Analyse sémantique automatique des adverbiaux de localisation temporelle : application à la recherche d'information et à l'acquisition de connaissances

Charles Teissèdre

▶ **To cite this version:**

Charles Teissèdre. Analyse sémantique automatique des adverbiaux de localisation temporelle : application à la recherche d'information et à l'acquisition de connaissances. Recherche d'information [cs.IR]. Université de Nanterre - Paris X, 2012. Français. NNT : . tel-00762440

HAL Id: tel-00762440

https://theses.hal.science/tel-00762440v1

Submitted on 7 Dec 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITE PARIS OUEST-NANTERRE LA DEFENSE

ECOLE DOCTORALE CONNAISSANCE, LANGAGE ET MODELISATION
LABORATOIRE MODYCO (MODELES, DYNAMIQUES, CORPUS) – UMR CNRS 7114

CONVENTION CIFRE N° 160/2009

THESE DE DOCTORAT

DISCIPLINE : SCIENCES DU LANGAGE
SPECIALITE : TRAITEMENT AUTOMATIQUE DES LANGUES

Analyse sémantique automatique des adverbiaux de localisation temporelle : application à la recherche d'information et à l'acquisition de connaissances

PRESENTEE PAR CHARLES TEISSEDRE

Sous la direction de Monsieur Jean-Luc Minel et Madame Delphine Battistelli

2012

Membres du jury :

Rapporteurs : Mme Nathalie Aussenac-Gilles, Directrice de recherche, IRIT (CNRS)
Mme Adeline Nazarenko, Professeur des Universités, Université Paris Nord
Examinateurs : M. Maarten de Rijke, Professeur des Universités, Universiteit van Amsterdam
M. Guy Lapalme, Professeur des Universités, Université de Montréal
Mme Florence Amardeilh, Directrice du département recherche de Mondeca
Co-Directeurs : M. Jean-Luc Minel, Professeur des Universités, Paris-Ouest Nanterre La Défense
Mme Delphine Battistelli, Maître de conférences (HDR), Université Paris Sorbonne

Remerciements

Ces quelques lignes pour remercier l'ensemble des personnes qui ont soutenu et contribué à ce projet de recherche seront bien courtes : je n'oublie pas qu'il doit beaucoup à celles et ceux qui y ont contribué et en ont fait, en un sens, un projet collectif.

Mes pensées vont bien sûr à Jean-Luc Minel, mon directeur de recherche, qui s'est toujours efforcé de réunir les conditions les plus favorables à l'avancée de ces travaux : par son soutien, son écoute, sa confiance, en m'accordant une grande autonomie, mais aussi en mobilisant des personnes et des ressources autour des problématiques liées à ces travaux, en suggérant, jamais en imposant, des méthodes de travail, des échéances progressives, des façons de valoriser et de poursuivre le travail accompli.

Mes pensées vont également à Delphine Battistelli, ma co-directrice de thèse, sans qui ce projet n'aurait pas vu le jour. Depuis son lancement jusqu'à son achèvement, à la fois au niveau des conditions de sa mise en place et au niveau scientifique, elle aura contribué à faire naître et progresser ce projet par son regard critique, par sa rigueur, par ses relectures attentives descendant jusqu'à ce qui pouvait de loin paraître des détails. A cela s'ajoute qu'elle a pris une part active pour que les travaux présentés ici puissent trouver à se poursuivre.

Je tiens à remercier aussi tout particulièrement Nathalie Aussenac-Gilles et Adeline Nazarenko, qui ont accepté d'être rapporteurs de cette thèse et m'ont grandement aidé à en améliorer la rédaction. Merci également à Maarten de Rijke et Guy Lapalme d'avoir bien voulu participer au jury de thèse.

Je remercie également chaleureusement Marcel Cori pour l'intérêt qu'il a manifesté envers ces travaux et le temps important qu'il leur a consacré.

Je tiens à remercier aussi Jean Delahousse, PDG de Mondeca au lancement de ce projet, pour son enthousiasme, ses encouragements, le dialogue constructif qu'il a engagé avec toutes les personnes impliquées et le pari audacieux qu'il a sans cesse renouvelé en faveur de la recherche au sein de l'entreprise.

L'équipe de Mondeca, du reste, a grandement contribué à ce que cette thèse prenne place dans un environnement favorable et propice à l'émulation. Merci donc à Florence Amardeilh, directrice du département recherche, qui a encadré cette thèse dans l'entreprise et a veillé à ce que les problématiques scientifiques et industrielles s'entrecroisent de façon constructive. Merci également à Pierre-Yves Vandenbussche, à Bernard Vatant, à Olivier Carloni, pour ne citer que ceux qui se sont le plus directement investis, chacun avec un regard différent, dans les problématiques liées à ces travaux.

C'est dans un environnement tout aussi favorable du côté de l'université de Paris Ouest Nanterre La Défense et du laboratoire MoDyCo qu'a pris place ce travail de thèse. Merci à celles et ceux qui m'ont aidé et soutenu, à Sofia Zidane, qui a contribué très activement au minutieux travail de

dépouillement des données pour l'évaluation de ces travaux, à Romain Loth aussi et, plus largement, à l'équipe des enseignants-chercheurs.

J'ai aussi à cœur de remercier l'équipe du L3i avec qui j'ai eu grand plaisir à travailler : Cyril Faucher, Frédéric Bertrand et Jean-Yves Lafaye.

Ces trois années de thèse, riches et intenses, où les heures et les jours de travail ont souvent débordé de leur cadre, ont aussi mêlé indirectement mes proches à ces travaux. Je les remercie pour leur patience et leur bienveillance.

Résumé

Cette thèse aborde la question de l'accès aux textes numériques, en particulier de l'accès à leur « contenu informationnel », vu sous l'angle de leur ancrage dans le temps. La visée de ces travaux est double : il s'agit de concilier une approche linguistique et une approche applicative, afin de contribuer à l'élaboration de nouveaux outils pour la fouille de textes, la recherche d'information et la gestion des connaissances - nouveaux outils en mesure de tirer parti de la sémantique d'un certain nombre d'informations temporelles exprimées dans les textes. Il s'agit ainsi à la fois de mettre en œuvre des systèmes d'interaction avec les utilisateurs et de parvenir à modéliser la sémantique des unités textuelles qui contribuent de façon saillante à l'ancrage dans le temps des situations décrites dans les textes : les *adverbiaux de localisation temporelle*.

L'étude linguistique vise ici à montrer que l'ingénierie des langues peut gagner à envisager d'une façon renouvelée les phénomènes de localisation temporelle, en cherchant à reconnaître et à mettre au jour les différentes opérations à l'œuvre dans les adverbiaux de localisation temporelle. En faisant émerger les valeurs sémantiques de ces opérations, nous montrons qu'il devient possible d'élaborer de nouveaux systèmes de recherche d'information, susceptibles de traiter des requêtes associant un critère calendaire avec un ensemble de mots-clés, telles que « *les universités au début du XIIe siècle* », par exemple. S'appuyant sur les outils développés en ce sens, on montre qu'il devient également possible d'interagir avec des données structurées décrivant des informations temporelles, à la fois pour les interroger et pour les enrichir de façon semi-automatique.

Mots-clés : Extraction d'informations temporelles ; Annotation sémantique des adverbiaux de localisation temporelle ; Recherche d'information ; Acquisition de connaissances

Abstract

This Ph. D. thesis addresses the issue of accessing the content of digital texts, in so as it is linked to the expression of temporal location. The aim of this work is twofold: it consists in conciliating a linguistic approach and an applied approach, to participate in the development of new tools for text mining, information retrieval and knowledge management - new tools that would able to take advantage of the semantics of temporal information expressed in texts. It is thus both about implementing interaction systems and modeling the semantics of one the most salient textual units that contributes to anchor in time the situations described in texts: *temporal locating adverbials*.

The linguistic analysis aims to show that language engineering can benefit from considering in a new way the phenomena of temporal location, by seeking to recognize and uncover the different operations at work in temporal locating adverbials. Pinpointing the semantic values of these operations, we show that it becomes possible to develop new Information Retrieval systems, able of processing queries involving both a calendar criterion and a set of keywords, such as "*universities in the early twelfth century*", for instance. We show that it also becomes possible to interact with structured data describing temporal information, both for search process and for data semi-automatic enrichment.

Keywords: Temporal Information Extraction; Semantic Annotation of Temporal Locating Adverbials; Information Retrieval; Knowledge Acquisition.

Table des matières

REMERCIEMENTS I

RESUME III

ABSTRACT V

TABLE DES MATIERES 1

LISTE DES FIGURES 5

LISTE DES TABLEAUX 7

CHAPITRE 1 : INTRODUCTION 9

1.1 D'ENJEUX INDUSTRIELS A DES ENJEUX SCIENTIFIQUES 9
1.2 L'ARTICULATION DU MEMOIRE DE THESE 13
1.3 LISTE DES PUBLICATIONS ASSOCIEES A CES TRAVAUX DE RECHERCHE 16

CHAPITRE 2 : LA LOCALISATION TEMPORELLE A L'ŒUVRE DANS LES TEXTES 19

2.1 NOMMER ET DESIGNER DES PERIODES DE TEMPS : DES DATES AUX EVENEMENTS ET DES EVENEMENTS AUX DATES 20
2.1.1 NOMS DE TEMPS : LES CHRONONYMES 22
2.1.2 DESIGNATIONS EVENEMENTIELLES 25
2.1.3 LES DATES FAISANT EVENEMENTS : LE CAS DES HEMERONYMES 26
2.1.4 LA REFERENCE TEMPORELLE EN QUESTION 27
2.2 TEMPS ET ENONCIATION 29
2.3 LES INDICES TEXTUELS CONTRIBUANT A DENOTER UNE OPERATION DE LOCALISATION TEMPORELLE 32
2.4 LA CLASSE DES « ADVERBIAUX ASPECTO-TEMPORELS » 36
2.5 LA CLASSE DES ADVERBIAUX DE LOCALISATION TEMPORELLE 41
2.5.1 LES DIFFERENTES FORMES POSSIBLES 41
2.5.2 POSITION DANS LA PHRASE ET PORTEE 43
2.5.3 FORMALISER LA SEMANTIQUE DES ADVERBIAUX DE LOCALISATION TEMPORELLE 45
2.6 BILAN DU CHAPITRE 47

CHAPITRE 3 : LES ADVERBIAUX DE LOCALISATION TEMPORELLE : UN OBJET POUR L'INGENIERIE DES LANGUES ET L'INGENIERIE DES CONNAISSANCES 49

3.1 TRAITEMENT AUTOMATIQUE DES LANGUES : QUEL(S) OBJET(S) D'ANALYSE ET QUEL(S) USAGE(S) ? 51
3.1.1 UN RAPIDE APERÇU HISTORIQUE 52
3.1.2 LES DIFFERENTS TYPES D'INFORMATIONS TEMPORELLES VISEES PAR LES SYSTEMES D'ANNOTATION 57
3.1.3 UN POINT SUR LES SYSTEMES D'ANNOTATION AUJOURD'HUI 63
3.2 LA MODELISATION DES PROPRIETES TEMPORELLES EN INGENIERIE DES CONNAISSANCES 68
3.2.1 LES ENTREPOTS DE DONNEES TEMPORELLES 68
3.2.2 LES VOCABULAIRES 70
3.2.3 L'INTERROGATION DES DONNEES TEMPORELLES 71
3.3 RECHERCHE D'INFORMATION ET INFORMATIONS TEMPORELLES 72
3.3.1 VARIETES DES INFORMATIONS ET VARIETES DES USAGES 72
3.3.2 EXPLOITER LES EXPRESSIONS TEMPORELLES PRESENTES DANS LES TEXTES 75
3.4 BILAN DU CHAPITRE 80

CHAPITRE 4 : LES ADVERBIAUX DE LOCALISATION TEMPORELLE : UNE PROPOSITION D'ANALYSE SEMANTIQUE 83

4.1 DEFINITION, TYPOLOGIE ET PORTEE DES ADVERBIAUX DE LOCALISATION TEMPORELLE 84
4.1.1 CONSIDERATIONS MORPHOSYNTAXIQUES ET SYNTAXIQUES 84
4.1.2 DEFINITION 85
4.2 OPERANDES ET OPERATEURS SEMANTIQUES : UNE REPRESENTATION FORMELLE DES ADVERBIAUX DE LOCALISATION TEMPORELLE 89
4.2.1 LA DEMARCHE D'ANALYSE 89
4.2.2 TYPOLOGIE DES BASES 94
4.2.3 LES ADVERBIAUX DE LOCALISATION UNAIRES 96
4.2.4 OPERATIONS DE FOCALISATION ET DE DEPLACEMENT 98
4.2.5 OPERATION DE REGIONALISATION 105
4.2.6 LES ADVERBIAUX DE LOCALISATION N-AIRES 106
4.3 BILAN DU CHAPITRE 109

CHAPITRE 5 : CALCULS SUR LES VALEURS DES ADVERBIAUX DE LOCALISATION TEMPORELLE 111

5.1 MOTIVATIONS 111
5.2 REPRESENTATION REFERENTIELLE DES ADVERBIAUX CALENDAIRES 113
5.2.1 L'ASSOCIATION D'UN INTERVALLE CALENDAIRE A UN ADVERBIAL DE LOCALISATION TEMPORELLE 113
5.2.2 LES ENSEMBLES D'INTERVALLES CALENDAIRES 123
5.3 APPLICATION DANS LE CADRE DE LA RECHERCHE D'INFORMATION SELON DES CRITERES CALENDAIRES 126
5.3.1 UN ALGORITHME POUR MESURER LA PERTINENCE DES ADVERBIAUX CALENDAIRES 127
5.3.2 LA PROBLEMATIQUE 127
5.3.3 LES CRITERES D'ADEQUATION 128
5.3.4 L'ATTRIBUTION D'UN SCORE D'ADEQUATION 129
5.3.5 ORDONNER LES REPONSES 131
5.4 BILAN DU CHAPITRE 132

CHAPITRE 6 : RESSOURCES POUR L'ACQUISITION DE CONNAISSANCES RELATIVES A LA LOCALISATION TEMPORELLE 135

6.1 UN SYSTEME D'ANNOTATION ET DE TRANSDUCTION DES ADVERBIAUX CALENDAIRES 136
6.1.1 QUEL OBJECTIF ? 136
6.1.2 ARCHITECTURE DU SYSTEME 136
6.1.3 IMPLEMENTATION DU MODELE OBJET DES ADVERBIAUX DE LOCALISATION TEMPORELLE 137
6.1.4 LE MODULE D'ANNOTATION 139
6.1.5 LE MODULE DE TRANSDUCTION VERS DES INTERVALLES CALENDAIRES 156
6.2 OUTILLER LA SAISIE D'INFORMATIONS TEMPORELLES COMPLEXES : UN CAS D'APPLICATION INDUSTRIEL 158
6.2.1 MOTIVATIONS ET ENJEUX 158
6.2.2 TKA, UN SYSTEME D'ASSISTANCE A LA SAISIE DE DATES ET HORAIRES D'OUVERTURE 160
6.3 BILAN DU CHAPITRE 168

CHAPITRE 7 : CASE, UN SYSTEME EXPERIMENTAL POUR LA RECHERCHE D'INFORMATION EXPLOITANT LES ADVERBIAUX CALENDAIRES PRESENTS DANS LES TEXTES 171

7.1 CASE : UN MOTEUR DE RECHERCHE EXPERIMENTAL 172
7.1.1 MOTIVATIONS 172
7.1.2 FONCTIONNALITES 173
7.1.3 ARCHITECTURE 177
7.1.4 FILTRAGE DES DOCUMENTS 180
7.1.5 MESURE DE LA PERTINENCE DES DOCUMENTS 181
7.1.6 LE PROBLEME DE LA VISUALISATION DES RESULTATS SUR UNE FRISE CHRONOLOGIQUE 184
7.1.7 QUELQUES-UNES DES LIMITES DU SYSTEME 184
7.2 EXPERIMENTATIONS AUTOUR DU SYSTEME CASE 186
7.2.1 LA RECHERCHE DOCUMENTAIRE ET L'EXPLORATION INTRA-DOCUMENTAIRE AVEC CASE 186
7.2.2 INTERROGER DES DONNEES STRUCTUREES : UN CAS D'UTILISATION 189
7.2.3 EN CONSULTANT, EN CONTRIBUANT : ENRICHIR DES DONNEES OUVERTES A MESURE QU'ON LES INTERROGE. UNE EXPERIMENTATION AUTOUR DES ŒUVRES D'ART DANS FREEBASE 193
7.3 BILAN DU CHAPITRE 196

CHAPITRE 8 : EVALUATION 197

8.1 LE SYSTEME D'ANNOTATION ET DE TRANSDUCTION VERS DES INTERVALLES CALENDAIRES 197
8.1.1 LE CORPUS D'EVALUATION 198
8.1.2 L'EXTRACTION 200
8.1.3 L'ANNOTATION 202
8.1.4 LA TRANSDUCTION SOUS LA FORME D'INTERVALLES CALENDAIRES 204
8.2 EVALUATION DE TKA, LE PROTOTYPE D'ASSISTANCE A LA SAISIE DE DATES ET HORAIRES D'OUVERTURE 206
8.2.1 LE PROTOCOLE D'EVALUATION 206
8.2.2 LES RESULTATS DE L'EVALUATION 209
8.3 EVALUATION DU SYSTEME DE RECHERCHE D'INFORMATION 211
8.3.1 METHODOLOGIE 211

8.3.2 EVALUATION DU MODELE D'ORDONNANCEMENT DES ADVERBIAUX CALENDAIRES 213
8.3.3 EVALUATION DU SYSTEME CASE 215
8.3.4 LIMITES DE L'EVALUATION 218

CHAPITRE 9 : BILAN ET PERSPECTIVES 221

9.1 BILAN 221
9.2 PERSPECTIVES 223
9.3 D'ENJEUX SCIENTIFIQUES A DES ENJEUX INDUSTRIELS 226

REFERENCES BIBLIOGRAPHIQUES 227

ANNEXE 1 : EXEMPLES DE REPRESENTATION DES ADVERBIAUX SOUS LA FORME D'UNE SUCCESSION D'OPERATIONS SUR UN REPERE TEMPOREL NOYAU 241

ANNEXE 2 : LE SCHEMA D'ANNOTATION CHRONOLOCATIONML 243

ANNEXE 3 : UNE FICHE DE SYNTHESE DES EVALUATIONS MANUELLES DE LA PERTINENCE D'UN ENSEMBLE D'ADVERBIAUX-CIBLES PAR RAPPORT A DES ADVERBIAUX-REQUETES 245

ANNEXE 4 : LE CORPUS DE REQUETES POUR L'EVALUATION DU SYSTEME CASE 247

Liste des figures

FIG. 1 : UNE REPRESENTATION DE L'ENONCE « HIER, IL A DIT QU'IL VIENDRAIT DEMAIN. » 30
FIG. 2 : LA SYNCHRONISATION ENTRE LE REFERENTIEL ENONCIATIF ET LE REFERENTIEL CALENDAIRE 30
FIG. 3 : UNE REPRESENTATION DE L'ENONCE « HIER, IL A DIT QU'IL VIENDRAIT DEMAIN. » 31
FIG. 4 : REPRESENTATION DE L'ENONCE « HIER, IL A DIT : "JE VIENDRAI DEMAIN.". » 32
FIG. 5 : REPRESENTATION DE LA STRUCTURE ASPECTUO-TEMPORELLE DE L'ENONCE « LA POLICE RECHERCHAIT LE COUPABLE DEPUIS TROIS JOURS » (GOSSELIN, 2005B) 35
FIG. 6 : TYPOLOGIE DES ADVERBIAUX ASPECTO-TEMPORELS (GOSSELIN, 2005B) 40
FIG. 7 : ARBRE DE DEPENDANCE SYNTAXIQUE D'UN ADVERBE DE LOCALISATION TEMPORELLE 42
FIG. 8 : COPIE D'ECRAN DU SERVICE GOOGLE VIEW:TIMELINE (11 MARS 2010) 75
FIG. 9 : COPIE D'ECRAN DE L'OUTIL TIME EXPLORER (MATTHEWS ET AL., 2010) 79
FIG. 10 : COPIE D'ECRAN DE L'OUTIL TIME-SURFER (LLORENS ET AL.,2010) 80
FIG. 11 : UNE ANALYSE SEMANTIQUE HOMOGENE 89
FIG. 12 : TYPOLOGIE DES BASES 94
FIG. 13 : LES BASES CALENDAIRES 94
FIG. 14 : LES UNITES CALENDAIRES 95
FIG. 15 : LES UNITES ORDINALES 95
FIG. 16 : DIAGRAMME D'OBJET POUR LA BASE CALENDAIRE 2E TRIMESTRE 2012 96
FIG. 17 : LE MODELE DES ADVERBIAUX DE LOCALISATION TEMPORELLE UNAIRE 97
FIG. 18 : DIAGRAMME D'OBJET POUR L'ADVERBIAL JUSQUE VERS LA FIN DES ANNEES 1920 97
FIG. 19 : DIAGRAMME D'OBJET POUR L'ADVERBIAL QUELQUES JOURS AVANT LA FIN DE SON MANDAT 98
FIG. 20 : LES OPERATIONS DE FOCALISATION ET DE DEPLACEMENT 99
FIG. 21 : LE MODELE DES DUREES 99
FIG. 22 : LA FOCALISATION GRANULAIRE 100
FIG. 23 : DIAGRAMME D'OBJET ASSOCIE A L'ADVERBIAL DEPUIS CETTE SEMAINE-LA 100
FIG. 24 : LA FOCALISATION PAR SUBDIVISION 101
FIG. 25 : DIAGRAMME D'OBJET ASSOCIE A L'ADVERBIAL QUELQUES JOURS AVANT LA FIN DU 2ND SEMESTRE 2011 101
FIG. 26 : LA FOCALISATION PAR SELECTION 101
FIG. 27 : DIAGRAMME D'OBJET ASSOCIE A L'ADVERBIAL LES 1ER ET 2E MERCREDIS DE CHAQUE MOIS 102
FIG. 28 : LE MODELE DES RANGS 102
FIG. 29 : DIAGRAMME D'OBJET DECRIVANT L'OPERATION DE FOCALISATION DANS L'ADVERBIAL LORS DES DEUX DERNIERES SEMAINES DE SON SEJOUR 103
FIG. 30 : LES OPERATIONS DE DEPLACEMENT 103
FIG. 31 : DIAGRAMME D'OBJET DECRIVANT L'OPERATION DE DEPLACEMENT DANS L'ADVERBIAL PRES DE TROIS JOURS AVANT 104
FIG. 32 : DIAGRAMME D'OBJET ASSOCIE A L'ADVERBIAL IL Y A UN MOIS 104
FIG. 33 : DIAGRAMME D'OBJET ASSOCIE A L'ADVERBIAL LE MOIS DERNIER 105
FIG. 34 : L'OPERATION DE REGIONALISATION 105
FIG. 35 : DIAGRAMME D'OBJET ASSOCIE A L'ADVERBIAL DEPUIS BIENTOT TROIS SEMAINES 106
FIG. 36 : LES TYPES D'ADVERBIAUX DE LOCALISATION TEMPORELLE 106
FIG. 37 : LES ADVERBIAUX BINAIRES 107
FIG. 38 : DIAGRAMME D'OBJET ASSOCIE A L'ADVERBIAL DE FIN MARS A MI-JUIN 107
FIG. 39 : LES ADVERBIAUX COMPOSES 108
FIG. 40 : DIAGRAMME D'OBJET ASSOCIE A L'ADVERBIAL DU LUNDI AU SAMEDI, DE 10H A 18H, SAUF LES 1ER ET 8 MAI 109
FIG. 41 : LE PROCESSUS FORMEL DE RECHERCHE 127
FIG. 42 : ILLUSTRATION DES CRITERES D'ADEQUATION ENTRE UNE REQUETE Q ET DIFFERENTES REPONSES POSSIBLES 128
FIG. 43 : ARCHITECTURE DU SYSTEME D'ANNOTATION ET DE TRANSDUCTION DES ADVERBIAUX DE LOCALISATION TEMPORELLE 137

FIG. 44 : IMPLEMENTATION DU MODELE OBJET DES ADVERBIAUX DE LOCALISATION TEMPORELLE (VUE SIMPLIFIEE) 138
FIG. 45 : LE MODULE D'ANNOTATION 140
FIG. 46 : ARCHITECTURE DU MODULE D'ANNOTATION (VUE SIMPLIFIEE) 141
FIG. 47 : DEUX EXEMPLES D'ENONCES OU L'ON VISUALISE LES SEGMENTS ANNOTES AVEC TIMEML (EN BLEU) ET CEUX ANNOTE AVEC CHRONOLOCATIONML (EN ROUGE) 147
FIG. 48 : COPIE D'ECRAN D'UN GRAPHE UNITEX POUR L'ANNOTATION DES ADVERBIAUX DE LOCALISATION TEMPORELLE 151
FIG. 49 : COPIE D'ECRAN D'UN SOUS GRAPHE UNITEX POUR L'ANNOTATION DES ADVERBIAUX DE LOCALISATION TEMPORELLE 152
FIG. 50 : COPIE D'ECRAN D'UN SOUS GRAPHE UNITEX POUR L'ANNOTATION DES ADVERBIAUX DE LOCALISATION TEMPORELLE 152
FIG. 51 : ARCHITECTURE DU MODULE DE TRANSDUCTION 157
FIG. 52 : EXEMPLE DE FORMULAIRE DE SAISIE DE DATES ET HORAIRES D'OUVERTURE/FERMETURE 160
FIG. 53 : ARCHITECTURE DU SERVICE WEB TKA 160
FIG. 54 : LES CLASSES DE L'ONTOLOGIE CHRONEX.OWL 162
FIG. 55 : L'INTERFACE UTILISATEUR DE TKA 165
FIG. 56 : LA ZONE DE SAISIE EN LANGAGE NATUREL DES PERIODES D'ACCES 166
FIG. 57 : VISUALISATION CORRESPONDANT A LA PERIODE D'ACCES « OUVERT DU MARDI AU SAMEDI, DE 10H A 19H ET LE DIMANCHE, DE 13H A 19H. FERME LES 1ER ET 8 MAI. » POUR LA SEMAINE DU 30 AVRIL AU 6 MAI 2012 167
FIG. 58 : PANNEAU DE CONTROLE 167
FIG. 59 : L'INTERFACE DU SYSTEME CASE (REQUETE « LAÏCITE VERS 1905 ») 174
FIG. 60: LA ZONE DE SAISIE DES REQUETES 174
FIG. 61 : VISUALISATION CHRONOLOGIQUE DES RESULTATS D'UNE RECHERCHE 175
FIG. 62: VISUALISATION CHRONOLOGIQUE DES RESULTATS D'UNE RECHERCHE : LA FENETRE D'AFFICHAGE D'UN RESULTAT 175
FIG. 63 : LA ZONE D'AFFICHAGE DE LA LISTE DES RESULTATS 176
FIG. 64 : LES COMPOSANTS D'INDEXATION ET DE RECHERCHE 178
FIG. 65 : ARCHITECTURE DU SERVICE WEB DE RECHERCHE 179
FIG. 66 : FILTRAGE DES PLUS PROCHES VOISINS DANS L'INDEX 180
FIG. 67 : COPIE D'ECRAN DE LA LISTE DES RESULTATS PROPOSES POUR LA REQUETE « UNIVERSITE AU DEBUT DU XIIE SIECLE » 188
FIG. 68 : COPIE D'ECRAN DE LA LISTE DES RESULTATS PROPOSES POUR LA REQUETE « UNIVERSITE AU DEBUT DU XIIE SIECLE » DANS LE CADRE D'UNE RECHERCHE INTRA-DOCUMENTAIRE 188
FIG. 69 : LA REQUETE MODIFIEE SUITE A UN DEPLACEMENT DE LA FRISE CHRONOLOGIQUE : L'ADVERBIAL « DES ANNEES 1530 AUX ANNEES 1570 » A ETE GENEREE AUTOMATIQUEMENT 189
FIG. 70 : COPIE D'ECRAN DE LA LISTE DES RESULTATS PROPOSES SUITE AU DEPLACEMENT DE LA FRISE CHRONOLOGIQUE 189
FIG. 71 : EXEMPLE DE GRAPHE RDF REPRESENTANT UN EVENEMENT 190
FIG. 72 : ARCHITECTURE DU SYSTEME D'INDEXATION DES EVENEMENTS 191
FIG. 73 : COPIE D'ECRAN DES RESULTATS DU MOTEUR POUR LA REQUETE « ROCK IN LONDON IN AUGUST 2008 » 192
FIG. 74 : COPIE D'ECRAN DE L'INTERFACE D'INTERROGATION DES ŒUVRES D'ART DE FREEBASE 194
FIG. 75 : SUGGESTION D'AJOUT D'UNE PROPRIETE TEMPORELLE DANS FREEBASE 195
FIG. 76 : DIAGRAMME D'OBJET ASSOCIE A L'ADVERBIAL HIER 241
FIG. 77 : DIAGRAMME D'OBJET ASSOCIE A L'ADVERBIAL CE JOUR-LA 241
FIG. 78 : DIAGRAMME D'OBJET ASSOCIE A L'ADVERBIAL DURANT LES DEUX DERNIERES SEMAINES DE LA CAMPAGNE ELECTORALE 242

Liste des tableaux

TABLEAU 1 : SCORES DES REPONSES POUR LA REQUETE EN 1980 131
TABLEAU 2 : EXEMPLE D'ORDONNANCEMENT D'UNE LISTE DE REPONSES POUR LA REQUETE DEPUIS 1980 132
TABLEAU 3 : LA COMPOSITION DU JEU DE DONNEES DE DERIVE 190
TABLEAU 4 : LA REPARTITION DES ADVERBIAUX UNAIRES DANS LE CORPUS D'EVALUATION 199
TABLEAU 5 : LA REPARTITION DES ADVERBIAUX N-AIRES DANS LE CORPUS D'EVALUATION 199
TABLEAU 6 : RESULTATS DE L'EVALUATION DU SYSTEME POUR LA RECONNAISSANCE DES ADVERBIAUX DE LOCALISATION TEMPORELLE (METHODE 1) 201
TABLEAU 7 : RESULTATS DE L'EVALUATION DU SYSTEME POUR LA RECONNAISSANCE DES ADVERBIAUX DE LOCALISATION TEMPORELLE (METHODE 2) 201
TABLEAU 8 : RESULTATS DE L'EVALUATION DU SYSTEME POUR L'ANNOTATION DES ADVERBIAUX DE LOCALISATION TEMPORELLE 203
TABLEAU 9 : RESULTATS DE L'EVALUATION DU SYSTEME POUR L'ANNOTATION DES RELATIONS ENTRE ADVERBIAUX DE LOCALISATION TEMPORELLE 203
TABLEAU 10 : RESULTATS DE L'EVALUATION DU SYSTEME POUR LA TRANSDUCTION DES ADVERBIAUX DE LOCALISATION TEMPORELLE VERS DES INTERVALLES CALENDAIRES 205
TABLEAU 11 : RESULTATS DE L'EVALUATION DU SYSTEME D'ASSISTANCE A LA SAISIE DE DATES ET HORAIRES D'OUVERTURE SUR L'ENSEMBLE DU CORPUS 210
TABLEAU 12 : RESULTATS DE L'EVALUATION DU SYSTEME D'ASSISTANCE A LA SAISIE DE DATES ET HORAIRES D'OUVERTURE SUR UN SOUS ENSEMBLE DU CORPUS OU LES EXPRESSIONS TESTEES CONTIENNENT AU MOINS UNE RELATION DE COMPOSITION ENTRE ADVERBIAUX 210
TABLEAU 13 : EXTRAIT D'UNE SYNTHESE D'UN CORPUS SOUMIS A DEUX EVALUATEURS POUR L'EVALUATION DU MODULE D'ORDONNANCEMENT DES ADVERBIAUX CALENDAIRES 214

Chapitre 1 : Introduction

1.1 D'enjeux industriels à des enjeux scientifiques

L'irruption des technologies de l'information et de la communication s'est accompagnée d'un accroissement considérable des ressources numériques disponibles, produites et consultées, dans les entreprises, les institutions et sur le Web. L'explosion quantitative de l'information, l'accumulation documentaire, qui continue encore d'aller croissant, amène à s'interroger sur les différentes stratégies à mettre en place pour accéder aux ressources numériques pertinentes pour un besoin donné. Ces nouvelles technologies, parallèlement, viennent remettre en question le statut établi du document : les supports numériques (physiques et logiques) sont de plus en plus variés et contribuent à renouveler le rapport qu'ont les utilisateurs/lecteurs avec le « contenu » des documents. Les interactions que permettent ces supports influent en effet directement sur leur réception.

Cherchant à répondre aux difficultés de l'accumulation documentaire qu'ils ont indirectement contribué à favoriser en permettant de rechercher dans de très vastes ensembles de ressources numériques, les systèmes de recherche d'information élargissent progressivement les fonctions d'interrogation et de filtrage qu'ils proposent. La plupart des outils grand-public offre ainsi par exemple des services de filtrage sur le type de documents (texte, images, vidéos, formats, etc.). Des critères de plus en plus nombreux sont pris en compte pour mesurer la pertinence d'un document : les moteurs de recherche s'appuient par exemple sur l'analyse des archives de requêtes, afin de distinguer plusieurs profils de requêtes et privilégier des documents « frais », c'est-à-dire publiés récemment sur le Web, lorsque les requêtes semblent porter sur un sujet d'actualité. Il y a toutefois peu de réalisations opérationnelles permettant de prendre en charge les informations temporelles ; non pas tant celles *autour* des documents (comme leur date de publication), mais celles exprimées *dans* les documents (Alonso *et al.*, 2007). Il y a également peu de réalisations opérationnelles sur la recherche intra-documentaire ou la navigation au sein d'un même document, - c'est un constat général - et, plus spécifiquement aussi, pour ce qui va nous retenir, peu de réalisations pour la

recherche intra-documentaire, vue à travers le prisme de la représentation chronologique qu'il est possible de construire à partir d'un texte.

Sous un autre angle, celui de la représentation des connaissances, dont les approches continuent d'évoluer avec l'émergence d'un Web de données (nous reviendrons sur l'enjeu des données ouvertes), des difficultés émergent lorsqu'il s'agit de représenter des informations temporelles.

De ce double constat vient l'intérêt manifesté par Mondeca et MoDyCo, respectivement l'entreprise et le laboratoire qui ont soutenu cette thèse CIFRE, d'entreprendre un projet de recherche à la croisée de l'ingénierie des langues et de l'ingénierie des connaissances, autour des informations temporelles.

Le croisement de ces deux disciplines, à l'articulation des sciences du langage et des sciences informatiques, est largement sous-jacent à la problématique industrielle sur laquelle Mondeca s'est positionné et qui forme l'un de ses axes stratégiques de développement. L'un de ces enjeux industriels est de faciliter la description du contenu des documents numériques, pour produire des données, des connaissances structurées, susceptibles d'être interrogées de façon riche, soit pour elles-mêmes, soit encore pour améliorer l'accès aux documents. De ce point de vue, les deux domaines que sont l'ingénierie des langues et l'ingénierie des connaissances ont partie liée avec celui de la recherche d'information, qui place au cœur du problème la question de l'exposition des connaissances et des façons d'interroger et de parcourir de vastes ensembles de ressources numériques.

Mondeca développe une série logicielle autour de la gestion de données structurées (*ITM* et *Smart Content Factory*), afin de valoriser des contenus à l'aide de technologies dites « sémantiques ». L'amélioration de la production et de l'accès aux ressources numériques et plus largement la gestion globale de la connaissance sont aujourd'hui des enjeux industriels et stratégiques majeurs. La naissance même du Web s'est accompagnée, chez ses concepteurs, d'une réflexion sur l'intérêt d'en faire un Web « sémantique », qui viserait à en rendre les ressources accessibles et manipulables, pour ses utilisateurs certes, mais aussi pour les agents logiciels. A cette fin, Tim Berners-Lee défendait dès 1994 l'intérêt d'un système de métadonnées formelles[1]. Au regard des technologies du Web sémantique au cœur de la suite logicielle développée par Mondeca, la temporalité occupe une position transverse, car elle n'est pas liée à un domaine ou un secteur d'activité en particulier. Dans la gestion des connaissances, les difficultés propres aux données temporelles relèvent aussi bien de leur modélisation, de leur extraction dans des corpus de documents, des raisonnements à mettre en œuvre pour lier entre elles ces informations, que de leur indexation en vue de proposer aux utilisateurs de nouveaux services de recherche et de navigation.

Sur un plan industriel, l'objectif de ce projet de recherche est ainsi d'ajouter à la suite logicielle développée par Mondeca des ressources permettant l'extraction d'informations temporelles dans les textes. Au niveau de la restitution des données, un des objectifs est de proposer de nouveaux outils pour l'accès aux informations temporelles exprimées par les textes. L'hypothèse qui sous-tend ces travaux est que la perspective linguistique est susceptible d'améliorer non seulement les ressources

[1] http://www.w3.org/Talks/WWW94Tim/

dédiées à l'annotation temporelle des textes – ce que l'on sait pouvoir en attendre -, mais aussi de renouveler les représentations formelles de la temporalité pour la représentation des connaissances – ce qui n'allait pas de soi et qui, rétrospectivement, aura été une des retombées inattendues de ce projet. On se sera ainsi aperçu qu'il peut être utile (1) de faire coexister deux modes de représentation, l'un qui découle directement de la manière dont on exprime dans la langue la localisation temporelle, l'autre associé aux standards normés de représentation des dates et (2) de faire dépendre des besoins applicatifs la façon dont on transpose la représentation des expressions langagières de localisation temporelle vers un modèle normé.

Un exemple simple permet d'illustrer un des apports de notre démarche. L'étude d'un objet archéologique ou d'une œuvre d'art ne permet pas toujours de les dater de façon très précise. Ainsi, les études du tableau *Saint Jean Baptiste* du Caravage semblent permettre d'estimer qu'il a été peint *vers 1600*. Sur les deux versions de *Saint Jérôme écrivant* par exemple, les textes se montrent également hésitants : les articles de Wikipédia qui leur sont consacrés précisent que la première version a été peinte *vers 1605-1606* ou bien encore, peut-être, *entre 1602 et 1604,* et la seconde *en 1607 ou 1608*. Ce type d'incertitude ne peut pas être exprimée avec les standards normés des dates (comme le standard le plus utilisé qu'est l'ISO 8601). En effet, la standardisation de l'expression des dates a précisément eu pour objectif de chasser toute « indétermination » dans leur représentation. Très contraints, ces modèles de représentation pour la datation posent des difficultés lorsqu'il s'agit de localiser un objet aux alentours d'une date (*vers 1600* ou *au début du XVIIe siècle*, par exemple) ou de manipuler des informations de granularité variable. Or, pour un archéologue, un historien, un journaliste, qui s'intéresse à des périodes de temps plus ou moins étendues, pouvoir jongler avec des granularités plus ou moins fines (jour, mois, décennie, siècle, etc.) et exprimer des dates aux contours imprécis est souvent une nécessité. D'autres exemples, plus simples encore ne peuvent pas être exprimés de façon satisfaisante dans les standards actuels : des expressions comme « dans les années 80 » ou « à la fin du XIIe siècle » n'ont pas de transposition directe dans les normes ISO de représentation des dates. On verra ainsi, à travers des exemples d'énoncés principalement issus de corpus de presse et de dépêches de l'Agence France Presse, que la représentation en langue de la localisation temporelle est beaucoup plus expressive que ne le sont les standards actuels.

Ce projet de recherche aborde donc en premier lieu la question de l'accès aux textes numériques, en particulier de l'accès à leur « contenu informationnel » en tant qu'il a partie liée à l'expression de la temporalité dans les textes. La visée du projet est double : il s'agit de concilier une approche conceptuelle et opératoire, du point de vue linguistique, et une approche applicative, dont l'objectif est de participer à l'élaboration de nouveaux outils pour la fouille de textes, la recherche d'information et la gestion des connaissances. Il s'agit ainsi à la fois de mettre en œuvre des systèmes d'interaction avec les utilisateurs et de parvenir à modéliser et représenter la sémantique des unités textuelles qui contribuent de façon saillante à la localisation temporelle dans les textes : les *adverbiaux de localisation temporelle*. Ces adverbiaux débordent largement les références temporelles que sont les dates, au sens où ils intègrent des prépositions (*avant fin 2011*), des locutions prépositionnelles (*à partir du 15 juillet*), mais aussi parce qu'ils recouvrent également des références non calendaires, liées à l'énonciation (*depuis l'an dernier*) ou liées à des « événements », qui, d'un point de vue morphosyntaxique, peuvent être exprimés par des noms ou des verbes (*depuis les élections ; jusqu'à ce que le puits de forage de BP soit déclaré "mort"*).

Dans ce cadre, la mise en œuvre de systèmes d'interaction avec les utilisateurs pour la recherche d'information et la gestion des connaissances suppose l'élaboration de modèles qui prennent en compte les usages envisagés et qui puissent représenter et rendre finement compte de divers types de structurations textuelles telles que les fait apparaître l'analyse linguistique de la temporalité. Le projet vise ainsi à contribuer au renouvellement de la ou des lecture(s) possible(s) des textes grâce au support numérique, via des critères liés à l'expression de la temporalité, en particulier de l'opération de *localisation temporelle* – opération qui contribue à ancrer sur un référentiel temporel les situations décrites dans les textes.

La représentation du temps telle qu'elle trouve à s'exprimer dans les textes comporte en outre inévitablement une dimension sociale, politique que la représentation calendaire, qui s'appuie sur un outil de mesure et de localisation issu d'un effort de rationalisation, ne saurait résumer, même si le calendrier comporte lui-aussi, dans l'histoire de sa constitution et de sa diffusion, les traces évidentes d'une composante politique. Un fil court du temps mesurable à l'histoire, dans lequel on peut lire la dynamique par laquelle se construisent, à travers les textes notamment, les représentations sociales du temps. Sur ce fil tient l'intérêt qui motive cette étude dont on espère qu'elle alimente la réflexion sur la production de nouveaux outils et de nouveaux modes d'accès au(x) texte(s), en s'efforçant de mettre en lumière l'articulation entre deux représentations : la représentation calendaire du temps d'une part et celle sous-jacente à l'expression de la localisation temporelle dans les textes d'autre part.

La perspective linguistique vient supporter, dans ce projet de recherche, la perspective applicative, en permettant de rassembler des éléments susceptibles de formaliser les opérations de localisation temporelle à l'œuvre dans les textes à travers l'analyse des adverbiaux. Les systèmes d'interaction que ces travaux visent à mettre en œuvre s'appuient sur ces opérations pour proposer de nouveaux modes d'accès à l'information, au sein d'un corpus ou d'un texte.

Nos travaux se situent à la croisée de deux grandes perspectives, qui chacune a donné lieu à des développements complémentaires : (1) la perspective linguistique, pour une part, et (2) la perspective de l'ingénierie des langues et des connaissances, d'autre part.

(1) Le problème de la localisation temporelle dans les textes est vu d'abord à travers le prisme linguistique, notamment de la linguistique énonciative et textuelle : comment les textes sont-ils structurés temporellement et avec quels phénomènes cette structuration temporelle interagit-elle ? Il est vu ensuite à travers le prisme de la linguistique formelle : il s'agit de voir quels modèles de représentation formelle adopter pour appréhender les opérations sémantiques à l'œuvre dans les adverbiaux de localisation temporelle. Au demeurant, l'analyse des indices textuels contribuant à ancrer temporellement les situations décrites dans les textes oppose aujourd'hui encore aux traitements automatiques une certaine résistance, en grande partie parce que ces indices recouvrent de nombreuses catégories morphosyntaxiques et, plus largement, une grande variété d'éléments linguistiques, qui en font un objet difficile à modéliser.

(2) L'étude linguistique vise aussi à montrer que l'ingénierie des langues peut gagner à envisager d'une façon renouvelée les phénomènes de localisation temporelle, en particulier pour l'annotation automatique des textes. La représentation formelle que l'on propose consiste à

décrire les adverbiaux de localisation temporelle sous la forme d'une succession d'opérations portant sur une référence temporelle noyau. En articulant cette représentation des adverbiaux, qui découle d'une analyse linguistique, et une représentation dite « référentielle » qui associe aux adverbiaux des valeurs calendaires, nous montrerons qu'il devient possible d'élaborer de nouveaux systèmes de recherche d'information, susceptibles de traiter des requêtes contenant des critères calendaires (en particulier, des requêtes associant un critère calendaire avec un ensemble de mots-clés, telles que « *la mode dans les années 80* »). Enfin, s'appuyant sur les outils que nous avons développés en ce sens, on montrera qu'il devient également possible d'interagir avec des données structurées, des bases de connaissances, à la fois pour les interroger et pour les enrichir de façon semi-automatique.

1.2 L'articulation du mémoire de thèse

Le chapitre 2 aborde la question de la localisation temporelle dans les textes dans une perspective linguistique et dresse un état de l'art des travaux relatifs à l'analyse des indices textuels contribuant à ancrer dans le temps les situations décrites dans les textes. Parmi ces indices, les adverbiaux de localisation temporelle occupent une place privilégiée en ce sens qu'ils peuvent effectuer un repérage aussi bien par rapport au référentiel du calendrier (*dès le début du XXe siècle*), un repérage relatif à un procès, éventuellement nominalisé (*depuis le début de la campagne électorale*), ou encore par rapport à l'instance de la parole (*dès demain*). On montrera que les liens entre les procès et les adverbiaux de localisation temporelle sont complexes, dans la mesure où ces derniers ne font que contribuer à les ancrer dans le temps. En outre, on montrera qu'il ne faut pas s'attendre à ce qu'un calcul systématique de l'ancrage, sur un calendrier, des situations décrites dans les textes livre une représentation fine de la structure temporelle des textes.

Le chapitre 3 aborde cette question à travers le prisme de l'ingénierie des langues et des connaissances et dresse un état de l'art des systèmes applicatifs et des ressources développées pour annoter les entités nommées que sont les dates ou encore les « expressions temporelles » dans les textes – pour reprendre les termes les plus souvent retenus - et les exploiter ensuite dans le cadre de l'ingénierie des connaissances et de la recherche d'information. Dans nombre de ces travaux, le rapport des adverbes de localisation temporelle à ce qu'ils déterminent, la question donc de leur *portée*, n'est pas véritablement prise en compte. Relevant davantage d'une approche « pragmatique » du problème que d'une approche linguistique, ils visent à expliciter le lien entre des « événements » - une notion qui soulève un problème définitoire souvent évacué - et des dates ou des durées. Le problème est que, ce faisant, est aussi évacuée la différence qu'il y a entre les adverbiaux de localisation temporelle dans les textes et les dates, dont le modèle découle de la représentation calendaire. Ce type d'approche suppose implicitement que l'on peut toujours associer une date à un adverbial de localisation temporelle. Ce sont pourtant deux modes d'indexation temporelle distincts, entre lesquels il n'y a pas toujours de transposition directe possible. Leur articulation doit donc être pensée comme étant problématique.

Parallèlement à ces travaux d'ingénierie des langues, la localisation temporelle constitue également une problématique en ingénierie des connaissances. Avec l'avènement des Linked Open Data

(réseaux de données « ouvertes », publiques et librement accessibles), de plus en plus de données structurées sous l'angle temporel, localisées dans le temps, sont accessibles, partagées, pour certaines enrichies de façon collaborative. Malgré cette multiplication des données, la prise en compte de la temporalité pose encore bien des difficultés en ingénierie des connaissances : (1) au niveau de la modélisation d'abord (comment représenter des informations temporelles de natures très variées), (2) au niveau des données elles-mêmes ensuite (comment faciliter l'enrichissement des Linked Open Data ? comment traiter des données parfois incomplètes ?), et enfin (3) au niveau de l'exploitation de ces données pour la recherche d'information (quels systèmes mettre en œuvre pour pouvoir traiter des requêtes contenant des critères temporels ?).

Pour ce qui regarde la recherche d'information, à l'heure actuelle, les moteurs de recherche ne tirent pas parti des expressions de localisation temporelle présentes dans les textes. Des expériences originales montrent néanmoins l'intérêt de cette démarche.

Le chapitre 4 vise à formaliser l'objet d'analyse qui nous retient en proposant une analyse sémantique des adverbiaux de localisation temporelle. A la suite des travaux de (Battistelli, 2009), on montrera qu'il est possible de représenter ces adverbiaux sous la forme d'une succession d'opérations sur un repère temporel noyau (opérations de régionalisation, de focalisation et de déplacement). Cette analyse s'attache à décrire les adverbiaux de localisation temporelle uniquement en fonction des unités linguistiques qui les composent (éventuellement de celles présentes dans leur contexte), indépendamment de la valeur calendaire que l'on peut associer au repère temporel noyau, indépendamment donc de ses « coordonnées temporelles ». Dans ce cadre d'analyse, que l'on fait nôtre et que l'on cherche à étendre, les adverbiaux de localisation temporelle sont traités uniformément, quelle que soit la nature de la référence au cœur de l'adverbial : qu'il s'agisse d'une base calendaire (*depuis 1987*, *peu après les années 20*), d'une base relative à l'énonciation (*hier*) ou à un procès (*quelques jours après les élections*). On étend également l'analyse pour pouvoir décrire les liens entre des adverbiaux composés (*tous les jours sauf le dimanche, de 9h à 19h, de mars à juillet*).

Dans le chapitre 5, nous présentons un processus formel permettant de transposer la représentation des adverbiaux qui découle d'une analyse linguistique vers une représentation sous la forme d'intervalles calendaires. Les intervalles calendaires présentent des propriétés calculatoires intéressantes pour les applications que l'on cherche à développer. On cherche ainsi à articuler le problème de la description des adverbiaux de localisation temporelle avec celui qui consiste à leur associer des valeurs calendaires pour des besoins applicatifs précis. En effet, les relations entre intervalles de temps ont des propriétés bien étudiées et formalisées en Intelligence Artificielle (Allen, 1981 ; Allen, 1983). Pour autant, nous montrerons qu'il peut être intéressant pour les applications de recherche d'information de ne pas s'en tenir à une description des relations entre intervalles en termes de relations dites d'Allen (inclusion, précédence, succession, chevauchement, etc.), mais de chercher à établir des mesures de similarité entre différents intervalles.

Nous décrivons ainsi une heuristique permettant d'établir des mesures de similarité entre intervalles calendaires. Cette mesure de similarité est exploitée dans le système expérimental de recherche d'information que l'on présentera. Ce système cherche à traiter des requêtes contenant l'expression de critères calendaires. La mesure de similarité vise ainsi à affecter un score de pertinence à

intervalle candidat (une réponse possible) par rapport à un intervalle source (une requête) : dit autrement, cette mesure vise à déterminer si un intervalle donné est pertinent par rapport à une « question » posée sous la forme d'un intervalle calendaire. Par exemple, pour un utilisateur qui s'intéresserait à ce qui, dans un corpus, est associé avec « *la fin du XVe siècle* », cette mesure vise à établir qu'un adverbial comme « *en 1497* » est plus « pertinent » que l'adverbial « *en 1430* ». Il s'agit donc de filtrer et de trier par pertinence un ensemble d'adverbiaux calendaires candidats par rapport à un adverbial formant une requête, en raisonnant à partir d'une représentation sous la forme d'intervalles calendaires.

Le chapitre 6 décrit les ressources et les applications que nous avons développées pour annoter les adverbiaux de localisation temporelle présents dans les textes et les transformer en objets manipulables par des agents logiciels. Nous présentons le schéma d'annotation permettant de décrire la sémantique des adverbiaux repérés dans les textes, ainsi que les ressources que nous avons développées pour automatiser leur annotation. Précisons que ces ressources visent en premier lieu les adverbiaux qui opèrent un ancrage sur le calendrier (ces adverbiaux dits *calendaires*, tels que « *depuis le début du XIXe siècle* », « *en 1920* », « *peu après les années 30* », sont une sous-catégorie des adverbiaux de localisation temporelle). On verra en effet que les adverbiaux dont le noyau est formé par une base non calendaire appellent des traitements spécifiques et présentent une plus grande complexité pour les traitements automatiques.

Une fois annotés, les adverbiaux calendaires extraits par le système sont transformés sous forme d'intervalles calendaires, à partir desquels il est possible de mettre en œuvre des raisonnements. Cependant, l'annotation automatique des textes étant un processus par nature imparfait, il arrive que les informations annotées ne le soient pas correctement ou pas complètement. L'acquisition de « connaissances » à partir des textes invite ainsi à réfléchir à des systèmes d'interactions permettant d'effectuer des contrôles sur les informations annotées. Nous illustrons cette problématique en présentant une expérimentation qui repose toujours sur cette idée : l'intérêt d'articuler les deux modes de représentation de la localisation temporelle, celle exprimée en langue et celle, normalisée, du calendrier. Cela permet de recourir à plusieurs stratégies, plusieurs interactions avec les utilisateurs, plusieurs manières de présenter une information, en passant d'une représentation pivot vers différents autres formats. On montrera ainsi à travers une expérimentation comment ces ressources peuvent être exploitées pour l'acquisition de connaissances, à travers un cas d'application industriel visant à faciliter la saisie, dans une base de connaissances, d'informations relatives à des dates et horaires d'ouverture.

Le chapitre 7 décrit un système de recherche d'information pour la recherche documentaire et la recherche intra-documentaire. Nous présenterons ainsi un moteur de recherche expérimental, le système CaSE (*Calendar Search Engine*), un prototype qui montre ce que pourrait être un moteur de recherche tirant parti des adverbiaux calendaires présents dans les textes. Le système CaSE est tout à la fois un moteur de recherche et un outil de visualisation des informations calendaires présentes dans un ensemble de documents. L'interface présente une frise chronologique au-dessus de la liste des documents formant la réponse à une requête. Il permet ainsi la fouille d'un corpus ou d'un texte sous l'angle calendaire. Ce prototype de moteur de recherche permet de traiter des requêtes associant des mots-clés et des critères calendaires : *la laïcité avant 1905*, *la peine de mort depuis les années 70*, etc.

Nous montrons à travers trois expérimentations que ce système peut prendre en charge non seulement des textes, mais aussi des données structurées, afin d'interroger et de faciliter l'enrichissement de bases de connaissances.

Nous présentons, dans le chapitre 8, une évaluation des différentes ressources et applications développées pour le traitement des adverbiaux de localisation temporelle dans les textes, en nous efforçant de montrer leur intérêt et leurs limites.

Nous présentons ainsi une évaluation des ressources pour le français et pour l'anglais permettant d'annoter automatiquement les adverbiaux calendaires dans les textes. Nous présentons également une évaluation du modèle de sélection et de tri par pertinence des adverbiaux calendaires pour la recherche d'information. L'évaluation montre que l'algorithme de sélection et de tri obtient des résultats proches de ceux obtenus par un tri effectué manuellement. Enfin, nous présenterons une évaluation plus générale du moteur CaSE, qui s'appuie sur ces deux composants (d'annotation et de tri). L'évaluation montre que, même avec un modèle de pertinence rudimentaire pour le traitement des mots-clés, le système produit des résultats encourageants qui confortent la démarche adoptée.

Le chapitre 9 dresse un bilan des acquis de ce projet de recherche et des perspectives qu'il a contribué à ouvrir. Le mémoire s'achèvera ainsi sur une synthèse critique des résultats obtenus à l'issue de ce projet de recherche, en montrant en quoi il a contribué à sa manière, à travers l'étude d'un phénomène linguistique précis, à la réflexion sur l'intérêt d'articuler les différentes disciplines que sont la linguistique, l'ingénierie des langues et l'ingénierie des connaissances. Nous évoquerons également plusieurs des limites de ces travaux, ainsi que les perspectives sur lesquelles ils pourraient ouvrir.

1.3 Liste des publications associées à ces travaux de recherche

Communications avec actes dans un congrès international

Battistelli D., Cori M., Minel J.-L. et **Teissèdre C.** (2012). Information Retrieval: Ranking Results according to Calendar Criteria. In *Proceedings of IPMU 2012*, July 9-13 2012, Catania, Italy (à paraître).

Vandenbussche P.-Y. et **Teissèdre C.** (2011). Events Retrieval Using Enhanced Semantic Web Knowledge. In *Proceedings of ISWC Workshop DeRiVE2011* (Challenge paper), October 23-27 2011, Bonn, Germany, pp. 112-116.

Battistelli D., Cori M., Minel J.-L. et **Teissèdre C.** (2011). Semantics of Calendar Adverbials for Information Retrieval. In *Proceedings of ISMIS 2011*, June 28-30 2011, Warsaw, Poland, pp. 622-631.

Faucher C., **Teissèdre C.**, Lafaye J.Y. et Bertrand F. (2010). Temporal Knowledge Acquisition and Modeling, In *Proceedings of EKAW 2010 (17th International Conference on Knowledge Engineering*

and Knowledge Management), 11-15 October, Lisbon (Portugal), volume 6317 of LNCS (LNAI), Springer-Verlag, pp. 271-280 (article court).

Teissèdre C., Battistelli D. et Minel J-L. (2010). Resources for Calendar Expressions Semantic Tagging and Temporal Navigation through Texts. In *Proceedings of LREC 2010 (7th International Conference on Language Resources and Evaluation)*, Malta. 19-21 mai, pp. 3572-3577.

Communications avec actes dans un congrès national

Teissèdre C., Battistelli D. et Minel J.-L. (2011). Recherche d'information et temps linguistique : une heuristique pour calculer la pertinence des expressions calendaires. In *Actes de TALN 2011*, 27 juin-1er juillet 2011, Montpellier, pp. 161-172.

Teissèdre C., Battistelli D. et Minel J.-L. (2010). Du texte au portail sémantique : cas d'utilisation lié à des données temporelles. In *Actes d'IC'2010 - 21es Journées Francophones d'Ingénierie des Connaissances*, Nîmes, 9-11 juin 2010, pp. 209-220.

Faucher C., Lafaye J.Y., Bertrand F. et **Teissèdre C.** (2010). Modélisation et reformulation d'expressions temporelles extraites de textes en langage naturel. In *Actes d'AFADL 2010 (Approches Formelles dans l'Assistance au Développement de Logiciels)*, 9-11 Juin, Poitiers (France), pp. 213-216 (article court).

Communications par affiche dans un congrès international ou national

Bittar A., Hagège C., Moriceau V. Tannier X. et **Teissèdre C**. (2012). Temporal Annotation: A Proposal for Guidelines and an Experiment with Inter-annotator Agreement. In *Proceedings of LREC 2012*, 21-27 May 2012, Istanbul, Turkey (à paraître).

Chapitre 2 : La localisation temporelle à l'œuvre dans les textes

Ce chapitre s'attache à décrire la façon dont la linguistique s'est emparée de la question de la localisation temporelle dans les textes : quels sont les indices qui permettent de rendre compte d'un ancrage dans le temps des situations décrites dans les textes ? Ceci pose en premier lieu la question de savoir en quoi précisément consiste cette opération de localisation dans le temps. Il faut également encore préciser de quel temps il est question. La linguistique générale distingue généralement trois types de repérage temporel, qui forment un continuum : un repérage qui s'effectue par rapport au référentiel du calendrier (à travers les expressions datatives), un repérage relatif à un procès (par rapport à un « événement ») et un repérage qui s'effectue par rapport à l'instance de la parole (par rapport à l'énonciation).

Dans ce chapitre, nous évoquerons les travaux de linguistes en analyse du discours, qui se sont intéressés en particulier aux deux premiers modes d'ancrages. Ils se sont en effet attachés à observer la façon dont des communautés discursives nomment et désignent des portions de temps.

La section suivante montre comment le troisième mode de repérage, lié à l'instance de la parole, permet d'établir une distinction entre temps linguistique et temps physique. Ce repérage fait en effet partie des concepts-clés de la linguistique énonciative qui s'intéresse à la façon dont les expressions temporelles déictiques signalent la présence de l'énonciateur dans le discours.

Dans les deux sections suivantes, il est question des unités textuelles qui participent de façon privilégiée à l'expression de la localisation temporelle. Elles reçoivent plusieurs dénominations, qui témoignent d'une absence de consensus sur la façon de les catégoriser : adverbes de temps, compléments circonstanciels de temps, syntagmes prépositionnels de temps, subordonnées temporelles, etc. En linguistique générale, du fait de leur parenté sous l'angle de leur fonction syntaxique, on regroupe parfois ces unités textuelles sous le terme générique d'adverbiaux temporels. Seule catégorie du discours qui mélange dans sa dénomination une fonction syntaxique

et une valeur sémantique (les grammaires distinguent généralement les adverbes de manière, de temps et de lieu), les adverbiaux posent des difficultés définitoires et forment une catégorie remise en cause par de nombreux linguistes. Les adverbiaux temporels peuvent être analysés à plusieurs niveaux : au niveau morphosyntaxique (comment ils sont construits ? quelles unités entrent dans leur composition ?), au niveau syntaxique (quelles fonctions occupent-ils ? sur quoi opèrent-ils ? à quoi sont-ils incidents ?) et au niveau sémantique (quelles distinctions opérer au sein de cette catégorie ?).

Nous montrerons l'intérêt de dégager, au sein des adverbiaux temporels, la catégorie des adverbiaux de localisation temporelle, en essayant de révéler leur homogénéité du point de vue des opérations sémantiques qui sont sous-jacentes à leur fonction de localisation temporelle.

2.1 Nommer et désigner des périodes de temps : des dates aux événements et des événements aux dates

La notion de localisation temporelle telle qu'elle est à l'œuvre dans les textes recouvre l'opération par laquelle les textes réfèrent, nomment et désignent des périodes de temps qui s'ancrent sur un référentiel de temps. En première approche, cet ancrage prend corps dans les expressions datatives et les désignations d'événements. A y regarder de près, ces deux catégories, événements et dates, entretiennent des rapports complexes et possèdent des frontières communes. Asseoir ces deux catégories sur un plan sémantique ne va pas sans poser de difficultés. Plutôt que d'en proposer d'emblée une définition, on abordera ces catégories, en rapportant des travaux qui se sont intéressés à la façon dont les textes nomment et désignent des portions de temps. Cette première approche permettra d'accéder au cœur même de la problématique, en introduisant les difficultés linguistiques liées à l'analyse des éléments qui entrent dans la composition des adverbiaux de localisation temporelle. Les travaux d'analyse du discours que l'on relate permettent également de poser les enjeux de l'observation de la façon dont les textes nomment et désignent des portions du temps et montrent l'intérêt qu'il peut y avoir à outiller cette observation.

Les textes, donc, peuvent référer à des dates et désigner des événements et ainsi ancrer sur un référentiel temporel les situations qu'ils décrivent. Dans ce processus, on peut observer la dynamique par laquelle des faits ou des dates sont constitués et reconnus comme faisant événements. Des linguistes se sont intéressés à la manière dont ces référents temporels circulent d'un texte à l'autre, de façon dialogique et diachronique : les désignations d'événements ont en effet des fortunes diverses, plus ou moins durables. Certaines dénominations se figent - ainsi d'un chrononyme comme *La Belle Epoque* ou d'un toponyme événementiel comme *Tchernobyl* -, d'autres ont une durée de vie plus « limitée », circonscrite dans le temps et dans les usages et présentent généralement une plus grande variabilité : d'un texte à l'autre, un même « événement » sera désigné de façons différentes. A côté des *noms de temps* qui se sont cristallisés, il faut donc compter les « *désignations d'événements* » (comme *la canicule de l'été 2003*, *la Vache Folle*, etc.) qui « *constituent des formes condensées de l'événement, rassemblant en même temps des faits et des discours et évoquant des images largement partagées par une société* » (Calabrese Steimberg, 2006).

Les désignations d'événements ne se résument donc pas à leur extension temporelle réduite à la représentation du temps physique, mais contribuent à organiser la mémoire sociale.

Il faut toutefois ici préciser de quels textes et de quelles manifestations discursives il est question. La question du/des genre(s) de textes au(x)quel(s) on s'intéresse pour observer ces phénomènes doit en effet être posée. Si l'opération de localisation temporelle peut s'observer dans tout type de textes, son importance (quantitative et significative) n'est pas partagée de façon équivalente par tous. On pourrait opposer en ce sens articles de presse et ouvrages d'histoire, d'un côté, et fictions de l'autre, par exemple. Pour autant, une fiction peut tout à fait faire référence à des faits historiques avérés et les présenter d'une façon tout à fait éclairante. Il n'y a donc pas de ligne de partage qui permettrait d'épouser les frontières entre des genres de textes sur la base de l'opération de localisation temporelle. Les études auxquelles nous faisons référence ici se sont surtout intéressées à la presse, à des discours ou écrits de politiques, à des ouvrages d'histoire. Cependant, il ne s'agit pas d'une limitation théorique, certains travaux conduits dans cette même veine se sont aussi attachés à constituer des corpus plus éclectiques, y incorporant des œuvres de fiction. On pense par exemple aux travaux sur le chrononyme Année de Plomb (*Anni di piombo*) de (Lettieri, 2010). Ainsi, lorsqu'il s'en trouve d'abondantes, les occurrences par lesquelles est désigné un « événement » méritent tout à fait d'être observées dans des ouvrages de fiction.

Le chrononyme *Anni di piombo* (Années de Plomb), adopté en Italie pour désigner la période qui fait suite aux mouvements contestataires de 1968 et 1969, a ainsi donné lieu à une analyse comparatiste intéressante : le corpus étudié est constitué de livres d'histoire, d'enquêtes journalistiques, d'essais, de témoignages et d'œuvres de fiction, dans lequel Lettieri observe que « *le flou qui persiste [d'un texte à l'autre] en ce qui concerne la date du début et celle de la fin de la période est le signe des enjeux politiques et historiographiques qui produisent l'arbitraire du découpage.* ». On touche ici le cœur du problème : les textes ancrent des situations sur un référentiel temporel externe à la langue, mais, ce faisant, ils n'exhibent pas nécessairement des coordonnées temporelles précises et stables qui désigneraient une portion bien délimitée du calendrier. Il faut donc prendre soin de distinguer la représentation calendaire du temps (qui s'appuie sur un outil de mesure normalisé) et les représentations temporelles que les textes construisent.

Ainsi, au-delà des aspects syntaxiques et sémantiques des adverbiaux temporels et des expressions, qui considérées de façon large, réfèrent au calendrier, des linguistes se sont intéressés à la façon dont le temps est (re)construit socialement dans les écrits, s'attaquant ainsi à l'analyse de sa perception, et notamment à la façon dont les textes élaborent des représentations temporelles en désignant et nommant des événements. Le découpage social du temps est une organisation construite avant tout par le discours (médiatique, scientifique, éducatif, etc.), afin de lui donner sens, de le penser comme histoire, ou tout au moins comme un cadre par rapport auquel il est loisible de se repérer. Dans ce cadre d'analyse, l'opération par laquelle s'opère le repérage temporel ne s'effectue pas d'emblée au moyen d'une référence au calendrier, mais d'abord par la sélection de faits saillants. Eriger ainsi un fait ou un ensemble de faits en événement, découper le temps en plusieurs périodes qui reçoivent un nom, revient à délimiter des portions de temps mémorables, dignes d'attention, porteuses de sens.

Les travaux que l'on présente ici suivent plusieurs directions qui vont de l'étude du processus de nominalisation à l'œuvre pour désigner des événements (construction linguistique de « désignations d'événements ») à l'étude de la circulation de ces dénominations à travers les textes.

2.1.1 Noms de temps : les chrononymes

Les *chrononymes*, terme proposé par (Büchi, 1996), sont des « noms propres de temps ». Pour (Van de Velde, 2000), ces noms propres participent, à côté des noms propres de personnes et de lieux, à la triade des déictiques sur laquelle repose la « référence ». En dépliant son analyse de la fonction des noms propres, Van de Velde est amenée à poser l'existence de « noms de temps » :

> « l'institution des noms propres procède d'une sorte de révolution, qui nous permet de nous mouvoir autour de choses immobiles, au lieu que ce soient les choses qui se meuvent autour de nous. Le sol originaire à partir duquel se constitue tout discours sur le monde est en effet essentiellement mobile et changeant, puisqu'il est constitué de l'ici-maintenant de *je*. Or, s'il est vrai que l'existence de ce point d'ancrage du langage dans le réel est la condition absolue de tout discours possible, en même temps sa mobilité constitue un obstacle majeur à la création d'un monde au sens d'une totalité objective accessible à tout *je* possible. Je proposerai donc de considérer l'institution des noms propres comme une sorte de révolution copernicienne à l'envers : la projection dans un ciel fixe de la triade de repères mobiles qui inaugure le langage. »
>
> (Van de Velde, 2000)

Au sein de cette catégorie des noms de temps, (Van de Velde, 2000) range les noms des jours, des années, des mois et les structures syntaxiques équivalentes. Elle observe que « les noms propres de temps, comme ceux des lieux, peuvent devenir aussi un moyen de référer à des évènements particuliers : *le 4 juillet*, *octobre 17*, *Septembre noir* ».

Comme le remarque (Calabrese Steimberg, 2006), les linguistes ne sont pas unanimes sur le fait qu'il faille ou non inclure dans la catégorie des chrononymes des désignations de faits historiques ou d'événements, si aucune référence lexicale dénotant une zone calendaire n'apparaît (noms de jours, de mois, d'année, etc.) :

> Pour sa part, Flaux considère comme des noms propres de temps *Mai 68* et *Saint-Barthélémy* (Flaux, 2000, p. 123), tandis que Leroy (2004, p. 173) hésite pour ce dernier entre chrononyme et praxonyme, c'est-à-dire entre nom propre de temps et nom de fait historique. La catégorie de praxonyme est d'ailleurs loin de faire l'unanimité : pour Bauer (1985), qui a proposé le concept, il comprend des faits historiques, des évènements culturels et des noms de maladies, mais d'autres chercheurs (Daille, Fourour, Morin, 2000, p. 122) proposent d'y ajouter les noms de périodes historiques (*le Paléolithique*).

A la suite de ces travaux, l'analyse des « noms de temps » a été étendue à d'autres phénomènes linguistiques, glissant vers l'analyse des processus de nominalisation ou de désignation des

événements et s'intéressant à la façon dont ces désignations plus ou moins figées et durables circulent d'un texte à l'autre.

> « Nous appelons chrononyme une expression, simple ou complexe, servant à désigner en propre une portion de temps que la communauté sociale appréhende, singularise, associe à des actes censés lui donner une cohérence, ce qui s'accompagne du besoin de la nommer. À côté des étiquettes strictement calendaires existe en effet tout un appareil de dénominations seul à même de permettre à une société de penser son histoire. »
>
> (Bacot *et al.*, 2008)

Le statut des « noms de temps » est tout de même plus difficile à établir que celui des noms de lieux et de personnes, en particulier parce qu'ils n'ont pas de véritable reconnaissance « administrative ». Par quoi il faut comprendre que ces noms relèvent davantage d'un partage social et culturel et ne peuvent exister sans « une dimension fortement politisée » :

> « Sur le plan sémantico-référentiel, si un chrononyme au sens strict, c'est-à-dire une étiquette fonctionnant en tant que nom propre, vise son référent de façon exclusive et stable, la stabilité référentielle qui est la sienne doit être – comme pour les autres noms propres – relativisée, et cela pour deux raisons. D'une part, les limites précises de l'empan historique couvert par la dénomination sont en général controversables, comme le montre le caractère indécis d'une expression du type l'Après-guerre, voire l'Entre-deux-guerres, ainsi que la difficulté de borner dans le temps les périodes auxquelles de grands événements politiques ou culturels ont laissé leur nom (la Révolution, les Lumières). D'autre part, l'agrégat des représentations colportées par le chrononyme varie avec les imaginaires sociaux : variations partisanes et variations chronologiques. »
>
> (Bacot *et al.*, 2008)

Un même chrononyme peut ainsi donner lieu à des interprétations différentes, à des constructions chronologiques différentes, des vues différentes du découpage temporel, donc des césures dans l'histoire dont le sens varie : la façon dont l'histoire est ressaisie dans les textes se lit aussi à travers ces découpages.

Sur un plan linguistique, une des caractéristiques qui permet également de reconnaître les chrononymes comme des noms propres est qu'ils peuvent être l'objet d'une antonomase (figure qui consiste à employer un nom propre pour désigner un nom commun) : on peut ainsi trouver dans les textes des expressions telles que « un 21-avril à l'envers », « l'Ancien Régime démographique », « le 11 septembre italien ».

Les chrononymes sont en outre constitutivement porteurs d'une indécision quant au référent qu'ils désignent : ils désignent certes une période, mais aussi un ensemble d'événements ou de situations et doivent se lire comme l'« *invention discursive d'un passé constitué comme un tout* » (Christin, 2008). Ainsi, Christin (2008) s'intéressant à la genèse et aux conditions historiques et langagières de la naissance de l'expression *Ancien Régime* montre qu'à l'origine l'expression ne renvoyait pas uniquement à un système ou un régime politique révolu, mais également, en arrière-plan, à un certain type d'ordre social, désignant dès son apparition une période révolue certes, mais aussi « *une*

menace présente » : « *les aristocrates et leurs complices, le clergé réfractaire », « la corruption des mœurs et les préjugés* ». L'expression *Nouveau Régime*, apparue un temps, n'a pas tenue, elle, dans la durée ; elle ne s'est pas sédimentée dans la mémoire collective, nationale.

La pérennisation de l'expression *Ancien Régime* tient aussi à celle d'un pouvoir opposé au régime de la monarchie absolue. Du reste, fait significatif, son apparition dans les textes est concomitante avec la tentative d'imposer un nouveau calendrier, le calendrier révolutionnaire. L'analyse de Christin (2008) montre également que l'extension temporelle désignée par l'expression et sa délimitation précise ont été polémiques, pour l'avènement de la période d'une part (la fin de l'époque féodale, l'absolutisme, la fin du XVIe siècle), mais aussi et surtout pour ce qui est de son achèvement (la conquête de la liberté en 1789 ou l'établissement de l'égalité en 1792).

> « il apparaît clairement que les chrononymes, tels que nous les avons définis ici, n'ont pas vocation à nommer avec précision les fractions du temps humain mais, parallèlement à la dénomination calendaire, à compenser l'inaptitude de celle-ci à construire à elle seule la mémoire historique. »
>
> « Durables ou fugitifs, naturalisés ou éristiques, aptes à cristalliser la durée en lui donnant du sens, les chrononymes sont les outils par excellence de la politisation du temps. »
>
> (Bacot *et al.*, 2008)

Ces différentes remarques soulèvent une question importante : de quelle « inaptitude » exactement serait frappée la « dénomination calendaire » au regard des noms de temps ou désignations d'événements ? En quoi la représentation calendaire est-elle insuffisante pour construire la « mémoire historique » ? Avec leur statut d'« évidence partagée », de référence socialement partagée, les chrononymes peuvent très bien se passer de précision dans « l'empan historique » qu'ils recouvrent : les coordonnées temporelles d'un chrononyme ne vont pas toujours de soi. Du reste, les chrononymes désignent des « portions du temps » en leur reconnaissant une cohérence : ils réfèrent ainsi implicitement à une région plus ou moins bien déterminée du calendrier, mais « en lui donnant du sens », en la constituant comme un tout.

> « l'événement n'est pas une entité homogène et lisse, bien délimitée, ni dans l'espace ni dans le temps, contrairement à ce qu'entendent certains théoriciens[2] un peu trop à la hâte. Prenons quelques exemples de ME [mots-événements] à circulation massive : *le 11 septembre* désigne plusieurs faits qui ont eu lieu ce jour-là, non pas un seul (l'attentat de New York et au Pentagone) ; *Awschwitz* ne désigne pas un événement à proprement parler mais toute une famille d'événements, et les connotations dépassent largement la liste de faits que le désignant résume ; les guerres, par exemple, n'ont pas toujours des limites temporelles précises, et d'autre part, les désignants toponymiques n'ont pas toujours des frontières claires, ce qui nous permet de dire qu'ils mobilisent des imaginaires géopolitiques très complexes : comme dans le cas du *Conflit au Proche-Orient*, qui peut parfois renvoyer au

[2] Voir, par exemple, la définition de l'historien Philippe Joutard : « On peut définir comme événement ce qui advient à une certaine date et un certain lieu » (Bevort *et al.* 1999 : 26).

> *Conflit israélo-palestinien*, au *conflit israélo-arabe* ou bien au conflit global où interviennent la plupart des pays de la région ».
>
> (Calabrese Steimberg, 2006)

Constitués comme des entités par le processus de nominalisation à l'œuvre dans les textes, les événements ont donc des contours flous dont il n'est pas toujours possible de préciser les coordonnées temporelles de façon univoque. Ceci est encore plus vrai des périodes de temps nommées comme le Moyen-Âge ou l'Ancien Régime. Cependant, l'observation conjointe, dans les textes, de l'ancrage calendaire et de l'ancrage à travers des chrononymes ou désignations d'événements doit permettre de faire émerger les représentations chronologiques construites à travers les textes. De là vient l'intérêt d'outiller la fouille des textes à la fois sous l'angle calendaire et thématique.

2.1.2 Désignations événementielles

À côté des noms de temps plus ou moins figés que sont les chrononymes, la nominalisation d'un événement est le fruit d'un processus plus ou moins achevé. Cette dynamique sociale et langagière est à l'œuvre de façon très visible dans les corpus de presse, où s'opère la transition entre le temps « court » de la médiatisation et le temps « long » de l'histoire, où se décante la multitude des événements médiatiques, avant de se réduire progressivement à des événements historiques mémorables, exemplaires ou symboliques.

La langue est du reste un bon observatoire de ce processus, qui, pour ce qui regarde les événements, va des désignations à forte variabilité vers une cristallisation dans des chrononymes quasiment invariables. Positionnant leur analyse en amont de cette opération, Calabrese Steimberg et Sophie Moirand, dans leurs travaux, s'intéressent à la dénomination des événements dans la presse, qui donne lieu à ces constructions singulières que sont les « mots-événements », tels que « vache folle », « poulets à la dioxine », « Tchernobyl », « syndrome des Balkans » :

> « Précédés d'un déterminant défini, ils correspondent de fait à des opérations de référence à des événements, ils fonctionnent comme des dénominations partagées. »
>
> (Moirand, 2001)

Ces désignations peuvent fonctionner comme des noms propres (ainsi des toponymes événementiels comme *Tchernobyl* ou *Fukushima*), comme des dates (appelées également *héméronymes* comme *le 11-septembre*, *le 21 avril*) ou encore comme des syntagmes nominaux (*la Canicule*). Ce qui distingue les noms communs désignant des événements, des toponymes – qui sont des noms propres – et des chrononymes – qui s'assimilent aux noms propres – « c'est qu'ils contiennent des sèmes d'événementialité et ont le pouvoir de créer l'événement sans expliciter les coordonnées événementielles, étant donné qu'ils ont une valeur en langue, une valeur sémantique qui décrit la nature de l'événement, mais qui se spécialise en discours pour désigner un événement particulier. […] Par exemple des « dénominations » telles que guerre, affaire ou conflit, deviennent des « désignations » d'événements uniques et concrets tels que : *la guerre de 14-18*, *l'affaire du voile* ou *le conflit au Proche Orient*, s'actualisant toujours avec un complément. »

Sophie Moirand, dans ses travaux, cherche les traces dans les discours de la façon dont un événement est construit comme tel, à travers des « moments discursifs » :

> « le terme désigne le surgissement dans les médias d'une production discursive intense et diversifiée à propos d'un même fait, par exemple les attentats du 11 septembre 2001, la surprise lors du premier tour de l'élection présidentielle en France le 21 avril 2002, le déclenchement de la guerre en Irak en 2003, la canicule de l'été 2003 »
>
> (Moirand, 2001)

Sophie Moirand s'attache ainsi à analyser ces *moments discursifs* particuliers tels qu'ils apparaissent dans les médias. Son analyse du cas des OGM (Moirand, 2001) vise à la fois à l'observation linguistique de la façon dont s'établit le débat sur ce thème (les termes utilisés, les configurations discursives dans lesquels ils apparaissent) et à l'observation des échanges auxquels il donne lieu sous la forme de « trajets mémoriels », interrogeant ainsi la dimension dialogique à l'œuvre dans ces moments discursifs, où un sème peut se détacher de son contexte initial d'apparition pour permettre de penser de nouveaux événements qui présentent des similarités. Ainsi du sème de la folie apparu initialement pour décrire la crise sanitaire liée à l'infection bovine (la vache folle), qui se retrouve par filiation utilisé dans des contextes différents :

- Alerte au soja **fou** (01/11/96, titre, la Une),
- La faux contre le colza **fou** (09/07/97, titre)
- Le maïs transgénique **affole** les étiquettes (15-16/11/97, titre)
- Le maïs transgénique rend **fou** le PS (04/12/1997, titre)
- En dépit d'un contexte sans précédent (chute de la Commission de Bruxelles, première guerre sur le continent depuis 1945, affaire de la dioxine), les candidats ne sont pas parvenus à faire vivre le débat (...]. Certes, le « **poulet fou** » confirme qu'il faut une Europe sanitaire mais laquelle? Une Europe plus démocratique, quand et comment? (12- 13/06/1999)

Ces travaux ont ceci d'intéressant qu'ils réfléchissent à la façon de mettre en regard des textes en observant la manière dont circulent des désignations d'événements, à la fois dans le temps et à travers les textes :

> « Ce quadrillage de ses observables mène Sophie Moirand à la croisée des deux dimensions du dialogisme bakhtinien: la pluriaccentuation du mot et la construction du discours par « tissage » et « faufilement » dans et avec d'autres discours. »
>
> (Adam, 2001)

2.1.3 Les dates faisant événements : le cas des héméronymes

Références datatives et désignations d'événements ont des frontières partagées. Une date peut ainsi désigner un événement (*Mai 68*, *octobre 17*, *Septembre noir*). (Calabrese Steimberg, 2008) propose le terme d'*héméronyme* (du grec héméra : jour) pour ces expressions qui mettent en avant les

coordonnées temporelles d'un événement. Le paradigme de cette catégorie est le *11 septembre*, en référence aux attentats de 2001[3].

