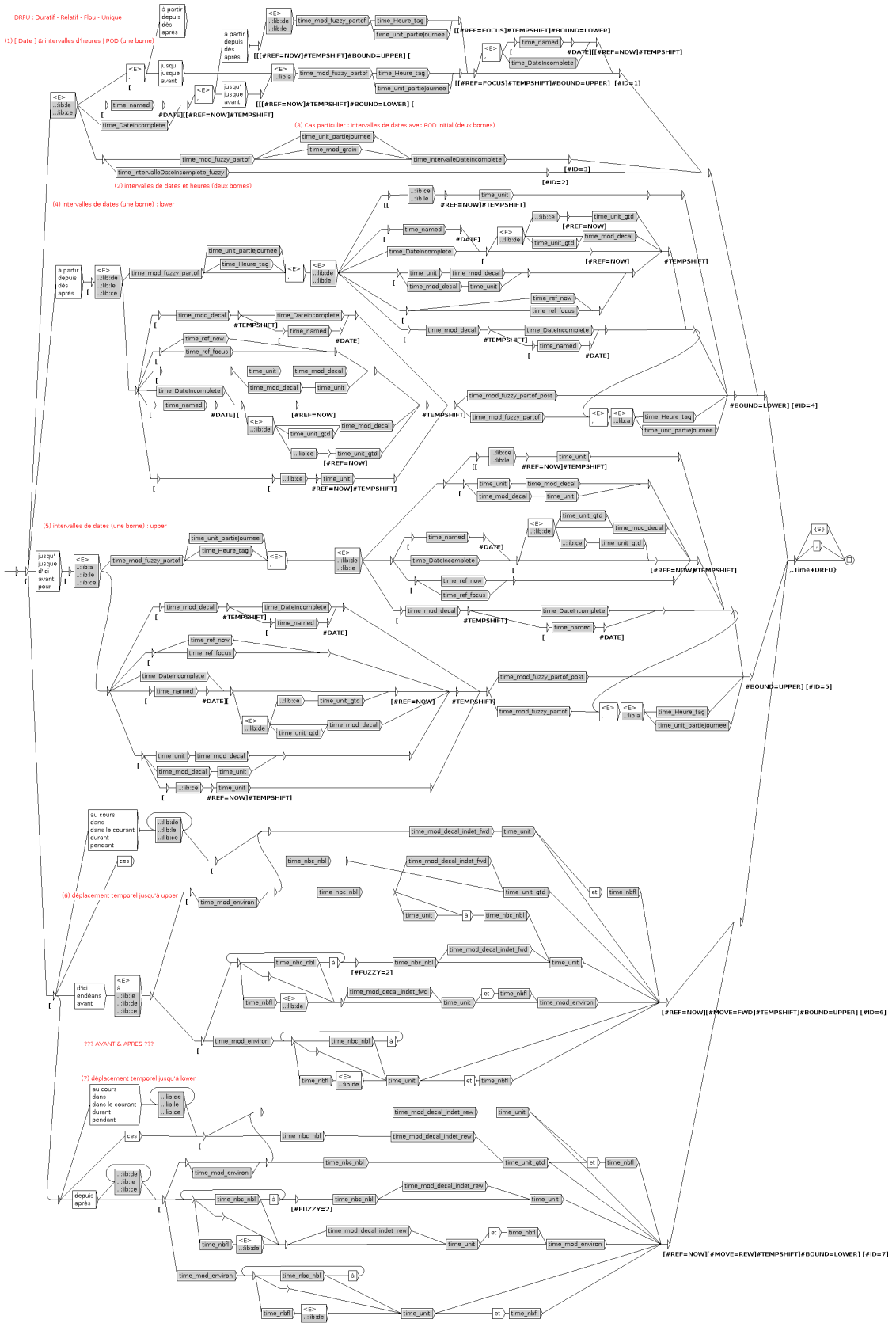
C.7 DAFU


DAFU : Duratif - Absolu - Flou - Unique
(1) Date & intervalles d'heures | POD (une borne)
(2) Intervalles de dates, heures et POD (deux bornes)
(3) Cas particulier : intervalles de dates avec POD initial (deux bornes)
remplacer par "la nuit" ?
la nuit de 9 au 10 novembre
(4) intervalles de dates (une borne) : lower
(5) intervalles de dates (une borne) : upper
pour éviter : ont atteint en février
44,53 milliards de USD
time_named
time_Date_lettres_annees
time_DateComplete
time_Heure_tag
time_mod_fuzzy_partof
time_unit_partiejournee
time_intervalleDateComplete
time_intervalleAnnees_fuzzy
time_intervalleDateComplete_fuzzy1
time_intervalleDateComplete_fuzzy2
time_Decennie
time_unit_partieAnnee
time_unit_partieordinaux_trans
#A_L
#A_C
#DATE
#BOUND=LOWER
#BOUND=UPPER
#BOUND=LOWER[#ID=4]
#A_C] [#GRAIN=YEAR] #DATE]
#A_L][#GRAIN=YEAR][#DATE][#BOUND=LOWER]
#A_L][#GRAIN=YEAR][#DATE][#BOUND=UPP
[#ID=1]
[#ID=2]
[#ID=3]
[#ID=5]
,,Time+DAFU}
à partir depuis dès après
jusqu' jusque d'ici avant pour
an année saison

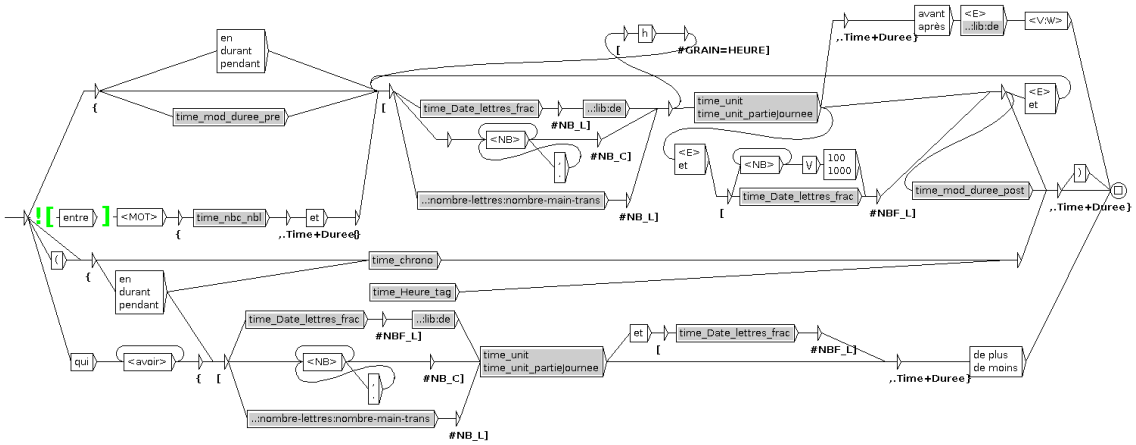
C.8 DRPU


DRPU : Duratif - Relatif - Précis - Unique
Désigne de manière indirecte une section précise de la ligne du temps délimitée par au moins une borne
(1) [Date] & intervalles d'heures | POD (une borne)
(2) [Date] & intervalle d'heures | POD (deux bornes)
(3) intervalles de dates et heures (deux bornes)
(4) Cas particulier : intervalles de dates avec POD initial (deux bornes)
(5) intervalles de dates (une borne) : lower
(6) intervalles de dates (une borne) : upper
(7) déplacement temporel jusqu'à upper
(8) déplacement temporel jusqu'à lower
time_unit_partiejournee
time_Heure_tag
time_named
time_DateIncomplete
time_mod_grain
time_intervalleHeurePOD_Ref
time_intervalleDateIncomplete
time_unit
time_mod_decal
time_unit_qtd
time_ref_now
time_ref_focus
time_nbc_nbl
time_mod_decal_indet_fwd
time_mod_decal_indet_rew
time_unit_letd
time_nbfl
#TEMPSHIFT
#DATE
#BOUND=LOWER] [#ID=5].Time+DRPU}
#BOUND=UPPER] [#ID=6]
[#REF=NOW][#MOVE=FWD]#TEMPSHIFT]#BOUND=UPPER] [#ID=7]
[#REF=NOW][#MOVE=REW]#TEMPSHIFT]#BOUND=LOWER] [#ID=8]

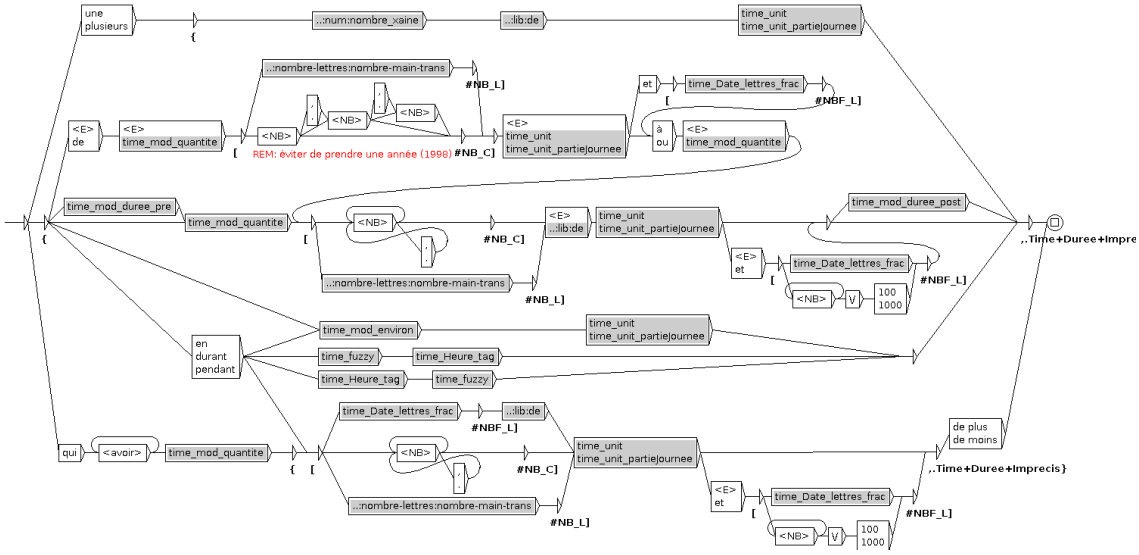
C.9 DRFU


DRFU : Duratif - Relatif - Flou - Unique
(1) [Date] & intervalles d'heures | POD (une borne)
(2) intervalles de dates et heures (deux bornes)
(3) Cas particulier : intervalles de dates avec POD initial (deux bornes)
(4) intervalles de dates (une borne) : lower
(5) intervalles de dates (une borne) : upper
(6) déplacement temporel jusqu'à upper
(7) déplacement temporel jusqu'à lower
??? AVANT & APRES ???
..Time+DRFU}
[#REF=NOW][#MOVE=FWD]#TEMPSHIFT]#BOUND=UPPER] [#ID=6]
[#REF=NOW][#MOVE=REW]#TEMPSHIFT]#BOUND=LOWER] [#ID=7]