> « L'héméronyme est une date qui désigne un évènement. Il faut distinguer les deux usages, car la date et le nom de l'évènement ont un comportement syntaxique spécifique. Le même syntagme peut être utilisé alternativement comme date (1) tout en se référant à l'évènement en question, ou comme héméronyme (2) :
>
> (1) La publicité se fera discrète aux États-Unis le 11 septembre. (lemonde.fr, 8 septembre 2002)
> United Airlines et American Airlines vont suspendre toute campagne ce jour-là. (11 septembre 2002)
> (2) La vérité sur le 11 mars - Le 11 mars dernier, l'Espagne a subi le pire attentat de son histoire, l'un des plus sanglants que le monde ait connus. (lemonde.fr, 31 mars 2004)
>
> L'héméronyme peut tout aussi bien être le nom d'une année, mais dans ce cas, sa circulation est forcément plus restreinte car il a besoin d'un contexte fort pour être actualisé en tant que désignant évènementiel, comme dans l'exemple (3) tiré d'un dossier sur la Guerre des six jours :
>
> (3) Les deux évènements qui continuent à structurer les consciences dans le monde arabe sont sûrement 1948 et 1967. (Courrier international, 31 mai-6 juin 2007) »

Calabrese Steimberg s'est attachée à observer la genèse de l'héméronyme, qui « ne nait pas tel quel, prêt à circuler » :

> « il est l'aboutissement d'un processus de condensation du sens et du syntagme (…) Lorsqu'on retrace ce processus de condensation, il est toujours possible de retrouver les éléments qui ont été effacés, et notamment le nom qui décrit la nature de l'évènement, de sorte que le désignant résultant n'est finalement, du point de vue syntaxique, qu'un complément déterminatif de ce nom. Dans *11 mars* et *11 septembre*, le nom effacé est « attentat(s) », comme « catastrophe » pour *Tchernobyl*, « affaire » pour *le voile*, etc. »

2.1.4 La référence temporelle en question

Sur un plan linguistique, les désignants événementiels forment ainsi un continuum qui va du nom propre au nom commun, « *en passant par le Np occasionnel et le Nc occasionnel* ». (Calabrese Steimberg, 2006) analyse la façon dont ces désignants gardent la « *mémoire de l'actualité* » : « *un Nc ne va pas garder la mémoire des faits de la même façon qu'un Np ou une date, étant donné qu'un Np a un référent plus stable -du moins en théorie- qu'un Nc. Exemple du* tsunami*, qui a longtemps fonctionné comme un Np, renvoyant au raz-de-marée de 2004 en Asie. Depuis que de nouveaux tsunami ont frappé d'autres lieux (les côtes indonésiennes en juillet 2006, le Japon en 2011), le désignant est souvent actualisé avec un article indéfini ou accompagné d'un circonstanciel de lieu, « signe d'un recadrage du référent », « d'une réactualisation »* ».

[3] Cet héméronyme est d'autant plus intéressant que cette date est « historique » à plus d'un titre. Si sa circulation en tant qu'héméronyme vient bien des attentats terroristes, elle a rappelé à la mémoire collective un autre événement, le coup d'état du 11 septembre 1973 au Chili.

> « un événement n'est pas une entité homogène et définie une fois pour toutes, mais bien un élément du discours protéiforme, capable de transmettre une série d'informations multiples, non pas uniquement sur des données du réel mais sur l'imaginaire de la société qui a créé cette désignation ».

Les désignants d'événements « *participent d'un processus de référenciation, donc de construction de la référence, en l'occurrence de l'événement* », « *ils le construisent en partie en gardant la mémoire événementielle* ».

Comme on l'a vu, dans la « *mémoire événementielle* » les coordonnées temporelles des désignations d'événements n'ont ainsi pas toujours une valeur prépondérante : les textes peuvent y référer davantage pour les représentations qu'ils véhiculent que pour désigner la période où ils se sont déroulés. En outre, que les désignations d'événements ne renvoient pas à des entités « *bien délimitées* » explique pourquoi il n'est pas nécessairement évident de faire coïncider les événements décrits dans différents textes avec des régions précisément définies sur le calendrier. Or on verra plus loin (cf. chapitre 3) que de nombreux travaux en Intelligence Artificielle et en Ingénierie des langues partagent pourtant cet objectif :

> « Une communauté importante de linguistes considère que le langage réfère in fine à des phénomènes du monde réel qui ont, en tant que tels, une extension temporelle, qu'il convient de faire coïncider par une série d'opérations cognitives, avec la représentation newtonienne du temps physique, celle de la droite géométrique. (…) Le traitement des relations temporelles entre procès – représentations linguistiques des objets du monde réel – nécessite en premier lieu le repérage dans le texte de ces procès, ainsi que les informations qui permettent de les situer localement, sinon dans le temps, au moins relativement les uns aux autres. (…) Ces travaux consistent en général à associer à tout objet une extension temporelle ponctuelle ou durative, voire une séquence de points et/ou d'intervalles, puis à traduire les informations temporelles concernant ces objets en relations binaires entre les extensions temporelles, comme un ensemble de relations atomiques possibles. »
>
> (Schwer, 2007)

Dans les différentes analyses linguistiques de la temporalité, (Saussure, 2003) rappelle qu'on distingue traditionnellement (1) des approches référentielles, pour lesquelles les temps verbaux, par exemple, ont pour rôle premier de référer à un moment du temps physique (de Reichenbach à la SDRT) et (2) la tradition sémantico-aspectuelle qui, de son côté, propose de faire dériver la référence temporelle des propositions, des caractéristiques intrinsèques des expressions linguistiques dénotant un procès. Ce qui se joue dans cette opposition, c'est la question de la référence et de la structuration interne du discours, que Saussure suggère de ne pas opposer frontalement, rappelant que la langue ne fait pas que *poser* un monde mais le *désigne* également. L'analyse de la localisation temporelle doit ainsi articuler deux représentations, celles d'un temps « externe » à la langue et un temps linguistique, organisé principalement autour du processus de l'énonciation.

2.2 Temps et énonciation

> « Ce qui en général caractérise l'énonciation est l'*accentuation de la relation discursive au partenaire*, que celui-ci soit réel ou imaginé, individuel ou collectif. Cette caractéristique pose par nécessité ce qu'on peut appeler le *cadre figuratif* de l'énonciation. Comme forme de discours, l'énonciation pose deux « figures » également nécessaires, l'une source, l'autre but de l'énonciation. C'est la structure du *dialogue*. Deux figures en position de partenaires sont alternativement protagonistes de l'énonciation. Ce cadre est donné nécessairement avec la définition de l'énonciation. »
>
> (Benveniste, 1970)

Tout discours engage nécessairement un énonciateur. Si la présence de l'énonciateur a tendance à s'oublier aisément au profit du « contenu » du message délivré, le processus énonciatif n'en demeure pas moins un processus qui engage du temps et qui, même, d'une certaine façon, dans son sillage, engage toute la temporalité linguistique. C'est, en effet, par rapport à ce temps de l'énonciation qu'un texte va pouvoir situer les constructions sémantiques et cognitives qu'il élabore : les différentes situations décrites dans les textes sont toutes, d'une manière ou d'une autre, agencées autour de ce processus, soit en lien avec lui, soit, au contraire, en rupture avec lui :

> « Ce que le temps linguistique a de particulier c'est qu'il est organiquement lié à l'exercice de la parole, qu'il se définit et s'ordonne comme fonction du discours. Ce temps a son centre – un centre, à la fois générateur et axial – dans le *présent* de l'instance de la parole. »
>
> (Benveniste, 1974)

Les déictiques, que la littérature scientifique s'est déjà abondamment employée à décrire, sont sans doute les premiers éléments auxquels on songe, lorsqu'on pense à l'ancrage du temps par rapport au moment de l'énonciation. En effet, si l'on cherche à positionner, sur un calendrier, la référence temporelle d'un énoncé comme « *hier, le président de la Cour de cassation s'est engagé à entamer une procédure…* », il faut pouvoir situer et résoudre en premier lieu le « moment » de l'énonciation.

Cependant, il n'y a pas que les déictiques qui opèrent un positionnement relatif à l'acte de l'énonciation. Les temps verbaux, en permettant de distinguer les ordres du réalisé, de la réalisation en cours et du non réalisé, sont aussi des modes d'expression d'un certain rapport aspectualisé au processus énonciatif.

Par-delà la distinction entre le temps linguistique et un temps « externe », dont la « *conceptualisation (…) a conduit à un temps mathématique newtonien, idéalisé et représentable par une ligne droite continue* » (Desclés, 1994), l'analyse temporelle d'un texte suppose de pouvoir distinguer les situations données comme réalisées, de celles qui sont potentielles, probables ou atemporelles (comme dans le champ de la définition). De ce point de vue, un texte peut être vu comme la construction imbriquée de différents référentiels temporels et modaux (Desclés, 1995) : on peut ainsi distinguer le référentiel énonciatif (où les procès sont situés par rapport au processus énonciatif), les référentiels externes (celui des systèmes calendaires), le référentiel des possibles, le référentiel du non actualisé, celui des vérités générales, etc.

Le temps linguistique s'organise, en premier lieu, par rapport au référentiel énonciatif, lié à l'acte ou au « processus » de l'énonciation, qui *prend* lui-même du temps et opère une rupture entre l'ordre du réalisé et celui du non réalisé.

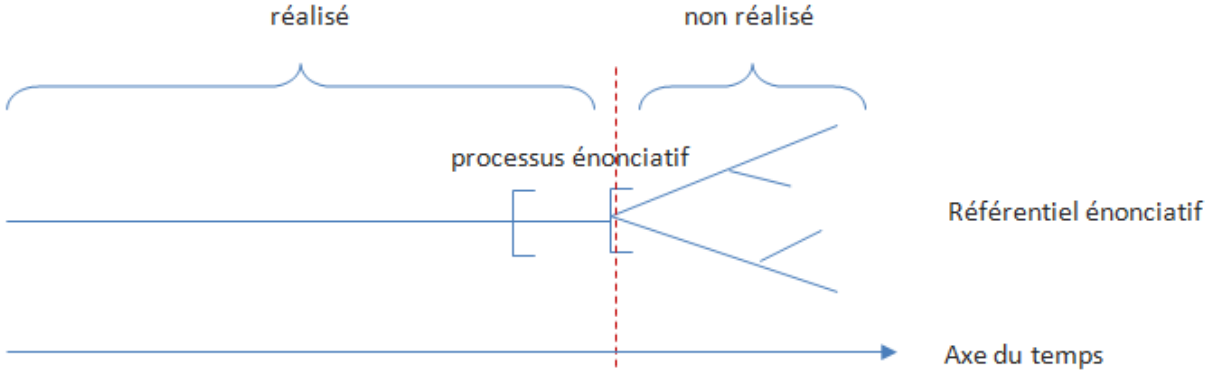


Fig. 1 : une représentation de l'énoncé « Hier, il a dit qu'il viendrait demain. »

Les situations décrites dans les textes peuvent néanmoins être ancrées sur un référentiel différent de celui de l'énonciation, soit en rupture avec le référentiel énonciatif comme dans les narrations (« il était une fois… »), soit, encore, dans un rapport de synchronisation entre un référentiel « externe » et le référentiel énonciatif.

Voici, par exemple, comment on pourrait représenter l'énoncé « *lundi dernier, le président s'était exprimé sur la question des retraites* », en distinguant un référentiel de temps « absolu » (ou référentiel calendaire) et un référentiel énonciatif (on emprunte ici à la modélisation proposée dans (Desclés, 1995) les représentations figuratives et iconiques du temps décomposé en différents référentiels) :

Fig. 2 : la synchronisation entre le référentiel énonciatif et le référentiel calendaire

Le référentiel énonciatif a pour centre l'acte d'énonciation (ou « processus énonciatif »). Dans les textes fortement citationnels, comme fréquemment dans la presse par exemple, plusieurs paroles sont parfois imbriquées dans le discours pris en charge par l'énonciateur. Ceci conduit à considérer non pas un mais plusieurs référentiels énonciatifs. (Battistelli et Chagnoux, 2007) montrent que si un texte présente nécessairement « *la trace d'au moins un acte d'énonciation, celui accompli par l'énonciateur principal qui prend en charge l'ensemble du discours* », en revanche « *certains segments ne sont pas complètement assumés par cet énonciateur, soit qu'un énonciateur second est convoqué au terme d'une citation, soit que ces segments sont assujettis à un certain degré de plausibilité ou d'intentionnalité.* ». Considérons par exemple cet extrait d'un article de presse :

> Les pertes déclarées par l'un ou l'autre des deux camps restent sujettes à caution. Vingt-sept soldats turcs et 240 rebelles kurdes auraient été tués, selon Ankara. Le PKK avance le chiffre de 130 soldats turcs tués, dont « beaucoup sont morts de froid », ainsi que cinq de ses combattants et douze blessés.
>
> Le Monde, 1 mars 2008

Le jeu des modalités montre que l'énonciateur ne prend pas en charge le contenu des paroles qu'il rapporte, mais seulement le fait de les citer : il incite même à une lecture prudente des paroles qu'il rapporte, en les déclarant « sujettes à caution », ayant recours à une modalité conditionnelle (« auraient été tués ») ainsi qu'à la citation directe – deux manières de souligner que le discours rapporté est à différencier du sien. En outre, le lecteur est amené à comprendre, à travers la juxtaposition de deux informations contradictoires, que l'énonciateur souhaite attirer son attention sur les sources mêmes des différentes paroles convoquées dans le texte.

Le repérage de cette intrication des voix, de ces « ruptures énonciatives », est important pour le calcul des déictiques et des anaphoriques. Considérons ainsi les deux exemples suivants :

*(1) Hier il a dit : « Je viendrai **demain**. ».*
*(2) Hier, il a dit qu'il viendrait **demain**.*

Dans le premier exemple, l'adverbe déictique « demain » opère un déplacement par rapport à l'acte d'énonciation d'un énonciateur *second*, celui dont la parole est rapportée. Il y a donc une rupture énonciative (cf. fig. 4). Dans le second exemple où la parole est rapportée de façon indirecte, la trame énonciative n'est pas rompue (cf. fig. 3). Le déictique, selon l'énoncé considéré, ne réfère donc pas au même référentiel énonciatif. Ceci est important, car la résolution de la référence par rapport au référentiel du temps externe sera différente : les deux énoncés ne réfèrent pas au même jour et ne se transposent donc pas de la même manière sur le calendrier.

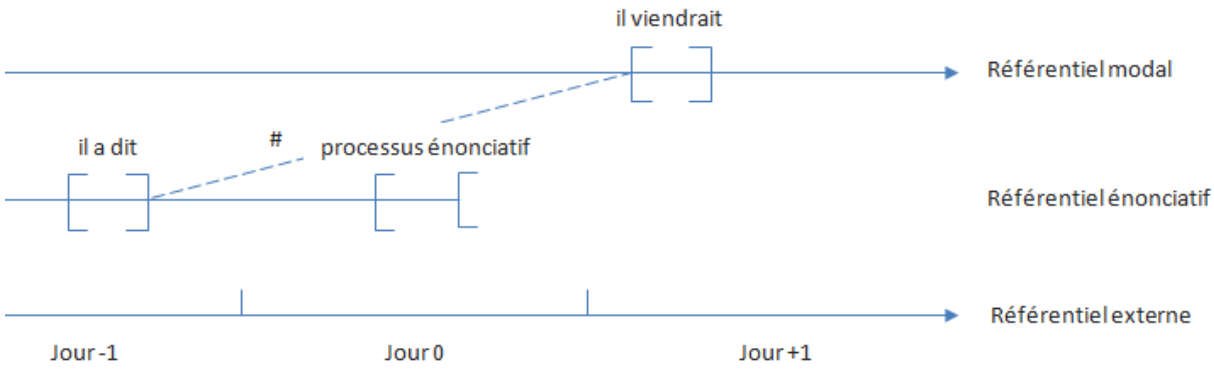


Fig. 3 : une représentation de l'énoncé « Hier, il a dit qu'il viendrait demain. »

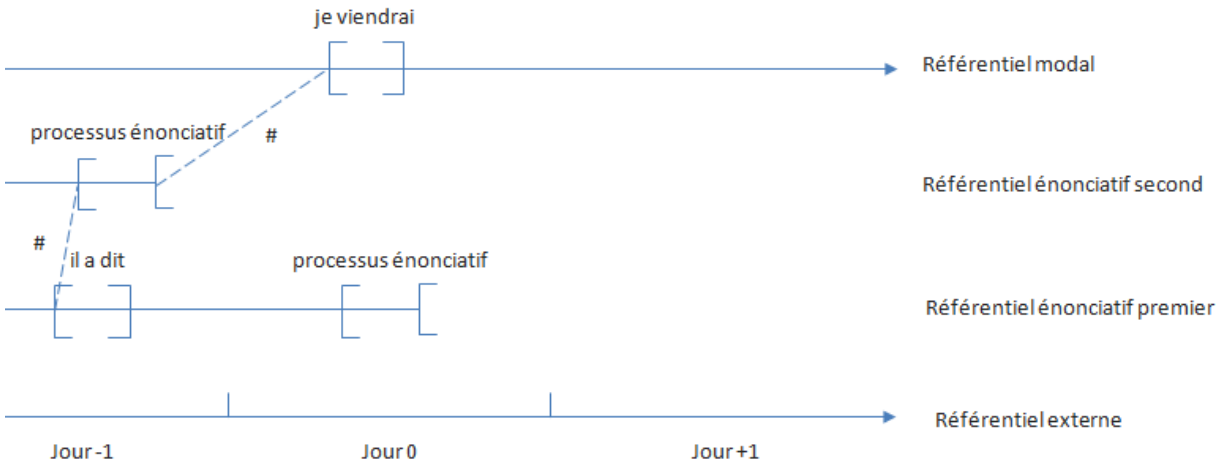


Fig. 4 : représentation de l'énoncé « Hier, il a dit : "Je viendrai demain.". »

Parvenir à repérer les ruptures énonciatives est donc important pour la résolution des références temporelles apparaissant dans le contexte d'un discours rapporté. Le repérage de cette dynamique énonciative peut s'appuyer sur différents éléments de structuration textuelle, comme la présence de changements dans la succession des temps verbaux, la présence de verbes introducteurs de discours rapportés ou encore d'éléments typographiques tels que les guillemets (Battistelli et Chagnoux, 2007).

En considérant les opérations de localisation temporelle dans les textes, on est donc amené à considérer différents types de repérage, par rapport à un référentiel externe, par rapport à l'acte d'énonciation (qui peut référer à d'autres actes d'énonciation) et par rapport à des événements. Sur un plan applicatif, pour notre problématique qui vise à tirer parti des informations de localisation temporelle dans les textes, la question se pose de savoir quelles unités textuelles contribuent de façon saillante à dénoter des opérations de localisation temporelle.

2.3 Les indices textuels contribuant à dénoter une opération de localisation temporelle

On a vu que certains marqueurs dans les textes (les adverbes de temps) pouvaient, ponctuellement, faire coïncider le référentiel énonciatif et le référentiel externe. Par-delà ce mode de référenciation chronologique à l'aide de renvois explicites à un référentiel calendaire, les situations décrites par les textes peuvent être temporellement ordonnées les unes par rapport aux autres. Cet ordonnancement peut ou bien être le seul fait de l'ordre syntagmatique et linéaire du texte (comme dans l'énoncé suivant : « *elle éteignit sa cigarette, s'empara de son sac, sortit par la porte de derrière…* ») ou bien être explicité par des marqueurs linguistiques comme « d'abord », « avant de », « puis », etc.

Si l'on représente les événements décrits par les textes à l'aide d'intervalles de temps, la question se pose ainsi de savoir comment, au niveau textuel, on doit positionner les intervalles les uns par

rapport aux autres, tout en leur associant une valeur aspecto-temporelle donnée. Cette valeur n'étant pas donnée de façon brute par les seuls temps verbaux, on ne saurait se limiter à leur repérage pour instruire le positionnement relatif des intervalles.

En première approche, les temps verbaux dénotent des relations temporelles avec l'acte d'énonciation (relations de postériorité, d'antériorité ou de concomitance). Cependant, la répartition qu'opèrent les grammaires entre les temps verbaux ne suffit pas à en justifier tous les usages : il n'y a pas biunivocité entre les temps verbaux, comme le présent, le passé et le futur, et des valeurs aspecto-temporelles établies et réglées de façon fixe. Le présent peut, par exemple, renvoyer à du futur, comme dans l'énoncé « *A partir de demain, il travaille à Rouen.* ».

Comme le remarque (Le Goffic, 1995) « *il ne va pas de soi que les temps soient fondamentalement « des machines à localiser dans le temps »* ». En effet, davantage que des procédés visant à ancrer des événements sur une ligne de temps, les temps verbaux, de concert avec d'autres marqueurs, renseignent sur le déroulement des procès envisagés et sur le point de vue à partir duquel ils sont décrits. Un même morphème de temps, tel que l'imparfait, peut ainsi conduire à différentes interprétations (Desclés, 2000). L'énoncé « *le train déraillait…* », peut, selon le contexte, ou bien signifier que le train a effectivement déraillé (« *malgré les efforts des cheminots, le train déraillait* ») ou bien que le train n'a pas déraillé (« *sans les efforts des cheminots, le train déraillait* »). Il y a ainsi une distinction à opérer et une articulation à prendre en compte entre la localisation temporelle proprement dite et le point de vue aspectuel à travers lequel un procès est considéré.

Des modifieurs peuvent opérer sur les verbes, et, de fait également, sur les valeurs aspectuelles des énoncés. Les adverbes de temps peuvent ainsi modifier l'aspectualité du procès décrit par le texte : c'est le cas des adverbes de durée (« depuis », « pendant », etc.) ou des adverbes de fréquence (« régulièrement », « fréquemment »). Voici deux exemples de ce phénomène :

1 : « Pierre a écrit son roman en une journée. »
2 : « Pierre a écrit son roman toute la journée d'hier. »

Dans l'exemple 1, l'adverbe de temps ouvert par la préposition « en » permet d'inférer que le roman est achevé. Dans le second exemple, rien ne permet de conclure que le roman est achevé : on peut uniquement inférer que l'acte d'écrire s'est déroulé sur une journée entière et qu'il est accompli, mais en revanche on ne peut rien dire de l'achèvement du but du procès, à savoir l'écriture complète d'un roman. En contexte, un temps verbal ne reçoit ainsi pas toujours la même interprétation aspecto-temporelle, c'est-à-dire que l'on ne peut pas toujours faire les mêmes inférences quant à l'achèvement ou l'accomplissement du procès décrit.

Si les circonstanciels de temps et les temps verbaux sont ainsi des éléments qui participent de façon très saillante à la détermination aspecto-temporelle des situations décrites dans les textes, à y regarder de près, on s'aperçoit que presque toutes les catégories grammaticales, et même certaines catégories lexicales, sont susceptibles d'y contribuer.

Les catégories grammaticales sont encodées dans des classes fermées de marqueurs, telles que les morphèmes de temps. Les informations portées par ces indices inclinent à associer telle ou telle

valeur aspecto-temporelle plutôt que telle autre. Les morphèmes du passé composé sont ainsi des indices forts de la valeur d'événement - événement compris comme un prédicat auquel est associée une valeur aspectuelle distincte du processus ou de l'état (Desclés, 1994). A eux seuls les morphèmes du passé composé ne suffisent cependant pas à attribuer la valeur d'événement, puisqu'un passé composé peut bien renvoyer à la valeur d'événement dans le contexte d'une phrase telle que « *Hier, il a déjeuné à 5h.* » , mais également à la valeur d'« état résultant » dans le contexte de la phrase « *Il a déjà déjeuné, pas besoin de l'inviter*. » (Desclés, 1994).

L'aspect lexical est, quant à lui, constitué par des classes ouvertes (non finies) de marqueurs. L'information aspectuelle peut être portée par un lexème verbal, par exemple : ainsi, le verbe « courir » est plutôt processuel, alors que le verbe « être » est plutôt enclin à prendre part à la construction de propositions statiques (état vs. processus ou événement). Il y a ainsi un continuum entre les composantes lexicale et grammaticale qu'il est difficile de séparer tout à fait : la valeur aspectuelle des énoncés est en effet la résultante d'une interaction entre ces différents éléments.

Les informations temporelles dénotées par les textes sont donc loin de ne se déposer que dans les références datatives ou même plus largement dans les circonstanciels de temps : elles impliquent des éléments très divers et très largement imbriqués, comme le souligne (Gosselin, 1996) :

> « 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 »

Le modèle de la Sémantique de la Temporalité (SdT) décrit dans (Gosselin, 1996 ; Gosselin, 2005a) montre en particulier que la relation entre les adverbiaux de localisation temporelle et le procès qu'ils contribuent à déterminer n'est pas une relation binaire entre deux intervalles, l'adverbial ne venant pas nécessairement déterminer un cadre temporel dans lequel le procès s'insèrerait :

> « Les structures aspectuo-temporelles utilisées dans le modèle SdT mettent en œuvre quatre types d'intervalles disposés sur l'axe temporel : l'intervalle d'énonciation [01, 02], l'intervalle du procès [B1, B2], l'intervalle de référence (ou de monstration) [I, II], et l'intervalle circonstanciel [ct1, ct2]. Alors que l'intervalle du procès ([B1, B2]) correspond à une opération de *catégorisation* (i.e. à la subsomption d'une série de changements et/ou de situations sous la détermination d'un procès), l'intervalle de référence ([I, II]) résulte d'une opération de *monstration* (il correspond à ce qui est perçu/montré du procès, par exemple à ce qui est asserté lorsque l'énoncé est assertif). Les intervalles circonstanciels sont marqués par les compléments de localisation temporelle (ex. *mardi dernier*) et les compléments de durée (ex. *pendant trois heures*). »
>
> (Gosselin, 2006)

L'énoncé « *La police recherchait le coupable depuis trois jours.* » donne ainsi lieu à la représentation présentée de la fig. 5.

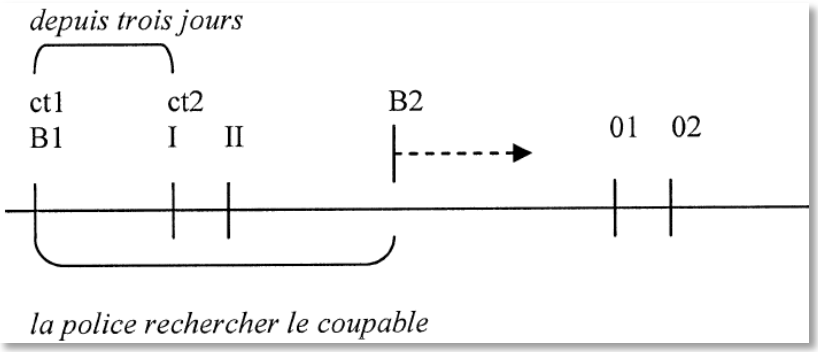


Fig. 5 : représentation de la structure aspectuo-temporelle de l'énoncé « La police recherchait le coupable depuis trois jours » (Gosselin, 2005b)

La fin de l'intervalle associé à l'adverbe de localisation temporelle (ct1, ct2) est concomitante avec le début de l'intervalle de référence (I, II), qui correspond à un anaphorique. L'intervalle associé au procès (B1, B2) précède l'intervalle associé à l'énonciation (01,02), mais rien ne permet d'en préciser l'achèvement, ce qui est représenté par la flèche en pointillés.

Ce qu'il ressort également de ce type de représentations sous forme d'intervalles (celle-ci comme celle plus haut empruntée à (Desclés, 1995)), c'est qu'elles sont parfois contraintes d'opérer une forme de surreprésentation : dans l'exemple précédent, la borne de fin de l'intervalle associé au procès (B2) est positionnée sur l'axe du temps entre l'intervalle de référence (I,II) et l'intervalle d'énonciation (01,02), mais cette position est problématique, car l'énoncé considéré ne dit rien de l'achèvement du procès par rapport à l'énonciation. L'ajout d'une flèche en pointillés vient signaler cette indétermination. De la même façon, l'extension des intervalles (leur durée) est également sur ou sous représentée par rapport à la zone temporelle que les adverbiaux de localisation temporelle peuvent désigner.

En revanche, ce que ces analyses et les représentations iconiques qui les accompagnent mettent en valeur, c'est que les valeurs aspecto-temporelles ne sont pas encodées uniquement dans des classes de marqueurs spécifiques, mais dans de très nombreux éléments de la phrase et même, plus largement, du texte, qu'il faut pouvoir analyser conjointement.

Cependant, s'ils ne sont pas les seuls éléments contribuant à ancrer temporellement les situations décrites dans les textes, les adverbiaux temporels et plus particulièrement ceux que l'on nomme ici « adverbiaux de localisation temporelle » – catégorie à laquelle on s'intéresse plus particulièrement dans ces travaux - fournissent un angle privilégié d'observation de l'opération de localisation temporelle dans les textes, car ils sont susceptibles de contribuer à ancrer les situations décrites dans les textes aussi bien par rapport à l'acte d'énonciation (ex. 1 à 3), par rapport au référentiel calendaire (ex. 4 à 6) que par rapport à un procès, éventuellement nominalisé (ex. 7 à 9).

> Ex. 1 : Vaclav Havel, le président tchèque qui souffre **depuis mardi** d'une affection virale des voies respiratoires, a dû être hospitalisé **hier soir**, après une «détérioration de l'état de santé», selon le porte-parole.
> Ex. 2 : Il n'y avait personne non plus à l'intérieur du hangar que le Secours populaire utilise **depuis un quart de siècle** pour l'accueil ainsi que la vente et le stockage de ses objets, meubles et vêtements.

Ex. 3 : Aux Pays-Bas, l'hypothèse de législatives anticipées **à l'automne** semblait tenir la corde hier, au lendemain de la démission du gouvernement du Premier ministre travailliste Wim Kok.

Ex. 4 : Hans-Joachim Klein, le complice présumé du terroriste Carlos dans la prise d'otages de l'OPEP **en 1975** à Vienne, devait être extradé hier soir vers l'Allemagne, a annoncé le Parquet de Francfort.
Ex. 5 : Ces contrats concernent l'aménagement du temps de l'activité de l'enfant, englobant les aménagements liés au temps périscolaire et extra-scolaire mis en place **depuis 1998**.
Ex. 6 : De plus, **dès l'année 2001**, la municipalité consciente des besoins exprimés, envisage d'engager un lourd programme d'investissements comprenant la restructuration des locaux de l'école maternelle et du centre de loisirs associé à l'école.

Ex. 7 : **Depuis le début de l'Intifada**, fin septembre 2000, les violences israélo-palestiniennes ont fait 2.994 morts, dont 2.243 Palestiniens et 695 Israéliens.
Ex. 8 : Ces propositions interviennent **avant que le bilan 2001 de la délinquance et de la criminalité soit rendu public** lundi matin au ministère de l'Intérieur.
Ex. 9 : Les deux otages français travaillant pour l'ambassade de France à Sanaa ont été libérés sains et saufs hier **après douze jours de captivité** entre les mains d'une tribu armée, du Yemen.

Susceptibles d'être formés à partir de références anaphoriques (cf. ex. 10 et 11), les adverbiaux de localisation temporelle contribuent également à la cohésion discursive.

Ex. 10 : La série, créée Par Garry Marshall -- qui s'illustrera **quelques années plus tard** en réalisant "Pretty Woman" avec Julia Roberts -- décrivait avec humour et candeur les années 50 et 60 américaines.

Ex. 11 : **Ce jour-là**, le Printemps proposera également une création autour des échanges musicaux entre la France et l'Afrique, baptisée "Yeke, Yeke".

Avant de s'intéresser à la façon dont ces adverbiaux sont analysés d'un point de vue syntaxique et sémantique, on s'attache dans un premier temps à les resituer dans la catégorie plus générique des adverbiaux temporels, une catégorie qui pose d'importantes difficultés d'analyse.

2.4 La classe des « adverbiaux aspecto-temporels »

Bien qu'assez intuitivement reconnaissable, la catégorie des adverbiaux temporels se révèle plus complexe à définir qu'il n'y paraît d'abord, à la fois sur le plan syntaxique et sur le plan sémantique. Ils font en effet davantage que répondre à la question *quand* - test à l'aide duquel de nombreuses grammaires constituent encore la catégorie.

Il est vrai que la catégorie des adverbes est elle-même problématique. Définie d'abord comme adjoint du verbe, la catégorie des adverbes peut en contexte être en position de complément d'autres catégories grammaticales. (Wilmet, 2003) remarque ainsi que les adverbes sont moins souvent définis par leur nature que par leur fonction : « *Traduisons en clair les manuels : l'adverbe est un mot invariable ou variable capable de complémenter le verbe et n'importe quoi. On glisse sans avertissement de la nature à la fonction adverbiale.* ». La catégorie des adverbes est également l'unique catégorie grammaticale à laquelle on adjoint également une caractéristique sémantique : les

manuels distinguent ainsi généralement les adverbes de temps, les adverbes de lieux et les adverbes de manière. Cette typologie traditionnelle est à rapprocher de celle des compléments circonstanciels, qui manifestement occupent une fonction syntaxique très similaire à celle des adverbes. Par ailleurs, la frontière entre adverbes et prépositions (qui construisent des syntagmes prépositionnels) n'est pas simple à tracer, comme en témoigne le numéro 157 de *Langue Française*, qui dresse un état des lieux des analyses linguistiques des prépositions. Mêlant un aspect sémantique et un aspect syntaxique, les grammaires, comme pour les adverbes, répartissent souvent les prépositions selon des fonctions différentes, distinguant ainsi des « prépositions spatiales » (sur, sous, à) et des « prépositions temporelles » (pendant, durant). En contexte toutefois, certaines prépositions sont susceptibles de recevoir des interprétations aussi bien spatiales que temporelles :

> « La plupart des prépositions sont polysémiques, et l'on ne peut réduire leur identité à une étiquette ramenant leur sens à l'expression d'une seule relation : celle de l'espace comme le signifie « préposition spatiale » (Vandeloise 1986) ou celle du temps comme l'indique « préposition temporelle », d'autant que cette étiquette ne concerne en fait pas la préposition elle-même mais l'interprétation qu'elle est susceptible de prendre selon le contexte : *à la maison* exprime le lieu dans *Le chien est retourné tout seul à la maison* mais est d'ordre plutôt temporel dans *À la maison, elle se met en robe de chambre* (« quand elle est à la maison », « dès qu'elle est chez elle ») ; le syntagme en sous est spatial dans *Le chat est caché sous le buffet* mais temporel dans *Cet événement a eu lieu sous la Révolution* et causal dans fondre sous la chaleur. »
>
> (Leeman, 2008)

Cette polysémie conduit (Aurnague *et al.*, 2001) à regrouper sous la catégorie des adverbiaux de localisation aussi bien des adverbiaux de localisation temporelle que spatiale, entre lesquels des rapprochements sont établis. Les auteurs illustrent cette question des liens entre adverbiaux spatiaux et temporels à l'aide d'un exemple éclairant, qui montre bien qu'un adverbial spatial est susceptible de revêtir une dimension temporelle :

> *On ne voyait rien du paysage. Il pleuvait à verse depuis Toulouse. À Cordes, la pluie se transforma en grêle, et, dix minutes plus tard / dix kilomètres plus loin, le tonnerre se mit à gronder.*

Polysémiques et parfois difficiles à catégoriser sur un plan sémantique, les adverbiaux temporels recouvrent également différentes catégories morphosyntaxiques, qui elles-mêmes reçoivent différentes dénominations. Adverbes de temps (*depuis, auparavant*), syntagmes prépositionnels ou compléments circonstanciels de temps (*depuis ce jour, avant les années 30*), subordonnées temporelles (*avant qu'il ne revienne*) sont autant de catégories dégagées par les grammaires pour circonscrire les unités textuelles qui contribuent à déterminer temporellement un procès. Ces catégories posent de sérieuses difficultés d'analyse, et ce d'abord parce qu'elles concentrent dans leur dénomination même un aspect syntaxique (c'est-à-dire un comportement particulier dans l'ordre syntagmatique) et un aspect sémantique (l'expression d'une caractéristique temporelle). De là des problèmes de délimitation entre des éléments linguistiques qui se comportent de la même manière d'un point de vue syntaxique, mais qui ne partagent pas les mêmes traits sémantiques et à

l'inverse des éléments qui diffèrent sur le plan syntaxique, mais sont à rapprocher sur le plan sémantique :

> « des adverbes temporels tels que *demain* et *parfois* ont peu de chose en commun : sémantiquement, *demain* réfère à un espace de temps situé par rapport au moment de la parole, alors que *parfois* quantifie, ne serait-ce que vaguement, la fréquence du procès en question, sans référence à une donnée extérieure au contenu de la phrase.
>
> On sait depuis longtemps que d'un point de vue syntaxique, les différences n'en sont pas moins grandes ; par contre, des adverbes sémantiquement proches (p. ex. *souvent/parfois*) peuvent diverger distributionnellement. Les critères sur lesquels repose la catégorie des adverbes de temps se réduisent donc à une idée relativement vague, et pour le moins inexplicable à partir de considérations purement linguistiques, de ce que pourrait impliquer de près ou de loin la notion de temps. »
>
> (Blumenthal, 1990)

Cette difficulté de catégorisation s'accroît encore lorsque l'on veut bien considérer que chacun des deux aspects (sémantique et syntaxique) ne forment pas des groupes aux frontières infranchissables, mais qu'au contraire il y a un continuum entre des éléments qui peuvent s'y ranger de façon typique et des éléments plus ou moins atypiques (loin des prototypes), mais dont on considère encore qu'ils sont encore polarisés par ces classifications.

Sur un plan sémantique, il faut distinguer des adverbiaux qui relèvent de façon prototypique de la catégorie des adverbiaux temporels, d'avec ceux qui sont à la frontière avec d'autres catégories d'adverbiaux, tels que les adverbes de localisation spatiale (*« Il pleuvait à verse depuis Toulouse. »*) ou les adverbes de manière (*« lentement », « doucement »*). Ainsi, les adverbiaux de manière, parce qu'ils véhiculent davantage qu'une représentation temporelle, se situent à la périphérie des adverbiaux temporels. Les adverbiaux de durée (« *des jours durant* »), les adverbiaux de fréquence (*« souvent », « quatre fois par jour », « un vendredi sur deux »*) et les adverbiaux de localisation temporelle appartiennent pleinement à la catégorie des adverbiaux temporels.

Répondant à cette difficulté classificatoire et sans limiter d'emblée son analyse aux adverbiaux liés d'abord sur un plan sémantique, Dan Van Raemdonck fournit une description riche des adverbes de temps, en affirmant un primat, sur le plan de l'analyse, du comportement syntaxique, la délimitation de sous-catégories sémantiques n'intervenant qu'en second lieu :

> « Nous proposons (…) d'inscrire la fonction adverbiale dans un système où toutes les fonctions sont définies à partir d'un même critère, l'incidence guillaumienne (relation entre un apport et un support de signification). Les compléments adverbiaux, quoique morphologiquement divers (du mot à la phrase), sont unifiés par la caractéristique de la fonction qu'ils ont en commun : l'incidence externe du second degré, la propriété qu'ils ont de porter syntaxiquement sur une relation entre deux termes. (…) Ces relations supports sont elles aussi multiples et se rencontrent à des niveaux divers de la phrase, allant du niveau supérieur, la relation prédicative, au niveau inférieur, infrasyntagmatique. Cependant, ce sont toujours des relations. »
>
> (Van Raemdonck, 2001)

L'originalité de son analyse tient dans ce qu'elle cherche à situer et répartir les adverbes de temps en fonction de leur portée, soit qu'ils déterminent une relation prédicative (*Hier*|*Le 15 mars*|*Après le départ de Marie, Pierre a ouvert la lettre*), une relation intraprédicative (*Pierre part demain*) ou une relation infrasyntagmatique (*Pierre écrit des lettres souvent compromettantes*). Van Raemdonck distingue également les adverbes qui déterminent un énoncé (*Ensuite, Pierre a ouvert la lettre*), de ceux qui déterminent l'énonciation-même (*Primo, Pierre n'a pas ouvert la lettre*).

À l'aide de différents critères qui ont tous en commun la détermination d'une relation, Van Raemdonck est conduit à regrouper dans la catégorie des adverbes de temps des adverbes dont l'analyse pose problème et qui bien souvent sont exclus de la catégorie des adverbiaux temporels : ainsi des adverbes dits de manière (*brusquement, lentement*), des adverbes de fréquence (*souvent*), par exemple, qui comportent bien une dimension sémantique temporelle, mais qui ne se limitent pas à celle-ci. Ainsi, certains adverbiaux, comme les adverbes de « cadre instrumentaux-manière » se situent *« à la frontière de deux catégories traditionnelles et intuitives, celle des compléments de manière et celle des compléments temporels.* »[4].

Sur un plan syntaxique, la catégorie des adverbes et des frontières qu'elle entretient avec d'autres constituants est ainsi l'objet d'épineux débats entre linguistes, que nous n'avons pas l'ambition ni les moyens de démêler ici. On retiendra donc pour la suite de cette étude, une acception large des compléments de temps, regroupés sous l'étiquette d'adverbial, qui pourra désigner aussi bien des adverbes (*depuis, avant, après*), des syntagmes prépositionnels (*durant la Révolution*) ou encore des subordonnées circonstancielles de temps (*depuis qu'il est parti*) et des syntagmes nominaux fonctionnant comme complément adverbial (une voiture volée la veille).

Comme on l'a vu, la catégorie des adverbiaux temporels pose également des problèmes définitoires sur un plan sémantique. Dans le modèle de la Sémantique de la Temporalité, (Gosselin, 2005b) oppose les circonstanciels de temps (*depuis ce jour, hier, le 15 mars, pendant ce temps*) qui « *construisent un intervalle circonstanciel sur l'axe temporel* », aux circonstanciels d'aspect (*souvent, trois fois, encore, déjà*), qui « *modifient les relations entre les intervalles construits à partir des autres marqueurs de l'énoncé (en particulier entre l'intervalle de référence et celui du procès)* ». Le schéma ci-dessous est une proposition de typologie des adverbiaux aspecto-temporels tirée de (Gosselin, 2005b) qui fait apparaître les *circonstants de localisation temporelle* comme une sous catégorie des *circonstants temporels*.

[4] Melis L. (1983). Les circonstants et la phrase. Louvain, P.U.L.

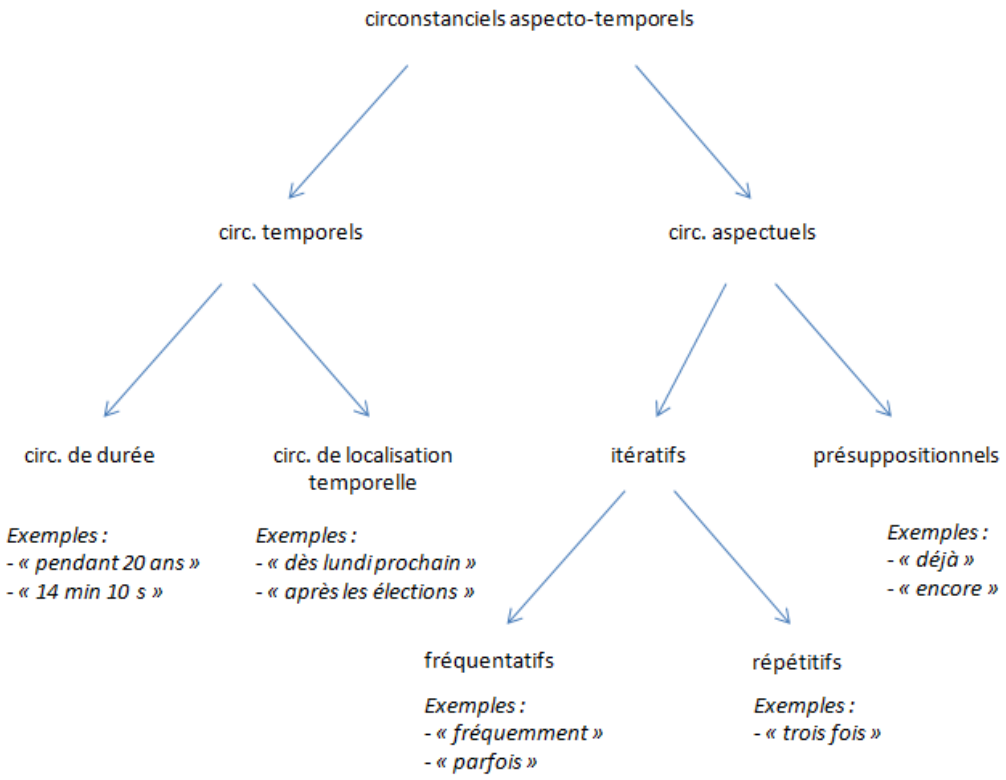


Fig. 6 : typologie des adverbiaux aspecto-temporels (Gosselin, 2005b)

Parmi les adverbiaux temporels, (Gosselin, 2005b) distingue les circonstanciels de durée (*pendant deux heures*), « *qui définissent la taille de l'intervalle circonstanciel sans le localiser autrement que par rapport au procès et/ou à l'intervalle de référence* » et les circonstanciels de localisation (*en 2002, mardi dernier, ce jour-là, avant son retour*), « *qui situent l'intervalle circonstanciel de façon plus ou moins précise et plus ou moins déterminée par rapport au calendrier (localisation absolue), à l'intervalle de l'énonciation (localisation déictique), ou à un autre procès (localisation relative) ; ce dernier type de localisation est caractéristique des subordonnées circonstancielles, dans lesquelles l'intervalle circonstanciel est situé par rapport au procès exprimé par la subordonnée.* ». Il y a cependant un continuum entre les adverbiaux de localisation et les adverbiaux de durée :

> « Signalons que la distinction entre « localisation » et « durée » est quelque peu trompeuse, laissant penser que l'on a des intervalles avec durée d'un côté, et sans durée de l'autre. Or tout intervalle a une durée, aussi infime soit-elle. L'accent n'est simplement pas mis sur la même caractéristique de l'intervalle selon les expressions : les adverbiaux de durée insistent sur sa taille, ceux de localisation sur sa situation. Les deux composantes, taille et situation, coexistent au sein des deux types d'intervalles, les adverbiaux les désignant devant donc être envisagés dans la perspective d'un continuum. »
>
> (Prévost, 2005)

Concernant les adverbiaux de localisation temporelle, l'opération de localisation temporelle consiste donc, dans ce cadre d'analyse, à situer un intervalle circonstanciel. Ceci revient à dire que l'intervalle circonstanciel et l'intervalle du procès ne se confondent pas : en ce sens, les adverbiaux de localisation temporelle n'ancrent pas un procès dans le temps, mais contribuent à le faire.

2.5 La classe des adverbiaux de localisation temporelle

S'ils présentent une grande diversité à la fois du point de vue de leur composition morphosyntaxique et au niveau des rôles qu'ils sont susceptibles d'occuper dans un énoncé, les adverbiaux de localisation temporelle présentent en revanche une relative homogénéité au niveau sémantique, comme le révèlent, bien qu'avec des points de vue différents, les propositions d'(Aurnague *et al.*, 2001), (Gagnon et Lapalme, 1996) et (Battistelli, 2009), qui s'attachent à formaliser la sémantique de ces adverbiaux.

2.5.1 Les différentes formes possibles

S'attachant à décrire la composition des adverbes de référence temporelle, (Borillo, 1998) distingue : (1) des adverbes simples ou composés (*demain*, *hier*, *désormais*, *bientôt*, *aussitôt*, *prochainement*, *plus tard*, *avant peu*, etc.), (2) des syntagmes prépositionnels (*à midi*, *dans l'après-midi*, *à l'aube*, *pendant la nuit*, *dans trois jours*, *en janvier*), (3) des formes nominales sans préposition (*une nuit*, *le lendemain*, *le 3^{e} jour, le jour suivant*, *une heure après*), et (4) des formes nominales sans déterminant (*lundi dernier*, *jour après jour*).

Au niveau morphosyntaxique, plusieurs catégories peuvent être ainsi rangées sous l'étiquette d'adverbial de localisation temporelle : des adverbes (cf. ex. 1 et 2), des syntagmes prépositionnels (cf. ex. 3 et 4), des subordonnées (cf. ex. 5 et 6), des syntagmes nominaux (cf. ex. 7 et 8) ou même des noms communs sans déterminant en position d'adverbial (cf. ex. 9 et 10).

> Ex. 1 : **Depuis**, il a gagné deux matches de Coupe Davis sur la surface, Roland-Garros en 2010 et 4 Masters 1000 (Monte-Carlo en 2011 et 2010, Rome et Madrid en 2010).
> Ex. 2 : Les avocats, qui commencent à se frotter aux nouvelles règles de la garde à vue applicables depuis vendredi, pensent pouvoir pousser plus loin l'exigence de garanties pour les droits de la défense en contestant les procédures judiciaires engagées **auparavant**.
>
> Ex. 3 : Mais **dès dimanche matin**, des tirs particulièrement intenses sur la porte ouest d'Ajdabiya indiquaient que les forces pro-Kadhafi étaient revenues à moins de 20 km de cette ville, poussant certains rebelles et les habitants restés dans la ville à fuir par centaines, selon un journaliste de l'AFP.
> Ex. 4 : Le taux de participation national n'était pas connu **à 18h15 GMT**, mais pour Helsinki il était de 75%.
>
> Ex. 5 : L'administration américaine a annoncé dimanche une nouvelle organisation des horaires de travail des contrôleurs aériens **après que plusieurs aiguilleurs du ciel se sont endormis durant leur vacation ces dernières semaines**.
> Ex. 6 : L'Isaf soutient depuis fin 2001 le gouvernement afghan dans sa lutte contre l'insurrection que mènent les talibans **depuis qu'ils ont été chassés du pouvoir**.
>
> Ex. 7 : Elle sera suivie **le dimanche** par une grande soirée reggae autour de Tiken Jah Fakoly, Alborosie ou encore Chinese Man.
> Ex. 8 : La colonie d'Itamar a été la cible **ces dernières années** d'une série d'attentats palestiniens, dont une attaque armée en juin 2002, qui avait fait quatre morts (une femme et trois enfants).

Ex. 9 : Son parti domine les sondages depuis des mois et il est crédité de 21,2% d'intentions de vote, selon la dernière enquête publiée **jeudi**.

Ex. 10 : Samedi matin, cinq militaires de l'Isaf avaient péri dans un attentat-suicide contre le quartier général de l'armée afghane dans l'est du pays, qui avait également tué quatre soldats afghans, l'une des attaques les plus meurtrières pour les forces de l'Otan depuis leur arrivée dans le pays **fin 2001**.

(Aurnague *et al.*, 2001) remarquent que derrière cette diversité des adverbiaux de localisation, il est possible de dégager une structure morphosyntaxique régulière, qui peut s'enrichir de façon récursive. Les adverbiaux de localisation prennent ainsi généralement la forme d'un syntagme prépositionnel qui peut connaître deux types de complémentation : un complément classique de la préposition (*avant les vacances* ; *depuis ce jour-là*) et un complément en position dite de « spécifieur » (*specifier*) (*peu avant ; quelques jours après la réunion*). Ces compléments peuvent dénoter ou bien des durées (*deux jours ; longtemps*) ou bien une ancre permettant un repérage temporel. Seuls les syntagmes exprimant des durées peuvent occuper la position de spécifieur, alors que la position de complément à droite de la préposition peut être occupée par des syntagmes exprimant ou bien une durée (*depuis trois jours*) ou bien une ancre temporelle (*depuis hier*). En outre, le complément à droite peut imbriquer de façon récursive un nouveau syntagme prépositionnel (cf. fig. 7).

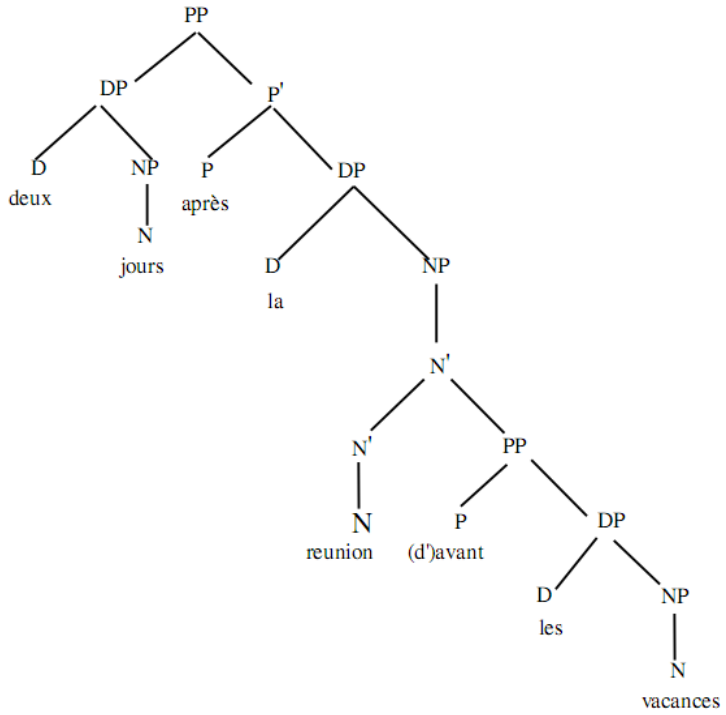


Fig. 7 : Arbre de dépendance syntaxique d'un adverbe de localisation temporelle

On a vu que certains adverbiaux sont formés par un syntagme nominal (*le 3 mai ; dimanche ; ces derniers jours*). Pour (Aurnague *et al.*, 2001), ces syntagmes nominaux doivent s'analyser comme des adverbiaux dont la préposition n'est pas marquée au niveau morphosyntaxique. Pour le cas des adverbes seuls (tels que les anaphoriques comme *depuis* ou *auparavant*), à l'inverse, c'est le complément du syntagme prépositionnel qui n'est pas marqué. Les auteurs remarquent également

que certaines prépositions (en particulier *jusque*) peuvent avoir comme complément un syntagme prépositionnel complet (*jusqu'après Noël*).

2.5.2 Position dans la phrase et portée

Les adverbiaux de localisation temporelle peuvent occuper différentes positions dans l'énoncé. Ils peuvent ainsi être adjoints d'un syntagme verbal (cf. ex. 1), mais ils peuvent également être adjoints d'un verbe seul (cf. ex. 2 et 3) ou d'une structure prédicative complète (cf. ex. 4), ou bien encore porter sur la relation entre un prédicat second et un nom (cf. ex. 5) et enfin porter même sur un prédicat nominalisé (cf. ex. 6).

> Ex. 1 : Cinq personnes ont été blessées **dimanche** dans la dispersion de manifestations en faveur de la liberté à Soueida, bastion des druzes syriens dans le sud du pays, ont affirmé des militants.
>
> Ex. 2 : La loi d'urgence date **de 1962** et elle est en vigueur depuis l'arrivée au pouvoir du parti Baas en 1963.
>
> Ex. 3 : Avec le gouvernement qu'il doit encore former, ce nouveau Premier ministre, Luc-Adolphe Tiao, 56 ans, ambassadeur du Burkina à Paris, va essayer de mettre fin à la colère de toutes les couches de la population, qui dure **depuis février** et s'est intensifiée depuis la fin de la semaine dernière.
>
> Ex. 4 : **Dimanche**, après deux jours d'entretiens à Nouméa, Marie-Luce Penchard a proposé une double solution "juridique et politique" pour sortir de "cette crise dangereuse pour la stabilité", avec un texte court modifiant la loi organique de 1999 empêchant les démissions à répétition du gouvernement néo-calédonien.
>
> Ex. 5 : La Bourse de Londres a terminé en hausse mardi, les investisseurs s'étant remis du choc provoqué **la veille** par la mise en garde de Standard and Poor's sur la dette américaine, et ayant salué des annonces du groupe de luxe Burberry.
>
> Ex. 6 : La banque d'affaires Goldman Sachs, dont le rôle pendant la grande crise financière **de l'automne 2008** est souvent vertement dénoncé, publie ses résultats **du 1er trimestre**.

Le cas de la position initiale

Susceptible de porter sur différentes portions d'un énoncé, les adverbiaux de localisation temporelle peuvent également porter sur des portions supérieures à l'énoncé. On parle alors de cadratifs temporels. Par cadratif, au sens de (Charolles, 1997), il faut entendre des éléments linguistiques qui ouvrent un cadre de cohérence forte dans un texte. Un cadratif temporel est ainsi un ancrage chronologique dont la portée dépasse la phrase où il apparaît, c'est-à-dire que toutes les situations décrites dans le cadre qu'il ouvre sont, d'une certaine façon, sous son aile.

> « les adverbiaux détachés en tête de phrase sont cadratifs dans le sens où ils sont à même de porter non seulement sur la proposition en tête de laquelle ils apparaissent mais sur une ou plusieurs autres figurant dans la suite. Les cadres qu'ils délimitent constituent des blocs

> informationnellement homogènes par rapport au critère signalé par l'adverbial. Ce critère fonctionne comme une sorte d'index que le lecteur ou l'auditeur doivent garder en mémoire pour le traitement de la phrase hôte de l'adverbial et au-delà, jusqu'à l'occurrence d'indices signalant que sa portée est terminée. »
>
> (Charolles, 2003)

(Le Draoulec et Pery-Woodley, 2005) montrent que pour être doté d'un potentiel cadratif, un adverbial temporel doit être en position initiale dans une phrase. En voici un exemple type :

> ***En juin 1992,*** *747 500 candidats se sont présentés à 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 élèves du privé sont nombreux, d'Orléans-Tours, Reims et Grenoble, ainsi que dans les Midis aquitain et méditerranéen*[5].

Dans ces exemples, l'encadrement, donc la forme d'indexation, s'établit vers l'avant. Le lien se fait entre l'expression indexante et la ou les propositions qui la suivent. Cette organisation temporelle peut, toutefois, agir de façon concurrentielle avec d'autres formes de structuration discursive. Comme le remarquent (Le Draoulec et Pery-Woodley, 2005), la structuration temporelle est en interaction parfois conflictuelle avec d'autres types de relations discursives, parmi lesquelles la relation de narration. Dans cet article, les auteurs observent que les cadratifs temporels, auxquels on a tendance à prêter une portée qui déborde la phrase, sont susceptibles, dans les faits, d'avoir une portée restreinte, s'ils entrent en conflit avec une relation de narration. Voici ainsi un exemple analysé par les auteurs, où l'adverbial temporel en position initiale n'a pas la portée extra-phrastique qu'ont habituellement les cadratifs :

> ***En 1933,*** *il [Klaus Mann] fonda à Amsterdam la revue antinazie "Die Sammlung". Il sillonna l'Europe pour mobiliser les intellectuels contre le fascisme, donna des conférences, écrivit des articles virulents contre le régime hitlérien, notamment dans le "Pariser Tageblatt", journal des Allemands antinazis en France, et collabora au cabaret satirique dirigé par sa soeur Erika, "Die Pfeffermühle" (Le Moulin à Poivre).* ***En 1938,*** *il se rendit en Espagne pour faire des reportages sur la guerre civile; il prit parti pour les Républicains dans ses articles très polémiques*[6].

Dans cet exemple, le premier adverbial en position initiale (« en 1933 ») pourrait ouvrir un cadre, qui ne se fermerait qu'avec l'expression calendaire suivante (« en 1938 »). Cependant, il ne s'agit pas d'un cadratif, car la phrase suivante présente une série d'événements qui ne se déroulent pas tous en 1933. (Le Draoulec et Pery-Woodley, 2005) montrent ainsi que la structuration temporelle entre ici en conflit avec la relation de narration qui se donne à lire dans la juxtaposition des passés simples (« sillonna », « écrivit » et « collabora ») et vient contrarier la portée usuelle de l'adverbial placé en position initiale dans la phrase précédente. L'organisation temporelle des textes opère ainsi en interaction avec d'autres éléments discursifs.

[5] Hérin, R. & Rouault, R., 1994. *Atlas de la France Scolaire de la Maternelle au Lycée*. Paris, La Documentation Française.
[6] « La résistance allemande au nazisme » (http://resistanceallemande.online.fr/).

Cette question de la portée des adverbiaux et plus généralement de leur façon de contribuer à la cohérence textuelle intéresse notre problématique, car comme on le verra, sans d'ailleurs qu'on réponde directement à cette difficulté, les systèmes de recherche d'information qui fonctionnent par extraction de segments textuels, tel que celui que l'on a mis en œuvre, sont confrontés au problème de la continuité référentielle[7] :

> « Isoler une connaissance de son cadre de validité revient à occulter une part essentielle de son contenu, au risque de lui faire perdre sa valeur opératoire. »
>
> (Jackiewicz et Minel, 2003)

Comme on l'a évoqué, cette continuité référentielle prend corps également à travers la trame référentielle construite par le jeu des anaphores, mais aussi dans celui de la dynamique énonciative : les adverbiaux anaphoriques réfèrent ainsi à des éléments déterminés antérieurement dans un texte (*auparavant*) et les adverbiaux déictiques peuvent renvoyer à différents référentiels énonciatifs.

Ainsi, si l'ancrage temporel qu'ils contribuent à déterminer peut dépendre de différents référentiels, s'ils peuvent occuper différentes fonctions syntaxiques et recouvrir différentes catégories morphosyntaxiques, en revanche, ce que mettent en lumière les analyses visant à formaliser la sémantique des adverbiaux de localisation temporelle, c'est qu'ils présentent une relative homogénéité au niveau sémantique.

2.5.3 Formaliser la sémantique des adverbiaux de localisation temporelle

(Aurnague *et al.*, 2001) décrivent une approche originale de la façon dont sont formés les adverbiaux de localisation (temporelle ou spatiale) - originale à la fois parce qu'elle distingue mais traite conjointement l'analyse syntaxique et sémantique des adverbiaux de localisation temporelle et parce qu'elle milite en faveur d'une analyse « relationnelle » plutôt que « référentielle » de ces adverbiaux (au moins lorsqu'ils sont en position d'adjoint d'un verbe). Ces travaux visent ainsi à formaliser une « sémantique compositionnelle des adverbiaux » :

> « Cette étude examine les propriétés des adverbiaux de localisation du français à plusieurs niveaux. La structure syntaxique de ces éléments est décrite de même que les interactions complexes entre position dans la phrase et contribution sémantique. En se focalisant sur la position d'adjoint du syntagme verbal, on montre que le contenu sémantique des marqueurs considérés est mieux saisi par une approche 'relationnelle' que par une approche 'référentielle'. »

[7] On verra que la navigation textuelle, par opposition aux systèmes qui fonctionnent par extraction de segments textuels offrent une réponse à cette difficulté, parce qu'elle permet d'accéder à des portions de textes sans les isoler de leur contexte. Si nos travaux, sur un plan applicatif, ne se sont pas insérés à ce stade dans ce paradigme, on verra qu'ils pourraient néanmoins tout à fait le faire.

Deux approches différentes semblent ainsi coexister, selon les (Aurnague *et al.*, 2001), l'une dite « référentielle » (qui considère le syntagme prépositionnel complet comme une référence temporelle) et l'autre dite « relationnelle », que les auteurs défendent, et qui considère le complément du syntagme comme la référence temporelle, la préposition fonctionnant, elle, comme relateur entre le procès et le complément.

La discussion, telle qu'elle est posée, consiste à savoir à quel niveau situer la référence temporelle désignée par les adverbiaux de localisation temporelle : faut-il par exemple considérer que l'adverbial « avant les vacances » constitue une référence temporelle ou bien faut-il considérer que seule l'expression « les vacances » constitue une référence temporelle, la préposition « avant » dénotant la relation entre le procès déterminé par l'adverbial et cette référence temporelle ? Décrit ainsi, le problème revient donc, d'une certaine façon, à décrire les liens entre deux intervalles de temps. Le premier intervalle de temps est constitué par l'événement (le procès) que l'adverbial de localisation vient déterminer. La question est de savoir comment est constitué le second intervalle : dans l'approche relationnelle décrite par (Aurnague *et al.*, 2001), le second intervalle est constitué non pas par l'adverbial en entier, mais par le complément du syntagme prépositionnel, la préposition qui vient clore le syntagme dénotant, elle, la sémantique de la relation entre les deux intervalles. Le travers associé à l'approche référentielle est qu'elle décrit toujours les liens entre les deux intervalles de temps sous la forme d'une relation d'inclusion :

> « *the referential approach (…) implies that the locating adverbials call for a unique locating relation, inclusion: the preposition operates as a function on the complement and provides the time or place in which the eventuality will be located via the inclusion.* ».
> (l'approche référentielle (…) suppose que l'adverbial de localisation n'appelle qu'un unique type de relation, l'inclusion : la préposition opère comme une fonction sur le complément et fournit le lieu ou le temps au sein duquel l'événement va être localisé via l'inclusion.)

L'approche référentielle suppose ainsi que l'adverbial de localisation détermine un cadre temporel dans lequel s'insère un procès, alors que, dans l'approche relationnelle, les relations entre les intervalles de temps associés au procès et au complément de l'adverbial sont plus riches. Sur un plan formel, une préposition est donc analysée comme une relation à plusieurs arguments, dont l'un est le procès (*eventuality*). La sémantique du complément de la préposition peut ensuite être décomposée à son tour, sachant que les adverbiaux de localisation peuvent être construit par enchâssement, de façon récursive (par ex., *deux jours après la réunion d'avant les vacances*).

Les auteurs précisent toutefois que cette analyse relationnelle vaut pour les adverbiaux en position d'adjoint au verbe. Ils soulignent ainsi indirectement l'importance de la nature des éléments sur lesquels portent les adverbiaux et des difficultés d'analyse qu'il peut y avoir lorsqu'il s'agit d'établir les liens entre un adverbial de localisation et ce à quoi il est adjoint.

Opposer ainsi une approche « relationnelle » et une approche « référentielle » n'est peut-être pas pour autant nécessaire, si l'on veut bien considérer que les adverbiaux ne désignent pas une fenêtre de temps durant laquelle les procès qu'ils contribuent à ancrer dans le temps se dérouleraient. Rappelons en effet, comme le montre (Gosselin, 1996 ; 2006), que les adverbiaux de localisation temporelle peuvent être représentés par des intervalles circonstanciels, mais que ces

intervalles circonstanciels ne se confondent pas avec l'intervalle de procès : ils contribuent, conjointement à d'autres unités discursives, à en déterminer l'ancrage (cf. section 2.3). Associer un intervalle de temps à l'ensemble d'un adverbial plutôt qu'à la référence temporelle qui en forme le noyau ne revient donc pas nécessairement à occulter l'intérêt d'une approche relationnelle.

(Gagnon et Lapalme, 1996) proposent également une heuristique formelle pour représenter la sémantique des adverbiaux de localisation temporelle. Cette heuristique - utilisée dans le cadre de la génération automatique de texte - propose de formaliser les adverbiaux de localisation sous la forme d'une combinaison d'opérateurs sémantiques en petit nombre permettant d'exprimer des relations entre des repères temporels. Par exemple, l'adverbial de localisation temporel « jusqu'à il y a trois jours » sera représenté par la combinaison d'opérateurs et de repères temporels suivante :

$$end([t_{loc}, \_, \_], before([t_2, \_, \_], [n, \_, \_], duration(3, day))$$

Cette représentation formelle peut être paraphrasée de la façon suivante : la zone temporelle durant laquelle s'achève le procès localisé (t_{loc}) chevauche un repère temporel (t_2) qui précède de trois jours ($duration(3, day)$) l'acte d'énonciation (n).

C'est dans la veine de cette approche que nos travaux s'inscrivent, en cherchant à représenter formellement les adverbiaux de localisation temporelle comme une succession d'opérations sémantiques agissant sur un repère temporel noyau (opération de régionalisation, de déplacement et de focalisation). L'analyse que l'on propose, décrite dans le chapitre 4, s'appuie sur la formalisation proposée par (Battistelli, 2009). On montrera qu'initialement limitée à l'analyse des adverbiaux calendaires (qui effectuent un repérage par rapport au calendrier), la modélisation peut être étendue à l'ensemble des types de repérage (déictique, anaphorique ou relatif à un procès).

2.6 Bilan du chapitre

On a vu que les adverbiaux temporels posaient des difficultés d'analyse aux linguistes à plusieurs niveaux : au niveau sémantique (quels critères définitoires permettent de circonscrire cette catégorie ?), au niveau de leur composition morphosyntaxique (les adverbiaux peuvent être constitués de différentes catégories morphosyntaxiques), et au niveau de l'analyse de leur portée, elle-même dépendante de critères dispositionnels. Les adverbiaux de localisation temporelle semblent néanmoins former une sous-classe relativement homogène de ces adverbiaux sur un plan sémantique, même s'ils sont susceptibles d'opérer une localisation à partir de repères de différentes natures (calendaires, déictiques, anaphoriques ou encore relatifs à un procès, éventuellement nominalisé).

On retrouve la trace de ces difficultés sur un plan terminologique, dans les différences de dénomination de l'objet d'analyse selon les disciplines. Ainsi, en ingénierie des langues, le terme d'« expressions temporelles » s'est progressivement imposé pour désigner les références temporelles présentes dans les textes. On a vu que cette approche tend à considérer les expressions temporelles comme des « entités nommées ». En linguistique, ce sont plus souvent des unités linguistiques prenant part à la structuration du discours qui sont considérées, articulant une notion

sémantique avec une fonction syntaxique : l'unité considérée est l'adverbial (avec toutes les difficultés liées à la définition précise de cette catégorie) et non la référence temporelle noyau - référence qu'il est difficile d'isoler en dehors de toute structure, à la fois sur un plan syntaxique et sémantique, parce qu'ainsi isolée, on perd la possibilité de poser le problème de la portée des adverbiaux, ou encore d'analyser de façon unifiée les différents types d'adverbiaux (adverbes de temps, subordonnées temporelles, syntagmes prépositionnels, par exemple, qui occupent manifestement une fonction syntaxique proche).

On a vu en outre, que les liens entre la désignation d'un événement et la détermination de ses coordonnées temporelles n'allaient pas de soi et ce d'abord parce qu'un événement n'est pas réductible à son ancrage temporel. Comme on va le voir dans le chapitre suivant, de nombreuses applications issues de travaux en ingénierie des langues et en Intelligence Artificielle cherchent à normaliser les expressions temporelles dans les textes, afin d'assigner une date ou un intervalle de dates au procès que ces expressions viennent déterminer. Or, très souvent, il n'est possible de représenter les événements et les adverbiaux de localisation temporelle sous la forme d'intervalles de temps qu'au prix d'une perte de sémantique ou d'une surdétermination de celle-ci. On verra que cette question devient centrale lorsqu'il s'agit de manipuler des représentations facilitant la mise en œuvre de raisonnements ou de calculs, puisqu'il devient alors important de pouvoir conserver différents niveaux de représentation selon les usages attendus (ou bien des intervalles ou bien des représentations formelles sous la forme d'une combinaison d'opérations sémantiques).

Chapitre 3 : Les adverbiaux de localisation temporelle : un objet pour l'ingénierie des langues et l'ingénierie des connaissances

Au cœur des programmes d'annotation automatique, la caractérisation de l'« *information temporelle* » constitue un enjeu, tant sur le plan descriptif (quelles sont les unités de la langue qui expriment une information temporelle ?) que sur le plan analytique (quels sont les niveaux de représentation et les stratégies calculatoires à mettre en œuvre pour appréhender la catégorie sémantique du temps ?).

Dans le champ de l'ingénierie des connaissances et de la recherche d'information, par rapport auquel se situent souvent ces programmes d'annotation automatique, l'information temporelle est la plupart du temps rapportée à ce qui permettrait la résolution d'une tâche en particulier : celle du calcul de l'ancrage calendaire de situations - souvent appelées « *événements* » - décrites dans les textes. La nature de cet ancrage, c'est-à-dire la façon dont on associe l'objet localisé dans le temps et les valeurs calendaires, pose toutefois des difficultés.

Dans ces applications, les expressions linguistiques référant explicitement à un calendrier (le calendrier grégorien, par exemple) constitue un champ d'investigation exploré de longue date. C'est ainsi dans le cadre de recherches entreprises autour de systèmes de questions/réponses qu'a été organisée, pour la première fois, en 2004, une tâche d'évaluation dévolue uniquement à la problématique du repérage et de la normalisation de ce type d'expressions : Time Expression Recognition and Normalization (TERN).

Cette volonté d'ancrer les situations décrites dans les textes sur le calendrier est également à l'origine de la démarche visant à proposer une standardisation de l'annotation sémantique des « expressions temporelles » (cf. en particulier (Schilder et Habel, 2001 ; Pustejovsky *et al.*, 2003 ; Ferro *et al.*, 2003 ; Saquete *et al.*, 2004)), dont ISO-TimeML est le résultat (Pustejovsky *et al.*, 2010).

L'objectif qui anime cette démarche est d'améliorer la performance des systèmes de recherche d'information, en sortant du paradigme d'un accès aux textes reposant uniquement sur l'analyse des mots-clés : ce qui est visé, c'est le *contenu* des textes (*access of information from texts through content rather than keywords* (Pustejovsky *et al.*, 2003)). En l'occurrence donc, le contenu temporel des textes, fait des événements et de leur ancrage temporel.

Cet effort s'est accompagné, au fil du temps, de l'élaboration de corpus annotés tels que ACE et TimeBank et, bien sûr aussi, d'un nombre croissant de systèmes automatiques pour annoter les « expressions temporelles » présentes dans les textes (cf. par exemple Mani et Wilson, 2000; Han *et al.*, 2006; Ahn *et al.*, 2007).

Pour autant, un constat s'impose : il y a aujourd'hui encore peu de réalisations opérationnelles du côté des systèmes de recherche d'information pour prendre en compte l'expression de critères temporels (Alonso *et al.*, 2007). Ceux-ci n'exploitent encore que très peu et très difficilement les informations temporelles exprimées dans les textes. L'unique réalisation grand-public d'envergure, le service *view:timeline* de Google a d'ailleurs été abandonné en 2011.

Du côté des systèmes de gestion des connaissances, les propriétés calendaires décrivant la localisation temporelle d'objets très divers sont des propriétés très largement répandues, au sens où l'on trouve une grande quantité d'informations localisées temporellement dans les bases de données et de connaissances. Pour ce qui nous retient ici, du point de vue de l'ingénierie des connaissances, le problème s'articule à la fois avec l'ingénierie des langues (il relève alors de l'exploitation des connaissances portées par les textes), mais aussi de la modélisation des propriétés temporelles elles-mêmes.

Dans ce chapitre, nous nous attacherons à décrire la façon dont l'ingénierie des langues s'est emparée de cet objet d'analyse désigné sous les termes d'« expressions temporelles » (section 3.1), avant de s'intéresser à deux domaines d'application dans lequel les systèmes d'annotation qui visent à extraire et décrire ces expressions sont ou pourraient être exploitées : l'ingénierie des connaissances (section 3.2) et la recherche d'information (section 3.3).

Nous souhaitons par la suite montrer comment, à partir d'une analyse linguistique, il est possible de faire émerger d'autres représentations de la sémantique des « expressions temporelles » que celles qui prédominent actuellement – termes auxquels on préfère ceux, moins ambigus, d'« adverbiaux de localisation temporelle » - et comment, à partir de ces représentations, il est possible de répondre à certaines des difficultés que rencontrent aujourd'hui les systèmes dédiés à l'acquisition de connaissances et à la recherche d'information, lorsqu'ils doivent gérer des informations temporelles.

3.1 Traitement Automatique des Langues : quel(s) objet(s) d'analyse et quel(s) usage(s) ?

S'efforçant de synthétiser les nombreux travaux menés dans le champ du traitement automatique des langues pour l'annotation des informations temporelles dans les textes, (Muller et Tannier, 2004) distinguent quatre types principaux de tâches assignées aux systèmes d'annotation :

(1) la détection de dates et de marqueurs temporels
(2) le repérage d'événements
(3) la datation d'événements
(4) la détermination de l'ordre des événements dans un texte

Ils soulignent que la première tâche, le repérage des dates (*les années 20 ; 10 juillet 2007*) et des marqueurs temporels (tels que « *avant* », « *après* », « *durant* »), semble aisé à résoudre à l'aide d'automates à états finis (Wilson *et al.*, 2001). La deuxième tâche, le repérage d'événements, pose, elle, des problèmes théoriques et conceptuels (on a vu dans le chapitre précédent que la définition d'un événement n'allait pas de soi), mais aussi technique, la catégorie impliquant de prendre en compte des verbes, des noms et des adjectifs. La troisième tâche pose des problèmes de deux ordres : d'abord, de nombreux événements ne sont pas associés à des dates dans les textes, et de plus les rapports entre les événements et les dates est lui-même problématique. Enfin l'annotation des relations entre événements présente également plusieurs niveaux de difficultés : la modélisation et le choix de ces relations et le consensus, en général très faible, entre les différents annotateurs à qui l'on confie une telle tâche.

(Muller et Tannier, 2004) soulignent également que si la première tâche, le repérage des dates et des marqueurs temporels, semble a priori aisée, elle peut néanmoins s'avérer plus complexe et nécessiter une analyse grammaticale plus profonde, s'il s'agit de considérer des adverbiaux et des subordonnées temporelles et non plus seulement des dates et des marqueurs de relation temporelle considérés isolément (*avant les années 2000 ; trois jours après le départ de Robert* ; *lors d'une messe dans la Basilique Saint-Pierre*) (Gagnon et Lapalme, 1996 ; Vazov, 2001). On ajoute qu'en s'attachant à décrire de telles unités textuelles, qui sont cohérentes d'un point de vue linguistique (il s'agit de considérer des *adverbiaux* plutôt que des *dates* et des *marqueurs* qui ne renvoient pas à de catégories morphosyntaxiques reconnues), alors le découpage des tâches tel qu'il est présenté (repérage des dates et des marqueurs temporels, repérage des événements, datation des événements) est contraint d'être revu : en effet, des adverbiaux temporels que l'on peut décrire d'une façon similaire d'un point de vue linguistique peuvent contenir des « dates », des « marqueurs temporels » ou des « événements » (*depuis 2008 ; depuis le début de la campagne électorale*). Nous reviendrons à plusieurs reprises dans ce chapitre sur cette redéfinition des tâches généralement assignées aux systèmes d'annotation des « expressions temporelles », car nous souhaitons défendre cette idée qu'une modélisation plus soucieuse de la cohérence de son objet d'analyse, d'un point de vue linguistique, doit permettre de répondre à plusieurs difficultés rencontrées par les applications qui utilisent ces systèmes d'annotation.

3.1.1 Un rapide aperçu historique

3.1.1.1 Faire émerger la structure temporelle des textes

Historiquement, la dernière tâche, celle qui consiste à ordonner les événements d'un texte et à décrire leurs relations est celle qui a d'abord fait l'objet de nombreux travaux. Ainsi (Hitzeman *et al.*, 1995), par exemple, dans le sillage des travaux alors très abondants sur les relations de discours, la cohérence et la cohésion textuelle (Hobbs, 1985 ; Dowty, 1986 ; Hovy, 1990 ; Lascarides et Asher, 1993), se sont attachés à faire émerger la structure temporelle des textes en prenant conjointement en compte les effets des temps verbaux, de l'aspect et des adverbiaux temporels. D'autres travaux (Webber, 1988) se sont attachés à déplier le jeu complexe des anaphores temporelles, à travers l'analyse des temps verbaux.

Dans le cadre de la RST (*Rhetorical Structure Theory*) (Mann et Thompson, 1988 ; Marcu, 2000), les relations de discours sont vues comme des liens sémantiques entre des unités de discours (des propositions, des phrases, ou encore des segments textuels plus vastes). Dans le cadre de la SDRT (Asher et Lascarides, 2003), ces relations opèrent sur les représentations du contenu des propositions (*Discourse Representation Structures*). Malgré de très nombreux travaux, l'annotation des relations de discours est restée problématique, du fait de l'ambiguïté des marqueurs de relation, mais aussi du fait qu'aucun consensus ne s'est jamais véritablement dégagé sur la définition et la nature des relations de discours à prendre en compte. On voit ainsi coexister des approches qui tendent à réduire le nombre des relations de discours décrites (Marcu et Echihabi, 2002 ; Saito *et al.*, 2006), ce qui a pour effet de les rendre plus ambigües, car elles couvrent alors davantage de phénomènes (cf. par exemple la relation de contraste qui désigne aussi bien des relations d'antithèse, de concession ou de rupture d'attente (« violated expectation »)) et des approches qui, au contraire, multiplient ces relations (cf. par exemple, Bärenfanger *et al.*, 2006).

D'autres travaux s'inscrivent dans cette lignée et s'attachent plus spécifiquement aux relations temporelles (Pustejovsky *et al.,* 2003b ; Mani et Pustejovsky, 2004), mais dans un esprit sensiblement différent, car il est moins question de lier entre elles des unités de discours, que des entités, les événements repérés dans les textes, auxquels on cherche à associer des coordonnées temporelles. Comme on le verra, l'objectif de ces travaux est de pouvoir répondre à des questions faisant intervenir des informations temporelles.

Les premières approches, qui procédaient à l'origine d'une démarche plus linguistique, moins directement opératoire, ont ainsi cédé progressivement le pas aux approches venues de la sphère de l'Intelligence Artificielle, orientées vers une finalité précise : nourrir des systèmes de Questions/Réponses et de recherche d'information.

3.1.1.2 Repérer et normaliser des dates

S'appuyant sur cette hypothèse qu'il est possible d'indexer et d'ordonner tous les événements d'un texte sur un calendrier, qu'il est possible de les situer sur un référentiel commun[8], tout un pan de la

[8] "Events in articles are naturally anchored in time within the narrative of a text." (Pustejovsky *et al.*, 2003)

recherche en ingénierie des langues consacré à l'analyse des expressions temporelles a fondé l'espoir de parvenir à extraire les connaissances temporelles des textes et de circuler dans les informations ainsi extraites de manière simplifiée et facilement échangeable. Une telle idée peut paraître d'autant plus raisonnable que les expressions datatives dans les textes semblent, en première approche, plutôt simples à appréhender. Il semble en effet possible de faire un tour presque exhaustif des différentes formes que peut revêtir l'expression des dates. La tâche de repérage semble ainsi pouvoir être menée à bout à relativement bon frais, à l'aide d'outils comme les automates, par exemple, qui reconstruisent les moules auxquels se conforment ces expressions (cf. par exemple les travaux de (Maurel, 1988) sur le repérage des dates en français).

Le repérage des entités nommées dans les textes (Chinchor et al, 1999), au nombre desquelles les dates sont souvent associées, vise à améliorer les outils de recherche d'information en facilitant l'accès au *contenu* des documents, en permettant l'identification des noms, des lieux ou des produits par exemple. C'est ainsi avec la 6e conférence MUC (*Message Understanding Conference*), en 1995, qu'apparaît pour la première fois une sous-tâche dédiée à l'identification des expressions temporelles dans les textes, s'ajoutant aux autres types d'entités nommées. La tâche est alors réduite au repérage de dates simples. Lors de la 7e conférence MUC (MUC-7, 1998), la sous-tâche dédiée aux expressions temporelles étend ses exigences aux expressions temporelles dites *relatives*, par opposition aux expressions *absolues*. Cette distinction, qui perdure jusqu'à aujourd'hui, était alors encore mal établie. Le terme « expression absolue » désignait alors les segments de textes définissant une zone temporelle précise (*10h20*, *minuit*, *10 octobre*), par opposition aux « expressions relatives » dont la zone temporelle pouvait varier selon le contexte (*hier*, *le mois dernier*). Avec des travaux comme ceux de (Setzer et Gaizauskas, 2000), progressivement la distinction s'affirme comme une distinction opératoire entre les expressions qui peuvent être placées sans ambiguïté sur un calendrier (*10 octobre 1999*) et celles qui requièrent d'autres informations pour pouvoir être situées sur un calendrier (en particulier, les expressions déictiques dont le repère temporel noyau est formé par le processus énonciatif).

Cette distinction étant posée, une tâche dite de *normalisation* se fait jour : elle consiste à résoudre la référence calendaire des expressions temporelles et à leur assigner une date dans un format standard, généralement le format ISO 8601. Pour les expressions absolues, la tâche paraît triviale. Elle est plus complexe avec les expressions relatives, qui peuvent être des déictiques (*hier*, *le mois dernier*) ou des anaphoriques (*la veille*, *l'année suivante*). Une idée tenace fait alors conjointement son apparition, celle qui considère que les expressions *relatives* se résolvent par rapport à la date de création d'un document (*Document Creation Date*). On mesure bien ici les différences de perspectives entre les approches de l'ingénierie des langues (qui viennent en grande partie de l'Intelligence Artificielle et dont procèdent les travaux que l'on mentionne) et les approches linguistiques qui font dépendre les expressions déictiques du référentiel énonciatif. En effet, à considérer que la résolution des expressions déictiques doit s'opérer à partir de la date du document, il devient impossible de résoudre ce type de références lorsqu'elles apparaissent dans le contexte d'un discours rapporté (on renvoie sur ce point à la section 2.2).

En outre, à y regarder de près, on peut observer que les systèmes d'annotation se cantonnent généralement à l'analyse non pas des adverbiaux de localisation temporelle, mais précisément uniquement des *expressions datatives*. Or une étude plus approfondie des segments textuels

permettant un ancrage d'un procès sur un référentiel temporel, montre que les dates ne recouvrent qu'une partie seulement des différents types d'adverbiaux de localisation temporelle. Dans les textes, en effet, les adverbiaux de localisation temporelle sont très variés : ils ne dénotent pas tous une succession simple de grains calendaires. Or, si les systèmes d'annotation automatique repèrent aisément les segments textuels identifiés comme des dates (« *le 3 avril 1986* »), en revanche, ils s'intéressent rarement à la fonction discursive qu'ils occupent, en général au sein des adverbes de localisation temporelle – adverbes dont la formation peut s'avérer très complexe (« *deux jours avant le Nouvel An* », « *dès le lendemain* », « *depuis des années* », « *depuis plus d'un siècle* », etc.). Les adverbiaux de localisation temporelle qui contribuent à ancrer des procès sur un référentiel temporel débordent ainsi largement l'expression des dates (Gagnon et Lapalme, 1996).

Ce type d'approches tend ainsi à considérer les « expressions temporelles » (aussi bien d'ailleurs que les « événements ») comme une forme particulière d'*entités nommées* : or, comme on l'a vu dans le chapitre précédent (cf. section 2.1), les noms de temps que l'on peut rapprocher d'un point de vue syntaxique et sémantique des noms propres forment des catégories très particulières (les chrononymes et les héméronymes) – catégories qui ne sauraient recouvrir celles des adverbiaux qui permettent de contribuer à l'ancrage dans le temps d'un procès. Or les « expressions datatives », le plus souvent, occupent une fonction adverbiale (*en juin, l'an passé, jusqu'aux années 80*, etc.).

On retrouve encore aujourd'hui la trace d'une telle approche dans TimeML, même si la catégorie regroupée sous le terme générique d'*expression temporelle* (TIMEX), s'est ouverte pour accueillir les déictiques, les anaphoriques, les durées et les fréquences (Ferro *et al.*, 2001 ; Mani et al, 2001).

3.1.1.3 Ancrer sur le calendrier les situations décrites dans les textes

Avec les conférences MUC-5, puis MUC-7, on voit aussi apparaître une nouvelle tâche qui invitait les participants à assigner une date à des événements prédéfinis en s'appuyant sur l'analyse de textes (l'annonce de fusion acquisition et le lancement de roquettes). Bien que la tâche fût alors très limitée, les systèmes évalués obtinrent des scores très faibles. La voie était malgré tout définitivement ouverte à l'extension du repérage des informations temporelles, qui, en plus des expressions datatives, intègre désormais les événements. On voit ainsi se multiplier les travaux sur les corpus de presse (Filatova and Hovy, 2001 ; Schilder et Habel, 01 ; Setzer, 2001 ; Setzer et Gaizauskas, 02), qui visent notamment à assigner des dates aux événements repérés.

C'est dans ce contexte qu'apparaît TimeML (Pustejovsky *et al.*, 2002 ; Pustejovsky *et al.*, 2003 ; Saurí *et al.*, 2006 ; Saurí et Pustejovsky, 2009 ; Pustejovsky *et al.*, 2010), un langage d'annotation des informations temporelles dans les textes. Issu de l'atelier TERQAS[9] qui lui-même prenait part au projet AQUAINT[10], ce langage d'annotation est conçu à l'origine pour améliorer les systèmes de questions/réponses, en les nourrissant d'informations temporelles caractérisant des événements, afin de permettre à ces systèmes de répondre à des questions faisant intervenir des critères

[9] http://www.timeml.org/site/terqas/index.html
[10] http://www-nlpir.nist.gov/projects/aquaint/

temporels. Voici trois exemples de questions fournies comme exemples de ce qu'on pourrait attendre de tels systèmes (Pustejovsky *et al.*, 2003) :

> a. Is Gates currently CEO of Microsoft? (*Gates est-il aujourd'hui PDG de Microsoft ?*)
> b. When did Iraq finally pull out of Kuwait during the war in the 1990s? (*Dans les années 1990, quand l'Irak s'est-il finalement retiré du Koweït durant la guerre ?*)
> c. Did the Enron merger with Dynegy take place? (*La fusion d'Enron avec Dynegy a-t-elle eu lieu ?*)

Le langage d'annotation TimeML s'est aujourd'hui très largement imposé dans la communauté scientifique. Les campagnes d'évaluation des systèmes d'annotation des informations temporelles s'appuient généralement sur cette spécification (TempEval-1 en 2007 (Verhagen *et al.*, 2007) et TempEval-2 en 2010 (Verhagen *et al.*, 2010)). TimeML a également servi de référence pour l'annotation du vaste corpus TimeBank (Pustejovsky *et al.*, 2003b)[11] et il est aujourd'hui en cours de certification pour devenir un standard (*ISO-TimeML* (Pustejovsky *et al.*, 2010)).

TimeML répond à plusieurs objectifs : (1) repérer des expressions temporelles et des événements, (2) décrire les références datatives sous une forme normalisée (c'est-à-dire les ramener au format ISO 8601), (3) assigner des dates aux événements décrits dans les textes (et donc lier des références temporelles à des événements), (4) décrire les liens qu'entretiennent les événements décrits dans les textes. Quatre grandes structures de données sont spécifiées par le langage : les balises EVENT (pour la description des événements), TIMEX3 (pour la description des expressions temporelles), SIGNAL (pour marquer les indices de relations entre intervalles de temps) et LINK (pour la description des relations de dépendances).

Dans TimeML, les expressions temporelles et les événements sont annotés comme des isolats, détachés de leur structure discursive : il ne s'agit pas en effet d'annoter des adverbes de temps et les procès sur lesquels ils portent, mais plutôt, dans l'esprit des systèmes d'annotation des entités nommées, des segments textuels isolés. Voici, par exemple, une phrase annotée avec TimeML (Hobbs et Pustejovsky, 2003) :

John left 2 days before the attack. (*John est parti deux jours avant l'attaque*.)

```
John
<EVENT eid="e1" class="OCCURRENCE">
left
</EVENT>
<MAKEINSTANCE eiid="ei1" eventID="e1" tense="PAST" aspect="PERFECTIVE"/>
<TIMEX3 tid="t1" type="DURATION" value="P2D" temporalFunction="false">
2 days
</TIMEX3>
<SIGNAL sid="s1">
before
<SIGNAL>
the
```

[11] http://timeml.org/site/timebank/timebank.html

<EVENT eid="e2" class="OCCURRENCE">
attack
</EVENT>
<MAKEINSTANCE eiid="ei2" eventID="e2" tense="NONE" aspect="NONE"/>
<TLINK eventInstanceID="ei1" signalID="s1" relatedToEventInstance="ei2"
relType="BEFORE"/>.

Dans l'exemple, l'expression « *two days before the attack* » (*deux jours avant l'attaque*) n'est pas analysée comme un adverbe de temps portant sur le procès « *John left* », si bien que rien n'est dit du rapport entre la durée « *two days* » et l'événement « *attack* », ni non plus entre l'adverbe et le prédicat « *left* ».

Isoler ainsi événements et expressions temporelles au sein des structures discursives en rend ensuite l'exploitation difficile pour les systèmes de recherche d'information. Ceci explique sans doute pourquoi les équipes qui travaillent à la spécification de TimeML en sont venues à s'intéresser aux « arguments » d'un événement (Pustejovsky *et al.*, 2006), rejoignant ainsi les efforts d'autres groupes de recherche qui cherchent à décrire les liens sémantiques entre propositions en leur assignant des « rôles sémantiques » (Palmer *et al.*, 2005).

On voit ainsi peu à peu s'ajouter aux premiers éléments visés par l'analyse de TimeML, la dimension textuelle, notamment la relation arguments/prédicats/déterminants du prédicat, afin de compléter l'analyse des expressions désignant des événements. Ainsi, si l'annotation des informations temporelles ne met l'accent à l'origine que sur certains aspects de la temporalité (les références au calendrier et les événements), très vite néanmoins l'ensemble des aspects du problème de l'analyse de la localisation temporelle évoqués dans le chapitre précédent finit par être impliqué : les liens avec le référentiel énonciatif pour la résolution des expressions déictiques, la prise en compte de la coréférence pour la résolution des anaphoriques, la délimitation des adverbiaux pour déterminer la relation qui lie adverbiaux et événements, la définition des « événements », etc.

Pour autant, les fonctions syntaxiques et sémantiques des expressions temporelles elles-mêmes ne sont pas véritablement abordées en tant que telles. Les adverbiaux de localisation temporelle ne sont pas, dans ces approches, analysés comme des segments textuels contribuant à ancrer un procès sur un référentiel temporel, mais on l'a vu, comme des formes d'*entités nommées* d'un genre particulier. Ces travaux ne s'attachent donc pas à décrire les rapports complexes qu'entretiennent les références temporelles et les événements ou situations décrites par les textes. Ainsi, qu'un événement (au sens de TimeML) ou une référence calendaire puisse, l'une comme l'autre, occuper une fonction adverbiale n'est pas pris en compte : dans l'exemple ci-dessus (« *John left two days before the attack* »), la référence au cœur de l'adverbial est un « événement » (« *two day before the attack* »), mais l'annotation ne permet pas de voir que cet événement contribue à situer temporellement le procès. L'ancrage temporel est en outre toujours ramené à un ancrage sur le référentiel calendaire. C'est pourquoi aussi les adverbiaux déictiques (*hier*, *en octobre dernier*) sont considérés comme des expressions « *sous-spécifiées* » (*Underspecified Temporal Expressions*) et non comme des expressions dont la référence temporelle est liée au processus énonciatif.

3.1.2 Les différents types d'informations temporelles visées par les systèmes d'annotation

Les unités textuelles visées par les systèmes d'annotation des informations temporelles se répartissent généralement en trois classes : celles qui dénotent des *expressions temporelles* (dates, durées, fréquence, etc.), celles qui dénotent des *événements* et enfin celles qui dénotent des *relations temporelles* entre les événements et entre les événements et les expressions temporelles.

Le repérage et l'extraction des informations temporelles posent un certain nombre de difficultés communes à toute tâche d'annotation dont les principales sont :

- la nature des marqueurs à prendre en compte,
- l'ambiguïté de ces marqueurs,
- l'identification des bornes initiales et finales,
- le choix des étiquettes à attribuer.

Elles présentent également des difficultés qui leur sont propres, comme la normalisation des expressions temporelles dites *relatives* (les déictiques et les anaphoriques). Là encore, l'ambiguïté des marqueurs présente une réelle difficulté, comme l'illustrent les deux exemples suivants :

> *(1) Le ministre est venu [3 minutes après son porte-parole] qui avait déjà annoncé la bonne nouvelle.*
> *(2) La réunion a commencé et [3 minutes après] son porte-parole qui avait déjà annoncé la bonne nouvelle est parti pour la capitale.*

Dans ces deux exemples tirés de (Vazov, 2001), l'analyse syntaxique conduit à annoter différemment les unités textuelles dénotant une référence temporelle, en dépit de leurs similitudes : les deux adverbiaux ne sont pas délimités de la même manière. Ainsi, dans le second exemple, l'adverbial temporel est anaphorique (la référence noyau est le prédicat « *a commencé* »), alors que dans le premier exemple, la référence noyau est le syntagme nominal « *son porte-parole* ». Il faut remarquer en outre que, sur un plan sémantique, il est problématique, dans cet exemple, d'annoter le syntagme « *son porte-parole* » comme un « événement » : le syntagme est une élision qui reprend de façon anaphorique le verbe *venir* (on pourrait paraphraser l'adverbial ainsi : *3 minutes après que son porte-parole est venu*). Cet exemple illustre bien la difficulté de l'annotation des événements et de la résolution des références temporelles dans les adverbiaux de localisation temporelle.

3.1.2.1 Les expressions temporelles

A quels critères précis reconnaît-on dans les textes une « expression temporelle » ? Dans la spécification TimeML, les expressions temporelles recouvrent des *dates*, des *durées* et des *agrégats*, ces notions étant essentiellement définies à l'aide d'exemples. S'ils suggèrent de rester compatibles avec cette spécification, les travaux de (Ehrmann et Hagège, 2009) soulignent les difficultés engendrées par certaines des propositions de TimeML pour l'annotation des expressions temporelles (TIMEX). Ces travaux visent ainsi à systématiser les critères permettant d'étiqueter un segment textuel comme *expression temporelle*. Le premier des critères définitoires pour caractériser une

expression temporelle est un critère d'ordre *sémantique* : une expression temporelle doit permette de répondre à l'une des questions suivantes : *Quand ? Combien de temps ? A quelle fréquence ?* Le second critère définitoire mêle un trait syntaxique et sémantique : une expression temporelle doit contenir au minimum une unité lexicale de la liste suivante :

- un patron temporel caractéristique, plus ou moins complet (20/02/2009, le 3)
- une unité de mesure temporelle (seconde, jour, année, etc.) ou un adverbe en –ment dérivé de ces unités
- un substantif nommant un élément calendaire (lundi, mars), une saison, une fête du calendrier
- un nom désignant un moment particulier de la journée (matinée, soir), ainsi que des noms génériques de périodes temporelles (époque, moment, instant, etc.)
- un indexical temporel : soit un adverbe de temps simple, soit un nom « situant les faits dans la durée par rapport au moment de la parole ou un autre repère » (Grévisse & Goose, 1986). Par exemple, *jadis* ou *hier* pour les adverbes, *veille* et *surlendemain* pour les noms.
- Un groupe prépositionnel complément d'un nom événementiel (quantité + unités de temps).
- Une expression de fréquence de type de *temps en temps, quelques fois, fréquemment*.
- Une expression construite avec un présentatif (il y a, cela fait) suivi d'un des éléments précédents.

Cette approche lexicale, qui s'attache à recenser les noms qui entrent dans la composition des adverbes de référence temporelle, est à rapprocher des travaux de (Borillo, 1998). L'originalité de cette approche, sur un plan syntaxique, est de faire remarquer que le problème de la délimitation des expressions temporelles a partie liée avec celui de leur caractérisation. Ainsi, selon la façon dont on délimite l'expression temporelle telle que « *pendant ces deux jours* », on peut la qualifier de DATE (si on retient l'ensemble de l'expression) ou de DUREE (si, comme le propose TimeML, on isole « *deux jours* », puisque les prépositions n'entrent pas dans les expressions temporelles). Les travaux de (Erhmann et Hagège, 2009) posent ainsi de façon originale la question de la délimitation des expressions temporelles, en la faisant dépendre de l'analyse syntaxique. Ainsi, pour ce qui est de la borne gauche, tous les dépendants syntaxiques font partie intégrante de l'expression temporelle, un choix différent de celui proposé dans TimeML, puisqu'il intègre les prépositions, alors que dans TimeML elles sont exclues et généralement annotées comme des marqueurs de relations temporelles (annotés avec la balise SIGNAL). Cette approche permet de différencier des expressions comme *il y a 3 ans* et *pendant 3 ans*, tout en permettant de garder la trace d'informations importantes pour la résolution des références temporelles nécessitant un calcul (*ce lundi* vs. *le lundi*).

Pragmatique, cette approche ne cherche toutefois pas à délimiter des adverbiaux complets, car cette tâche s'avère complexe lorsqu'il s'agit de repérer des subordonnées ou des adverbiaux enchâssés (cf. ex. 2 ci-dessous). Ainsi, au niveau de la borne droite, sont intégrés dans les expressions temporelles les adjectifs, adverbes, prépositions ou noms faisant partie des listes de syntagmes noyaux détaillés plus haut.

Ex. 1 : Il part au printemps de l'année prochaine.
Ex. 2 : Au troisième jour de sa visite d'état en Chine, …

Il semble qu'en poursuivant la démarche jusqu'au bout, l'ensemble de l'adverbial de l'exemple 2 devrait être considéré comme formant une expression temporelle, ce qui semble plus cohérent d'un

point de vue syntaxique et sémantique. Cela signifierait néanmoins, qu'il ne s'agirait plus d'opposer *événement* et *expression temporelle*, mais plutôt d'analyser les adverbiaux temporels et les objets sur lesquels ils portent, conduisant de ce fait à s'éloigner du modèle d'annotation spécifié dans TimeML. Pour autant, et au-delà des difficultés techniques propres à cette approche plus linguistique, elle ne serait pas incompatible avec l'annotation des événements et des expressions temporelles, mais il s'agirait alors, par-delà ces annotations, de faire apparaître des éléments de la structuration discursive.

Les expressions absolues

Répondant à un critère opératoire, les expressions temporelles absolues (désignées comme des dates « explicites » ou « concrètes », dans TimeML) sont celles qui ne nécessitent aucun élément additionnel d'analyse (*additional information*) : elles pointent sans ambiguïté sur une zone précise du calendrier. Si elles sont simples à identifier, leur délimitation n'en est pas moins problématique, selon qu'on se limite à y voir des dates isolées ou que l'on souhaite au contraire récupérer l'ensemble de l'adverbe dans lequel la référence calendaire s'insère. Cette dernière approche, qui prend en compte leur fonction syntaxique et sémantique, conduit ainsi à exclure certaine expressions absolues lorsqu'elles apparaissent comme sujet d'une phrase (« *ce 14 juillet 1989 fut rude* »).

Les expressions relatives

La normalisation d'une expression temporelle relative (une expression déictique ou anaphorique) consiste à calculer la valeur effective de sa référence calendaire : ainsi, normaliser une expression telle que « *demain* » consiste à lui associer une valeur calendaire, mettons, *le 12 octobre 2010*, par exemple. La valeur calendaire associée est généralement représentée dans un format entièrement spécifié, comme le format ISO 8601. Cet aspect du problème est un peu passé sous silence dans TimeML, qui ne souhaite pas favoriser un algorithme particulier de normalisation au détriment d'un autre. Orienté vers une finalité précise, nourrir des systèmes de questions/réponses, le langage d'annotation ne décrit pas la sémantique propre à des expressions comme « *hier* » : seule la valeur calendaire effective pointée par l'expression importe.

Cependant, la modélisation des expressions temporelles prend une large part dans la capacité qu'ont ensuite les systèmes à traiter les expressions relatives qui renvoient à l'acte d'énonciation : il faut en effet pouvoir s'appuyer sur une description fine des éléments linguistiques qui entrent dans la composition des expressions temporelles pour pouvoir les « normaliser ». En outre, l'approche retenue dans TimeML suppose implicitement que l'on peut toujours associer une date à une expression temporelle. Or la représentation temporelle dans les textes n'est pas sémantiquement équivalente à la représentation calendaire. Il est des cas en effet où la valeur calendaire associée à une expression donnée ne peut être qu'une proposition de normalisation, qui doit pouvoir être ajustée en fonction des besoins ou du contexte. Pour des expressions telles que « *fin juin* », « *vers la mi-mars* », « *l'hiver prochain* », il n'y a pas de transposition équivalente dans le modèle ISO 8601. Par exemple, s'il est possible de dire avec vraisemblance que *le 29 juin* est bien inclus dans la zone temporelle dénotée par une expression comme *la fin du mois de juin*, en revanche, selon le contexte,

il se peut que *le 19 juin* le soit ou ne le soit pas : il n'y a donc pas de transposition univoque de cette expression vers la représentation calendaire.

Les durées

Les durées sont le plus souvent, comme les autres expressions temporelles, annotées comme des entités nommées, non pas des adverbes (*pendant trois mois* ou *durant deux ans*). TimeML précise ainsi qu'elles sont composées d'une valeur numérique et d'une unité calendaire cardinale (*3 mois, 2 ans*). Les durées qui font intervenir des quantifications non numériques (*quelques années*) ou des grains non calendaires (*peu de temps, en quelques instants*) sont plus difficiles à appréhender.

Comme on l'a vu, la délimitation des segments textuels annotés comme des durées peut être problématique, car elles peuvent intervenir dans des adverbiaux de localisation temporelle dont on perd alors la cohérence (*deux mois avant le début de l'année*).

Les itératifs

Les itératifs (annotés à l'aide de la balise SET) sont définis très succinctement dans le guide d'annotation de TimeML (Saurí *et al.*, 2006). Ils semblent recouvrir essentiellement les fréquences et faire écho à la notion de récurrence dans la norme ISO 8601, dans laquelle on peut définir des fréquences du type, *3 fois par mois*. S'efforçant de clarifier cette catégorie, (Erhmann et Hagège, 2009) proposent de la requalifier en la désignant comme l'ensemble des expressions présentant un « ancrage multiple » sur le calendrier (*agrégats*).

S'inscrivant dans le sillage des travaux très complets sur les itératifs réunis dans le rapport Ogre (Bécher *et al.*, 2005), (Lebranchu, 2011) et (Lebranchu et Mathet, 2011) ont cherché à étendre le standard TimeML, afin de pouvoir traiter plus finement les phénomènes d'itérations. Le modèle des phénomènes d'itération proposé dans (Mathet, 2007) articule les notions d'itérateurs et de sélection. La notion d'itérateur est divisée en quatre classes : (1) les itérateurs calendaires (*tous les jeudis*, *tous les weekends*), (2) les itérateurs fréquentiels (*souvent, rarement, parfois*), (3) les itérateurs quantificationnels (*3 fois, à cinq reprises*), et (4) les itérateurs événementiels qui renvoient à certaines subordonnées temporelles (*Lorsque John enseigne à l'IUT, il dépose d'abord les enfants à l'école.*) (cf. (Lebranchu et Mathet, 2011)). La notion de sélecteur vise à modéliser la sélection de certains motifs itérés au sein d'un ensemble plus vaste :

> Dans « Nous sommes allés sept fois à la montagne. Parfois, des amis nous ont accompagnés », le procès modèle *aller à la montagne* est répété par le déclencheur quantificationnel [sept fois]. Puis s'opère une sélection [Parfois], rattachée à l'itération du procès précédent, qui se voit enrichie du procès modèle *des amis nous accompagner*.
>
> (Lebranchu et Mathet, 2011)

Dans cet exemple, l'itératif quantificationnel « sept fois » précise combien de fois s'est répété le procès *aller à la montagne* et l'itérateur fréquentiel « parfois » opère une sélection dans cet ensemble de répétition du même procès. Cette modélisation permet d'affiner dans TimeML la

représentation des itératifs, en distinguant différents types d'expressions itératives et en précisant comment elles interagissent entre elles et avec les procès considérés. On verra également par la suite que les itératifs peuvent se composer entre eux pour former des circonstanciels complexes, dont il est intéressant de faire ressortir les liens (*tous les jours, sauf le dimanche, du 1er avril au 30 septembre*).

Les expressions composées

(Erhmann et Hagège, 2009) évoquent également le problème des expressions temporelles composées et de la délimitation des unités minimales qui les composent. Pour déterminer si plusieurs expressions temporelles forment une expression complexe ou bien si elles doivent être analysées comme des expressions juxtaposées, les auteurs proposent une méthode pour décider de la scission ou de l'unification des expressions temporelles imbriquées. Deux critères sont retenus pour décider de la séparation en plusieurs expressions temporelles : si les combinaisons des expressions minimales avec le procès syntaxique sont toutes à la fois (1) syntaxiquement valides et (2) sémantiquement logiques (validité vériconditionnelle), alors il y a plusieurs expressions temporelles. Dans l'énoncé « Il est parti pendant deux jours avant Noël », l'expression peut ainsi être divisée en deux expressions, alors que dans l'énoncé « Il est tombé deux jours avant Noël » le critère 1 étant violé, l'expression est indivisible (*Il est tombé 2 jours).

Dans ces travaux cependant, rien n'est dit des liens qui peuvent exister entre les expressions séparées à l'aide de ces critères : nous verrons qu'il est possible de caractériser les liens entre les éléments constitutifs des adverbiaux composés (cf. section 4.2.6.2). La caractérisation de ces liens est importante, car elle permet de préciser la façon dont se réalise l'ancrage sur le calendrier : une expression comme *tous les dimanches de 14h à 17h, du 15 août au 15 septembre* renvoie à des ensembles d'intervalles, qu'on ne peut déterminer correctement qu'en prenant en compte l'ensemble de l'expression.

3.1.2.2 Les événements

Sous l'étiquette « événement », TimeML invite à annoter des verbes conjugués, des adjectifs et des noms qui correspondent à des événements ou des états, en leur ajoutant des attributs comme la catégorie d'un événement (perception, état, verbe aspectuel, etc.), le temps des verbes (présent, passé, futur), des informations sur l'aspectualité (imperfectif, progressif), sur la présence d'une négation, sur la modalité et sa cardinalité, si l'événement se produit plusieurs fois.

TimeML définit de façon très lâche la catégorie des événements. Y sont ainsi définis comme événements des situations qui ont lieu ou qui surviennent (« situations that *happen* or *occur* »). Ces événements peuvent être ponctuels ou s'étendre sur une certaine durée.

Cette définition a minima des événements peut s'avérer problématique et laisse dans l'ombre ce qu'on attend de leur annotation : s'agit-il de repérer des faits saillants qui organisent la mémoire sociale et sur lesquels il est intéressant de réunir des informations, pour pouvoir répondre à des questions à leur sujet ou constituer des bases de connaissances ? ou s'agit-il de dégager l'ensemble

des structures prédicatives des textes ? On peut se demander en effet si dégager la structuration temporelle des textes nécessite les mêmes traitements que la mise en œuvre de systèmes de recherche d'information.

Comme on l'a vu (cf. section 2.1), des travaux en analyse du discours sur des corpus de presse montrent que les désignations d'événements, au sens de faits saillants organisant la mémoire sociale, font l'objet de jeux langagiers très productifs. Ainsi, dans l'énoncé suivant, il faut comprendre que « poulet fou » renvoie à une crise sanitaire qui fait écho à celle de la « vache folle » : « C*ertes, le « poulet fou » confirme qu'il faut une Europe sanitaire mais laquelle?* »). Annoter ce type d'information pour nourrir des systèmes de recherche d'information nécessiterait des procédures très spécifiques qui sont encore à définir. Dégager la structure temporelle des textes (telle que l'ordonnancement des procès et leur ancrage calendaire) n'est qu'une des stratégies devant permettre de traiter ce problème.

Par ailleurs, dans TimeML, la question de la délimitation des événements à la surface des textes est traitée d'une façon très pragmatique : il ne s'agit pas d'annoter des syntagmes complets ou des structures prédicatives, mais seulement le terme considéré comme la tête (*head word*) d'un événement. C'est aussi, en un sens, la même approche qui est retenue pour l'annotation des expressions temporelles, puisque seule est annotée la référence calendaire noyau et non l'adverbe. Ainsi, dans une expression telle que « la 2nde guerre mondiale », seul sera annoté comme événement le terme « guerre ». Ceci engendre parfois des difficultés. (Bittar et Danlos, 2009) mentionnent notamment le cas des verbes supports pour lesquels il n'est pas évident de savoir s'il faut considérer le verbe comme tête de l'événement ou bien s'il n'est pas plus pertinent de choisir son objet (*Marie a subi une agression*). Pour répondre à ces difficultés, les concepteurs de TimeML entendent progressivement étendre l'annotation aux arguments des événements (Pustejovsky *et al.*, 2006) : vu sous l'angle linguistique, cela consiste donc à essayer de faire émerger les structures prédicatives.

3.1.2.3 Les relations entre événements et expressions temporelles

Dater les événements

Si TimeML vise a priori l'annotation de tous les événements des textes, certains semblent toutefois présenter un intérêt plus grand que les autres, compte-tenu des finalités du projet d'annotation : ceux qui sont voisins d'une expression temporelle, car il semble alors possible de les ancrer sur le calendrier. Ce problème, présenté comme l'assignation d'une date à un événement, n'est pas trivial : d'abord parce que les textes n'ancrent pas l'ensemble des situations qu'ils décrivent sur le référentiel calendaire, mais aussi parce que les circonstanciels de localisation, même calendaires, ne sont pas toujours réductibles à des dates (*depuis 2004* vs. *2004*). En outre, les liens qui peuvent unir un procès et un adverbial de localisation temporelle peuvent nécessiter de prendre en compte conjointement l'intervalle de l'énonciation, celui du procès, celui d'un intervalle de référence, et celui de l'intervalle circonstanciel (cf. section 2.3 et en particulier les travaux de (Gosselin, 1996 ; Gosselin, 2005a)).

Ordonner les événements entre eux

On a vu que les travaux récents autour de cette question (Mani et Pustejovsky, 2004 ; Muller et Tannier, 2004 ; Bejan et Harabagiu, 2010) rejoignaient, de façon renouvelée, les travaux plus anciens sur les relations de discours. Ce n'est toutefois plus tant sous la forme de relations de discours que la structuration temporelle est appréhendée, mais davantage sous la forme de relations entre intervalles de temps. Chaque événement est en effet considéré comme un intervalle de temps dont on peut décrire les liens avec les autres événements (i.e. avec d'autres intervalles de temps) en termes de relation d'Allen (Muller et Tannier, 2004).

Le langage d'annotation TimeML repose sur un parti pris fort de distinguer très nettement les expressions temporelles et les événements, alors que les deux notions, comme on l'a vu, peuvent entretenir des rapports complexes. Un événement peut entrer dans la composition d'un adverbe de temps (« *John left two days before the attack* »). Un adverbial de localisation temporelle peut en outre fournir des informations sur la perspective à travers laquelle l'événement noyau est considéré : un grain calendaire peut ainsi permettre une focalisation sur une zone temporelle plus spécifique comme dans le circonstanciel suivant « *au troisième jour de sa visite en Chine* ». On formulera par la suite plusieurs propositions pour tenir compte et décrire ces informations (cf. chapitre 4).

3.1.3 Un point sur les systèmes d'annotation aujourd'hui

3.1.3.1 Les campagnes d'évaluation

La nature des exigences retenues et proposées par les campagnes d'évaluation permet de dresser un état des lieux des technologies mises en œuvre pour y répondre ; elle permet ainsi de mesurer ce que l'on sait faire aujourd'hui et ce qui oppose encore des résistances, ainsi que les systèmes qui sont le mieux à même de prendre en charge les tâches évaluées.

TempEval-2, la dernière campagne en date (Verhagen *et al.*, 2010) pour évaluer les moteurs d'annotation des informations temporelles, s'appuie sur TimeML, mais ne couvre pas l'ensemble des tâches auxquelles le langage de spécification aspire. Elle prend part à une campagne d'annotation plus vaste, SemEval. La campagne a fixé des objectifs jugés raisonnables eu égard à la campagne antérieure de 2007 (Verhagen *et al.*, 2007), tout en prolongeant et élargissant l'expérimentation précédente. La nature même des tâches proposées à l'évaluation donne déjà une mesure des progrès accomplis par les systèmes de traitement automatique des textes.

Historiquement, les conférences MUC (Messages Understanding Conferences) ont été pionnières dans l'évaluation des moteurs d'annotation. Organisées par le DARPA et le NOSC (Naval Ocean System Center), les sept conférences étalées de 1987 à 1998 avaient pour objectif de mesurer et de dynamiser les efforts de recherche dans le traitement automatique de « messages ». Ces conférences ont introduit les critères d'évaluation désormais classiques (la mesure des taux de précision et de rappel), obtenus en comparant les résultats fournis par les systèmes automatiques testés avec ceux fournis par les experts ayant préparé les données à tester. La partie assignée à l'annotation des informations temporelles, regroupées avec les entités nommées, visait essentiellement les expressions temporelles (et parmi elles, essentiellement les dates).

MUC (Grishman et Sundheim, 1996) et ACE (Doddington *et al.*, 2004) sont les premières campagnes d'évaluation où les tâches de reconnaissance et d'extraction des événements apparaissent. Cependant, le périmètre des événements couverts était limité à des domaines particuliers, et donc à un vocabulaire restreint (des messages de la marine américaine, des récits d'attentats terroristes en Amérique du Sud, des messages concernant des joint-ventures). En outre, il s'agissait de remplir automatiquement des champs de formulaires prédéfinis (des modèles à trou des informations à capter). Avec TimeML et les campagnes TempEval, on a vu apparaître les premières expérimentations pour évaluer l'annotation des événements de façon indépendante d'un domaine (rappelons toutefois que TimeML vise en premier lieu l'annotation de corpus de presse).

Les entités visées aujourd'hui ont donc été élargies aux événements (ce qui est déjà beaucoup) et les campagnes d'évaluation commencent à faire entrer des éléments de structuration textuelle (sur des phénomènes encore très locaux), en proposant aux outils en lice de capter des phénomènes qui jouent au niveau phrastique ou au niveau de phrases contigües.

TempEval 2 définit ainsi plusieurs tâches :

A. Repérage des expressions temporelles telles qu'elles sont définies par le modèle TimeML TIMEX3. En plus du repérage, la tâche demande à ce qu'elles soient typées (DATE, TIME, DURATION ou SET) et que leur valeur calendaire soit déterminée (tâche de normalisation)
B. Repérage des événements tels qu'ils sont définis par le modèle TimeML EVENT[12]. Il était également demandé que soit déterminées les valeurs des attributs CLASS, TEMPS, ASPECT, POLARITE et MODALITE.
C. Déterminer la relation temporelle entre un événement et une expression temporelle d'une même phrase (si l'événement « domine » syntaxiquement l'expression temporelle ou qu'ils appartiennent tous deux à une phrase nominale.
D. Déterminer la relation temporelle entre un événement et la date de création du document (DCT).
E. Déterminer la relation temporelle entre deux événements principaux dans deux phrases consécutives.
F. Déterminer la relation temporelle entre deux événements lorsqu'un événement « domine » syntaxiquement un second événement.

3.1.3.2 Les systèmes d'annotation

Les deux modules d'analyse sémantique décrits dans (UzZaman et Allen, 10) semblent résumer très bien l'état de l'art aujourd'hui pour ce qui concerne le repérage et l'analyse sémantique des expressions temporelles et des événements dans les textes. Il va même plus loin que les objectifs fixés par la campagne d'évaluation TempEval-2 (Verhagen *et al.*, 2010), à laquelle il a participé, puisqu'il permet aussi de décrire des éléments de structuration prédicative.

Un des intérêts majeurs de ces deux outils, c'est qu'ils fonctionnent sur des textes bruts, là où la campagne d'évaluation permettait d'utiliser ponctuellement des annotations prédéfinies, en

[12] Notons que le taux d'accord inter-annotateurs pour cette tâche est de 64% pour l'identification des « événements nominaux » et de 80% pour les « événements verbaux ».

particulier pour la sous-tâche consistant à détecter des relations entre événements et entre événements et expressions temporelles. En effet, afin d'isoler l'évaluation de cette sous tâche en la séparant bien des autres, le repérage des expressions temporelles et des événements était fourni.

Cette expérimentation semble ainsi la plus aboutie, en particulier quant à la caractérisation des événements (aspect, modalité, polarité, etc.). Elle repose sur une approche hybride qui mêle des ressources linguistiques (des patrons développés à la main) pour extraire des événements et des expressions temporelles à partir des « formes logiques » produites en sortie d'un outil de classification, qui repose, lui, sur des méthodes d'apprentissage automatique. Ce premier module, le parseur TRIPS (Allen *et al.*, 2008), fournit la structure prédicative ou « forme logique » après analyse des textes[13]. Il s'appuie sur un étiqueteur morphosyntaxique ainsi que sur une ontologie linguistique, qui décrit des types et des rôles sémantiques, ou *Formes Logiques* (actes de langage, etc.) reliant des arguments syntaxiques et sémantiques. Le parseur repose sur la Logique de Réseau Markovienne (*Markov Logic Network*), qui, par apprentissage automatique, attribue des poids à des formules de logique du premier ordre pour la classification.

Voici un exemple de sortie après analyse par le parseur TRIPS de la phrase « He fought in the war » (*Il participa à la guerre*) :

```
(SPEECHACT V1 SA-TELL :CONTENT V2)
(F V2 (:* FIGHTING FIGHT) :AGENT V3 :MODS
(V4) :TMA ((TENSE PAST)))
(PRO V3 (:* PERSON HE) :CONTEXT-REL HE)
(F V4 (:* SITUATED-IN IN) :OF V2 :VAL V5)
(THE V5 (:* ACTION WAR))
```

Des patrons conçus manuellement sont alors appliqués sur la sortie du parseur pour annoter des événements. Ces patrons sont indépendants des langues, dans la mesure où ils travaillent directement sur les formes logiques.

Voici un exemple de règle qui capte des syntagmes nominaux exprimant un événement :

```
((THE ?x (? type SITUATION-ROOT))
  -extract-noms>
(EVENT ?x (? type SITUATION-ROOT)
  :pos NOUN :class OCCURRENCE ))
```

Comme le mot « guerre » (war) est de type action et appartient à la catégorie des « situations » dans l'ontologie TRIPS, la règle va l'annoter comme un événement en reconnaissant le motif de la forme logique :

```
(THE V5 (:* ACTION WAR))
```

[13] Le système peut être testé à l'adresse suivante :
http://www.cs.rochester.edu/research/cisd/projects/trips/parser/cgi/web-parser-xml.cgi

Le produit de l'annotation à ce stade est le suivant :

<EVENT eid=V2 word=FIGHT
pos=VERBAL ont-type=FIGHTING
class=OCCURRENCE tense=PAST
voice=ACTIVE aspect=NONE
polarity=POSITIVE
nf-morph=NONE>
<RLINK eventInstanceID=V2
ref-word=HE
ref-ont-type=PERSON
relType=AGENT>
<SLINK signal=IN
eventInstanceID=V2
subordinatedEventInstance=V5
relType=SITUATED-IN>
<EVENT eid=V5 word=WAR pos=NOUN
ont-type=ACTION
class=OCCURRENCE
voice=ACTIVE
polarity=POSITIVE
aspect=NONE tense=NONE>

Ainsi, l'outil extrait des événements et les annote avec le standard TimeML.

Le second ensemble d'outils, TRIOS, chapote le premier, et inclut ainsi le parseur, l'annotateur, et des post-traitements. S'appuyant également sur des méthodes de classification automatique, le module décrit les événements conformément aux balises spécifiées dans TimeML. Cette opération de classification requiert toutefois, à la différence de la suite TRIPS, des corpus annotés pour s'entraîner.

L'outil annote également les expressions temporelles et décrit les liens entre les événements annotés et ces expressions temporelles. Lorsqu'il dispose de suffisamment d'informations, le système permet de normaliser les expressions temporelles en s'appuyant sur la DCT (*Document Creation Time*) selon la méthode suggérée par (Pustejovsky, 2004). Il s'agit ainsi d'associer une valeur calendaire à d'expressions telles que « le mois dernier » (*last month*), « dimanche » (*Sunday*), « aujourd'hui » (*today*). Testé dans la campagne d'évaluation TempEval-2 (Verhagen *et al.*, 2010), l'analyseur TRIPS a ainsi obtenus de bons scores sur l'annotation des expressions temporelles (tâche A, extraction : score : précision 0,85, rappel 0,85 ; normalisation : typage : 0,94, valeur : 0,76) et des événements (tâche B, score : précision 0,80, rappel 0,74).

Le système a également cherché à identifier des relations temporelles, mais a obtenu sur ces tâches des scores plus faibles :

- entre un événement et une expression temporelle dans une même phrase (tâche C, score : précision 0,63, rappel 0,52),
- entre un événement et la date de création du document (tâche D, score : précision 0,76, rappel 0,69)

- entre les principaux événements de deux phrases adjacentes (tâche E, score : précision 0,58, rappel 0,50)
- entre deux événements, dans une configuration où l'un des deux est en rapport de subordination à l'autre d'un point de vue syntaxique (tâche F, score : précision 0,59, rappel 0,54).

La suite d'outils présentée produit donc des résultats satisfaisants sur le repérage, l'annotation et la caractérisation des événements et des expressions temporelles, mais une frontière résiste encore, celle de l'analyse de la structuration textuelle.

D'autres systèmes, très nombreux, visent à résoudre un sous ensemble des tâches proposées dans Temp-Eval 2 en s'appuyant sur le schéma d'annotation TimeML. Le système Evita (Saurí *et al.* 2005), par exemple, prend en entrée un texte annoté à l'aide d'un étiqueteur morpho-syntaxique. Sur la tâche d'identification des événements dans les textes, le système obtient un taux de précision 74.03% un taux de rappel de 87.31% (F-score : 80.12%). L'application GUTime (Mani et Wilson, 2000) annote pour sa part des expressions temporelles (TIMEX) conformément au schéma d'annotation TimeML et normalise leur valeur. Les auteurs rapportent que le système obtient un F-score de 85% sur la tâche de repérage des expressions temporelles et de 82% pour leur normalisation.

D'autres travaux (Bejan et Harabagiu, 2010) se sont attaqués à la résolution des co-références des désignations d'événements, une tâche qui n'était pas proposée dans Temp-Eval 2. Il s'agit d'identifier les différentes mentions d'un même événement au sein d'un texte ou dans un corpus, en s'appuyant sur la description des arguments de la structure prédicative.

3.1.3.3 Ressources pour le français

(Bittar, 2008) décrit une chaîne d'annotation des informations temporelles au sens de TimeML pour le français testée avec les corpus proposés dans une campagne d'évaluation. La chaîne s'appuie sur des traitements à l'aide de grammaires d'extraction (ou automates à états finis). Le système obtient des scores comparables aux outils élaborés pour l'anglais sur les expressions temporelles (F-score 83%). La tâche de normalisation obtient en revanche des scores un peu plus faibles (49% des expressions relatives ont été correctement normalisées). L'annotation des événements, conformément aux exigences de TempEval-1, vise à la fois leur repérage et leur typage et la description de plusieurs de leurs attributs (polarité, aspect, modalité de certaines constructions verbales). L'approche retenue repose sur des patrons lexicaux (noms et verbes) auxquels s'ajoutent quelques règles de désambigüisation au niveau des chunks. Evalué sur un corpus constitué de 10 articles du journal Le Monde, l'annotateur présente un F-score d'environ 76%.

(Parent *et al.* 2008) présentent également un système d'annotation pour le français qui utilise le schéma d'annotation TimeML et qui obtient des résultats comparables. Il a été évalué sur un corpus contenant 544 expressions temporelles annotées à la main. Pour l'annotation des événements, le système obtient un F-score autour de 70%. Pour ce qui est de la reconnaissance et de l'annotation des expressions temporelles, l'application montre un taux de précision de 83% pour un taux de rappel de 79% (soit un F-score de 81%). Pour ce qui est de la normalisation des expressions relatives, le système obtient un F-score de 50%.

Le système décrit dans (Weiser, 2010) obtient lui aussi des scores comparables, mais sur des tâches néanmoins différentes, qui relèvent d'un domaine d'application précis : les informations temporelles relatives au tourisme, comme les dates et horaires d'ouverture. Ce système d'annotation repose sur une modélisation *ad hoc*, qui s'écarte de TimeML, car, comme on l'a vu, ce type d'informations est difficilement et incomplètement pris en charge par le standard ; elles font en effet intervenir des itératifs et des expressions composées.

3.2 La modélisation des propriétés temporelles en ingénierie des connaissances

Après cette synthèse des travaux relatifs à l'annotation des informations temporelles dans les textes, on souhaite désormais faire une incursion du côté de l'ingénierie des connaissances, afin de voir comment les données temporelles y sont modélisées. Ceci est d'autant plus intéressant que l'acquisition de connaissances à partir des textes vise notamment à nourrir des bases de connaissances. DBPedia (Auer *et al.*, 2007)[14] et Freebase[15], deux bases de connaissances encyclopédiques, sont ainsi en grande partie enrichies à partir de textes. Cela nous conduit donc du côté des usages attendus de l'annotation des adverbiaux de localisation temporelle dans les textes. On a mentionné les systèmes de Questions/Réponses, qui sont à l'origine des travaux menés autour de TimeML. Ces annotations peuvent ainsi également permettre de constituer des bases de données structurées, en s'appuyant sur les textes, vus comme des sources d'information « non structurées » (*unstructured information*).

Avec l'avènement des *Linked Open Data* et l'ouverture des données publiques, de plus en plus de données structurées sous l'angle temporel (des données localisées dans le temps), sont mises à disposition de tous. En dépit de cette multiplication des données accessibles, la prise en compte de la temporalité pose des difficultés : (1) au niveau de la modélisation d'abord (comment représenter des informations temporelles de natures très variées ?), (2) au niveau des données elles-mêmes ensuite (comment faciliter l'enrichissement des bases de connaissances ? comment traiter des données parfois incomplètes ? comment les relier entre elles ?), et enfin (3) au niveau de l'exploitation de ces données pour la recherche d'information (quels systèmes mettre en œuvre pour pouvoir traiter des requêtes contenant des critères temporels ?).

3.2.1 Les entrepôts de données temporelles

Les données temporelles accessibles sur le Web sont celles qu'on trouve dans ce qu'on nomme les « données ouvertes ». L'expression couvre deux phénomènes de natures très différentes, d'un côté les *linked open data* (Bizer *et al.*, 2009), de l'autre le mouvement en faveur de l'ouverture des données publiques. Il faut bien mesurer la différence de l'un à l'autre, dans la mesure où les données publiques ne sont pas – ou le sont encore très peu - publiées dans des formats du Web Sémantique, alors que les *linked open data* reposent précisément sur ce format de données. Leur point commun

[14] http://dbpedia.org/About
[15] http://www.freebase.com/

tient ainsi dans le partage et la mise à disposition de ces données pour tous. Sous un autre angle néanmoins, la question est aussi de trouver des moyens innovants d'exploiter et de structurer davantage ces vastes entrepôts de données. La perspective temporelle trouve ici sa place.

L'intérêt pour les informations temporelles peut se mesurer aux nombres de données chrono-localisées qui enrichissent peu à peu le « nuage » des données ouvertes (*cloud*) : LODE pour *Linking Open Description of Events*, qui se présente à la fois comme un vocabulaire commun de représentation des événements et un entrepôt de données sur les événements historiques (cf. Shaw *et al.*, 2009) ; OWL-Time[16], une ontologie formelle du système calendaire (Hobbs et Pan, 2004) ; Event Ontology[17], un vocabulaire de représentation des événements (Raimond *et al.*, 2007) ; Timeline Ontology[18], un vocabulaire représentant les frises chronologiques et les objets qu'elles manipulent (Raimond et Abdallah, 2006) ; iCal Ontology[19], une transposition au format RDF des spécifications du standard iCalendar pour les agendas numériques ; ou encore Freebase time/Event et DBPedia Events, par exemple. L'objectif du mouvement en faveur des données interconnectées (*Linked Open Data*) est de publier et de lier entre eux des ensembles de données (*data sets*) sur le Web, à l'aide d'URI déréférençables pour identifier des documents sur le Web et des objets du monde « réel » (*real-world objects*). Il s'agit de lier entre eux ces objets, ces informations et ces vocabulaires.

Du côté des données publiques, après des initiatives d'acteurs non étatiques promouvant l'accès libre de leurs ressources (comme l'ouverture de banques de données génomiques en biologie, par exemple), plusieurs Etats[20], dans un mouvement initié par les Etats-Unis en 2009, suivi par le Royaume-Uni en 2010, puis par la France en 2011, s'attachent à rendre publiques de nombreuses données collectées par les institutions et collectivités. Ces données sont toutefois souvent diffusées sous des formats qui en rendent la réexploitation et l'accès difficile (des fichiers pdf, des tableurs, des documents textes). Un des objectifs d'Etalab[21], la mission publique française chargée de l'ouverture des données publiques, est ainsi de faciliter leur exploitation en invitant des acteurs divers (des secteurs publics et privés) à développer des services qui en facilitent l'accès et la réutilisation et permettent de les mettre en relation. Un des enjeux consiste notamment à poser, par-dessus ces données et documents, une structure formelle, qui peut être exploitée par des agents logiciels. Là encore, la perspective temporelle a son importance : exemple parmi d'autres, on trouve dans les données du projet Etalab les horaires d'ouverture d'un grand nombre de musées décrits en langage naturel. Structurer ces informations doit permettre de les rendre interrogeables.

L'annotation des textes a donc sa place dans cette démarche, dans la mesure où elle permet de passer d'une représentation textuelle à une représentation structurée, formelle, que les agents logiciels peuvent ensuite pouvoir manipuler.

[16] http://www.w3.org/TR/owl-time/
[17] http://purl.org/NET/c4dm/event.owl#
[18] http://motools.sourceforge.net/timeline/timeline.html
[19] http://www.w3.org/2002/12/cal/icaltzd
[20] source : François Bancilhon, *Data Publica*, séminaire du 28 mars 2012 au Collège de France, chaire « Informatique et sciences numériques »
[21] http://www.etalab.gouv.fr/

3.2.2 Les vocabulaires

Les vocabulaires spécifiques aux objets temporels sont nombreux (on en a mentionné plusieurs) et ont des vocations diverses : modéliser des événements, modéliser le système calendaire, les fuseaux horaires, les frises chronologiques, les agendas numériques, etc. Mais on trouve aussi des informations temporelles dans des ontologies dites « génériques »[22], telles que DOLCE ou DOLCE+DnS Ultralite (DUL) (Scherp *et al.*, 2009), CIDOC (Doerr, 2003) ou ABC (Lagoze et Hunter, 2001), par exemple. Pour schématiser, il y a deux types principaux d'objets : les objets localisés dans le temps (généralement appelés des *événements*) et les objets permettant cette localisation (généralement des *valeurs calendaires*).

Il y a plusieurs façons de modéliser les valeurs calendaires, qui vont de la plus simple (les dates du type xsd:date qui couvrent les valeurs du type jour-mois-année) à la plus aboutie, OWL-Time (Hobbs et Pan, 2006). OWL-Time se cantonne toutefois délibérément à la formalisation du système calendaire et ne s'avance qu'avec précaution sur le terrain qui consiste à mettre en lien l'objet localisé dans le temps et ces valeurs calendaires. L'ontologie distingue toutefois quatre relations possibles : (i) *atTime* (x a lieu à l'instant *t*), (ii) *during* (x a lieu sur tout l'intervalle de temps *T*), (iii) *holds* (x est vrai à l'instant *t* ou sur un intervalle de temps *T* ou encore sur un ensemble d'intervalles de temps) et (iv) *timespan* (x a lieu à l'intérieur d'une fenêtre de temps, qui peut être un intervalle ou une séquence d'intervalles).

(Shaw *et al.*, 2009) dressent une synthèse des différentes façons de modéliser les événements, en particulier les événements historiques, leur objectif étant de créer un vocabulaire en mesure de lier les différents vocabulaires existants. Ils montrent en particulier qu'il y a presqu'autant de définitions des événements que de modèles. Dans le cadre de nos travaux, plutôt qu'aux événements en tant que tel, on s'intéresse à la localisation temporelle, et donc à la façon dont des objets de natures variées peuvent être associés à des périodes de temps.

Comme le rapportent (Shaw *et al.*, 2009), dans le cadre du Web Sémantique, deux approches coexistent, dont l'une consiste à relier directement un événement avec des valeurs calendaires, au moyen de propriétés de type *datatype* pointant sur des littéraux RDF (*RDF literals*) représentant des dates (qui sont ainsi typées à l'aide des schémas XML *xsd:date* ou *xsd:dateTime*). Une autre approche, plus complexe et plus fine, introduit des classes pour représenter les intervalles de temps et recourt à des propriétés de type *object properties* pour relier des instances d'événements avec ces classes. Les instances d'intervalles de temps peuvent ensuite être liées à des valeurs calendaires à l'aide de propriétés *datatype*.

ABC, CIDOC, et Event Ontology utilisent par exemple cette seconde méthode : ainsi ABC et CIDOC introduisent des classes pour les intervalles de temps et Event Ontology réemploie la classe *TemporalEntity* empruntée à OWL-Time. DUL autorise les deux approches : les dates associées à un événement peuvent être directement reliées à lui au moyen de la propriété datatype *hasEventDate*,

[22] Les ontologies génériques, dites aussi de « haut niveau », sont souvent opposées aux ontologies de domaine, au sens où elles décrivent des concepts abstraits subsumant les concepts existants dans les différents domaines.

mais un intervalle calendaire peut également être associé à un événement à l'aide de la propriété *isObservableAt* pointant sur une instance de la classe *TimeInterval*.

Il est de nombreux cas de figure où la localisation temporelle ne pose pas de difficultés particulières et appelle un traitement simple. Associer directement un objet à des dates évite alors d'avoir à manipuler trop d'abstractions : il est plus simple de filtrer et trier les objets ainsi datés à l'aide de routines standards pour manipuler et comparer des dates. Mais le revers de cette simplicité, c'est qu'elle ne permet pas de représenter des relations complexes : les modèles des dates des schémas xsd:date et xsd:dateTime ne sont pas adaptées pour représenter les bornes de certains intervalles de temps, par exemple pour représenter une propriété temporelle comme celle décrite par l'expression « *du début des années 80 jusque vers le milieu des années 90* ». Cette approche ne permet pas non plus d'associer une période de référence à un objet, lorsqu'il s'agit moins de le dater que de donner une indication sur la période dans laquelle il a pris place (« *entre mars et avril* »). Dans ce cas, une relation représentant l'inclusion est nécessaire. Enfin, cette approche ne permet pas non plus de modéliser des phénomènes de récurrence.

Ainsi, d'un point de vue général, il y a différentes façons de localiser dans le temps un objet, selon le niveau de granularité et de précision à partir duquel on se place, selon qu'il s'agit de donner une période de référence ou au contraire des coordonnées très précises ou encore selon les connaissances dont on dispose : en effet, il y a des événements que l'on ne sait pas dater avec précision. On montrera par la suite qu'il peut être intéressant pour modéliser la localisation temporelle en ingénierie des connaissances d'observer la façon dont les situations décrites dans les textes sont localisées dans le temps, en particulier à travers l'analyse des adverbiaux. On formulera en effet des propositions pour représenter la sémantique des adverbiaux de localisation temporelle : on s'attachera alors à montrer que ces représentations peuvent être réexploitées pour l'acquisition de connaissances à partir des textes.

3.2.3 L'interrogation des données temporelles

Si les moteurs de recherche classiques n'exploitent que peu les informations temporelles, en revanche les outils d'interrogation des bases de données (SQL) et des bases de connaissances (SPARQL, par exemple) les exploitent de façon très courante. Ces outils permettent de filtrer des données en fonction des dates qui leur sont associées, à l'intérieur d'une fenêtre de temps. Plusieurs types de requêtes sont possibles : soit l'utilisateur récupère toutes les données auxquelles une date précise est associée, soit toutes celles qui sont associées à une date supérieure ou inférieure à celle précisée dans la requête, soit encore toutes celles auxquelles sont associées une date comprise entre une borne de début et une borne de fin spécifiées dans la requête. On trouve ce genre d'approche dans les outils de réservation de billet de train ou d'avion par exemple.

Les requêtes sont généralement saisies dans des formulaires, même si d'autres interfaces sont possibles, comme celle proposée sur le site oldmapsonline.com, un outil de consultation de cartes géographiques anciennes, où il est possible d'ajuster graphiquement la fenêtre de temps du filtrage sur une règle dont les bornes sont mobiles.

Compte-tenu de la différence des formalismes sous-jacents et de la nature des informations stockées, on distingue généralement le Web de contenu (*Web of content*), composés de documents multimédias et de textes et le Web de données (*Web of data*), composés de données et de métadonnées formelles. Au niveau des mécanismes d'interrogation, les approches sont en effet sensiblement opposées : du côté des outils de fouille documentaire, la recherche produit des documents classés par pertinence, du côté des outils de fouille de données, la recherche filtre les résultats. Si ce sont des univers certes distincts, des ponts existent toutefois entre les deux, puisque des métadonnées formelles peuvent intégrer des pages Web (RDF-A, microformats) et décrire du contenu multimédia.

hCalendar est un de ces microformats. C'est une représentation sémantique XHTML du format iCalendar permettant d'exposer dans des pages Web, des informations relatives à des événements au sens des agendas numériques (pour proposer l'ajout d'un événement dans l'agenda de l'utilisateur, par exemple).

Par ailleurs, les moteurs de recherche plein texte proposent de plus en plus fréquemment des facettes de navigation, qui s'appuient sur les métadonnées associées aux documents (par exemple pour filtrer les documents en fonction de leur format ou de leur « fraîcheur »). Ces facettes de navigation permettent d'affiner les résultats d'une recherche. Une continuité se dessine ainsi entre les outils de recherche sur les mots-clés et les outils de filtrage et de navigation à partir de métadonnées, articulant ainsi des données structurées et des données textuelles.

(Alonso *et al.*, 2009 ; Campos *et al.*, 2009) proposent ainsi une chaîne de traitements permettant de regrouper des documents dans des facettes chronologiques (cluster), exploitant les expressions temporelles présentes dans ces documents. Les résultats présentés à l'issue d'une recherche peuvent ensuite être filtrés à l'aide de ces facettes chronologiques. C'est dans une démarche voisine, qui consiste à exploiter les expressions temporelles présentes dans les documents, que nos travaux s'inscrivent.

3.3 Recherche d'Information et informations temporelles

3.3.1 Variétés des informations et variétés des usages

Des équipes de recherche ont cherché à mesurer quantitativement l'importance des recherches sur le Web qui comportent l'expression d'un critère calendaire. S'appuyant sur l'analyse des archives des requêtes soumises aux moteurs de recherche, (Nunés *et al.*, 2008) rapportent qu'environ 1,5% des requêtes sur le web contiennent l'expression de critères temporels (*peinture italienne XVe siècle, coupe du monde 1998,* etc.). (Metzler *et al.*, 2009) évaluent pour leur part à 7% le nombre de requêtes qui renvoient « *implicitement* » à une période précise, sans toutefois que cette période soit exprimée dans la requête. Ainsi, par exemple, les auteurs considèrent qu'une requête telle que « *coupe du monde de football Allemagne* » a toutes les chances de référer à la coupe du monde de 2006. Les auteurs rangent ce type de requêtes dans la catégorie des requêtes présentant un critère temporel « implicite ». Ceci revient à considérer que toute requête relative à un événement est une « requête temporelle » et pose donc le problème de savoir ce qu'il faut entendre par ces termes :

pour éviter toute ambiguïté, nous isolons, pour notre part, dans cet ensemble, les requêtes contenant l'expression d'un critère calendaire, car elles nécessitent, comme on le verra, la mise en œuvre de traitements spécifiques.

Les mesures fournies par ces auteurs valent pour les recherches sur le Web, tout utilisateur confondu et indépendamment de tout domaine. (Berberich *et al.*, 2010) précisent qu'il faut également compter des usages plus spécifiques des systèmes de recherche d'information, où l'expression de critères calendaires dans une requête pourrait revêtir une grande importance : ou bien pour des domaines précis (l'actualité, le sport) ou bien pour des utilisateurs experts (journalistes, historiens).

Par ailleurs, il est une donnée que les chiffres rapportés plus haut masquent. Les unités textuelles qui contribuent à ancrer dans le temps les situations décrites dans les textes sont aujourd'hui traitées par les moteurs de recherche comme des mots-clés dont la sémantique n'est pas exploitée. Une recherche sur un intervalle de temps (mettons « *de 1750 à 1800* ») ne ramène que des résultats où les termes mêmes de la recherche apparaissent : on pourra ainsi trouver des adverbiaux tels que « *en 1750* » ou « *en 1800* », mais pas des adverbiaux tels que « *peu après 1763* » ou « *de 1755 à 1799* », parce qu'il faudrait que l'outil puisse inférer que les zones temporelles qu'ils dénotent sont incluses dans celle dénotée par la requête. De même, les moteurs ne sont pas en mesure de rapprocher une adverbiaux comme « *en 1965* » avec une requête qui porterait sur « *les années 60* ». Les utilisateurs, par expérience, n'ignorent pas que les moteurs échouent devant ce type de requêtes. Il est vraisemblable qu'ils adaptent leur façon d'interroger les systèmes d'information à ce qu'ils savent pouvoir en attendre. On peut donc penser que si les moteurs étaient susceptibles de prendre en charge l'expression de critères calendaires, l'usage de ce type de requêtes se répandrait davantage.

L'idée d'introduire des critères temporels dans les modèles de pertinence pour la recherche d'information n'est pas une idée neuve. De nombreux travaux portent sur la façon dont il serait possible d'articuler un modèle de pertinence lié aux propriétés temporelles de pages Web (*time sensitive page rank*) avec les autres facteurs permettant de déterminer la pertinence d'une page Web. Pour autant, la plupart de ces approches (par exemple, Diaz et Jones, 2004 ; Asur et Buehrer, 2009 ; Kanhabua et Nørvag, 2010 ; Chen *et al.*, 2011) ne tiennent compte que des métadonnées *autour* des pages Web (comme leur date de publication, leur taux de modifications, etc.) plutôt que des expressions de localisation temporelle présentes *à l'intérieur* même des documents.

C'est là toutefois une problématique qui semble aujourd'hui intéresser grandement les moteurs de recherche, comme en témoigne le projet Google *view:timeline* (bien qu'il ait été abandonné en 2011), mais aussi les travaux de (Alonso *et al.*, 2010), du côté de Microsoft, et encore ceux de (Matthews *et al.*, 2010), du côté de Yahoo!.

Ainsi des initiatives récentes, encore expérimentales, se font jour avec pour objectif d'établir un modèle de pertinence en mesure de traiter des requêtes associant des mots-clés et des critères calendaires en tirant parti des expressions temporelles présentes dans les textes.

En outre, depuis peu, avec les avancées récentes sur le terrain de l'acquisition d'informations temporelles dans les textes dont les résultats commencent à être exploitables, le champ des applications s'étend progressivement à d'autres initiatives originales, telles que la construction

automatique de chronologies pour explorer et visualiser le contenu de corpus de presse (Alonso et al, 2010 ; Llorens *et al.*, 2010 ; Matthews *et al.*, 2010).

Les informations temporelles peuvent être exploitées à différents niveaux pour la recherche d'information. Il convient ainsi de distinguer des informations qui relèvent de *métadonnées* (qui peuvent donner lieu également à des recherches contenant l'expression de critères calendaires), de celles qui peuvent être présentes au sein même des documents (les expressions de localisation temporelle). C'est aux secondes que nous nous intéresseront plus particulièrement, même si certaines des premières ont également donné lieu à des recherches qui valent d'être évoquées, parce qu'elles abordent la problématique sous un angle intéressant.

Au nombre de ces dernières, la caractérisation temporelle des requêtes, qui permet comparativement d'en dresser des profils, a donné lieu à d'abondants travaux (Jones et Diaz, 2007 ; Asur et Buehrer, 2009 ; Chen *et al.*, 2011, par exemple). Explorant les archives des recherches sur le Web (*logs*), ces travaux distinguent ainsi les requêtes en fonction de leur distribution dans le temps : par exemple, des requêtes apparaissent plus fréquemment à intervalles réguliers, à l'occasion de fêtes ou d'événements sportifs récurrents, alors que certaines requêtes se distinguent par leur soudaine apparition en grand nombre, qui bouleverse leur distribution régulière dans le temps : elles sont en général liées à l'actualité et appellent des résultats « *frais* », c'est-à-dire des pages Web fraîchement publiées. Google, par exemple, a mis en place un tel algorithme, nommé *queries deserve freshness*, pour repérer ces requêtes. D'autres requêtes à l'inverse ont une distribution homogène dans le temps, ce qui tend à en faire des requêtes *atemporelles*, au sens où elles n'appellent pas des résultats publiés récemment : pour une requête sur Platon, par exemple, un document publié récemment n'aura pas nécessairement plus de pertinence qu'un document plus ancien.

Pour établir des profils de pages Web et de requêtes, ces travaux s'appuient sur les archives des requêtes (Metzler *et al.*, 2009 ; Chen *et al.*, 2011), sur le taux de modification des documents (Elsas et Dumais, 2010) ou encore sur les données sur les clics (Asur et Buehrer, 2009), c'est-à-dire sur la sélection par les utilisateurs des pages de résultats effectivement consultées à l'issue d'une recherche. Dans cette même veine, les travaux de (Kanhabua et Nørvag, 2010) visent à associer une période temporelle (une année) à une recherche thématique contenant un critère temporel « *implicite* ».

Concernant le traitement des requêtes définissant explicitement un filtre temporel, les travaux de (Nørvag, 2004 ; Berberich *et al.*, 2007) se sont intéressés à l'exploration de bases documentaires qui évoluent dans le temps, l'idée étant de donner accès aux versions successives des documents. C'est là une problématique cruciale pour la fouille d'archives, comme celle du Web, par exemple. Les requêtes, dans ces systèmes, sont composées de mots-clés et de dates, qui correspondent aux dates de création ou de modification des documents. Ces systèmes fouillent les bases documentaires et filtrent les documents qui remplissent à la fois le critère thématique et la contrainte temporelle.

Les travaux de (Jong *et al.*, 2005) portent eux sur l'évolution diachronique des langues, face à laquelle les systèmes de recherche d'information ne sont pas adaptés, leurs ressources ne contenant pas de liens entre les termes actuels des langues et leurs variations historiques.

Ces pistes de recherches ne visent toutefois pas à exploiter les expressions temporelles présentes dans les documents. C'est à cette piste que nous nous intéressons désormais et dont on rapporte ici différents résultats de recherche.

3.3.2 Exploiter les expressions temporelles présentes dans les textes

Sur ce terrain, au nombre des initiatives les plus visibles des moteurs de recherche, il nous faut citer celle, désormais abandonnée, de Google, le service *view:timeline*, qui tirait parti des expressions calendaires présentes dans les documents. Ce service permettait de visualiser les résultats d'une recherche thématique sur une frise chronologique, qui présentait la distribution des mots-clés dans le temps (cf. fig. 8). Cette distribution visait à rendre compte de la fréquence à laquelle un ensemble de mots-clés était associé à une date.

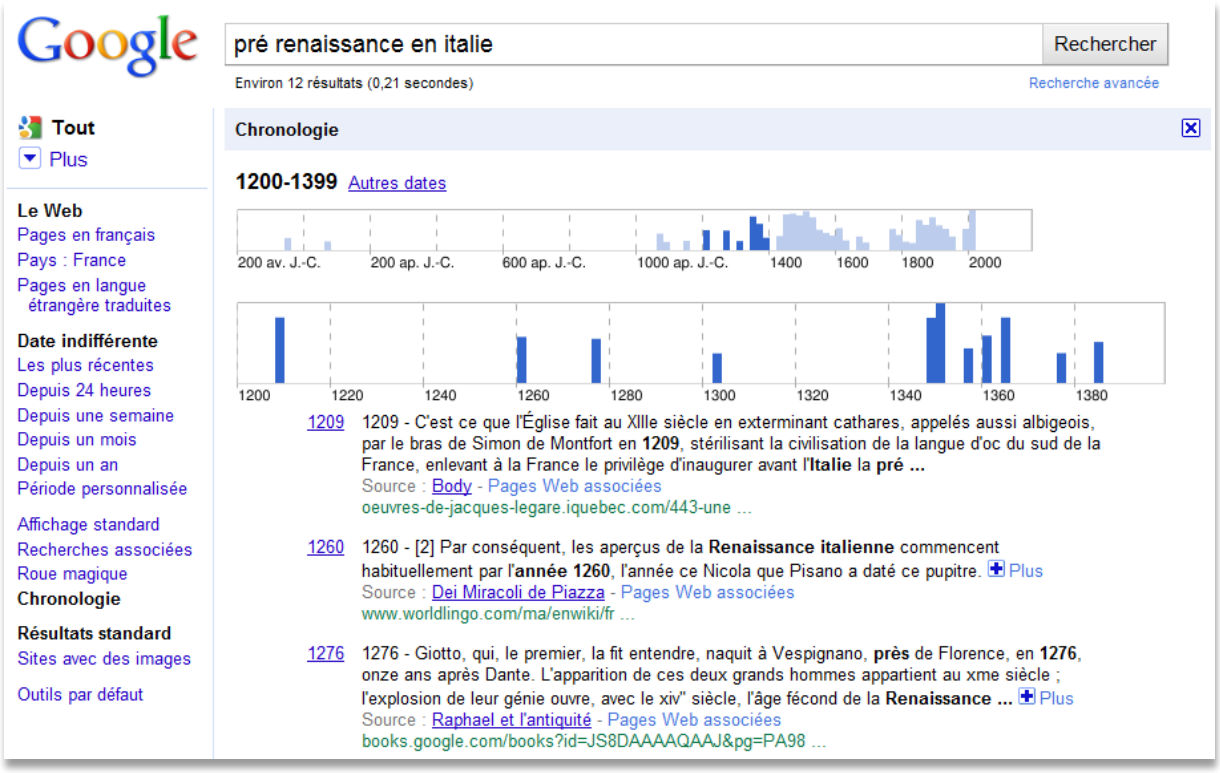


Fig. 8 : copie d'écran du service Google view:timeline (11 mars 2010)

Seule une sous-partie des expressions calendaires rencontrées dans les textes était indexée et donnait lieu à une analyse (peu ou prou, les expressions de la forme jour/mois/année, mois/année ou année). Les expressions calendaires étaient donc systématiquement réduites à une représentation rudimentaire : ainsi, une expression telle que « *de 1815 à 1871* » n'est pas analysée comme formant une zone temporelle bornée à gauche et à droite, mais plutôt comme deux dates.

Si ce service tirait parti des expressions calendaires présentes dans les textes (entendues dans le sens très restrictif de *dates*), il ne proposait pas à proprement parler de formuler des critères temporels dans les requêtes. Le service permettait néanmoins indirectement de le faire, en filtrant les résultats sur une fenêtre de temps à partir de la frise chronologique. Très pragmatique, c'était une piste

intéressante, car elle permettait d'éviter les risques de mauvaise analyse d'une expression calendaire dans une requête – mauvaise analyse qui conduit ensuite à fausser les résultats. En effet, pouvoir exploiter les expressions calendaires présentes dans les textes suppose de les avoir repérées par des traitements automatiques, qui, comme on l'a vu, peuvent générer du bruit. Cette approche permettait aussi de jongler avec le niveau de granularité attendu dans les résultats, en faisant varier les unités de temps.

Cependant, l'abandon du service s'explique vraisemblablement par ceci qu'il n'a pas emporté l'adhésion des utilisateurs de Google, constat face auquel on ne peut que formuler des hypothèses. On en avance deux, qui nous semblent expliquer en partie au moins ce en quoi le service ne répondait pas pleinement à ce qu'on pourrait attendre d'un système de recherche d'information qui tiendrait compte des expressions calendaires dans les textes.

En réduisant systématiquement les expressions calendaires à des dates, on trahit une grande partie de leur sémantique, jusqu'au possible contre-sens : un exemple de contre-sens consiste à considérer que l'expression « *avant 1905* » peut être représentée par la date *1905*. La représentation des expressions calendaires retenue ne permettait pas non plus d'analyser finement des expressions comme « *à la fin des années* 30 » ou « *de 1925 à 1930* ».

Par ailleurs, le service de Google view:timeline n'était pas à proprement parler un service de recherche documentaire (*text retrieval*), dans la mesure où la pertinence d'un document n'était pas évaluée en tant que telle : le service permettait de filtrer des extraits de documents remplissant à la fois les critères thématiques et la contrainte temporelle, mais le critère temporel n'entrait pas dans la mesure de la pertinence d'un document : il servait uniquement à écarter des documents, non à les ordonner.

Cette approche, qui consiste à considérer le critère temporel comme un filtre ou une fenêtre de sélection est celle qui est le plus souvent retenue dans les systèmes de recherche d'information qui permettent d'exprimer une contrainte temporelle : cette approche se retrouve en effet dans les bases de données où elle est très répandue (pour réserver un billet de train ou d'avion pour ne prendre qu'un exemple). Comme le remarquent (Arikan *et al.*, 2009), ce type d'approche ne permet pas de mesurer la proximité entre les expressions temporelles présentes dans les documents et celle exprimée dans la requête.

Une autre approche est possible, qui consiste à tenir compte des caractéristiques du critère temporel exprimé dans une requête et d'essayer d'en mesurer la proximité avec les expressions temporelles présentes dans les textes. (Berberich *et al.*, 2010), à la suite des travaux présentés dans (Arikan *et al.*, 2009), ont ainsi développé un modèle de pertinence très fin pour analyser les critères temporels dans les requêtes et les expressions temporelles repérées dans les documents. Ils associent un quadruplet à chaque expression temporelle. Ce quadruplet est constitué d'une borne de début et d'une borne de fin, mais aussi, afin de prendre en compte ce qu'ils nomment *l'incertitude* (*uncertainty*) sur la précision de l'intervalle, ils ajoutent encore de deux autres bornes, destinées à modéliser l'incertitude sur le début et la fin de l'intervalle. Ces bornes sont comprises entre les bornes de début et de fin. Cette incertitude porte sur le fait de savoir si toute la zone temporelle associée à une expression calendaire doit être couverte par les réponses apportées par le système de

recherche d'information ou bien si seule une partie de cette surface peut l'être[23]. Par exemple, l'expression « *en 1998* » désigne-t-elle toute la période qui va du 1er janvier 1998 au 31 décembre 1998 (comme dans l'assertion suivante *en 1998, Bill Clinton était président des Etats-Unis*) ou bien n'est-ce qu'une fenêtre de temps approximative (comme dans l'assertion suivante *La France a gagné la coupe du monde en 1998*) ?

Dit autrement, sur le plan de la recherche d'information, cela signifie qu'une réponse couvrant toute la zone temporelle associée à une requête sera plus pertinente dans le premier cas, et qu'une réponse de granularité inférieure peut en revanche tout à fait être pertinente dans le second cas.

Ceci renvoie à nouveau, du reste, au rapport qu'entretiennent les adverbiaux de localisation temporelle avec le prédicat qu'ils viennent déterminer. (Klein, 1994) rappelle ainsi que les adverbiaux peuvent ne déterminer qu'une fenêtre de temps (*frame*) dans laquelle le prédicat peut prendre place : l'adverbial n'assigne pas une zone temporelle durant laquelle le prédicat est vrai (*event time*), mais détermine une zone temporelle de référence (*reference time*). C'est bien en effet le rôle de l'adverbial *en 1998* dans l'énoncé *La France a gagné la coupe du monde de football en 1998*.

Ces travaux sont intéressants en ce sens qu'ils visent à établir des critères pour établir la pertinence de documents qui dépendent de la sémantique des expressions temporelles. (Le Parc-Lacayrelle *et al.*, 2007) présentent également une méthode pour calculer le degré de pertinence d'une expression temporelle par rapport à une expression requête. Ils proposent de faire dépendre la mesure de pertinence des rapports de durée entre les intervalles associés aux expressions temporelles et de la distance entre les centres (« centroïdes ») des intervalles comparés. Ils distinguent ainsi trois scores :

- Un score de *précision*, égal au rapport de la durée commune aux intervalles comparés par rapport à la durée de l'intervalle associé à l'expression dont on souhaite mesurer la pertinence ;
- Un score de « *signification » (significance*), égal au rapport de la durée commune aux intervalles comparés par rapport à la durée de l'intervalle associé à l'expression requête ;
- Un score de *distance*, égal au rapport entre les centroïdes des intervalles associés à l'expression testée et à l'expression requête.

Ces scores sont agrégés ensuite pour associer une mesure de pertinence temporelle. Nos travaux s'inscrivent dans une démarche similaire (cf. section 5.3). En outre, (Le Parc-Lacayrelle *et al.*, 2007) évoquent également la difficulté qu'il y a à combiner les mesures de pertinence associées aux critères thématiques d'une requête (les mots-clés) avec celles associées aux critères temporels. Nous formulerons également des propositions pour résoudre le problème de leur combinaison (cf. section 7.1.5).