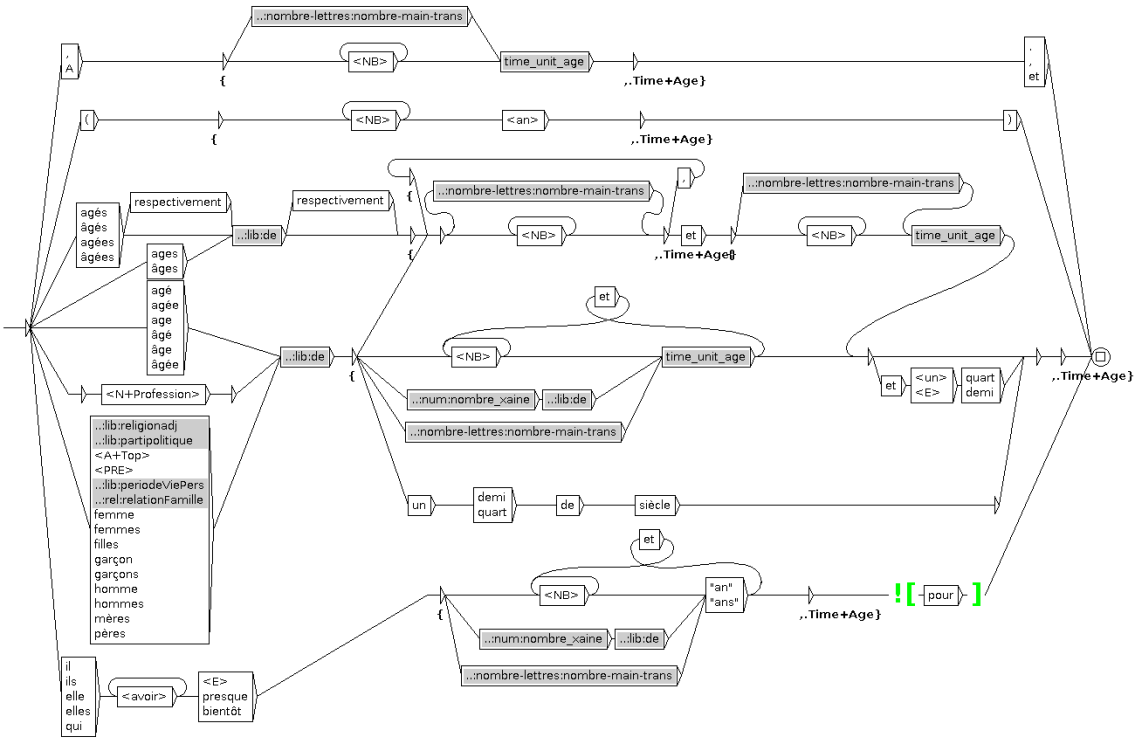
C.10 Durée



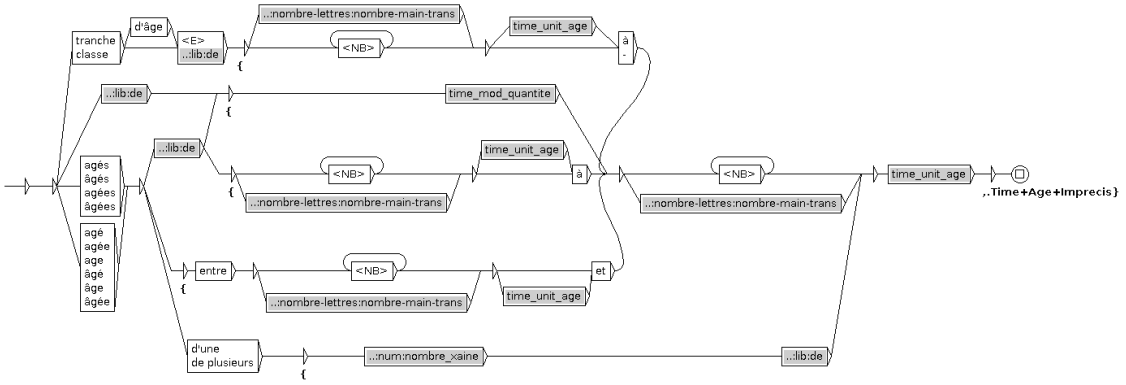
C.11 Durée imprécise



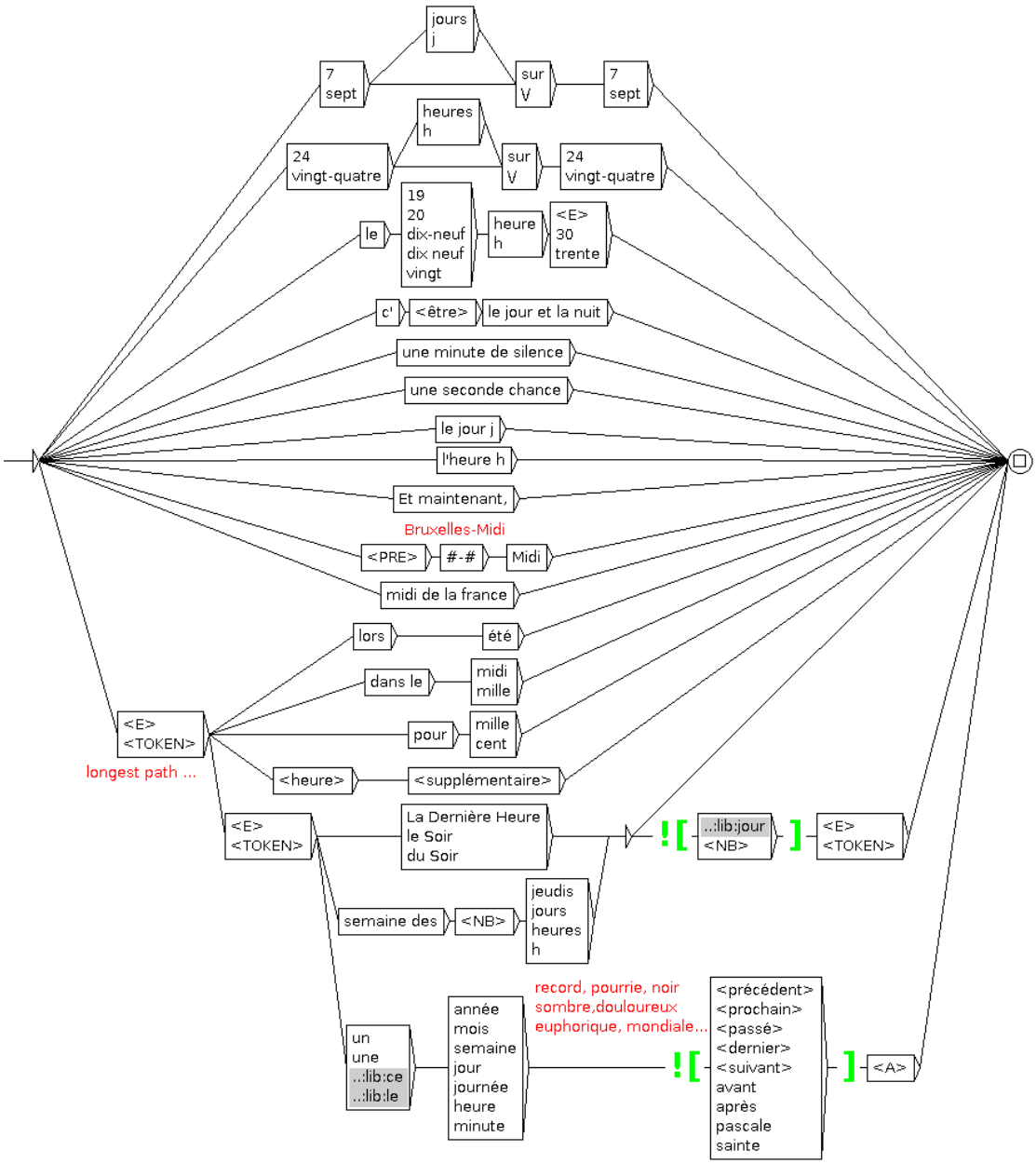
C.12 Age



C.13 Age imprécis



C.14 Graphe d'exclusion



ANNEXE D

STATISTIQUES DÉTAILLÉES : DISTRIBUTION DES EXPRESSIONS TEMPORELLES

D.1 Introduction

Les statistiques fournies dans cette annexe reprennent la distribution des expressions temporelles telles que reconnues par les transducteurs. Il n'y a donc pas d'étape de validation. Elles incluent donc un taux d'erreur comparable à celui constaté lors de l'évaluation (Section 7.9).

Les corpus proposés sont ceux déjà utilisés dans cette thèse, c'est-à-dire les corpus *News* et *Parlementaire* (Section 7.9.1). Ceux-ci étant relativement limités en nombre de textes, les statistiques ont également été produites pour six mois du journal Le Soir (avril 1998 à septembre 1998).

Pour chaque corpus, les statistiques de distribution sont proposées d'une manière résumée (c'est-à-dire distinguées au niveau de la catégorie générale), et sont présentées par fréquence décroissante. Les statistiques détaillées reprennent l'ensemble des catégories à tous les niveaux, et sont classées selon l'ordre de présentation des graphes (voir Annexe C). Pour le copus du journal Le Soir, pour des raisons de place, nous ne rapportons que les sous-catégories, sans mentionner l'ensemble des signatures rencontrées.

Une rapide analyse et comparaison des résultats obtenus sur chaque corpus (Figure D.1) montre une certaine régularité. Les catégories PAPU et PRPU apparaissent en effet toujours dans les trois premières places. D'une manière générale PRPU est la plus fréquente.

Le corpus *Parlementaire* est quelque peu différent à cet égard, en raison de sa composition particulière (entre autres des textes de loi dans lesquels il est

fréquent de voir apparaître une date complète, par exemple « loi du jj.mm.aaaa sur tel sujet »).

Entre le corpus *News* et le corpus plus étendu du journal Le Soir, les différences ne sont pas énormes, surtout pour les cinq premières places. Celles-ci sont attribuée au même sous-ensemble de catégories, et l'ordre de celles-ci est identique à deux inversions près (entre les catégories 2 et 3 d'une part, 4 et 5 d'autre part).

D'une manière générale, la distribution observée pour le journal Le Soir est mieux répartie, ce qui peut s'expliquer par la plus grande quantité de textes pris en compte.

Enfin, en ce qui concerne la plus faible proportion observée pour la catégorie PAPU dans le corpus *News*, il faut tenir compte du fait que les dates d'émission des articles (toujours de type PAPU) ont été enlevées du corps du texte pour être placées en tant que *métadonnée* dans l'en-tête du texte. Ces dates constituent une différence de 365 expressions PAPU qui auraient pu être identifiées en plus de 179 actuelles. Dans le corpus Le Soir, la date d'émission des textes n'est pas facilement identifiable et, lorsqu'elle existe, a donc été conservée dans le corps du texte.

	Copus *News*	Corpus *Parlementaire*	Journal Le Soir (6 mois)
1.	Time+PRPU : 1.256 (47.56%)	Time+PAPU : 1.922 (53.24%)	Time+PRPU : 145.750 (35.49%)
2.	Time+Duree : 491 (18.59%)	Time+PAFU : 343 (9.5%)	Time+PAPU : 71.169 (17.33%)
3.	Time+PAPU : 179 (6.78%)	Time+PRPU : 311 (8.61%)	Time+Duree : 47.076 (11.46%)
4.	Time+PAFU : 153 (5.79%)	Time+Duree : 306 (8.48%)	Time+PRFU : 35.011 (8.52%)
5.	Time+PRFU : 149 (5.64%)	Time+DAPU : 247 (6.84%)	Time+PAFU : 31.081 (7.57%)
6.	Time+Age : 108 (4.09%)	Time+DRFU : 125 (3.46%)	Time+DRPU : 23.382 (5.69%)
7.	Time+DRFU : 87 (3.29%)	Time+Duree+Imprecis : 106 (2.94%)	Time+DRFU : 18.475 (4.5%)
8.	Time+DAPU : 78 (2.95%)	Time+DRPU : 80 (2.22%)	Time+Duree+Imprecis : 13.111 (3.19%)
9.	Time+DRPU : 67 (2.54%)	Time+PRFU : 76 (2.11%)	Time+DAPU : 11.684 (2.84%)
10.	Time+Duree+Imprecis : 55 (2.08%)	Time+Age+Imprecis : 35 (0.97%)	Time+Age : 10.584 (2.58%)
11.	Time+DAFU : 12 (0.45%)	Time+Age : 34 (0.94%)	Time+DAFU : 2.010 (0.49%)
12.	Time+Age+Imprecis : 6 (0.23%)	Time+DAFU : 25 (0.69%)	Time+Age+Imprecis : 1.362 (0.33%)
	Total = 2.641 (100%)	Total = 3.610 (100%)	Total = 410.695 (100%)

Tableau D.1 : Synthèse de la distribution des expressions temporelles selon le corpus.