[23] (Gosselin, 2005b) rappelle ainsi qu'il est « important, lorsqu'on analyse les phénomènes temporels dans les textes, de préciser quelles relations sont contraintes et quelles relations restent indéterminées, l'indétermination relative étant une propriété essentielle de la sémantique des textes. (…) Par exemple, en énonçant : *Il y a dix minutes (quand je suis sorti), il pleuvait*, le locuteur n'indique en rien si le procès (la pluie) a cessé ou non au moment de l'énonciation. Le procès a certes une partie passée, mais il se peut très bien qu'il se poursuive dans le présent et même dans le futur. La seule information sûre, c'est qu'au moment de référence, situé dans le passé, et localisé grâce au circonstanciel, le procès était en cours (aspect inaccompli). »

Parmi les travaux qui cherchent à tirer parti des expressions calendaires dans les textes, ceux de (Baeza-Yates, 2005) partent d'une idée initiale qui consiste à observer comment, dans le passé, les textes référaient à l'avenir : le cas d'application considéré s'appliquait à un corpus d'archive de presse dont les références aux années à venir étaient extraites, aussi bien les références déictiques comme *d'ici 30 ans* que les références absolues telles que *en 2020*.

Dans le sillage de ces travaux (Matthews *et al.*, 2010) présentent un démonstrateur *Time Explorer*[24] (cf. fig. 9) permettant de faire des requêtes par mots-clés sur le corpus d'archive du *New York Times*[25]. A ces requêtes par mots-clés, l'utilisateur peut associer deux types de filtres temporels, l'un effectuant des recherches portant sur des références au passé, l'autre des recherches portant sur des références au futur : les expressions temporelles sont ainsi rangées en deux catégories, selon la date de publication de l'article où elles sont présentes. L'outil permet ensuite d'explorer les références à l'avenir et au passé associées à un ensemble de mots-clés, en navigant sur la frise chronologique présentée au-dessus des résultats. Il est également possible de filtrer les documents d'archive en fonction de leur date de publication. En revanche, l'outil ne permet pas d'exprimer de critères temporels directement dans la requête.

Un des aspects intéressants de ce projet est sa façon de présenter les résultats, qui diffère de celle des moteurs de recherche grand public : au lieu de courts extraits de textes segmentés autour des mots-clés retrouvés (*snippets*), le démonstrateur *Time Explorer* présente le plus souvent des phrases complètes, où figurent les termes recherchés et l'expression calendaire qui leur est associée. Cela renvoie à deux problèmes auxquels doivent faire face les systèmes de recherche d'information, l'un générique, l'autre spécifique au traitement des expressions calendaires. Le premier a trait à la segmentation des textes : comment, dans la liste des résultats, isoler de leur contexte d'interprétation des extraits de textes, tout en s'assurant qu'ils restent intelligibles ? L'autre concerne la portée des adverbiaux temporels[26] : comment s'assurer que l'expression calendaire présente dans l'extrait fourni comme résultat est bien en rapport avec les mots-clés eux aussi présents dans cet extrait ? La solution pragmatique adoptée (présenter des phrases) évite d'aborder frontalement les problèmes de la résolution des anaphores ou de la relation entre les adverbes et les objets sur lequel ils portent – deux problèmes qui nécessitent des analyses linguistiques profondes complexes à opérationnaliser.

[24] Le démonstrateur est accessible à l'adresse suivante : http://fbmya01.barcelonamedia.org:8080/future-nyt/
[25] http://ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T19
[26] On renvoie sur ce sujet à la section 2.4.

Fig. 9 : Copie d'écran de l'outil Time Explorer (Matthews et al., 2010)

(Llorens *et al.*,2010) présentent également un outil, *Time-Surfer*, qui s'appuie sur les expressions temporelles présentes dans les textes annotées à l'aide du langage d'annotation TimeML (cf. fig. 9). Cet outil construit de façon automatique une chronologie qui regroupe dans des cercles de tailles d'importance variable les événements (au sens de TimeML) qui se sont produits à une même date. Ce système n'est pas un outil de recherche documentaire (les chronologies sont construites à partir d'un seul texte). L'outil permet de représenter de façon originale les informations temporelles d'un texte. En ce sens, il relève davantage d'un outil de fouille intra-documentaire.

Dans cette approche, conformément à TimeML, les événements, lorsque l'outil y parvient, se voient assigner une date : ainsi, dans l'exemple de la fig. 10, les événements *formed* et *came* sont les deux événements associés à la date du *24 octobre 1945* qui a généré un cercle sur la frise chronologique. De ce point de vue, cette démarche s'efforce de résoudre la question de la portée de l'adverbial temporel (*on 24 October 1945*), même si, comme on l'a vu plus haut, la démarche qui anime TimeML ne pose pas les termes du problème de cette façon-là.

L'outil propose également une recherche par mots-clés, mais alors la recherche ne se limite pas aux événements qui sont déterminés par un adverbial temporel : comme dans *Time Explorer* (Matthews *et al.*, 2010), les termes de la recherche peuvent apparaître n'importe où dans la phrase.

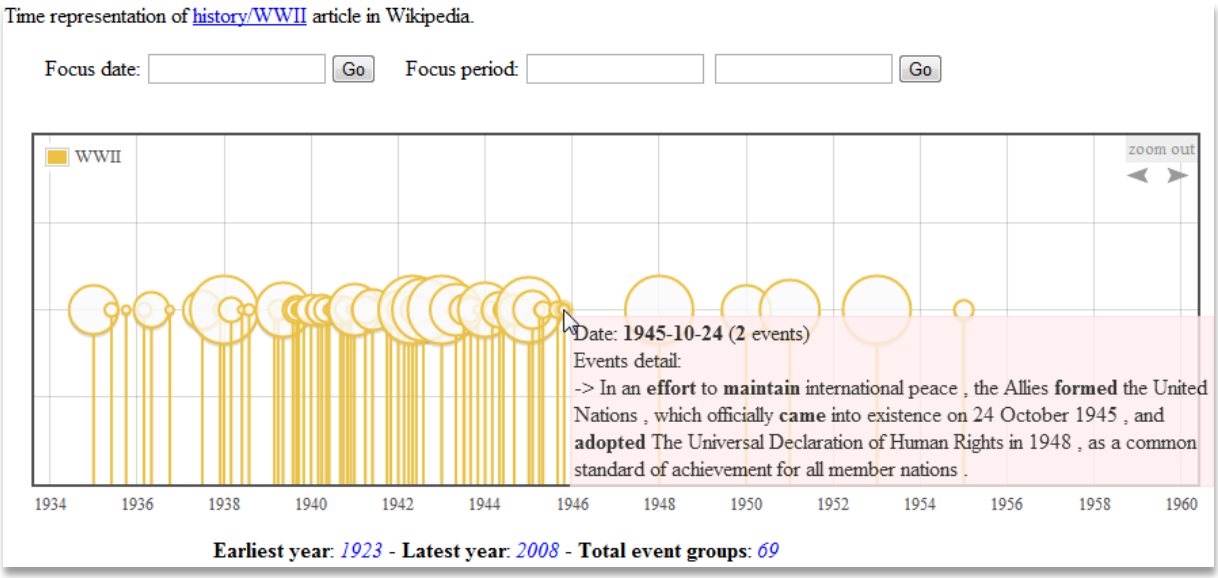


Fig. 10 : copie d'écran de l'outil Time-Surfer (Llorens et al.,2010)

Encore largement expérimentales, ces différentes applications n'ont pas encore donné lieu à des systèmes de recherche d'information opérationnels exploitant les adverbiaux de localisation temporelle dans les textes.

3.4 Bilan du chapitre

Au niveau de la recherche d'information temporelle, on a ainsi vu que deux démarches coexistent, l'une consistant à considérer le critère temporel d'une requête comme un simple filtre, l'autre consistant à mesurer la proximité entre les expressions temporelles présentes dans les textes et une expression présente dans une requête (Berberich *et al.*, 2010). S'appuyant sur une modélisation de la sémantique des adverbiaux de localisation temporelle, c'est dans cette dernière démarche que nos travaux s'inscrivent comme on le verra dans le chapitre 5 (cf. section 5.3). On essaiera ainsi de montrer l'intérêt d'articuler une représentation qui tient compte de la sémantique de ces adverbiaux avec une représentation calculable, destinée à faciliter la mise en œuvre d'applications pour l'ingénierie des connaissances et pour la recherche d'information.

Sur le plan de l'ingénierie des langues, au niveau de l'annotation des expressions datatives et des événements dans les textes, on a vu que la définition des objets d'analyse et l'objectif généralement assigné aux systèmes d'annotation n'allaient pas de soi et méritaient d'être questionnés. Nous souhaitons montrer à présent l'intérêt d'une approche qui ne réduirait pas les « expressions temporelles » à des formes d'entités nommées d'un genre particulier, mais qui s'intéresserait plutôt à caractériser la sémantique d'une catégorie morphosyntaxique qui contribue à l'ancrage des procès dans le temps : la catégorie des adverbiaux de localisation temporelle. Il s'agira ensuite de montrer que la description de ces adverbiaux gagne à ne pas être considérée comme une sous-tâche intermédiaire pour les systèmes d'annotation, mais précisément comme leur objectif central, à l'inverse de la démarche adoptée dans TimeML, où il n'est pas question de représenter la sémantique des expressions datatives en s'appuyant sur une analyse linguistique : tout au plus une telle description intermédiaire peut-elle éventuellement servir au calcul d'une valeur calendaire qui,

elle, sera inscrite dans les annotations. La question de la « normalisation » de ces expressions (qui consiste à leur associer une valeur calendaire) relève donc, selon nous, d'une autre problématique qu'il s'agit d'isoler et d'articuler avec celle de l'annotation sémantique des adverbiaux de localisation temporelle.

Comme on l'a brièvement remarqué, si l'on considère des *adverbiaux* plutôt que des unités textuelles considérées comme des entités nommées (des dates et des événements), on en vient à rapprocher des expressions jusque-là analysées différemment (*depuis fin juin* et *depuis la fin de la campagne électorale*, par exemple). C'est partant de cette démarche que l'on cherche à circonscrire notre objet d'analyse, les *adverbiaux de localisation temporelle*, qui contribuent à localiser dans le temps les situations décrites dans les textes.

Chapitre 4 : Les adverbiaux de localisation temporelle : une proposition d'analyse sémantique

Nous détaillons, dans ce chapitre, une proposition de modélisation des *adverbiaux de localisation temporelle* qui découle d'une analyse linguistique. Cette modélisation s'attache à mettre en lumière la façon dont ils déterminent un ancrage temporel à partir d'un repère temporel noyau. La représentation que l'on en propose les décrit sous la forme d'une succession d'opérations agissant sur ce repère temporel initial.

Cette démarche procède d'une hypothèse forte avancée par (Battistelli, 2009), que l'on reprend ici en l'étendant, pour montrer que non seulement elle s'applique aux *adverbiaux calendaire* (qui opèrent un ancrage sur le référentiel du calendrier), mais qu'elle vaut également pour l'ensemble des adverbiaux de localisation temporelle (les adverbiaux de localisation déictiques, anaphoriques, ou encore les adverbiaux contenant une référence à un « événement »). Cette hypothèse porte sur les liens qu'entretiennent les fonctions syntaxique et sémantique des adverbiaux de localisation temporelle. Ces adverbiaux peuvent être représentés sous la forme d'une séquence toujours identique d'opérations sémantiques, même si ces opérations ne sont pas toujours marquées au niveau morphosyntaxique et, ajoute-t-on, même si la nature du repère temporel noyau peut varier : l'ancrage peut s'effectuer par rapport au référentiel du calendrier (*en 1977*), par rapport au référentiel de l'énonciation (*l'an prochain*) ou par rapport à un « événement » (*dès son retour*).

Notre approche se distingue de celles qui prédominent (au moins en ingénierie des langues) par ceci qu'elle ne s'intéresse pas, en premier lieu, à la valeur calendaire pointée par les adverbiaux de localisation temporelle, mais plutôt à la façon dont les marqueurs linguistiques dont ils sont formés viennent déterminer un ancrage temporel à partir d'un repère temporel noyau. Ces interactions sont représentées sous la forme d'opérations sémantiques (opération de focalisation, de déplacement et

de régionalisation). Il ne s'agit donc pas, à ce niveau d'analyse, de déterminer des « coordonnées temporelles » (comme on parle de coordonnées géographiques).

A nos yeux, donc, la projection de la référence temporelle dénotée par les adverbiaux de localisation temporelle sur un axe du temps (en particulier sur le référentiel calendaire) constitue l'objet d'une autre tâche, liée à des besoins applicatifs et pour laquelle nous formulerons également des propositions (cf. chapitre 5).

Dans ce chapitre, nous proposons une définition des adverbiaux de localisation temporelle, ainsi qu'une typologie qui découle de la nature de leur repère temporel noyau (section 4.1). Nous détaillons ensuite les opérations à l'aide desquelles nous cherchons à décrire la façon dont les adverbiaux de localisation temporelle déterminent un ancrage temporel (section 4.2).

4.1 Définition, typologie et portée des adverbiaux de localisation temporelle

4.1.1 Considérations morphosyntaxiques et syntaxiques

Comme on l'a vu dans le chapitre 2 (cf. section 2.4), la catégorie des adverbiaux – et, partant, des adverbiaux temporels - est une catégorie grammaticale problématique. Du point de vue morphosyntaxique, elle est susceptible de recouvrir des syntagmes prépositionnels (*sous l'Ancien Régime*), des subordonnées (*depuis qu'il est parti*) ou même encore de groupes nominaux (*il conduit une voiture volée la veille*). Chacune de ces catégories morphosyntaxiques peut former un *adverbial de localisation temporelle*, dès lors qu'elle permet de déterminer un ancrage temporel.

En outre, on a vu dans le chapitre 2 (cf. section 2.4), qu'un adverbial de localisation temporelle peut déterminer une relation prédicative (ex. 1), une relation intraprédicative (ex. 2), un prédicat nominalisé (ex. 3) ou encore la relation entre un prédicat second et le syntagme nominal auquel il est rapporté (ex. 4).

> Ex. 1 : **Fin novembre**, la Tunisie a signé avec l'Espagne un protocole d'accord en vue de collaborer dans le domaine de la lutte contre le dopage.
> Ex. 2 : Le dopage compte parmi les principales causes de "la détérioration du jeu sain", a estimé Samir Laabidi, ministre tunisien de la Jeunesse et des Sports, **vendredi** à Tunis (…).
> Ex. 3 : Relancées le 2 septembre à Washington, les négociations de paix directes sont suspendues depuis la reprise **fin septembre** de la construction dans les colonies juives à Jérusalem-Est annexée et en Cisjordanie.
> Ex. 4 : L'entourage familial du président Zine El-Abidine Ben Ali est une "quasi-mafia" et le régime tunisien "n'accepte ni critique ni conseil", affirment des télégrammes confidentiels américains obtenus par WikiLeaks et révélés **mardi soir** par le quotidien Le Monde.

On rappelle également, que des travaux sur les adverbes occupant une fonction de cadratifs (Charolles, 1997 ; Le Draoulec et Pery-Woodley, 2005) ont montré que les adverbes de localisation

temporelle sont également susceptibles de déterminer un cadre temporel dans lequel sont inclus plusieurs procès (cf. section 2.5.2) : ces adverbes de cadre portent ainsi sur des portions de texte supérieures à la phrase et même, potentiellement, sur tout un texte, telle qu'une dépêche par exemple (cf. ex. 5).

> Ex. 5 : JERUSALEM, **2 jan 2011** (AFP)
> Cinq Palestiniens suspectés d'avoir projeté un tir de roquette sur le principal stade de football de Jérusalem-ouest, ont été arrêtés, a annoncé dimanche la police israélienne.

4.1.2 Définition

Au niveau sémantique, les adverbiaux de localisation temporelle opèrent un ancrage sur un référentiel temporel, permettant ainsi de *situer* un « intervalle circonstanciel » par rapport au calendrier, à l'énonciation ou à un autre procès. Ils contribuent ainsi à déterminer temporellement un procès. Le terme d'« intervalle circonstanciel » est emprunté ici au modèle de la Sémantique de la Temporalité (Gosselin, 1996, 2005), qui distingue quatre types d'intervalles (cf. section 2.3) : l'intervalle d'énonciation, l'intervalle de procès, l'intervalle de référence et l'intervalle circonstanciel. On montrera que l'intervalle circonstanciel à l'aide duquel on peut représenter un adverbial de localisation temporelle se construit d'abord à partir d'une référence temporelle noyau qui est l'opérande d'une succession d'opérations sémantiques.

Un *adverbial de localisation temporelle*, en première approche, sera ainsi défini comme un adverbial permettant de situer un intervalle circonstanciel ou bien sur le référentiel calendaire (cf. ex 1 à 3), ou bien par rapport à un procès (cf. ex 4 à 6) ou bien par rapport au processus énonciatif (cf. 7 à 9).

> Ex. 1 : Les manifestations sont proscrites en Algérie en vertu de l'état d'urgence en vigueur **depuis 1992**.
> Ex. 2 : « La légitimité ne peut être bâtie sur la répression ni sur le déni des droits politiques et sociaux », assure M. Hamzaoui. « Personne ne peut accepter cela **au XXIe siècle** et les Arabes ne sont pas une exception ».
> Ex. 3 : **A la fin des années 2000** son régime est décrit comme "autoritaire" par les organisations de défense des droits de l'Homme.

> Ex. 4 : **depuis la fuite de Zine El Abidine Ben Ali**, qui s'est réfugié vendredi en Arabie Saoudite, les Tunisiens exigent dans la rue la dissolution du Rassemblement constitutionnel démocratique (RCD), qui règne sur la Tunisie depuis son indépendance en 1956.
> Ex. 5 : Rached Ghannouchi, qui a quitté Londres dans la matinée en compagnie notamment d'une de ses filles, s'était déclaré "très heureux" **juste avant son départ**.
> Ex. 6 : Quatre voyages sont concernés entre le 15 avril et le 6 mai inclus, **au début de la saison touristique** en Tunisie, indique Apollo, filiale suédoise du géant suisse du tourisme Kuoni, dans un communiqué.

> Ex. 7 : "Il est mort **hier** à 19h00", a déclaré à l'AFP Mme Belhassen.

Ex. 8 : Le Premier ministre luxembourgeois Jean-Claude Juncker s'est dit "très choqué" **jeudi** par la publication par WikiLeaks de documents secrets américains lorsque leur contenu menace la vie des gens, tout en comprenant la nécessité de télégrammes diplomatiques.
Ex. 9 : "Ce qu'il faut noter, c'est que si les pays importateurs de pétrole de la région ont connu un ralentissement relativement modéré de la croissance **l'année dernière**, ce taux de croissance, (…), est en deçà du niveau nécessaire pour permettre la création d'emplois suffisants pour absorber les nouveaux entrants sur le marché du travail (…)", a poursuivi le porte-parole.

Selon la nature du référentiel ou du repère temporel impliqué dans l'opération de localisation, on distingue ainsi trois types d'adverbiaux de localisation temporelle : les adverbiaux *calendaires* (ex. *dès les années 1930*), les adverbiaux *relatifs à un procès* (ex. *dès son retour*) ou *à un nom de temps* (ex. *avant la Belle Epoque*) et les adverbiaux *déictiques* (ex. *depuis le siècle prochain*). Ces distinctions appellent toutefois plusieurs remarques et posent des difficultés de différentes natures.

D'abord, il est des adverbiaux qui, considérés isolément, ne permettent pas de déterminer la nature du repère temporel impliqué dans l'opération de localisation. On pense ici aux *adverbes de localisation temporelle anaphoriques* : pour connaître la nature de la référence temporelle noyau d'un adverbe anaphorique, il est nécessaire de remonter le fil du discours (cf. ex 10 à 12).

Ex. 10 : La journée de lundi devrait être cruciale pour le nouveau pouvoir qui a appelé les fonctionnaires à reprendre "impérativement" le travail **ce jour-là**.
Ex. 11 : "Quand j'avais 17 ans, ce qui fait un bout de temps maintenant, je suis allé à Paris pour voir Bud Powell (réputé pianiste de be-bop). **A l'époque** les amateurs anglais de jazz devaient "aller en France" pour en écouter, souligne-t-il.
Ex. 12 : « Ce fut en commémoration de la belle conduite du maréchal Lefebvre pendant ce siège, qu'il reçut le titre héréditaire de duc de Dantzick, par lettres-patentes du 28 mai de la même année 1807. **Pendant *la suivante***, *il alla* commander en Espagne un corps d'armée composé de 3 divisions, à la tête duquel il gagna la bataille de Durago le 30 octobre (…) »

Les adverbes anaphoriques sont donc des adverbiaux de localisation temporelle qu'il n'est pas possible de typer plus finement, sans avoir une connaissance approfondie du contexte discursif dans lequel ils apparaissent : la référence anaphorique peut désigner un adverbial relatif à un procès (ex. 11, où la référence noyau correspond à la subordonnée temporelle « Quand j'avais 17 ans »), un adverbial déictique (ex. 10, où la référence noyau correspond à l'adverbe déictique « de lundi ») ou encore un adverbial calendaire (ex. 12 : « de la même année 1807 »).

Par ailleurs, à y regarder de près, il semble y avoir un continuum entre les trois types d'adverbes de localisation que l'on a distingués. Dans l'exemple 13 ci-dessous, on pourrait être tenté de ranger l'adverbe *à la fin des années 60* dans la catégorie des adverbiaux calendaires, dans la mesure où il opère un ancrage sur le référentiel du calendrier. Pour autant, l'adverbial est étroitement lié à son contexte d'énonciation : il est nécessaire en effet de disposer d'informations sur ce contexte pour déterminer qu'il s'agit des *années 1960*[27]. Il est à noter d'ailleurs que cet exemple est extrait d'une

[27] Le titre du roman V. Hugo « *93* » est à cet égard exemplaire : il désigne l'année 1793. Ceci appelle une remarque : une même expression, selon son contexte d'apparition, pourra être considérée comme un déictique ou un anaphorique.

dépêche AFP de janvier 2011. Compte-tenu de son contexte énonciatif, on peut dire que l'adverbial « *à la fin des années 60* » est une forme de déictique, au sens où il détermine un intervalle circonstanciel à partir d'une référence au processus énonciatif : on pourrait ainsi la paraphraser de la façon suivante *à la fin des années 60 du siècle dernier*. Les exemples 13 à 15 regroupent des adverbiaux déictiques qui contiennent également des unités du calendrier :

> Ex. 13 : Après un bref passage en France, il rentre en Tunisie **à la fin des années 60** et découvre avec effroi une société lancée sur la voie de la laïcité et où les femmes ont obtenu l'interdiction de la polygamie et de la répudiation.
> Ex. 14 : Trois tour-opérateurs suédois ont annoncé lundi l'annulation de quinze vols charters vers la Tunisie en raison des troubles dans le pays, affectant près de 2.000 séjours **entre février et mai**.
> Ex. 15 : Ce sommet constitue la première réunion des chefs d'Etat arabes depuis le départ **vendredi dernier**, sous la pression populaire, du président tunisien Zine El Abidine Ben Ali, après 23 ans de règne.

Dans chacun des exemples 13 à 15 des unités *ordinales* du calendrier apparaissent (*années 60*, *février*, *mai*, *vendredi*). Ces adverbiaux ont donc partie liée avec le référentiel calendaire. Ils sont à rapprocher des suivants (ex. 14 à 16) où des unités *cardinales* du calendrier apparaissent (jours, semaine, année) :

> Ex. 14 : Plusieurs pays arabes -Algérie, Egypte, Mauritanie- ont connu **ces derniers jours** une série d'immolations par le feu, semblables au geste d'un jeune vendeur ambulant tunisien mi-décembre, qui avait marqué le début de la révolte ayant renversé le président Ben Ali.
> Ex. 15 : Il est "nécessaire que justice soit faite", a-t-elle souligné, annonçant qu'une mission d'évaluation des droits de l'homme se rendrait en Tunisie **la semaine prochaine**.
> Ex. 16 : "Ce qu'il faut noter, c'est que si les pays importateurs de pétrole de la région ont connu un ralentissement relativement modéré de la croissance **l'année dernière**, ce taux de croissance, équivalent à environ 4%, est en deçà du niveau nécessaire pour permettre la création d'emplois suffisants pour absorber les nouveaux entrants sur le marché du travail (…)", a poursuivi le porte-parole.

S'il est très fréquent que des unités calendaires participent à la composition des adverbiaux déictiques, pareillement, on peut trouver des unités calendaires dans des adverbiaux anaphoriques (cf. ex. 17 et 18) ou dans des adverbiaux de localisation relatifs à un procès (cf. ex. 19 et 20) :

> Ex. 17 : Et pourtant les propos échangés **ce 18 juin-là**, quel écho ils ont aujourd'hui !
> Ex. 18 : **L'année suivante**, un tribunal militaire de Tunis le condamne avec d'autres responsables religieux à la prison à vie pour "complot" contre le président.
>
> Ex. 19 : Une minute de silence sera observée au début de chaque cours **le jour de la rentrée** en mémoire des victimes du soulèvement populaire (…).
>
> Ex. 20 : La lettre de candidature à une fonction au sein du Bureau de la CIHM doit parvenir au secrétaire général CIHM **à la fin du mois d'avril de l'année des élections quinquennales statutaires**.

D'un point de vue opératoire, on distingue souvent les adverbiaux qui nécessitent un calcul pour déterminer la zone du calendrier à laquelle ils réfèrent (ce sont les *« expressions »* dites « *relatives »*), et ceux qui, au contraire, désignent directement une zone du calendrier (ce sont les « *expressions »* dites *« absolues »*). Cette distinction masque en général une visée qui n'est pas explicitée comme telle, qui cherche à opérer une transposition de tous les adverbiaux de localisation temporelle sur le référentiel calendaire : les adverbiaux qui ne sont pas directement transposables sur ce référentiel devraient ainsi être considérés comme relatifs à autre chose.

La présence d'unités du calendrier dans un adverbial de localisation temporelle, aussi bien d'unités ordinales, que d'unités cardinales, ne suffit pas à en faire des adverbiaux calendaires, c'est-à-dire de adverbiaux déterminant un ancrage sur le référentiel du calendrier. Ceci peut sembler contre-intuitif et revient à dire que la présence d'unité calendaire n'est pas ce qui détermine qu'un adverbial de localisation temporelle est un *adverbial calendaire* : on verra plus loin que ce qui permet de typer les adverbiaux de localisation temporelle est leur repère temporel noyau.

L'analyse sémantique vise ici à décrire la succession des opérations qui agissent sur ces repères temporels noyaux, qu'on appellera désormais des *Bases* (base *calendaire, déictique* ou base *relative à un procès*).

Ainsi, recouvrant différentes catégories morphosyntaxiques occupant une fonction de complément adverbial, les adverbiaux de localisation temporelle forment une catégorie homogène au niveau sémantique, leur fonction étant de contribuer à l'ancrage temporel d'un procès, en déterminant un « intervalle circonstanciel » (cf. fig. 11) (sur cette notion d' « intervalle circonstanciel » empruntée à (Gosselin, 2006), on renvoie à la discussion section 2.3). Les adverbiaux de localisation temporelle sont de différents types selon qu'ils déterminent un ancrage temporel par rapport au calendrier, à l'énonciation, à un procès ou par rapport à un repère temporel antérieur dans un texte, repris de façon anaphorique.

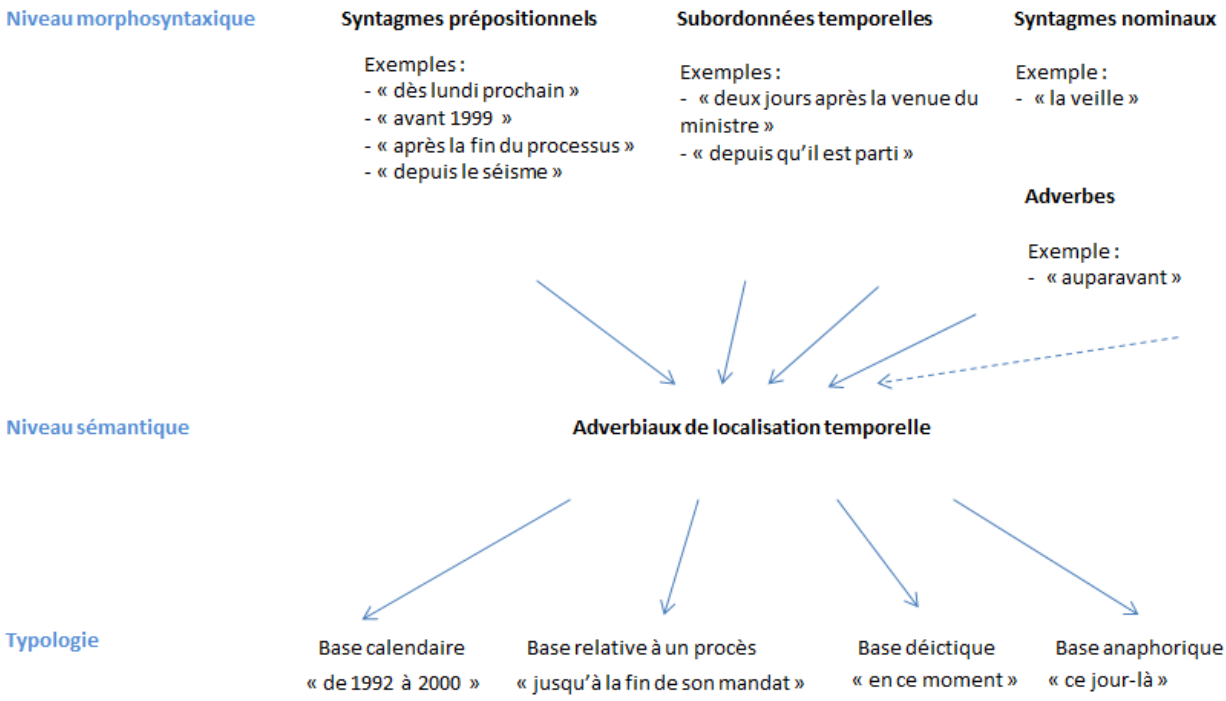


Fig. 11 : une analyse sémantique homogène

4.2 Opérandes et opérateurs sémantiques : une représentation formelle des adverbiaux de localisation temporelle

4.2.1 La démarche d'analyse

Notre analyse s'attache à décrire les adverbiaux de localisation temporelle sous la forme d'une succession d'opérations sur un repère temporel noyau ou *Base*. Elle traite uniformément les adverbiaux de localisation temporelle, quelle que soit la nature de ce repère initial : qu'il s'agisse d'une base calendaire (ex. 1), d'une base déictique (ex. 2), d'une base relative à un procès, éventuellement nominalisé (ex. 3), ou d'une base anaphorique (ex. 4) :

Ex 1 : « **Neuf ans après le 21 avril 2002**, le PS ne semble pas… »
Ex 2 : « Il est avec sa famille en Syrie et retournera en Autriche **dans les prochains jours**. »
Ex 3 : « L'arrestation était intervenue **quelques jours après une visite de travail du président syrien Bachar al-Assad en Autriche**. »
Ex 4 : « Elle est **depuis lors** en proie à des attaques meurtrières et constitue un défi sécuritaire majeur pour le gouvernement du président ivoirien »

A ce niveau d'analyse, conformément à la proposition de (Battistelli *et al.*, 2008), on distingue trois grands types d'opérateurs : les opérateurs de *Régionalisation*, les opérateurs de *Focalisation* et les opérateurs de *Déplacement*. Les opérandes (repères temporels noyaux) au cœur des adverbiaux de localisation temporelle sont nommés des *Bases*.

Les *Opérations de Régionalisation* sont marquées par des prépositions ou des locutions prépositionnelles comme *vers, aux alentours de, depuis, avant*, qui agissent sur le résultat des opérations successives portant une base et permettent de déterminer la région pointée par l'adverbial sur un axe du temps (cf. ex. 5 à 7 où sont surlignés les marqueurs de l'opération).

> Ex. 5 : Certains membres de l'opposition se sont demandé si l'abrogation du programme n'était pas un moyen pour le gouvernement d'éviter la tenue d'un référendum sur le nucléaire les 12 et 13 juin en supprimant sa raison d'être **avant de reproposer la construction de centrales** dans un an.
> Ex. 6 : Le Burkina attend la nomination du gouvernement du nouveau Premier ministre, Luc-Adolphe Tiao, dont la tâche principale va être de tenter de juguler des mouvements de colère multiples, dont ceux de militaires et de jeunes, qui durent **depuis deux mois**.
> Ex. 7 : Revendiquant la paternité du "commando invisible" - des insurgés qui **dès janvier** avait mis en échec les forces pro-Gbagbo dans le nord d'Abidjan -, il interpelle le pouvoir.

L'*Opération de Focalisation* encode les changements de « zoom » (focal) opérés par rapport à la granularité du repère temporel noyau (cf. ex. 8 à 10).

> Ex. 8 : Une usine-pilote en bordure du salar a produit **dès fin 2009** des échantillons de carbonate de lithium, métal mou au fort potentiel électrochimique, utilisé pour les batteries de voiture mais aussi la verrerie ou la médecine.
> Ex. 9 : A Montpellier dimanche, l'ancien Toulousain a collé à l'encéphalogramme de l'équipe: totalement plat **jusqu'au milieu de la seconde période**, où il a perdu beaucoup de ballons et fait preuve d'agressivité mal placée, il a ensuite retrouvé du palpitant pour finir en trombe.
> Ex. 10 : Benoît XVI a déclaré avoir senti le couperet d'une "guillotine" **le jour de son élection** en 2005, et Jean Paul Ier, mort après 33 jours de pontificat en 1978, "ne voulait pas accepter" la lourde charge, selon une confidence du cardinal autrichien Franz König.

Les *Opérations de Déplacement* décrivent les décalages opérés par rapport à la zone temporelle pointée initialement par la base (cf. ex. 11 à 13).

> Ex. 11 : "Comme le président l'a dit **la semaine dernière**, s'attaquer à la situation budgétaire actuelle est largement dans nos capacités en tant que pays", a ajouté Mary Miller, secrétaire adjointe au Trésor chargée des marchés financiers.
> Ex. 12 : Il va être chargé de former un nouveau gouvernement qui devra tenter de mettre un terme aux divers mouvements de contestation, souvent violents, notamment de soldats et de jeunes, qui touchent le Burkina Faso **depuis deux mois**.
> Ex. 13 : Les débats **des prochains mois** seront largement consacrés aux questions de financement et à la place respective des assurances privées et de la solidarité nationale dans le nouveau système de prise en charge de la dépendance.

(Battistelli *et al.*, 08) ajoutaient à ces trois grands opérateurs (Régionalisation, Focalisation et Déplacement), l'opérateur de Pointage, afin d'encoder la nature de l'ancrage sur le référentiel calendaire (pointage d'une zone unique ou multiple du calendrier, pointage de type anaphorique). Les valeurs de cette opération, qui n'apparaît plus dans notre proposition de modélisation, sont redistribuées dans le typage des Bases : le type d'une Base (calendaire, déictique, relative à un procès ou anaphorique), renseigne sur la nature de la référence temporelle qui forme le noyau d'un adverbial de localisation temporelle.

(Battistelli *et al.*, 2008 ; Battistelli, 2009 ; Battistelli, 2011) proposent une formalisation de la sémantique des *adverbiaux calendaires* ou *datatifs* dits « simples » (unaires) qui, sous sa forme générique, peut être présentée comme suit :

OpRegionalisation+ (OpFocalisation/Déplacement+ (OpPointage (Base Calendaire))[28]

L'hypothèse avancée est que cette succession d'opérations se retrouve, toujours identique, dans l'ensemble des adverbiaux calendaires unaires, même si elles ne sont pas nécessairement marquées au niveau morphosyntaxique et même si elles peuvent ne pas modifier la région calendaire désignée par leur opérande. Cette représentation distingue ainsi quatre opérateurs : les opérateurs de régionalisation, de focalisation, de déplacement et de pointage.

> « Un opérateur désigne un ensemble d'instances lexicales possibles liées par une relation d'équivalence. Le critère de regroupement est lié à l'effet qu'un opérateur produit sur une EC [Expression Calendaire]. Ainsi, nous considérons que les expressions « à l'aube des années 1980 » et « au début des années 1980 » désignent des zones utiles identiques (les expressions « à l'aube de » et « au début de » sont donc considérées comme deux instances d'un même opérateur). De même, « pendant » et « en » désignent la même région calendaire, même si elles ne désignent pas la même relation entre la zone calendaire et le procès concerné. Rappelons que notre but est de situer relativement les unes aux autres les EC [Expression Calendaire], et non les procès qu'elles contribuent à situer temporellement. »
>
> (Battistelli *et al.*, 2008)

Les valeurs associées à chacun de ces opérateurs peuvent ainsi s'exprimer en langue par différents marqueurs : la régionalisation de type « Depuis » décrit ainsi la valeur d'une opération qui peut s'exprimer par la préposition « depuis », mais également par la locution prépositionnelle « à partir de », par exemple. Une opération peut ainsi avoir plusieurs valeurs, dont une valeur ID, qui vaut pour l'identité et qui représente une opération qui n'est pas marquée au niveau morphosyntaxique. On présente ci-dessous quelques exemples d'analyse d'adverbiaux calendaires tirés de (Battitstelli *et al.*, 2008) :

Ex. 14 : mai 1974
Modélisation algébrique :
Régionalisation ID (Focalisation/Déplacement ID (base calendaire : mai 1974))

Ex. 15 : à partir de 1958
Modélisation algébrique :
Régionalisation Début (Focalisation/Déplacement ID (base calendaire : 1958))

Ex. 16 : trois jours avant Noël 1994
Modélisation algébrique :
Régionalisation ID (Déplacement (jour,-3) (base calendaire : Noël 1994))

Ex. 17 : au milieu de la fin des années 1990

[28] Les opérations de régionalisation, de focalisation et de déplacement peuvent s'enchâsser récursivement, alors que l'opération de pointage ne peut apparaître qu'une fois.

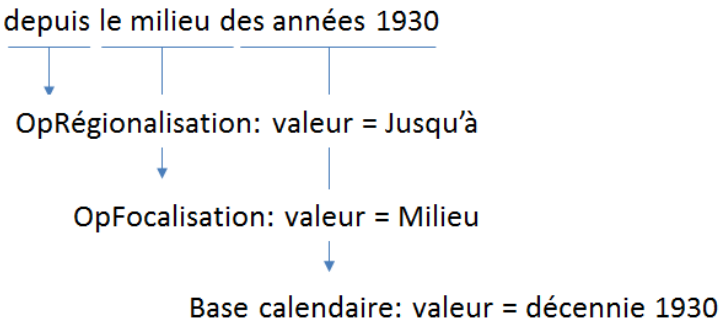
Modélisation algébrique :
Régionalisation ID (Focalisation Milieu (Focalisation Fin (base calendaire : années 1990)))

Nos travaux ont visé à généraliser cette analyse pour l'étendre à l'ensemble des adverbiaux de localisation temporelle. La formalisation initiale a également été enrichie de sorte à pouvoir traiter des adverbiaux dits « composés », c'est-à-dire des adverbiaux formés à partir de plusieurs adverbiaux de localisation temporelle plus simples (*dans la nuit de mardi à mercredi, vers 2h du matin*, par exemple).

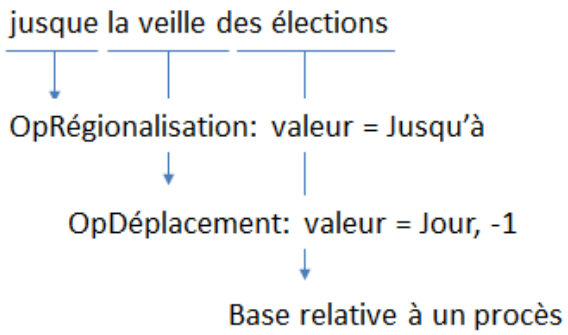
La proposition de modélisation que l'on détaille ici cherche à décrire les adverbiaux de localisation temporelle à l'aide d'une liste fermée d'opérations sémantiques. Cette description dépend uniquement des instances lexicales présentes dans les adverbiaux de localisation temporelle et, éventuellement, de celles présentes dans leur contexte : on ne s'intéresse pas, à ce niveau d'analyse, à leur valeur calendaire (quelle date leur associer ?). Ces instances lexicales permettent de déterminer la valeur associée à chacune des opérations sémantiques à l'aide desquelles on représente les adverbiaux de localisation temporelle (cf. ex 18 et 19). Une opération peut toutefois ne pas être marquée au niveau morphosyntaxique (cf. l'opération de Régionalisation dans l'exemple 20) et une même instance lexicale peut déterminer conjointement la valeur de plusieurs opérations (cf. les opérations de déplacement et de focalisation dans l'exemple 20).

L'ordre séquentiel des instances lexicales qui déterminent les valeurs associées aux opérations ne suit donc pas nécessairement celui de la succession des opérations à l'aide desquelles on représente les adverbiaux.

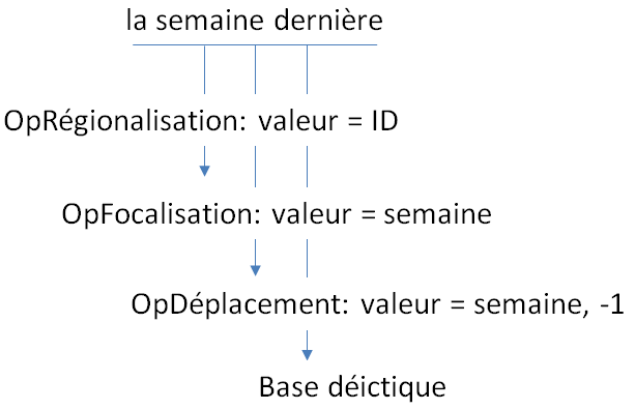
Ex. 18 :



Ex. 19 :

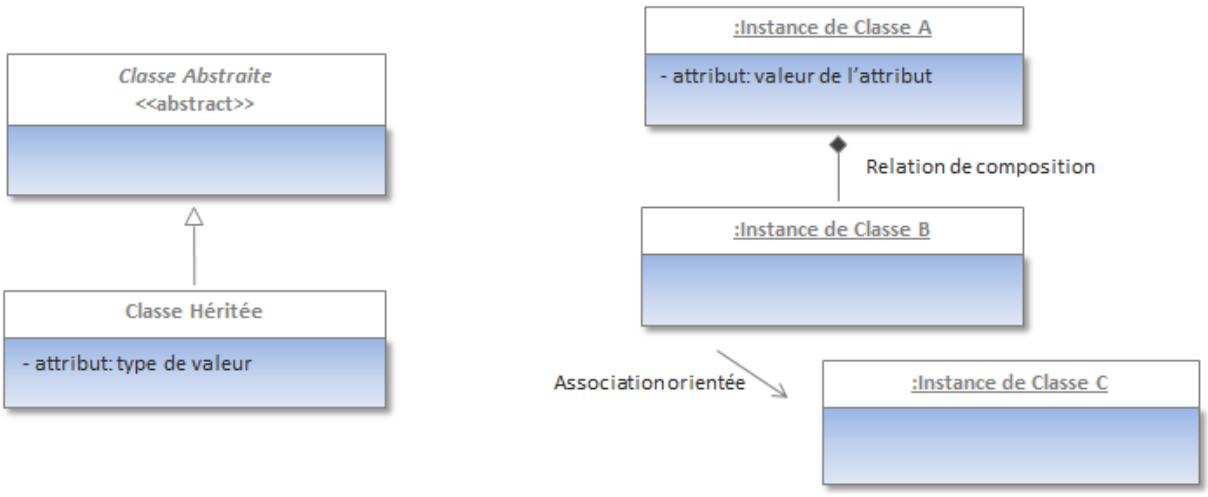


Ex. 20 :



On présente par la suite cette proposition de formalisation au format UML, car c'est sur ces modèles que repose en grande partie l'opérationnalisation des traitements automatiques développés pour manipuler les instances de cette représentation formelle des adverbiaux de localisation temporelle.

UML (pour *Unified Modeling Language* ou « langage de modélisation unifié ») est un langage de modélisation graphique couramment utilisé pour le développement logiciel. Il peut servir notamment, comme ici, pour représenter des structures de données, sous la forme d'objets reliés entre eux. Nous avons ici essentiellement recours aux concepts de classes, de relations entre classes et aux diagrammes d'objets.



Une classe décrit le type, les propriétés et attributs d'un ensemble d'objets. Les éléments de cet ensemble sont les instances de la classe. Les relations entre les classes sont représentées par des arcs. Elles permettent de décrire des relations sémantiques entre deux classes. En particulier, on peut représenter des relations d'héritage, de composition et des associations dont on peut préciser la sémantique. Les classes peuvent être liées entre elles grâce au mécanisme d'héritage qui permet de mettre en évidence des relations de parenté. Une classe abstraite (<<abstract>>) est une classe qui ne peut pas être instanciée : elle sert de dénominateur commun à d'autres classes qui en dérivent (les classes héritées). Les diagrammes d'objets permettent eux de représenter les instances

d'une classe. Les valeurs d'un attribut d'une classe peuvent être typées et contraintes par une liste de valeurs (<<enumeration>>).

4.2.2 Typologie des Bases

Comme on l'a vu, le modèle permet de distinguer plusieurs types de repères temporels noyaux (*Base*) au cœur des adverbiaux de localisation temporelle (cf. fig. 12) : les bases calendaires (*en 1871*), les bases relatives à un procès (*la veille des élections*), les bases déictiques (*le mois prochain*) et les bases anaphoriques (*depuis ce jour*).

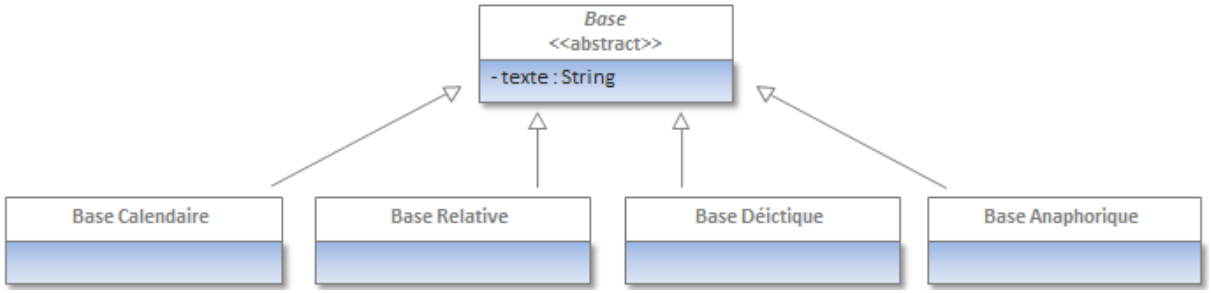


Fig. 12 : typologie des bases

Les *Bases Calendaires* sont abondamment décrites dans les divers travaux d'ingénierie des langues qui s'attachent à analyser les expressions temporelles. En effet, le plus souvent l'accent est mis sur cette base calendaire à laquelle est parfois réduite la catégorie des expressions temporelles. Les *Bases Calendaires* sont composées d'une *Unité Calendaire* au moins. Elles peuvent néanmoins en contenir plusieurs (*les 2, 3 et 4 février 2012*).

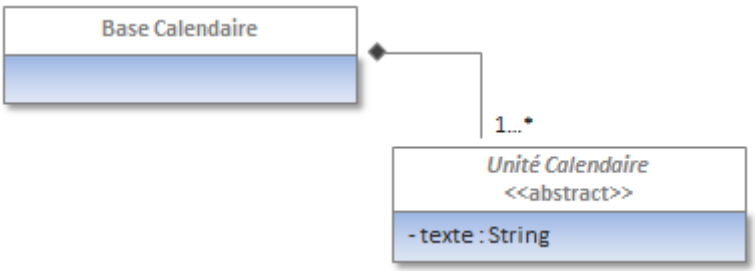


Fig. 13 : les bases calendaires

Au sein du système du calendrier, il convient de distinguer les unités *ordinales* (unités de datation) et les unités *cardinales* (unités de durée)[29].

[29] On renvoie à ce sujet à (Battistelli *et al.*, 2006) : « les systèmes de datation actuels possèdent la double dimension de tout système de mesure : ordinale en lecture longitudinale (à une unité donnée) et cardinale en lecture verticale (c'est-à-dire en passant d'une unité à une autre). »

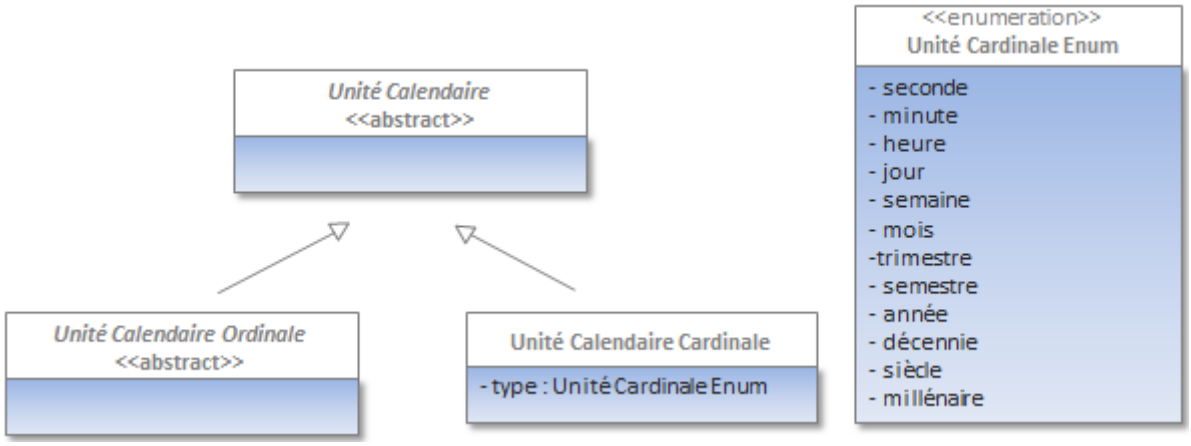


Fig. 14 : les unités calendaires

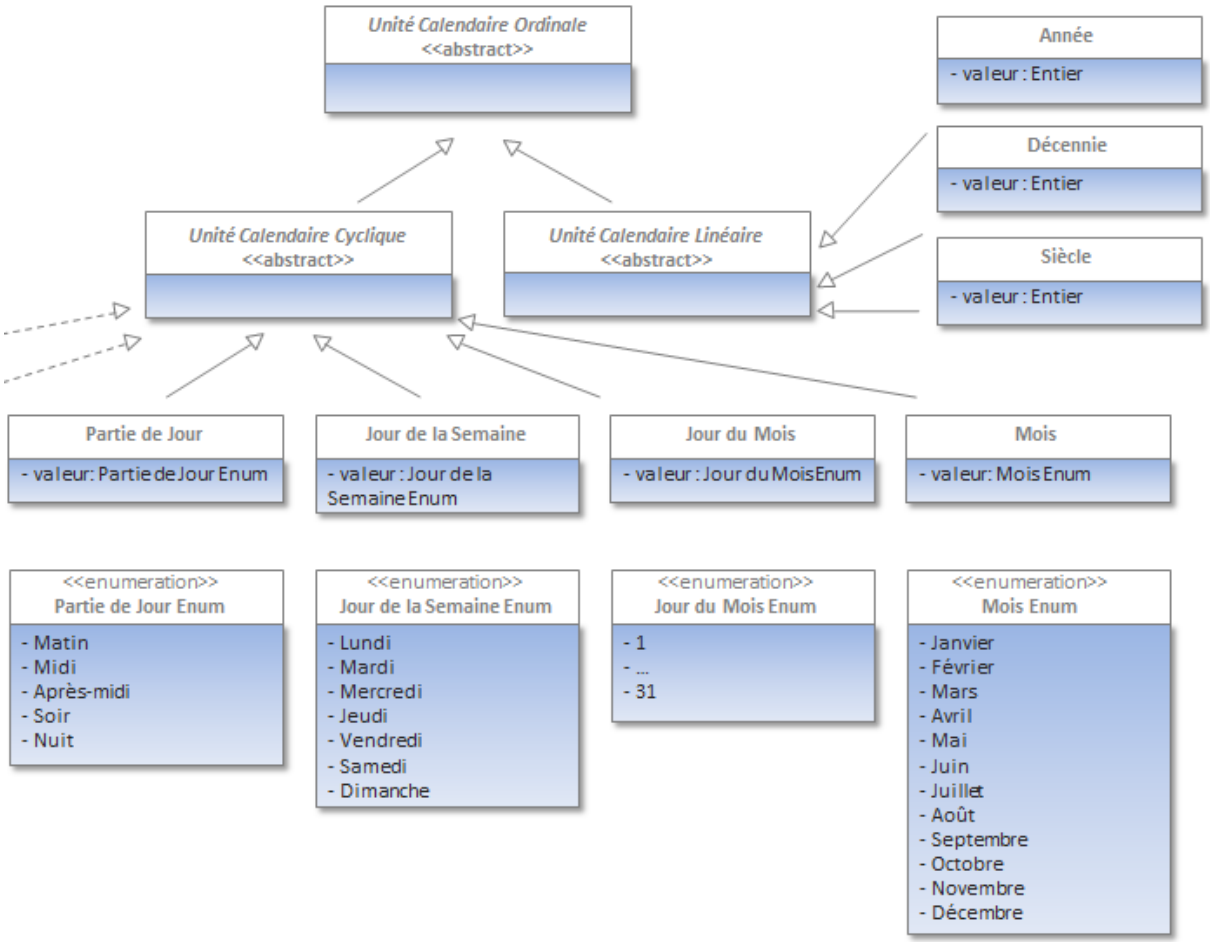


Fig. 15 : les unités ordinales[30]

Les unités linéaires déterminent une zone unique du calendrier (une année précise, une décennie ou encore un millénaire, par exemple) : elles renvoient à la dimension linéaire du calendrier dont elles pointent une zone bien identifiée (*en 2002, depuis octobre 2005, avant les années 1920)*.

[30] Ce modèle peut être étendu et raffiné pour tenir compte du nombre de jours des différents mois ou encore pour pouvoir exprimer des granularités plus fines (minutes, heures, secondes).

A côté de ces unités qui renvoient à la linéarité du calendrier, il faut compter celles qui renvoient à sa dimension cyclique (les unités *cycliques*), qui, elles, désignent plusieurs portions du calendrier (cf. 1 à 3).

> Ex. 1 : Ouverture : **tous les jours de 10 heures à 22 heures, sauf le mardi**.
> Ex. 2 : Nous en avons l'habitude **au début de chaque mois**, l'institut NPD nous permet d'avoir un aperçu du mois écoulé aux USA en nous offrant le classement des 10 titres les plus vendus (…).
> Ex. 3 : **Chaque année, avant la fin du premier trimestre**, un catalogue d'actions est proposé à tous les établissements de l'académie.

Ces distinctions renvoient à la double dimension du système calendaire, à la fois cyclique et linéaire[31]. La fig. 16 présente un diagramme d'objet associé à la *Base Calendaire* « *2e trimestre 2012 ».*

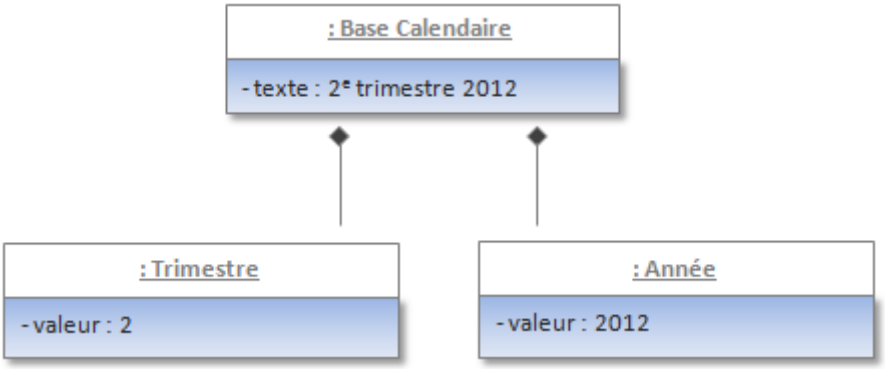


Fig. 16 : diagramme d'objet pour la Base Calendaire 2e trimestre 2012

4.2.3 Les adverbiaux de localisation unaires

Au cœur de notre proposition de modélisation des *Adverbiaux de Localisation Temporelle Unaire* (cf. fig. 17), on retrouve presqu'inchangée la formalisation proposée dans (Battistelli *et al.*, 2008). Le modèle des *Adverbiaux de Localisation Temporelle Unaire* permet de généraliser la notion d'*Adverbial Calendaire*, qui recouvrait peu ou prou les adverbes dont le noyau est formé par une base dite « calendaire ».

[31] La dimension cyclique n'est pas liée uniquement au système calendaire : un événement peut également être cyclique (« ***À chaque parution de notre revue***, *nous vous offrons un texte plus étoffé à approfondir.* »). A ce sujet on renvoie à (Battistelli *et al.*, 2006) qui, du fait de leur caractéristique cyclique, tend à considérer ce type d'itération comme formant des grains calendaires.

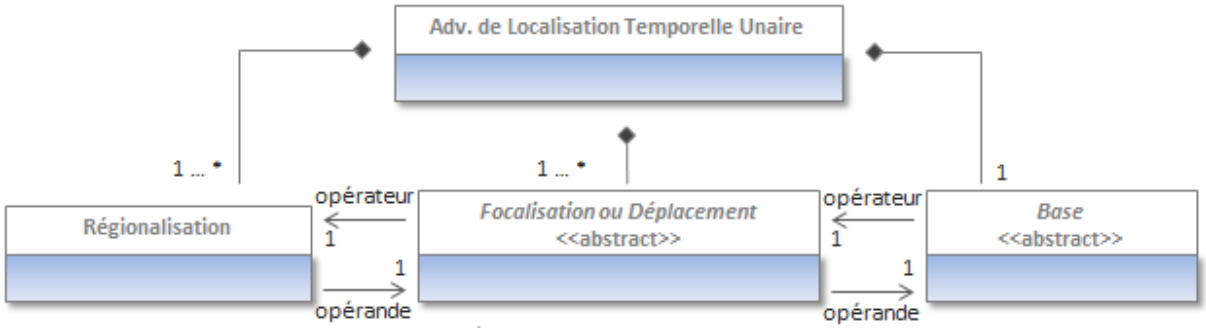


Fig. 17 : le modèle des adverbiaux de localisation temporelle unaire

Cette modélisation consiste à décrire les adverbiaux de localisation temporelle sous la forme d'une succession d'opérations qui agissent sur un repère temporel initial (Base). Il s'agit d'associer à ces opérations des valeurs qui dépendent des marqueurs linguistiques présents dans les adverbiaux. Le modèle descriptif que l'on reprend ici pour l'enrichir distingue ainsi des opérations de *Focalisation* (« *au début du siècle* », « *vers la mi-mars* », « *durant la seconde moitié de l'année* », etc.), de *Déplacement* (« *deux jours avant sa venue* ») et de *Régionalisation* (« *avant cette période* », « *depuis mars 2009* », « *jusqu'à cette date* », etc.). On décrira plus bas les valeurs associées à chacune de ces opérations.

Plusieurs opérations de même nature sont susceptibles de s'emboîter[32]. Dans l'exemple ci-dessous (cf. fig. 18), qui figure un diagramme d'objet pour l'adverbial *jusque vers la fin des années 1920*, on voit ainsi que le modèle permet de représenter des adverbiaux où deux *Opérations de Régionalisation* sont enchâssées (*jusque* et *vers*).

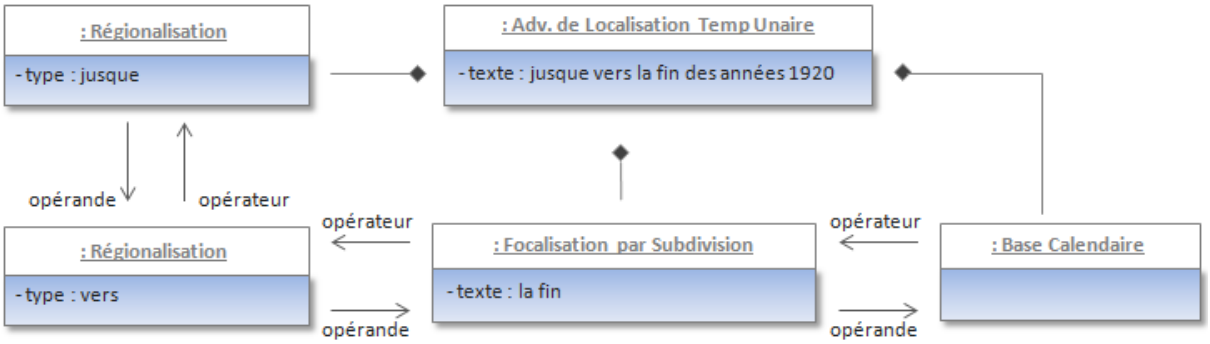


Fig. 18 : diagramme d'objet pour l'adverbial jusque vers la fin des années 1920

Les opérations de régionalisation, de focalisation ou déplacement ne sont pas nécessairement marquées au niveau morphosyntaxique : ainsi dans le diagramme d'objet ci-dessous, qui figure l'analyse de l'expression *quelques jours avant la fin de son mandat*, l'*Opération de Régionalisation*

[32] Il y aurait cependant une grammaire fine à dégager pour contraindre davantage le modèle, qui est plus permissif que la langue, car il semble qu'il ne soit pas possible d'enchaîner tout type d'opérateur, ni non plus de le faire dans un ordre libre (**depuis vers jusque la mi-mars*). En outre, l'*Opération de Déplacement* ne semble pas pouvoir former une expression avec toutes les valeurs possibles des *Opérateurs de Régionalisation* (**vers/depuis deux jours avant les élections*).

n'est pas marquée ; elle laisse l'expression inchangée. Le modèle prévoit ainsi la valeur ID, qui vaut pour l'identité et qui permet d'encoder ces opérations neutres.

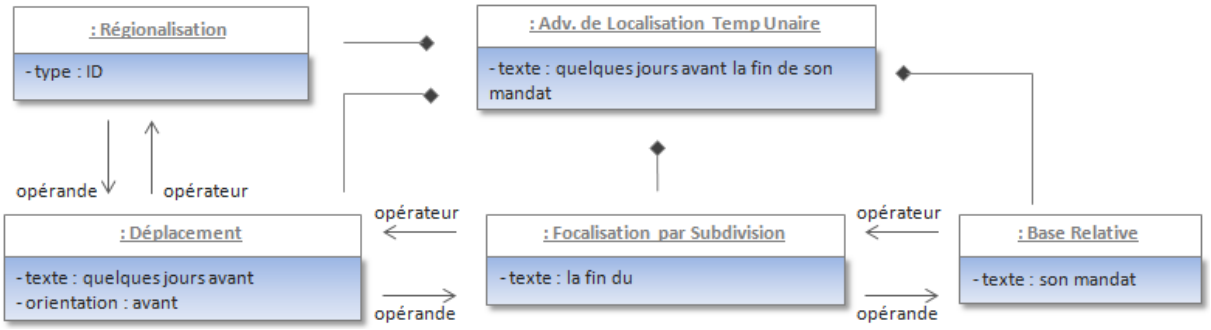


Fig. 19 : diagramme d'objet pour l'adverbial quelques jours avant la fin de son mandat

4.2.4 Opérations de Focalisation et de Déplacement

Les Opérations de Déplacement et de Focalisation sont placées au même niveau, car elles peuvent s'imbriquer l'une dans l'autre, récursivement.

Les valeurs des *Opérations de Déplacement* décrivent différentes informations en relation avec la façon dont un déplacement s'opère à partir de la zone temporelle pointée par la base. Ainsi, pour des adverbiaux tels que « *deux jours avant Noël* » ou « *la semaine prochaine* », depuis le repère temporel noyau (Noël d'un côté, le processus énonciatif de l'autre), un déplacement s'effectue vers l'*avant* dans la première expression, vers l'*après* pour la seconde. Comme pour toutes les opérations (déplacement, focalisation et régionalisation), l'*Opération de Déplacement* peut laisser l'expression inchangée.

L'*Opération de Focalisation* encode l'éventuel changement de focal (« zoom ») opéré par rapport à la granularité du repère temporel noyau, comme pour les expressions suivantes : « *vers la fin de l'année* », ou « *au début des années 60* ».

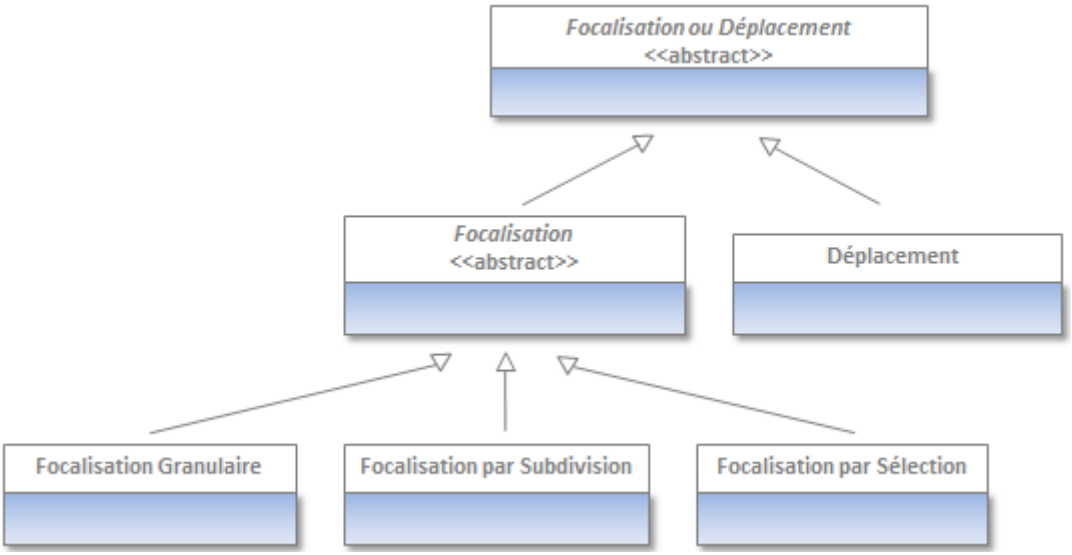


Fig. 20 : les Opérations de Focalisation et de Déplacement

Les opérations de déplacement et de focalisation peuvent toutes deux contenir l'expression de durée. Les *Durées* (fig. 21) font intervenir des *Unités Calendaires Cardinales* (« *près de 3 jours avant* »), des quantités (*« lors des deux dernières semaines précédant les élections »)* et des opérateurs dits de comparaison (« *près de trois mois après* »).

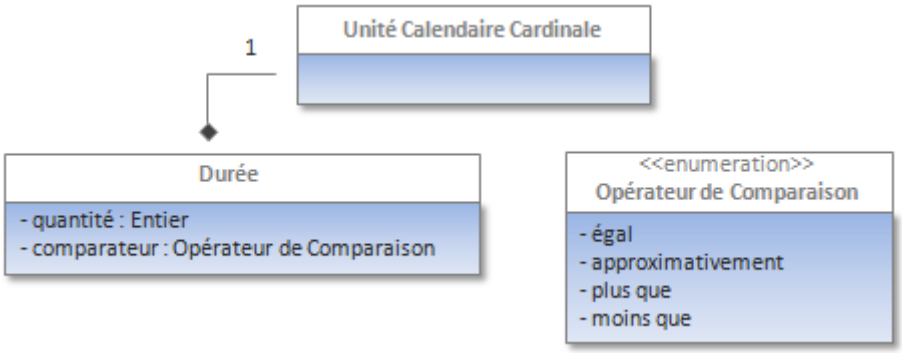


Fig. 21 : le modèle des Durées

Le modèle pourrait être raffiné pour pouvoir tenir compte de quantités imprécises (« depuis quelques jours ») ou encore pour pouvoir prendre en charge l'expression fine de degrés dans les opérations de comparaison (« *un peu plus d'une semaine après* », par exemple).

4.2.4.1 Opération de Focalisation

Les opérations de *Focalisation* peuvent être de plusieurs types : on distingue ainsi la focalisation granulaire (*cette semaine-là*), la focalisation par subdivision (*jusque mi-août*) et la focalisation par sélection (*la 1e semaine de l'année*). La focalisation permet également l'expression d'une certaine granularité ou *Unité Calendaire Cardinale* (*ce jour-là* ; *l'année des élections*). Cette opération peut également être composée d'une *Unité Calendaire Ordinale* (*ce vendredi-là* ; *le dimanche du vote*).

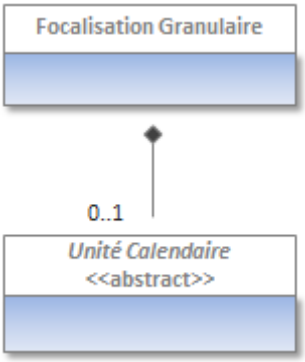


Fig. 22 : la Focalisation Granulaire

Dans l'exemple ci-dessous (cf. fig. 23), qui figure le diagramme d'objet associé à l'adverbial *depuis cette semaine-là*, la focalisation décrit la granularité associée à l'anaphore (le grain *semaine*).

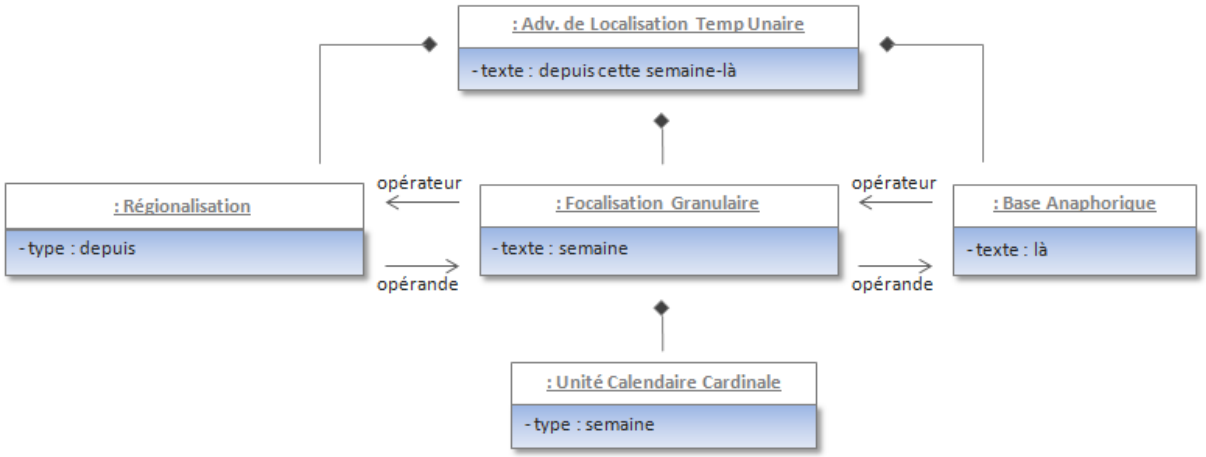


Fig. 23 : diagramme d'objet associé à l'adverbial depuis cette semaine-là

La Focalisation par Subdivision consiste à réduire l'empan de la zone temporelle désignée par l'opérande à une portion de celle-ci (« *à la fin des années 80* », « *durant le dernier tiers de son mandat* »).

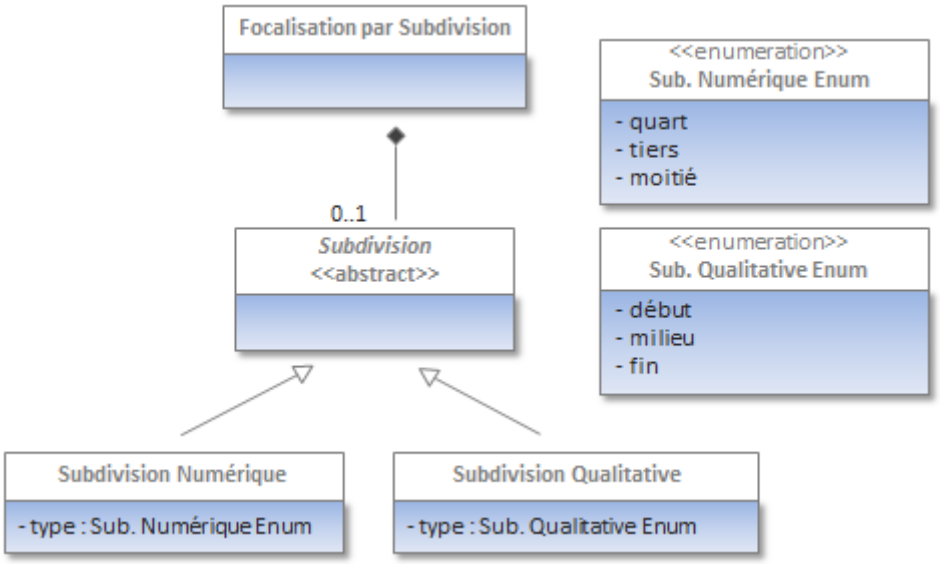


Fig. 24 : la Focalisation par Subdivision

Dans l'exemple ci-dessous (cf. fig. 25), qui figure le diagramme d'objet associé à l'adverbial *quelques jours avant la fin du semestre 2011*, la focalisation réduit l'empan de la zone temporelle associée à l'opérande (*2nd semestre 2011*) à la fin de celle-ci.

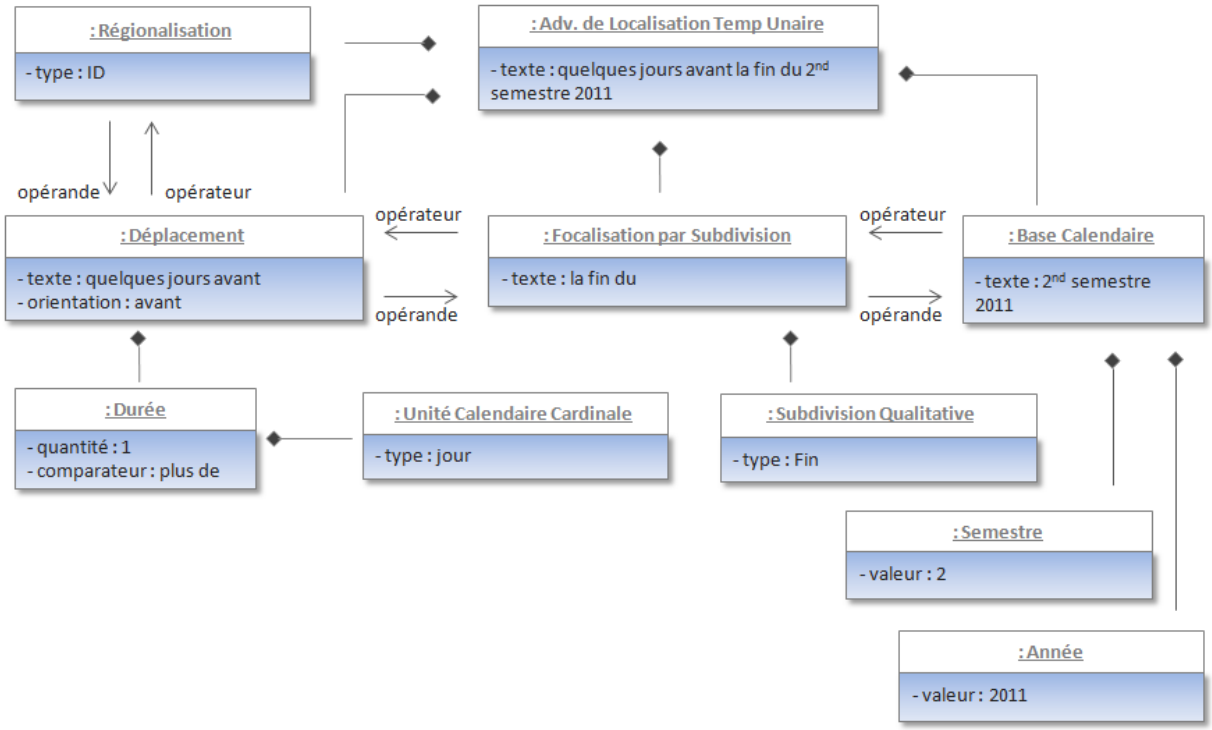


Fig. 25 : diagramme d'objet associé à l'adverbial quelques jours avant la fin du 2nd semestre 2011

La *Focalisation par Sélection* est constituée d'un *Rang* (« *depuis le premier jour de son déplacement* ») et d'une *Durée* (« *dès les deux premières semaines de son mandat* ») ou d'une *Unité Calendaire* (« *jusqu'à la dernière minute* »).

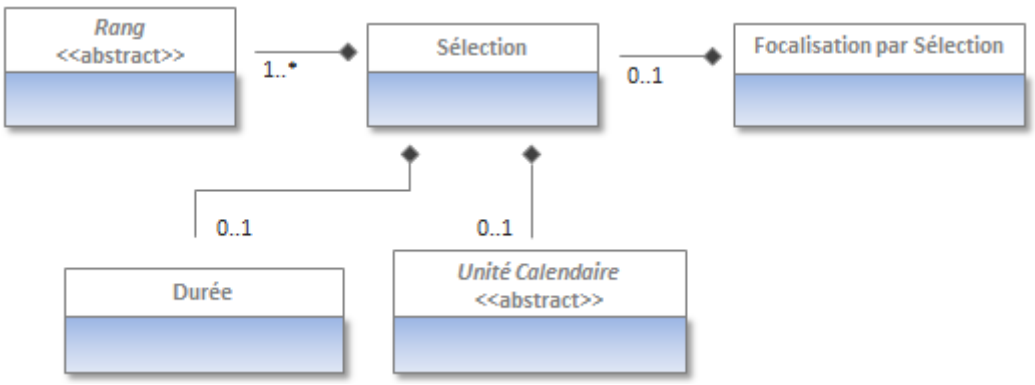


Fig. 26 : la Focalisation par Sélection

Dans l'exemple ci-dessous (cf. fig. 27), qui figure le diagramme d'objet associé à l'adverbial *les 1er et 2e mercredis de chaque mois*, la focalisation permet de sélectionner deux portions de la zone temporelle associée à l'opérande (*chaque mois*).

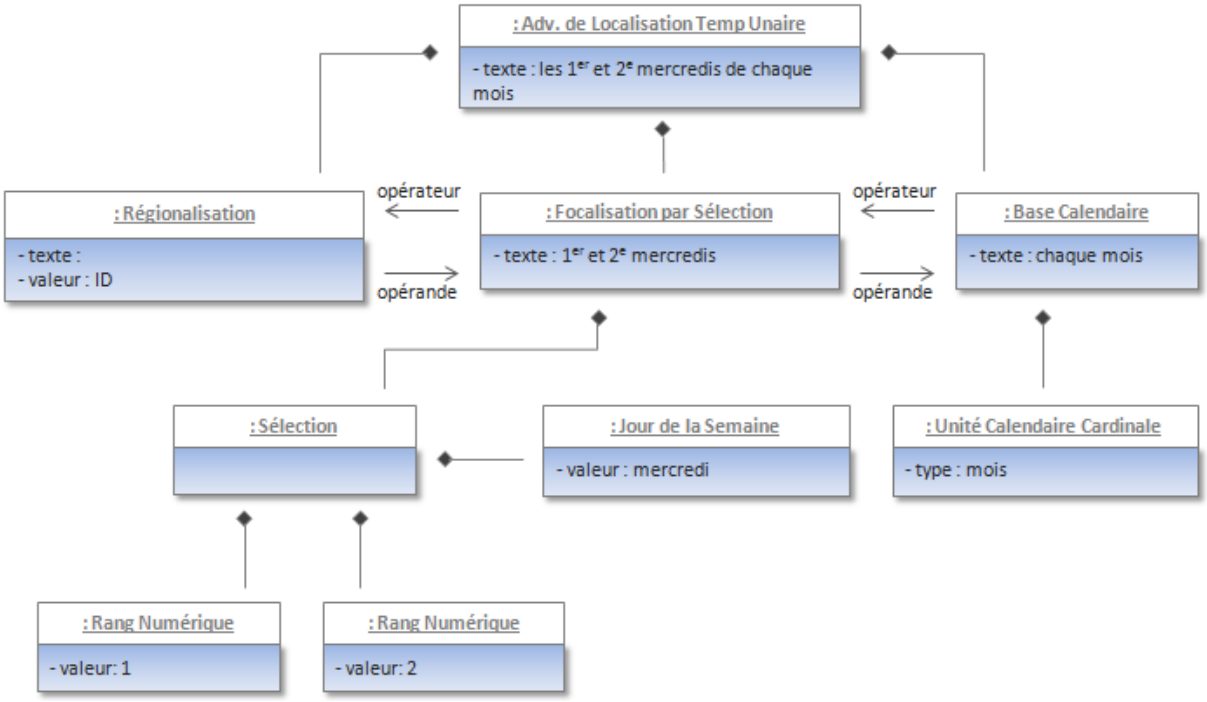


Fig. 27 : diagramme d'objet associé à l'adverbial les 1er et 2e mercredis de chaque mois

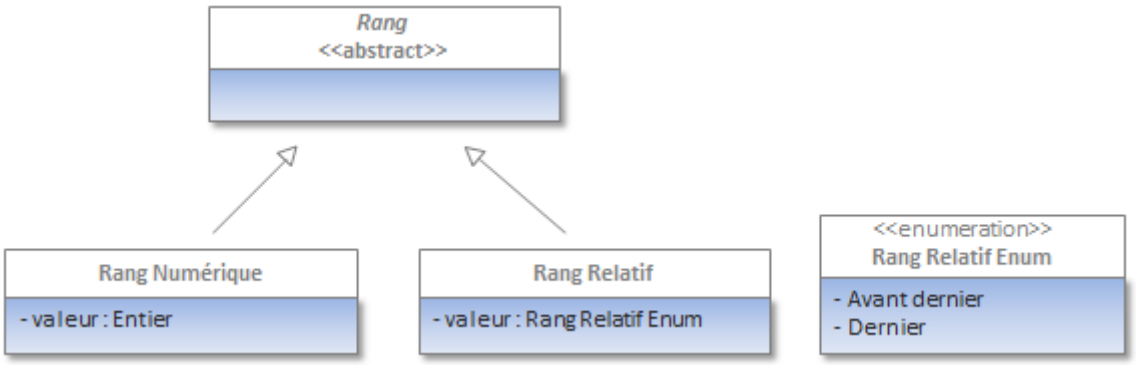


Fig. 28 : le modèle des Rangs

La figure 29 montre la façon de décrire l'*Opération de Focalisation* pour l'adverbial *lors des deux dernières semaines de son séjour*.