D.2 Corpus News

La distribution détaillée est reprise ci-dessous.

```
Time+PAPU : 179 (6.78%)

  Time+PAPU_1 : 70 (2.65%)
    Time+PAPU~DATE~ID.1 : 70 (2.65%)

  Time+PAPU_2 : 1 (0.04%)
    Time+PAPU~DATE~ID.2 : 1 (0.04%)

  Time+PAPU_4-8 : 99 (3.75%)
    Time+PAPU~DATE~HEURE~ID.4-8 : 1 (0.04%)
    Time+PAPU~DATE~ID.4-8 : 98 (3.71%)

  Time+PAPU_9 : 9 (0.34%)
    Time+PAPU~DATE-SEASON~ID.9 : 9 (0.34%)


Time+PAFU : 153 (5.79%)

  Time+PAFU_9-11 : 153 (5.79%)
    Time+PAFU~FUZZY.1~DATE~ID.9-11 : 135 (5.11%)
    Time+PAFU~FUZZY_PARTOF~DATE~ID.9-11 : 5 (0.19%)
    Time+PAFU~PARTOF.BEG~DATE~ID.9-11 : 6 (0.23%)
    Time+PAFU~PARTOF.END~DATE~ID.9-11 : 6 (0.23%)
    Time+PAFU~PARTOF.MID~DATE~ID.9-11 : 1 (0.04%)


Time+PRPU : 1256 (47.56%)

  Time+PRPU_1 : 1071 (40.55%)
    Time+PRPU~TARGET.PART-DATE-NAMED~TEMPSHIFT-REF.NOW~ID.1 : 8 (0.3%)
    Time+PRPU~TARGET.PART-DATE~TEMPSHIFT-NB.1-REF.NOW-MOVE.FWD~ID.1 : 3 (0.11%)
    Time+PRPU~TARGET.PART-DATE~TEMPSHIFT-NB.1-REF.NOW-MOVE.REW~ID.1 : 17 (0.64%)
    Time+PRPU~TARGET.PART-DATE~TEMPSHIFT-REF.NOW~ID.1 : 943 (35.71%)
```

```
  Time+PRPU~TARGET.PART-DATE~TEMPSHIFT-REF.NOW~TARGET.PART-GRAIN.POD~ID.1 : 78 (2.95%)
  Time+PRPU~TARGET.PART-DATE~TEMPSHIFT-REF.NOW~TARGET.PART-GRAIN.POD~TARGET.PART-HEURE~ID.1 : 1 (0.04%)
  Time+PRPU~TARGET.PART-DATE~TEMPSHIFT-REF.NOW~TARGET.PART-HEURE~ID.1 : 18 (0.68%)
  Time+PRPU~TARGET.PART-GRAIN.POD~TARGET.PART-DATE-NAMED~TEMPSHIFT-REF.NOW~ID.1 : 2 (0.08%)
  Time+PRPU~TARGET.PART-GRAIN.POD~TARGET.PART-DATE~TEMPSHIFT-REF.NOW~ID.1 : 1 (0.04%)

Time+PRPU_2 : 57 (2.16%)
  Time+PRPU~TARGET.PART-GRAIN.POD~TEMPSHIFT-REF.FOCUS~ID.2 : 6 (0.23%)
  Time+PRPU~TARGET.PART-HEURE~TEMPSHIFT-REF.FOCUS~ID.2 : 51 (1.93%)

Time+PRPU_4 : 5 (0.19%)
  Time+PRPU~TEMPSHIFT-NBF_L-GRAIN.HOUR-REF.FOCUS-MOVE.FWD~ID.4 : 1 (0.04%)
  Time+PRPU~TEMPSHIFT-NB_L-GRAIN.DAY-REF.FOCUS-MOVE.FWD~ID.4 : 1 (0.04%)
  Time+PRPU~TEMPSHIFT-NB_L-GRAIN.DAY-REF.FOCUS-MOVE.REW~ID.4 : 1 (0.04%)
  Time+PRPU~TEMPSHIFT-NB_L-GRAIN.HOUR-REF.FOCUS-MOVE.FWD~ID.4 : 2 (0.08%)

Time+PRPU_5 : 71 (2.69%)
  Time+PRPU~TEMPSHIFT-REF.NOW-GRAIN.DAY-MOVE.FWD-NB.1~ID.5 : 4 (0.15%)
  Time+PRPU~TEMPSHIFT-REF.NOW-GRAIN.DAY-MOVE.NO-NB.0~ID.5 : 15 (0.57%)
  Time+PRPU~TEMPSHIFT-REF.NOW-GRAIN.DAY-MOVE.REW-NB.1~ID.5 : 2 (0.08%)
  Time+PRPU~TEMPSHIFT-REF.NOW-GRAIN.DAY-MOVE.REW-NB.1~ID.5~TARGET.PART-DATE : 1 (0.04%)
  Time+PRPU~TEMPSHIFT-REF.NOW-GRAIN.UNDEF-MOVE.NO-NB.0~ID.5 : 49 (1.86%)

Time+PRPU_6 : 7 (0.27%)
  Time+PRPU~TEMPSHIFT-REF.FOCUS-GRAIN.DAY-MOVE.FWD-NB.1~ID.6 : 1 (0.04%)
  Time+PRPU~TEMPSHIFT-REF.FOCUS-GRAIN.DAY-MOVE.REW-NB.1~ID.6 : 6 (0.23%)

Time+PRPU_7 : 27 (1.02%)
  Time+PRPU~TEMPSHIFT-GRAIN.DAY-NB.1-REF.FOCUS-MOVE.FWD~ID.7 : 1 (0.04%)
  Time+PRPU~TEMPSHIFT-GRAIN.POD-NB.1-REF.FOCUS-MOVE.FWD~ID.7 : 1 (0.04%)
  Time+PRPU~TEMPSHIFT-GRAIN.WEEK-NB.1-REF.NOW-MOVE.REW~ID.7 : 6 (0.23%)
  Time+PRPU~TEMPSHIFT-GRAIN.YEAR-NB.1-REF.FOCUS-MOVE.REW~ID.7 : 1 (0.04%)
  Time+PRPU~TEMPSHIFT-GRAIN.YEAR-NB.1-REF.NOW-MOVE.FWD~ID.7 : 4 (0.15%)
  Time+PRPU~TEMPSHIFT-GRAIN.YEAR-NB.1-REF.NOW-MOVE.REW~ID.7 : 13 (0.49%)
  Time+PRPU~TEMPSHIFT-NB.1-REF.NOW-MOVE.REW-GRAIN.MIN~ID.7 : 1 (0.04%)

Time+PRPU_8 : 18 (0.68%)
  Time+PRPU~TARGET.PART-GRAIN.DAY~TEMPSHIFT-REF.FOCUS~ID.8 : 1 (0.04%)
```

```
    Time+PRPU~TARGET.PART-GRAIN.DAY~TEMPSHIFT-REF.NOW~ID.8 : 2 (0.08%)
    Time+PRPU~TARGET.PART-GRAIN.MONTH~TEMPSHIFT-REF.NOW~ID.8 : 1 (0.04%)
    Time+PRPU~TARGET.PART-GRAIN.WEEK~TEMPSHIFT-REF.NOW~ID.8 : 4 (0.15%)
    Time+PRPU~TARGET.PART-GRAIN.WE~TEMPSHIFT-REF.NOW~ID.8 : 4 (0.15%)
    Time+PRPU~TARGET.PART-GRAIN.YEAR~TEMPSHIFT-REF.NOW~ID.8 : 6 (0.23%)


Time+PRFU : 149 (5.64%)

  Time+PRFU_1-2 : 8 (0.3%)
    Time+PRFU~TEMPSHIFT-REF.NOW-MOVE.FWD-FUZZY.2-GRAIN.WEEK~ID.1-2 : 1 (0.04%)
    Time+PRFU~TEMPSHIFT-REF.NOW-MOVE.REW-FUZZY.2-GRAIN.MONTH~ID.1-2 : 1 (0.04%)
    Time+PRFU~TEMPSHIFT-REF.NOW-MOVE.REW-FUZZY.2-GRAIN.YEAR~ID.1-2 : 1 (0.04%)
    Time+PRFU~TEMPSHIFT-REF.NOW-MOVE.REW-NB_L-GRAIN.WEEK~ID.1-2 : 1 (0.04%)
    Time+PRFU~TEMPSHIFT-REF.NOW-MOVE.REW-NB_L-GRAIN.YEAR~ID.1-2 : 4 (0.15%)

  Time+PRFU_3-4 : 19 (0.72%)
    Time+PRFU~TEMPSHIFT-FUZZY.2-GRAIN.DAY-MOVE.FWD-REF.FOCUS~ID.3-4 : 2 (0.08%)
    Time+PRFU~TEMPSHIFT-FUZZY.2-GRAIN.DAY-MOVE.REW-REF.FOCUS~ID.3-4 : 1 (0.04%)
    Time+PRFU~TEMPSHIFT-FUZZY.2-GRAIN.HOUR-MOVE.FWD-REF.FOCUS~ID.3-4 : 2 (0.08%)
    Time+PRFU~TEMPSHIFT-FUZZY.2-GRAIN.MIN-MOVE.FWD-REF.FOCUS~ID.3-4 : 1 (0.04%)
    Time+PRFU~TEMPSHIFT-FUZZY.2-GRAIN.MONTH-MOVE.FWD-REF.FOCUS~ID.3-4 : 1 (0.04%)
    Time+PRFU~TEMPSHIFT-FUZZY.2-GRAIN.YEAR-MOVE.FWD-REF.FOCUS~ID.3-4 : 1 (0.04%)
    Time+PRFU~TEMPSHIFT-FUZZY.2-NB_L-GRAIN.DAY-MOVE.REW-REF.FOCUS~ID.3-4 : 1 (0.04%)
    Time+PRFU~TEMPSHIFT-FUZZY.2-NB_L-GRAIN.MONTH-MOVE.FWD-REF.FOCUS~ID.3-4 : 2 (0.08%)
    Time+PRFU~TEMPSHIFT-FUZZY.2-NB_L-GRAIN.WEEK-MOVE.REW-REF.FOCUS~ID.3-4 : 1 (0.04%)
    Time+PRFU~TEMPSHIFT-NB_C-GRAIN.MONTH-MOVE.FWD-REF.FOCUS~ID.3-4 : 1 (0.04%)
    Time+PRFU~TEMPSHIFT-NB_L-GRAIN.MONTH-MOVE.FWD-REF.FOCUS~ID.3-4 : 1 (0.04%)
    Time+PRFU~TEMPSHIFT-NB_L-GRAIN.WEEK-MOVE.FWD-REF.FOCUS~ID.3-4 : 1 (0.04%)
    Time+PRFU~TEMPSHIFT-NB_L-GRAIN.WEEK-MOVE.REW-REF.FOCUS~FOCTEMP.PART-DATE-NAMED~ID.3-4 : 1 (0.04%)
    Time+PRFU~TEMPSHIFT-NB_L-GRAIN.YEAR-MOVE.FWD-REF.FOCUS~ID.3-4 : 2 (0.08%)
    Time+PRFU~TEMPSHIFT-NB_L-GRAIN.YEAR-MOVE.REW-REF.FOCUS~ID.3-4 : 1 (0.04%)

  Time+PRFU_5 : 4 (0.15%)
    Time+PRFU~FUZZY.1~PARTOF.END~TEMPSHIFT-GRAIN.WEEK-NB.1-REF.NOW-MOVE.REW~ID.5 : 1 (0.04%)
    Time+PRFU~PARTOF.BEG~TEMPSHIFT-GRAIN.CENT-NB.1-REF.NOW-MOVE.REW~ID.5 : 1 (0.04%)
    Time+PRFU~PARTOF.BEG~TEMPSHIFT-GRAIN.WEEK-NB.1-REF.NOW-MOVE.FWD~ID.5 : 1 (0.04%)
    Time+PRFU~PARTOF.END~TEMPSHIFT-GRAIN.YEAR-NB.1-REF.NOW-MOVE.REW~ID.5 : 1 (0.04%)
```

```
Time+PRFU_6 : 31 (1.17%)
  Time+PRFU~FUZZY.1~PARTOF.BEG~TARGET.PART-GRAIN.POD~TEMPSHIFT-REF.NOW~ID.6 : 5 (0.19%)
  Time+PRFU~FUZZY.1~PARTOF.BEG~TARGET.PART-GRAIN.WEEK~TEMPSHIFT-REF.NOW~ID.6 : 1 (0.04%)
  Time+PRFU~FUZZY.1~PARTOF.END~TARGET.PART-GRAIN.POD~TEMPSHIFT-REF.NOW~ID.6 : 1 (0.04%)
  Time+PRFU~FUZZY.1~PARTOF.END~TARGET.PART-GRAIN.WEEK~TEMPSHIFT-REF.NOW~ID.6 : 1 (0.04%)
  Time+PRFU~FUZZY.1~PARTOF.END~TARGET.PART-GRAIN.YEAR~TEMPSHIFT-REF.NOW~ID.6 : 1 (0.04%)