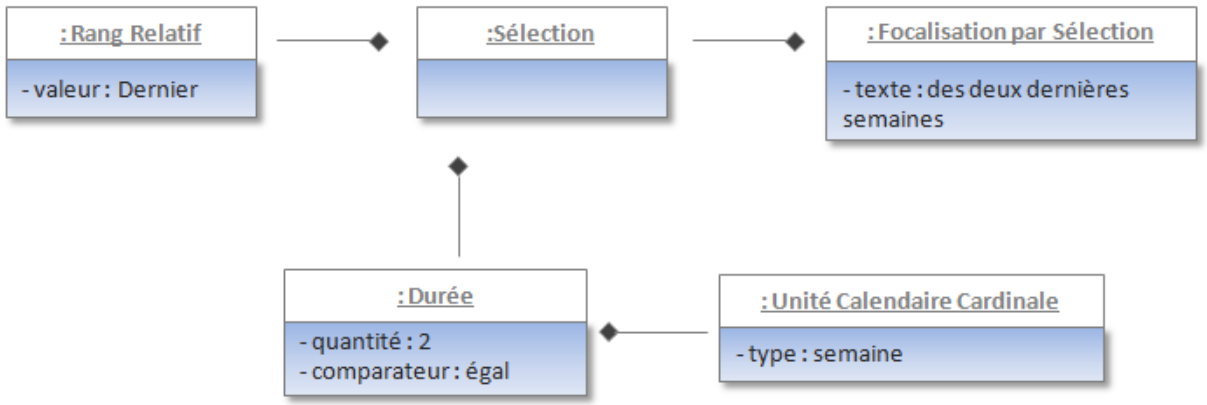


Fig. 29 : diagramme d'objet décrivant l'Opération de Focalisation dans l'adverbial lors des deux dernières semaines de son séjour

4.2.4.2 Opération de Déplacement

Les Opérations de Déplacement peuvent contenir l'expression d'un *Rang* (*«le dernier mois avant la fin de ce délais »)*, d'une *Durée* (« *deux mois avant la fin* ») ou encore d'une *Unité Calendaire Ordinale* (« *jusqu'au vendredi qui suit son arrivée* »). Elles contiennent toujours l'expression d'une orientation, dont les valeurs *avant* et *après* sauf si l'opération laisse l'expression inchangée ; dans ce la valeur est ID.

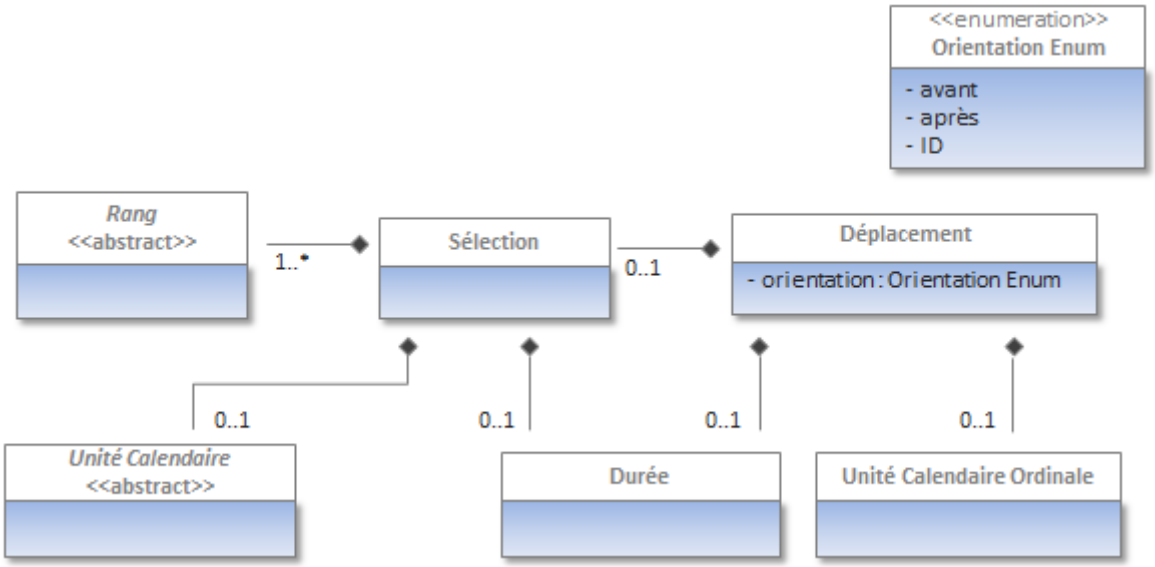


Fig. 30 : les opérations de Déplacement

La figure 31 montre la façon de décrire l'*Opération de Déplacement* dans l'expression *près de trois jours avant*.

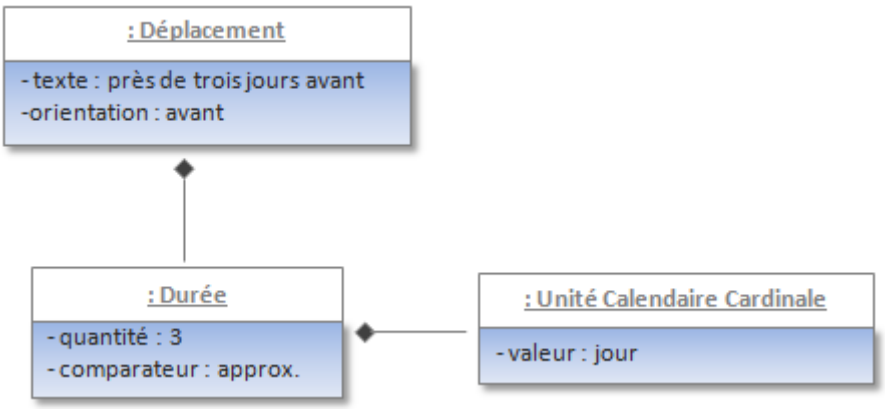


Fig. 31 : diagramme d'objet décrivant l'Opération de Déplacement dans l'adverbial près de trois jours avant

Le jeu des opérations de focalisation et de déplacement permet de rendre compte de la différence entre les adverbiaux « il y a un mois » (déplacement vers l'avant d'un mois, sans qu'une focalisation permette de déterminer la granularité à laquelle on se place (cf. fig. 32)) et « le mois dernier » (focalisation à l'échelle du mois opérant sur la base déictique, objet d'une opération de déplacement d'un mois vers l'avant (cf. fig. 33)).

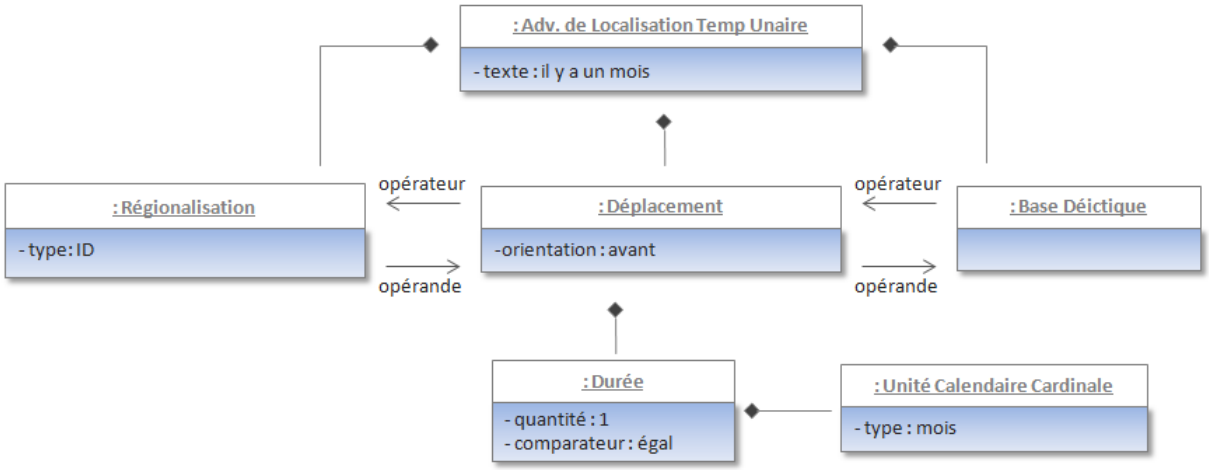


Fig. 32 : diagramme d'objet associé à l'adverbial il y a un mois

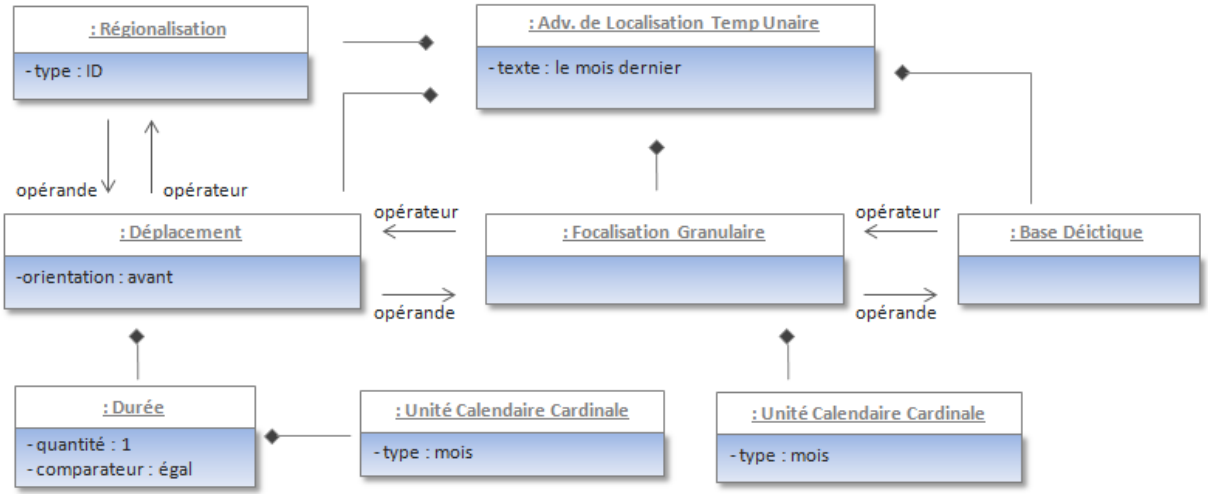


Fig. 33 : diagramme d'objet associé à l'adverbial le mois dernier

4.2.5 Opération de régionalisation

Les valeurs d'une *Opération de Régionalisation* permettent d'encoder la sémantique des prépositions et locutions prépositionnelles comme *vers, aux alentours de, depuis, avant*, qui agissent sur le résultat des opérations successives portant sur une base et déterminant la région pointée sur l'axe du temps. L'*Opération de Régionalisation* spécifie ainsi la région temporelle résultante (« *depuis la fin du siècle* », « *avant la mi-août* », etc.).

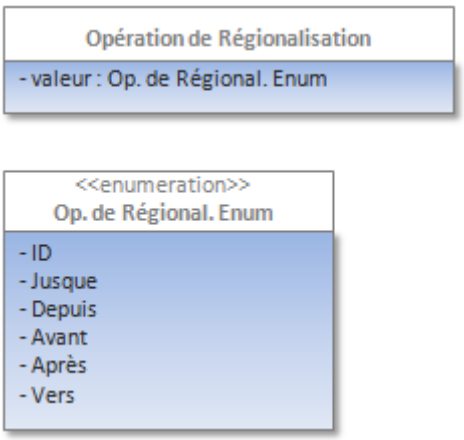


Fig. 34 : l'Opération de Régionalisation

L'exemple ci-dessous (fig. 35) illustre ainsi une analyse complète pour l'adverbial *« depuis bientôt trois semaines »*.

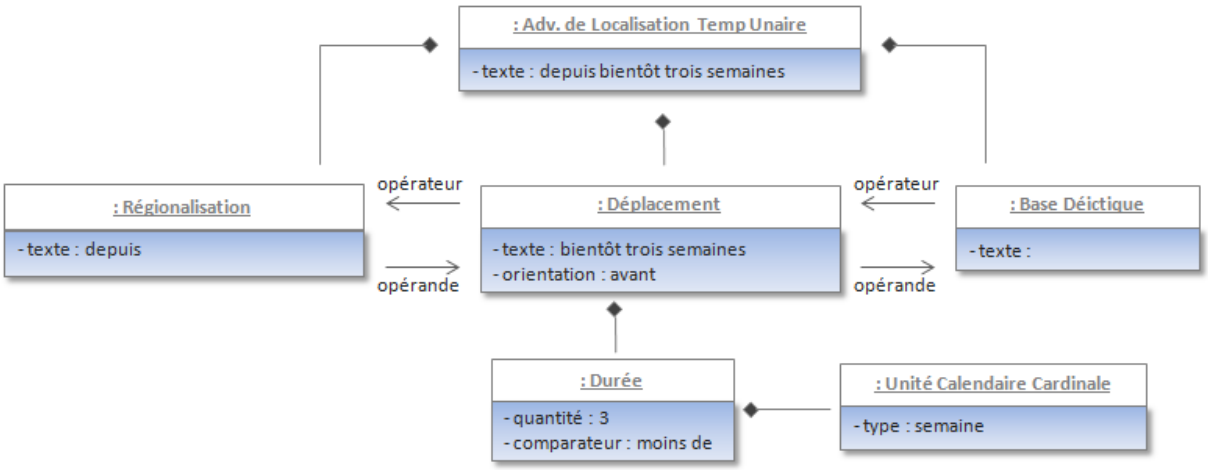


Fig. 35 : diagramme d'objet associé à l'adverbial depuis bientôt trois semaines

Dans les adverbiaux de localisation unaire, Focalisation, Déplacement et Régionalisation forment ainsi les trois grands types d'opérateurs qui agissent sur l'opérande que constitue une Base.

4.2.6 Les adverbiaux de localisation n-aires

Les *Adverbiaux de Localisation Temporelle* peuvent être de trois types (cf. fig. 36) : soit des adverbiaux unaires, soit des adverbiaux binaires, soit des adverbiaux composés[33].

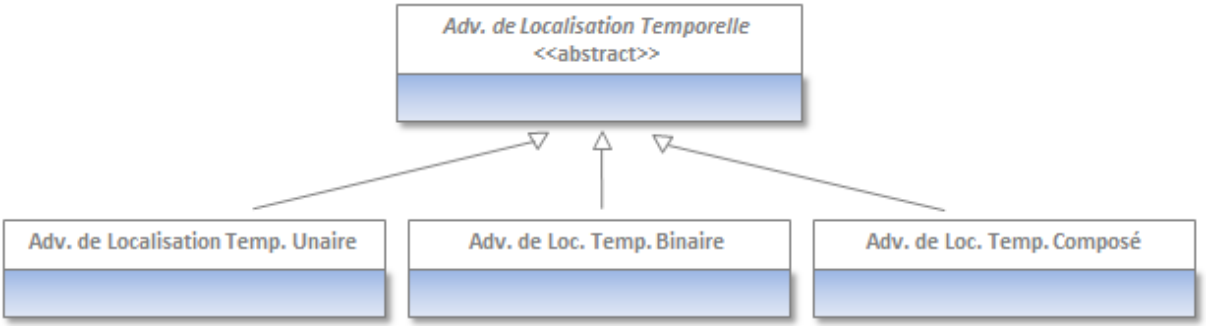


Fig. 36 : les types d'adverbiaux de localisation temporelle

Les adverbiaux unaires pointent sur un unique repère temporel (Base). Unique ici ne signifie pas que la base ne renvoie qu'à une zone unique d'un axe de temps : en effet, la base au cœur d'un adverbial unaire peut être itérative et donc désigner plusieurs zones sur l'axe du temps (*peu avant le début de chaque mois ; dès qu'il se réveille*).

Les *Adverbiaux de Localisation Binaires* sont composés de deux bases (par exemple : « *de la fin du mois d'août jusqu'à la reprise des cours* »). Les adverbiaux composés sont formés par l'association de plusieurs adverbiaux de localisation (*de mars à avril, dès 19h*).

[33] On verra par la suite que les adverbiaux binaires peuvent être analysés comme la résultante d'une opération de composition entre des adverbiaux de localisation temporelle unaires.

4.2.6.1 Les Adverbiaux de localisation Binaires

Les *Adverbiaux Binaires* forment une sous-catégorie des *Adverbiaux de Localisation Temporelle*. Composés de deux *Bases*, l'une formant le début, l'autre la fin, les *Adverbiaux Binaires* peuvent présenter, autour de chacune des bases, l'ensemble des opérations dégagées dans la modélisation des adverbiaux unaires. En ce sens, on peut ainsi considérer les *Adverbiaux Binaires* comme la résultante d'une *Opération de Composition* entre deux expressions de localisation unaires. Toutefois, les valeurs des *Opérations de Régionalisation* y sont contraintes : les *Adverbiaux Binaires* peuvent être de deux formes, soit ils respectent le schéma du type *depuis… jusqu'à…* (cf. ex. 1 et 2) soit celui du type *entre… et…* (cf. ex. 3 et 4). Dans le premier cas, les Opérations de Régionalisation qui composent l'adverbial ont respectivement pour valeur *depuis* et *jusque*, dans le second cas ces valeurs sont ID.

Ex. 1 : De fin mars à début avril 2013
Ex. 2 : depuis la fin des années 30 jusque vers la seconde guerre mondiale
Ex. 3 : entre avril et mai
Ex. 4 : entre fin 2010 et le renouvellement complet du Sénat

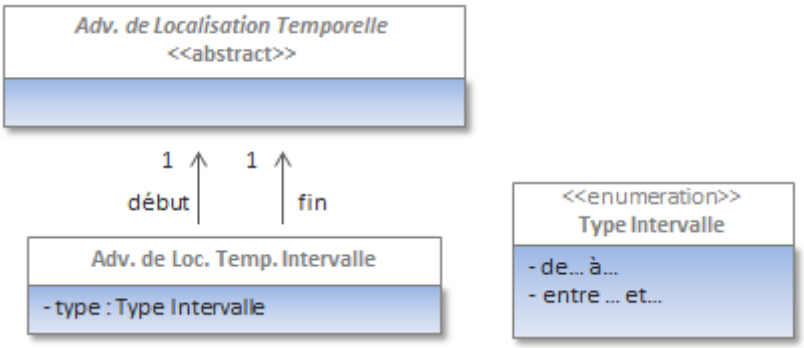


Fig. 37 : les Adverbiaux Binaires

Le diagramme ci-dessous est une représentation synthétique de l'analyse de l'adverbial *de fin mars à mi-juin*.

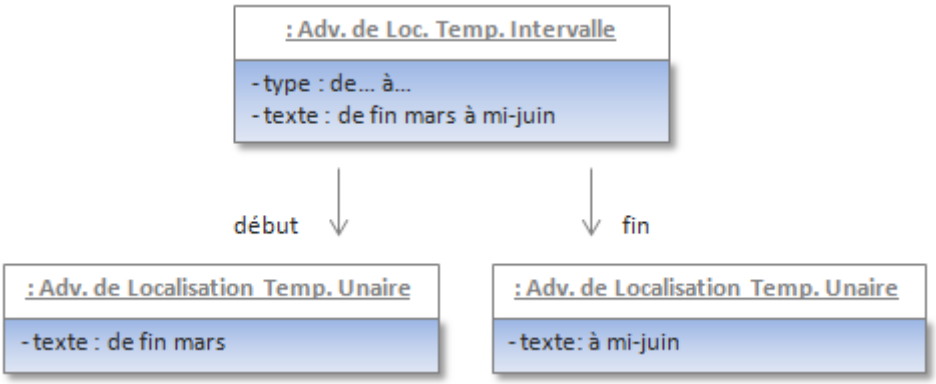


Fig. 38 : diagramme d'objet associé à l'adverbial de fin mars à mi-juin

4.2.6.2 Les Opérations de Composition

Les adverbes de localisation temporelle se déclinent en adverbes unaires (« simples ») et en adverbes composés, formés par l'association d'un adverbe de localisation temporelle avec d'autres adverbes temporels. On a vu en effet qu'un adverbe pouvait porter sur un autre adverbe pour former un nouvel adverbe. Les adverbiaux de localisation temporelle peuvent ainsi se combiner avec des adverbiaux de durée (ex. 1 à 3), des adverbiaux fréquentatifs ou répétitifs (ex. 4 à 6), ou avec d'autres adverbiaux de localisation temporelle (ex. 7 à 9).

Ex. 1 : à partir de 8h, pendant 2h
Ex. 2 : après son arrivée, en une heure
Ex. 3 : cette année-là, en deux mois

Ex. 4 : deux fois par semaine, à partir du 1er août
Ex. 5 : depuis les élections, deux fois par jour
Ex. 6 : hier, à raison d'une fois par heure

Ex. 7 : du 1er au 30 janvier, les lundis, mardis et jeudis, de 18h à 20h
Ex. 8 : pendant son séjour, tous les matins
Ex. 9 : le lendemain, à 8h

A ce stade, notre proposition de modélisation ne couvre que les associations entre adverbiaux de localisation temporelle (ex. 7 à 9). Les adverbiaux qui entrent dans la composition de ce type d'association sont susceptibles d'occuper différentes fonctions et entretenir différents liens : fonction de spécification (« *vendredi, à 10h* »), lien de concaténation (« *en mars et en septembre* »), lien de disjonction exclusive (« *à son retour ou un peu après* ») ou encore le rôle d'exception (« *toute l'année, hors périodes de vacances scolaires* »).

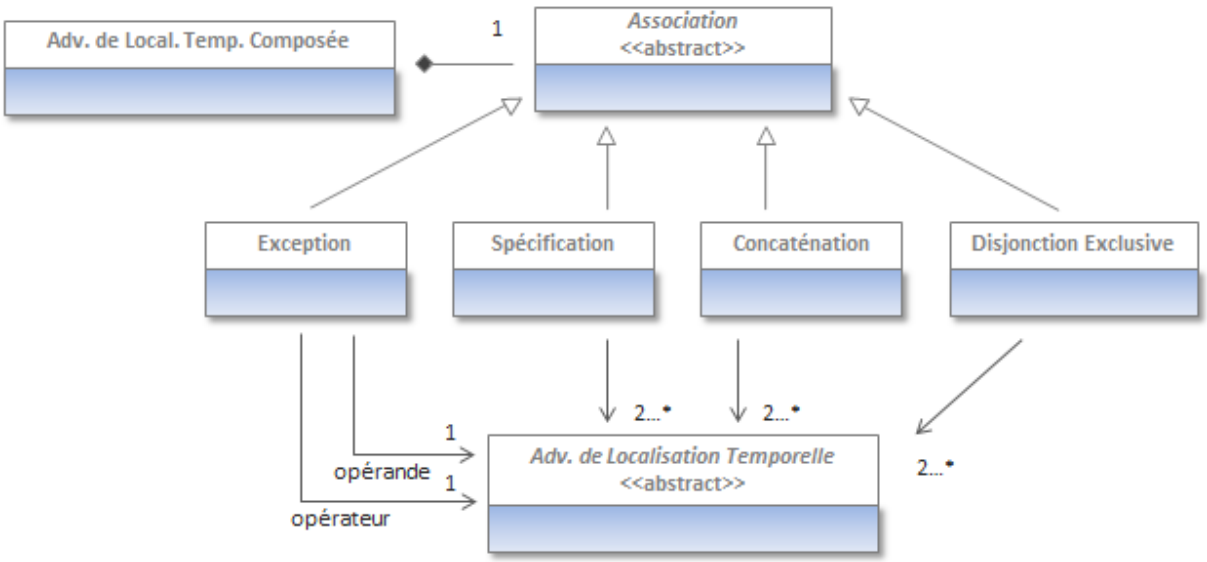


Fig. 39 : les adverbiaux composés

L'exemple ci-dessous (fig. 40) montre de façon synthétique l'analyse proposée pour l'adverbial composée *du lundi au samedi, de 10h à 18h, sauf les 1er et 8 mai*.

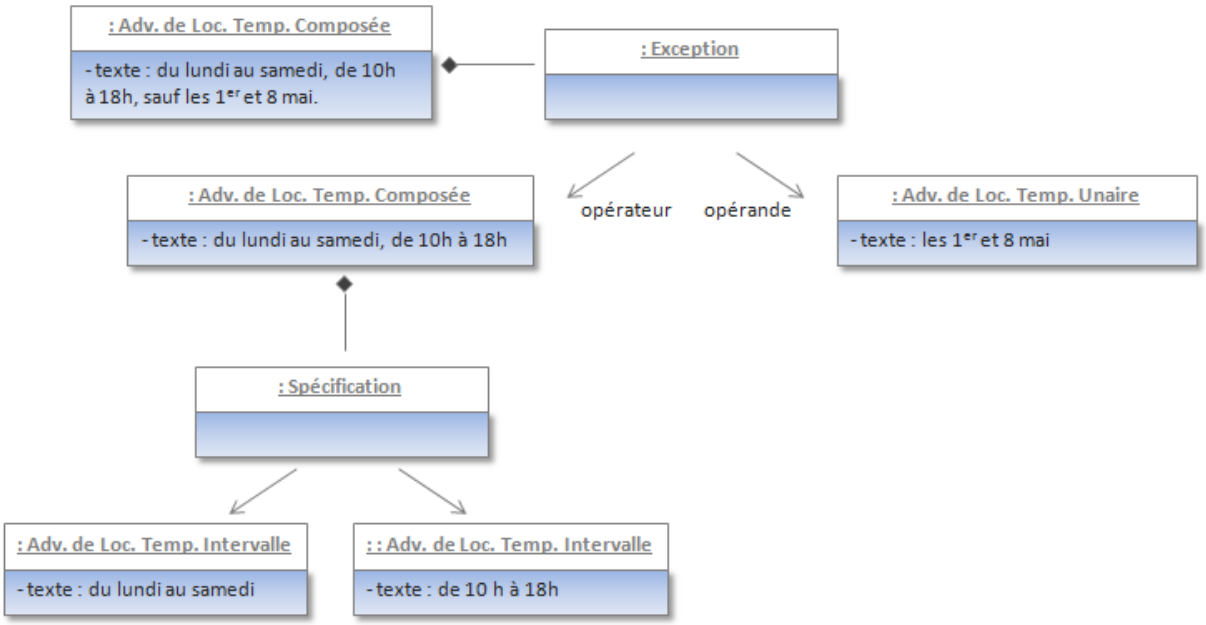


Fig. 40 : diagramme d'objet associé à l'adverbial du lundi au samedi, de 10h à 18h, sauf les 1er et 8 mai

D'autres exemples d'analyse sont présentés en annexe (cf. annexe 1 pour une analyse des adverbes *hier*, *ce jour-là*, *les deux dernières semaines de la campagne électorale*).

4.3 Bilan du chapitre

Pour représenter la sémantique des adverbiaux de localisation temporelle, conformément à la proposition de (Battistelli, 2009), notre proposition de modélisation distingue plusieurs opérateurs agissant sur un repère temporel noyau (opérateurs de Régionalisation, de Focalisation et de Déplacement). Ces opérateurs sont susceptibles de prendre différentes valeurs qui dépendent des instances lexicales constitutives des adverbiaux et qui traduisent la façon dont l'ancrage se définit, de proche en proche, à partir d'un ou plusieurs repères initiaux.

Nous avons généralisé et étendu cette représentation, ce qui nous a conduit à dégager la catégorie des adverbiaux de localisation temporelle. Il devient alors possible de décrire non seulement des adverbiaux calendaires, mais plus généralement tout type d'adverbial de localisation temporelle, indépendamment de la nature de repère temporel noyau. Nous avons également étendu cette proposition afin de décrire des adverbiaux binaires, ainsi que des adverbiaux formés par la composition de plusieurs adverbiaux.

Nous avons tenté de dégager une partie des différentes valeurs possibles des différents opérateurs à l'aide desquels on peut décrire les adverbiaux de localisation temporelle. Ce travail demanderait à

être poursuivi, afin de pouvoir représenter la façon dont l'ancrage est déterminé dans certains cas que nos propositions en l'état ne permettent pas de traiter de façon satisfaisante. En effet, ces valeurs pourraient être affinées pour pouvoir par exemple décrire des granularités temporelles indéterminées qui s'interprètent en fonction de leur contexte (*quelques temps*, *durant une longue période, peu avant*).

On a également cherché à décrire la façon dont un ancrage temporel est défini lorsque plusieurs adverbiaux contribuent conjointement à le déterminer : nous avons ainsi formulé quelques propositions pour formaliser les liens entre ces adverbiaux. Là encore, nos propositions doivent être affinées, car elles ne décrivent jusqu'ici que des associations entre des adverbiaux complets, or il semble qu'il peut être utile de décrire des liens entre un adverbial et une partie seulement d'un autre adverbial. Par exemple, dans une expression telle que « depuis la fin des Jeux Olympiques, en mars », l'adverbial « en mars » entretient un lien avec le syntagme nominal « la fin des Jeux Olympiques » plutôt qu'avec l'ensemble de l'adverbial (« depuis la fin des Jeux Olympiques »). Cette question renvoie au problème consistant à déterminer la portée des adverbiaux.

Enfin, la représentation formelle que l'on a spécifiée se veut indépendante des langues, même si pour valider pleinement cette hypothèse, il faudrait encore la tester plus avant : dans le cadre de ces travaux, elle a été testée sur le français et l'anglais, mais une étude sur les adverbiaux de localisation temporelle en allemand paraît conforter cette hypothèse (Moreau-Moquay, 2012).

Dans le chapitre suivant nous formulons des propositions pour associer aux opérations dégagées ici des transformations qui s'appliquent à des intervalles de temps, nommés *Intervalles Calendaires*. Ces transformations permettent d'associer un intervalle calendaire (éventuellement un ensemble d'intervalles calendaires) à un adverbial de localisation temporelle. On verra par la suite que ce processus de transduction permet d'obtenir une représentation sur laquelle il est possible de s'appuyer pour pouvoir ensuite proposer des applications dédiées à l'acquisition des connaissances et à la recherche d'information.

Chapitre 5 : Calculs sur les valeurs des adverbiaux de localisation temporelle

L'analyse des adverbiaux de localisation temporelle présentée dans le chapitre précédent permet de les représenter sous la forme d'une succession d'opérations sur un repère temporel noyau. Cette représentation, qui procède d'une démarche linguistique, vise à rendre compte de la façon dont ces unités textuelles déterminent un ancrage temporel, indépendamment de toute visée opératoire. Dans ce chapitre, on s'attache à articuler cette représentation avec une représentation présentant des propriétés calculatoires bien étudiées (celles des intervalles temporels), afin de pouvoir interagir avec des formats de représentation standard.

Après avoir exposé les motivations de notre démarche sur un plan applicatif (section 5.1), nous présentons dans la section 5.2 un mécanisme de transduction permettant d'associer un intervalle calendaire (ou un ensemble d'intervalles) à un adverbial de localisation temporelle. Nous montrerons ensuite comment ce mécanisme peut être exploité pour mettre en œuvre des systèmes de recherche d'information susceptibles de prendre en charge des requêtes contenant un critère calendaire (section 5.3). Nous présenterons dans ce cadre une heuristique permettant de comparer entre eux des adverbiaux de localisation temporelle, en leur associant une mesure de pertinence.

5.1 Motivations

Comme nous le verrons dans les chapitres suivants (6 et 7) où nous présenterons différentes applications implémentant les représentations formelles présentées dans les chapitres 4 et 5, le mécanisme permettant de passer d'une représentation sous la forme d'une succession d'opérations sur un repère temporel noyau vers des intervalles calendaires présente un intérêt à la fois pour l'ingénierie des connaissances, en particulier pour l'acquisition de connaissances à partir de textes et la recherche d'information.

Sur le plan de la gestion des connaissances, l'objectif est de pouvoir maintenir des bases de connaissances en sortant d'un modèle de représentation très contraint des informations temporelles, mais aussi de pouvoir mieux exploiter la connaissance portée par les textes, ce qui implique d'en avoir une représentation qui lui soit la plus fidèle possible. Il s'agit notamment de pouvoir associer des propriétés temporelles à différents objets d'une base de connaissances (pouvoir représenter par exemple qu'une œuvre a été peinte *aux alentours des années 1450*).Comme on l'a vu, dans les bases de données et les bases de connaissances, il n'est pas possible aujourd'hui de représenter et de manipuler des informations temporelles présentant un certain degré d'indétermination, ni non plus de jouer finement avec des niveaux de granularité différents. On montrera également que notre approche permet de traiter des informations temporelles complexes comme les dates et horaires d'ouverture, qui font intervenir des propriétés itératives (cf. section 6.2).

Pour la recherche d'information, la difficulté du traitement des informations temporelles tient dans ceci qu'il faut pouvoir en mesurer la pertinence, afin de filtrer et d'ordonner les informations qui répondent à une requête. Le problème consiste à trouver des critères à l'aide desquels on peut évaluer la pertinence d'adverbiaux aussi variés que « avant 2009 », « en mars 2009 », « de mi 2009 à fin 2011 », « aux alentours de 2009 », qui font intervenir des opérations de régionalisation, de focalisation et de déplacement.

L'heuristique que l'on présente dans la section 5.3 vise à déterminer des critères de filtrage et d'ordonnancement des adverbiaux de localisation temporelle, afin de pouvoir répondre à des requêtes contenant des critères calendaires : le type de système de recherche d'information que l'on souhaite mettre en œuvre vise en effet à traiter des requêtes telles que la laïcité en France avant 1905 ou la peinture italienne au début du XVe siècle, par exemple. Laissant de côté pour un temps le problème du traitement conjoint de critères thématiques (la laïcité, la peinture italienne) et de critères calendaires (avant 1905, au début du XVe siècle) (cf. section 7.1.5), on s'en tient d'abord à la résolution du problème qui consiste à répondre au critère calendaire (cf. section 5.3).

La démarche que l'on adopte cherche à maintenir et articuler deux représentations des adverbiaux de localisation temporelle, l'une fonctionnelle (une transposition formelle de la représentation sémantique décrite dans le chapitre 4), l'autre référentielle, que l'on va décrire ici. Cela nécessite de définir un processus de transduction de la première vers la seconde, dont on s'attache ici à décrire les étapes.

Il faut noter d'emblée que ce processus repose sur des informations de nature calendaire et que donc ne sont traités, dans la phase de transduction, que les adverbiaux présentant une base calendaire. Cela signifie que les bases des adverbiaux de localisation temporelle déictique ou relatif à un procès doivent avoir été transposées sous la forme d'un intervalle de dates, que l'on nomme désormais *Intervalle Calendaire*. C'est là une problématique de recherche à part entière que ce projet n'a pas abordé frontalement. Ces problèmes relèvent de la normalisation des adverbiaux déictiques (pouvoir dire par exemple que « lundi prochain » renvoie au lundi 16 avril 2012) et de l'assignation de coordonnées temporelles à des bases relatives à un procès ou à un chrononyme (il s'agit, par

exemple, pour un adverbial tel que « après le pléistocène », d'associer à la référence *pléistocène* un intervalle calendaire)[34].

On présente ici les résultats d'un atelier de travail dédié à la temporalité dont une partie importante a fait l'objet de publications collectives (Teissèdre *et al.*, 2011 ; Battistelli *et al.*, 2011b ; Battistelli *et al.*, 2012). Cette proposition de formalisation repose ainsi sur un travail collectif qui doit beaucoup aux propositions de Marcel Cori, Professeur à Paris-Ouest Nanterre La Défense. Cette formalisation a été enrichie ici pour pouvoir traiter également les adverbiaux calendaires itératifs (*tous les lundis, de 10h à 19h*, par exemple).

5.2 Représentation référentielle des adverbiaux calendaires

Nous présenterons ici une méthode permettant de transformer la *représentation fonctionnelle* des adverbiaux de localisation temporelle dont le noyau est formé par une base calendaire, vers une *représentation référentielle* sous la forme d'*Intervalles Calendaires* (notés *IC*). La représentation fonctionnelle des adverbiaux de localisation temporelle est une façon d'encoder le résultat de l'analyse sémantique présentée dans le chapitre précédent.

5.2.1 L'association d'un Intervalle Calendaire à un adverbial de localisation temporelle

On considère dans un premier temps les opérations appliquées à une *Base Calendaire Absolue*. On verra plus loin que les *Bases Calendaires Périodiques* appellent des transformations différentes qui génèrent des ensembles d'intervalles. On a vu que l'on pouvait représenter les adverbiaux de localisation temporelle comme la résultante de l'application successive des opérateurs de focalisation/déplacement et de régionalisation. Cette succession d'opérations peut être encodée sous la forme d'une *Expression Fonctionnelle*. A chacune de ces opérations, on fait correspondre des transformations appliquées à l'*Intervalle Calendaire* associé à la base calendaire de l'adverbial. Ce processus de transduction nous permet d'associer un *Intervalle Calendaire* à chaque expression fonctionnelle.

A ce stade de nos travaux, seule une sous-partie des valeurs associées aux opérations sémantiques à l'aide desquelles sont décrits les adverbiaux calendaires sont prises en compte. La démarche doit toutefois pouvoir être étendue pour les prendre progressivement toutes en compte. Sont ainsi prises en compte les valeurs suivantes :

- Opération de Focalisation = {Focalisation ID, Focalisation Début, Focalisation Fin, Focalisation Milieu}.
- Opération de Déplacement = {Déplacement (u, n) ; avec $n \in \mathbb{Z}$ et où u est une *Unité Calendaire Cardinale*}.
- Opération de Régionalisation = {Régionalisation ID, Régionalisation Avant, Régionalisation Après, Régionalisation Jusque, Régionalisation Depuis}

[34] On peut envisager, dans ce second cas de figure, que ce type d'information est fourni par une base de connaissances.

- Opération de Composition = {Composition Binaire, Composition Spécification, Composition Concaténation, Composition Exception}

Les valeurs *Focalisation ID* et *Régionalisation ID* permettent d'encoder les opérations de focalisation et de régionalisation qui laissent les intervalles calendaires inchangés : ces opérations ne donnent pas lieu à des transformations. On aura ces deux valeurs ID par exemple si l'on analyse une adverbial tel que « *en 2002* », où la régionalisation et la focalisation ne modifient pas l'intervalle calendaire associé à la base *2002*.

On peut associer une expression fonctionnelle à la représentation d'un adverbial de localisation temporelle sous la forme d'une succession d'opérations sur une base calendaire. On définit une *Expression Fonctionnelle* (EF) de la façon suivante :

(i) Si α est une *Base Calendaire Absolue* ou une EF et si $\in Opération\ de\ Focalisation \cup Opération\ de\ Déplacement \cup Opération\ de\ Régionalisation,$ alors $\Omega\ (\alpha)$ est une EF.

(ii) Si α et β sont deux EFs et si $\Omega \in Opération\ de\ Composition,$ alors $\Omega\ (\alpha, \beta)$ est une EF.

Les exemples 1 à 4 illustrent la façon dont on peut associer une *Expression Fonctionnelle* aux adverbiaux calendaires conformément à la représentation qui découle de notre proposition d'analyse sémantique :

Ex. 1 : aux débuts des années 30
EF1 : Régionalisation ID(
Focalisation Début(
Base Calendaire(
décennie : 1930)))

Ex. 2 : trois mois avant le début de l'année 1985
EF2 : Régionalisation ID(
Déplacement(mois,-3)(
Focalisation Début(
Base Calendaire(
année : 1985))))

Ex. 3 : jusque trois mois avant le début des années 30
EF3 : Régionalisation Jusque(
Déplacement(mois,-3)(
Focalisation Début(
Base Calendaire(
décennie : 1930))))

Ex. 4 : de fin 2007 à début mars 2009
EF4 : Composition Binaire(

Régionalisation Depuis(
Focalisation Fin(
Base Calendaire(
année : 2007))),
Régionalisation Jusque(
Focalisation Début(
Base Calendaire(
mois : 3, année : 2009))))

Dans un premier temps, on s'attache à décrire le processus de transduction pour les adverbiaux calendaires unaires (ex 1 à 3). On décrit ensuite ce processus pour les opérations de *Composition Binaire,* à savoir celles qui décrivent des adverbiaux binaires, comme dans l'exemple 4. Les transformations associées à ces opérations ne génèrent qu'un unique *Intervalle Calendaire*. On verra ensuite que les autres opérateurs de composition ainsi que les adverbiaux dont le noyau est formé par une *Base Calendaire Périodique* nécessitent de considérer des ensembles d'*Intervalles Calendaires*.

5.2.1.1 Unités Calendaires

Nous considérons un ensemble fini d'unités calendaires $U = \{u, v, w, \dots\}$. Par exemple : $\{millénaire, siècle, décennie, année, mois, jour, \dots\}$. Une date i représentée à l'unité u sera notée u_i. A chaque unité u est associée une séquence infinie de dates :

$$S(u) = \langle \dots, u_{-n}, \dots, u_{-1}, u_0, u_1, \dots, u_m \rangle$$

$S(u)$ décrit ainsi la succession des dates conformément à une unité donnée u. Par exemple, si u correspond à l'unité $mois$, $S(u)$ correspondra à la séquence suivante :

$$S(u) = \langle \dots, 2010/11, 2010/12, 2011/01, 2011/02, \dots \rangle$$

On considère une relation d'ordre entre les dates et entre les unités considérées : si une unité u est inférieure à v, on notera $u < v$, si une date i précède une date j, on notera $i < j$. Par exemple, on a $jour < année$.

On définit deux applications $début_{u \to v}$ et $fin_{u \to v}$, de telle sorte que pour une date i :

$$\forall i \; début_{u \to v}(i) < fin_{u \to v}(i)$$

Pour deux dates i et j, si $i < j$, on a :

$$fin_{u \to v}(i) < début_{u \to v}(j)$$

Si $début_{u \to v}(i) = j$ et $fin_{u \to v}(i) = k$, cela signifie que v_j est le début de u_i conformément à v et que v_k est la fin de u_i conformément à v. En particulier, pour chaque unité u et pour chaque date i, on a $début_{u \to u}(i) = i$ et $fin_{u \to u}(i) = i$.

Ainsi, pour une date i correspondant à l'année 1997, on a : $début_{année \to mois}(i) = j$, où j correspond à 1997/01 et $fin_{année \to mois}(i) = k$, où k correspond à 1997/12.

5.2.1.2 Les Intervalles Calendaires

Un $Intervalle\ Calendaire$ (ou IC) est défini par une paire ordonnée d'éléments pris dans une des séquences $S(u)$: $\langle u_i, u_j \rangle$ (où $i \leq j$). Une autre notation possible est : $\langle i, j, u \rangle$. u_i représente la date de début de l'intervalle IC, u_j représente la date de fin de l'intervalle IC et u correspond à l'unité considérée.

Les cas particuliers où $i = -\infty$ et $j = +\infty$ sont également pris en compte, afin de pouvoir transposer des adverbiaux calendaires comme « *depuis 1998* » ou « *jusque dans les années 20* ». Le cas d'un IC vide, noté $\emptyset$, est également pris en compte.

A chaque IC $\langle i, j, u \rangle$ d'unité u et pour chaque unité v inférieure à u, il est possible d'associer un IC qui est son *image* conformément à l'unité v :

$$t_{u \to v}(\langle i, j, u \rangle) = \langle début_{u \to v}(i), fin_{u \to v}(j), v \rangle$$

Ainsi, par exemple, l'image de l'IC $\langle 1995/03, 1996/05, mois \rangle$, conformément à l'unité $jour$ est l'IC $\langle 1995/03/01, 1996/05/31, jour \rangle$.

5.2.1.3 Propriétés des Intervalles Calendaires

Considérons deux IC A et B, dont les unités sont respectivement u et v. Soit w la plus petite des deux unités u et v, on a : $t_{u \to w}(A) = \langle i, j, w \rangle$, $t_{v \to w}(B) = \langle k, l, w \rangle$.

L'*intersection* de A et de B forme un IC qui se définit de la façon suivante : $A \cap B = \langle max\ (i, k), min(j, l), w \rangle$, sauf si $max\ (i, k) > min(j, l)$. Dans ce cas, l'intersection est vide : $A \cap B = \emptyset$.

On dira de A qu'il est *inclus* dans B (ou que B *contient* A) si et seulement si $i \geq k$ et $j \leq l$.

Par exemple, on dira par exemple de l'IC $A = \langle 1980, 1980, année \rangle$ qu'il contient l'IC $B = \langle 1987, 1987, année \rangle$. On dira également que A contient l'IC $C = \langle 1987/01, 1987/12, mois \rangle$, qui correspond à l'image de B à l'unité $mois$.

A est *égal* à B si et seulement si A est inclus dans B et B est inclus dans A.

La *longueur relative* de A et de B (avec $B \neq \emptyset$) correspond à la valeur suivante :

$$rl(A/B) = \frac{j - i + 1}{l - k + 1}$$

Par exemple, pour les $IC\ A = \langle 1980, 1989,\ année\rangle$ et $B = \langle 1980, 1984,\ année\rangle$, on a :

$$rl(A/B) = \frac{1989 - 1980 + 1}{1984 - 1980 + 1} = \frac{10}{5} = 2$$

Si $A = \emptyset, rl(A/B) = 0$ pour chaque $B \neq \emptyset$.

Si B est un intervalle infini et $A \neq \emptyset$, alors on aura $rl(A/B) = \varepsilon$, où ε est une valeur supérieure à 0, mais inférieure à l'ensemble des autres nombres positifs.

Si A et B sont des intervalles infinis :

- Si A est strictement inclus dans B, alors $rl(A/B) = 1 - \varepsilon$
- Si B est strictement inclus dans A, alors $rl(A/B) = 1 + \varepsilon$
- SI B est égal à A, alors $rl(A/B) = 1$

5.2.1.4 Opérations sur les Bases Calendaires Absolues

A chacune des opérations sémantiques formant des adverbiaux calendaires absolus, on fait correspondre une transformation qui permet de lui associer un *Intervalle Calendaire*.

(1) A chaque *Base Calendaire Absolue* d'unité u, on associe un $IC\ \langle i, j, u\rangle$ dont la date de début est égale à la date de fin. Par exemple :

Ex1 : Janvier 1985 : $\langle 1985/01, 1985/01, mois\rangle$
Ex 2 : 10 janvier 1985 : $\langle 1985/01/10, 1985/01/10, jour\rangle$
Ex 3 : Années 80 : $\langle 198\_, 198\_, décennie\rangle$

(2) Considérons une $EF\ \alpha$ à laquelle un $IC\ \langle i, j, u\rangle$ est associé.

(2.1) Si Ω est un *Opérateur de Focalisation*, on associe un IC à $\Omega\ (\alpha)$ pour chaque unité v strictement inférieure à u de la façon suivante :

On définit un coefficient τ compris entre 0 et 1/2. Ce coefficient influe sur les bornes de l'intervalle associé à un adverbial calendaire. Ainsi, par exemple, de sa valeur peut dépendre le fait que l'intervalle associé à l'adverbial « *le 18 juin 1997* » est ou non inclus dans l'intervalle associé à « *la fin du mois de juin 1997* ». La valeur de τ peut dépendre du type

d'expression considéré : *au début de, à l'aube, au tout début de*, etc. Pour les exemples suivants, la valeur de τ est fixée à 1/3.

Les transformations associées aux Opérations de Focalisation qui ont pour valeur début, milieu et fin, nécessitent de considérer une unité inférieure à celle de l'IC initial sur lequel s'applique l'opération. On peut donc représenter l'IC résultant de la transformation à différentes unités v, dès lors que celles-ci sont inférieures à l'unité u de l'IC sur lequel s'applique l'opération de focalisation.

Remarquons par ailleurs que les expressions modélisées par les Opérations de Focalisation Début, Milieu et Fin présentent une ambiguïté qui ne peut être levée éventuellement qu'en contexte. En effet, « *la fin du mois de juin* » peut désigner une plage de temps plus ou moins longue ou bien le moment précis où le mois prend fin. Il est des cas où cette ambiguïté est levée, en particulier dans des adverbiaux tels que « *deux jours avant la fin du mois de juin* », où il est clair que la fin ne peut pas désigner une plage de temps étendue. On verra que ce dernier cas de figure est pris en compte ici, mais qu'il implique d'associer d'autres transformation aux Opérations de Focalisation.

A l'exception de l'identité (*Focalisation ID*), qui ne produit pas de transformation, les opérateurs de focalisation produisent les transformations suivantes :

(2.1.1) *Focalisation Début*

A l'IC $\langle i, j, u\rangle$ associé à l'*Expression Fonctionnelle*, on fait correspondre l' IC suivant pour chaque unité v inférieure à u :

$$\langle début_{u \to v}(i), début_{u \to v}(i) + \lfloor \tau(fin_{u \to v}(j) - début_{u \to v}(i) + 1)\rfloor, v\rangle$$

Grâce à une fonction « plancher »[35], on obtient toujours des nombres entiers. Le résultat de la transformation sera donc différent selon l'unité prise en considération.

On a vu que pour appliquer la transformation associée à l'opération de focalisation, il est nécessaire de considérer une unité inférieure à celle de l'IC sur lequel s'applique l'opération.

Le résultat de l'Opération de Focalisation Début appliquée à l'IC $A = \langle 198\_, 198\_, décennie\rangle$ représentant les « années 80 » peut par exemple être représenté à l'unité $année$ ou à l'unité $mois$.

Par exemple, l'image de A conformément à l'unité $année$ est $\langle 1980, 1989, année\rangle$. Telle que définie, à l'unité $année$, la transformation associée à l'Opération de Focalisation Début produit le résultat suivant :

[35] $\lfloor x\rfloor$ correspond à la fonction « plancher » de x, autrement dit à la partie entière par défaut.

$$\langle 1980, 1980 + \lfloor \tau(1989 - 1980 + 1) \rfloor, année \rangle = \langle 1980, 1983, année \rangle$$

L'image de A conformément à l'unité $mois$ est $\langle 1980/01, 1989/12, mois \rangle$. Telle que définie, à l'unité $mois$, la transformation associée à l'Opération de Focalisation Début appliquée à A produit le résultat suivant :

$$\langle 1980/01, 1980/01 + \lfloor \tau(1989/12 - 1980/01 + 1) \rfloor, mois \rangle = \langle 1980/01, 1983/04, mois \rangle$$

(2.1.2) *Focalisation Fin*

A l'IC $\langle i, j, u \rangle$ associé à l'*Expression Fonctionnelle*, on fait correspondre l'IC suivant pour chaque unité v inférieure à u :

$$\langle fin_{u \to v}(j) - \lfloor \tau(fin_{u \to v}(j) - début_{u \to v}(i) + 1) \rfloor, fin_{u \to v}(j), v \rangle$$

Appliquée à l'IC $\langle 1997, 1997, année \rangle$, la transformation associée à la Focalisation Fin nécessite de considérer des unités inférieures à l'année. Par exemple, la transformation produit les IC suivants aux unités $mois$ et $jour$:

$\langle 1997/09, 1997/12, mois \rangle$
$\langle 1997/09/01, 1997/12/31, jour \rangle$

(2.1.3) *Focalisation Milieu*

A l'IC $\langle i, j, u \rangle$ associé à l'*Expression Fonctionnelle*, on fait correspondre l' IC suivant pour chaque unité v inférieure à u :

$\langle début_{u \to v}(i) + \lfloor \tau(fin_{u \to v}(j) - début_{u \to v}(i) + 1) \rfloor, fin_{u \to v}(j) -$
$\lfloor \tau(fin_{u \to v}(j) - début_{u \to v}(i) + 1) \rfloor, v \rangle$

Appliquée à l'IC $\langle 2011/06, 2011/06, mois \rangle$, la transformation associée à la Focalisation Milieu produit l'IC suivant à l'unité $jour$:

$\langle 2011/06/10, 2011/06/20, jour \rangle$

(2.2) Si Ω est un *Opérateur de Déplacement* où v est une unité de U inférieure ou égale à u, on associe un IC à Ω (α) de la façon suivante :

(2.2.1) *Déplacement Avant*

Soit le Déplacement Avant $(v, -n)$. A l'IC $\langle i, j, u \rangle$ associé à l'*Expression Fonctionnelle*, on fait correspondre l'IC suivant :

$\langle début_{u \to v}(i) - n, début_{u \to v}(i) - n, v \rangle$.

Prenons l'exemple du Déplacement Avant présent dans la représentation formelle que l'on associe à l'adverbial *« trois mois avant les années 80 »*. L'image de l'$IC\ A = \langle 198\_,198\_, décennie\rangle$ – qui correspond à la base calendaire *« années 80 »* -, conformément à l'unité $mois$ est $\langle 1980/01, 1989/12, mois\rangle$. Appliquée à A, la transformation associée au Déplacement Avant $(mois, -3)$ produit l'IC suivant :

$$\langle 1980/01 - 3\ mois, 1980/01 - 3\ mois, mois\rangle = \langle 1979/10, 1979/10, mois\rangle$$

Comme on l'a évoqué plus haut, les Opérations de Focalisation Début, Milieu et Fin présentent une ambiguïté. Par exemple, l'adverbial *« fin 1992 »* peut désigner ou bien une plage de temps plus ou moins étendue, située dans la dernière partie de l'année 1992, ou bien le moment précis où elle s'achève. La conjonction d'une opération de *Déplacement* avec une opération de *Focalisation* constitue un cas particulier où cette ambiguïté est levée et où la Focalisation désigne un « moment » précis. Dans ces cas, aux Opérations de Focalisation, on fait correspondre les IC suivants :

(1) Focalisation Début :

$$\langle début_{u \to v}(i), début_{u \to v}(i), v\rangle$$

Par exemple, appliquée à l'$IC = \langle 1972{,}1972, année\rangle$, à l'unité $mois$, la transformation associée à la Focalisation Début produit l'$IC = \langle 1972/01{,}1972/01, mois\rangle$.

(2) Focalisation Fin :

$$\langle fin_{u \to v}(i), fin_{u \to v}(i), v\rangle$$

Par exemple, appliquée à l'$IC = \langle 1972/08{,}1972/08, mois\rangle$, à l'unité $jour$ la transformation associée à la Focalisation Fin produit l'$IC = \langle 1972/08/31{,}1972/08/31, jour\rangle$.

(3) Focalisation Milieu :

$$\langle début_{u \to v}(i) + \lfloor \tau(fin_{u \to v}(j) - début_{u \to v}(i) + 1)/2 \rfloor, début_{u \to v}(i) + \lfloor \tau(fin_{u \to v}(j) - début_{u \to v}(i) + 1)/2 \rfloor, v\rangle$$

Par exemple, appliquée à l'$IC = \langle 1981/09{,}1981/09, mois\rangle$, à l'unité $jour$ la transformation associée à la Focalisation Milieu produit l'$IC = \langle 1981/09/15{,}1981/09/15, mois\rangle$.

Prenons l'exemple du Déplacement Avant présent dans notre représentation de l'adverbial *« trois mois avant la fin des années 80 »*. L'image de l'$IC\ A = \langle 198\_,198\_, décennie\rangle$ – qui correspond à la base calendaire *« années 80 »* -, conformément à l'unité $mois$ est $\langle 1980/01, 1989/12, mois\rangle$. Dans ce cas, comme il y a conjonction d'un Déplacement avec une Focalisation, on applique la transformation particulière définie plus haut pour la Focalisation Fin. Appliquée à A, la transformation associée à la Focalisation Fin produit l' l'$IC\ B = \langle 1989/12,1989/12, mois\rangle$. Appliquée à B, la transformation associée au Déplacement Avant $(mois, -3)$ produit l'IC suivant :

$$\langle 1989/09, 1989/09, mois\rangle.$$

(2.2.2) *Déplacement Après*

Soit le Déplacement Après $(v, +n)$. A l'IC $\langle i, j, u\rangle$ associé à l'*Expression Fonctionnelle*, on fait correspondre l' IC suivant :

$$\langle fin_{u\to v}(j) + n, fin_{u\to v}(j) + n, v\rangle.$$

Prenons l'exemple du Déplacement Après présent dans notre représentation de l'adverbial *« trois mois après l'année 1804 »*. L'image de l'$IC\ A = \langle 1804,1804, année\rangle$ – qui correspond à la base calendaire *« année 1804 »* - , conformément à l'unité $mois$ est $\langle 1804/01, 1804/12, mois\rangle$. Appliquée à A, la transformation associée au Déplacement Après $(mois, +3)$ produit l'IC suivant :

$$\langle 1804/12 + 3, 1804/12 + 3, mois\rangle = \langle 1805/03, 1805/03, mois\rangle.$$

(2.3) Si Ω est un *Opérateur de Régionalisation*, on associe un IC à $\Omega\ (\alpha)$ de la façon suivante :

(2.3.1) *Régionalisation Avant*

A $\langle i, j, u\rangle$, on associe l'intervalle $\langle -\infty, i - 1, u\rangle$.

A l'adverbial « *avant mai 68* », on associe ainsi, à l'unité *mois,* l'IC suivant :

$$\langle -\infty, 1968/04, mois\rangle.$$

(2.3.2) *Régionalisation Après*

A $\langle i, j, u\rangle$, on associe l'intervalle $\langle j + 1, +\infty, u\rangle$.

A l'adverbial « *après les années 20* », on associe ainsi, à l'unité *année*, l'IC suivant :

$$\langle 1930, +\infty, année\rangle.$$

(2.3.3) *Régionalisation Jusqu'à*

A $\langle i, j, u \rangle$, on associe l'intervalle $\langle -\infty, j, u \rangle$.

En appliquant successivement les opérations de *Focalisation*, de *Déplacement* et de *Régionalisation*, on obtient ainsi, pour l'adverbial « *jusqu'à trois mois avant le début des années 80* », à l'unité *mois*, l'IC suivant :

$$\langle -\infty, 1979/10, mois \rangle.$$

(2.3.4) *Régionalisation Depuis*

A $\langle i, j, u \rangle$, on associe l'intervalle $\langle i, +\infty, u \rangle$.

A l'adverbial « *depuis le 16 mars 2011* », on associe ainsi, à l'unité *jour*, l'IC suivant :

$$\langle 2011/03/16, +\infty, jour \rangle.$$

Rappelons à nouveau que les adverbiaux de localisation temporelle et les intervalles calendaires sont deux modes d'indexation temporelle différents. En ce sens, le processus de transduction de l'un vers l'autre n'est qu'une heuristique. En effet, le processus de transduction implique aussi une surdétermination dans la représentation résultante : par exemple, la transduction d'un adverbial tel que « *à partir de 22h* » sous la forme d'un intervalle dont la borne de fin est infinie est une transposition infidèle, car, dans l'adverbial, cette borne n'est pas déterminée. On verra plus loin que les pôles associés aux intervalles apportent une réponse partielle à cette difficulté, sans la résoudre complètement (cf. section 5.3.5). Une voie possible pour prendre en compte cette incertitude quant à la détermination d'une borne pourrait consister à introduire deux autres bornes, c'est-à-dire à considérer deux intervalles, conformément à la proposition de (Berberich *et al.*, 2010) (cf. section 3.3.2).

Par ailleurs, cette transposition laisse dans l'ombre les liens complexes que peuvent entretenir les adverbiaux et les procès qu'ils contribuent à déterminer. Par exemple, dans l'énoncé « *M. Blair a donné* ***jusqu'au 30 juin*** *aux responsables catholiques et protestants pour débloquer ce dossier* », l'intervalle associé à l'adverbial « jusqu'au 30 juin » devrait avoir pour début l'intervalle associé au procès « a donné ». Sur les rapports complexes entre les intervalles associés au procès et ceux associés aux adverbiaux, on renvoie aux travaux de L. Gosselin (1996, 2005a et 2005b) mentionnés dans la section 2.3).

5.2.1.5 Le cas des Adverbiaux de localisation Binaires

Les transformations associées aux *adverbiaux de localisation binaires*, lorsqu'ils sont de type c*alendaires absolus*, génèrent un unique *Intervalle Calendaire*.

Considérons deux *Expressions Fonctionnelles* α et β auxquelles deux *Intervalles Calendaires* sont associés : $\langle i_1, j_1, u\rangle$ et $\langle i_2, j_2, v\rangle$. Soit w la plus grande unité inférieure à u et v. A l'opération Composition Binaire(α, β), on associe l'IC suivant :

$$\langle début_{u \to w}(i_1), fin_{v \to w}(j_2), w\rangle$$

Considérons par exemple l'adverbial « *de la fin de l'année 2007 au début du mois de mars 2009* ».

En appliquant la transformation associée à la Focalisation Fin sur la base calendaire « année 2007 », on obtient pour l'unité $mois$, l'$IC\ A = \langle 2007/09, 2007/12, mois\rangle$. A l'unité $jour$, l'image de A est $\langle 2007/09/01, 2007/12/31, jour\rangle$.

En appliquant la transformation associée à la Focalisation Début sur la base calendaire « mars 2009 », on obtient, pour l'unité $jour$, l'$IC\ B = \langle 2009/03/01, 2009/03/10, jour\rangle$.

En appliquant la transformation associée à l'Opération de Composition Binaire, on obtient ainsi l'IC suivant :

$$\langle 2007/09/01, 2009/03/10, jour\rangle.$$

5.2.2 Les ensembles d'Intervalles Calendaires

$\mathrm{S(u)}$ décrit la succession des dates conformément à l'unité u. Un ensemble d' IC est défini par un ensemble de paires ordonnées d'éléments pris dans une des séquences de $\mathrm{S(u)}$. Un ensemble peut contenir une infinité de paires ordonnées.

A chaque ensemble d'IC $\{\langle \mathrm{i_0, j_0, u}\rangle, \dots, \langle \mathrm{i_n, j_n, u}\rangle\}$ d'unité u, il est possible d'associer un ensemble d'IC équivalent (son image) pour chaque unité v inférieure à u :

$$\mathrm{f_{u \to v}}(\{\langle \mathrm{i_0, j_0, u}\rangle, \dots, \langle \mathrm{i_n, j_n, u}\rangle\}) =$$
$$\{\langle \mathrm{début_{u \to v}(i_0), fin_{u \to v}(j_0), v}\rangle, \dots, \langle \mathrm{début_{u \to v}(i_0), fin_{u \to v}(j_0), v}\rangle\}$$

Ainsi, par exemple, l'image de l'ensemble d'IC $\{\langle 1995, 1996, année\rangle, \langle 1998, 2003, année\rangle\}$ conformément à l'unité mois est l'ensemble d'IC suivant :

$$\{\langle 1995/01, 1996/12, mois\rangle, \langle 1998/01, 2003/12, mois\rangle\}.$$

Il est désormais possible de définir le complémentaire d'un IC $\langle \mathrm{i, j, u}\rangle$. Le complémentaire d'IC forme l'ensemble d'IC :

$$\{\langle -\infty, \mathrm{i} - 1, \mathrm{u}\rangle, \langle \mathrm{j} + 1, +\infty, \mathrm{u}\rangle\}.$$

Par exemple, le complémentaire de l'IC $\langle 1995, 1996, année\rangle$ est l'ensemble :

$$\{\langle -\infty,\ 1994, année\rangle, \langle 1997, +\infty, année\rangle\}.$$

5.2.2.2 Propriétés des ensembles d'Intervalles Calendaires

L'union de A et de B, notée $A \cup B$, forme un ensemble d'IC qui résulte de l'union de chacun des IC de A avec chacun des IC de B, puis de chacun des IC obtenus et ce récursivement jusqu'à obtenir un ensemble d'éléments mutuellement disjoints (dont l'intersection est vide).

On dira de A qu'il est inclus dans B, si et seulement si chacun des IC de A est inclus dans chacun des IC de B.

A est égal à B si et seulement si A est inclus dans B et B est inclus dans A.

L'intersection de A et de B, notée $A \cap B$, est l'ensemble qui résulte de l'intersection de chacun des IC de A avec chacun des IC de B.

Le complémentaire de A, noté $\overline{A}$, est l'ensemble formé par l'intersection des complémentaires de chacun des IC constitutifs de A.

5.2.2.3 Opération sur les Bases Calendaires Périodiques

A chacune des opérations sémantiques à l'aide desquelles on représente un adverbial de localisation unaire de *Base Calendaire Périodique*, décrit sous la forme d'une *Expression Fonctionnelle*, on fait correspondre une transformation qui permet de lui associer un ensemble d'*Intervalles Calendaires*.

(1) A chaque *Base Calendaire Périodique* d'unité u, on associe un ensemble infini d'IC $\{\dots, \langle i_{-n}, j_{-n}, u\rangle, \dots, \langle i_0, j_0, u\rangle, \dots, \langle i_n, j_n, u\rangle, \dots\}$ dont la date de début est égale à la date de fin.

Par exemple :

Ex1 : Janvier : $\{\dots, \langle 1985/01, 1985/01, mois\rangle, \langle 1986/01, 1986/01, mois\rangle, \dots\}$
Ex 2 : 10 janvier : $\{\dots, \langle 1985/01/10, 1985/01/10, jour\rangle, \langle 1986/01/10, 1986/01/10, jour\rangle, \dots\}$
Ex 3 : chaque année : $\{\dots, \langle 1985, 1985, année\rangle, \langle 1986, 1986, année\rangle, \dots\}$

(2) Les transformations associées aux opérations de focalisation, de déplacement et de régionalisation s'appliquent successivement à chacun des IC de l'ensemble d'IC associé à une *Base Calendaire Périodique*[36].

[36] Il semble toutefois que seule la Régionalisation ID s'applique aux adverbiaux constitués d'une *Base Calendaires Périodiques*. Dans une expression telle que « avant le début de chaque mois », la préposition *avant* semble marquer un déplacement plutôt qu'une régionalisation.

5.2.2.4 Opérations de Composition

Les transformations associées aux *Opérations de Composition* décrites dans cette section génèrent un ensemble d'*Intervalles Calendaires*.

(1) Composition Binaire

Considérons deux *Expressions Fonctionnelles* α et β auxquelles deux ensembles d'*Intervalles Calendaires* sont associés : $\{\langle i_1, j_1, u\rangle, \ldots, \langle i_n, j_n, u\rangle\}$ et $\{\langle k_1, l_1, u\rangle, \ldots, \langle k_n, l_n, v\rangle\}$, respectivement d'unité u et v. Ces ensembles doivent être de même taille (les ensembles infinis étant admis). Soit w la plus grande unité inférieure à u et v. A l'opération Composition Binaire(α, β), on associe l'ensemble d'IC suivant :

$$\{\langle début_{u \to w}(i_1), fin_{v \to w}(l_1), w\rangle, \ldots, \langle début_{u \to w}(i_n), fin_{v \to w}(l_n), w\rangle\}$$

Par exemple, on représentera l'adverbial « *de lundi à vendredi* » par l'ensemble :

$$\{\ldots, \langle 2012/04/09, 2012/04/13, jour\rangle, \langle 2012/04/16, 2012/04/20, jour\rangle, \ldots\}.$$

(2) Composition Spécification

Considérons deux *Expressions Fonctionnelles* α et β auxquelles deux ensembles d'*Intervalles Calendaires* sont associés : A d'unité u et B d'unité v. Soit w la plus grande unité inférieure à u et v. Si les applications $t_{u \to w}$ et $t_{v \to w}$ produisent respectivement les ensembles C et D, à l'opération Composition Spécification(α, β), on associe l'ensemble d'IC $C \cap D$.

Par exemple, on représentera l'adverbial « *tous les vendredis, de 15h à 17h* » par l'ensemble :

$$\{\ldots, \langle 2012/04/13{:}\,15, 2012/04/13{:}\,17, jour\rangle, \langle 2012/04/20{:}\,15, 2012/04/20{:}\,17, jour\rangle, \ldots\}.$$

(3) Composition Concaténation

Considérons deux *Expressions Fonctionnelles* α et β auxquelles deux ensembles d'*Intervalles Calendaires* sont associés : A d'unité u et B d'unité v. Soit w la plus grande unité inférieure à u et v. Si les applications $t_{u \to w}$ et $t_{v \to w}$ produisent respectivement les ensembles C et D, à l'opération Composition Concaténation(α, β), on associe l'ensemble d'IC $C \cup D$.

Par exemple, on représentera l'adverbial « *tous les vendredis ainsi que les 1er mai* » par l'ensemble :

$$\{\ldots,\langle 2012/04/28, 2012/04/28, jour\rangle,\langle 2012/05/01, 2012/05/01, jour\rangle,$$
$$\langle 2012/05/05, 2012/05/05, jour\rangle,\ldots\}.$$

(4) Composition Exception

Considérons deux *Expressions Fonctionnelles* α et β auxquelles deux ensembles d'*Intervalles Calendaires* sont associés : A d'unité u et B d'unité v. Soit w la plus grande unité inférieure à u et v. Si les applications $t_{u\to w}$ et $t_{v\to w}$ produisent respectivement les ensembles C et D, à l'opération Composition Exception(α, β), on associe l'ensemble d'$IC\ C \cap \overline{D}$.

Par exemple, on représentera l'adverbial « *tous les jours sauf les 1er mai* » par l'ensemble :

$$\{\ldots,\langle 2012/04/30, 2012/04/30, jour\rangle,\langle 2012/05/02, 2012/05/02, jour\rangle,$$
$$\langle 2012/05/03, 2012/05/03, jour\rangle,\ldots\}.$$

Dans cette section, nous avons décrit un processus de transduction qui permet de passer d'une représentation fonctionnelle vers une représentation référentielle sous forme d'*Intervalles Calendaires*. Le mécanisme de cette transduction étant posé, on peut présenter désormais une heuristique qui s'appuie sur le format des intervalles calendaires pour sélectionner et trier par pertinence des adverbiaux calendaires. Cette heuristique est utilisée dans le système de recherche d'information présenté dans le chapitre 7.

5.3 Application dans le cadre de la recherche d'information selon des critères calendaires

Les travaux présentés ici cherchent ainsi à répondre à la problématique de la recherche d'information selon des critères calendaires. Considérons par exemple un corpus relatif à l'histoire des Etats-Unis. Un utilisateur pourrait s'intéresser par exemple à des informations relatives à « *la prohibition au début des années 30* ». Il pourrait y avoir plusieurs réponses plus ou moins pertinentes à cette requête dans les documents du corpus. Une des réponses les plus pertinentes pourrait par exemple être « *En 1931, peu avant la fin de la Prohibition, Madden a quitté le milieu de la contrebande* ». Pour autant, une autre réponse telle que « *Vers la fin des années 20, l'agent Eliot Ness du Bureau en charge de la Prohibition ouvre une enquête sur Capone et ses activités* » peut aussi présenter un intérêt pour l'utilisateur. Dans cette dernière réponse, la référence temporelle désignée par l'adverbial n'entre pourtant pas dans la fenêtre de temps définie par le critère calendaire de la requête « *au début des années 30* ». Elle en est toutefois très proche. C'est là tout l'intérêt de l'approche que l'on présente, qui ne cherche pas uniquement à filtrer les adverbiaux qui réfèrent à des périodes incluses dans la fenêtre de temps définie dans une requête : on cherche ici à définir des critères permettant d'évaluer la *proximité* entre différents adverbiaux de localisation temporelle. L'inclusion n'est qu'un de ces critères, important certes, mais pas le seul.

Dans un premier temps, laissant de côté la problématique du traitement et de l'analyse des mots-clés tels que « *prohibition* », ainsi que de ce qu'il convient d'appeler une *réponse* à la requête (doit-il s'agir d'une phrase, d'un document, d'une portion d'un document ?), on s'intéresse ici au problème

qui consiste à relier les informations calendaires exprimées dans la requête avec les adverbiaux calendaires que l'on est susceptible de trouver dans un corpus. On décrit ainsi une heuristique permettant de sélectionner et d'ordonner par pertinence un ensemble d'adverbiaux calendaires.

5.3.1 Un algorithme pour mesurer la pertinence des adverbiaux calendaires

Le modèle d'ordonnancement que l'on présente est une mesure de similarité qui s'appuie sur un processus d'analyse des adverbiaux calendaires présents dans un corpus de textes. Ne sont traités à ce stade que les adverbiaux dont la base est une base calendaire absolue.

La chaîne de traitements mise en œuvre, d'un point de vue formel, est la suivante : il s'agit d'analyser les adverbiaux calendaires d'un corpus pour les décrire sous la forme d'une représentation fonctionnelle, qui est alors traduite sous la forme d'une représentation référentielle, selon la procédure décrite dans la section 5.2.1. La même chaîne de traitements permet d'analyser les requêtes des utilisateurs pour en extraire le critère calendaire. Le processus de filtrage et d'ordonnancement des adverbiaux calendaires s'appuie alors sur la représentation référentielle pour établir des mesures de similarité.

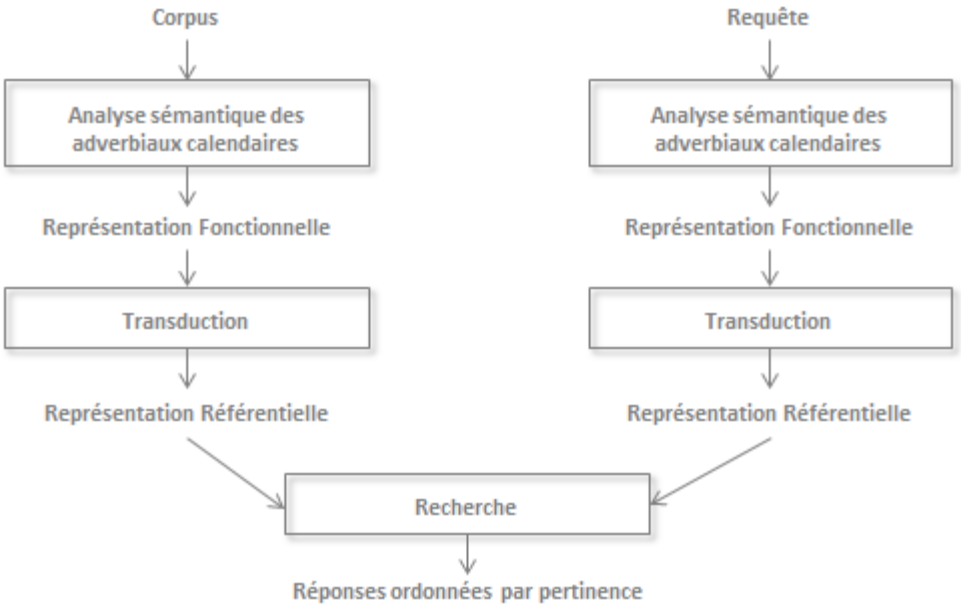


Fig. 41 : le processus formel de recherche

5.3.2 La problématique

Considérons un ensemble d'adverbiaux calendaires traduits sous la forme d'intervalles calendaires, conformément à la transformation présentée dans la section 5.2.1 :

$$A = \{A_1, A_2, \dots, A_n\}$$

Le critère calendaire exprimé dans une requête est également traduit sous la forme d'un Intervalle Calendaire, nommé Q. L'objectif de l'algorithme est d'extraire un sous-ensemble $A(Q) =$

$\{A_{i_1}, A_{i_2}, \ldots, A_{i_p}\}$ dans l'ensemble A et de l'ordonner du plus pertinent au moins pertinent. Pour évaluer la pertinence des adverbiaux calendaires, on considère en premier lieu un critère d'adéquation, et, si nécessaire, en cas d'égalité au niveau du critère d'adéquation, un critère d'ordre.

5.3.3 Les critères d'adéquation

On établit trois types de critères d'adéquation, du meilleur au moins bon (cf. fig. 42) : (1) l'*Egalité*, (2) l'*Inclusion* (de A_i dans Q) et (3) l'*Inclusion Inverse* (inclusion de Q dans A_i) ou l'*Intersection*. Dans le cas d'une *Intersection Vide*, on dira ainsi qu'aucun critère d'adéquation n'est pas satisfait. Ces critères peuvent être décrits sous la forme de relations d'Allen (Allen, 1983).

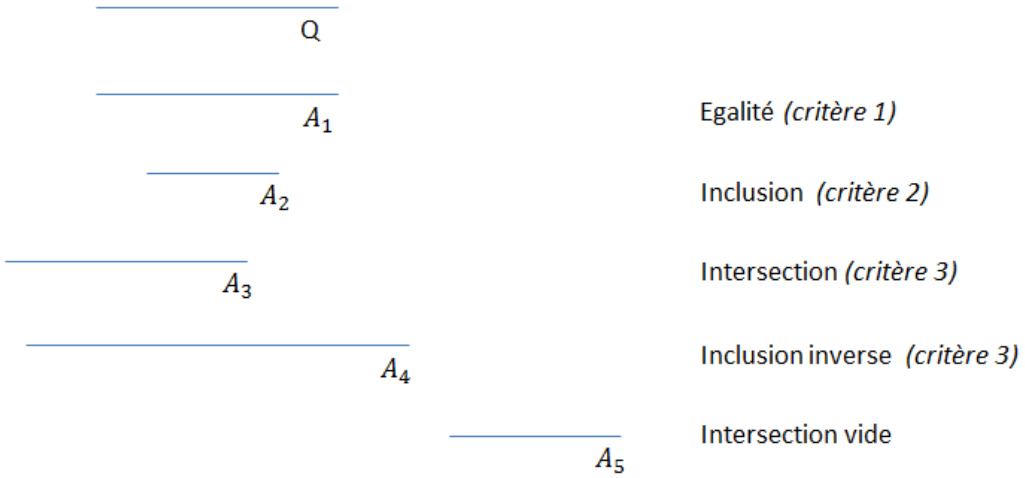


Fig. 42 : illustration des critères d'adéquation entre une requête Q et différentes réponses possibles

(1) Egalité

Si un élément A_i est égal à Q, il répond à la requête de la meilleure façon possible.

(2) *Inclusion*

Si un élément A_i est inclus dans Q, il répond également à la requête. En termes de relation d'Allen, il s'agit des relations suivantes : A_i est inclus dans Q (*during*) ; A_i est inclus dans Q et leurs bornes de début sont concomitantes (*starts*) ; A_i est inclus dans Q et leurs bornes de fin sont concomitantes (*ends*).

> **Exemple** :
> Si Q représente l'adverbial « *en 1980* » et A_1, l'adverbial « *de mars à mai 1980* », A_1 est inclus dans Q (on peut également se reporter aux ex. A'_1 et A''_1 du tableau 1 de la section suivante).

(3) *Inclusion Inverse* et *Intersection* :

Si Q est inclus dans un élément A_i, c'est-à-dire si Q est inclus dans A_i (*during*), ou si Q est inclus dans A_i et leurs bornes de début sont concomitantes (*starts*), ou encore si Q est inclus dans A_i et leurs bornes de fin sont concomitantes (*ends*), on dit alors que A_i contient Q. On parle alors d'*Inclusion Inverse*.

Exemple :
si A_1 représente l'adverbial « *de mars à mai 1980* » et A_2, l'adverbial « *de 1978 à 1982* », A_1 est alors une meilleure réponse que A_2 pour la requête Q « *en 1980* », parce que A_1 est inclus dans Q, alors que A_2 inclut Q (on peut également se reporter aux ex. A'_2 et A''_2 du tableau 1).

Si A_i et Q se chevauchent partiellement (*overlaps*), on parle d'intersection.

Exemple :
Si A_3 représente l'adverbial « *de novembre 1979 à mai 1980* » et Q l'adverbial « *en 1980* », alors il y a intersection entre A_3 et Q (on peut également se reporter aux ex. A'_3 et A''_3 du tableau 1).

L'*Intersection* n'est pas considérée comme étant un meilleur ou un moins bon critère d'adéquation que l'*Inclusion Inverse*.

Dans tous les autres cas de figure, il n'y a pas de recouvrement entre Q et A_i : leur intersection est vide ; A_i ne satisfait donc aucun des critères d'adéquation. On verra, néanmoins, que certains éléments de A qui ne satisfont aucun des critères d'adéquation peuvent être conservés dans l'étape de filtrage (cf. section 5.3.5), en prenant en compte une mesure de distance entre les intervalles considérés. Cette mesure permet ainsi d'apprécier la proximité entre les intervalles Q et A_i en cas d'intersection vide (ils peuvent être adjacents ou plus ou moins distants)

5.3.4 L'attribution d'un score d'adéquation

L'attribution d'un score d'adéquation permet d'ordonner une liste de réponses par pertinence, lorsque la requête forme un intervalle fini et lorsque les réponses satisfont les critères d'*Egalité*, d'*Inclusion*, d'*Inclusion Inverse* ou d'*intersection*. De façon générale, plus l'intersection entre l'IC associé à une réponse A_i et l'IC associé à une requête Q est grande, meilleure sera la réponse. Cette intersection est notée $(A_i \cap Q)$.