  Time+PRFU~FUZZY.1~TARGET.PART-GRAIN.MONTH~TEMPSHIFT-REF.NOW~ID.6 : 1 (0.04%)
  Time+PRFU~FUZZY.1~TARGET.PART-GRAIN.POD~TEMPSHIFT-REF.NOW~ID.6 : 13 (0.49%)
  Time+PRFU~FUZZY.2~TARGET.PART-GRAIN.POD~TEMPSHIFT-REF.NOW~ID.6 : 1 (0.04%)
  Time+PRFU~PARTOF.BEG~TARGET.PART-GRAIN.WEEK~TEMPSHIFT-REF.NOW~ID.6 : 1 (0.04%)
  Time+PRFU~PARTOF.BEG~TARGET.PART-GRAIN.YEAR~TEMPSHIFT-REF.NOW~ID.6 : 2 (0.08%)
  Time+PRFU~PARTOF.END~TARGET.PART-GRAIN.MONTH~TEMPSHIFT-REF.NOW~ID.6 : 1 (0.04%)
  Time+PRFU~PARTOF.END~TARGET.PART-GRAIN.POD~TEMPSHIFT-REF.NOW~ID.6 : 1 (0.04%)
  Time+PRFU~PARTOF.END~TARGET.PART-GRAIN.WEEK~TEMPSHIFT-REF.NOW~ID.6 : 1 (0.04%)
  Time+PRFU~PARTOF.END~TARGET.PART-GRAIN.YEAR~TEMPSHIFT-REF.NOW~ID.6 : 1 (0.04%)

Time+PRFU_7 : 56 (2.12%)
  Time+PRFU~FUZZY.1~TARGET.PART-DATE~TEMPSHIFT-NB.1-REF.NOW-MOVE.FWD~ID.7 : 1 (0.04%)
  Time+PRFU~FUZZY.1~TARGET.PART-DATE~TEMPSHIFT-NB.1-REF.NOW-MOVE.REW~ID.7 : 8 (0.3%)
  Time+PRFU~FUZZY.1~TARGET.PART-DATE~TEMPSHIFT-REF.NOW~ID.7 : 28 (1.06%)
  Time+PRFU~PARTOF.BEG~TARGET.PART-DATE-SEASON~TEMPSHIFT-REF.NOW~ID.7 : 1 (0.04%)
  Time+PRFU~PARTOF.BEG~TARGET.PART-DATE~TARGET.PART-GRAIN.POD~TEMPSHIFT-REF.NOW~ID.7 : 3 (0.11%)
  Time+PRFU~PARTOF.BEG~TARGET.PART-DATE~TEMPSHIFT-REF.NOW~ID.7 : 6 (0.23%)
  Time+PRFU~PARTOF.END~TARGET.PART-DATE~TEMPSHIFT-NB.1-REF.NOW-MOVE.REW~ID.7 : 1 (0.04%)
  Time+PRFU~PARTOF.END~TARGET.PART-DATE~TEMPSHIFT-REF.NOW~ID.7 : 6 (0.23%)
  Time+PRFU~PARTOF.MID~TARGET.PART-DATE~TEMPSHIFT-REF.NOW~ID.7 : 2 (0.08%)

Time+PRFU_9-10 : 4 (0.15%)
  Time+PRFU~FUZZY.1~PARTOF.END~TARGET.PART-GRAIN.POD~TARGET.PART-DATE~TEMPSHIFT-REF.NOW~ID.9-10 : 1 (0.04%)
  Time+PRFU~TARGET.PART-DATE~TEMPSHIFT-REF.NOW~FUZZY.1~PARTOF.BEG~TARGET.PART-GRAIN.POD~ID.9-10 : 2 (0.08%)
  Time+PRFU~TARGET.PART-DATE~TEMPSHIFT-REF.NOW~PARTOF.MID~TARGET.PART-GRAIN.POD~ID.9-10 : 1 (0.04%)

Time+PRFU_11-12 : 27 (1.02%)
  Time+PRFU~TARGET.PART-DATE~TEMPSHIFT-REF.NOW~FUZZY.2~TARGET.PART-HEURE~ID.11-12 : 10 (0.38%)
  Time+PRFU~TARGET.PART-DATE~TEMPSHIFT-REF.NOW~TARGET.PART-GRAIN.POD~FUZZY.2~TARGET.PART-HEURE~ID.11-12 : 2 (0.08%)
  Time+PRFU~TEMPSHIFT-REF.FOCUS~FUZZY.2~TARGET.PART-HEURE~ID.11-12 : 15 (0.57%)
```

```
Time+DAPU : 78 (2.95%)

  Time+DAPU_3 : 34 (1.29%)
    Time+DAPU~BOUND.LOWER~BOUND.UPPER~ID.3 : 19 (0.72%)
    Time+DAPU~ID.3~BOUND.LOWER~BOUND.UPPER : 15 (0.57%)

  Time+DAPU_5 : 35 (1.33%)
    Time+DAPU~BOUND.LOWER~ID.5 : 35 (1.33%)

  Time+DAPU_6 : 9 (0.34%)
    Time+DAPU~BOUND.UPPER~ID.6 : 9 (0.34%)


Time+DAFU : 12 (0.45%)

  Time+DAFU_2 : 2 (0.08%)
    Time+DAFU~FUZZY.1~BOUND.LOWER~BOUND.UPPER~ID.2 : 2 (0.08%)

  Time+DAFU_4 : 2 (0.08%)
    Time+DAFU~BOUND.LOWER~ID.4 : 2 (0.08%)

  Time+DAFU_5 : 8 (0.3%)
    Time+DAFU~BOUND.UPPER~ID.5 : 8 (0.3%)


Time+DRPU : 67 (2.54%)

  Time+DRPU_1 : 6 (0.23%)
    Time+DRPU~BOUND.LOWER~ID.1 : 5 (0.19%)
    Time+DRPU~BOUND.UPPER~ID.1 : 1 (0.04%)

  Time+DRPU_2 : 1 (0.04%)
    Time+DRPU~DATE~TEMPSHIFT-REF.NOW~BOUND.LOWER~BOUND.UPPER~ID.2 : 1 (0.04%)

  Time+DRPU_3 : 10 (0.38%)
    Time+DRPU~BOUND.LOWER~BOUND.UPPER~ID.3 : 10 (0.38%)
```

```
  Time+DRPU_5 : 26 (0.98%)
    Time+DRPU~BOUND.LOWER~ID.5 : 26 (0.98%)

  Time+DRPU_6 : 17 (0.64%)
    Time+DRPU~BOUND.UPPER~ID.6 : 17 (0.64%)

  Time+DRPU_7 : 1 (0.04%)
    Time+DRPU~BOUND.UPPER~ID.7 : 1 (0.04%)

  Time+DRPU_8 : 6 (0.23%)
    Time+DRPU~BOUND.LOWER~ID.8 : 6 (0.23%)


Time+DRFU : 87 (3.29%)

  Time+DRFU_3 : 9 (0.34%)
    Time+DRFU~FUZZY.1~GRAIN.POD~BOUND.LOWER~BOUND.UPPER~BOUND.LOWER~BOUND.UPPER~ID.3 : 1 (0.04%)
    Time+DRFU~FUZZY.1~GRAIN.POD~BOUND.LOWER~BOUND.UPPER~ID.3 : 8 (0.3%)

  Time+DRFU_4 : 4 (0.15%)
    Time+DRFU~BOUND.LOWER~ID.4 : 4 (0.15%)

  Time+DRFU_5 : 6 (0.23%)
    Time+DRFU~BOUND.UPPER~ID.5 : 6 (0.23%)

  Time+DRFU_6 : 11 (0.42%)
    Time+DRFU~BOUND.UPPER~ID.6 : 11 (0.42%)

  Time+DRFU_7 : 57 (2.16%)
    Time+DRFU~BOUND.LOWER~ID.7 : 57 (2.16%)


Time+Duree : 491 (18.59%)

    Time+Duree~H_C~H_M_C~GRAIN.MIN : 279 (10.56%)
    Time+Duree~H_C~H_M_C~GRAIN.MIN~H_S : 51 (1.93%)
    Time+Duree~NB_C : 1 (0.04%)
    Time+Duree~NB_C~GRAIN.DAY : 3 (0.11%)
```

```
    Time+Duree~NB_C~GRAIN.HEURE~NB_C~GRAIN.MIN~NB_C~GRAIN.SEC : 3 (0.11%)
    Time+Duree~NB_C~GRAIN.HOUR : 10 (0.38%)
    Time+Duree~NB_C~GRAIN.MIN : 10 (0.38%)
    Time+Duree~NB_C~GRAIN.MIN~NB_C~GRAIN.SEC : 10 (0.38%)
    Time+Duree~NB_C~GRAIN.MIN~NB_C~GRAIN.SEC~NBF_L : 2 (0.08%)
    Time+Duree~NB_C~GRAIN.MONTH : 6 (0.23%)
    Time+Duree~NB_C~GRAIN.SEC : 9 (0.34%)
    Time+Duree~NB_C~GRAIN.WEEK : 1 (0.04%)
    Time+Duree~NB_C~GRAIN.YEAR : 34 (1.29%)
    Time+Duree~NB_L~GRAIN.DAY : 14 (0.53%)
    Time+Duree~NB_L~GRAIN.HOUR : 3 (0.11%)
    Time+Duree~NB_L~GRAIN.MIN : 1 (0.04%)
    Time+Duree~NB_L~GRAIN.MONTH : 10 (0.38%)
    Time+Duree~NB_L~GRAIN.MONTH~NBF_L : 2 (0.08%)
    Time+Duree~NB_L~GRAIN.POD : 2 (0.08%)
    Time+Duree~NB_L~GRAIN.WEEK : 6 (0.23%)
    Time+Duree~NB_L~GRAIN.YEAR : 29 (1.1%)
    Time+Duree~NB_L~GRAIN.YEAR~NBF_L : 5 (0.19%)


Time+Duree+Imprecis : 55 (2.08%)

    Time+Duree+Imprecis~FUZZY.2~GRAIN.DAY : 6 (0.23%)
    Time+Duree+Imprecis~FUZZY.2~GRAIN.DEC : 1 (0.04%)
    Time+Duree+Imprecis~FUZZY.2~GRAIN.HOUR : 3 (0.11%)
    Time+Duree+Imprecis~FUZZY.2~GRAIN.MIN : 2 (0.08%)
    Time+Duree+Imprecis~FUZZY.2~GRAIN.MONTH : 3 (0.11%)
    Time+Duree+Imprecis~FUZZY.2~GRAIN.WEEK : 3 (0.11%)
    Time+Duree+Imprecis~FUZZY.2~GRAIN.YEAR : 3 (0.11%)
    Time+Duree+Imprecis~GRAIN.YEAR : 6 (0.23%)
    Time+Duree+Imprecis~NB_C~GRAIN.HOUR : 1 (0.04%)
    Time+Duree+Imprecis~NB_C~GRAIN.MIN : 2 (0.08%)
    Time+Duree+Imprecis~NB_C~GRAIN.SEC : 1 (0.04%)
    Time+Duree+Imprecis~NB_C~GRAIN.YEAR : 4 (0.15%)
    Time+Duree+Imprecis~NB_C~NB_C~GRAIN.DAY : 3 (0.11%)
    Time+Duree+Imprecis~NB_C~NB_C~GRAIN.MONTH : 2 (0.08%)
    Time+Duree+Imprecis~NB_C~NB_C~GRAIN.YEAR : 1 (0.04%)
    Time+Duree+Imprecis~NB_L~GRAIN.HOUR : 1 (0.04%)
```

```
    Time+Duree+Imprecis~NB_L~GRAIN.MIN : 3 (0.11%)
    Time+Duree+Imprecis~NB_L~GRAIN.MONTH : 4 (0.15%)
    Time+Duree+Imprecis~NB_L~GRAIN.MONTH~NBF_L : 1 (0.04%)
    Time+Duree+Imprecis~NB_L~GRAIN.YEAR : 3 (0.11%)
    Time+Duree+Imprecis~NB_L~NB_L~GRAIN.DAY : 1 (0.04%)
    Time+Duree+Imprecis~NB_L~NB_L~GRAIN.YEAR : 1 (0.04%)

Time+Age : 108 (4.09%)

Time+Age+Imprecis : 6 (0.23%)
```