Pour mesurer la pertinence d'une réponse, on considère deux critères : le critère du *rappel*, qui correspond à la part de A_i incluse dans Q, par rapport à la zone de temps couverte par Q, et le critère de la *précision*, qui correspond à la part de A_i incluse dans Q, par rapport à la zone de temps couverte par A_i. On définit donc deux quantités[37] :

$$rappel(A_i/Q) = rl((A_i \cap Q)/Q)$$

[37] Ces deux mesures entrent également dans la mesure de similarité proposée par (Le Parc Lacayrelle *et al.*, 2007) (cf. section 3.3.2).

$$précision(A_i/Q) = rl((A_i \cap Q)/A_i)$$

Ces deux mesures, bien que sous d'autres dénominations, entrent également dans la mesure de similarité proposée par (Le Parc Lacayrelle *et al.*, 2007) (cf. section 3.3.2).

On rappelle que la mesure $rl(A/B)$ correspond à la longueur relative d'un *Intervalle Calendaire* A par rapport à un *Intervalle Calendaire* B. On renvoie à la section 5.2.1.3 pour le détail du calcul de cette mesure.

Pour toute réponse satisfaisant le critère d'*Egalité*, la *précision* et le *rappel* équivalent à 1. Pour toute réponse satisfaisant le critère d'*Inclusion* (A_i inclus dans Q), la *précision* équivaut à 1. C'est le cas, par exemple, des réponses A_1 *« de mars à mai 1980 »* et A'_1 *« de février à novembre 1980 »* pour la requête *« en 1980 »* (cf. tableau 1). Pour toute réponse satisfaisant le critère d'*Inclusion Inverse* (A_i incluant Q), le *rappel* équivaut à 1. C'est le cas, par exemple, des réponses A'_2 « *d'octobre 1979 à mars 1981* » et A_2 « *de 1978 à 1982* » pour la requête *« en 1980 »* (cf. tableau 1).

Les mesures de *précision* et de *rappel* ne sont toutefois pas d'égale importance. En effet, une réponse satisfaisant le critère d'*Intersection* (cf. par exemple la réponse A_3 *« de novembre 1979 à mai 1980 »*) peut présenter un bon score de *précision*, mais un score nécessairement plus faible de *rappel* qu'une réponse satisfaisant le critère d'*Inclusion Inverse* (cf. par exemple la réponse A_2 *« de 1978 à 1982 »*). Dès lors, on introduit un coefficient, α, inférieur à 1, pour minimiser l'importance du *rappel* par rapport à la *précision*. On attribue ainsi un *score* (A_i/Q) pour une réponse A_i relativement à une requête Q de la façon suivante :

$$score\ (A_i/Q) = \frac{précision(A_i/Q) + \ \alpha\ rappel(A_i/Q)}{1 + \ \alpha}$$

Nos expérimentations nous ont conduits à fixer α à 0,4, mais ce score peut être ajusté. Compte-tenu du mode de calcul du score d'adéquation, il est possible qu'une réponse satisfaisant le critère d'*Intersection* ou d'*Inclusion Inverse* obtienne un meilleur score qu'une réponse satisfaisant le critère d'*Inclusion*. Ainsi A'_2 (« *d'octobre 1979 à mai 1981* ») obtient un meilleur score que A''_1 (« *le 25 mai 1980* ») pour la requête « *en 1980* ». Dans le cas où aucun critère d'adéquation n'est satisfait, le $score\ (A_i/Q)$ est égal à 0.

Le tableau 1 présente les scores attribués à différents adverbiaux susceptibles de répondre à la requête « *en 1980* ». On peut y voir notamment que les intervalles calendaires infinis sont également des réponses possibles et que les scores qu'ils sont susceptibles d'obtenir peuvent être comparés à ceux des intervalles calendaires finis (cf. ex A'_3 et A''_3).

	Réponses à la requête *en 1980*	**IC correspondant, à l'unité jour**	**Score**
$\boldsymbol{A_0}$	*en 1980*	<1980/01/01, 1980/12/31, jour>	1
$\boldsymbol{A'_1}$	*de février à novembre 1980*	<1980/02/01, 1980/11/30, jour>	0,952
$\boldsymbol{A_1}$	*de mars à mai 1980*	<1980/03/01, 1980/05/31, jour>	0,785

A'_2	*d'octobre 1979 à mars 1981*	<1979/10/01, 1981/03/31, jour>	0,762
A''_1	*le 25 mai 1980*	<1980/05/25, 1980/05/25, jour>	0,715
A_3	*de novembre 1979 à mai 1980*	<1979/10/01, 1980/05/31, jour>	0,629
A_2	*de 1978 à 1982*	<1978/01/01,1982/01/01, jour>	0,428
A''_3	*depuis janvier 1980*	<1980/01/01, $+\infty$, jour>	0,285
A'_3	*depuis mai 1980*	<1980/05/01, $+\infty$, jour>	0,190
A''_2	*de juillet 1980 à juin 2010*	<1980/07/01, 2010/06/30, jour>	0,154

Tableau 1 : Scores des réponses pour la requête en 1980

5.3.5 Ordonner les réponses

Le score d'adéquation permet d'ordonner des réponses dont l'intersection avec l'intervalle formé par la requête est non vide. Nous cherchons désormais à ordonner les réponses pour lesquelles cette intersection est vide (par exemple, une réponse telle que *« en 1979 »* pour la requête *« en 1980 »*). Ces réponses sont considérées comme moins bonnes que les précédentes ; elles peuvent néanmoins être ordonnées entre elles. Nous cherchons également à ordonner des réponses pour une requête formant un intervalle infini (par exemple, une requête telle que *« depuis 1980 »*). Pour cela, on introduit une mesure de *distance,* qui dépend des *pôles* des intervalles considérés et dont on détaille ici la méthode de calcul.

Pour chaque intervalle calendaire, un *pôle* est déterminé. Pour $\langle i, +\infty, u\rangle$, le pôle est i, pour $\langle -\infty, j, u\rangle$, le pôle est j. Aussi pour un adverbial tel que « *depuis les années 80* », le pôle correspond au début de l'intervalle calendaire associé à la base calendaire *années 80* ; pour un adverbial tel que « *jusqu'en mars 2007* », le pôle correspond à la fin de l'intervalle calendaire associé à la base calendaire *mars 2007*.

Si $\langle i, j, u\rangle$ est obtenu à la suite de la transformation associée à la *Focalisation Début*, alors le pôle est i : ainsi, pour une expression telle que « *au début de l'année 2011* », le pôle correspond ainsi *au 1er janvier 2011,* si l'intervalle calendaire associé à l'adverbial est exprimé à l'unité *jour*. S'il est obtenu à la suite d'une fonction *Focalisation Fin*, alors le pôle est j ; dans les autres cas, le pôle est $\lfloor (i + j)/2 \rfloor$. Ainsi, pour l'adverbial « *dans les années 60* », le pôle correspondra au milieu des années 60, soit le *15 juin 1965* si l'intervalle calendaire associé à l'adverbial est exprimé à l'unité *jour*.

La *distance* entre deux intervalles calendaires A et B de même unité u est définie comme la valeur absolue de la différence entre deux pôles : $|p\hat{o}le(A) - p\hat{o}le(B)|$.

Si deux réponses ont le même score, elles sont ordonnées d'après leur distance par rapport à la requête. Cette mesure permet par ailleurs d'ordonner, parmi les résultats présentés, des réponses dont le score d'adéquation est nul.

On a également recours à la mesure de distance lorsqu'il s'agit d'ordonner des réponses par rapport à une requête dont l'intervalle calendaire est infini. En effet, si Q forme un intervalle infini, on ne considère pas la valeur du score d'adéquation, mais seulement la mesure de *précision* définie plus

haut. Enfin, si deux réponses ont la même valeur de *précision*, celle dont la distance est la plus faible est favorisée. Ce comportement est illustré dans le tableau 2 ci-dessous. Les réponses sont présentées de façon ordonnée, de la meilleure à la moins bonne. ε correspond au symbole introduit dans la section 5.2.1.3 – il représente une valeur supérieure à 0, mais inférieure à tous les autres nombres positifs.

Réponses à la req. *depuis 1980*	**IC correspondant, à l'unité année**	**Précision**	**Distance**
depuis 1980	<1980, +∞, année>	1	0 an
en 1982	<1982, 1982, année>	1	2 ans
depuis 1983	<1983, +∞, année>	1	3 ans
de 1983 à 1986	<1983, 1986, année>	1	4 ans
depuis 1978	<1978, +∞, année>	$1-\varepsilon$	2 ans
depuis 1975	<1975, +∞, année>	$1-\varepsilon$	5 ans
de 1979 à 1981	<1979, 1981, année>	0,666	0 an
jusqu'en 1984	<−∞, 1984, année>	ε	4 ans
jusqu'en 1975	<−∞, 1975, année>	0	5 ans

Tableau 2 : Exemple d'ordonnancement d'une liste de réponses pour la requête depuis 1980

Ce modèle d'ordonnancement permet ainsi de trier un ensemble d'adverbiaux calendaires pour une requête donnée en comparant les intervalles calendaires que le processus de transduction décrit dans la section 5.2 permet de leur associer.

5.4 Bilan du chapitre

Nous avons présenté dans ce chapitre un processus formel permettant d'associer un intervalle calendaire ou un ensemble d'intervalles calendaires à un adverbial de localisation temporelle de base calendaire. Cette transposition est une étape dans le processus permettant de mesurer la pertinence des adverbiaux calendaires pour la recherche d'information.

Nous avons en ce sens également présenté une heuristique pour filtrer et ordonner par pertinence un ensemble d'adverbiaux calendaires par rapport à une requête exprimée sous la forme d'un critère calendaire.

Cette heuristique, qui consiste à comparer deux à deux des *Intervalles Calendaires*, ne peut traiter que des requêtes désignant une zone unique sur le calendrier : elle ne prend donc pas en charge des requêtes qui désigneraient plusieurs zones du calendrier. Par exemple, le système ne sait pas traiter une requête telle que *« tous les lundis »*.

On présentera dans le chapitre 7 l'implémentation de cet algorithme sous la forme d'un composant qui prend part à un système de recherche d'information expérimental. Ce système peut prendre en charge des requêtes combinant des critères thématiques (des mots-clés) et des critères calendaires.

L'heuristique de mesure de la pertinence a fait l'objet d'une évaluation présentée dans la section 8.3.2 du chapitre 8.

Les chapitres 6 et 7 présentent différents composants permettant d'opérationnaliser le traitement des adverbiaux de localisation temporelle, à la fois pour les annoter, les transformer et les comparer entre eux. Ces composants sont exploités dans les prototypes d'applications que l'on présente pour l'acquisition de connaissances et la recherche d'information.

Chapitre 6 : Ressources pour l'acquisition de connaissances relatives à la localisation temporelle

Calculables, mais moins expressifs que les représentations en langue de la localisation temporelle, les intervalles calendaires sont communément utilisés dans les systèmes de gestion des connaissances. Nous nous efforcerons de montrer à travers différentes expérimentations, en quoi il peut être utile de faire coexister deux modes de représentation formelle des adverbiaux de localisation temporelle, l'une qui découle d'une analyse linguistique (présentée dans le chapitre 4) et qui représente les adverbiaux sous la forme d'une succession d'opérations sur une référence temporelle noyau et l'autre qui les représente sous la forme d'intervalles calendaires (présentée dans le chapitre 5).

Nous décrivons ainsi dans ce chapitre un ensemble de ressources logicielles comprenant :

- un système d'annotation des adverbiaux de localisation temporelle qui les repère dans les textes et les encode sous la forme d'une succession d'opérations sémantiques (cf. section 6.1.4),
- des librairies Java permettant de transposer cette représentation formelle pour la traduire sous la forme d'intervalles calendaires (cf. section 6.1.5).

Nous montrerons comment cet ensemble de ressources peut être utilisé pour répondre à un cas d'application industriel. Nous décrirons ainsi un Web Service développé afin de faciliter la saisie d'informations temporelles complexes : les dates et horaires d'accessibilité d'un site ou d'un service (cf. section 6.2).

6.1 Un système d'annotation et de transduction des adverbiaux calendaires

6.1.1 Quel objectif ?

Comme l'ensemble des ressources développées, le système présenté ici a surtout valeur de démonstration : l'objectif est en effet de montrer l'intérêt de la démarche théorique qui consiste à représenter les adverbiaux de localisation temporelle sous la forme d'une succession d'opérations agissant sur un repère temporel noyau plutôt que d'abord sous la forme d'une valeur calendaire. Il s'agit également de montrer la faisabilité des applications qui procèderaient d'une telle démarche. Le parti-pris a été de n'utiliser, pour l'implémentation des démonstrateurs que l'on présente, que des solutions *open-source* ou sous licence libre, même si des solutions propriétaires étaient disponibles[38].

Le système d'annotation présenté dans cette section n'a pas vocation à prendre en charge l'ensemble des tâches généralement dévolues aujourd'hui aux systèmes d'annotation d'informations temporelles : il n'annote pas les « événements », par exemple, et ne s'attache pas à normaliser systématiquement les « expressions temporelles », pour reprendre les termes communément employés en ingénierie des langues. On a vu en effet que les systèmes d'annotation des expressions temporelles cherchent plus fréquemment à leur associer une valeur calendaire qu'à les décrire en s'appuyant sur une analyse linguistique (cf. section 3.1.1.2). Notre démarche distingue volontairement la tâche d'annotation des adverbiaux de localisation temporelle de celle qui consiste à leur associer une valeur calendaire.

6.1.2 Architecture du système

Le système d'annotation est composé de trois modules : (1) une implémentation du modèle objet des adverbiaux de localisation temporelle, (2) un module qui annote une sous-partie des adverbiaux de localisation temporelle dans les textes et (3) un module qui transpose les adverbiaux calendaires vers des intervalles calendaires.

[38] La plateforme XIP développée par Xerox, par exemple, eut été une solution possible pour développer nos ressources dédiées à l'annotation des adverbiaux de localisation temporelle. Cette plateforme, qui permet de coupler différents traitements d'analyse des textes (analyse morphologique, analyse syntaxique et analyse sémantique) a du reste déjà été utilisée pour l'annotation d'informations temporelles (Hagège et Tannier, 2008).

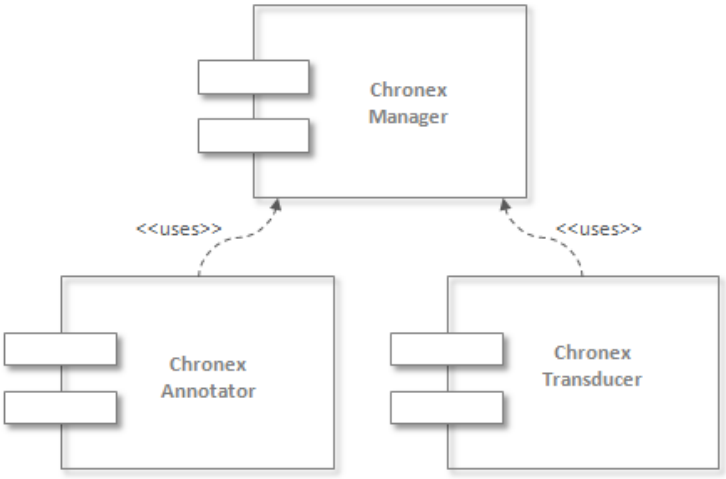


Fig. 43 : architecture du système d'annotation et de transduction des adverbiaux de localisation temporelle

Développé intégralement en Java, le système prend en entrée des textes encodés dans différents formats et produit en sortie des textes enrichis par des annotations représentées sous la forme de balises XML : c'est le rôle du module *Chronex Annotator*. Il offre également la possibilité de manipuler ensuite ces annotations sous la forme d'instances d'objets qui peuvent être de deux types, selon les besoins et les contextes d'utilisation : ou bien des objets décrivant des adverbiaux de localisation temporelle sous la forme d'une succession d'opérations sur une référence temporelle noyau (*Chronex Manager*) ou bien des intervalles calendaires (générés par le module *Chronex Transducer*).

6.1.3 Implémentation du modèle objet des adverbiaux de localisation temporelle

Le module central est un modèle objet pour manipuler les adverbiaux de localisation temporelle sous la forme d'objets les représentant sous la forme d'une succession d'opérations conformément à l'analyse présentée dans le chapitre 4 (cf. section 4.2). Ce module fournit plusieurs méthodes pour accéder aux valeurs associées aux différentes opérations sémantiques dégagées pour décrire ces adverbiaux.

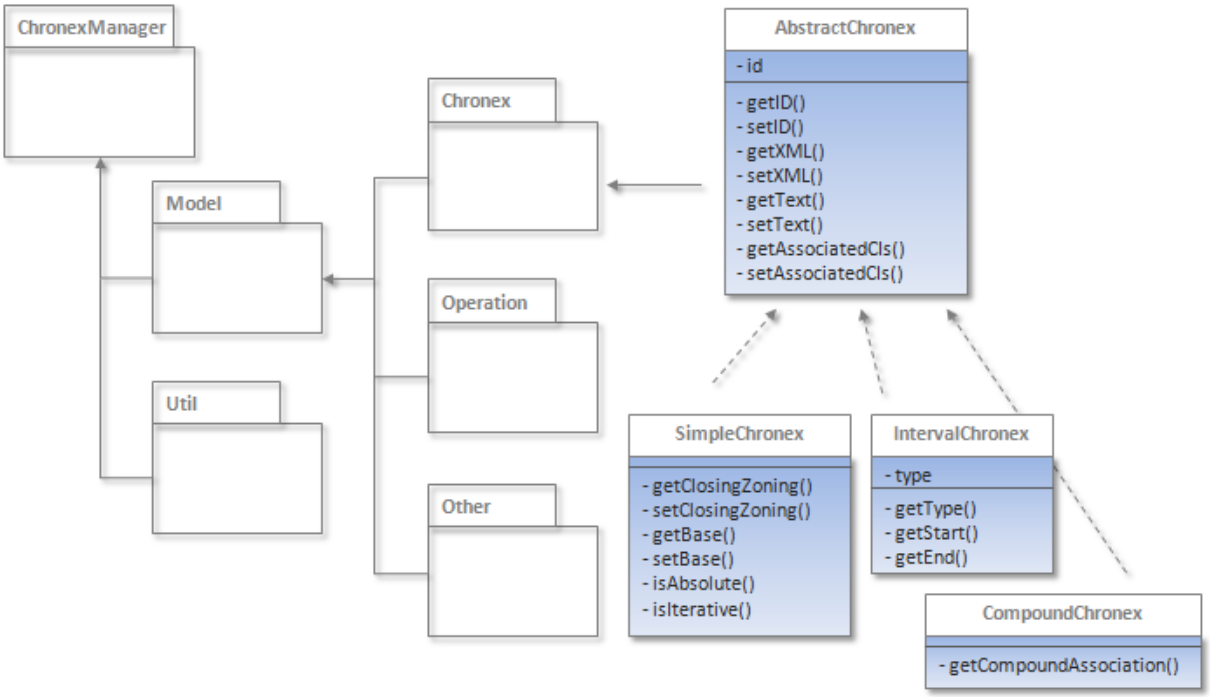


Fig. 44 : implémentation du modèle objet des adverbiaux de localisation temporelle (vue simplifiée)

La figure 44 présente l'architecture du module implémentant le modèle objet associé à la description des adverbiaux de localisation temporelle. Conformément à la représentation formelle présentée dans le chapitre 4, les adverbiaux peuvent être de trois types : unaires (*SimpleChronex*), binaires (*IntervalChronex*) ou composés (*CompoundChronex*). Le module permet d'explorer la succession des opérations qui opèrent au sein des instances d'objet décrivant les adverbiaux de localisation temporelle, en instanciant les différentes opérations à l'aide desquelles sont décrits les adverbiaux.

La nature de l'ancrage des adverbiaux de localisation temporelle – ancrage unique ou itératif (*isAbsolute(), isIterative()*) - détermine le type de traitements qu'on leur associe durant le processus de transduction.

La package *Util* contient une classe *Serializer* permettant d'exporter les instances d'objets représentant les adverbiaux de localisation temporelle dans différents formats :

- *toRDF()* : la méthode permet d'exporter au format RDF les instances d'un objet décrivant un adverbial. Cette méthode prend son sens dans l'architecture plus générale où l'ensemble des composants s'insère. Elle permet en particulier d'exporter les instances vers une base de connaissances. Ce format de stockage est présenté dans la section 6.2.2.2.
- *toTXT()* : cette méthode génère une expression linguistique correspondant à une instance du modèle. Cette méthode peut être utile dans le cas où l'instance manipulée ne provient pas d'une annotation : on verra section 6.2 qu'il est en effet aussi possible d'instancier le modèle à partir d'un agenda numérique éditable. Dans ce cas, il peut être utile pour les éditeurs de la base de connaissances d'avoir une représentation en langage naturel correspondant à l'information saisie.

6.1.4 Le module d'annotation

6.1.4.1 L'architecture du module d'annotation

Le module d'annotation permet d'ajouter dans un texte des informations décrivant les segments textuels identifiés comme étant des adverbiaux de localisation temporelle. Les annotations produites, des balises XML, décrivent ces adverbiaux sous la forme d'une succession d'opérations sur un repère temporel noyau, conformément à la démarche d'analyse présentée dans le chapitre 4.

L'annotation s'appuie sur un ensemble de lexiques et de patrons (ou grammaires locales), implémentés manuellement sous la forme de transducteurs Unitex. Unitex est une plateforme logicielle open source dédiée à l'analyse textuelle automatique (Paumier, 2002). L'analyse textuelle s'effectue à l'aide de ressources de différents types : des dictionnaires électroniques, des grammaires permettant de représenter des phénomènes linguistiques en s'appuyant sur un formalisme proche de celui des automates à états finis et des tables de lexique-grammaire. L'outil dispose d'une interface graphique permettant à un linguiste de concevoir des grammaires d'annotation. L'interface de la plateforme permet de produire simplement ces règles (appelées aussi patrons d'annotation), même si leur maintenance devient difficile à mesure qu'elles s'accumulent[39]. Nous verrons en effet qu'une des difficultés que nous avons rencontrée a été de maintenir ces règles à mesure qu'évoluait notre proposition de représentation formelle des adverbiaux : chaque modification de cette représentation obligeait à modifier les règles en les parcourant une à une, un travail qui devient fastidieux lorsque leur nombre et leur complexité augmente.

Le schéma d'annotation effectivement utilisé pour décrire les adverbiaux de localisation temporelle est une DTD héritée d'un projet dans lequel s'inscrivait le module d'annotation à l'origine : ce schéma opérationnel respecte le format attendu dans la plateforme Navitext (Couto et Minel, 2006a et 2006b) : initialement, le module d'annotation a en effet été conçu pour alimenter ce système de navigation textuelle (Teissèdre *et al.*, 2010), avant que nos expérimentations nous conduisent à concevoir et développer un système de recherche d'information permettant de répondre à des requêtes exprimant des critères calendaires (cf. chapitre 7). Le schéma d'annotation est donc différent du schéma *ChronolocationML* qui a été formalisé a posteriori, afin de prendre en compte les évolutions apportées au fil des expérimentations à la représentation formelle des adverbiaux de localisation temporelle présentée dans le chapitre 4. Ce dernier schéma d'annotation, décrit dans la section suivante (6.1.4.2), vise ainsi à refléter l'ensemble du modèle formel dans sa version finale. Pour se conformer à ce dernier schéma, les règles d'annotation devront être adaptées par la suite.

[39] Il faut en outre mentionner une limite de la plateforme, qui ne permet pas directement de réexploiter et manipuler les annotations produites en sortie. De ce point de vue, elle est moins performante que d'autres plateformes de traitement automatique des langues telles que Gate (Cunningham *et al.*, 2002), qui elle permet de combiner des traitements Java avec le processus d'annotation. Le choix d'Unitex s'explique ici en partie par ceci que l'on a réexploité en premier lieu des ressources déjà existantes issues du projet ANR Blanc CONIQUE. D'autres choix auraient été possibles, aussi bien au niveau des logiciels de développement des ressources (à l'aide de la plateforme Gate, par exemple) qu'au niveau du type même des ressources développées. Il aurait par exemple été possible de recourir à des méthodes d'apprentissage automatique, bien qu'elles n'offrent pas de contrôle fin et complet sur les règles d'annotation générées : elles fonctionnent en effet comme des « boîtes noires ».

Le module d'annotation est alimenté en entrée par des textes, qui peuvent être de différents formats (HTML, XML ou du texte brut). En sortie, ces textes sont enrichis de balises XML décrivant les adverbiaux de localisation temporelle repérés dans les textes sous la forme d'une succession d'opérations sur un repère temporel noyau.

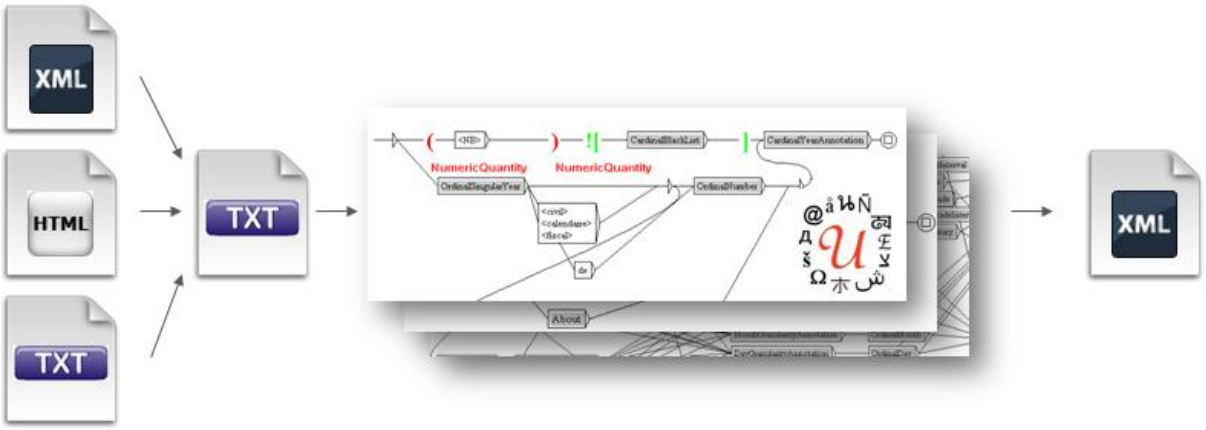


Fig. 45 : le module d'annotation

Le processus d'annotation repose sur une analyse dite de *surface*. Il ne s'appuie pas sur une analyse morphosyntaxique profonde, qui, en général, au moins dans les solutions libres existantes, impacte de façon significative les temps de traitements. Cette approche légère permet d'annoter des corpus volumineux (cf. section 7.2.1). Son défaut en revanche, est que parfois, faute d'informations suffisantes, les annotations produites sont incomplètes, au sens où elles ne permettent pas d'instancier toutes les valeurs nécessaires à la description d'un adverbial de localisation temporelle conformément à notre proposition d'analyse. Nous verrons dans le chapitre dédié à l'évaluation de ces ressources que cette approche peut parfois générer également des erreurs dans la description des adverbiaux. Le plus souvent, il faudrait, pour compléter les annotations manquantes, disposer d'informations relatives aux temps des verbes. On verra plus loin que ces informations doivent pouvoir être apportées par d'autres systèmes d'annotation (on renvoie sur ce point à la discussion sur la complémentarité avec TimeML, cf. section 6.1.4.3).

Le module d'annotation est composé de plusieurs *packages*, dont on présente les principales classes (cf. fig. 46).

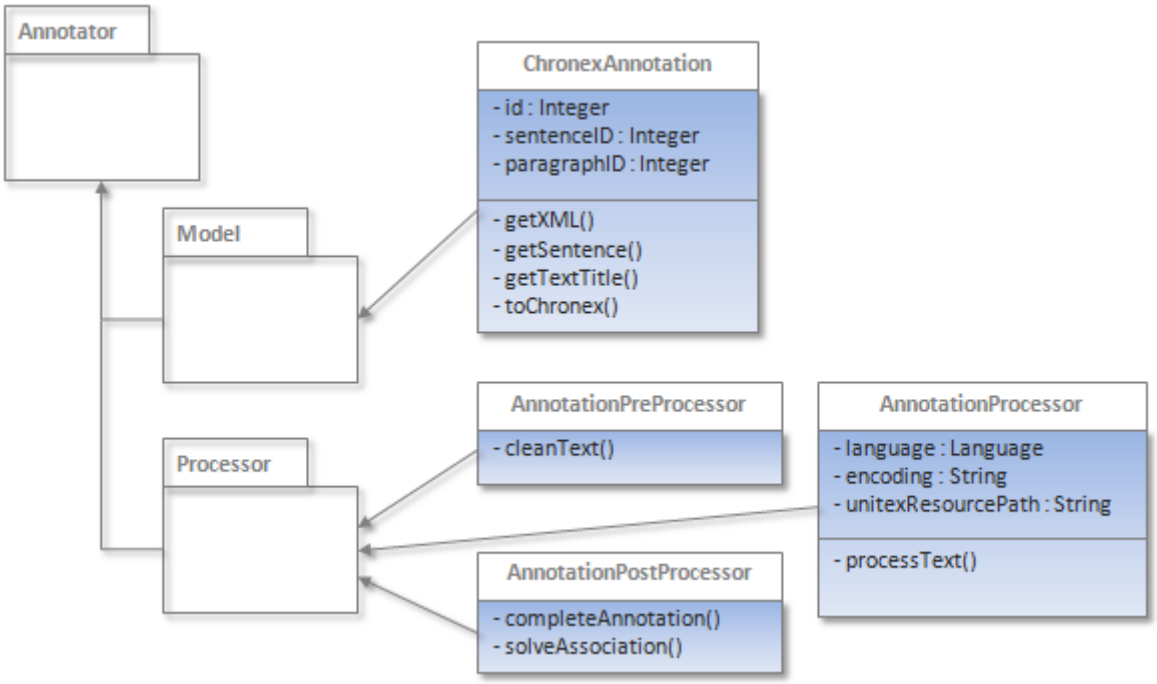


Fig. 46 : architecture du module d'annotation (vue simplifiée)

La classe *AnnotationPreProcessor* permet d'effectuer des prétraitements sur les textes en entrée, en particulier le nettoyage des balises XML ou HTML.

La classe *AnnotationProcessor* exécute les traitements pour l'annotation des textes. Elle pilote notamment le lancement d'une série d'instructions automatisées en ligne de commande (*batch*), qui appellent des programmes de la libraire Unitex pour :

- le prétraitement des textes, en particulier la normalisation des formes contractées ;
- la segmentation des textes en phrases ;
- l'annotation des textes à l'aide de dictionnaires de langue ;
- le lancement des lexiques et patrons d'extraction des adverbiaux de localisation temporelle ;
- le formatage des sorties

La classe *AnnotationProcessor* produit une première analyse des adverbiaux de localisation temporelle, qui ne permet pas toujours d'en instancier complètement le modèle objet. La classe *AnnotationPostProcessor* effectue ainsi plusieurs post-traitements sur les textes annotés, afin de compléter certaines annotations. Ces traitements s'exécutent à l'appel des méthodes suivantes :

5.1 *completeAnnotation()* : cette méthode effectue différents traitements : en particulier l'ajout d'identifiants uniques aux annotations produites et l'ajout de nouvelles annotations pour expliciter l'analyse des unités factorisées par la coordination (*entre janvier et mars 2011*). Pour que l'analyse d'un adverbial tel que « *les 7 et 8 janvier* » soit complète, elle doit produire la même analyse que pour l'adverbial « *les 7 janvier et 8 janvier* ». Les post-traitements ont ainsi à charge de reporter les informations factorisées dans la coordination.

- *solveAssociation()* : cette méthode permet de déterminer la nature des relations qui lient des adverbiaux unaires au sein des adverbiaux composés. Les traitements raisonnent à partir d'annotations contigües (en prenant notamment en compte la nature des unités calendaires impliquées), pour déterminer la nature de leur relation, s'il y a lieu.

Le second post-traitement vise à dégager, à l'aide d'un ensemble de règles, les opérations de composition entre adverbiaux. Pour l'expression « *ouvert tous les jours sauf le lundi et le 1er mai, de 9h à 19h* », il faut ainsi que le système analyse la relation entre l'adverbial de localisation binaire qui définit les horaires (*de 9h à 19h*) comme une relation de spécification, qui n'est pas liée à l'exception (*sauf le lundi et le 1er mai*), mais à l'adverbial composé (*tous les jours sauf le lundi et le 1er mai*).

Les règles permettant de déterminer les relations entre adverbiaux dégagent d'abord les associations de type *Concaténation*, puis de type *Exception*. Pour ce faire, l'outil compare la granularité et la nature des adverbiaux calendaires et s'appuie sur les indices lexicaux et typographiques dénotant ces associations (« et », « sauf », « hormis », etc.). En présence d'un indice d'une association de Concaténation (« et » ou une virgule) et tant qu'ils sont de même granularité (celle du jour dans « *sauf le lundi et le 1er mai* »), l'outil reconnaît que l'association est une concaténation (entre « *le lundi* » et « *le 1er mai* »). La présence d'un indicateur d'une association de type *Exception* permet ensuite d'associer l'adverbial « *le lundi et le 1er mai* » avec l'adverbial « *tous les jours* ». A chaque fois qu'une association entre deux adverbiaux est détectée, ces adverbiaux sont alors considérés comme formant un adverbial composé. Ces règles sont très dépendantes des indices lexicaux et typographiques : on verra dans le chapitre dédié à l'évaluation des ressources d'annotation qu'elles peuvent échouer si ces indices ne sont pas présents (cf. section 8.2.2).

Une règle permet ensuite de déterminer les associations de type *Spécification*. Si la granularité de deux adverbiaux diffère (*tous les jours sauf le lundi et le 1er mai* vs. *de 9h à 19h*), l'outil interprète ce changement de granularité comme une marque de la présence d'une association de type *Spécification*. Les différentes règles s'appuient ainsi toujours sur la granularité pour définir de proche en proche les relations de concaténation et d'exception (qui associent des adverbiaux de même granularité) et les relations de spécification (qui associent des adverbiaux de granularités différentes).

Le module d'annotation contient également un modèle objet pour manipuler les annotations produites (*ChronexAnnotation*). Il permet d'extraire les annotations au format XML et d'instancier le modèle objet des adverbiaux de localisation temporelle. Il permet également de récupérer les éléments de contexte dans lequel un adverbial est présent (la phrase, le paragraphe et le titre du document).

6.1.4.2 Le schéma d'annotation ChronolocationML

Défini a posteriori, à l'issu de nos travaux sur la description et l'annotation des adverbiaux, le langage *ChronolocationML* définit un jeu de balises XML pour annoter les adverbiaux de localisation

temporelle dans les textes. Ce schéma ne correspond donc pas au schéma opérationnel utilisé par le composant d'annotation : il reflète plus complètement le modèle formel de description des adverbiaux de localisation temporelle présenté dans le chapitre 4, bien que les contraintes exprimées soient moins fortes et qu'un certain nombre de simplifications aient été introduites pour en faciliter l'utilisation. Le schéma se veut ainsi permissif : il n'oblige pas à produire une description complète des adverbiaux de localisation temporelle.

Conformément à notre démarche, l'objectif de ce schéma est de faire émerger les opérations sémantiques à l'aide desquelles on décrit les adverbiaux de localisation temporelle. C'est pourquoi le standard *TimeML* ne pouvait pas être directement être exploité dans nos travaux, même si, comme on va le voir, *ChronolocationML* peut se présenter comme un prolongement de ce dernier. Les unités discursives que l'on souhaite annoter sont à la fois plus larges (des circonstanciels et des adverbiaux et non pas des dates) et ne recouvrent pas toujours les distinctions établies dans *TimeML*. En effet, la séparation franche établie par *TimeML* entre expressions temporelles et événements ne permet pas de ranger l'ensemble des adverbiaux de localisation temporelle dans la première de ces deux classes : on l'a vu, un événement au sens de *TimeML* peut former le noyau d'un adverbial de localisation temporelle (par exemple : « *dès le lendemain du match* »).

Conformément à la représentation des adverbiaux (annotés par la balise CHRONEX), le schéma permet de décrire les opérations qui entrent dans leur composition : les opérations de régionalisation (ZONING), de focalisation (ZOOMING) et de déplacement (SHIFTING) qui agissent sur une base (BASE).

On détaille les valeurs possibles des attributs de chaque élément de la DTD (notation BNF). Les liens entre les opérateurs et leurs opérandes, pour les opérations de régionalisation (ZONING), de focalisation (ZOOMING) et de déplacement (SHIFTING), sont marqués par l'attribut *operand* qui pointe sur leurs identifiants.

CHRONEX :

attributes ::= cid
cid ::= ChronexID
ChronexID ::= c<integer>

ZONING :

attributes ::= zonid type
zonid ::= ZoningID
ZoningID ::= zon<integer>
type ::= (before | after | until | since | around | ID)
operand ::= CDATA
{ operand ::= (zon<integer> | zm<integer> | s<integer> | b<integer>)}

ZOOMING :

attributes ::= zmid [subdivision] [quantity] [rank] [comparisonOperator]
zmid ::= ZoomingID
ZoomingID ::= zm<integer>
subdivision ::= (quarter | third | half | beginning | mid | end)

rank ::= CDATA
{ rank ::= (integer | last but one | last)}
quantity ::= CDATA
{quantity ::= integer }
comparisonOperator ::= CDATA
{ comparisonOperator ::= (less_than | more_than | approximatively) }
operand ::= CDATA
{ operand ::= (zm<integer> | s<integer> | b<integer>)}

SHIFTING :
attributes ::= sid [orientation] [quantity] [rank] [comparisonOperator]
sid ::= ShiftingID
ShiftingID ::= s<integer>
orientation ::= CDATA
{ orientation ::= (before | after) }
rank ::= CDATA
{ rank ::= (integer | last but one | last)}
quantity ::= CDATA
{quantity ::= integer }
comparisonOperator ::= CDATA
{ comparisonOperator ::= (less_than | more_than | approximatively) }
operand ::= CDATA
{ operand ::= (zm<integer> | s<integer> | b<integer>)}

BASE :
attributes ::= bid type [isIterative]
bid ::= BaseID
BaseID ::= b<integer>
isIterative ::= <boolean>
type ::= CDATA
{ type ::= (calendar | deictic | anaphoric | relative | unknown)}

CALENDAR_UNIT :
attributes ::= type [value]
type ::= CDATA
{ type ::= (second | minute | hour | day | partofDay | monthDay | weekDay | weekend | week | month | term | half-year | season | year | decade | century | millenium}
value ::= CDATA
{ value ::= (integer | spring | summer | autumn | winter)}

Les liens entre les deux adverbiaux unaires formant un adverbial de localisation binaire sont marqués par la balise INTERVAL. Le début et la fin de l'adverbial de localisation binaire sont marqués à l'aide des identifiants des adverbiaux unaires (ChronexID).

INTERVAL :
attributes ::= iid type start end
iid ::= IntervalID
IntervalID ::= i<integer>

type ::= CDATA
{ type ::= (between | fromTo)}
start ::= c<integer>
end ::= c<integer>

Les opérations de composition entre plusieurs adverbiaux de localisation temporelle sont marquées par la balise COMPOSITION_LINK. Les adverbiaux qui entrent dans la composition des adverbiaux composés peuvent être : des adverbiaux unaires (CHRONEX), des adverbiaux de localisation binaires (INTERVAL) ou des adverbiaux composés (COMPOSITION_LINK), chacun ayant un identifiant unique.

COMPOSITION_LINK :
attributes ::= clid type [operator] [operand] [list]
clid ::= CompositionLinkID
CompositionLinkID ::= cl<integer>
type ::= CDATA
{ type ::= (concatenation | specification | exception)}
operator ::= CDATA
{ operator ::= (c<integer> | il<integer> | cl<integer>)}
operand ::= CDATA
{ operand ::= (c<integer> | il<integer> | cl<integer>)}
list ::= CDATA
{ list ::= (' (c<integer> | il<integer> | cl<integer>) ') *}

La balise COMPOSITION_SIGNAL permet d'annoter des marqueurs de composition (*excepté les lundis 5 et 6 mai* ; *le 6, 7 et 8 février et les 12 et 13 mars*).

COMPOSITION_SIGNAL :
attributes ::= type [value]
type ::= CDATA
{ type ::= (exception | concatenation | specification) }

Enfin, le schéma permet de marquer, à l'aide de la balise CHRONOLOC_LINK, le lien entre un adverbial de localisation et l'objet sur lequel il porte (par exemple, un événement au sens de *TimeML*).

CHRONOLOC_LINK :
attributes ::= chonlid operator operand
chronlid ::= ChronolocationLinkID
ChronolocationLinkID ::= chronl<integer>
operator ::= CDATA
{ operator ::= (c<integer> | il<integer> | cl<integer>)}
operand ::= eiid

La DTD complète est présentée dans l'annexe 3.

Voici deux exemples d'annotation se conformant au schéma :

Exemple 1 :

La France "est restée dans une sorte d'aveuglement idéologique" jugeant "que Ben Ali était une rempart contre l'intégrisme islamique et contre l'immigration clandestine", expliquait cette semaine Vincent Geisser, spécialiste du monde arabe et musulman.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE ChronolocationML SYSTEM "Chronolocation-V1.dtd">
<ChronolocationML>
  <TEXT>
La France "est restée dans une sorte d'aveuglement idéologique" jugeant "que Ben Ali était une rempart contre l'intégrisme islamique et contre l'immigration clandestine", expliquait <CHRONEX cid="c1"><ZOOMING zmid="zm1" operand="b1"><CALENDAR_UNIT type="week">cette semaine</CALENDAR_UNIT></ZOOMING> <BASE bid="b1" type="deictic"/></CHRONEX> Vincent Geisser, spécialiste du monde arabe et musulman.
  </TEXT>
</ChronolocationML>
```

Dans cet exemple, on voit que l'opération de régionalisation n'est pas marquée : afin d'alléger les annotations, les opérations qui ne produisent pas de transformation peuvent ainsi ne pas être annotées. Le modèle permet toutefois, si nécessaire, de les préciser : il était aussi possible ici d'insérer l'annotation <ZONING zonid='zon1' type='ID' operand='zm1'/>.

Exemple 2 :

Ouvert de 10h à 19h, du lundi au samedi.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE ChronolocationML SYSTEM "Chronolocation-V1.dtd">
<ChronolocationML>
  <TEXT>
    Ouvert <INTERVAL iid ="il1" type="fromTo" start="c1" end="c2"><CHRONEX cid="c1"><ZONING zonid="zon1" type="since" operand="b1">de</ZONING> <BASE bid="b1" isIterative="true" type="calendar"><CALENDAR_UNIT type="hour">10h</CALENDAR_UNIT></BASE></CHRONEX> <CHRONEX cid="c2"><ZONING zonid="zon2" type="until" operand="b2">à</ZONING> <BASE bid="b2" isIterative="true" type="calendar"><CALENDAR_UNIT type="hour">19h</CALENDAR_UNIT></BASE></CHRONEX></INTERVAL>, <INTERVAL iid ="il2" type="fromTo" start="c3" end="c4"><CHRONEX cid="c3"><ZONING zonid="zon3" type="since" operand="b3">du</ZONING> <BASE bid="b3" isIterative="true" type="calendar"><CALENDAR_UNIT type="weekDay">lundi</CALENDAR_UNIT></BASE></CHRONEX> <CHRONEX cid="c4"><ZONING zonid="zon4" type="until" operand="b4">au</ZONING> <BASE bid="b4" isIterative="true" type="calendar"><CALENDAR_UNIT type="weekDay">samedi</CALENDAR_UNIT></BASE></CHRONEX></INTERVAL>.
  </TEXT>
  <COMPOSITION_LINK clid="cl1" type="specification" operator="il1" operand="il2"/>
</ChronolocationML>
```

Dans cet exemple, les deux adverbiaux de localisation binaires (*de 10h à 19h* et *du lundi au samedi*) sont annotés directement dans le texte. La relation de spécification entre ces adverbiaux est marquée sous le texte à l'aide de la balise COMPOSITION_LINK.

6.1.4.3 Complémentarité avec TimeML

Si les deux approches sont différentes, au sens où *TimeML* vise à annoter des entités isolées puis à les relier, et où, pour notre part, on cherche à annoter des unités discursives complètes (les adverbiaux de localisation temporelle), il est possible néanmoins de les coupler. Notre schéma d'annotation peut en effet aussi se présenter comme une structure posée par-dessus des annotations conformes à la spécification de *TimeML* (cf. fig. 47).

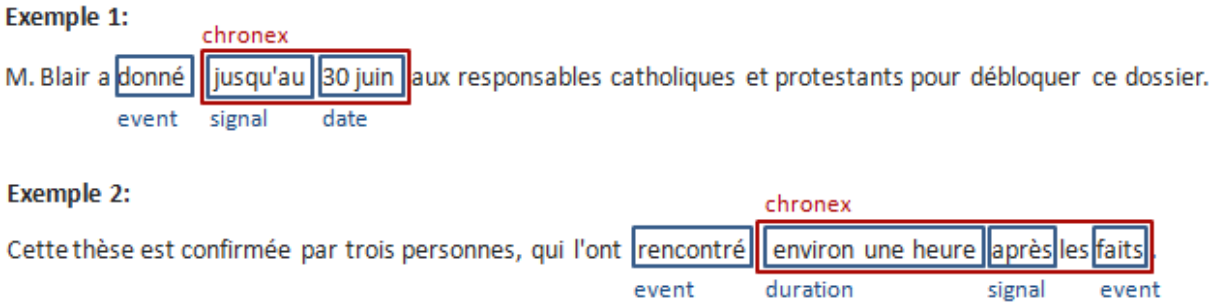


Fig. 47 : deux exemples d'énoncés où l'on visualise les segments annotés avec TimeML (en bleu) et ceux annoté avec ChronolocationML (en rouge)

Dans les deux exemples ci-dessus, on voit que les annotations au format *ChronolocationML* viennent par-dessus les marqueurs de relations temporelles (balise SIGNAL dans TimeML) et par-dessus les dates ou événements qui entrent dans la composition des adverbiaux de localisation temporelle.

Il est donc possible de mettre en œuvre un système d'annotation se conformant à la fois au schéma d'annotation des adverbiaux de localisation temporelle *ChronolocationML* et à *TimeML*. Ceci serait d'autant plus pertinent que les annotations produites par les systèmes qui se conforment à TimeML pourraient apporter des informations permettant de compléter les valeurs qui manquent parfois en sortie de notre système d'annotation pour décrire les adverbiaux de localisation temporelle : ceci à la fois pour l'annotation des adverbiaux déictiques et des adverbiaux de localisation temporelle dont le noyau est formé par une base relative à un procès ou encore pour établir la portée des adverbiaux.

En effet, notre système d'annotation, qui repose sur une analyse en surface, ne dispose pas toujours d'informations suffisantes pour compléter l'analyse de certains adverbiaux dont la base est déictique : ainsi, dans l'exemple 1 ci-dessus (cf. fig. 47), le système ne sait pas déterminer si le déplacement du déictique « *jusqu'au 30 juin* » est orienté avant ou après le processus énonciatif. Pour l'analyse des adverbiaux relatifs à un procès, les annotations de *TimeML*, si elles sont complètes, permettent de capter la référence au cœur de l'adverbial, comme dans l'exemple 2 (cf. fig. 47). Enfin, en s'appuyant sur les annotations de *TimeML* décrivant les liens entre événements et entre événements et expressions temporelles, il devient également possible, théoriquement, de délimiter le lien entre l'adverbe de localisation temporelle et le(s) prédicat(s) sur le(s)quel(s) il porte.

Inversement, l'annotation des adverbiaux permet d'exploiter la sémantique des opérations de focalisation et de déplacement dont *TimeML* ne tire pas parti, en particulier dans les adverbiaux relatifs à un procès (*à la mi-mandat*, *au troisième jour de sa visite*, par exemple, pour la focalisation ou encore, *quelques jours avant son arrivée*, pour les opérations de déplacement). Ainsi, dans

l'exemple 2, le segment « *environ une heure* » est annoté comme une durée. Les liens avec les événements désignés par le terme « *faits* » (« après les faits ») ne sont cependant pas décrits. En outre, l'annotation des opérations de composition entre adverbiaux unaires définies dans *ChronolocationML* doit permettre de mieux décrire les adverbiaux de localisation binaires (cf. l'exemple 3 ci-dessous : *d'ici le 30 juin*) et les adverbiaux composés (*dans la nuit de vendredi à samedi, vers minuit ; tous les jours sauf le dimanche, de 10h à 19h*).

Exemple 3:

L'exemple 4 présente une annotation simplifiée où l'on voit comment les schémas *ChronolocationML* et *TimeML* peuvent s'imbriquer. Pour faciliter la lecture, les attributs des éléments SIGNAL, EVENT, et TIMEX3 ne sont pas tous marqués. Les annotations provenant de *ChronolocationML sont surlignées*, celles provenant de *TimeML* sont en italique.

Exemple :

Europe Ecologie Les Verts (EE-LV) a affirmé mardi sa solidarité avec la mobilisation citoyenne en cours depuis plusieurs semaines dans la région de Sidi Bouzid (centre-ouest de la Tunisie).

<TEXT>
Europe Ecologie Les Verts (EE-LV) a *<EVENT eiid="ei1">*affirmé*</EVENT>*
<CHRONEX cid="c1">
<SHIFTING sid="s1" orientation="before" operand="b1">
<CALENDAR_UNIT type="weekDay">
*<TIMEX3 type=DATE>*mardi *</TIMEX3>*
</CALENDAR_UNIT>
</SHIFTING>
<BASE bid="b1" type="deictic"/>
</CHRONEX>
sa solidarité avec la *<EVENT eiid="ei2">*mobilisation*</EVENT>* citoyenne en cours
<CHRONEX cid="c2">
<ZONING zonid="z2" type="since" operand="s2">
*<SIGNAL>*depuis*</SIGNAL>*
</ZONING>
<TIMEX3 type=DURATION>
<SHIFTING sid="s2" orientation="before" quantity="1" comparisonOperator="more_than" operand="b2">plusieurs
<CALENDAR_UNIT type="week">semaines</CALENDAR_UNIT>
</SHIFTING>
</TIMEX3>
<BASE bid="b2" type="deictic"/>
</CHRONEX>
dans la région de Sidi Bouzid (centre-ouest de la Tunisie).
</TEXT>
<CHRONOLOC_LINK chronlid="chron1" operator="c1" operand="ei1"/>
<CHRONOLOC_LINK chronlid="chron2" operator="c2" operand="ei2"/>
<TLINK eventInstanceID ="ei1" lid="l1" relType="IS_INCLUDED" relatedToTime="t1"/>

Techniquement, si l'on devait réexploiter le système d'annotation que l'on a développé pour opérationnaliser cette double annotation (*TimeML* et *ChronolocationML*), cela nécessiterait un parseur pour récupérer les annotations au format *TimeML* produites par un autre système - par exemple (Bittar, 2010) pour le français. Il faudrait ensuite réinjecter dans le texte de nouvelles annotations correspondant aux informations inférées grâce à l'exploitation des deux jeux d'annotation. Une solution plus simple consisterait à développer un système unifié qui annoterait conjointement les segments textuels visés par *TimeML* et par *ChronolocationML*. C'est cette dernière solution qui est à l'étude dans le cadre du projet ANR Chronolines[40].

6.1.4.4 Les limites du système d'annotation

Les adverbiaux relatifs à un procès

Les adverbiaux relatifs à un procès sont des circonstanciels de temps où figure ce qui est le plus souvent appelé un *événement* dans les travaux en ingénierie des langues qui s'intéressent aux informations temporelles dans les textes (*deux jours avant les élections*, *lorsqu'il est arrivé*). Le système ne prenant pas en charge l'annotation des événements, ces adverbiaux ne sont pas ou pas complètement annotés. Seuls quelques-uns des éléments qui rentrent dans leur composition sont parfois annotés, en particulier les déplacements (*deux jours avant les élections*).

Lorsque ces adverbiaux correspondent à des subordonnées temporelles, ils peuvent être particulièrement ardus à annoter, en particulier pour la délimitation de la borne de fin de l'unité textuelle, alors même qu'il peut y avoir plusieurs adverbiaux, subordonnées ou incises enchâssées (*Ainsi, à la veille de la première tournée dans la région du secrétaire d'État américain, Madeleine Albright,…*). On renvoie sur ce sujet à la discussion abordée dans la section 3.1.2.1 et aux travaux de (Ehrmann et Hagège, 2009). TimeML contourne cette difficulté en n'annotant que la tête d'un événement (*head word*) (*tournée*, dans l'exemple précédent, « *à la veille de la première tournée* »).

Les adverbiaux de base déictique

Les annotations produites en sortie du module sont parfois incomplètes, du fait de l'absence d'analyse morphosyntaxique et syntaxique. C'est par exemple le cas pour certains déictiques, pour lesquels les valeurs décrivant l'orientation d'un déplacement ne sont pas renseignées : le plus souvent le temps des verbes dont dépendent les adverbiaux déictiques pourrait permettre de renseigner cette orientation. Ainsi, dans l'énoncé *« Une chute de tension artérielle **a été** à l'origine de l'hospitalisation **mardi** de l'ancien président égyptien Hosni Moubarak »,* l'orientation du déictique peut être inférée du passé composé. Des solutions libres existent pour effectuer ce type d'analyse, comme l'outil Freeling (Padró *et al.*, 2010) pour l'anglais ou encore celui développé par l'équipe d'Alpage pour le français (Denis et Sagot, 2010), mais elles pourraient entraîner des traitements coûteux en temps, puisqu'il faudrait enchaîner une analyse morphosyntaxique, une analyse syntaxique des énoncés et l'analyse sémantique des adverbiaux. Une solution pourrait être de ne

[40] http://chronolines.fr/

recourir à ce type d'analyse que localement, dans les cas où la description sémantique est incomplète à l'issue du processus d'annotation des adverbiaux déictiques.

On a vu par ailleurs avec TimeML que la normalisation des déictiques est considérée comme une sous-tâche, intermédiaire, de l'annotation. De notre point de vue, il s'agit plutôt d'une tâche d'une autre nature, dans la mesure où adverbiaux de localisation temporelle et valeurs calendaires sont deux modes distincts d'indexation temporelle : il ne s'agit donc pas d'une *normalisation*, mais plutôt d'une *transduction*. La transposition de l'une vers l'autre ne semble donc pas devoir ressortir d'une tâche qui consiste à décrire la sémantique des adverbiaux de localisation temporelle. Du reste, la transduction des adverbiaux de base calendaire périodique produit potentiellement des ensembles d'intervalles infinis. C'est pourquoi ici l'association de valeurs calendaires aux adverbiaux de localisation temporelle est prise en charge par le module de transduction, suite à l'annotation.

Le module de transduction est en mesure d'associer un intervalle calendaire aux adverbiaux déictiques dont la description est complète à l'issue du processus d'annotation (complète, au sens de la représentation formelle que nous proposons, c'est-à-dire que toutes les valeurs en sont renseignées) et lorsqu'on peut associer une valeur calendaire au processus énonciatif.

Ainsi, si l'on associe une valeur calendaire à une base déictique donnée (une base déictique désigne toujours le processus énonciatif), par exemple, s'il est possible de lui associer la valeur calendaire *23 avril 2012*, on peut alors associer une valeur calendaire à un adverbial comme *lundi dernier*, parce que l'annotateur est en mesure, même sans analyse profonde, de déterminer que le déplacement par rapport à la base déictique s'effectue vers l'arrière.

Les adverbiaux de base anaphorique

Retrouver la référence d'un anaphorique relève d'une tâche d'analyse textuelle complexe et difficile à opérationnaliser. Une analyse de surface n'est pas adaptée à cette tâche, qui fait partie de la problématique plus générale de la résolution des coréférences.

Les adverbiaux composés

Ici encore, notre approche s'écarte des propositions de TimeML où il demeure une hésitation devant l'analyse de certaines expressions temporelles : par exemple, dans le corpus annoté FR-Timebank, des expressions comme *le samedi 5 juin* *à 17 h* sont parfois annotées comme deux expressions, parfois comme une seule. Notre système les annote comme deux adverbiaux entretenant une relation de spécification (entre l'adverbial *le samedi 5 juin* et l'adverbial *à 17h*). TimeML considère également qu'il y a deux expressions temporelles dans l'adverbial *Entre 1990 et 1999* (« 1990 » et « 1999 ») ou encore dans l'adverbial *d'ici la fin de l'année* (« ici » et « fin de l'année »), alors que nous les analysons comme des adverbiaux de localisation binaires.

Les adverbiaux composés ne sont pas toujours identifiés comme tels en sortie du processus d'annotation. Leur analyse doit donc encore être complétée à deux niveaux : (1) au niveau des adverbiaux unaires qui rentrent dans leur composition et dont la description est incomplète du fait

de tournures elliptiques et (2) au niveau de la description des liens qui unissent les adverbiaux unaires. Comme on l'a vu, les méthodes *completeAnnotation* et *solveAssociation* de la classe *AnnotationPostProcessor* sont respectivement en charge de ces traitements.

6.1.4.5 Ressources pour le français

Les ressources pour l'annotation des adverbiaux de localisation temporelle en français sont constituées d'un ensemble de transducteurs Unitex, 70 au total.

Un graphe principal (cf. fig. 48) est subdivisé en différents sous-graphes de la façon suivante : des sous-graphes pour les adverbiaux unaires et composés, puis, dans chacun de ces sous-graphes, de nouveaux sous-graphes correspondant à la succession des opérations dans les adverbiaux de localisation temporelle (régionalisation, focalisation et déplacement) (cf. fig. 49 et 50), enfin des graphes pour les différentes bases possibles et les unités qui entrent dans leur composition.

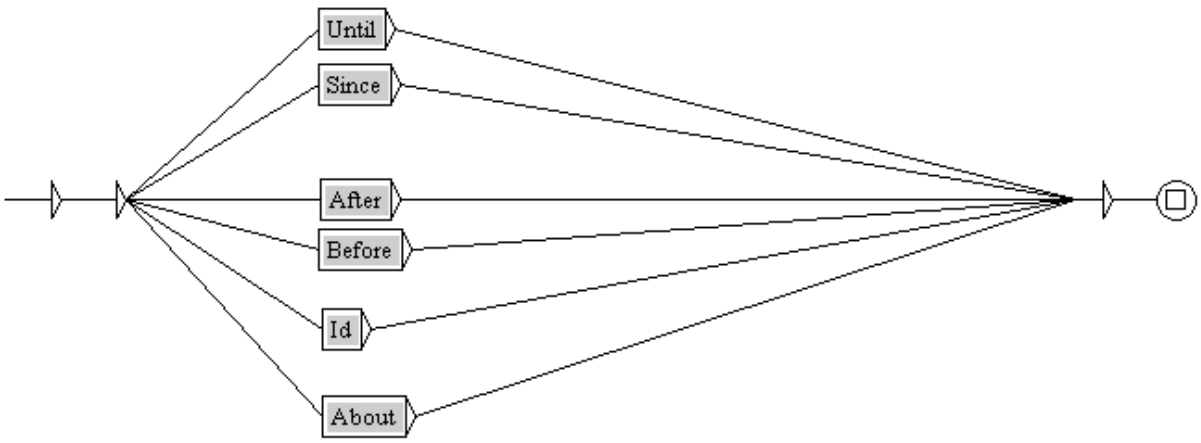


Fig. 48 : copie d'écran d'un graphe Unitex pour l'annotation des adverbiaux de localisation temporelle

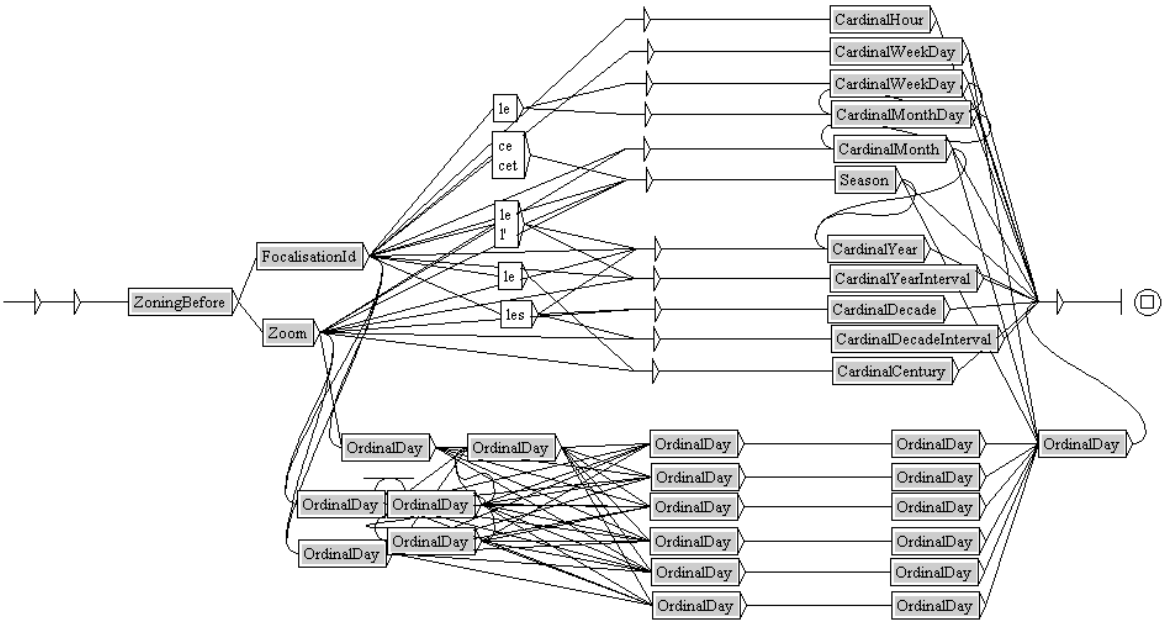


Fig. 49 : copie d'écran d'un sous graphe Unitex pour l'annotation des adverbiaux de localisation temporelle

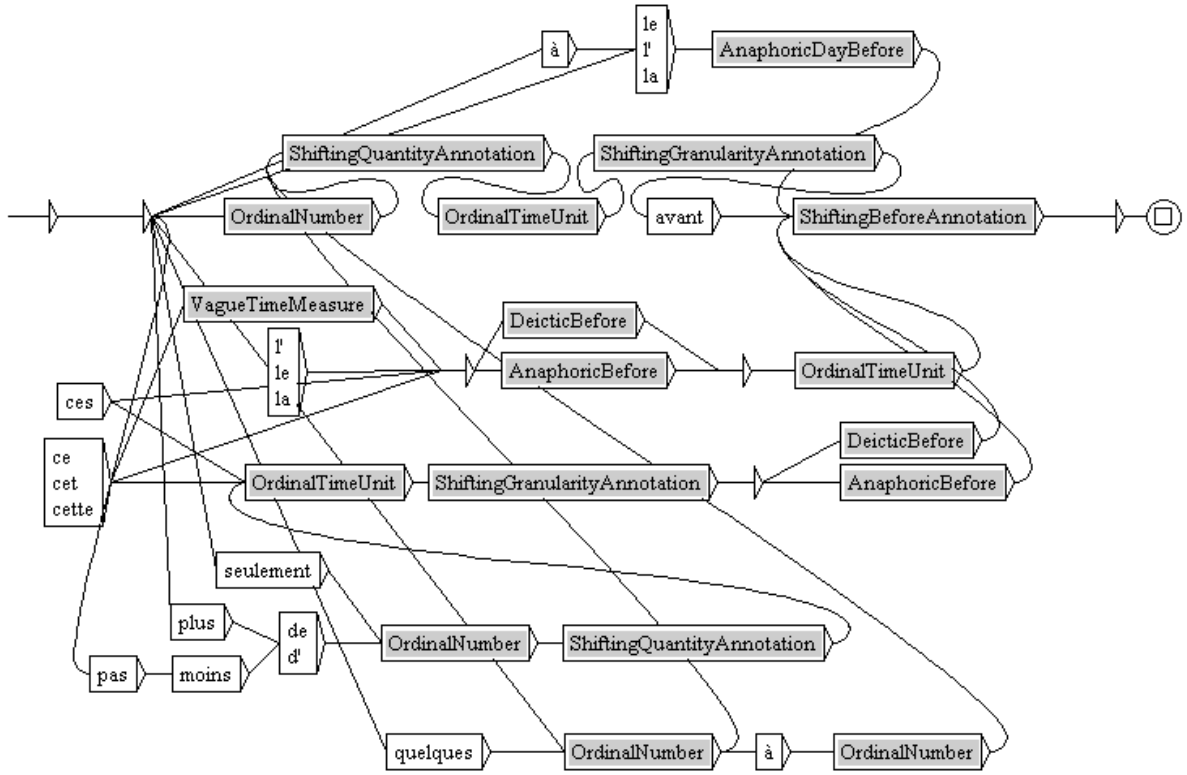


Fig. 50 : copie d'écran d'un sous graphe Unitex pour l'annotation des adverbiaux de localisation temporelle

Le développement de ces patrons d'annotation s'est fait progressivement pour améliorer les performances du système, l'objectif étant de réduire le bruit (les erreurs d'annotation) et le silence (l'absence non souhaitée d'une annotation), à mesure que le système était testé sur de nouveaux corpus (corpus de dépêches AFP, corpus littéraire, corpus d'articles de Wikipedia, corpus d'articles de

l'Est Républicain, etc.). On fournit dans la section 8.1 des éléments d'évaluation du système qui permet d'en mesurer les performances et les limites.

Alors qu'une bonne pratique eut été d'élaborer les ressources d'annotation après avoir arrêté un modèle stable de représentation des unités textuelles à annoter, dans notre cas de figure, le travail de modélisation des adverbiaux de localisation temporelle a accompagné l'effort de constitution des ressources pour les annoter.

Construites à partir d'un ensemble de ressources préexistantes, héritées de projets antérieurs (cf. section 6.1.4.1), elles ont en effet été enrichies en parallèle du travail visant à formaliser la description des adverbiaux de localisation temporelle. Elles ont donc évolué en même temps que cette description, entraînant une difficulté croissante dans leur maintenance : les modifications du modèle formel ont nécessité de mettre à jour régulièrement l'ensemble des règles d'annotation. Le schéma ChronolocationML est en outre, comme on l'a vu, un schéma construit a posteriori pour refléter ce modèle formel. Le schéma opérationnel utilisé par les ressources d'annotation s'appuie sur une DTD héritée du projet initial dans lequel ces ressources étaient utilisées. On rappelle également que des post-traitements sont effectués sur les textes annotés pour compléter l'analyse des adverbiaux de localisation temporelle (en particulier pour déterminer les associations entre ces adverbiaux). Pour que les annotations produites par notre système se conforment au schéma ChronolocationML, il faudrait donc ajouter un post-traitement permettant de sérialiser, dans ce format, les instances d'objet associés aux adverbiaux.

L'élaboration progressive de notre objet d'analyse a donc entraîné des modifications dans le jeu des règles d'annotation : initialement cantonnées à l'annotation des adverbiaux de base calendaire, les ressources ont dû ensuite permettre de capter d'autres types d'adverbiaux, en particulier les adverbiaux déictiques et les adverbiaux dont le repère temporel noyau est formé par un base calendaire périodique (« tous les lundis »). En effet, comme on l'a vu (cf. section 4.1), notre travail a consisté à généraliser la description initiale des « expressions calendaires », en partant du constat que la description de ces expressions sous la forme d'une succession des opérations sémantiques peut être généralisée à l'ensemble des adverbiaux de localisation temporelle.

Un tel travail de maintenance et d'adaptation des ressources d'annotation, coûteux, n'est pas facilité par le logiciel Unitex (au moins dans la version 2.0 utilisée), dans la mesure où il ne fournit pas de ressources pour s'assurer de la non-régression des performances du système, lorsqu'une règle ou un ensemble de règles est modifié[41]. La mise en place de tests de non-régression aurait eu portant un double avantage : celui de constituer progressivement un corpus de référence de plus en plus volumineux et celui d'avoir un meilleur contrôle sur les effets entraînés par les modifications des ressources d'annotation.

Une autre limite liée au logiciel utilisé pour développer les ressources d'annotation vient de la gestion des annotations produites en sortie du système : Unitex ne permet pas de revenir en arrière pour éliminer une annotation produite à un stade intermédiaire des traitements, ni non plus de

[41] Des ressources pour constituer des tests de non-régression sont annoncées pour les prochaines versions de la plate-forme Unitex (cf. Grf2test).

mettre en place simplement une cascade de transducteurs (un transducteur produisant une sortie utilisée ensuite par un autre transducteur et ce de façon récursive). Cette limitation implique de décrire dans un unique graphe (divisé éventuellement en sous-graphes) l'ensemble de l'unité textuelle à reconnaître et des éléments qui la composent, alors qu'une annotation décomposée en plusieurs passes permettrait de reconnaître des unités intermédiaires, en posant au besoin des annotations provisoires, susceptibles d'être conservées ou supprimées par les étapes suivantes d'annotation.

La grande variabilité des unités textuelles susceptibles de former des adverbiaux de localisation temporelle et la grande variabilité de leur agencement ont pour effet de rendre parfois les patrons d'annotation complexes, d'autant qu'ils sont, comme on l'a vu, imbriqués dans différents sous-graphes (cf. fig. 49 et 50). Une annotation découpée en plusieurs phases aurait permis d'alléger les graphes, facilitant ainsi leur maintenance.

Quelques exemples de sortie du système

Pour des raisons de lisibilité, les exemples présentés ont été modifiés pour qu'ils se conforment au schéma ChronolocationML, mais on souligne à nouveau que le système d'annotation s'appuie sur une autre DTD héritée d'un projet antérieur (cf. section 6.1.4.1).

Exemple 1 :
Outre la réforme du système monétaire mondial, objet d'un "premier séminaire" fin mars en Chine, ses explications ont également porté sur la lutte contre la volatilité des prix des matières premières.

Outre la réforme du système monétaire mondial, objet d'un "premier séminaire"
<CHRONEX cid="c1">
<ZOOMING zmid="zm1" subdivision="end" operand="s1">fin</ZOOMING>
<SHIFTING sid="s1" operand="b1">mars</SHIFTING>
<BASE bid="b1" type="deictic"/>
</CHRONEX>
en Chine, ses explications ont également porté sur la lutte contre la volatilité des prix des matières premières.

Dans l'exemple 1, on voit que l'analyse du déplacement (SHIFTING) n'est pas complète : son orientation n'est pas précisée.

Exemple 2 :
Jong a inscrit son 13e but pour Bochum, club avec lequel il est en contrat jusqu'en juin 2012.

Jong a inscrit son 13e but pour Bochum, club avec lequel il est en contrat
<CHRONEX cid="c1">
<ZONING zonid="zon1" type="until" operand="b1">jusqu'en</ZONING>
<BASE bid="b1" type="calendar">
<CALENDAR_UNIT type="month" value="6">juin</CALENDAR_UNIT>
<CALENDAR_UNIT type="year" value="2012">2012</CALENDAR_UNIT>
</BASE>
</CHRONEX>.

Exemple 3 :

L'Institut de médecine légale colombien, qui parraine aussi la manifestation, a indiqué que rien qu'entre janvier et mars 2011, 5.715 morts violentes ont été enregistrées en Colombie.

L'Institut de médecine légale colombien, qui parraine aussi la manifestation, a indiqué que rien qu'
<INTERVAL iid="i1" type="between" start="c1" end="c2">entre
<CHRONEX cid="c1">
<BASE bid="b1" type="calendar">
<CALENDAR_UNIT type="month" value="1">janvier</CALENDAR_UNIT>
<CALENDAR_UNIT type="year" value="2011"/>
</BASE>
</CHRONEX> et
<CHRONEX cid="c2">
<BASE bid="b2" type="calendar">
<CALENDAR_UNIT type="month" value="3">mars</CALENDAR_UNIT>
<CALENDAR_UNIT type="year" value="2011">2011</CALENDAR_UNIT>
</BASE>
</CHRONEX>
</INTERVAL>, 5.715 morts violentes ont été enregistrées en Colombie.

Dans l'exemple 3, on peut noter que le système a reporté sur la base formant le début de l'adverbial de localisation binaire (« *janvier* » dans « *entre janvier et mars 2011* ») l'information factorisée par la coordination, à savoir la mention de l'année 2011.

6.1.4.6 Ressources pour l'anglais

Les graphes Unitex implémentés pour l'anglais n'annotent que les adverbiaux calendaires itératifs et absolus. Ces ressources, moins complètes que celles développées pour le français, ont été développées pour les besoins des expérimentations menées dans le cadre du projet RMM2 (cf. section 6.2) et de celles menées autour du moteur de recherche expérimental décrit dans le chapitre 7.

Si les ressources pour l'anglais couvrent donc une partie moins importante des adverbiaux de localisation temporelle que les ressources pour le français (en particulier les adverbiaux déictiques et anaphoriques ne sont pas repérés), elles ont néanmoins été construites sur le même modèle. Elles respectent la même logique de découpage des graphes. Ces graphes sont au nombre de 68. Ils ont été constitués à partir d'un travail de repérage manuel effectué sur les différents corpus utilisés dans le cadre des expérimentations autour de l'acquisition de connaissances sur les dates et horaires d'accessibilité d'un service et autour de la recherche d'information : ils ont ainsi été essentiellement développés et testés à partir corpus d'articles de presse, d'articles de Wikipedia et sur des corpus d'énoncés définissant des dates et horaires d'ouverture de musées.

Quelques exemples de sortie du système

Pour des raisons de lisibilité, les exemples présentés ont été modifiés pour qu'ils se conforment au schéma ChronolocationML, mais on souligne à nouveau que le système d'annotation s'appuie sur une autre DTD héritée d'un projet antérieur (cf. section 6.1.4.1).

Exemple 1 :

By the 1770s the Thirteen Colonies contained two and half million people.

<CHRONEX cid="c1">
<ZONING zonid="zon1" type="around" operand="b1">By</ZONING>
<BASE bid="b1" type="calendar">
<CALENDAR_UNIT type="decade" value="1770">the 1770s</CALENDAR_UNIT>
</BASE>
</CHRONEX> the Thirteen Colonies contained two and half million people.

Exemple 2 :
By the end of the 19th century a few western states had granted women full voting rights, though women had made significant legal victories, gaining rights in areas such as property and child custody.

<CHRONEX cid="c1">
<ZONING zonid="zon1" type="around" operand="zm1">By</ZONING>
<ZOOMING zmid="zm1" operand="b1" subdivision="end">the end of</ZOOMING>
<BASE bid="b1" type="calendar">
<CALENDAR_UNIT type="century" value="19">19th century</CALENDAR_UNIT>
</BASE>
</CHRONEX> a few western states had granted women full voting rights, though women had made significant legal victories, gaining rights in areas such as property and child custody.

Les ressources pour l'annotation des adverbiaux de localisation temporelle présents dans des textes en langue française ou anglaise permettent de les décrire sous la forme d'une représentation formelle qui peut ensuite être exploitée à différentes fins. En particulier, nous présentons dans la section suivante, l'implémentation d'un module permettant de transposer cette représentation vers une représentation sous la forme d'intervalles calendaires. Une évaluation des ressources d'annotation est présentée dans le chapitre 8 (cf. sections 8.1.2 et 8.1.3).

6.1.5 Le module de transduction vers des intervalles calendaires

Le module de transduction est une implémentation de l'algorithme présenté dans la section 5.2. Il raisonne à partir d'instances d'objet décrivant des adverbiaux de localisation temporelle et produit des instances d'*Intervalles Calendaires*.

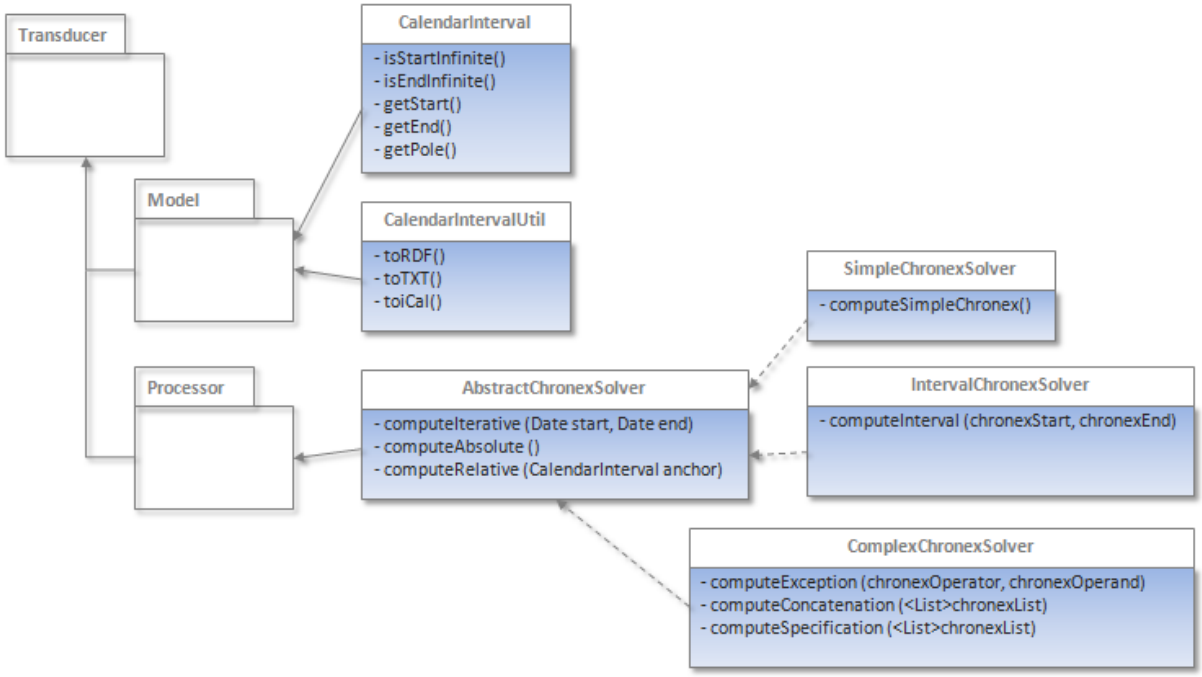


Fig. 51 : architecture du module de transduction

Le module comprend une implémentation objet des Intervalles Calendaires (*CalendarInterval*). La classe *CalendarInterval* permet notamment de récupérer les dates de début et de fin de l'intervalle au format ISO 8601 (sauf si ces bornes sont infinies), ainsi que le pole associé à l'intervalle. Ce pôle est également une date au format ISO 8601, qui peut-être la borne de début, la borne de fin ou bien correspondre au milieu de l'intervalle calendaire (cf. section 5.3.5). Cette valeur est utilisée par le système de recherche d'information que l'on présentera dans le chapitre suivant.

Le raisonneur transpose les instances d'objet associées aux adverbiaux de localisation sous la forme d'un Intervalle Calendaire ou d'un ensemble d'Intervalles Calendaires, à l'appel des méthodes suivantes :

- *computeAbsolute()* : conformément à la procédure de transduction présentée dans la section 5.2.1, cette méthode associe un intervalle calendaire à un adverbial calendaire absolu.
- *computeIterative()* : conformément à la procédure de transduction présentée dans la section 5.2.2, cette méthode associe un ensemble d'intervalles calendaires à un adverbial calendaire itératif. L'implémentation prévoit que le raisonnement s'effectue sur une fenêtre de temps précisée en argument. Il faut donc spécifier en paramètre une date de début et une date de fin. L'ensemble d'intervalles ainsi obtenu n'est donc qu'une projection de l'adverbial transposé sur une fenêtre de temps donnée.
- *computeRelative()* : cette méthode permet d'associer un intervalle calendaire aux adverbiaux déictiques dont la description est complète et pour laquelle on fournit un intervalle calendaire de référence, qui correspond à la valeur calendaire associée au processus énonciatif. En théorie, elle peut fonctionner aussi sur les adverbiaux de localisation relatifs à un procès ou les adverbiaux anaphoriques, si on peut également associer un intervalle calendaire à leurs bases.

La classe *Util* contient des méthodes pour formater les instances d'intervalles calendaires :

- *toRDF()* : la méthode permet d'exporter au format RDF les instances d'un intervalle calendaire. Cette méthode prend son sens dans l'architecture plus générale où l'ensemble des composants s'insère ; elle permet en particulier d'exporter les instances vers une base de connaissances ou vers un moteur d'indexation.
- *toiCal()* : la méthode permet d'exporter au format iCalendar les instances d'un intervalle calendaire, afin notamment de pouvoir interagir avec des agendas numériques.
- *toTXT()* : cette méthode génère une expression en langage naturel correspondant à une instance du modèle d'intervalle calendaire.