D.3 Corpus Parlementaire

La distribution détaillée est reprise ci-dessous.

```
Time+PAPU : 1922 (53.24%)

  Time+PAPU_1 : 364 (10.08%)
    Time+PAPU~DATE~ID.1 : 364 (10.08%)

  Time+PAPU_2 : 5 (0.14%)
    Time+PAPU~DATE~ID.2 : 5 (0.14%)

  Time+PAPU_3 : 1 (0.03%)
    Time+PAPU~DATE~ID.3 : 1 (0.03%)

  Time+PAPU_4-8 : 1535 (42.52%)
    Time+PAPU~DATE~ID.4-8 : 1535 (42.52%)

  Time+PAPU_9 : 17 (0.47%)
    Time+PAPU~DATE-POY~ID.9 : 15 (0.42%)
    Time+PAPU~DATE-SEASON~ID.9 : 2 (0.06%)


Time+PAFU : 343 (9.5%)
```

```
  Time+PAFU_9-11 : 343 (9.5%)
    Time+PAFU~FUZZY.1~DATE-POY~ID.9-11 : 5 (0.14%)
    Time+PAFU~FUZZY.1~DATE-SEASON~ID.9-11 : 2 (0.06%)
    Time+PAFU~FUZZY.1~DATE~ID.9-11 : 303 (8.39%)
    Time+PAFU~FUZZY.1~PARTOF.END~DATE~ID.9-11 : 2 (0.06%)
    Time+PAFU~FUZZY.2~PARTOF.END~DATE~ID.9-11 : 1 (0.03%)
    Time+PAFU~FUZZY.2~PARTOF.MID~DATE~ID.9-11 : 1 (0.03%)
    Time+PAFU~FUZZY_PARTOF~DATE~ID.9-11 : 13 (0.36%)
    Time+PAFU~PARTOF.BEG~DATE~ID.9-11 : 4 (0.11%)
    Time+PAFU~PARTOF.END~DATE-POY~ID.9-11 : 1 (0.03%)
    Time+PAFU~PARTOF.END~DATE~ID.9-11 : 11 (0.3%)


Time+PRPU : 311 (8.61%)

  Time+PRPU_1 : 91 (2.52%)
    Time+PRPU~TARGET.PART-DATE~TEMPSHIFT-GRAIN.YEAR-REF.NOW~ID.1 : 1 (0.03%)
    Time+PRPU~TARGET.PART-DATE~TEMPSHIFT-NB.1-REF.NOW-MOVE.FWD~ID.1 : 3 (0.08%)
    Time+PRPU~TARGET.PART-DATE~TEMPSHIFT-NB.1-REF.NOW-MOVE.REW~ID.1 : 2 (0.06%)
    Time+PRPU~TARGET.PART-DATE~TEMPSHIFT-REF.NOW~ID.1 : 85 (2.35%)

  Time+PRPU_2 : 8 (0.22%)
    Time+PRPU~TARGET.PART-GRAIN.POD~TEMPSHIFT-REF.FOCUS~ID.2 : 4 (0.11%)
    Time+PRPU~TARGET.PART-HEURE~TEMPSHIFT-REF.FOCUS~ID.2 : 4 (0.11%)

  Time+PRPU_4 : 6 (0.17%)
    Time+PRPU~TEMPSHIFT-NB_C-GRAIN.DAY-REF.FOCUS-MOVE.FWD~ID.4 : 1 (0.03%)
    Time+PRPU~TEMPSHIFT-NB_L-GRAIN.DAY-REF.FOCUS-MOVE.FWD~ID.4 : 2 (0.06%)
    Time+PRPU~TEMPSHIFT-NB_L-GRAIN.DAY-REF.FOCUS-MOVE.REW~ID.4 : 1 (0.03%)
    Time+PRPU~TEMPSHIFT-NB_L-GRAIN.HOUR-REF.FOCUS-MOVE.FWD~ID.4 : 1 (0.03%)
    Time+PRPU~TEMPSHIFT-NB_L-GRAIN.HOUR-REF.FOCUS-MOVE.REW~ID.4 : 1 (0.03%)

  Time+PRPU_5 : 170 (4.71%)
    Time+PRPU~TEMPSHIFT-REF.NOW-GRAIN.DAY-MOVE.NO-NB.0~ID.5 : 14 (0.39%)
    Time+PRPU~TEMPSHIFT-REF.NOW-GRAIN.DAY-MOVE.REW-NB.1~ID.5 : 3 (0.08%)
    Time+PRPU~TEMPSHIFT-REF.NOW-GRAIN.UNDEF-MOVE.NO-NB.0~ID.5 : 153 (4.24%)
```

```
  Time+PRPU_6 : 12 (0.33%)
    Time+PRPU~TEMPSHIFT-REF.FOCUS-GRAIN.DAY-MOVE.FWD-NB.1~ID.6 : 1 (0.03%)
    Time+PRPU~TEMPSHIFT-REF.FOCUS-GRAIN.DAY-MOVE.REW-NB.1~ID.6 : 4 (0.11%)
    Time+PRPU~TEMPSHIFT-REF.FOCUS-GRAIN.DAY-MOVE.REW-NB.1~ID.6~FOCTEMP.PART-DATE-NAMED : 7 (0.19%)

  Time+PRPU_7 : 16 (0.44%)
    Time+PRPU~TEMPSHIFT-GRAIN.DAY-NB.1-REF.FOCUS-MOVE.FWD~ID.7 : 2 (0.06%)
    Time+PRPU~TEMPSHIFT-GRAIN.YEAR-NB.1-REF.FOCUS-MOVE.FWD~ID.7 : 1 (0.03%)
    Time+PRPU~TEMPSHIFT-GRAIN.YEAR-NB.1-REF.FOCUS-MOVE.REW~ID.7 : 6 (0.17%)
    Time+PRPU~TEMPSHIFT-GRAIN.YEAR-NB.1-REF.NOW-MOVE.FWD~ID.7 : 1 (0.03%)
    Time+PRPU~TEMPSHIFT-GRAIN.YEAR-NB.1-REF.NOW-MOVE.REW~ID.7 : 6 (0.17%)

  Time+PRPU_8 : 8 (0.22%)
    Time+PRPU~TARGET.PART-GRAIN.DAY~TEMPSHIFT-REF.NOW~ID.8 : 6 (0.17%)
    Time+PRPU~TARGET.PART-GRAIN.YEAR~TEMPSHIFT-REF.NOW~ID.8 : 2 (0.06%)


Time+PRFU : 76 (2.11%)

  Time+PRFU : 1 (0.03%)
    Time+PRFU~PARTOF.END~TEMPSHIFT-REF.FOCUS-GRAIN.DAY-MOVE.REW-NB.1~ : 1 (0.03%)

  Time+PRFU_1-2 : 3 (0.08%)
    Time+PRFU~TEMPSHIFT-REF.NOW-MOVE.REW-FUZZY.2-GRAIN.WEEK~ID.1-2 : 1 (0.03%)
    Time+PRFU~TEMPSHIFT-REF.NOW-MOVE.REW-FUZZY.2-GRAIN.YEAR~ID.1-2 : 2 (0.06%)

  Time+PRFU_3-4 : 22 (0.61%)
    Time+PRFU~TEMPSHIFT-FUZZY.2-GRAIN.YEAR-MOVE.REW-REF.FOCUS~ID.3-4 : 1 (0.03%)
    Time+PRFU~TEMPSHIFT-FUZZY.2-NB_L-GRAIN.DAY-MOVE.FWD-REF.FOCUS~ID.3-4 : 1 (0.03%)
    Time+PRFU~TEMPSHIFT-FUZZY.2-NB_L-GRAIN.MONTH-MOVE.FWD-REF.FOCUS~ID.3-4 : 1 (0.03%)
    Time+PRFU~TEMPSHIFT-NB_C-GRAIN.MONTH-MOVE.FWD-REF.FOCUS~ID.3-4 : 1 (0.03%)
    Time+PRFU~TEMPSHIFT-NB_C-GRAIN.YEAR-MOVE.FWD-REF.FOCUS~ID.3-4 : 1 (0.03%)
    Time+PRFU~TEMPSHIFT-NB_L-GRAIN.MONTH-MOVE.FWD-REF.FOCUS~ID.3-4 : 7 (0.19%)
    Time+PRFU~TEMPSHIFT-NB_L-GRAIN.MONTH-MOVE.REW-REF.FOCUS~ID.3-4 : 2 (0.06%)
    Time+PRFU~TEMPSHIFT-NB_L-GRAIN.YEAR-MOVE.FWD-REF.FOCUS~ID.3-4 : 6 (0.17%)
    Time+PRFU~TEMPSHIFT-NB_L-GRAIN.YEAR-MOVE.REW-REF.FOCUS~ID.3-4 : 1 (0.03%)
    Time+PRFU~TEMPSHIFT-NB_L-GRAIN.YEAR-NBF_L-MOVE.REW-REF.FOCUS~ID.3-4 : 1 (0.03%)
```

```
  Time+PRFU_5 : 3 (0.08%)
    Time+PRFU~FUZZY.1~TEMPSHIFT-GRAIN.MONTH-NB.1-REF.FOCUS-MOVE.FWD~ID.5 : 1 (0.03%)
    Time+PRFU~FUZZY.1~TEMPSHIFT-NB.1-REF.NOW-MOVE.REW-GRAIN.DEC~ID.5 : 2 (0.06%)

  Time+PRFU_6 : 22 (0.61%)
    Time+PRFU~FUZZY.1~PARTOF.END~TARGET.PART-GRAIN.WEEK~TEMPSHIFT-REF.NOW~ID.6 : 2 (0.06%)
    Time+PRFU~FUZZY.1~TARGET.PART-GRAIN.POD~TEMPSHIFT-REF.NOW~ID.6 : 6 (0.17%)
    Time+PRFU~FUZZY.1~TARGET.PART-GRAIN.WEEK~TEMPSHIFT-REF.NOW~ID.6 : 1 (0.03%)
    Time+PRFU~FUZZY.1~TARGET.PART-GRAIN.YEAR~TEMPSHIFT-REF.NOW~ID.6 : 8 (0.22%)
    Time+PRFU~FUZZY.2~PARTOF.END~TARGET.PART-GRAIN.YEAR~TEMPSHIFT-REF.NOW~ID.6 : 1 (0.03%)
    Time+PRFU~PARTOF.BEG~TARGET.PART-GRAIN.YEAR~TEMPSHIFT-REF.NOW~ID.6 : 2 (0.06%)
    Time+PRFU~PARTOF.END~TARGET.PART-GRAIN.MONTH~TEMPSHIFT-REF.NOW~ID.6 : 1 (0.03%)
    Time+PRFU~PARTOF.END~TARGET.PART-GRAIN.YEAR~TEMPSHIFT-REF.NOW~ID.6 : 1 (0.03%)

  Time+PRFU_7 : 24 (0.66%)
    Time+PRFU~FUZZY.1~TARGET.PART-DATE-SEASON~TEMPSHIFT-REF.NOW~ID.7 : 5 (0.14%)
    Time+PRFU~FUZZY.1~TARGET.PART-DATE~TEMPSHIFT-NB.1-REF.NOW-MOVE.FWD~ID.7 : 1 (0.03%)
    Time+PRFU~FUZZY.1~TARGET.PART-DATE~TEMPSHIFT-NB.1-REF.NOW-MOVE.REW~ID.7 : 1 (0.03%)
    Time+PRFU~FUZZY.1~TARGET.PART-DATE~TEMPSHIFT-REF.NOW~ID.7 : 7 (0.19%)
    Time+PRFU~FUZZY.1~TEMPSHIFT-NB.1-REF.NOW-MOVE.REW~TARGET.PART-GRAIN.WEEK-DATE~ID.7 : 1 (0.03%)
    Time+PRFU~FUZZY.2~PARTOF.BEG~TARGET.PART-DATE-SEASON~TEMPSHIFT-REF.NOW~ID.7 : 1 (0.03%)
    Time+PRFU~FUZZY.2~PARTOF.END~TARGET.PART-DATE~TEMPSHIFT-REF.NOW~ID.7 : 1 (0.03%)
    Time+PRFU~FUZZY.2~TARGET.PART-DATE~TEMPSHIFT-REF.NOW~ID.7 : 1 (0.03%)
    Time+PRFU~PARTOF.BEG~TARGET.PART-DATE~TEMPSHIFT-GRAIN.YEAR-REF.NOW~ID.7 : 2 (0.06%)
    Time+PRFU~PARTOF.BEG~TARGET.PART-DATE~TEMPSHIFT-REF.NOW~ID.7 : 1 (0.03%)
    Time+PRFU~PARTOF.END~TARGET.PART-DATE~TEMPSHIFT-REF.NOW~ID.7 : 3 (0.08%)

  Time+PRFU_11-12 : 1 (0.03%)
    Time+PRFU~TEMPSHIFT-REF.FOCUS~PARTOF.END~TARGET.PART-HEURE~ID.11-12 : 1 (0.03%)


Time+DAPU : 247 (6.84%)

  Time+DAPU_3 : 131 (3.63%)
    Time+DAPU~BOUND.LOWER~BOUND.UPPER~ID.3 : 102 (2.83%)
    Time+DAPU~ID.3~BOUND.LOWER~BOUND.UPPER : 29 (0.8%)

  Time+DAPU_5 : 62 (1.72%)
```

```
    Time+DAPU~BOUND.LOWER~ID.5 : 62 (1.72%)

  Time+DAPU_6 : 54 (1.5%)
    Time+DAPU~BOUND.UPPER~ID.6 : 54 (1.5%)


Time+DAFU : 25 (0.69%)

  Time+DAFU_2 : 6 (0.17%)
    Time+DAFU~FUZZY.1~BOUND.LOWER~BOUND.UPPER~ID.2 : 5 (0.14%)
    Time+DAFU~PARTOF.END~BOUND.LOWER~BOUND.UPPER~ID.2 : 1 (0.03%)

  Time+DAFU_4 : 6 (0.17%)
    Time+DAFU~BOUND.LOWER~ID.4 : 6 (0.17%)

  Time+DAFU_5 : 13 (0.36%)
    Time+DAFU~BOUND.UPPER~ID.5 : 13 (0.36%)


Time+DRPU : 80 (2.22%)

  Time+DRPU_3 : 42 (1.16%)
    Time+DRPU~BOUND.LOWER~BOUND.UPPER~ID.3 : 42 (1.16%)

  Time+DRPU_5 : 15 (0.42%)
    Time+DRPU : 4 (0.11%)
    Time+DRPU~BOUND.LOWER~ID.5 : 11 (0.3%)

  Time+DRPU_6 : 9 (0.25%)
    Time+DRPU~BOUND.UPPER~ID.6 : 9 (0.25%)

  Time+DRPU_7 : 14 (0.39%)
    Time+DRPU~BOUND.UPPER~ID.7 : 14 (0.39%)


Time+DRFU : 125 (3.46%)

  Time+DRFU_3 : 1 (0.03%)
```

```
    Time+DRFU~FUZZY.1~BOUND.LOWER~BOUND.UPPER~ID.3 : 1 (0.03%)

  Time+DRFU_5 : 1 (0.03%)
    Time+DRFU~BOUND.UPPER~ID.5 : 1 (0.03%)

  Time+DRFU_6 : 31 (0.86%)
    Time+DRFU~BOUND.UPPER~ID.6 : 31 (0.86%)

  Time+DRFU_7 : 92 (2.55%)
    Time+DRFU~BOUND.LOWER~ID.7 : 92 (2.55%)


Time+Duree : 306 (8.48%)

    Time+Duree~H_C~H_M_C~GRAIN.MIN : 14 (0.39%)
    Time+Duree~H_C~H_M_C~GRAIN.MIN~H_S : 2 (0.06%)
    Time+Duree~NB_C~GRAIN.DAY : 7 (0.19%)
    Time+Duree~NB_C~GRAIN.HOUR : 35 (0.97%)
    Time+Duree~NB_C~GRAIN.HOUR~NB_C~GRAIN.MIN : 1 (0.03%)
    Time+Duree~NB_C~GRAIN.MONTH : 30 (0.83%)
    Time+Duree~NB_C~GRAIN.WEEK : 15 (0.42%)
    Time+Duree~NB_C~GRAIN.YEAR : 50 (1.39%)
    Time+Duree~NB_C~GRAIN.YEAR~NB_C~GRAIN.DAY : 2 (0.06%)
    Time+Duree~NB_L~GRAIN.DAY : 15 (0.42%)
    Time+Duree~NB_L~GRAIN.HOUR : 2 (0.06%)
    Time+Duree~NB_L~GRAIN.MONTH : 31 (0.86%)
    Time+Duree~NB_L~GRAIN.SEC : 2 (0.06%)
    Time+Duree~NB_L~GRAIN.WEEK : 11 (0.3%)
    Time+Duree~NB_L~GRAIN.YEAR : 88 (2.44%)
    Time+Duree~NB_L~POY~GRAIN.QYEAR : 1 (0.03%)


Time+Duree+Imprecis : 106 (2.94%)

    Time+Duree+Imprecis~FUZZY.2~GRAIN.DAY : 8 (0.22%)
    Time+Duree+Imprecis~FUZZY.2~GRAIN.DEC : 1 (0.03%)
    Time+Duree+Imprecis~FUZZY.2~GRAIN.HOUR : 16 (0.44%)
    Time+Duree+Imprecis~FUZZY.2~GRAIN.MONTH : 2 (0.06%)
```

```
    Time+Duree+Imprecis~FUZZY.2~GRAIN.YEAR : 31 (0.86%)
    Time+Duree+Imprecis~GRAIN.YEAR : 1 (0.03%)
    Time+Duree+Imprecis~NB_C~GRAIN.DAY : 2 (0.06%)
    Time+Duree+Imprecis~NB_C~GRAIN.DAY~NB_C~GRAIN.MONTH : 1 (0.03%)
    Time+Duree+Imprecis~NB_C~GRAIN.DAY~NB_C~GRAIN.YEAR : 1 (0.03%)
    Time+Duree+Imprecis~NB_C~GRAIN.HOUR : 7 (0.19%)
    Time+Duree+Imprecis~NB_C~GRAIN.MONTH~NB_C~GRAIN.YEAR : 1 (0.03%)
    Time+Duree+Imprecis~NB_C~GRAIN.YEAR : 3 (0.08%)
    Time+Duree+Imprecis~NB_C~NB_C~GRAIN.WEEK : 1 (0.03%)
    Time+Duree+Imprecis~NB_C~NB_C~GRAIN.YEAR : 2 (0.06%)
    Time+Duree+Imprecis~NB_L~GRAIN.DAY : 5 (0.14%)
    Time+Duree+Imprecis~NB_L~GRAIN.DAY~NB_L~GRAIN.YEAR : 2 (0.06%)
    Time+Duree+Imprecis~NB_L~GRAIN.MONTH : 2 (0.06%)
    Time+Duree+Imprecis~NB_L~GRAIN.MONTH~NB_L~GRAIN.YEAR : 4 (0.11%)
    Time+Duree+Imprecis~NB_L~GRAIN.WEEK : 1 (0.03%)
    Time+Duree+Imprecis~NB_L~GRAIN.YEAR : 8 (0.22%)
    Time+Duree+Imprecis~NB_L~NB_L~GRAIN.YEAR : 7 (0.19%)


Time+Age : 34 (0.94%)


Time+Age+Imprecis : 35 (0.97%)
```

D.4 Journal Le Soir

Le corpus est constitué des articles des journaux Le Soir, d'avril 1998 à septembre 1998 (six mois).

La distribution détaillée est reprise ci-dessous.

```
Time+PAPU : 71169 (17.33%)
  Time+PAPU_1 : 8108 (1.97%)
  Time+PAPU_2 : 849 (0.21%)
  Time+PAPU_3 : 762 (0.19%)
  Time+PAPU_4-8 : 60835 (14.81%)
```

```
  Time+PAPU_9 : 615 (0.15%)

Time+PAFU : 31081 (7.57%)
  Time+PAFU : 31 (0.01%)
  Time+PAFU_1-5 : 2 (0%)
  Time+PAFU_6-8 : 1 (0%)
  Time+PAFU_9-11 : 31047 (7.56%)

Time+PRPU : 145750 (35.49%)
  Time+PRPU_1 : 75399 (18.36%)
  Time+PRPU_2 : 10919 (2.66%)
  Time+PRPU_3 : 17 (0%)
  Time+PRPU_4 : 1692 (0.41%)
  Time+PRPU_5 : 35815 (8.72%)
  Time+PRPU_6 : 2252 (0.55%)
  Time+PRPU_7 : 8587 (2.09%)
  Time+PRPU_8 : 11069 (2.7%)

Time+PRFU : 35011 (8.52%)
  Time+PRFU : 35 (0.01%)
  Time+PRFU_1-2 : 7631 (1.86%)
  Time+PRFU_11-12 : 2967 (0.72%)
  Time+PRFU_3-4 : 6001 (1.46%)
  Time+PRFU_5 : 849 (0.21%)
  Time+PRFU_6 : 4681 (1.14%)
  Time+PRFU_7 : 12277 (2.99%)
  Time+PRFU_9-10 : 570 (0.14%)

Time+DAPU : 11684 (2.84%)
  Time+DAPU_1 : 1 (0%)
  Time+DAPU_2 : 1 (0%)
  Time+DAPU_3 : 3250 (0.79%)
  Time+DAPU_4 : 22 (0.01%)
  Time+DAPU_5 : 5509 (1.34%)
  Time+DAPU_6 : 2901 (0.71%)

Time+DAFU : 2010 (0.49%)
  Time+DAFU_2 : 378 (0.09%)
```

```
  Time+DAFU_3 : 49 (0.01%)
  Time+DAFU_4 : 323 (0.08%)
  Time+DAFU_5 : 1260 (0.31%)

Time+DRPU : 23382 (5.69%)
  Time+DRPU_1 : 2567 (0.63%)
  Time+DRPU_2 : 4097 (1%)
  Time+DRPU_3 : 3814 (0.93%)
  Time+DRPU_4 : 81 (0.02%)
  Time+DRPU_5 : 6022 (1.47%)
  Time+DRPU_6 : 5012 (1.22%)
  Time+DRPU_7 : 832 (0.2%)
  Time+DRPU_8 : 957 (0.23%)

Time+DRFU : 18475 (4.5%)
  Time+DRFU_1 : 87 (0.02%)
  Time+DRFU_2 : 125 (0.03%)
  Time+DRFU_3 : 707 (0.17%)
  Time+DRFU_4 : 840 (0.2%)
  Time+DRFU_5 : 1982 (0.48%)
  Time+DRFU_6 : 1849 (0.45%)
  Time+DRFU_7 : 12885 (3.14%)

Time+Duree : 47076 (11.46%)

Time+Duree+Imprecis : 13111 (3.19%)

Time+Age : 10584 (2.58%)

Time+Age+Imprecis : 1362 (0.33%)
```

Bibliographie

ABNEY, S. (1996). Partial parsing via finite-state cascades. *In Workshop on Robust Parsing, 8th European Summer School in Logic, Language and Information*, pages 8–15.

ADAFRE, S. F. et de RIJKE, M. (2005). Feature engineering and Post-Processing for temporal expression recognition using conditional random fields. *In Proceedings ACL-2005 Workshop on Feature Engineering*.

AFNOR (1981). Règles d'établissement des thésaurus monolingues. Norme homologuée NF Z47-100.

AFNOR (1993). Information et documentation - principes généraux pour l'indexation des documents. Norme homologuée NF Z47-102.

AHN, D., ADAFRE, S. F. et de RIJKE, M. (2005). Extracting temporal information from open domain text : A comparative exploration. *Journal of Digital Information Management*, 3(1):14–20.

AHN, D., RANTWIJK, J. et de RIJKE, M. (2007). A cascaded machine learning approach to interpreting temporal expressions. *In Proceedings of NAACL HLT 2007*, pages 420–427, Rochester, New York. Association for Computational Linguistics.

AÏT-MOKHTAR, S., CHANOD, J. et ROUX, C. (2002). Robustness beyond shallowness : incremental deep parsing. *Nat. Lang. Eng.*, 8(3):121–144.

ALLEN, J. (1984). Towards a general theory of action and time. *Artificial Intelligence*, 23(2):123–154.

ALLEN, J. (1991). Time and time again : The many ways to represent time. *International Journal of Intelligent Systems*, 6(4):341–355.

ALONSO, O., GERTZ, M. et BAEZA-YATES, R. (2007). On the value of temporal information in information retrieval. *ACM SIGIR Forum*, 41(2):35–41.

ALONSO, O., GERTZ, M. et BAEZA-YATES, R. (2009). Clustering and exploring search results using timeline constructions. *In CIKM 2009*, pages 97–106, Hong Kong.

AMGHAR, T., BATTISTELLI, D. et CHARNOIS, T. (2002). Reasonning on Aspectual-Temporal information in french within conceptual graphs. *In Proceeding of the 14th IEEE International Conference on Tools with Artificial Intelligence (ICTAI2002)*, pages 315–322, Los Alamitos, CA, USA. IEEE Computer Society.

ANANIADOU, S. et MCNAUGHT, J. (2006). Introduction to text mining in biology. *In Text Mining for Biology and Biomedicine*, pages 1–12. Artech House Books.

ARNAULD, A. et LANCELOT, C. (1810). *Grammaire générale et raisonnée de Port-Royal*. Bossange et Masson.

ARONSON, A. R. (2001). Effective mapping of biomedical text to the UMLS metathesaurus : the MetaMap program. *Proceedings of the AMIA Symposium*, pages 17–21.

ARONSON, A. R. et LANG, F. (2010). An overview of MetaMap : historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

ASSADI, H. et BEAUDOUIN, V. (2002). Comment utilise-t-on les moteurs de recherche sur internet ? *Réseaux*, 116(6):171–198.

BAEZA-YATES, R. et RIBEIRO-NETO, B. (1999). *Modern Information Retrieval*. Addison Wesley, 1st édition.

BAGLEY, P. (1968). *Extension of programming language concepts*. Philadelphia : University City Science Center.

BAPTISTA, J. et GUITART, D. C. (2002). Compound temporal adverbs in portuguese and in spanish. *In* RANCHHOD, E. et MAMEDE, N. J., éditeurs : *Proceedings of the Third International Conference on Natural Language Processing (PorTAL 2002)*, volume 2389 de *Lecture Notes in Computer Science, Advances in Natural Language Processing*, pages 133 – 136. Springer-Verlag, Faro, Potugal.

BATTISTELLI, D., MINEL, J. et SCHWER, S. (2006). Représentation des expressions calendaires dans les textes : vers une application à la lecture assistée de biographies. *TAL*, 47(3):11–37.

BEAUZÉE, N. (1767). *Grammaire Générale*. J. Barbou, Paris.

BELL, A. (1998). The discourse structure of news stories. *In* BELL, A. et GARRETT, P., éditeurs : *Approaches to Media Discourse*, pages 65–103. Malden, Oxford.

BENVENISTE, E. (1974). *Problèmes de linguistique générale, tome 2*. Gallimard.

BERNERS-LEE, J. H. T. et LASSILA, O. (2001). The semantic web. *Scientific American*, pages 29–37.

BESTOUGEFF, H. et LIGOZAT, G. (1989). *Outils logiques pour le traitement du temps*. Masson.

BETTINI, C., JAJODIA, S. G. et WANG, S. X. (2000). *Time Granularities in Databases, Data Mining and Temporal Reasoning*. Springer-Verlag New York, Inc.

BHOGAL, J., MACFARLANE, A. et SMITH, P. (2007). A review of ontology based query expansion. *Information Processing & Management*, 43(4):866–886.

BILHAUT, F. (2007). Analyse thématique automatique fondée sur la notion d'univers de discours. *Discours*, 1.

BILHAUT, F., DUMONCEL, F., ENJALBERT, P. et HERNANDEZ, N. (2007). Indexation sémantique et recherche d'information interactive. *In Actes de la Quatrième Conférence Francophone en Recherche d'Information et Applications (CORIA)*, pages 65–76, Saint-Etienne, France.

BILHAUT, F., HO-DAC, M., BORILLO, A., CHARNOIS, T., ENJALBERT, P., DRAOULEC, A. L., MATHET, Y., MIGUET, H., PÉRY-WOODLEY, M. et SARDA, L. (2003). Indexation discursive pour la navigation intradocumentaire : cadres temporels et spatiaux dans l'information géographique. *In Actes de la 10e Conférence Traitement Automatique du Langage Naturel (TALN'03)*, pages 315–320, Batz-sur-Mer.

BIPM (2006). Le système international d'unités (SI). Rapport technique 8e édition, Bureau International des Poids et Mesures - Organisation intergouvernementale de la Convention du Mètre, Paris, France.

BIPM (2010). International atomic time. http ://www.bipm.org/en/scientific/tai/tai.html.

BITTAR, A. (2008). Annotation des informations temporelles dans des textes en français. *In Actes de la 12e édition de RECITAL*, pages 11–20, Avignon, France.

BITTAR, A. (2009). Annotation of events and temporal expressions in french texts. *In Computational Linguistics in the Netherlands 19*, Groningen, Pays-Bas.

BITTAR, A. (2010). *Construction d'un TimeBank du français : un corpus de référence annoté selon la norme ISO-TimeML*. Thèse de doctorat, Université Paris Diderot (Paris 7), Paris, France.

BOGURAEV, B., CASTANO, J., GAIZAUSKAS, R., INGRIA, R., KATZ, G., JESSICA, J. L., MANI, I., PUSTEJOVSKY, J., SANFILIPPO, A., SEE, A., SETZER, A., SAURI, R., STUBBS, A., SUNDHEIM, B., SYMONENKO, S. et VERHAGEN, M. (2005). TimeML 1.2.1. a formal specification language for events and temporal expressions. Rapport technique 1.2.1.

BORILLO, A. (1983). Les adverbes de référence temporelle dans la phrase et dans le texte. *DRLAV : Revue de linguistique*, 29:109–131.

BORILLO, A. (1986). Lexique et syntaxe : Les emplois adverbiaux des noms de temps. *In Actes du séminaire "Lexique et Traitement Automatique des Langues"*, Toulouse, France. Université Paul Sabatier.

BORILLO, A. (1988). L'expression de la durée : construction des noms et des verbes de mesure temporelle. *Linguisticae Investigationes*, XII(2):363–396.

BORILLO, A. (1998). Les adverbes de référence temporelle comme connecteurs temporels de discours. *In* VOGELEER, S., BORILLO, A., VETTERS, C. et VUILLAUME, M., éditeurs : *Temps et discours*, volume 99 de *Bibliothèque des cahiers de l'institut de linguistique de Louvain (BCILL)*, pages 131–145. Peeters, Louvain-la-Neuve.

BORILLO, A., BRAS, M., LE DRAOULEC, A., VIEU, L., MOLENDIJK, A., de SWART, H., VERKUYL, H., VET, C. et VETTERS, C. (2004). Tense, connectives and discourse structure. *In* CORBLIN, F. et de SWART, H., éditeurs : *Handbook of French Semantics*, CSLI Lecture Notes, pages 309–348. CSLI Publications, Standford.

BOUGHANEM, M., TAMINE-LECHANI, L., MARTINEZ, J., CALABRETTO, S. et CHEVALLET, J. P. (2006). Un nouveau passage à l'échelle en recherche d'information. *Ingénierie des Systèmes d'Information (ISI)*, 11(4):9–35.

BRIN, S. et PAGE, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117.

CALABRESE STEIMBERG, L. (2008). Les héméronymes. ces évènements qui font date, ces dates qui deviennent évènements. *Mots. Les langages du politique*, 88(3):115–128.

CASELLI, T., IDE, N. et BARTOLINI, R. (2008). A bilingual corpus of inter-linked events. *In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech (Morocco).

CEUSTERS, W., MICHEL, C., PENSON, D. et MAUCLET, E. (1994). Semi-automated encoding of diagnoses and medical procedures combining ICD-9-CM with computational-linguistic tools. *Ann Med Milit Belg*, 8(2):53–58.

CHAROLLES, M. (1997). L'encadrement du discours : univers, champs, domaines et espaces. *Cahier de Recherche Linguistique*, 6:1–73.

CHAROLLES, M., DRAOULEC, A. L., PERRY-WOODLEY, M. et SARDA, L. (2005). Temporal and spatial dimensions of discourse organisation. *Journal of French Language Studies*, 15(02):115–130.

CHAUMIER, J. et DEJEAN, M. (2003). Recherche et analyse de l'information textuelle. *Documentaliste-Sciences de l'Information*, 40(1):14.

CHENG, Y., ASAHARA, M. et MATSUMOTO, Y. (2007). Constructing a temporal relation tagged corpus of chinese based on dependency structure analysis. *In Temporal Representation and Reasoning, 14th International Symposium on*, pages 59–69.

CHINCHOR, N. (1997). MUC-7 named entity task definition. *In Message Understanding Conference Proceedings MUC-7*.

COMMISSION EUROPÉENNE (2010). Europe 2020 - une stratégie pour une croissance intelligente, durable et inclusive. Communication de la Commission COM(2010) 2020, Bruxelles.

COMRIE, B. (1976). *Aspect*. Cambridge University Press.

COMRIE, B. (1985). *Tense*. Cambridge University Press.

CONSTANT, M. (2003). *Grammaires locales pour l'analyse automatique de textes : méthodes de construction et outils de gestion*. Thèse de doctorat en informatique, Université de Marne-la-Vallée.

CONSTANT, M., NAKAMURA, T. et PAUMIER, S. (2002). L'héritage des gènes MG - localisation d'auxiliaires en français. *In Actes du 21e colloque international Grammaires et Lexiques Comparés*, Bari.

CORTES, C. et VAPNIK, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

COSTERMANS, J. et BESTGEN, Y. (1991). The role of temporal markers in the segmentation of narrative discourse. *CPC/ European Bulletin of Cognitive Psychology*, 11:349–370.

CUI, H., WEN, J., NIE, J. et MA, W. (2002). Probabilistic query expansion using query logs. *In Proceedings of the 11th international conference on World Wide Web*, pages 325–332, Honolulu, Hawaii, USA. ACM.

CULIOLI, A. (1983). A propos de quelque. *In* FISHER, S. et FRANCKEL, J., éditeurs : *Linguistique, énonciation. Aspects et détermination*, pages 21–29. EHESS.

CUNNINGHAM, H., MAYNARD, D., BONTCHEVA, K. et TABLAN, V. (2002). GATE : a framework and graphical development environment for robust NLP tools and applications. *In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*.

DA SYLVA, L. (2004). Les traitements documentaires automatiques et le passage du temps. *In Actes du colloque EBSI-ENSSIB « Le numérique : Impact sur le cycle de vie du document »*, Montréal.

DA SYLVA, L. (2006). Thésaurus et systèmes de traitement automatique de la langue. *Documentation et bibliothèques*, 52(2):149–156.

DALBIN, S. (2007). Thésaurus et informatique documentaires. *Documentaliste-Sciences de l'Information*, 44(1):42.

DAVID, P. A. et FORAY, D. (2002). Une introduction à l'économie et à la société du savoir. *Revue internationale des sciences sociales*, 171(1):13–28.

DAWSON, R. (2010). Launch of newspaper extinction timeline for every country in the world. http ://-rossdawsonblog.com/weblog/archives/2010/10/launch_of_newsp.html, accédé le 17/11/2010.

DE SAINT-OURS, A. (2010). Variations sur une définition. *La Recherche*, 442(juin 2010):44–45.

DESCLÉS, J. (1991). Archétypes cognitifs et types de procès. *In* FUCHS, C., éditeur : *Les typologies de procès*, numéro 29 de Travaux de linguistique et de philologie, pages 171–195. Klincksieck, Paris.

DESCLÉS, J., CARTIER, E., JACKIEWICZ, A. et MINEL, J. (1997). Textual processing and contextual exploration method. *In CONTEXT'97*, pages 189–197, Rio de Janeiro.

DIKI-KIDIRI, M. (2007). *Comment assurer la présence d'une langue dans le cyberespace ?* UNESCO, Organisation des Nations Unies pour l'éducation, la science et la culture. Programme Information pour tous (PIPT). Division de la société de l'information, Secteur de la communication et de l'information.

FAIRON, C. et SENELLART, J. (1999). Réflexions sur la localisation, l'étiquetage, la reconnaissance et la traduction d'expressions linguistiques complexes. *In Actes de la conférence TALN 1999*, Cargese, France.

FAIRON, C. et WATRIN, P. (2003). From extraction to indexation. collecting new indexation keys by means of IE techniques. *In Proceedings of the Workshop on Finite-State Methods in Natural Language Processing (EACL-2003)*, pages 113–118.

FARKAS, R. et SZARVAS, G. (2008). Automatic construction of rule-based ICD-9-CM coding systems. *BMC Bioinformatics*, 9(Suppl 3):S10–S10.

FERRARI, S., BILHAUT, F., WILDÖCHER, A. et LAIGNELET, M. (2005). Une plate-forme logicielle et une démarche pour la validation de ressources linguistiques sur corpus : application à l'évaluation de la détection automatique de cadres temporels. *In Actes des 4èmes Journées de Linguistique de Corpus*, pages 187–196, Lorient, France.

FERRET, O., GRAU, B., MINEL, J. et PORHIEL, S. (2001). Repérage de structures thématiques dans des textes. *In Actes de la 8ème conférence sur le Traitement Automatique des Langues Naturelles (TALN2001)*, pages 163–172, Tours, France.

FERRO, L., GERBER, L., MANI, I., SUNDHEIM, B. et WILSON, G. (2005). TIDES 2005 standard for the annotation of temporal expressions. Rapport technique, MITRE.

FIKES, R. et MAKARIOS, S. (2004). KANI time ontology. KSL Technical Report KSL-04-05, Knowledge System Laboratory, Stanford University, Stanford.

FILATOVA, E. et HOVY, E. (2001). Assigning Time-Stamps to Event-Clauses. *In Proceedings of the 2001 ACL Workshop on Temporal and Spatial Information Processing*, pages 1–8, Toulouse, France. Association for Computational Linguistics.

FRANCIS, W. et KUCERA, H. (1982). *Frequency Analysis of English Usage. Lexicon and Grammar.* Houghton Mifflin.

FRIEDMAN, C., SHAGINA, L., LUSSIER, Y. et HRIPCSAK, G. (2004). Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association*, 11(5):392–402.

FUCHS, C. (1991). Les typologies de procès : un carrefour théorique interdisciplinaire. *In* FUCHS, C., éditeur : *Les typologies de procès*, Travaux de linguistique et de philologie, pages 9–17. Klincksieck, Paris.

GAILLARD, B., BOURAOUI, J., de NEEF, E. G. et BOUALEM, M. (2010). Query expansion for cross language information retrieval improvement. *In Fourth International Conference on Research Challenges in Information Science (RCIS)*, pages 337–342.