Les trois modules qui prennent part au système d'annotation (le module pour manipuler des adverbiaux de localisation temporelle sous la forme d'objet, le module d'annotation et le module de transduction) permettent ainsi d'extraire des informations temporelles dans les textes et de les rendre manipulables par des agents logiciels. Le système d'annotation des adverbiaux de localisation temporelle est ainsi exploité par deux systèmes expérimentaux que nous avons mis en œuvre, afin d'illustrer l'intérêt de l'approche : l'un pour la recherche d'information, l'autre pour l'acquisition de connaissances relatives à des dates et horaires d'ouverture, que l'on présente dans la section suivante.

6.2 Outiller la saisie d'informations temporelles complexes : un cas d'application industriel

6.2.1 Motivations et enjeux

Si le Web Sémantique tel qu'il est envisagé par (Berners-Lee *et al.*, 2001) doit s'entendre comme un dispositif permettant aux machines d'assister plus efficacement les utilisateurs pour l'accès aux ressources sur le Web, dans le mouvement inverse, la puissance d'interrogation qu'il offre pour l'utilisateur final s'accompagne également d'une complexité croissante pour ceux qui ont à charge de modéliser, de maintenir et d'alimenter les bases de connaissances au cœur de cette infrastructure.

En ce sens, l'assistance à l'alimentation manuelle des bases de connaissances répond en partie à cette problématique. En s'appuyant sur des traitements linguistiques, ainsi que sur des processus de raisonnement pour construire des données structurées à partir du texte libre, il devient possible de faciliter la saisie d'informations complexes. C'est dans le cadre d'une coopération avec un industriel, Relaxnews, partenaire du projet de recherche RMM2[42], que le prototype présenté ici a été développé. Cette entreprise produit et diffuse des fils de dépêches relatifs au tourisme et aux loisirs. Une des difficultés que rencontre Relaxnews concerne la saisie et le stockage d'informations temporelles complexes, en l'occurrence les dates et horaires dits *d'accessibilité*. On présente ici une solution logicielle, TKA (*Temporal Knowledge Acquisition*), destinée à faciliter la saisie et le stockage

[42] http://www.rmm2.org/

de ce type d'informations, qui peuvent décrire les horaires d'ouverture d'un magasin, la programmation d'un théâtre ou d'un cinéma, par exemple.

Le recours aux techniques de l'ingénierie des langues pour faciliter l'alimentation d'ontologies à partir de textes (Bontcheva et Cunningham, 2003) et plus généralement la constitution de ressources termino-ontologiques (Bourigault *et al.*, 2004) est un des moyens de tirer parti des textes pour en extraire des connaissances. L'objectif est ici d'en faciliter l'industrialisation pour un cas d'usage précis, en ne faisant porter l'analyse que sur des portions restreintes et bien identifiées de la langue et en couplant l'ensemble avec des outils de normalisation et de raisonnement permettant un contrôle de l'information saisie. Il ne s'agit donc pas d'extraire des informations dans des textes, mais de proposer une solution d'assistance au peuplement manuel d'ontologies pour les informations concernant les dates et horaires d'accessibilité d'un lieu.

Dans ce cas d'application, l'objectif est double. Il s'agit, d'un côté, de fournir des outils pour renseigner de façon simple, dans une base de connaissances, des périodes d'accès et, d'un autre, d'offrir des outils pour les interroger à travers un portail : quand un site touristique, un musée, un restaurant est-il ouvert ? Quel jour et à quelle heure ont lieu les séances d'une pièce de théâtre ? On présentera cette problématique de la recherche d'information temporelle dans le chapitre suivant, mais voici le type de requêtes que l'on souhaiterait pouvoir traiter :

« musées ouverts le weekend du 1er mai »
« supermarchés ouverts le dimanche à Nanterre »
« le restaurant R est-il ouvert ce soir ? »

Dans ce cadre, la notion d'accessibilité recouvre aussi bien l'accès à un musée, à un restaurant, la programmation d'un festival ou de séances de cinéma. Désormais, nous désignerons par « période d'accessibilité » tout énoncé renvoyant à des propriétés temporelles caractérisant l'accessibilité d'un lieu, dont voici plusieurs exemples extraits sur les pages Web de différents musées :

> Ex. 1 : Ouvert du mardi au samedi de 10h à 19h et le dimanche de 13h à 19h, sauf les jours fériés suivants : 1er jan, dimanche et lundi de Pâques, 1er et 8 mai.
> Ex. 2 : Ouvert de 10h à 4h et le vendredi et samedi de 10h à 5h.
> Ex. 3 : Horaires : Du lundi au jeudi : 9h à 21h30. Vendredi et samedi : de 9h à 22h30. Dimanche : de 9h à 18h30. Fermé le 1er janvier.
> Ex. 4 : Ouvert tous les jours de 10h à 19h jusqu'à dimanche 3 janvier. A partir du lundi 4 janvier : ouvert le mercredi, samedi et dimanche : 10h-19h ; le vendredi : 14h-19h. Fermé le 1er janvier.

La capacité qu'offre la langue de condenser, dans des formules brèves, des périodes dont la définition en extension peut être coûteuse en termes de stockage (définitions itératives vs. définitions en extension), explique pourquoi la représentation sémantique des adverbiaux itératifs est une bonne alternative aux représentations standards et normées du temps calendaire, pour définir les propriétés temporelles relatives à l'accessibilité d'un site. En effet, le système proposé permet de stocker, selon les besoins, ou bien des périodes définies en intension ou bien des intervalles calendaires définis en extension. Le système que nous présentons évite par ailleurs aux

opérateurs d'avoir à saisir ces informations à travers des formulaires qui découpent l'information en tiroirs multiples (cf. la fig. 52 ci-dessous qui présente un exemple de ce type de formulaire[43] :

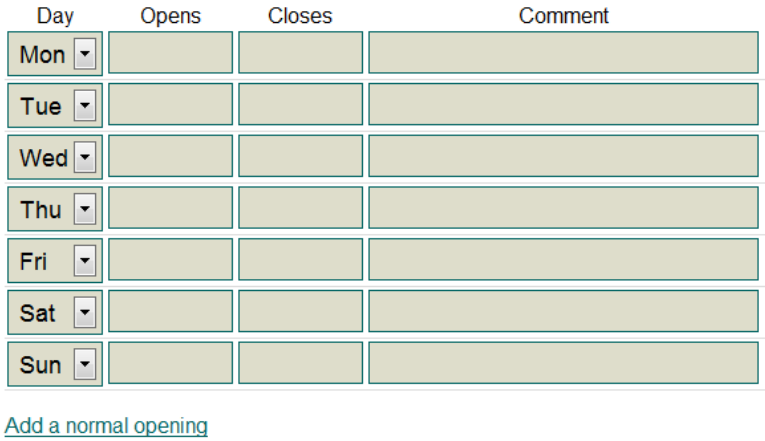


Fig. 52 : exemple de formulaire de saisie de dates et horaires d'ouverture/fermeture

Ce mode de saisie, à travers plusieurs formulaires, rend difficile et longue la saisie d'horaires, simples à exprimer en langue (par exemple, *« le dimanche de 20h à 2h du matin » ; « du 15 mars au 30 septembre, ouvert du lundi au vendredi de 10h à 19h »*), parce que le découpage des informations oblige à multiplier les créneaux horaires définis.

6.2.2 TKA, un système d'assistance à la saisie de dates et horaires d'ouverture

6.2.2.1 Architecture

Le système que l'on présente permet à des opérateurs de saisir ou copier ce type d'informations en langage naturel via un service Web. Les informations saisies sont alors annotées. Afin que les utilisateurs puissent contrôler l'analyse et si nécessaire modifier les informations saisies, le système projette les résultats de l'analyse sur un agenda numérique éditable.

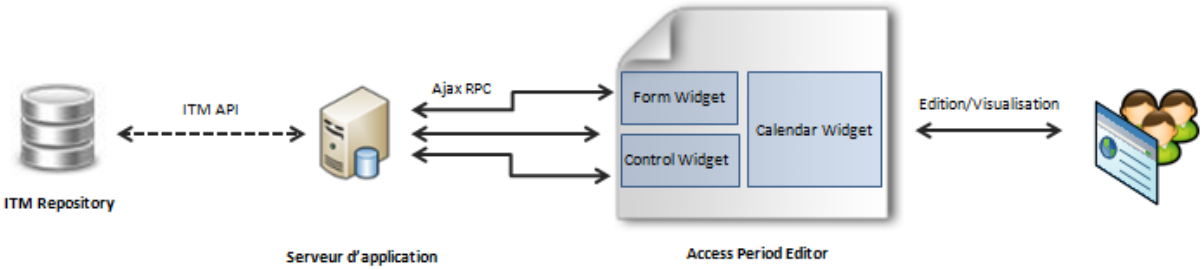


Fig. 53 : architecture du service Web TKA

L'application Web, un module GWT[44], est composée d'une partie *serveur*, responsable de l'interrogation de la base de connaissances et des échanges avec le système d'annotation, et d'une partie *client*. Cette dernière est constituée d'agents logiciels (*widgets*) qui génèrent et injectent

[43] http://opening-times.co.uk/f/new
[44] https://developers.google.com/web-toolkit/

dynamiquement du code HTML dans la page Web de l'application. Chaque widget gère l'affichage d'un composant : (1) le premier (*Form Widget*) correspond à l'interface de saisie des dates et horaires d'ouverture en langage naturel, (2) le second (*Control Widget*) au panneau de contrôle permettant de lancer l'annotation et de valider les informations saisies et (3) le troisième contient la fenêtre d'affichage d'un agenda numérique avec lequel il est possible d'interagir.

L'annotation des informations saisies par les opérateurs s'appuie sur le système présenté dans la section précédente, enrichi par de nouveaux graphes Unitex pour repérer les prédicats dénotant l'ouverture ou la fermeture d'un site. On a vu que le système permettait d'associer des ensembles d'intervalles aux adverbiaux dont la base calendaire est absolue ou itérative et d'exporter les instances d'objet associées aux adverbiaux dans différents formats : iCalendar pour l'affichage sur un agenda numérique et RDF pour le stockage.

L'application interagit ainsi avec le système d'annotation, afin de traduire dans des formats structurés les énoncés définissant des périodes d'accessibilité. L'application projette le résultat de leur analyse sur un agenda numérique éditable, afin que l'utilisateur puisse vérifier la cohérence de la transformation et éventuellement modifier, ajouter ou corriger les informations. La synchronisation entre les deux représentations, textuelles et calendaires, est maintenue par l'applicatif : une modification sur l'agenda numérique entraîne la modification du texte et inversement.

Cette synchronisation permet ainsi deux modes de saisie, qui ont chacun leur avantage respectif : la saisie en langage naturel permet de définir les horaires sous une forme ramassée, l'agenda numérique permettant lui d'ajouter et de modifier facilement les informations.

Le prototype a vocation à s'insérer dans une architecture logicielle, décrite dans (Noël et Azémard, 2008), où une base de connaissances (ici, celle qui contient les données sur l'accessibilité des sites) est couplée à un moteur de recherche. Pour que les moteurs de recherche puissent traiter des requêtes contenant des critères temporels, on verra que les définitions de périodes d'accessibilité doivent être exprimées sous la forme d'ensembles d'intervalles calendaires, à la fois pour l'indexation et l'interrogation. En effet, la transposition des périodes d'accessibilité sous la forme d'intervalles calendaires ne peut pas être opérée à la volée, car la procédure affecterait trop les temps de réponse. Elle est donc opérée en amont : dans cette architecture, la transposition des adverbiaux calendaires sous la forme d'intervalles calendaires s'effectue au moment de l'export de la base pour l'indexation par des moteurs de recherche. La base de connaissances, en revanche, stocke l'information sur les périodes d'accessibilité sous la forme d'une représentation conforme à la représentation sémantique exprimée dans un format RDF, ce qui permet de ne conserver dans la base qu'un réseau sémantique de taille réduite.

6.2.2.2 Le format de stockage RDF

Les instances d'objet associées aux adverbiaux de localisation temporelle sont ainsi stockées dans un format OWL/RDF (chronex.owl) qui se conforme à la représentation décrite dans le chapitre 4. L'ontologie définit neuf classes :

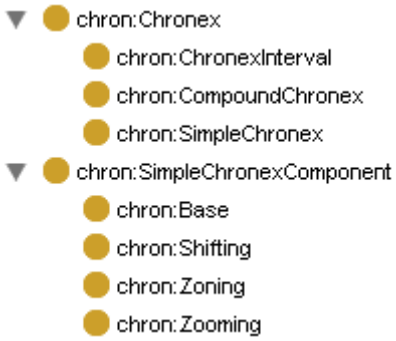


Fig. 54 : les classes de l'ontologie chronex.owl

Pour alléger le réseau sémantique des instances, les associations entre adverbiaux de localisation temporelle ne sont pas réifiées : elles sont définies sous la forme d'*Object Properties* (*exceptionOperator*, *exceptionOperand*, *specification* et *concatenation*) permettant d'associer un adverbial composé (*CompoundExpression*) à des adverbiaux de localisation temporelle (*Chronex*).

Les unités calendaires (heure, jour du mois, mois, année, etc.) sont représentées sous la forme de *Datatype Properties* pointant sur un ensemble de valeurs contraintes.

Les relations entre un opérande (une des classes des *SimpleChronexComponent*) et son opérateur sont modélisées à l'aide des *Object Properties zoningOperator*, *zoomingOperator*, *shiftingOperator* et *baseOperator*. Ces propriétés permettent de remonter le fil des opérations agissant sur une Base, reliée elle-même directement à l'adverbial unaire dont elle est la référence noyau (*SimpleChronex*) par l'*Object Property* symétrique *base*.

Afin de pouvoir représenter l'accessibilité d'un site, on a défini une classe *AccessPeriod* (Période d'Accès) et les deux *Object Properties accessibleDuring* (accessible durant) et *notAccessibleDuring* (non accessible durant) pointant sur un objet de type *Chronex*. Ces relations ne sont pas définies dans l'ontologie *chronex.owl* elle-même, mais dans l'ontologie de domaine qui l'utilise, définissant des périodes d'accès. Par exemple, l'énoncé « *Ouvert du mardi au samedi, de 10h à 19h et le dimanche, de 13h à 19h. Fermé les 1er et 8 mai.* » sera modélisé sous la forme de trois adverbiaux de localisation temporelle (la description RDF de ces adverbiaux est présentée ci-dessous) : (1) CompoundChronex_1 : *du mardi au samedi, de 10h à 19h* ; (2) CompoundChronex_2 : *le dimanche, de 13h à 19h* ; (3) SimpleChronex_3 : *les 1er et 8 mai*. On définit deux Périodes d'Accès : la première est reliée aux deux premiers adverbiaux (CompoundChronex_1 et CompoundChronex_2) par la propriété *accessibleDuring* et au troisième (SimpleChronex_3) par la propriété *notAccessibleDuring*.

Représentation RDF de l'adverbial composé *« du mardi au samedi, de 10h à 19h »*

```
<chron:Base rdf:about="#Base_4">
  <chron:dayOfWeek rdf:datatype="&xsd;int">2</chron:dayOfWeek>
  <chron:label rdf:datatype="&xsd;string">mardi</chron:label>
</chron:Base>
<chron:Base rdf:about="#Base_5">
  <chron:dayOfWeek rdf:datatype="&xsd;int">6</chron:dayOfWeek>
  <chron:label rdf:datatype="&xsd;string">samedi</chron:label>
```

```
</chron:Base>

<chron:SimpleChronex rdf:about="#SimpleChronex_4">
   <chron:label rdf:datatype="&xsd;string">mardi</chron:label>
   <chron:base rdf:resource="#Base_4"/>
</chron:SimpleChronex>
<chron:SimpleChronex rdf:about="#SimpleChronex_5">
   <chron:label rdf:datatype="&xsd;string">samedi</chron:label>
   <chron:base rdf:resource="#Base_5"/>
</chron:SimpleChronex>

<chron:ChronexInterval rdf:about="#ChronexInterval_1">
   <chron:label rdf:datatype="&xsd;string">du mardi au samedi</chron:label>
   <chron:end rdf:resource="#SimpleChronex_4"/>
   <chron:start rdf:resource="#SimpleChronex_5"/>
</chron:ChronexInterval>

<chron:Base rdf:about="#Base_6">
   <chron:hour rdf:datatype="&xsd;int">10</chron:hour>
   <chron:label rdf:datatype="&xsd;string">10h</chron:label>
</chron:Base>
<chron:Base rdf:about="#Base_7">
   <chron:hour rdf:datatype="&xsd;int">19</chron:hour>
   <chron:label rdf:datatype="&xsd;string">19h</chron:label>
</chron:Base>

<chron:SimpleChronex rdf:about="#SimpleChronex_6">
   <chron:base rdf:resource="#Base_6"/>
   <chron:label rdf:datatype="&xsd;string">de 10h</chron:label>
</chron:SimpleChronex>
<chron:SimpleChronex rdf:about="#SimpleChronex_7">
   <chron:base rdf:resource="#Base_7"/>
   <chron:label rdf:datatype="&xsd;string">&#224; 19h</chron:label>
</chron:SimpleChronex>

<chron:ChronexInterval rdf:about="#ChronexInterval_2">
   <chron:label rdf:datatype="&xsd;string">de 10h &#224; 19h</chron:label>
   <chron:end rdf:resource="#SimpleChronex_6"/>
   <chron:start rdf:resource="#SimpleChronex_7"/>
</chron:ChronexInterval>

<chron:CompoundChronex rdf:about="#CompoundChronex_2">
   <chron:label rdf:datatype="&xsd;string">du mardi au samedi, de 10h &#224; 19h</chron:label>
   <chron:specification rdf:resource="#ChronexInterval_1"/>
   <chron:specification rdf:resource="#ChronexInterval_2"/>
</chron:CompoundChronex>
```

Représentation RDF de l'adverbial composé *« le dimanche, de 13h à 19h »*

```
<chron:Base rdf:about="#Base_8">
   <chron:dayOfWeek rdf:datatype="&xsd;int">6</chron:dayOfWeek>
   <chron:label rdf:datatype="&xsd;string">dimanche</chron:label>
</chron:Base>

<chron:SimpleChronex rdf:about="#SimpleChronex_8">
```

```
  <chron:label rdf:datatype="&xsd;string">dimanche</chron:label>
  <chron:base rdf:resource="#Base_8"/>
</chron:SimpleChronex>

<chron:Base rdf:about="#Base_9">
  <chron:hour rdf:datatype="&xsd;int">13</chron:hour>
  <chron:label rdf:datatype="&xsd;string">13h</chron:label>
</chron:Base>
<chron:Base rdf:about="#Base_10">
  <chron:hour rdf:datatype="&xsd;int">19</chron:hour>
  <chron:label rdf:datatype="&xsd;string">19h</chron:label>
</chron:Base>

<chron:SimpleChronex rdf:about="#SimpleChronex_9">
  <chron:label rdf:datatype="&xsd;string">de 13h</chron:label>
  <chron:base rdf:resource="#Base_9"/>
</chron:SimpleChronex>

<chron:SimpleChronex rdf:about="#SimpleChronex_10">
  <chron:label rdf:datatype="&xsd;string">à 19h</chron:label>
  <chron:base rdf:resource="#Base_10"/>
</chron:SimpleChronex>

<chron:ChronexInterval rdf:about="#ChronexInterval_3">
  <chron:label rdf:datatype="&xsd;string">de 13h &#224; 19h</chron:label>
  <chron:end rdf:resource="#SimpleChronex_10"/>
  <chron:start rdf:resource="#SimpleChronex_9"/>
</chron:ChronexInterval>

<chron:CompoundChronex rdf:about="#CompoundChronex_3">
  <chron:label rdf:datatype="&xsd;string">le dimanche, de 13h &#224; 19h</chron:label>
  <chron:specification rdf:resource="#ChronexInterval_3"/>
  <chron:specification rdf:resource="#SimpleChronex_8"/>
</chron:CompoundChronex>
```

Représentation RDF de l'adverbial *« les 1er et 8 mai »*

```
<chron:Base rdf:about="#Base_3">
  <chron:dayOfMonth rdf:datatype="&xsd;int">1</chron:dayOfMonth>
  <chron:month rdf:datatype="&xsd;int">5</chron:month>
  <chron:dayOfMonth rdf:datatype="&xsd;int">8</chron:dayOfMonth>
  <chron:label rdf:datatype="&xsd;string">les 1er et 8 mai</chron:label>
</chron:Base>
<chron:SimpleChronex rdf:about="#SimpleChronex_3">
  <chron:label rdf:datatype="&xsd;string">les 1er et 8 mai</chron:label>
  <chron:base rdf:resource="#Base_3"/>
</chron:SimpleChronex>
```

Représentation RDF de l'énoncé définissant l'accessibilité d'un service *« Ouvert du mardi au samedi, de 10h à 19h et le dimanche, de 13h à 19h. Fermé les 1er et 8 mai. »*

```
<ap:AccessPeriod rdf:about="#AccessPeriod_1">
```

```
   <ap:label rdf:datatype="&xsd;string">Ouvert du mardi au samedi, de 10h &#224; 19h et le dimanche, de 13h
&#224; 19h. </ap:label>
   <ap:accessibleDuring rdf:resource="#CompoundChronex_2"/>
   <ap:accessibleDuring rdf:resource="#CompoundChronex_3"/>
</ap:AccessPeriod>

<ap:AccessPeriod rdf:about="#AccessPeriod_1">
   <ap:label rdf:datatype="&xsd;string">Fermé les 1er et 8 ma.i</ap:label>
   <ap:notAccessibleDuring rdf:resource="#SimpleChronex_3"/>
</ap:AccessPeriod>
```

6.2.2.3 Fonctionnalités et Interfaces Utilisateurs

L'interface Web d'édition (cf. fig. 55), disponible en français et en anglais est répartie en trois zones (widgets) : une zone pour la saisie en langage naturel des périodes d'accessibilité (cf. fig. 56), (2) une zone pour l'affichage des périodes d'accessibilité sur un agenda numérique (cf. fig. 57) et (3) une zone de contrôle pour lancer l'annotation et stocker les périodes d'accès dans la base de connaissances (cf. fig. 58).

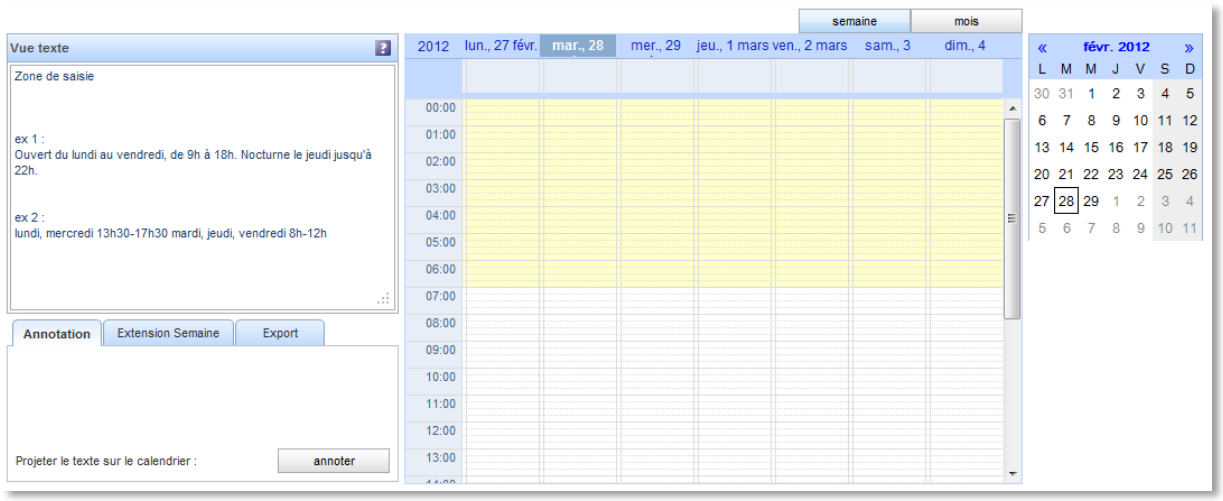


Fig. 55 : l'interface utilisateur de TKA

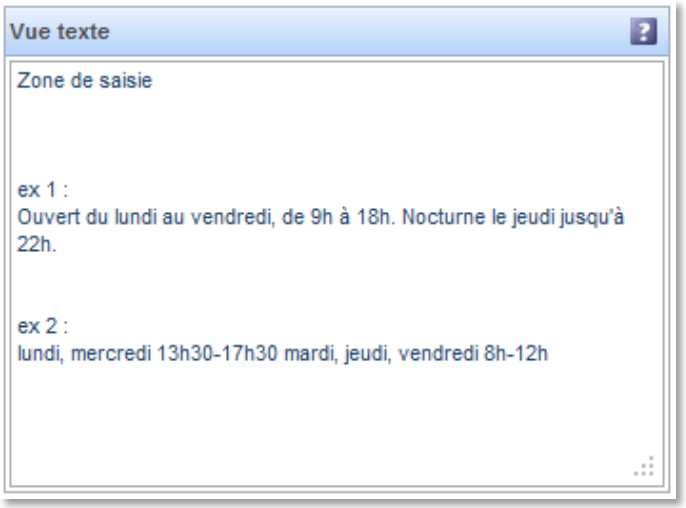


Fig. 56 : la zone de saisie en langage naturel des périodes d'accès

Le calendrier où sont projetés les résultats de l'analyse est éditable : un utilisateur peut ainsi par exemple y modifier un horaire ou y ajouter un jour d'ouverture ou de fermeture. Le texte initialement saisi est alors automatiquement mis à jour. Un opérateur peut ainsi, à travers le calendrier, préciser que le samedi 1er mai est un jour de fermeture. La phrase suivante « *Fermé samedi 1 mai 2010.* » est alors automatiquement générée et ajoutée à l'énoncé initial.

L'utilisateur peut donc jongler entre deux manières de définir des périodes d'accessibilité, qui chacune renvoie à un modèle qui lui est propre : le texte et le calendrier. Les informations ajoutées directement sur le calendrier définissent toujours des périodes d'accès non itératives (*fermé le 1er mai 2012,* par exemple) : on ne peut pas, via le calendrier, préciser par exemple qu'un site est fermé le 1er mai chaque année ; pour cela, il faut modifier directement le texte (« *Fermé le 1er mai* »). Après la saisie, les données sont stockées au format RDF défini plus haut.

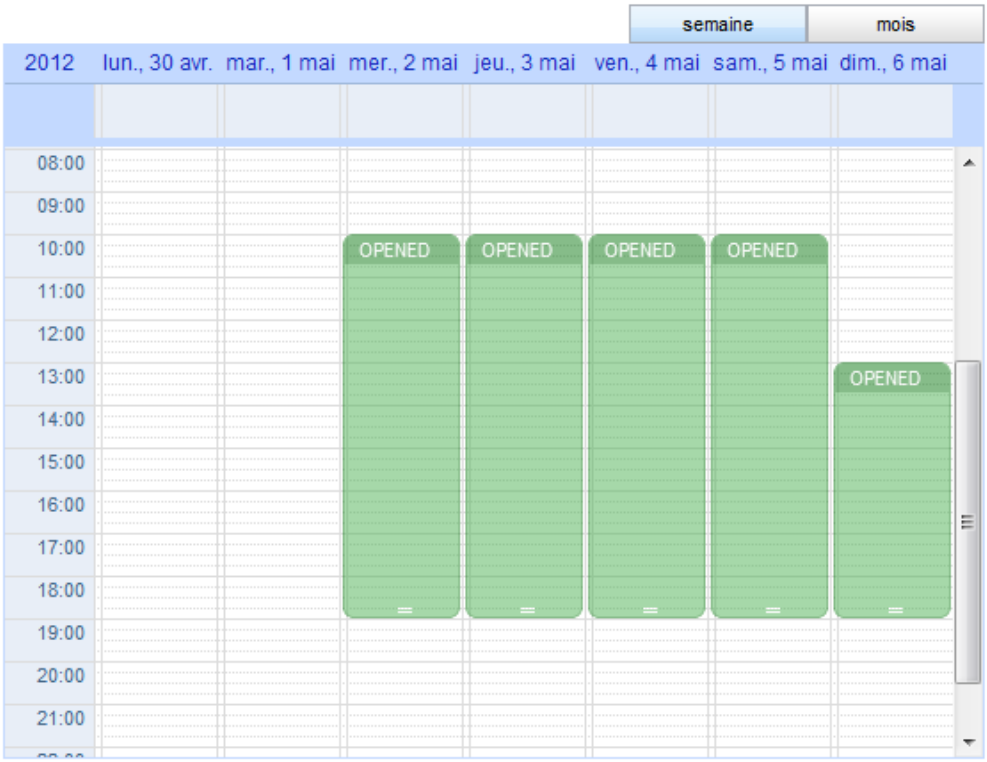


Fig. 57 : Visualisation correspondant à la période d'accès « Ouvert du mardi au samedi, de 10h à 19h et le dimanche, de 13h à 19h. Fermé les 1er et 8 mai. » pour la semaine du 30 avril au 6 mai 2012

Le panneau de contrôle (cf. fig. 58) contient trois onglets : l'un pour exécuter l'annotation du texte saisi, l'autre pour étendre sur une fenêtre de temps des horaires définis directement sur l'agenda et le troisième pour l'export et le stockage des informations.

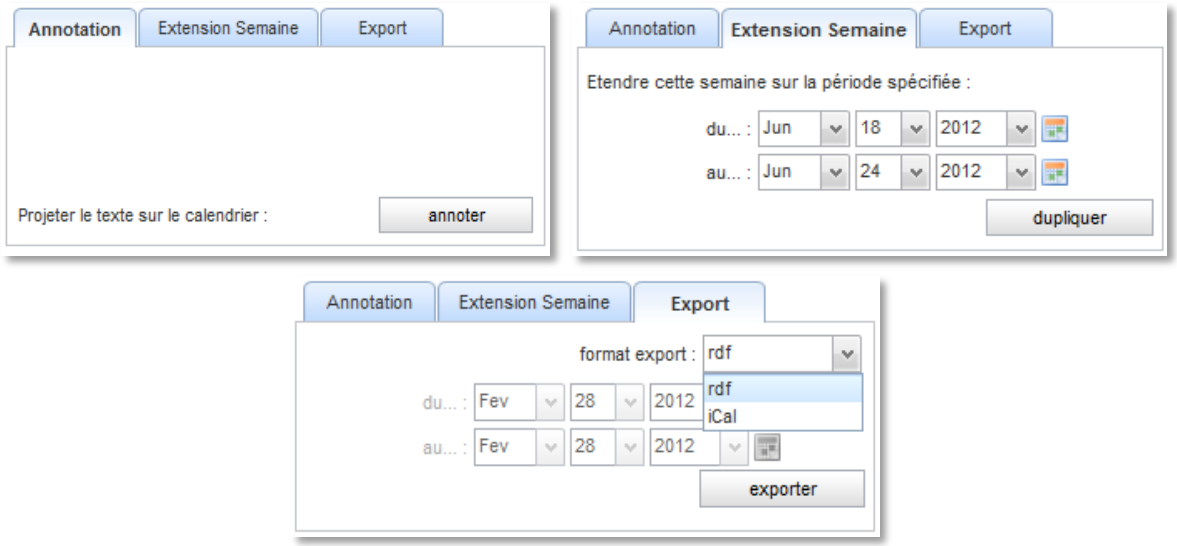


Fig. 58 : panneau de contrôle

Il est possible également de définir les horaires d'accessibilité pour une semaine type, en utilisant l'agenda numérique, et d'étendre ensuite cette semaine sur une fenêtre de temps donnée. Un texte est alors généré qui reflète l'information saisie sur l'agenda. Ce mode de saisie est plus simple

lorsqu'il s'agit par exemple de définir les horaires de séances de cinéma. Voici un exemple de texte généré après une saisie sur l'agenda numérique, étendue ensuite sur une plage d'un mois du 30 avril 2012 au 30 mai 2012 :

> du 30 avril 2012 au 30 mai 2012 :
> - lundi, de 10h à 11h30, de 13h à 14h30
> - mercredi, de 12h à 13h30, de 15h à 16h30.

Ce prototype montre ainsi comment il est possible d'exploiter la sémantique des adverbiaux de localisation temporelle dans un cas d'application industriel. Le système permet de faciliter la saisie d'informations complexes, en évitant aux opérateurs de passer par de nombreux formulaires. Il permet aussi de stocker l'information dans un réseau sémantique réduit et d'exporter ensuite ces connaissances dans d'autres formats (iCalendar notamment) pour interagir avec des systèmes d'indexation et de recherche d'information.

Le système peut s'intégrer dans une chaîne de traitement semi-automatique, où seules les définitions de dates et horaires d'accessibilité dont l'analyse est incohérente feraient l'objet d'une intervention manuelle. La détection des incohérences dans l'analyse devient alors centrale, car elle permet d'isoler deux sous-ensembles, celui des énoncés définissant des périodes d'accessibilité bien transformés et ne nécessitant pas d'intervention manuelle et celui des énoncés mal transformés nécessitant une intervention manuelle. Il s'agit alors de privilégier la précision du système, afin qu'il délivre des informations sûres.

Les incohérences d'analyse peuvent être détectées à différents niveaux : durant la phase d'annotation, durant la phase d'instanciation d'un réseau sémantique ou encore durant la phase de transduction vers le format iCalendar. Nous avons ainsi mis en œuvre quelques règles simples pour détecter de telles incohérences : (1) repérage des portions de texte non annotées contenant des informations calendaires, (2) repérage de structures RDF non-conformes, (3) repérage des sorties iCalendar où des plages horaires se chevauchent.

L'industrialisation du prototype appelle ainsi des développements dans deux directions principales : l'amélioration de la détection des incohérences et l'interfaçage avec une base de connaissances, afin que le système puisse prendre en compte des informations faisant appel à des connaissances externes (pour résoudre les jours fériés variables, pour prendre en compte la notion de saisonnalité ou encore associer des dates précises aux périodes de vacances, par exemple).

6.3 Bilan du chapitre

Nous avons présenté dans ce chapitre une série de composants logiciels permettant d'effectuer des traitements sur les adverbiaux de localisation temporelle. Le composant d'annotation permet de repérer et d'annoter dans des textes en français et en anglais certains des adverbiaux de localisation temporelle (en premier lieu les adverbiaux calendaires et les adverbiaux déictiques). L'annotation s'effectue au moyen de grammaires Unitex. Nous avons également spécifié un langage d'annotation,

ChronolocationML (cf. annexe 2), reflétant le modèle formel qui représente les adverbiaux sous la forme d'une succession d'opérations sémantiques sur une base. A partir des annotations ou métadonnées posées dans les textes, il est possible d'instancier un modèle objet des adverbiaux de localisation temporelle et de les transposer sous la forme d'intervalles calendaires.

On a vu, sur un cas d'utilisation précis, en présentant un système facilitant la saisie de dates et horaires d'accessibilité d'un service, comment cet ensemble d'outils pouvait être utilisé dans une application dédiée à l'enrichissement d'une base de connaissances.

Le chapitre suivant décrit un moteur de recherche expérimental qui s'appuie également sur cet ensemble d'outils pour offrir des services de recherche d'information en mesure de traiter des requêtes combinant des mots-clés et des critères calendaires.

Chapitre 7 : CaSE, un système expérimental pour la Recherche d'Information exploitant les adverbiaux calendaires présents dans les textes

On présente dans ce chapitre le système *CaSE* (*Calendar Search Engine*), un moteur de recherche expérimental permettant d'exprimer des requêtes associant des critères thématiques et des critères calendaires (section 7.1). S'appuyant sur le système d'annotation et de transduction des adverbiaux de localisation temporelle décrit dans la section 6.1, ainsi que sur l'algorithme de filtrage et d'ordonnancement d'intervalles calendaires décrit dans la section 5.2.1, l'expérimentation menée avec ce système de recherche d'information illustre une des façons dont il est possible d'exploiter notre proposition de représentation formelle des adverbiaux calendaires présents dans les textes, pour faciliter la recherche documentaire.

Le système expérimental de recherche d'information que l'on présente dans ce chapitre permet de traiter des requêtes exprimant des critères calendaires « absolus » (*en 1920, depuis la fin du XIXe siècle, dans les années 80*). Ce moteur de recherche permet de traiter potentiellement tout type de corpus et de documents, même si la pertinence d'une recherche temporelle dépend de la nature des textes. Nous présenterons ainsi une expérimentation menée à partir d'un corpus d'articles de Wikipédia.

A travers différentes expérimentations, nous nous attacherons ainsi à montrer l'intérêt de tenir compte des adverbiaux calendaires pour établir la pertinence de documents dans le cadre de la recherche d'information (section 7.2). On montrera également à travers deux expérimentations qu'il est possible de s'appuyer sur la même infrastructure pour interagir non plus seulement avec des textes, mais avec des données structurées auxquelles sont associées des propriétés temporelles (cf. section 7.2.2 et 7.2.3).

7.1 CaSE : un moteur de recherche expérimental

7.1.1 Motivations

Les systèmes actuels de Recherche d'Informations sur le Web, le plus souvent, ne sont pas en mesure de répondre de façon satisfaisante à des requêtes contenant l'expression de critères calendaires (cf. section 3.3). Le critère calendaire pourrait pourtant utilement intervenir dans le calcul de la pertinence d'un document. On souhaite ainsi pouvoir traiter des requêtes comme celles qui suivent :

Ex. 1 : Festival de musique aux alentours de la mi-août
Ex. 2 : gastronomie au début du XIXe siècle
Ex. 3 : peine de mort depuis les années 70

Si les systèmes de bases de données spécialisés peuvent permettre de fournir des réponses à de telles requêtes en filtrant la recherche sur une fenêtre de temps, pour autant elles n'offrent pas de moyen d'ordonner les résultats selon des critères calendaires. En effet, le plus souvent les réponses fournies dans le cadre de ces systèmes sont uniquement celles qui répondent favorablement au test d'inclusion dans la période recherchée. La très grande majorité des systèmes de recherche d'information et de gestion des connaissances qui offrent la possibilité d'exprimer des requêtes temporelles, considère ainsi le critère calendaire comme un filtre, c'est-à-dire comme l'expression d'une fenêtre de temps dans laquelle les résultats doivent cadrer (*time span queries*). Dans ces systèmes, les requêtes calendaires peuvent prendre diverses formes : l'expression d'un intervalle de dates avec une borne de début et de fin, ou encore la sélection d'une facette temporelle (permettant par exemple de préciser que les documents recherchés doivent avoir été publiés *dans les dernières 24 heures*, *il y a moins d'une semaine*, *dans le mois*, etc.). Formellement, chacune de ces approches est identique et consiste à filtrer des documents et non pas à les trier suivant des critères calendaires. Il s'agit ainsi de ne garder que les documents (ou les données) auxquelles les propriétés temporelles associées sont incluses ou qui au moins chevauchent l'intervalle de dates associé à la requête ou au filtre. En outre, ces systèmes n'exploitent pas les adverbiaux calendaires présents dans les documents, mais plus souvent des métadonnées associées aux documents (comme leur date de publication).

A l'exception des travaux de (Arikan *et al.*, 2009) et (Berberich *et al.*, 2010) (cf. section 3.3.2), la mise en relation entre les critères calendaires exprimés dans une requête et les « expressions temporelles » présentes dans les documents n'est donc pas, le plus souvent, regardée comme un potentiel critère de pertinence ou de tri dans les systèmes de recherche d'information. Ces derniers travaux toutefois ne prennent en compte que les *dates* présentes dans les textes, sans analyser les opérations sémantiques qui opèrent dans les adverbiaux calendaires : les deux adverbiaux « *en 1906* » et « *depuis 1906* » sont ainsi analysés de la même manière, réduits à leur référence temporelle noyau, l'année 1906. On souhaite ainsi montrer qu'une approche linguistique plus fine doit permettre de progresser sur le terrain de la recherche d'information temporelle. Il s'agit, par exemple, de tenir compte de la sémantique associée à la préposition « depuis » dans l'adverbial « *depuis 1906* », représentée, dans le modèle que l'on propose, comme une *Opération de Régionalisation* : la valeur associée à cette préposition influe ainsi sur le résultat d'une recherche.

Ceci est rendu possible par l'heuristique de transduction présentée dans le chapitre 5 (cf. section 5.2.1) : à chaque valeur d'une opération sémantique est associée une transformation particulière, aussi l'intervalle calendaire associé à l'adverbial « *depuis 1906* » est-il différent de l'intervalle calendaire associé à l'adverbial « *en 1906* ».

L'expérience illustre en outre l'intérêt d'une approche fondée sur la notion de pertinence ou de similarité temporelle. On a vu dans la section 5.3 qu'il était possible d'associer des scores de pertinence pour des ensembles d'intervalles calendaires-cibles par rapport à un intervalle-source. Cette mesure de pertinence peut être combinée avec d'autres critères pour établir la pertinence globale d'un document ou d'un fragment de document. On peut, en particulier, la combiner avec la mesure de pertinence associée aux mots-clés, afin de pouvoir traiter des requêtes à la fois thématiques et calendaires.

On rappelle que, compte-tenu du modèle de pertinence adopté, qui compare deux à deux des intervalles calendaires, le système ne prend en charge que des requêtes dont le critère temporel désigne une zone unique du calendrier : en effet, l'algorithme de comparaison entre intervalles calendaires n'est pas en mesure de comparer des ensembles d'intervalles tels que ceux que l'on peut associer aux adverbiaux dénotant plusieurs zones du calendrier (« *tous les lundis* », «*en 1980 et en 1987* »). Le système peut donc traiter des requêtes telles que « *jusqu'au début du XXe siècle* » ou « *le 10 août 1852* », mais pas « *tous les lundis au mois d'octobre* ».

7.1.2 Fonctionnalités

Le système CaSE est un outil de recherche documentaire (un moteur de recherche), qui permet de filtrer et d'ordonner par pertinence des documents (des pages Web, par exemple). La mesure de la pertinence dépend à la fois des mots-clés et des critères calendaires exprimés dans les requêtes.

CaSE permet également d'explorer un document (recherche *intra*-documentaire). L'utilisateur peut ainsi sélectionner un document de la liste des résultats proposés par le système et restreindre sa recherche au seul contenu de ce document. De ce point de vue, le système CaSE répond à l'une des limitations des moteurs de recherche actuels, qui permettent de trouver des documents pertinents, mais pas d'explorer leur contenu (Couto, 2006 ; Couto et Minel, 2006 ; Bilhaut, 2006).

Les requêtes soumises sur la page d'accueil du système peuvent être de trois types : (1) elles peuvent combiner des mots-clés et un critère calendaire (par exemple, *laïcité avant 1905*), (2) ou bien ne contenir que des mots-clés - ce qui permet d'observer à quelles périodes de temps le critère thématique est associé dans le corpus indexé (par exemple, *laïcité*), (3) ou bien encore ne contenir que l'expression d'un critère calendaire (par exemple, *avant 1905*) – ce qui permet de retrouver dans un corpus ou un document toutes les références pertinentes proche du critère calendaire. On verra qu'il est également possible d'affiner sa requête en interagissant avec une frise chronologique.

L'interface est divisée en trois parties (cf. fig. 59) : (1) une zone pour la saisie des requêtes (cf. fig. 60), (2) une zone pour la visualisation et la navigation sur une frise chronologique (cf. fig. 61 et 62) et (3) une zone où est affichée la liste ordonnée des résultats (cf. fig. 63).

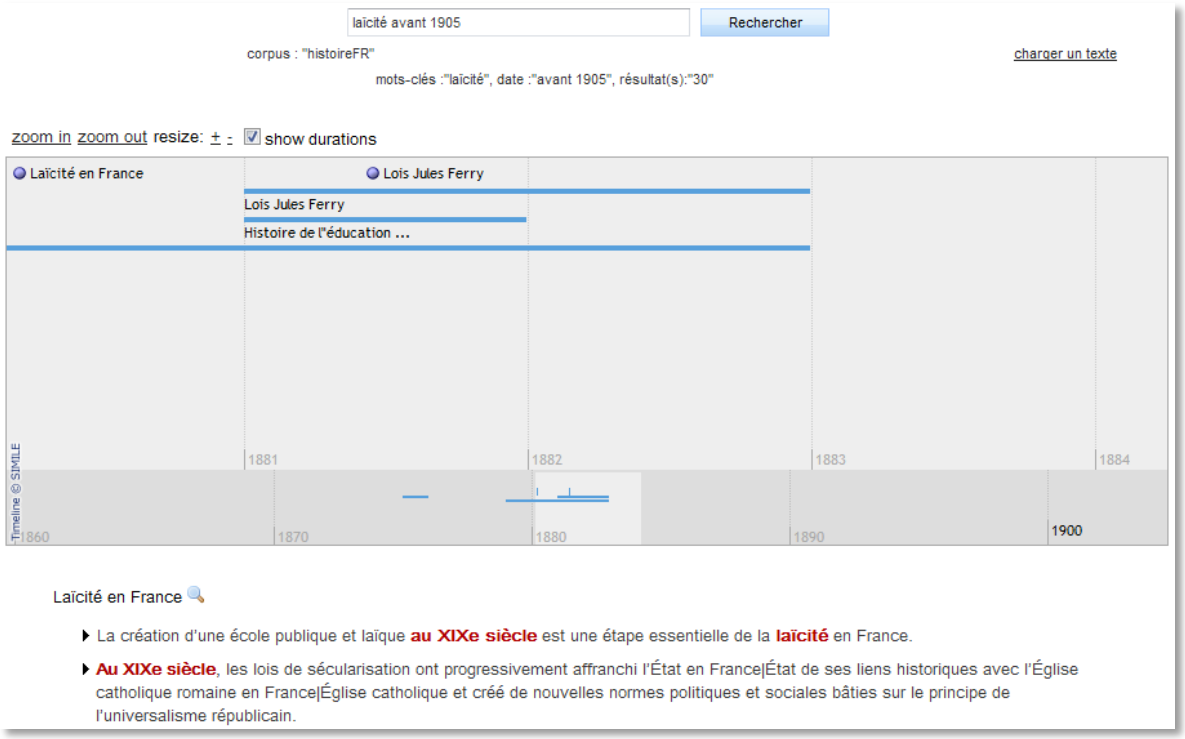


Fig. 59 : l'interface du système CaSE (requête « laïcité vers 1905 »)

(1) La zone de saisie des requêtes

Fig. 60: la zone de saisie des requêtes

Cette zone contient un champ pour la saisie des requêtes, sous lequel sont affichées plusieurs informations :

- le nom du corpus exploré,
- l'analyse de la requête, qui sépare la partie thématique d'une requête (les mots-clés) et l'éventuel critère calendaire,
- le nombre de résultats retournés pour une requête donnée

(2) La zone de visualisation et de navigation sur une frise chronologique

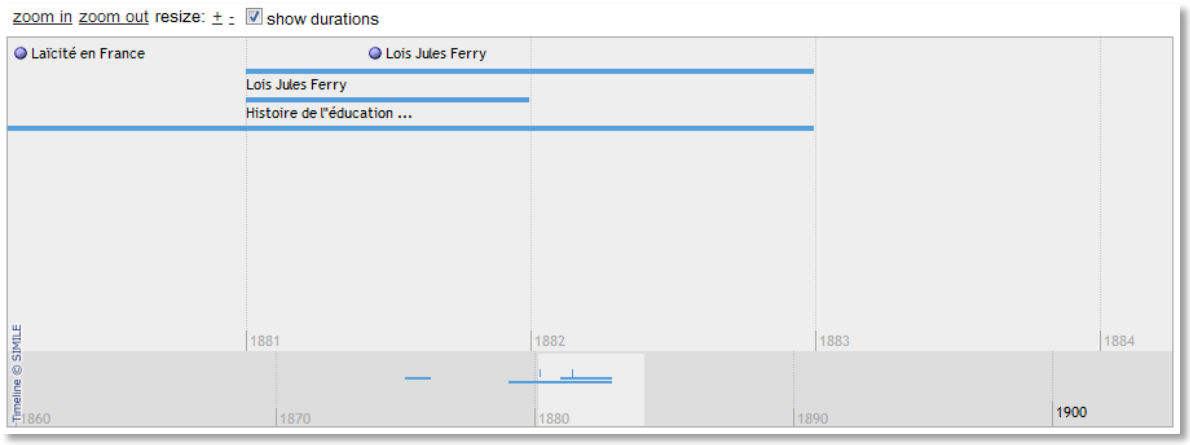


Fig. 61 : visualisation chronologique des résultats d'une recherche

Cette zone présente une frise chronologique et les différentes fonctionnalités de paramétrage qui l'accompagnent. Il est ainsi possible de paramétrer l'échelle de la frise chronologique, sa taille sur l'écran, ainsi que le mode de présentation des éléments figurants sur la frise. L'utilisateur peut ainsi choisir d'afficher des durées ou des points pour représenter les adverbiaux calendaires repérés dans les documents. Les adverbiaux peuvent ainsi être représentés par des lignes continues (ce qui peut parfois surcharger la frise chronologique) ou bien par des points (dans ce cas, ce sont les *pôles* associés aux adverbiaux qui servent d'ancre (sur cette notion de *pôle*, on renvoie à la section 5.3.5).

L'interface permet de sélectionner un résultat sur la frise. Une fenêtre s'affiche alors (cf. fig. 62).

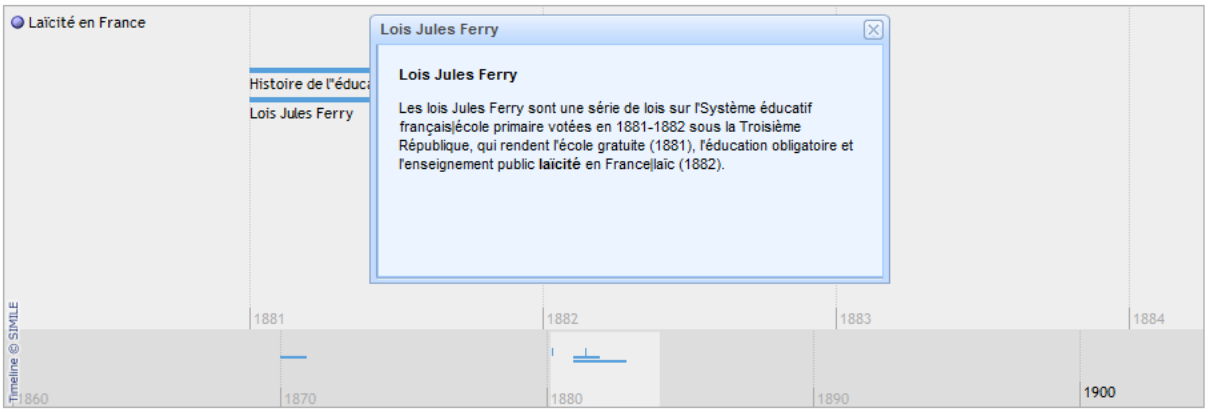


Fig. 62: visualisation chronologique des résultats d'une recherche : la fenêtre d'affichage d'un résultat

Il est possible également de se déplacer sur la frise chronologique, ce qui a pour effet de relancer automatiquement des requêtes sur le serveur. Les résultats de la frise et de la liste des documents sont alors mis à jour à la volée.

Les interactions avec la frise chronologique permettent ainsi d'affiner les critères calendaires en fonction des résultats obtenus. Les déplacements sur la frise déclenchent également la génération automatique d'un adverbial calendaire associé à la fenêtre de temps visualisée sur la frise

chronologique. Cet adverbial vient remplacer le critère qui avait pu être initialement exprimé dans la requête : on illustre ce comportement dans le scenario d'utilisation décrit dans la section 7.2.1. Ceci permet de préserver la cohérence des liens entre les termes de la requête et les résultats affichés.

(3) La zone d'affichage de la liste des résultats

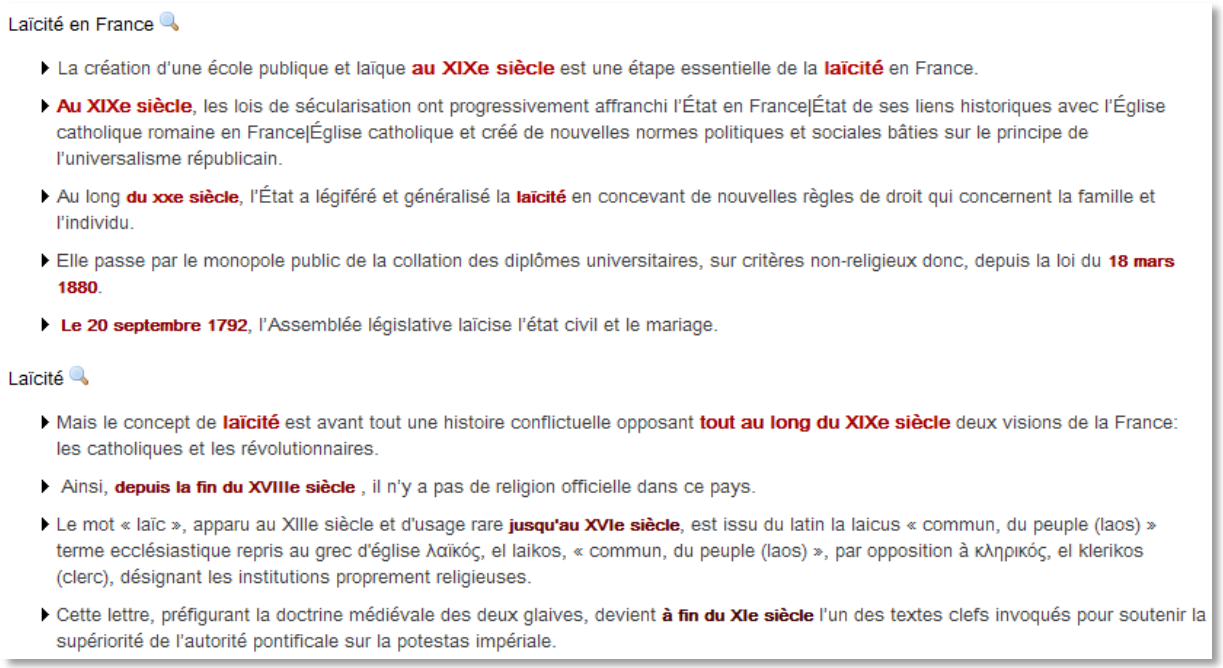


Fig. 63 : la zone d'affichage de la liste des résultats

Cette zone présente une liste de documents ordonnés par pertinence. Cette liste est paginée. Pour chaque document de la liste, le système affiche :

- son titre (un lien hypertexte vers le document d'origine),
- une loupe (qui permet de basculer sur le mode d'exploration intra-documentaire),
- une sélection des extraits les plus pertinents du document (le nombre de ces extraits est paramétrable).

A la place des courts extraits de documents (*snippets*) fournis en général par les moteurs de recherche, le système affiche ainsi les phrases les plus pertinentes extraites d'un document.

Lorsque le système bascule en mode d'exploration *intra-documentaire*, les résultats sont présentés sous la forme d'une liste de phrases ordonnées par pertinence (au lieu de la liste ordonnée de documents pour la recherche documentaire). Les interactions proposées pour la recherche documentaire sont également disponibles pour la recherche intra-documentaire : recherche par mots-clés et/ou critères calendaires et navigation à l'aide de la frise chronologique. La recherche intra-documentaire est à rapprocher de la navigation textuelle, dont l'objectif est de permettre à l'utilisateur/lecteur de parcourir un texte selon différents points de vue (Couto, 2006 ; Couto et Minel, 2006). Ici toutefois, la navigation est pensée dans sa continuité avec la recherche documentaire : il s'agit d'un service Web (non d'un logiciel dédié) qui permet, une fois un document

sélectionné, de le parcourir sous l'angle temporel, avant éventuellement d'accéder au contenu intégral du document.

En revanche, en l'état actuel des développements, le moteur ne permet pas de naviguer dans un document en gardant en vue l'intégralité du texte, puisqu'il fonctionne par extraction de phrases. A cet égard, les fonctionnalités du système pourraient être enrichies, afin de répondre aux problématiques propres à la navigation textuelle, en particulier à la question de l'accès en contexte aux informations pertinentes. L'architecture technique est cependant compatible avec ce type d'approche : il serait par exemple possible de mettre en place une visualisation qui surlignerait les passages les plus pertinents d'un document plutôt que de les extraire.

7.1.3 Architecture

Le système CaSE, intégralement développé en Java, est formé de deux composants : un moteur d'indexation et un service Web de recherche. L'indexation et la recherche sur l'index s'appuient sur les librairies standard de Lucene[45]. En ce sens, le moteur peut être présenté comme une surcouche ou une extension de Lucene.

Lucene est un moteur de recherche open-source développé en Java qui permet d'indexer et d'effectuer des recherches dans des textes numériques. Lors de la phase d'indexation, Lucene établit un index destiné à faciliter ensuite l'accès à ces documents et à leur contenu : cet index consiste en une liste de descripteurs auxquels est associée une liste de documents ou de parties de documents. Durant l'étape d'indexation, Lucene effectue plusieurs traitements d'analyse linguistique. L'une d'elle consiste à réduire les mots apparaissant dans un document à leur racine (en retirant les flexions de nombre et de genre ou en réduisant un verbe conjugué à son radical, par exemple). Cette opération, appelée stemming, permet d'établir la fréquence des termes d'un document, en cumulant leurs occurrences, indépendamment de leurs variations morphosyntaxiques. L'idée sous-jacente est qu'un terme qui apparaît souvent dans un texte représente un concept important. Chaque document ou chaque passage de document est ainsi représenté sous la forme d'un vecteur, dont les coordonnées représentent les fréquences des termes.

L'indexation automatique cherche donc à représenter les documents en fonction des termes qui semblent représenter le mieux leur contenu informationnel. Pour ce faire, Lucene recourt à une mesure classique de l'importance d'un terme contenu dans un document. Cette mesure repose sur une formule de pondération nommée TF-IDF (Term Frequency-Inverse Document Frequency). Cette mesure statistique permet d'évaluer l'importance d'un terme contenu dans un document, par rapport à sa distribution dans le corpus indexé. Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document ; il varie également en fonction de la fréquence du mot dans le corpus.

Lors de la phase de recherche d'information, le système effectue des rapprochements entre la requête d'un usager et le contenu de l'index pour établir une liste de réponses ordonnées par pertinence. Pour mesurer la pertinence d'un document, Lucene compare les termes de la requête

[45] http://lucene.apache.org/core/

avec ceux des documents. Si une requête contient le terme T, un document a d'autant plus de chances d'y répondre que la fréquence du terme au sein du document (TF) est grande. Toutefois, si le terme T est lui-même très fréquent au sein du corpus, c'est-à-dire qu'il est présent dans de nombreux documents, il est considéré comme étant peu discriminant. La pertinence d'un terme augmente ainsi en fonction de sa rareté au sein du corpus. La présence dans un document d'un terme recherché qui est rare dans le corpus indexé fait ainsi croître le « score » de ce dernier.

Notre objectif a donc été de compléter cette mesure de pertinence d'un document par une autre mesure, qui tient compte de la similarité entre un critère calendaire exprimé dans une requête et les adverbiaux calendaires présents dans un document. Nous avons ainsi développé deux extensions à Lucene : un composant d'indexation (Indexer) et un composant pour la recherche d'information temporelle (Searcher) (cf. fig. 64).

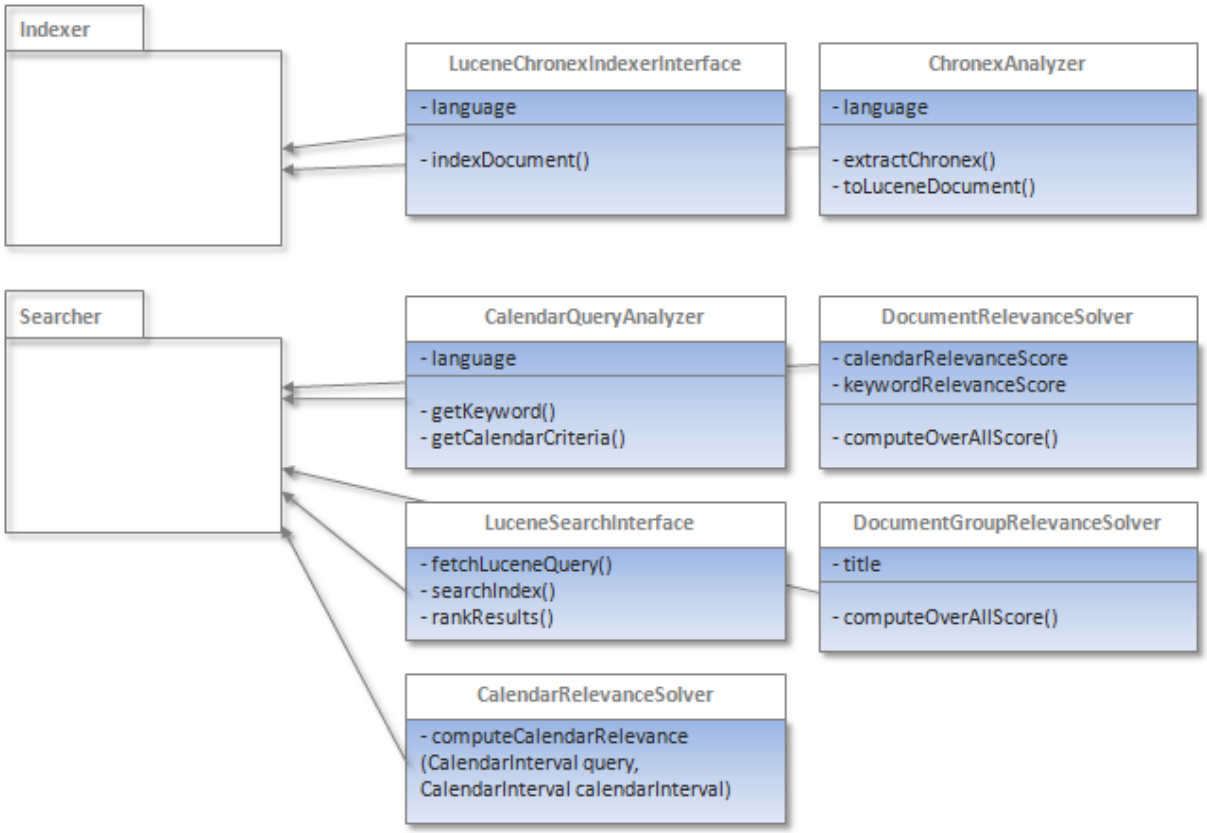


Fig. 64 : les composants d'indexation et de recherche

(1) Le moteur d'indexation

Le composant d'indexation appelle le système d'annotation présenté dans la section 6.1 pour extraire, dans un document, les phrases contenant des adverbiaux calendaires. Les adverbiaux annotés sont transposés sous la forme d'intervalles calendaires. La segmentation des textes en phrases est opérée par Unitex, lors de la phase d'annotation des textes.

Pour chacun des documents indexés différents champs sont renseignés dans l'index généré : le titre du document, son url, la phrase où apparaît l'adverbial, l'unité textuelle correspondant à l'adverbial, l'intervalle calendaire qui lui a été associé, le pôle attribué à cet intervalle, ainsi que les annotations qui représentent la sémantique de l'adverbial sous la forme d'une succession d'opérations sur une base calendaire. Pour le traitement des mots-clés, les documents indexés sont analysés avec les outils linguistiques de Lucene (*FrenchAnalyzer* pour le français et *StandardAnalyzer* pour l'anglais) : ces outils découpent les unités textuelles en *tokens* (séparant les mots, les signes de ponctuation et les chiffres) et réduisent à leur radical les mots présents dans les documents.

(2) Le moteur de recherche

Le moteur de recherche est un service Web (un module GWT[46]). Il repose sur une architecture client-serveur : le client contient l'IHM (interface homme-machine) et le serveur exécute les traitements pour l'analyse des requêtes, l'interrogation de l'index, la sélection et le tri par pertinence des résultats.

L'IHM contient différents *widgets* qui correspondent aux zones décrites plus hauts (zone de saisie, zone de visualisation et de navigation sur une frise chronologique et zone d'affichage de la liste des résultats). La frise chronologique est un composant Javascript open-source : SIMILE timeline[47].

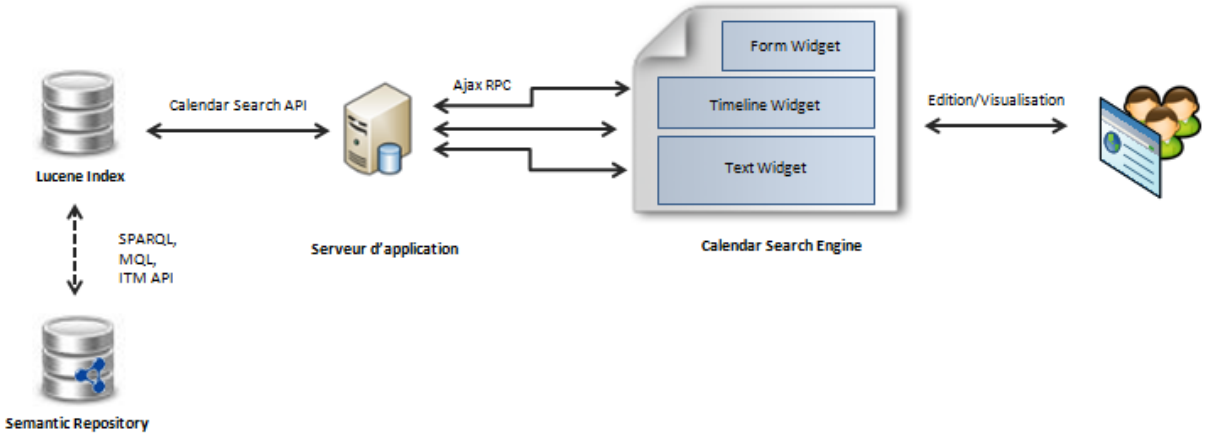


Fig. 65 : architecture du service Web de recherche

Lorsqu'une requête est soumise au système CaSE, le serveur appelle le système d'annotation, afin de récupérer l'éventuel critère calendaire qui a pu être exprimé dans la requête. Il génère ensuite une série de requêtes sur l'index qui permettent d'effectuer un premier filtrage des résultats.

L'ensemble des résultats ainsi obtenu est ensuite ordonné par pertinence. La hiérarchisation des résultats tient compte du score de pertinence associé aux mots-clés (ce score est fourni par Lucene) et du score associé aux adverbiaux calendaires. Ce dernier score est obtenu grâce au composant *CalendarRelevanceSolver*, qui implémente l'algorithme de filtrage et d'ordonnancement des adverbiaux calendaires présenté dans la section 5.3.

[46] https://developers.google.com/web-toolkit/
[47] http://www.simile-widgets.org/timeline/

7.1.4 Filtrage des documents

Le processus de recherche se déroule donc en deux étapes successives :

1. Filtrage des K plus proches voisins de la requête dans l'index, par rapport au critère thématique et au critère calendaire,
2. Ordonnancement des résultats filtrés en fonction de leur pertinence relativement à la requête.

Ainsi, dans l'architecture retenue, le calcul du score de la pertinence temporelle est opéré après une première étape de filtrage, car il nécessite une comparaison deux à deux des intervalles calendaires. Pour éviter des traitements coûteux qui demanderaient de balayer tout l'index pour comparer les informations temporelles stockées et la requête, le parti pris consiste donc à filtrer d'abord les résultats en fonction du pôle associé aux intervalles calendaires. Pour rappel (cf. section 5.3.5), un *pôle* est attribué à chaque intervalle calendaire. Par exemple, pour l'adverbial « *depuis les années 80* », le pôle correspond au début de l'intervalle qui lui est associé ; pour l'adverbial « *jusqu'en mars 2007* », le pôle correspond à l'intervalle qui lui est associé (la fin du mois de mars 2007). Le filtrage dans l'index ne tient donc compte que de la date associée à ce pôle.

Le filtrage permet de récupérer dans l'index les K plus proches voisins de la requête (cf. fig. 66) (soit les K résultats a priori les plus intéressants), où K correspond au nombre de résultats maximum susceptibles d'être traités dans un temps raisonnable : ce nombre peut donc varier en fonction des performances de la machine sur laquelle le système fonctionne. Dans les expériences menées, ce nombre est fixé à mille. Pour obtenir le nombre de résultats attendu, on règle progressivement un intervalle autour du pôle attribué à l'intervalle associé à la requête.

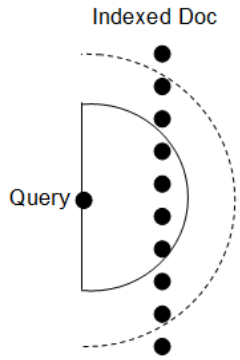


Fig. 66 : filtrage des plus proches voisins dans l'index

L'étape suivante consiste à évaluer la pertinence des résultats issus de l'étape de filtrage.

7.1.5 Mesure de la pertinence des documents

La mesure de la pertinence des documents fait intervenir plusieurs modèles de pertinence associés entre eux afin d'obtenir un score final.

(1) Mesure de la pertinence pour des requêtes combinant des mots-clés et un critère calendaire

Pour les requêtes contenant un critère calendaire et un critère thématique, la pertinence d'une phrase candidate P_i dépend du score associé aux mots-clés et du score associé à l'adverbial calendaire extrait dans la phrase (noté $scoreCal$).

Le modèle de pertinence pour les adverbiaux calendaires présenté dans le chapitre 5 (cf. section 5.3) permet d'ordonner un ensemble d'adverbiaux-cibles par rapport à un adverbial-requête : plus précisément, il permet d'attribuer un score de pertinence à chaque intervalle associé à un adverbial. Le moteur Lucene attribue pour sa part un score de pertinence, qui est lui fonction des mots-clés présents dans la phrase.

Considérons un ensemble de phrases filtrées par le moteur de recherche :

$$P = \{P_1, P_2, \dots, P_n\}$$

L'objectif est d'ordonner les éléments de P du plus pertinent au moins pertinent compte-tenu de la requête soumise au moteur de recherche. Les scores de pertinence attribués pour les mots-clés sont divisés en deux classes, (1) l'une correspondant aux résultats qui contiennent tous les mots-clés, (2) l'autre aux résultats qui n'en contiennent qu'une partie seulement.

On obtient ainsi deux sous-ensembles de P, qui sont alors ordonnés en fonction du score associé aux adverbiaux calendaires. Pour les phrases du premier ensemble, la pertinence d'une phrase P_i équivaut à $scoreCal$.

$$Pert(P_i) = scoreCal.$$

Pour les résultats appartenant au second ensemble, ce score est minoré par un facteur α, compris entre 0 et 1. La pertinence d'une phrase P_j appartenant à cette catégorie est ainsi :

$$Pert(P_j) = \alpha \times scoreCal.$$

Nos expérimentations nous ont conduits à fixer α à 0,02, soit un facteur de pondération faible destiné à minimiser le score des phrases peu pertinentes sous l'angle des mots-clés. Cette façon très empirique de faire dépendre la mesure de la pertinence d'une phrase du score associé aux mots-clés est rudimentaire en l'état et demanderait à être affinée. Ce modèle de pertinence est celui utilisé pour l'exploration d'un texte : le système présente une ainsi liste ordonnée de phrases pour une requête restreinte à un seul texte (cf. le scenario d'utilisation décrit dans la section 7.2.1).

(2) Mesure de la pertinence pour des requêtes ne contenant que des mots-clés

Pour mesurer la pertinence des phrases par rapport à des requêtes qui ne contiennent que des mots-clés, le système filtre un ensemble de phrases P, qui correspond à l'ensemble des K phrases les plus pertinentes sous l'angle des mots-clés. Cet ensemble P est ordonné selon la fréquence d'apparition des adverbiaux calendaires qu'il contient. On obtient ainsi une liste des adverbiaux calendaires qui sont le plus fréquemment associés aux mots-clés de la requête.

Par exemple, pour le corpus d'articles de Wikipédia relatifs à l'histoire de France que l'on présente plus bas dans la section dédiée aux expérimentations autour du système Case (cf. section 7.2.1.1), aux mots-clés « *élection présidentielle* » et « *Saint-Bartélemy* », le système associe les listes d'adverbiaux calendaires suivantes :

```
élection présidentielle ---> en 2002 (fréquence : 40)
élection présidentielle ---> en 1965 (fréquence : 29)
élection présidentielle ---> 1981 (fréquence : 27)
élection présidentielle ---> en 2007 (fréquence : 26)
élection présidentielle ---> Le 21 avril 2002 (fréquence : 24)


saint-barthélemy ---> le 24 août 1572 (fréquence : 19)
saint-barthélemy ---> En 1572 (fréquence : 15)
saint-barthélemy ---> (1553-1615) (fréquence : 5)
saint-barthélemy ---> durant l'été 1610 (fréquence : 5)
saint-barthélemy ---> fin août 1572 (fréquence : 4)
```

Le premier élément de la liste ainsi obtenue (*en 2002*, dans le premier exemple, *le 24 août 1572*, dans le second) est alors considéré comme le critère calendaire de la requête. Ceci revient à considérer que la valeur calendaire la plus fréquemment associée à une requête permet de la désambiguïser. Le système peut alors produire une requête combinant à la fois des mots-clés et un critère calendaire et appliquer la mesure de pertinence décrite précédemment. Dans les exemples ci-dessus, les requêtes générées sont ainsi « *élection présidentielle en 2002* » et « *Saint-Barthélemy le 24 août 1572* ».

Remarquons que cette liste de fréquence pourrait être présentée à l'utilisateur afin qu'il ait lui-même la possibilité de désambiguïser sa recherche : elle pourrait en effet lui permettre d'affiner sa requête initiale, sous la forme d'une liste de suggestions[48].

[48] C'est du reste à ce type d'informations que s'est intéressée jusqu'à présent la plupart des équipes travaillant sur la problématique temporelle dans le cadre de la recherche d'information. Le service de Google view:timeline, par exemple, permettait de visualiser la fréquence d'apparition d'un ensemble de mots-clés à côté d'une date. On renvoie à ce sujet aux travaux mentionnés dans la section 3.3.1, notamment aux travaux sur l'analyse temporelle des requêtes qui permettent d'en distinguer différents profils (Diaz et Jones, 2005 ; Asur et Buehrer, 2009 ; Chen *et al.*, 2011) : bien qu'ils analysent les archives des requêtes soumises aux moteurs de recherche (*logs*) plutôt que des textes, la démarche consiste bien à analyser les fréquences de cooccurrences entre des mots-clés et des expressions temporelles.

(3) Des phrases aux textes : mesure globale de la pertinence d'un document

Jusqu'ici, le système associe des scores de pertinence à un ensemble P de phrases. Dans le cadre d'une recherche documentaire, ces scores de pertinence doivent être associés, afin d'obtenir un score de pertinence global pour chaque document. On cherche donc à ordonner un ensemble de documents :

$$D = \{Doc_1, Doc_2, \dots, Doc_n\}$$

A ce stade, les documents sont représentés comme des ensembles de phrases ordonnées par pertinence : $Doc_i = \{P_{i_1}, P_{i_2}, \dots, P_{i_{m-1}}, P_{i_m}\}$, où P_{i_1} correspond à la phrase ayant obtenue le meilleur score de pertinence et P_{i_m} le plus faible.

En première approche, le score de pertinence global d'un document Doc_i correspond au score de pertinence de P_{i_1} $(Pert(P_{i_1}))$, soit la phrase qui a obtenu le meilleur score de pertinence. Afin de valoriser les documents contenant plusieurs phrases pertinentes, on associe à ce score la somme pondérée de l'ensemble des autres phrases du document de la façon suivante :

$$Pert(Doc_i) = Pert(P_{i_1}) + \sum_{j=2}^{m} \left(1 + Pert\left(P_{i_j}\right)\right) \times \varepsilon$$

ε correspond au symbole introduit dans la section 5.2.1.3 – il représente une valeur supérieure à 0, mais inférieure à l'ensemble des autres nombres positifs. Le score ainsi obtenu permet de valoriser un document Doc_i par rapport à un document Doc_j, lorsque $Pert\ (P_{i_1}) = Pert\ (P_{j_1})$, mais que la somme pondérée des autres phrases de Doc_i produit une valeur supérieure à celle de Doc_j.

Par exemple, pour la requête « *abolition de l'esclavage* », le système retourne un ensemble de documents dont les deux suivants :

Doc_a « Décret d'**abolition de l'esclavage du 27 avril 1848** » :

> P_{a_1} : Le deuxième décret d'**abolition de l'esclavage** en France a été signé **le 27 avril 1848** par Lamartine.
> P_{a_2} : **Le 4 mars 1848**, le décret, rédigé par Schoelcher, abolissant **l'esclavage** et créant la Commission d'**abolition de l'esclavage** chargée de préparer l'émancipation, est signé par le gouvernement provisoire de la toute jeune République.

Doc_b « Alsace » :

> P_{b_1} : Victor Schoelcher, homme de gauche d'origine alsacienne, est nommé président de la commission d'**abolition de l'esclavage**, il est l'initiateur du décret **du 27 avril 1848** abolissant définitivement **l'esclavage** dans l'empire colonial français.

Dans cet exemple, $\mathrm{Doc_a}$ est plus pertinent que $\mathrm{Doc_b}$, bien que les phrases $\mathrm{P_{a_1}}$ et P_{b_1} soient de pertinence égale.

Le facteur de pondération ε permet d'éviter de valoriser un document Doc_i par rapport un document Doc_k, lorsque $Pert\ (P_{i_1})$ est inférieur à $Pert\ (P_{k_1})$ de façon significative.

On obtient ainsi un score de pertinence global pour chaque document : il est alors possible de les ordonner du plus pertinent au moins pertinent. C'est sur ce modèle que s'appuie le système CaSE dans le cas d'une recherche documentaire.

7.1.6 Le problème de la visualisation des résultats sur une frise chronologique

Le choix d'une présentation des résultats sur une frise chronologique a découlé presque naturellement de la mise place la chaîne d'indexation et de recherche, le système permettant de disposer des informations nécessaires à ce type d'affichage, à savoir une liste d'intervalles calendaires répondant à une requête. Cependant, des problèmes propres à ce type d'outils de visualisation et aux contraintes de l'outil choisi (SIMILE Timeline) ont fait surface, enrichissant et complexifiant la question de la recherche d'information temporelle.

En effet, au niveau de l'implémentation, les données nécessaires à l'affichage sur la frise chronologique SIMILE Timeline (exprimées en Javascript) sont gérées du côté du client, c'est-à-dire par le navigateur de l'utilisateur. Le nombre de résultats doit donc être limité, car plus ils sont nombreux, plus le temps de chargement est long. De façon plus générale, pour qu'une frise chronologique demeure lisible, il faut éviter de présenter un nombre trop important de résultats.

Cette limitation fait qu'il n'est pas possible de charger l'ensemble des résultats retournés par le système de recherche d'information : on limite donc l'affichage aux K résultats les plus pertinents, où K correspond au maximum de résultats susceptibles d'être affichés et chargés en un temps raisonnable (environ 150). Pour contourner cette contrainte limitative, un mécanisme dynamique de génération de requêtes a été mis en œuvre, afin de pouvoir charger de nouveaux résultats à mesure que l'utilisateur se déplace sur la frise.

Par ailleurs, ce type de visualisation présente deux difficultés : (1) celle du choix de l'échelle de temps (dans les expériences menées, elle est calculée dynamiquement en fonction de la durée la plus fréquente parmi les 20 premiers résultats) et (2) le choix de la date où positionner le centre de la frise : il ne peut être fonction uniquement de la requête, car si les résultats proposés à l'utilisateur ne correspondent pas exactement à la requête, ils risquent de ne pas apparaître sur la frise. Cette date est donc fonction du résultat le plus pertinent. Enfin, il est parfois nécessaire de masquer certains résultats qui ne sont pas pertinents au regard de l'échelle sélectionnée, parce qu'ils sont d'une granularité supérieure.

7.1.7 Quelques-unes des limites du système

A ce stade, le prototype développé présente des limites à plusieurs niveaux dont nous listons ici les principales :

(1) Les ressources du système d'annotation qui analysent les critères calendaires exprimés dans les requêtes ont été conçues initialement pour l'annotation des textes. Or le fonctionnement des moteurs de recherche inclinent les utilisateurs à saisir des mots-clés et non des expressions en langage naturel : par exemple, un utilisateur saisira plus naturellement « *1997* » que « *en 1997* ». Il faudrait donc prévoir des ressources spécifiques pour l'annotation des requêtes. Ces ressources pourraient au demeurant être développées à l'aide d'outils plus simples qu'Unitex (reposant, par exemple, sur des expressions régulières) : exprimés sous la forme de mots-clés, les critères calendaires présentent moins de variations que les adverbiaux calendaires. Ceci serait d'autant plus pertinent que l'annotation des requêtes avec Unitex affecte de façon significative les temps de réponse du système CaSE.