GIRARD, G. (1747). *Les vrais principes de la langue françoise, ou, La parole réduite en méthode, conformément aux loix de l'usage*. Le Breton.

GIRAULT, S. (2007). *Recherche sur les marques temporelles et aspectuelles dans les organisations narratives*. Thèse de doctorat, Université de Caen/Basse-Normandie.

GOLDSTEIN, I., ARZRUMTSYAN, A. et UZUNER, O. (2007). Three approaches to automatic assignment of ICD-9-CM codes to radiology reports. *Proceedings of AMIA Annual Symposium*, pages 279–83.

GOSSELIN, L. (1996). *Sémantique de la temporalité en français*. Champs linguistiques. De Boeck, Bruxelles.

GOSSELIN, L. (2005). *Temporalité et modalité*. Champs linguistiques. De Boeck, Bruxelles.

GOSSELIN, L. et FRANÇOIS, J. (1991). Les typologies de procès : des verbes aux prédications. *In* FUCHS, C., éditeur : *Les typologies de procès*, Travaux de linguistique et de philologie, pages 19–86. Klincksieck, Paris.

GREENSLADE, R. (2010). Newspapers will be irrelevant in 12 years. http ://www.guardian.co.uk/media/greenslade/2010/aug/24/newspapers-crowdsourcing, accédé le 17/10/2010.

GRISHMAN, R. (1997). Information extraction : Techniques and challenges. *In* PAZIENZA, J. S. M. T. et CARBONELL, J. G., éditeurs : *Information Extraction : A Multidisciplinary Approach to an Emerging Information Technology*, pages 10–27.

GRISHMAN, R. et SUNDHEIM, B. (1996). Message understanding conference-6 : a brief history. *In Proceedings of the 16th conference on Computational linguistics - Volume 1*, pages 466–471, Copenhagen, Denmark. Association for Computational Linguistics.

GROSS, M. (1986). *Grammaire transformationnelle du français. 3 - Syntaxe de l'adverbe*. Asstril, Paris.

GROSS, M. (1989). The use of finite automata in the lexical representation of natural language. volume 377 de *Lecture Notes in Computer Science*, pages 34–50. Springer Verlag, Berlin.

GROSS, M. (1997). The construction of local grammars. *In* ROCHE, E. et SCHABÈS, Y., éditeurs : *Finite-State Language Processing*, pages 329–354. The MIT Press.

GROSS, M. (1999). A bootstrap method for constructing local grammars. *In Contemporary Mathematics. Proceedings of the Symposium*, Belgrad, University of Belgrad.

GROSS, M. (2000). Lemmatization of compound tenses in english. *Lingvisticae Investigationes*, 22(2):71–122.

GROSS, M. (2002). Les déterminants numéraux, un exemple : les dates horaires. *Langages*, 36(145): 21–37.

GROSSMAN, D. A. et FRIEDER, O. (2004). *Information Retrieval : Algorithms and Heuristics (The Information Retrieval Series), 2nd ed.* Springer.

GRUBER, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220.

GUARINO, N. (1998). Formal ontology and information systems. *In* GUARINO, N., éditeur : *Proceedings of the 1st International Conference on Formal Ontologies in Information Systems, FOIS'98*, pages 3–15. IOS Press.

GUELFI, N., PRUSKI, C. et REYNAUD, C. (2007). Les ontologies pour la recherche ciblée d'information sur le web : une utilisation et extension d'OWL pour l'expansion de requêtes. *In 18èmes journées francophones d'Ingénierie des Connaissances, IC'2007*, pages 61–72, Grenoble, France.

GUILLAUME, G. (1964). *Langage et science du langage*. Nizet-Presses.

HACIOGLU, K., CHEN, Y. et DOUGLAS, B. (2005). *Automatic Time Expression Labeling for English and Chinese Text*, pages 548–559.

HAGÈGE, C. et TANNIER, X. (2008). XTM : a robust temporal text processor. volume 4919 de *Lecture Notes in Computer Science*, pages 231–240. Springer-Verlag.

HAJEK, P. (2010). Fuzzy logic. *In* ZALTA, E. N., éditeur : *The Stanford Encyclopedia of Philosophy (Fall 2010 Edition)*. http ://plato.stanford.edu/archives/fall2010/entries/logic-fuzzy, accédé le 17/01/2011.

HAN, B. et LAVIE, A. (2004). A framework for resolution of time in natural language. *ACM Transactions on Asian Language Information Processing Special Issue on Spatial and Temporal Information Processing*, 3(1):11–32.

HARRIS, Z. S. (1954). Distributional structure. *Word*, 10(23):146–162.

HOBBS, J., ALLEN, J., FIKES, J., HAYES, P., MCDERMOTT, D., NILES, I., PEASE, A., TATE, A., TYSON, M. et WALDINGER, R. (2002). A DAML ontology of time.

HOBBS, J. et PAN, F. (2004). An ontology of time for the semantic web. *ACM Transactions on Asian Language Information Processing*, 3(1):66–85.

HOWE, J. (2006). The rise of crowdsourcing. *Wired*, (14.06).

IKEDA, M., SETA, K. et MIZOGUCHI, R. (1997). Task ontology makes it easier to use authoring tools. *In IJCAI'97 Proceedings of the 15th international joint conference on Artifical intelligence*, pages 342–347.

ISO (1986). Guidelines for the establishment and development of monolingual thesauri. ISO 2788.

ISO (2004). ISO 8601 :2004 - Éléments de données et formats d'échange - Échange d'information - représentation de la date et de l'heure. Rapport technique, International Organization for Standardization, Geneva, Switzerland.

JANG, S. B., BALDWIN, J. et MANI, I. (2004). Automatic TIMEX2 tagging of korean news. *ACM Transactions on Asian Language Information Processing (TALIP)*, 3(1):51–65.

JANSEN, B. et EASTMAN, C. (2003). The effects of search engines and query operators on top ranked results. *In Proceedings ITCC 2003. International Conference on Information Technology : Coding and Computing*, pages 135–139, Las Vegas, NV, USA.

JESPERSEN, O. (1971). *La philosophie de la grammaire*. Gallimard.

JOHO, H., COVERSON, C., SANDERSON, M. et BEAULIEU, M. (2002). Hierarchical presentation of expansion terms. *In Proceedings of the 2002 ACM symposium on Applied computing*, pages 645–649, Madrid, Spain. ACM.

KENNY, A. (1963). States, performances, activities. *In Action, Emotion and Will*. Routledge & Kegan Paul.

KEVERS, L. (2006). L'information biographique : modélisation, extraction et organisation en base de connaissances. *In* MERTENS, P., FAIRON, C., DISTER, A. et WATRIN, P., éditeurs : *Verbum ex machina, actes de la 13eme conférence sur le traitement automatique des langues naturelle (TALN06)*, pages 680–689. Presses Universitaires de Louvain.

KEVERS, L. et FAIRON, C. (2007). Vers une base de connaissances biographique : extraction d'information et ontologie. *In* NOIRHOMME-FRAITURE, M. et VENTURINI, G., éditeurs : *Actes des cinquièmes journées Extraction et Gestion des Connaissances*, volume RNTI-E-9 de *Revue des Nouvelles Technologies de l'Information*, pages 373–378, Namur, Belgique. Cépaduès-Éditions 2007.

KEVERS, L., MANTRACH, A., FAIRON, C., BERSINI, H. et SAERENS, M. (2010). Classification supervisée hybride par motifs lexicaux étendus et classificateurs SVM. *In Actes des 10ème Journées internationales d'analyse des données textuelles*, Rome, Italy.

KEVERS, L. et MEDORI, J. (2010). Symbolic classification methods for patient discharge summaries encoding into ICD. *In* LOFTSSON, H., RÖGNVALDSSON, E. et HELGADÓTTIR, S., éditeurs : *Proceedings of 7th International Conference on NLP, IceTAL 2010,Reykjavik, Iceland, 16-18 August 2010*, volume 6233 de *Lecture Notes in Computer Science, Advances in Natural Language Processing*, pages 197–208. Springer.

KITTREDGE, R. et LEHRBERGER, J. (1982). *Sublanguage : Studies of Language in Restricted Semantic Domains*. Walter de Gruyter.

KLEIN, E. (2010). Insaisissable. *La Recherche*, 442(juin 2010):40.

KOWALSKI, R. et SERGOT, M. (1986). A logic-based calculus of events. *New Gen. Comput.*, 4(1): 67–95.

KYUNG-SUN, K., SEI-CHING, J. S., SOO-JIN, P., XIAOHUA, Z. et POLPARSI, J. (2006). Facet analysis of categories used in web directories : a comparative study. *In World Library and Information Congress : 72nd of IFLA General Conference and Council*.

LA DOCUMENTATION FRANÇAISE (2007). Internet dans le monde : Lutte contre la fracture numérique dans le monde. http ://www.ladocumentationfrancaise.fr/dossiers/internet-monde/fracture-numerique.shtml, accédé le 17/11/2010.

LACROIX, C. (2010). Technologies de l'information. *In Statistiques de la culture : chiffres clés 2010*. La Documentation Française, France.

LE MEUR, H. (2010). Carlo Rovelli : «il faut oublier le temps». *La Recherche*, 442(juin 2010):41–43.

LE PARC-LACAYRELLE, A., GAIO, M. et SALLABERRY, C. (2007). La composante temps dans l'information géographique textuelle. *Document numérique*, 10(2007/2):129–148.

LECUIT, E., MAUREL, D., DUSKO, V. et CVETANA, K. (2009). Temporal expressions : Comparisons in a multilingual corpus. *In Proceedings of the 4th Language & Technology Conference*, pages 531–535, Poznań, Poland.

LEVY, M. et JOUYET, J. (2006). L'économie de l'immatériel. la croissance de demain. Rapport de la commission sur l'économie de l'immatériel, Paris.

LI, W., WONG, K. et YUAN, C. (2001). A model for processing temporal references in chinese. *In Proceedings of the workshop on Temporal and spatial information processing - Volume 13*, pages 1–8. Association for Computational Linguistics.

LYONS, J. (1980). *Sémantique linguistique*. Larousse.

MANGUINHAS, H., MARTINS, B., BORBINHA, J. et VACA, W. L. S. (2009). The DIGMAP geo-temporal web gazetteer service. *In Proceedings of Third Internacional Workshop Digital Approaches to Cartographic Heritage*, volume 4, Issue 1 de *e-Perimetron*, pages 9–24, Barcelona, Espagne.

MANI, I., PUSTEJOVSKY, J. et GAIZAUSKAS, R. (2005). *The Language of Time : A Reader*. Oxford University Press, USA.

MANI, I. et SCHIFFMAN, B. (2005). Temporally anchoring and ordering events in news. *In* PUSTEJOVSKY, J. et GAIZAUSKAS, R., éditeurs : *Event Recognition in Natural Language*. John Benjamins, Amsterdam.

MANI, I. et WILSON, G. (2000). Robust temporal processing of news. *In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 69–76, Hong Kong. Association for Computational Linguistics.

MARTINEAU, C., NAKAMURA, T., VARGA, L. et VOYATZI, S. (2009). Annotation et normalisation des entités nommées. *Arena Romanistica*, 4:234–243.

MASCHERIN, L. (2007). *Analyse morphosémantique de l'aspectuo-temporalité en français. Le cas du préfixe RE-*. Thèse de doctorat, Université de Nancy 2.

MAUREL, D. (1990). Description par automate des dates et des adverbes apparentés. *Mathématiques et sciences humaines*, (109):5–16.

MAUREL, D. (2008). Prolexbase : Une base de données lexicale de noms propres pour le tal. *In Actes du Colloque Lexicographie et informatique : bilan et perspectives*, pages 137–144, Nancy, France.

MAUREL, D. et MOHRI, M. (1994). French temporal expressions : Recognition, parsing and real computation. *In Reflections on the Future of text*, pages 33–41, Waterloo, Ontario, Canada.

MCCARTHY, J. et HAYES, P. (1969). Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence*, 4:463–502.

MCDERMOTT, D. (1982). A temporal logic for reasoning about processes and plans. *Cognitive Science*, 6(2):101–155.

MEDELYAN, O. et WITTEN, I. H. (2005). Thesaurus-Based index term extraction for agricultural documents. *In Proc. of the 6th Agricultural Ontology Service (AOS) workshop at EFITA/WCCA 2005*, Vila Real, Portugal.

MEDELYAN, O. et WITTEN, I. H. (2006). Thesaurus based automatic keyphrase indexing. *In Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 296–297, Chapel Hill, NC, USA. ACM.

MEDORI, J. (2008). From free text to ICD : development of a coding help. *In Proceedings of the First Louhi Workshop on Text and Data Mining of Health Documents*, Turku, Finland.

MEDORI, J. (2010). Machine learning and features selection for semi-automatic ICD-9-CM encoding. *In Proceedings of the Second Louhi Workshop on Text and Data Mining of Health Documents*, Los Angeles, USA.

MEYER, P. (2009). Julian day numbers. http ://www.hermetic.ch/cal_stud/jdn.htm, accédé le 17/11/2010.

MOENS, M. (2006). *Information Extraction : Algorithms and Prospects in a Retrieval Context*. Springer, 1 édition.

MOLDOVAN, D. I. et MIHALCEA, R. (2000). Using WordNet and lexical operators to improve internet searches. *IEEE Internet Computing*, 4(1):34–43.

MOURELATOS, A. P. D. (1978). Events, processes, and states. *Linguistics and Philosophy*, 2(3):415–434.

MULLER, P. et TANNIER, X. (2004). Annotating and measuring temporal relations in texts. *In Proceedings of the 20th international conference on Computational Linguistics*, Geneva, Switzerland. Association for Computational Linguistics.

NAKAMURA, T. (2006). Package VAUX. http ://monge.univ-mlv.fr/~nakamura/VAUXTN2.html.

NÉDELLEC, C., VETAH, M. A. et BESSIÈRES, P. (2001). Sentence filtering for information extraction in genomics, a classification problem. *In Principles of Data Mining and Knowledge Discovery*, volume 2168/2001 de *Lecture Notes in Computer Science*, pages 326–337. Springer, Berlin / Heidelberg.

NÉVÉOL, A., MORK, J. G., ARONSON, A. R. et DARMONI, S. J. (2005). Evaluation of french and english MeSH indexing systems with a parallel corpus. *AMIA Annual Symposium Proceedings*, 2005:565–569.

NÉVÉOL, A., ROGOZAN, A. et DARMONI, S. (2006). Automatic indexing of online health resources for a french quality controlled gateway. *Information Processing and Management : an International Journal*, 42(3):695–709.

NILES, I. et PEASE, A. (2001). Towards a standard upper ontology. *In Proceedings of the international conference on Formal Ontology in Information Systems - Volume 2001*, pages 2–9, Ogunquit, Maine, USA. ACM.

NISO (2004). *Understanding Metadata*. NISO Press.

NIST (2006). The ACE 2007 (ACE07) evaluation plan. evaluation of the detection and recognition of ACE entities, values, temporal expressions, relations and events. Rapport technique 1.3, National Institute of Standards and Technology.

NOJGAARD, M. (1992). *Les adverbes français : essai de description fonctionnelle*. Munksgaard, Copenhagen.

NUNES, S., RIBEIRO, C. et DAVID, G. (2008). Use of temporal expressions in web search. *In* MACDONALD, C., OUNIS, I., PLACHOURAS, V., RUTHVEN, I. et WHITE, R. W., éditeurs : *Proceedings of the 30th European Conference on IR Research, ECIR 2008*, volume 4956 de *Lecture Notes in Computer Science*, page 2008, Glasgow, UK. Springer Berlin Heidelberg.

NÚNEZ, R. et SWEETSER, E. (2006). With the future behind them : Convergent evidence from aymara language and gesture in the crosslinguistic comparison of spatial construals of time. *Cognitive Science*, 30(3):401–450.

O'CONNOR, M. J. et DAS, A. K. (2010). A lightweight model for representing and reasoning with temporal information in biomedical ontologies. *In Proceedings of International Conference on Health Informatics (HEALTHINF)*, Valencia, Spain.

O'CONNOR, M. J., SHANKAR, R. D., PARRISH, D. B. et DAS, A. K. (2009). Knowledge-data integration for temporal reasoning in a clinical trial system. *International Journal of Medical Informatics*, 78, Supplement 1:S77–S85.

OGDEN, C. K. et RICHARDS, I. A. (1969). *The meaning of meaning : a study of the influence of language upon thought and of the science of symbolism*. International library of psychology,

philosophy and scientific method. Routledge & Kegan Paul, London, 10th ed., 7th impr. of the 1923 ed. édition.

OHLBACH, H. J. (2000). About real time, calendar systems and temporal notions. *In* BARRINGER, H., FISHER, M., GABBAY, D. et GOUGHS, G., éditeurs : *Advances in temporal logic*, numéro 16 de Applied Logic Series, pages 319–338. Kluwer Academic Publishers.

PAKHOMOV, S. V., BUNTROCK, J. D. et CHUTE, C. G. (2006). Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *JAMIA*, 13(5):516–525.

PALACIO, D., CABANAC, G., SALLABERRY, C. et HUBERT, G. (2010). Cadre d'évaluation de systèmes de recherche d'information géographique apport de la combinaison des dimensions spatiale, temporelle et thématique. *In Actes de INFORSID'10 : 28e congrès de l'INFormatique des Organisations et Systèmes d'Information et de Décision*, pages 245–260, Marseille : France.

PAPROCKA-PIOTROWSKA, U. et DEMAGNY, A. (2004). L'acquisition du lexique verbal et des connecteurs temporels dans les récits de fiction en français l1 et l2. *Langages*, 155(3):52–75.

PARENT, G., GAGNON, M. et MULLER, P. (2008). Annotation d'expressions temporelles et d'événements en français. *In Actes de la 15e Conférence sur le Traitement Automatique des Langues Naturelles*, pages 39–48, Avignon, France.

PASCA, M. (2008). Towards temporal web search. *In Proceedings of the 23rd ACM Symposium on Applied Computing (SAC-2008)*, pages 1117–1121.

PASCAL, B. (1838). Pensées. *In Moralistes français*, pages 22–145. Firmin Didot frères, Paris.

PAUMIER, S. (2003). *De la reconnaissance de formes linguistiques à l'analyse syntaxique*. Thèse de doctorat, Université de Marne-la-Vallée.

PAUMIER, S. (2008). *Unitex 2.0 User Manual*.

PEREIRA, S., NEVEOL, A., KERDELHUÉ, G., SERROT, E., JOUBERT, M. et DARMONI, S. J. (2008). Using multi-terminology indexing for the assignment of MeSH descriptors to health resources in a french online catalogue. *AMIA Annual Symposium Proceedings*, pages 586–590.

PEREIRA, S., NÉVÉOL, A., MASSARI, P., JOUBERT, M. et DARMONI, S. (2006). Construction of a semi-automated ICD-10 coding help system to optimize medical and economic coding. *Studies in Health Technology and Informatics*, 124:845–850.

PERSON, C. (2004). *Traitement automatique de la temporalité du récit : implémentation du modèle linguistique SdT*. Thèse de doctorat, Université de Caen.

PESTIAN, J. P., BREW, C., MATYKIEWICZ, P., HOVERMALE, D. J., JOHNSON, N., COHEN, K. B. et DUCH, W. (2007). A shared task involving multi-label classification of clinical free text. *In Proceedings of the Workshop on BioNLP 2007 : Biological, Translational, and Clinical Language Processing*, pages 97–104, Prague, Czech Republic. ACL.

PIÉRARD, S., DEGAND, L. et BESTGEN, Y. (2004). Vers une recherche automatique des marqueurs de la segmentation du discours. *In* PURNELLE, G., FAIRON, C. et DISTER, A., éditeurs : *Le poids des mots. Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles (JADT04)*, pages 859–864, Louvain-la-Neuve. Presses Universitaires de Louvain.

PINCHON, J. (1969). Problèmes de classification. les adverbes de temps. *Langue Française*, 1(1):74–81.

PORTER, M. F. (1997). *An algorithm for suffix stripping*, pages 313–316. Morgan Kaufmann Publishers Inc.

PORTINE, H. (1995). Repérages et rôle de la géométrie dans l'analyse des temps verbaux. l'exemple de beauzée. *Mathématiques et Sciences Humaines*, 130:5–26.

POULIQUEN, B., STEINBERGER, R. et IGNAT, C. (2003). Automatic annotation of multilingual text collections with a conceptual thesaurus. *In Proceedings of the Workshop 'Ontologies and Information Extraction' at the Summer School 'The Semantic Web and Language Technology - Its Potential and Practicalities' (EUROLAN'2003)*, pages 9–28, Bucharest, Romania.

PUSTEJOVSKY, J., BELANGER, L., CASTANO, J., GAIZAUSKAS, R., HANKS, P., INGRIA, B., KATZ, G., RADEV, D., RUMSHISKY, A., SANFILIPO, A., SAURÍ, R., SETZER, A., SUNDHEIM, B. et VERHAGEN, M. (2002). NRRC summer workshop on temporal and event recognition for question answering systems. Rapport technique 0.1.0.

PUSTEJOVSKY, J., CASTANO, J., INGRIA, R., SAURI, R., GAIZAUSKAS, R., SETZER, A., KATZ, G. et RADEV, D. (2003a). TimeML : robust specification of event and temporal expressions in text. *In IWCS-5 Fifth International Workshop on Computational Semantics*.

PUSTEJOVSKY, J., HANKS, P., SAURI, R., SEE, A., GAIZAUSKAS, R., SETZER, A., RADEV, D., SUNDHEIM, B., DAY, D., FERRO, L. et LAZO, M. (2003b). The TIMEBANK corpus. *In Proceedings of Corpus Linguistics 2003*, pages 656, 647.

QUERLER, N. L. (1996). *Typologie des modalités*. Presses Universitaires de Caen.

RANGANATHAN, S. R. (1967). *Prolegomena to Library Classification*. Asia Publishing House, New York.

REICHENBACH, H. (1947). *Elements of Symbolic Logic*. Macmillan.

REICHENBACH, H. (2005). The tenses of verbs. *In The Language of Time : A Reader*, pages 71–78. Oxford University Press, USA.

RILOFF, E. et LEHNERT, W. (1994). Information extraction as a basis for high-precision text classification. *ACM Trans. Inf. Syst.*, 12(3):296–333.

ROBERTSON, S. E., WALKER, S., BEAULIEU, M. et WILLETT, P. (1998). Okapi at TREC-7 : automatic ad hoc, filtering, VLC and interactive track. *In Proceedings of Seventh Text Retrieval Conference (TREC-7)*, pages 253–264.

SALTON, G. et LESK, M. E. (1968). Computer evaluation of indexing and text processing. *Journal of the ACM*, 15(1):8–36.

SALTON, G. et MCGILL, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc.

SAURÍ, R., LITTMAN, J., KNIPPEN, B., GAIZAUSKAS, R., SETZER, A. et PUSTEJOVSKY, J. (2006). TimeML annotation version 1.2.1. Rapport technique.

SCHILDER, F. et HABEL, C. (2001). From temporal expressions to temporal information : Semantic tagging of news message. *In Proceedings of ACL'01 workshop on temporal and spatial information processing*, pages 65–72, Toulouse, France. Association for Computational Linguistics.

SCHMID, H. (1994). Probabilistic Part-of-Speech tagging using decision trees. *In International Conference on New Methods in Language Processing*, Manchester, UK.

SCHWER, S. (2009a). Représentation du temps, relations temporelles et théories des temps verbaux. HAL-SHS, http ://halshs.archives-ouvertes.fr/halshs-00403655/fr/ (accédé le 05/08/2009).

SCHWER, S. (2009b). Systèmes des temps verbaux de l'indicatif du français à travers les grammaires de Port-Royal (1660), de l'Abbé girard et (1747) et de Nicolas Beauzée (1767). *In Actes des 16èmes rencontres interdisciplinaires sur les systèmes complexes naturels et artificiels*, Rochebrune. http ://gemas.msh-paris.fr/dphan/rochebrune09/papiers/SchwerSylviane.pdf (accédé le 05/08/2009).

SEBASTIANI, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

SETZER, A. (2001). *Temporal Information in Newswire Articles : An Annotation Scheme and Corpus Study*. PhD thesis, University of Sheffield.

SHAH, N. H., BHATIA, N., JONQUET, C., RUBIN, D., CHIANG, A. P. et MUSEN, M. A. (2009). Comparison of concept recognizers for building the open biomedical annotator. *BMC Bioinformatics*, 10(Suppl 9):S14.

SHAH, U., FININ, T., JOSHI, A., COST, R. S. et MATFIELD, J. (2002). Information retrieval on the semantic web. *In Proceedings of the eleventh international conference on Information and knowledge management*, pages 461–468, McLean, Virginia, USA. ACM.

STEFANOWSKI, J. et WEISS, D. (2003). Carrot2 and language properties in web search results clustering. *In Advances in Web Intelligence*, volume 2663/2003 de *Lecture Notes in Computer Science*, page 955. Springer, Berlin / Heidelberg.

STRÖTGEN, J., GERTZ, M. et POPOV, P. (2010). Extraction and exploration of spatio-temporal information in documents. *In Proceedings of the 6th Workshop on Geographic Information Retrieval*, pages 1–8, Zurich, Switzerland. ACM.

SUCHANEK, F. M., KASNECI, G. et WEIKUM, G. (2007). Yago : a core of semantic knowledge. *In Proceedings of the 16th international conference on World Wide Web*, pages 697–706, Banff, Alberta, Canada. ACM.

UMBRICH, J. et BLOHM, S. (2008). Exploring the knowledge in semi structured data sets with rich queries. *In Proceedings of the Workshop on Semantic Search (SemSearch)*, page 89–101, Tenerife, Spain.

VAN SLYPE, G. (1987). *Les langages d'indexation : conception,construction et utilisation dans les systèmes documentaires*. Systèmes d'Information et de Documentation. Les éditions d'organisation, Paris.

VAZOV, N. (2001). A system for extraction of temporal expressions from french texts based on syntactic and semantic constraints. *In Proceedings of the workshop on Temporal and spatial information processing*, volume 13, pages 1–8. Association for Computational Linguistics.

VENDLER, Z. (1957). Verbs and times. *The Philosophical Review*, 66(2):143–160.

VET, C. (2007). The descriptive inadequacy of reichenbach's tense system : A new proposal. *In* de SAUSSURE, L., MOESCHLER, J. et PUSKAS, G., éditeurs : *Tense, Mood and Aspect. Theoretical and Descriptive Issues*, volume 17 de *Cahiers Chronos*, pages 7–26. Rodopi édition.

VETTERS, C. (1996). *Temps, aspect et narration*. Numéro 106 de Faux titre. Rodopi.

VICENTE-DIEZ, M. et MARTINEZ, P. (2009). Temporal semantics extraction for improving web search. *In Database and Expert Systems Application, 2009. DEXA '09. 20th International Workshop on*, pages 69–73.

VICENTE-DÍEZ, M. T., SAMY, D. et MARTÍNEZ, P. (2008). An empirical approach to a preliminary successful identification and resolution of temporal expressions in spanish news corpora. *In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

VOORHEES, E. M. (1994). Query expansion using lexical-semantic relations. *In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 61–69, Dublin, Ireland. Springer-Verlag New York, Inc.

WEISER, S. (2010). *Repérage et typage d'expressions temporelles pour l'annotation sémantique automatique de pages Web. Application au e-tourisme.* Thèse de doctorat, Université Paris Ouest Nanterre La Défense, Paris, France.

WILDÖCHER, A. et BILHAUT, F. (2007). La plate-forme LinguaStream. *In Autour des langues et du langage : perspective pluridisciplinaire*, pages 447–454. Presses Universitaires de Grenoble.

WILMET, M. (1997). *Grammaire critique du français*. Duculot.

WILSON, G., MANI, I., SUNDHEIM, B. et FERRO, L. (2001). A multilingual approach to annotating and extracting temporal information. *In Proceedings of the workshop on Temporal and spatial information processing - Volume 13*, pages 1–7. Association for Computational Linguistics.

YAROWSKY, D. (1993). One sense per collocation. *In Proceedings of the workshop on Human Language Technology*, pages 266–271, Princeton, New Jersey. Association for Computational Linguistics.

ZADEH, L. A. (1965). Fuzzy sets. *Information and Control*, 8(3):338–353.

ZAKON, R. H. (2010). Hobbes' internet timeline (v.10). http ://www.zakon.org/robert/internet/timeline/, accédé le 17/11/2010.

ZHOU, Q. et FIKES, R. (2002). A reusable time ontology. *In Proceeding of the AAAI Workshop on Ontologies for the Semantic Web*.

ZINS, C. (2002). Models for classifying internet resources. *Knowledge organization*, 29(1):20–28.

ZIPF, G. K. (1949). *Human behavior and the principle of least effort*. Hafner, New York.

ZWEIGENBAUM, P. et CONSORTIUM MENELAS (1995). Menelas : Coding and information retrieval from natural language patient discharge summaries. *In* LAIRES, M., LADEIRA, M. et CHRISTENSEN, J., éditeurs : *Advances in Health Telematics*, pages 82–89. IOS Press.