(2) Par ailleurs, les ressources du système d'annotation, comme on l'a vu, ne transposent pas tous les adverbiaux de localisation temporelle vers des intervalles calendaires. A ce stade, seuls les adverbiaux dont la base est calendaire sont ainsi pris en compte pour l'indexation. Cependant, le système CaSE est indépendant des ressources dédiée à l'annotation : tout autre système d'annotation en mesure d'associer des intervalles calendaires à unités textuelles peut être exploité. Une expérience a été menée en ce sens avec un système d'annotation des expressions temporelles en espagnol décrit dans (Etcheverry Méndes, 2010)[49].

(3) Dans sa version actuelle, le système indexe des fragments de textes, en l'occurrence des phrases, et non des textes entiers. Les textes ne sont « reconstitués » qu'au moment de la mise en forme des résultats : les phrases les plus pertinentes sont alors regroupées en fonction du titre du document qui leur est associé. Ceci permet de présenter à l'utilisateur des phrases plutôt que des extraits courts non contigus et permet aussi à l'utilisateur d'effectuer des recherches intra-documentaires. Cette approche répond au problème de la lisibilité des résultats proposés, mais surtout permet de limiter la recherche des mots-clés à un contexte voisin de celui où apparaît un adverbial calendaire. Cependant, cette heuristique ne garantit pas que l'adverbial et le critère thématique soient liés. Par exemple, pour la requête « *chocolat au 18e siècle* », on obtient parmi d'autres le résultat suivant qui n'est pas pertinent, dans la mesure où l'adverbial « en 1720 », qui a contribué à faire remonter cet extrait parmi les résultats, n'est pas lié syntaxiquement au terme « chocolaterie » :

> La ville fut également le lieu de création des usines de sirop Teisseire **en 1720**, de la fabrique de pâtes Lustucru en 1824, de la **chocolaterie** Cémoi en 1920.

Il s'agit donc là d'une simplification rudimentaire du problème complexe, du point de vue de la linguistique textuelle et de l'ingénierie des langues, du calcul de la portée des adverbiaux, c'est-à-dire de la façon dont ils prennent part à la structuration du discours (Van Raemdonck, 2001 ; Charolles et Vigier, 2005 ; La Draoulec et Pery Woodley, 2005).

[49] Ce projet a été réalisé dans le cadre d'une coopération inter-universitaire avec l'institut d'informatique InCO de l'Université de Montevideo (programme Ecos-Sud 28 80).

7.2 Expérimentations autour du système CaSE

Nous décrivons ici trois expérimentations menées autour du système CaSE, chacune visant à illustrer différents aspects des possibilités ouvertes par cette approche de la recherche d'information temporelle :

(1) La première de ces expérimentations a été menée sur des corpus de textes (essentiellement des corpus d'articles de Wikipedia) (cf. section 7.2.1). L'objectif était de montrer comment les moteurs de recherche pourraient tirer parti d'une représentation de la sémantique des adverbiaux calendaires qui découle d'une analyse linguistique pour mettre en œuvre des services dédiés à la recherche documentaire ou intra-documentaire en mesure de traiter des requêtes temporelles.
(2) La seconde expérimentation a été menée sur des données structurées décrivant des événements liés à l'univers de la musique (cf. section 7.2.2). L'objectif était de montrer que le système peut interagir avec des données structurées exprimées dans les formats propres au Web Sémantique.
(3) Enfin, la troisième expérimentation a été menée sur des données structurées provenant de Freebase relatives à des œuvres d'art (cf. section 7.2.3). L'objectif était de montrer qu'il peut être intéressant de mettre en œuvre des systèmes qui permettent aux utilisateurs d'enrichir des bases de connaissances à mesure qu'ils les consultent, tirant parti à la fois de ressources structurées au sens du Web Sémantique et de ressources textuelles.

7.2.1 La recherche documentaire et l'exploration intra-documentaire avec CaSE

Cette expérimentation a pour but de dérouler un scénario d'utilisation du système CaSE qui articule la recherche documentaire et la recherche intra-documentaire. Elle vise à montrer l'intérêt et la faisabilité de la démarche adoptée pour mettre en œuvre un système de recherche d'information temporelle. On rappelle que le système ne peut traiter que des requêtes temporelles qui couvrent une plage unique sur le calendrier (en 2009, dans les années 1820, depuis le 12^{e} siècle, etc.).

Le démonstrateur mis en œuvre pour ce scenario est accessible en ligne aux adresses suivantes :

- Pour la version française : http://client1.mondeca.com/TemporalQueryModule/?locale=fr
- Pour la version anglaise : http://client1.mondeca.com/TemporalQueryModule/?locale=en

7.2.1.1 La constitution des corpus pour les démonstrateurs

Les deux corpus indexés pour cette expérimentation (l'un en français sur l'histoire de France, l'autre en anglais, sur l'histoire des Etats-Unis) sont constitués d'articles de Wikipédia[50]. Les textes ont été collectés grâce à un ensemble de requêtes sur le *endpoint* SPARQL[51] de DBpedia[52], en agrégeant les

[50] http://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Accueil_principal
[51] http://dbpedia.org/sparql
[52] http://dbpedia.org/About

liens sortant d'un petit nombre de pages de catégorisation de Wikipédia (« France », « Histoire de France », etc.), sur trois niveaux de profondeur. L'objectif était de constituer automatiquement des corpus suffisamment volumineux et cohérents pour permettre des requêtes libres sur une thématique donnée. Constitués de façon automatique sans intervention humaine de contrôle, ces corpus ne sont bien évidemment pas exhaustifs. Ils ne contiennent pas non plus que des documents en rapport avec leur thème. Cependant, ils délimitent un périmètre thématique assez ouvert pour permettre de tester le système sur des requêtes très variées.

(1) La version française du démonstrateur permet d'explorer un corpus réunissant près de 16 000 articles de Wikipédia sur l'histoire de France. Il contient près de 280 000 adverbiaux calendaires.

Voici quelques exemples de requêtes :

- laïcité avant 1905
- peine de mort depuis les années 70
- travaux paris vers le milieu du XIXe siècle
- Villers-Cotterêts
- Saint-Barthélemy

(2) La version anglaise du démonstrateur permet d'explorer un corpus réunissant plus de 55 000 articles de Wikipedia sur l'histoire des Etats-Unis. Il contient près de 840 000 adverbiaux calendaires.

Voici quelques exemples de requêtes :

- prohibition at the beginning of the 30s
- desegregation since the mid-50s
- earthquake between 1980 and 1990
- Cuban Missile Crisis
- United States Constitution by the end of the 18th century
- Herbert Hoover
- Normandy landings

7.2.1.2 Un scenario d'utilisation

Nous présentons ici un scenario d'usage, afin d'illustrer les fonctionnalités du moteur de recherche :

1) Un utilisateur soumet la requête suivante au système CaSE « *université au début du XIIe siècle* ». Le système produit les résultats suivants :

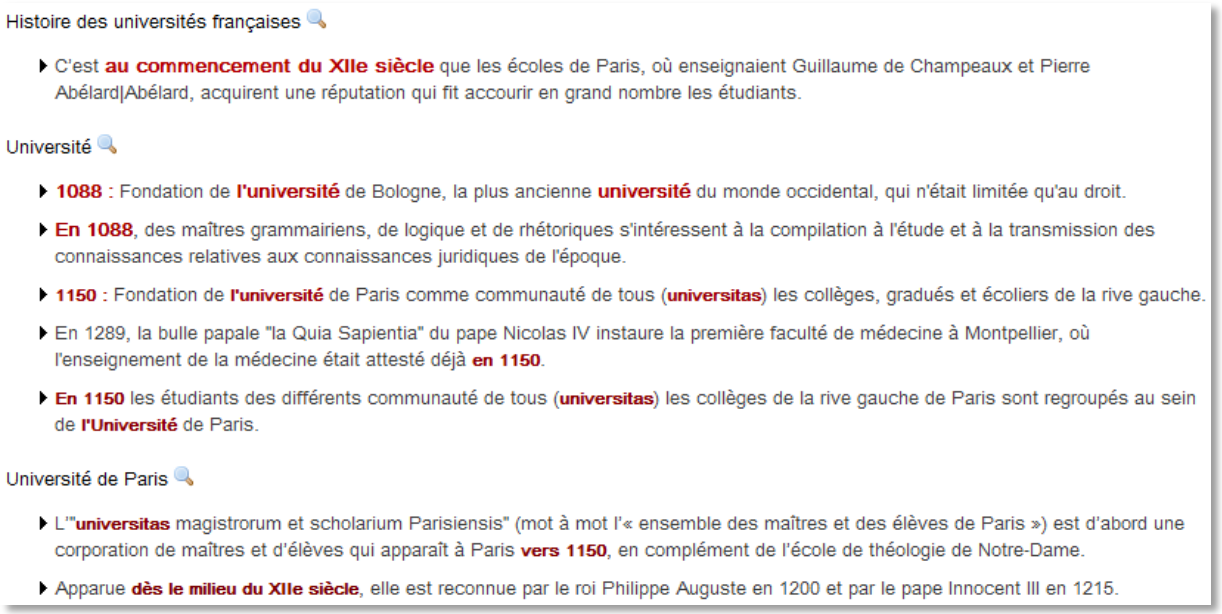


Fig. 67 : copie d'écran de la liste des résultats proposés pour la requête « université au début du XIIe siècle »

2) Parmi ces résultats, l'utilisateur peut par exemple sélectionner le second document (qui a pour titre *Université*) afin de le parcourir sous la perspective calendaire. Le système restreint alors la recherche initiale à ce seul document. Les phrases extraites du document pour cette recherche sont ordonnées par pertinence :

Fig. 68 : copie d'écran de la liste des résultats proposés pour la requête « université au début du XIIe siècle » dans le cadre d'une recherche intra-documentaire

3) L'utilisateur peut alors changer l'échelle de la frise chronologique, déplacer la fenêtre visualisée et la positionner, par exemple, aux alentours de 1550. Le système modifie alors la liste des résultats et accole à la requête initiale un nouvel adverbial calendaire correspondant à la fenêtre de temps visualisée (« *des années 1530 aux années 1570* ») :

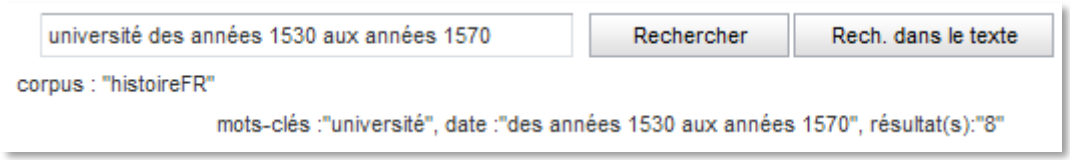


Fig. 69 : la requête modifiée suite à un déplacement de la frise chronologique : l'adverbial « des années 1530 aux années 1570 » a été générée automatiquement

Fig. 70 : copie d'écran de la liste des résultats proposés suite au déplacement de la frise chronologique

L'adverbial généré est fonction de l'échelle de la frise chronologique. En l'occurrence, dans le scenario décrit, elle correspond à une échelle décennale. En mettant à jour le critère calendaire exprimé dans la requête, le système permet de préserver la cohérence entre la requête affichée (« *université des années 1530 aux années 1570* ») et les résultats proposés, aussi bien sur la frise chronologique que dans la liste des phrases affichées.

Le système permet ainsi de basculer d'une recherche documentaire vers une recherche intra-documentaire, pour parcourir un document. Il est ainsi possible d'explorer un corpus ou un document sous l'angle calendaire, avant éventuellement d'accéder à la page Web d'un document donné.

7.2.2 Interroger des données structurées : un cas d'utilisation

Le but de cette seconde expérimentation décrite dans (Vandenbussche et Teissèdre, 2011) est de montrer que le système d'indexation peut également interagir avec des données structurées contenant des propriétés calendaires et donc intégrer aussi les infrastructures logicielles qui s'appuient sur le Web Sémantique. Ici, les informations indexées par le moteur sont des données RDF fournies dans le cadre d'un atelier, DeRiVE 2011[53] (*Detection, Representation, and Exploitation of Events in the Semantic Web*), dont l'objectif était d'explorer différentes pistes pour exploiter des

[53] http://semanticweb.cs.vu.nl/derive2011/Home.html

informations relatives à des événements. Le jeu de données (*dataset*) décrit en l'occurrence des événements en lien avec la musique, tels que des annonces de concerts ou de festivals.

Le jeu de données de DeRiVE a été développé dans le cadre du projet EventMedia[54] (Troncy *et al.*, 2010). Il est constitué d'environ 100 000 descriptions d'événements provenant des trois sources d'informations : le site Web de *Last.fm* consacré à la musique et deux autres sites consacrés aux loisirs, *upcoming.yahoo.com* et *eventful.com*.

	Events	Actors	Venues
Eventful	37.647	6.543	5.173
Upcoming	13.114	0	15.857
Last.fm	57.258	50.151	13.516
Links between the Challenge datasets	~700	~2500	~2000
Links to external datasets	~300 (DBpedia)	~11500 (DBpedia)	~300 (DBpedia) ~2000 (GeoNames)

Tableau 3 : La composition du jeu de données de DeRiVE

Les données décrivent ainsi des événements en lien avec la musique et fournissent également des informations sur les « *agents* » impliqués dans ces événements (des artistes, des groupes de musique), ainsi que sur les lieux où ils se déroulent (salles de concert, bars, café-concert, etc.). Le format des données se conforme au schéma du LODE[55] dont on a déjà évoqué les travaux (cf. section 3.2.2). Les événements décrits se sont étalés sur une période allant de 2007 à 2009. L'ensemble du jeu de données regroupe plus de 1 800 000 triplets RDF. Les informations calendaires sur les événements sont ou bien des dates « simples » ou bien des intervalles de dates. La fig. 71 présente un exemple de graphe RDF représentant un événement :

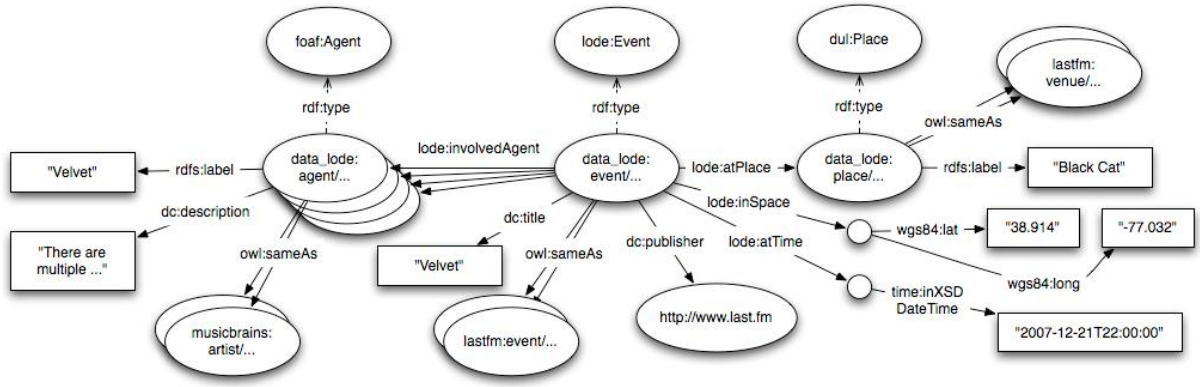


Fig. 71 : exemple de graphe RDF représentant un événement

Les différentes parties de ce schéma ne sont pas systématiquement instanciées dans les données RDF. Les données provenant du site *Upcoming events*, par exemple, ne contiennent pas de description d'« agents ».

[54] http://eventmedia.cwi.nl/
[55] http://linkedevents.org/ontology

Le principal objectif de l'expérience que l'on rapporte ici est de montrer comment on peut tirer parti de ces données, tout en masquant leur complexité à l'utilisateur. Pour les rendre simples à interroger, un unique champ de recherche est proposé à l'utilisateur, qui peut saisir des requêtes combinant des mots-clés, des lieux et des critères calendaires. Le système expérimental pourrait ainsi servir de point d'entrée pour la recherche d'événements musicaux. Le scenario d'utilisation considéré consiste à rechercher des événements qui se produisent à une certaine période, dans un endroit déterminé.

Afin de disposer des informations nécessaires à ce type de recherche (qui combine mots-clés, critères géographiques et critères calendaires), il a fallu enrichir le jeu de données initial. En particulier, il a fallu géo-localiser certains des événements incomplètement décrits. Parmi les informations susceptibles d'enrichir les données, nous avons également montré (Vandenbussche et Teissèdre, 2011) qu'il était possible de s'appuyer sur les données ouvertes (*Linked Open Data*) pour obtenir des informations sur les « agents » (ajout de visuels, liens vers des articles de Wikipédia sur un artiste ou d'un groupe). Les données ont été enrichies de façon automatique et donc non-contrôlée. Ceci signifie que tout terme polysémique (un groupe de musique portant un nom de ville ou un nom commun par exemple) est susceptible de produire des résultats erronés. Une opérationnalisation plus avancée du prototype nécessiterait de mettre en œuvre une chaîne de traitements semi-automatique pour le contrôle et la validation des informations collectées. Nous montrerons plus loin comment une chaîne de cette nature peut être implémentée (cf. section 7.2.3).

Dans ce cadre, le système CaSE a été intégré dans une architecture plus vaste et couplé avec un entrepôt de données RDF (*triplestore*), enrichi durant la phase de collecte et de pré-traitement des données (cf. fig. 72).

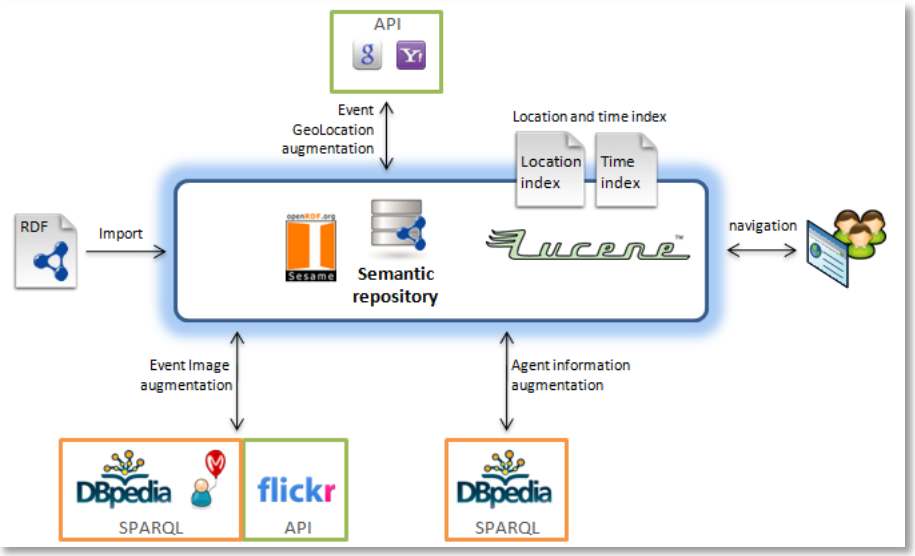


Fig. 72 : architecture du système d'indexation des événements

Les données RDF initiales sont chargées dans un entrepôt de données Sesame[56]. Cet entrepôt permet d'importer des données RDF, de les analyser, de les interroger et de les mettre à jour, via des API et des requêtes SPARQL.

A l'aide de requêtes SPARQL, il est ainsi possible d'interroger cet entrepôt pour récupérer les données sur lesquelles on souhaite ajouter des informations. Les données incomplètes sont alors enrichies à l'aide de requêtes sur le endpoint SPARQL de DBPedia (pour obtenir des liens vers des articles de Wikipedia relatifs aux artistes), des requêtes via les API de Flickr pour obtenir des visuels et des requêtes via les API de Google et de Yahoo! pour obtenir les coordonnées géographiques permettant de positionner sur une carte le lieu des événements. Les informations ainsi obtenues sur les artistes, les lieux et les événements sont alors mises à jour dans l'entrepôt de données. Les données à indexer sont ensuite extraites du triplestore par de nouvelles requêtes SPARQL : un adapteur Java que nous avons développé parse les données RDF extraites et fait alors correspondre les données à indexer avec les champs d'indexation de Lucene.

Lors de la phase d'interrogation, les requêtes soumises au système sont analysées de telle sorte que les mots-clés, les critères calendaires et les noms de lieux soient séparés. L'extraction des informations relatives aux lieux s'appuie sur un dictionnaire de noms de villes et de pays, dont les entrées ont été collectées pendant le pré-traitement des données. Le traitement des informations géographiques est donc rudimentaire et pourrait être considérablement amélioré. Il n'avait ici que vocation à illustrer qu'il est possible de croiser des informations spatiales, temporelles et thématique. Les résultats que renvoie le système sont présentés sur une frise chronologique (cf. fig. 73).

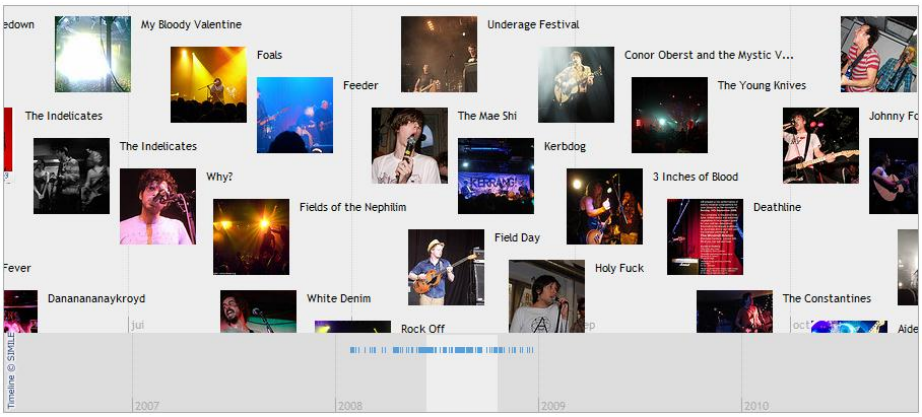


Fig. 73 : Copie d'écran des résultats du moteur pour la requête « rock in London in August 2008 »

Cette application expérimentale pourrait servir de socle pour un service permettant la recherche d'événements. L'approche étant générique et n'étant pas spécifiquement liée aux données de l'atelier DeRiVE, elle peut être étendue, comme on va le voir, à d'autres types de données ancrées temporellement.

[56] http://www.openrdf.org/about.jsp

7.2.3 En consultant, en contribuant : enrichir des données ouvertes à mesure qu'on les interroge. Une expérimentation autour des œuvres d'art dans Freebase

7.2.3.1 Motivations

La production de savoirs, l'édition collaborative, la contribution aux bases de connaissances ouvertes et libres sont des processus auxquels les utilisateurs ne participent pas nécessairement. Il y a ainsi des profils différents de contributeurs dans l'univers de l'Open-Source et des données ouvertes : la grande majorité des contributeurs participent de façon unique ou très épisodique, alors qu'un petit nombre contribue de façon très active, constituant parfois même l'essentiel des données produites (Lerner et Tirole, 2002). Le Web 2.0 a montré l'intérêt de partager l'effort de production des connaissances. Cependant, avec l'avènement des *Linked Open Data* et d'un Web de données ouvertes, le processus de contribution se complexifie, les contributeurs devant manipuler des données et des formats de types différents et non plus seulement du texte. Si le paradigme collaboratif est essentiel à la réussite des bases de connaissances à vocation encyclopédique telles que Freebase[57] - le pendant de Wikipedia dans le cadre du Web de données -, il se heurte donc néanmoins à l'expertise requise de la part des contributeurs.

Nous souhaitons montrer par l'expérience décrite ici qu'il doit être possible d'amener davantage d'utilisateurs non experts à contribuer, de la façon la plus simple possible, à mesure qu'ils consultent les données et selon un point de vue donné. Il s'agit ainsi de rapprocher deux démarches habituellement distinctes : la consultation et la contribution. Les propositions que l'on formule visent à amener les utilisateurs à consulter une base de connaissances et à y contribuer dans un même geste, sur des tâches simples.

Dans les résultats d'une recherche, l'outil soumet ainsi aux utilisateurs des suggestions d'ajout, qu'il leur suffit ensuite éventuellement de valider. Ces suggestions pourraient être fonction du type de recherche effectuée : une recherche sous l'angle géographique, une recherche sous l'angle temporel, une recherche sous l'angle des liens entre personnes, etc., donneraient lieu à des suggestions d'ajouts différentes (suggestions d'ajout de propriétés géographiques, de propriétés temporelles, de liens entre entités, etc.). Dans ce cadre, nous nous sommes attachés à la recherche d'information sous l'angle temporel, mais l'approche doit être reproductible et extensible à d'autres types de propriétés.

L'expérimentation menée pour illustrer l'intérêt de la démarche a consisté à indexer des données de Freebase, une base de connaissances ouverte et collaborative, s'appuyant sur le formalisme du Web Sémantique. A chacune des « entrées » de Freebase (*topics*) sont associés un court descriptif (du texte libre, qui provient fréquemment de Wikipédia) et diverses informations. Le format de ces informations est contraint par le ou les modèles associé(s) à chaque entrée de l'encyclopédie. Pour une œuvre d'art, par exemple, le modèle propose de renseigner le ou les artistes qui l'ont créée, la date à laquelle elle a été créée, etc. Parmi les informations réunies dans la base de connaissances Freebase, on trouve donc en particulier des informations de localisation temporelle : des

[57] http://www.freebase.com/

événements, des périodes de l'histoire, des œuvres d'art qui peuvent être datés ou encore des dates de naissance et de décès pour des personnes.

L'enrichissement de Freebase étant collaboratif et non contraint, les informations effectivement renseignées pour une entrée donnée sont souvent parcellaires. Les entrées de cette base de connaissances ne sont en effet pas toutes décrites avec le même degré de détail. Ainsi, par exemple, près d'un tiers des événements et des œuvres d'art répertoriés ne sont pas datés. Ces informations ne sont donc pas interrogeables sous l'angle temporel : elles ne peuvent pas par exemple être positionnées sur la frise chronologique proposée dans Freebase pour explorer certains contenus.

L'extension de système CaSE développée pour cette expérimentation permet d'interroger les données de Freebase décrivant des œuvres d'art décrites dans Freebase. La figure 74 présente le résultat d'une recherche proposée par le système CaSE pour la requête « *italian painting at the beginning of the 15th century* » (*peinture italienne au début du 15^{e} siècle*).

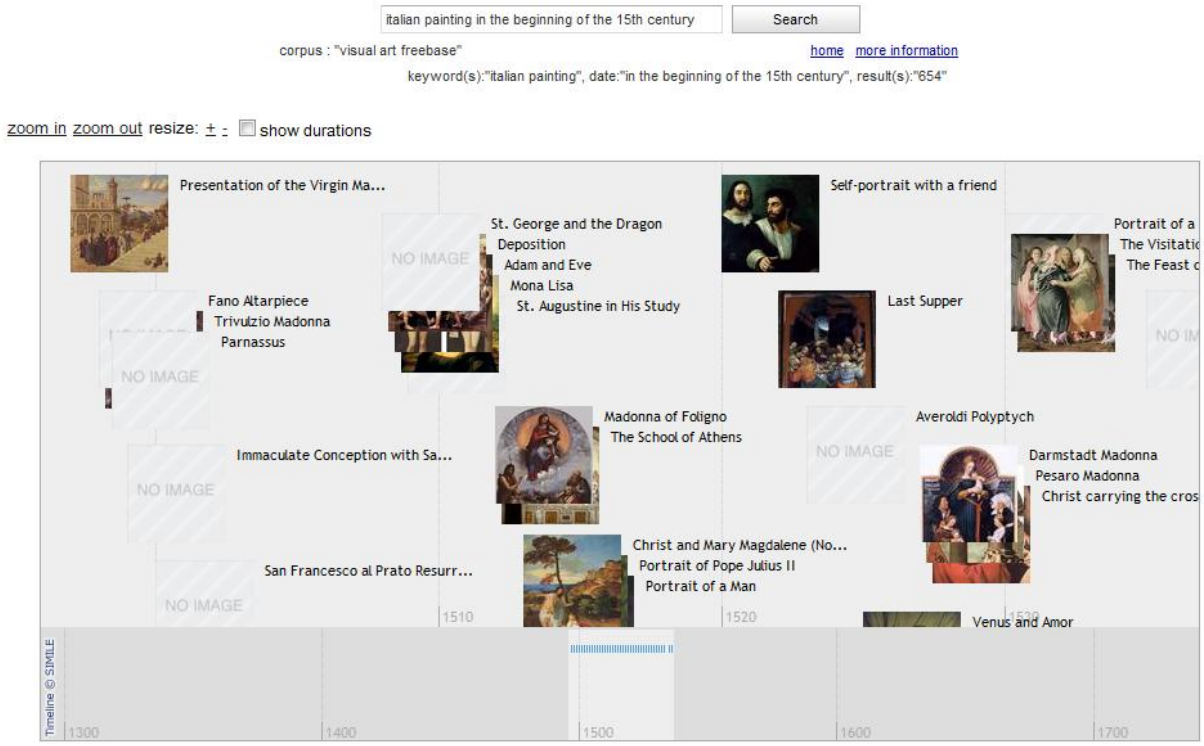


Fig. 74 : copie d'écran de l'interface d'interrogation des œuvres d'art de Freebase

Le système permet également d'interroger et de visualiser certaines des données relatives à des œuvres d'art qui pourtant ne précisent pas leur date de création. Sur un ensemble d'un peu plus de 16 600 œuvres d'art décrites dans Freebase, près de 30% n'étaient pas datées au 6 juin 2011.

7.2.3.2 Constitution de l'index

Freebase propose aux développeurs un langage d'interrogation de son entrepôt de données. Ce langage, MQL (Metaweb Query Language), a une syntaxe simple au format JSON. Elle consiste

notamment à décrire sous la forme d'un motif à trous les données que l'on souhaite récupérer. Les requêtes s'appuient sur la structure des données à récupérer. Le langage de requête permet également d'ajouter, de modifier ou de supprimer des données dans l'entrepôt. A l'aide d'une série de requêtes MQL, il est ainsi possible de récupérer l'ensemble des données décrivant des œuvres d'art dans Freebase. Ce sont ces données qui ont été indexées par le système CaSE.

Lors de la phase d'indexation, les données relatives aux œuvres d'art dont la date de création n'est pas renseignée font l'objet d'une analyse qui vise à extraire les adverbiaux calendaires éventuellement présents dans le descriptif (le texte) qui accompagne chacune des œuvres : le système d'annotation extrait les adverbiaux calendaires et le module de transduction leur associe une valeur calendaire. Cette valeur peut alors être présentée à l'utilisateur comme une suggestion d'ajout d'une propriété décrivant la date de création d'une œuvre.

7.2.3.3 Enrichissement des données

L'hypothèse qui est faite est qu'il est vraisemblable que, dans les adverbiaux annotés, l'un d'eux corresponde à la date de création de l'œuvre. Les adverbiaux ainsi annotés sont indexés par le moteur de recherche CaSE sous la forme d'informations candidates. Un avantage de ce processus est qu'il devient possible d'exposer, au sein des résultats d'une recherche, des connaissances non encore validées. Sur la frise chronologique, les œuvres auxquelles des suggestions sont associées sont signalées par le mot-clé « [contrib ?] » (cf. fig. 75).

Par exemple, une requête telle que « Caravaggio » ramène plusieurs résultats, parmi lesquels « *Saint Francis in Ecstasy* » (L'*extase* de *Saint François d'Assise)*. A cette œuvre aucune date n'est associée dans Freebase. Lors des pré-traitements, le système d'annotation a repéré dans le descriptif accompagnant cette œuvre l'adverbial calendaire « *from 1595* ». L'interface propose à l'utilisateur d'ajouter la date 1595 dans Freebase comme « date de création de l'œuvre ».

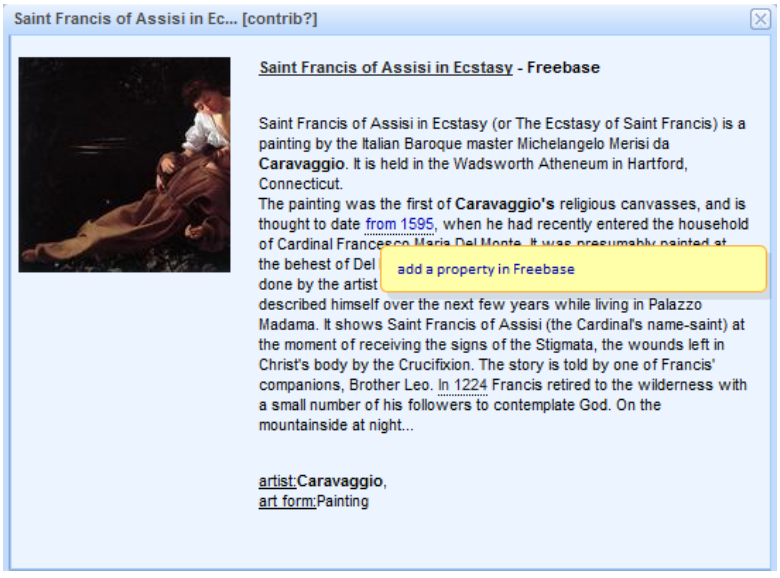


Fig. 75 : suggestion d'ajout d'une propriété temporelle dans Freebase

En validant cette proposition, l'utilisateur enrichit de façon simple les données de Freebase, sans avoir à passer par l'interface d'édition. Lors de la validation d'une suggestion d'ajout, une requête MQL est générée permettant d'ajouter une propriété à l'objet concerné. Parallèlement, l'index Lucene du démonstrateur est lui aussi mis à jour.

L'expérience, conduite avec quelques utilisateurs seulement, a permis d'ajouter des dates à près de 1 350 œuvres d'art décrites dans Freebase.

Cette expérience montre ainsi comment il est possible de tirer parti de l'annotation des adverbiaux calendaires pour la recherche d'information et la gestion de bases de connaissances. Elle montre également, dans l'optique d'un enrichissement collaboratif des données ouvertes, qu'il peut être intéressant de permettre l'ajout de connaissances structurées dans une base de connaissances à mesure même que l'on cherche et consulte des informations.

7.3 Bilan du chapitre

Encore faiblement exploités pour la recherche d'information, les adverbiaux de localisation temporelle peuvent fournir des éléments pertinents de sélection et de tri des documents. Le système de recherche d'information que l'on a présenté montre qu'il est possible de les exploiter tout en s'insérant dans les technologies open-source existantes.

On a montré en outre, que l'approche retenue, qui consiste à évaluer la pertinence relative d'un intervalle calendaire-cible par rapport à un intervalle calendaire-requête peut être exploitée à la fois pour la recherche documentaire et pour la recherche dans des données structurées. Cette approche peut être encore raffinée, afin d'exploiter plus finement la sémantique des adverbiaux de localisation temporelle, en enrichissant le modèle des intervalles calendaires à partir duquel est évaluée la pertinence d'un document.

Par ailleurs, pour la recherche intra-documentaire, il serait intéressant d'enrichir l'interface de présentation des résultats et de développer des fonctionnalités permettant à l'utilisateur/lecteur de visualiser en contexte les informations pertinentes : l'outil pourrait alors s'intégrer aux nouveaux supports dédiés à la lecture numérique et à la navigation textuelle.

Dans le chapitre suivant, nous présentons une évaluation des différents composants développés pour le traitement des adverbiaux de localisation temporelle, qui montrent que notre approche, qui procède initialement d'une démarche linguistique, présente un intérêt à la fois pour l'ingénierie des langues, pour l'acquisition de connaissances et pour la recherche d'information lorsqu'il s'agit de traiter des informations temporelles dans les textes.

Chapitre 8 : Evaluation

Pour mesurer les performances des composants développés dans ce cadre pour l'annotation, l'acquisition de connaissances et la recherche d'information, nous avons mis en œuvre différents protocoles d'évaluation, s'appuyant sur des métriques standard. L'évaluation se découpe en plusieurs étapes, l'idée étant, lorsque c'est possible, d'évaluer chacun des composants isolément, avant d'évaluer des chaînes de traitements où plusieurs composants interviennent.

On évalue ainsi dans ce chapitre :

- le système d'annotation des adverbiaux de localisation temporelle, en distinguant les tâches d'extraction, d'annotation et de transduction vers des intervalles calendaires ;
- le prototype d'acquisition de connaissances pour la saisie de dates et horaires d'ouverture ;
- le prototype de moteur de recherche CaSE, en isolant d'abord le module de sélection et de tri par pertinence des adverbiaux calendaires, avant d'évaluer le système dans son ensemble.

8.1 Le système d'annotation et de transduction vers des intervalles calendaires

On a vu que le système d'annotation des adverbiaux de localisation temporelle que l'on a développé ne permet pas, dans sa forme actuelle, de remplir intégralement le programme définit par le schéma d'annotation *ChronolocationML*.

Le schéma d'annotation reflète notre proposition d'analyse des adverbiaux de localisation temporelle présentée dans le chapitre 4 : il doit permettre, en théorie, d'annoter et de décrire tout type d'adverbiaux de localisation temporelle, conformément à la description sous la forme d'une

succession d'opérateurs sémantiques. Pour autant, comme on l'a vu, ce schéma, comme la modélisation dont il découle, doivent eux aussi encore être affinés à plusieurs niveaux.

Le système d'annotation ne reconnaît et ne décrit donc pas l'ensemble des adverbiaux de localisation temporelle, mais seulement une sous-partie d'entre eux (cf. 6.1.4.4) : les adverbiaux dont le repère temporel noyau est constitué d'une base calendaire (*« depuis mai 2008 »*), d'une base déictique (« dès demain ») ou d'une base anaphorique (*« ce jour-là »*). Les adverbiaux dont le repère temporel noyau est formé d'une base relative à un procès (*« quelques jours avant qu'il ne reconnaisse les faits »*) ne sont pas (ou incomplètement) repérés.

Pour évaluer un système d'annotation, on recourt généralement à trois mesures classiques, le taux de rappel (les informations pertinentes repérées ou annotées rapportées à l'ensemble des informations pertinentes), le taux de précision (les informations pertinentes repérées ou annotées rapportées à l'ensemble des informations repérées ou annotées), et une F-mesure (une moyenne harmonique combinant de façon pondérée les deux taux, pouvant ou bien accorder une importance plus grande à l'un des deux taux ou bien les mettre sur un pied d'égalité). Ces mesures se calculent de la façon suivante :

$$Rappel = \frac{Informations\ pertinentes\ repérées\ ou\ annotées}{Informations\ pertinentes}$$

$$Précision = \frac{Informations\ pertinentes\ repérées\ ou\ annotées}{Informations\ repérées\ ou\ annotées}$$

$$F - Mesure = \frac{(1 + \alpha^2) * Précision * Rappel}{(\alpha^2 * Précision) + Rappel}$$

Nous utilisons ici ces mesures, afin d'évaluer trois tâches distinctes : (i) la tâche de reconnaissance des unités textuelles visées (la bonne délimitation des unités reconnues), (ii) l'annotation (la justesse des annotations associées à ces unités) et (iii) la transduction des adverbiaux calendaires annotés vers des intervalles calendaires. La $F - Mesure$ calculée ne privilégie ni le rappel, ni la précision, qui sont pondérés de façon égale ($\alpha = 1$) : il s'agit donc de la mesure dite F_1.

8.1.1 Le corpus d'évaluation

Pour des raisons pratiques, nous avons utilisé le corpus FR-TimeBank (Bittar, 2010) pour l'évaluation du système d'annotation. Ce corpus est constitué de 109 articles de l'Est Républicain. Comme il est annoté conformément au langage TimeML, il est possible d'en extraire l'ensemble des « expressions temporelles » au sens de ce standard (les balises *Timex*). Une fois extraites, nous avons isolé manuellement, dans cet ensemble, les adverbiaux de localisation temporelle sur lesquels le système devait être testé : cette étape était nécessaire dans la mesure où les unités visées par TimeML ne sont pas tout à fait les mêmes que celles que visent notre système d'annotation. Contrairement à TimeML, le système d'annotation ne cherche à isoler ni les événements, ni les durées. Au sein des

adverbiaux de localisation temporelle, il s'attache à décrire les adverbiaux constitués d'une base calendaire, déictique ou anaphorique.

Un autre intérêt du corpus d'évaluation choisi est que la date de publication des articles est fournie : cette information est importante, car elle permet d'opérer (et donc d'évaluer) la transduction de certains adverbiaux déictiques vers des intervalles calendaires.

Pour évaluer le système, nous avons donc comparé les sorties du système avec celles qui ont été produites manuellement. Faute de ressources, nous avons nous-même effectué l'ensemble de ces étapes : une évaluation plus approfondie, plus rigoureuse aussi, aurait exigé l'intervention de plusieurs personnes n'ayant pas participé au développement des règles d'annotation et dont on aurait pu comparer les résultats en mesurant l'accord inter-annotateur. On décrira plus loin un tel protocole, que nous avons pu mettre en place grâce à l'intervention de plusieurs évaluateurs, pour évaluer les performances du moteur de recherche expérimental.

Les tableaux 4 et 5 fournissent des indications sur la nature du corpus, en détaillant le nombre des adverbiaux dits « unaires » (désignant un unique repère temporel), des adverbiaux binaires (désignant deux repères) et des adverbiaux composés (à partir d'adverbiaux unaires ou binaires). Pour les adverbiaux unaires, nous précisons la nature du repère temporel noyau (base calendaire absolue ou périodique, base déictique ou base anaphorique) (cf. tableau 4). Au total, le corpus contient 489 adverbiaux de localisation temporelle à annoter (sont exclus les adverbiaux de localisation dont la base est relative à un procès puisque le système ne les annote pas). Parmi ceux-ci, 85 sont des adverbiaux composés (cf. tableau 5).

	adverbiaux de localisation temporelle unaires
Total	**386**
bases calendaires absolues	30
bases déictiques	211
bases anaphoriques	130
bases calendaires périodiques	15

Tableau 4 : La répartition des adverbiaux unaires dans le corpus d'évaluation

	adverbiaux de localisation temporelle n-aires
Total	**103**
adverbiaux binaires	18
adverbiaux composés	85

Tableau 5 : La répartition des adverbiaux n-aires dans le corpus d'évaluation

Comme toute évaluation des systèmes d'annotation, les mesures globales obtenues dépendent étroitement de la nature du corpus (Maynard *et al.*, 2002 ; Jacques et Aussenac-Gilles, 2006) : la variété, la complexité et la fréquence des adverbiaux de localisation temporelle peut grandement varier d'un corpus à l'autre. Nous fournissons donc plus loin le détail des évaluations pour chaque type d'adverbiaux de localisation temporelle, afin de pouvoir mesurer plus finement les capacités du système, tout en ayant présent à l'esprit que, même ainsi détaillés, les chiffres que l'on fournit valent pour ce corpus - ou ce type de corpus -.

8.1.2 L'extraction

Pour la tâche d'extraction, il s'agit d'évaluer la capacité d'un système à identifier des unités textuelles pertinentes et à les délimiter correctement, en observant en particulier les frontières des unités pertinentes repérées. On peut ainsi distinguer plusieurs cas de figure (Freitag, 1998) : un adverbial peut être parfaitement délimité par le système d'annotation ou bien l'être imparfaitement (soit la délimitation proposée par le système déborde l'unité visée, soit elle ne la couvre qu'incomplètement). La mesure des taux de précision et de rappel peut ainsi être raffinée, de sorte à pondérer les résultats en fonction de la nature des erreurs : l'absence d'annotation d'une unité pertinente peut par exemple être considérée comme une erreur plus importante qu'une mauvaise délimitation.

Les résultats calculés dans le tableau 6 ont été obtenus sans distinction faite de la nature des erreurs (méthode 1). Les résultats calculés dans le tableau 7 ont été obtenus en pondérant selon un facteur de 0,5 les erreurs considérées comme étant de moindre importance (méthode 2). Ces erreurs sont celles liées à la mauvaise délimitation des unités textuelles reconnues, qui est considérée comme une erreur moindre que leur non reconnaissance (silence) ou que l'annotation d'une unité textuelle qui n'aurait pas dû l'être (bruit). Remarquons que ces mesures auraient pu encore être raffinées, en fonction de l'importance des informations laissées de côté dans le cas des annotations mal délimitées : une mauvaise délimitation peut être à l'origine d'une mauvaise analyse sémantique, alors que d'autres erreurs, plus bénignes, n'empêchent pas que la sémantique d'un adverbial soit correctement décrite conformément au modèle de représentation adopté. Cependant, un tel degré de raffinement ne semblait pas nécessaire dans la mesure où l'on évalue par ailleurs la qualité des annotations : on isole donc la problématique de la description sémantique de celle de la reconnaissance des unités textuelles visées. Cependant, afin de pouvoir fournir des chiffres détaillés pour chaque catégorie d'adverbial, le typage des adverbiaux est également évalué à ce niveau : nous avons ainsi évalué la capacité du système à extraire chacun des types d'adverbiaux.

	F-Mesure	précision	rappel
Total	**0,89**	**0,92**	**0,86**
Adv. unaires de base calendaire absolue	0,84	0,84	0,83
Adv. unaires de base déictique	0,93	0,93	0,92
Adv. unaires de base anaphorique	0,86	0,82	0,90
Adv. unaires de base calendaire périodique	0,83	0,75	0,92
Adverbiaux binaires	0,85	0,85	0,84

Tableau 6 : Résultats de l'évaluation du système pour la reconnaissance des adverbiaux de localisation temporelle (méthode 1)

	F-Mesure	précision	rappel
Total	**0,91**	**0,95**	**0,88**
Adv. unaires de bases calendaires absolues	0,91	0,96	0,87
Adv. unaires de bases déictiques	0,96	0,98	0,93
Adv. unaires de bases anaphoriques	0,85	0,89	0,82
Adv. unaires de bases calendaires périodiques	0,84	0,90	0,79
Adverbiaux binaires	0,89	0,92	0,87

Tableau 7 : Résultats de l'évaluation du système pour la reconnaissance des adverbiaux de localisation temporelle (méthode 2)

Globalement, il ressort de ces mesures que la tâche de reconnaissance est plutôt bien effectuée par le système. Voici cependant quelques exemples d'erreurs les plus fréquentes du système :

- Exemples d'adverbiaux non extraits : « à une date ultérieure », « dans la soirée », « 1950 ». Pour les deux premiers exemples, la non reconnaissance de l'adverbial vient d'une déficience du jeu de règles d'annotation (les adverbiaux ne sont pas reconnus parce qu'aucune règle ne les décrits). Dans le dernier cas, l'année n'est pas reconnue, car la règle d'annotation reconnaissant les années exigent un indice plus fort de la présence d'un adverbial (comme une préposition par exemple (« en 1950 ») : si un tel choix pèse sur le taux de rappel, en revanche il améliore le taux de précision. En effet, la succession de 4 chiffres représentant une année peut tout aussi bien correspondre à un nombre et non à une valeur calendaire. La constitution d'un jeu de règles d'annotation exige ainsi souvent, comme dans ce cas, de trouver un équilibre entre le taux de rappel et le taux de précision. Compte-tenu des applications où le système est utilisé, nous avons eu tendance à privilégier le taux de précision sur celui de rappel.
- Exemples d'unités textuelles extraites non pertinentes (les unités textuelles isolées par le système sont surlignées) : « Chiffre : le 6 », « autour de Mars », « Journal du Dimanche »,

« après le 1er saut ». A ce niveau, le bruit vient le plus souvent de l'ambiguïté des marqueurs (« Mars » peut désigner un mois ou la planète, « le 6 » peut désigner le 6 d'un mois ou bien, en l'occurrence, le chiffre… 6). Là encore, les marqueurs ambigus qui occasionnent trop fréquemment du bruit pourraient être délibérément retirés du jeu de règles d'annotation. Par ailleurs, le dernier exemple montre les limites de nos ressources, qui n'effectuent pas d'analyse morphosyntaxique : en effet, en reconnaissant que « le 1er saut » forme un syntagme nominal, le système aurait pu détecter que « après le 1er » ne formait pas un adverbial complet.

- Exemples d'adverbiaux mal délimités (les unités textuelles isolées par le système sont surlignées) : « ce même lundi-après midi », « jusqu'à tard dans la nuit », « A l'Aube de l'an 2000 », « d'ici à novembre », « du mercredi soir au lundi matin ». Ces mauvaises délimitations traduisent ici encore une déficience au niveau des règles d'annotation qui doivent pouvoir être améliorées. Les erreurs sur les adverbiaux de localisation binaires, comme dans le dernier exemple, viennent le plus souvent du fait que leurs constituants ont été reconnus individuellement, mais pas conjointement.

Les évaluations concernant les adverbiaux de base calendaire périodiques sont statistiquement peu significatives, compte-tenu du faible nombre de ces adverbiaux dans le corpus (15 au total). Cependant, le plus souvent c'est moins la délimitation de ces adverbiaux qui pose des difficultés au système que le fait de les typer correctement, car ils peuvent être confondus avec des adverbiaux déictiques. Par exemple dans l'énoncé : « *Pendant les mois d'été, et particulièrement* ***entre le 15 juillet et le 15 août****, le camping affiche complet.* », le système considère que l'adverbial binaire est composé de deux adverbiaux de base déictique, alors qu'il s'agit de bases calendaires périodiques.

8.1.3 L'annotation

Il s'agit d'évaluer la capacité du système à décrire les adverbiaux de localisation temporelle conformément à l'analyse formelle que l'on propose. A ce niveau, le système peut produire plusieurs types d'erreurs. On peut ainsi distinguer le cas des adverbiaux non annotés (le silence), le cas où des adverbiaux ont été annotés et n'auraient pas dû l'être (le bruit), le cas de ceux dont la description est incomplète et, enfin, les cas où le système décrit mal la sémantique des adverbiaux par rapport au modèle proposé. Dans cette évaluation, ces erreurs ne sont pas pondérées, car une mauvaise analyse par le système fausse les traitements ultérieurs : on considère donc que toutes les erreurs sont d'égale importance.

Pour l'évaluation, le système d'annotation a été paramétré en mode robuste, c'est-à-dire qu'il a été paramétré pour fournir une analyse complète des adverbiaux de localisation temporelle, même si, dans certains cas, il aurait pu produire plusieurs analyses ou s'arrêter à des analyses incomplètes.

	F-Mesure	précision	rappel
Total	**0,84**	**0,87**	**0,82**
Adv. unaires de bases calendaires absolues	0,92	0,97	0,88
Adv. unaires de bases déictiques	0,86	0,88	0,84
Adv. unaires de bases anaphoriques	0,84	0,88	0,80
Adv. unaires de bases calendaires périodiques	0,44	0,47	0,41
Adverbiaux binaires	0,81	0,83	0,79

Tableau 8 : Résultats de l'évaluation du système pour l'annotation des adverbiaux de localisation temporelle

Les adverbiaux de base calendaire absolue sont généralement bien annotés et posent peu de difficultés. Parmi les erreurs les plus fréquentes du système, on compte essentiellement :

a. des erreurs liées à des annotations incomplètes, en particulier lorsqu'il s'agit de déterminer l'orientation des opérations d'une opération de déplacement. Par exemple, le système n'ayant pas d'information sur le contexte des adverbiaux qu'il annote, il n'est pas en mesure de savoir si un adverbial déictique tel que « en mars » renvoie à mars dernier ou au mois de mars à venir. Il faudrait pour cela disposer d'informations contextuelles (le plus souvent le temps verbal permettrait de renseigner cette information. Cette erreur (15 au total) représente près de 30% des erreurs d'annotation.
b. des erreurs de typage des adverbiaux (18 erreurs) : on a vu en effet qu'il pouvait y avoir une ambiguïté entre les adverbiaux itératifs et déictiques.

On a par ailleurs évalué l'annotation des relations entre les adverbiaux de localisation temporelle visées par le système d'annotation. Ce corpus ne présente pas de relation de disjonction (ex. : « en mars ou en avril », ni de relation d'exception (« tous les jours sauf le dimanche ».

	F-mesure	précision	Rappel
Total	**0,92**	**0,95**	**0,89**
dont relations de spécification	0,86	0,91	0,81
dont relations de concaténation	0,97	0,98	0,96

Tableau 9 : Résultats de l'évaluation du système pour l'annotation des relations entre adverbiaux de localisation temporelle

Le système ne sait repérer et décrire que les relations entre des adverbiaux contigus. Dans les textes cependant, ces relations peuvent porter entre des éléments qui ne sont pas immédiatement voisins, ce qui explique un taux de rappel légèrement inférieur pour les relations de spécification (par exemple dans l'énoncé suivant : « *Dure loi des séries pour les canalisations en fonte grise de Nancy*

qui datent de la création du réseau, à la fin du XIXe siècle : deux ruptures ***en 97*** *(****en janvier*** *rue Saint-Dizier,* ***en octobre*** *rue de Metz) (…).* ».

8.1.4 La transduction sous la forme d'intervalles calendaires

Comme on l'a vu dans le chapitre 6, le système d'annotation est associé à des ressources logicielles développées pour effectuer différents traitements sur les annotations produites. Ces ressources permettent d'instancier des objets et de transposer éventuellement, lorsque c'est possible, les adverbiaux de localisation temporelle sous la forme d'intervalles calendaires. Cette tâche est généralement désignée, en ingénierie des langues comme la *normalisation des expressions temporelles* : il s'agit par exemple d'affecter une valeur calendaire à des unités textuelles telles que « lundi prochain », « dans deux jours » ou « au début des années 1960 ». Or, on le rappelle, contrairement à l'approche dominante, la transduction des adverbiaux de localisation temporelle sous la forme d'intervalles calendaires ne nous semble pas devoir faire directement partie de la tâche d'annotation des textes, en particulier parce que cette transduction produit une représentation calendaire distincte de la représentation formelle à l'aide de laquelle on décrit les adverbiaux de localisation temporelle. On isole donc, dans cette évaluation, la tâche de transduction, d'autant qu'elle peut produire des ensembles d'intervalles infinis (par exemple s'il s'agit d'affecter des valeurs calendaires à un adverbial tel que « tous les lundis »). En outre, il s'agit pour le système de transcrire un mode d'indexation temporelle vers un autre, entre lesquels il n'y a pas toujours d'équivalence. Par exemple, il n'y a pas une façon unique de transcrire les adverbiaux « *à la fin des années 80* » ou « *depuis un an* », sous la forme d'intervalles calendaires. La transduction de ces adverbiaux dépend en effet de la granularité considérée, alors que celle-ci est indéterminée dans les expressions sources : sur un plan formel, compte-tenu de nos propositions, cela revient à déterminer quelle transformation doit être associée à l'opération de focalisation fin dans « fin des années 80 » ? il s'agit également de savoir à quelle granularité on doit représenter l'intervalle calendaire associé à la base déictique dans l'adverbial « depuis un an » ? s'agit-il d'une granularité *jour* (depuis un an par rapport à aujourd'hui) ? d'une granularité *mois* (depuis un an par rapport à ce mois-ci) ? Il n'en reste pas moins qu'on peut évaluer la conformité de cette transduction avec ce qui était attendu de la part du système, compte-tenu de l'heuristique formelle de transduction que l'on a présentée (cf. section 5.2). Ce que l'évaluation détermine à ce niveau est donc l'ensemble de la chaîne, à savoir la capacité du système à affecter des valeurs calendaires à des adverbiaux, lorsque cette tâche est possible.

Dans un premier temps, dans le protocole d'évaluation mis en place pour mesurer les performances du système, on n'évalue le processus de transduction que dans les cas où il produit un intervalle calendaire unique. On évaluera le cas où le processus de transduction doit produire des ensembles d'intervalles plus loin, lorsqu'il s'agira d'évaluer le système d'assistance à la saisie des dates et horaires d'ouverture. Plus exactement, afin de pouvoir établir des comparaisons avec d'autres systèmes, l'évaluation est limitée au périmètre délimité traditionnellement par les campagnes d'évaluation (en particulier TempEval-2 (Verhagen *et al.*, 2010) ; on renvoie à la description de ces campagnes dans la section 3.1.3.1) : il s'agit d'affecter des valeurs calendaires aux adverbiaux contenant l'expression d'unités calendaires. Certains adverbiaux, tels que « dans quelques temps », par exemple, ne sont donc pas « normalisés ». L'évaluation a donc porté sur les adverbiaux unaires ou binaires de base déictique composés d'unités calendaires (203 au total) et sur les adverbiaux de

base calendaire absolue (38 au total) : nous avons ainsi comparé manuellement l'intervalle calendaire produit par le système pour chacun de ces adverbiaux, à celui qui devait être produit si la chaîne d'annotation et de transduction s'était déroulée correctement.

Pour les adverbiaux déictiques, la transduction était rendue possible dans cette évaluation, car les dépêches étant datées, il était possible d'associer une valeur calendaire au processus énonciatif[58]. Ceci signifie que l'intervalle calendaire associé à la base déictique de ces adverbiaux est systématiquement celui correspondant à la date du jour de publication de l'article dans lequel ils apparaissent.

	F-Mesure	précision	rappel
Total	**0,86**	**0,89**	**0,82**
Adv. unaires de bases calendaires absolues	0,94	0,97	0,91
Adv. unaires de bases déictiques	0,90	0,88	0,91
Adverbiaux binaires	0,90	0,86	0,95

Tableau 10 : Résultats de l'évaluation du système pour la transduction des adverbiaux de localisation temporelle vers des intervalles calendaires

Comme on pouvait s'y attendre dans la mesure où certaines des erreurs d'annotation se répercutent ensuite sur la transduction, il se dégage que la transduction des adverbiaux déictiques présente plus de difficultés que celle des adverbiaux de base calendaire absolue. En effet, si l'orientation de l'opération de déplacement est mal décrite dans un adverbial déictique tel que « Lundi » dans l'énoncé « Lundi, le maire a annoncé… », alors le processus de transduction génère un intervalle calendaire correspondant erroné (celui qui correspondrait à l'adverbial « Lundi prochain » au lieu de « Lundi dernier », par exemple). Les erreurs sur les adverbiaux calendaires sont dues au fait que certains d'entre eux n'ont pas été extraits (cf. section 8.1.2).

Même si, on le rappelle, l'évaluation telle qu'elle a pu être menée présente des limites (qui tiennent essentiellement à ceci que nous les avons nous-mêmes effectuées), il ressort que les performances semblent comparables à celles d'autres systèmes, tels que ceux évalués dans la campagne TempEval-2 (Verhagen *et al.*, 2010) (cf. section 3.1.3.1), aussi bien sur la reconnaissance, l'annotation et la transduction des adverbiaux de localisation temporelle. Cependant, en dehors des ordres de grandeur qui donnent des indications sur les tâches qui posent des difficultés, le système que l'on présente n'est pas directement comparable aux systèmes qui se conforment à TimeML, parce qu'ils ne poursuivent pas tout à fait les mêmes objectifs : les unités textuelles visées ne sont pas identiques, la description sémantique qui leur est associée diffère et la transduction n'est pas toujours équivalente à la tâche de normalisation dans TimeML.

[58] On rappelle toutefois que cette méthode présente des limites, en particulier lorsqu'il s'agit d'analyser des adverbiaux déictiques dans du discours rapporté (cf. section 2.2).

8.2 Evaluation de TKA, le prototype d'assistance à la saisie de dates et horaires d'ouverture

8.2.1 Le protocole d'évaluation

Tel qu'il a été conçu, TKA est un outil d'assistance à la saisie et non un système d'acquisition automatique. Cependant, on n'évalue pas directement ici l'aide fournie par le système, ce qui demanderait un retour d'expérience des utilisateurs sur un système qui fonctionnerait en condition opérationnelle. Une telle évaluation ne pouvait pas être mise en œuvre à ce stade, l'outil n'ayant pas été testé en production. S'il fallait faire l'évaluation en ce sens, on pourrait par exemple mesurer le gain de productivité (le temps passé avec l'outil et sans l'outil) pour saisir un ensemble de données décrivant des dates et horaires d'accessibilité. L'apport fourni par les interactions avec les utilisateurs n'est donc pas évalué en tant que tel.

Nous décrivons ici les résultats d'une évaluation visant à mesurer la capacité du système à annoter et transformer des dates et horaires définissant l'accessibilité d'un service. Nous avons défini en ce sens une série de tâches à effectuer par le système. Le système devait, sans intervention humaine, annoter les énoncés définissant des dates et horaires d'accessibilité, transformer ensuite les annotations produites dans le format de stockage RDF décrit dans la section 6.2.2.2 et enfin transposer les annotations vers une représentation sous la forme d'ensemble d'intervalles calendaires, sur une fenêtre de temps d'une semaine.

Les ressources d'annotation, ainsi que le format de représentation RDF ont été conçus pour décrire des adverbiaux de localisation temporelle. On rappelle que ces ressources ont été étendues pour pouvoir traiter les énoncés décrivant l'ouverture ou la fermeture d'un site ou d'un service.

Pour un énoncé tel que « *ouvert les lundis et mardis, de 10h à 19h* », le système devait générer des annotations conformes à la description formelle des adverbiaux sous la forme d'une succession d'opérations (étape 1), sérialiser ensuite les annotations vers le format RDF destiné au stockage des informations dans une base de connaissances (étape 2) et générer un ensemble d'intervalles calendaires décrivant les plages de temps où le service concerné est accessible (étape 3). Les plages de temps décrivant l'accessibilité du service sont représentées au format iCalendar sur une fenêtre de temps allant du 30 avril 2012 au 6 mai 2012.

Pour l'exemple « *ouvert les lundis et mardis, de 10h à 19h* », le système doit ainsi produire les informations suivantes :

Etape 1 : annotation des adverbiaux de localisation temporelle

Pour des questions de lisibilité, l'annotation est présentée au format ChronolocationML, mais on rappelle que le schéma opérationnel du module d'annotation repose sur une autre DTD. Les informations annotées n'en sont pas moins les mêmes, seules diffèrent les balises.

```
<TEXT>
        <ACCESSIBILITY_SIGNAL type="isOpened">ouvert</ ACCESSIBILITY_SIGNAL >
        <CHRONEX cid="c1">
                <BASE bid="b1" isIterative="true" type="calendar">
```

```
                    <CALENDAR_UNIT type="weekDay">les lundis</CALENDAR_UNIT>
                </BASE>
            </CHRONEX>
            <COMPOSITION_SIGNAL type="concatenation">et</COMPOSITION_SIGNAL>
            <CHRONEX cid="c2">
                <BASE bid="b2" isIterative="true" type="calendar">
                    <CALENDAR_UNIT type="weekDay">mardis</CALENDAR_UNIT>
                </BASE>
            </CHRONEX>
            <COMPOSITION_SIGNAL type="specification">,</COMPOSITION_SIGNAL>
            <INTERVAL iid ="il1" type="fromTo" start="c3" end="c4">
                <CHRONEX cid="c3">
                    <ZONING zonid="zon1" type="since" operand="b1">de</ZONING>
                    <BASE bid="b3" isIterative="true" type="calendar">
                        <CALENDAR_UNIT type="hour">10h</CALENDAR_UNIT>
                    </BASE>
                </CHRONEX>
                <CHRONEX cid="c4">
                    <ZONING zonid="zon2" type="until" operand="b2">&#224;</ZONING>
                    <BASE bid="b4" isIterative="true" type="calendar">
                        <CALENDAR_UNIT type="hour">18h</CALENDAR_UNIT>
                    </BASE>
                </CHRONEX>
            </INTERVAL>
</TEXT>
<COMPOSITION_LINK clid="cl1" type="concatenation" operator="c1" operand="c2"/>
<COMPOSITION_LINK clid="cl2" type="specification" operator="cl1" operand="il2"/>
<ACCESSIBILITY aid="a1" type="isOpened" chronex="cl2"/>
```

Etape 2 : sérialisation des sorties au format RDF

```
<chron:Base rdf:about="#Base_1">
  <chron:dayOfWeek rdf:datatype="&xsd;int">1</chron:dayOfWeek>
  <chron:label rdf:datatype="&xsd;string">lundis</chron:label>
</chron:Base>
<chron:Base rdf:about="#Base_2">
  <chron:dayOfWeek rdf:datatype="&xsd;int">2</chron:dayOfWeek>
  <chron:label rdf:datatype="&xsd;string">mardis</chron:label>
</chron:Base>

<chron:SimpleChronex rdf:about="#SimpleChronex_1">
  <chron:label rdf:datatype="&xsd;string">lundis</chron:label>
  <chron:base rdf:resource="#Base_1"/>
</chron:SimpleChronex>
<chron:SimpleChronex rdf:about="#SimpleChronex_2">
  <chron:label rdf:datatype="&xsd;string">mardis</chron:label>
  <chron:base rdf:resource="#Base_2"/>
</chron:SimpleChronex>

<chron:CompoundChronex rdf:about="#CompoundChronex_1">
  <chron:label rdf:datatype="&xsd;string">lundis et mardis</chron:label>
  <chron:concatenation rdf:resource="#SimpleChronex_1"/>
  <chron:concatenation rdf:resource="#SimpleChronex_2"/>
</chron:CompoundChronex>
```

```
<chron:Base rdf:about="#Base_3">
   <chron:hour rdf:datatype="&xsd;int">10</chron:hour>
   <chron:label rdf:datatype="&xsd;string">10h</chron:label>
</chron:Base>
<chron:Base rdf:about="#Base_4">
   <chron:hour rdf:datatype="&xsd;int">18</chron:hour>
   <chron:label rdf:datatype="&xsd;string">18h</chron:label>
</chron:Base>

<chron:SimpleChronex rdf:about="#SimpleChronex_3">
   <chron:label rdf:datatype="&xsd;string">de 10h</chron:label>
   <chron:base rdf:resource="#Base_3"/>
</chron:SimpleChronex>
<chron:SimpleChronex rdf:about="#SimpleChronex_4">
   <chron:label rdf:datatype="&xsd;string">&#224; 18h</chron:label>
   <chron:Base rdf:resource="#Base_4"/>
</chron:SimpleChronex>

<chron:ChronexInterval rdf:about="#ChronexInterval_1">
   <chron:label rdf:datatype="&xsd;string">de 10h &#224; 18h</chron:label>
   <chron:start rdf:resource="#SimpleChronex_3"/>
   <chron:end rdf:resource="#SimpleChronex_4"/>
</chron:ChronexInterval>

<chron:CompoundChronex rdf:about="#CompoundChronex_2">
   <chron:label rdf:datatype="&xsd;string">les lundis et mardis, de 10h &#224; 18h</chron:label>
   <chron:specification rdf:resource="#CompoundChronex_1"/>
   <chron:specification rdf:resource="#ChronexInterval_1"/>
</chron:CompoundChronex>

<ap:AccessPeriod rdf:about="#AccessPeriod_1">
   <ap:label rdf:datatype="&xsd;string">ouvert les lundis et mardis, de 10h &#224; 18h</ap:label>
   <ap:accessibleDuring rdf:resource="#CompoundChronex_2"/>
</ap:AccessPeriod>
```

Etape 3 : sérialisation des sorties au format iCalendar

Le système produit un ensemble d'intervalles calendaires représentés au format iCalendar précisant les plages où le service concerné est accessible pour la semaine du 30 avril 2012 au 6 mai 2012.

```
BEGIN:VCALENDAR
CALSCALE:GREGORIAN
PRODID:-//Mondeca 2010//EN
VERSION:2.0
        BEGIN:VEVENT
        DTSTAMP:20120607T091134Z
                DTSTART:20120430T100000
                DTEND:20120430T190000
        SUMMARY:OPENED
        UID:20120607T091134Z-1@mondeca

        BEGIN:VEVENT
        DTSTAMP:20120607T091134Z
```

```
        DTSTART:20120501T100000
        DTEND:20120501T190000
    SUMMARY:OPENED
    UID:20120607T091135Z-2@mondeca
END:VEVENT
END:VCALENDAR
```

Pour évaluer l'outil, 150 dates et horaires d'ouverture de musées ont été extraits d'un ensemble de données mises en ligne par Etalab, dans lequel on trouve la description en langage naturel de plus de 12 000 dates et horaires d'ouverture de musées. On a écarté du sous corpus d'évaluation initial les dates et horaires qui font intervenir des références insolubles (on a pu trouver par exemple un énoncé précisant qu'un musée est fermé « certains jours fériés ») ou qui mêlent des informations sur plusieurs objets (ouverture d'un musée et ouverture des caisses, par exemple), ou encore qui décrivent des informations contradictoires ou mal saisies (« *14h90* »). Après nettoyage du corpus d'évaluation initial, on obtient un corpus exploitable pour l'évaluation de 118 énoncés définissant des dates et horaires d'ouverture de musées.

Là encore, il faut avoir à l'esprit que le choix du corpus influe considérablement sur la qualité mesurée du système : par exemple, les dates et horaires des musées sont rarement saisonniers (ou en tout cas, cette saisonnalité est décrite précisément, et non juste par les mentions de « hautes » et « basses saisons ») : or la notion de saisonnalité n'est pas prise en compte par le système.

L'évaluation du système TKA a consisté à vérifier la cohérence des sorties produites, compte-tenu des énoncés fournis en entrée. La mesure d'un taux de rappel fait moins sens dans ce cadre, puisque seuls des énoncés pertinents sont soumis au système (ce dont on s'est assuré en nettoyant le corpus d'évaluation). L'évaluation des trois tâches successives (annotation, transduction au format RDF et transduction au format iCalendar sur une fenêtre de temps donnée) a ainsi été mesurée en divisant le nombre de sorties conformes aux attentes produites par le système par le nombre d'énoncés du corpus de test. Ici encore, faute de ressources, nous avons nous-mêmes effectués ces évaluations. Cependant, au moins pour l'étape de transduction où le système produit des intervalles calendaires, définissant en extension, sur une semaine, les plages d'accessibilité d'un service, la vérification ne devait pas pouvoir donner lieu, théoriquement, à des différences d'appréciation d'un évaluateur à l'autre.

8.2.2 Les résultats de l'évaluation

Le corpus présente un assez grand nombre de cas « simples » (66 au total) ne contenant qu'un adverbial (par exemple, un adverbial décrivant une plage horaire telle que « *de 10h à 19h* »). La complexité des énoncés donnés en entrée du système est donc extrêmement variable. Pour tenir compte de cette complexité dans l'évaluation, nous avons isolé du corpus une sous-partie considérée comme étant plus « complexe », cette complexité étant définie par la présence d'au moins une relation de composition entre deux adverbiaux (par exemple, la relation de spécification dans entre les adverbiaux « *du lundi au vendredi* » et « *de 10h à 19h* » pour l'énoncé « *ouvert du lundi au vendredi, de 10h à 19h* »). On produit ainsi les mesures d'évaluation pour le corpus de test dans son

intégralité (cf. tab. 11) et pour la sous partie de ce corpus où les énoncés contiennent au moins une relation de composition entre adverbiaux calendaires (52 énoncés au total sur 118) (cf. tab. 12).

Tâches	**scores**
Annotation	0,89
Transduction format RDF	0,85
Transduction format iCal	0,84

Tableau 11 : Résultats de l'évaluation du système d'assistance à la saisie de dates et horaires d'ouverture sur l'ensemble du corpus

Tâches	**scores**
Annotation	0,75
Transduction format RDF	0,67
Transduction format iCal	0,63

Tableau 12 : Résultats de l'évaluation du système d'assistance à la saisie de dates et horaires d'ouverture sur un sous ensemble du corpus où les expressions testées contiennent au moins une relation de composition entre adverbiaux

Sur le corpus de test dans son intégralité, le système génère des erreurs d'annotation dans 10% des cas, des erreurs de transformation au format RDF dans 15% des cas et des erreurs de transduction au format iCalendar dans 16% des cas environ. Le tableau 2 montre que les taux d'erreurs sont directement fonction de la complexité des énoncés soumis au système.

Les erreurs d'annotation relèvent des grammaires d'annotation et des post-traitements sur ces annotations destinés à les compléter. Comme bien souvent pour ce type de ressources, les règles sont très sensibles aux erreurs typographiques (« *Du 01/04 au 30/06 et 01/09 au 30/09* » au lieu de « *Du 01/04 au 30/06 et du 01/09 au 30/09* ») et aux variations lexicales. Or les énoncés définissant les horaires d'accessibilité d'un service sont très souvent rédigés dans un style télégraphique (« *1e quinz sept.* » par exemple), ce qui peut créer parfois des ambiguïtés. Prenons par exemple, le cas de l'énoncé suivant sur lequel le système a produit une annotation erronée : « *De mai à sept. lundi, mercredi, jeudi, vendredi: 14h00 - 18h00 Samedi, dimanche : 10h00 - 12h30 et 14h00 - 18h00* ». Les erreurs d'annotation de cet énoncé viennent du jeu de règles permettant de déterminer les associations entre adverbiaux, qui dépendent étroitement des indices lexicaux et typographiques.

Dans cet énoncé, les règles permettent de déterminer que l'adverbial « *lundi, mercredi, jeudi, vendredi* » est un adverbial composé formé d'adverbiaux associés par une relation de Concaténation. Elles permettent également de déterminer que cet adverbial composé est lié à l'adverbial « *14h00-18h00* » par une association de type Spécification. De même, ces règles permettent de déterminer que « *Samedi, dimanche : 10h00 – 12h30* » forme un adverbial composé. Le problème vient de l'absence d'indice d'une association de Concaténation entre les adverbiaux composés « *lundi,*

mercredi, jeudi, vendredi: 14h00 - 18h00 », d'une part, et « *Samedi, dimanche : 10h00 - 12h30 et 14h00 - 18h00* », d'autre part. En l'absence d'indice lexical, cette association n'est pas déterminée, si bien qu'ensuite l'adverbial « *De mai à sept.* » n'est lié par une association de Spécification qu'à l'adverbial « *lundi, mercredi, jeudi, vendredi: 14h00 - 18h00* » et pas à l'adverbial « *Samedi, dimanche : 10h00 – 12h30* ». Dit autrement, cela revient à considérer que le musée dont on décrit les horaires est ouvert toute l'année les samedis et dimanches de 10h à 12h30, et une partie seulement de l'année, les lundis, mercredis, jeudis et vendredis.

De façon attendue, les erreurs d'annotation se répercutent sur les traitements ultérieurs (la transformation au format RDF et la transduction vers des intervalles calendaires sur une fenêtre de temps donnée). En revanche, la différence des taux d'erreurs entre l'annotation et ces traitements ultérieurs s'expliquent par des problèmes au niveau logiciel.

De façon générale, ces résultats montrent qu'une intervention manuelle sur l'énoncé d'origine est nécessaire, sur ce corpus, dans un peu plus de 17% des cas. Sur le sous ensemble du corpus de test dont les expressions soumises au système contiennent au moins une relation de composition, le système génère des données conformes aux attentes dans 75% des cas. Ceci signifie qu'une intervention manuelle sur le texte d'origine serait nécessaire dans 25% des cas. Il ne faut pas conclure de ces mesures que le système permette de se passer d'une intervention humaine sur les cas où il réussit chacune des étapes du processus d'acquisition, puisqu'il faut encore repérer les cas d'échec. Les premières règles que nous avons mises en œuvre pour détecter les incohérences permettent de les repérer dans 74% des cas.

8.3 Evaluation du système de recherche d'information

8.3.1 Méthodologie

L'évaluation du système de recherche d'information est menée en plusieurs étapes : on évalue d'abord l'algorithme de sélection et de tri par pertinence des adverbiaux calendaires présenté dans la section 5.3 (on évalue son implémentation), puis le moteur de recherche dans son ensemble.

Pour évaluer les systèmes de recherche d'information, en particulier les moteurs de recherche qui ordonnent des documents par pertinence, (Järvelin et Kekäläinen, 2002) décrivent un protocole d'évaluation qui tend à s'imposer, s'appuyant sur un ensemble gradué de mesures de pertinence que l'on peut associer aux documents (*graded relevance*). Ce protocole d'évaluation a été pensé à l'origine afin d'affiner la mesure plus traditionnelle de la précision moyenne (*Mean Average Precision* ou MAP), qui ne permet que de différencier des documents pertinents par rapport à des documents non pertinents, sans graduation. L'idée de l'introduction d'une graduation est qu'il est utile de pouvoir préciser qu'un document peut-être plus ou moins pertinent. Le protocole décrit à l'origine par (Järvelin et Kekäläinen, 2002) et repris dans (Büttcher *et al.*, 2010) définit pour ce faire une mesure (*Normalized Discounted Cumulative Gain* ou nDCG) qui consiste à comparer la liste ordonnée de documents produite par le système évalué avec une liste obtenue à partir de la catégorisation manuelle des documents, rangés chacun sur une échelle graduée de pertinence.

(Najork *et al.*, 2007), par exemple, proposent ainsi une échelle de six gradations : pour une requête donnée, un document peut obtenir une mesure évaluant son intérêt comme étant maximal (*definitive*), très pertinent (*excellent*), pertinent (*good*), relativement pertinent (*fair*), non pertinent (*bad*) ou encore être évalué comme pénalisant la qualité des résultats (*detrimental*).

Pour produire la mesure de gains cumulés normalisés (nDCG) à partir d'une échelle de graduation et d'évaluation manuelle, il est nécessaire de calculer un vecteur idéal. Souvent ce vecteur est composé de l'ensemble ordonné des dix meilleurs résultats dans l'ensemble des résultats testés pour chaque requête. En comparant ce vecteur idéal avec celui associé aux résultats produits par le moteur, on obtient la mesure que l'on peut noter *nDCG@10*.

Prenons le cas de figure où les dix premiers documents retournés par le moteur sont ordonnés dans l'ordre suivant (à côté de l'ordre du document figure le score obtenu à partir des évaluations manuelles) :

doc 1 => score pertinence : 2
doc 2 => score de pertinence : 3
doc 3 => score de pertinence : 0
doc 4 => score de pertinence : 1
doc 5 => score de pertinence : 0
doc 6 => score de pertinence : 3
doc 7 => score de pertinence : 2
doc 8 => score de pertinence : 0
doc 9 => score de pertinence : 1
doc 10 => score de pertinence : 2

Le vecteur de gain (Gain Vector) associé aux résultats produits par le système est le suivant :

G = <2 ; 3 ; 0 ; 1 ; 0 ; 3 ; 2 ; 0 ; 1 ; 2>

Le vecteur de gains cumulés (Cumulative Gain vector) s'obtient comme suit :

$$CG[k] = \sum_{i=1}^{k} G[i]$$

On a donc :

CG = <2 ; 5 ; 5 ; 6 ; 6 ; 9 ; 11 ; 11 ; 12 ; 14>

Une fonction de réduction ("*discount*") est ensuite appliquée à chaque rang pour pénaliser les documents de faible rang, reflétant ainsi l'effort supplémentaire des utilisateurs pour les atteindre. Le gain cumulé réduit se calcule ainsi, à l'aide d'une fonction de réduction : $log_2\,(1+i)$:

$$DCG[k] = \sum_{i=1}^{k} \frac{G[i]}{log_2(1+i)}$$

On a donc, pour l'exemple :

DCG = <2,00 ; 3,89 ; 3,89 ; 4,32 ; 4,32 ; 5,39 ; 6,06 ; 6,06 ; 6,36 ; 6,94>

Enfin, la dernière étape consiste à normaliser le vecteur réduit de gains cumulés à partir d'un vecteur « idéal ». Le vecteur idéal reflète l'ordre qui maximise les gains cumulés à chaque niveau : il correspond au vecteur que l'on peut associer à la liste ordonnée de documents qu'un moteur idéal reproduisant le comportement des évaluateurs aurait produit. Mettons que ce vecteur idéal soit réparti comme suit :

G' = <3 ; 3 ; 3 ; 2 ; 2 ; 2 ; 1 ; 1 ; 1 ; 0>

Le vecteur idéal cumulé est :

CG' = <3 ; 6 ; 9 ; 11 ; 13 ; 15 ; 16 ; 17 ; 18 ; 18>

Le vecteur idéal réduit est :

DCG' = <3,00 ; 4,89 ; 6,39 ; 7,25 ; 8,03 ; 8,74 ; 9,07 ; 9,39 ; 9,69 ; 9,69>

La normalisation s'obtient de la façon suivante :

$$nDCG[k] = \frac{DCG[k]}{DCG'[k]}$$

Le résultat de l'évaluation est ainsi :

nDCG = <0,67 ; 0,80 ; 0,61 ; 0,60 ; 0,54 ; 0,62 ; 0,67 ; 0,65 ; 0,66 ; 0,72>

Au dixième résultat (nDCG@10), la mesure est donc 0,72.

8.3.2 Evaluation du modèle d'ordonnancement des adverbiaux calendaires

La performance du module de sélection et de tri par pertinence des adverbiaux calendaires utilisé dans le système CaSE a été évaluée à l'aide des mesures dites de Précision Moyenne (*Mean Average Precision*) et des Gains Cumulés Réduits et Normalisés (*Normalized Discounted Cumulative Gain*).

Cette évaluation considère le module de sélection et de tri par pertinence des adverbiaux calendaires à la fois indépendamment :

- du moteur d'annotation : on s'est assuré que tous les adverbiaux présents dans le corpus d'évaluation étaient annotés correctement par le moteur d'annotation,
- du moteur de recherche : on n'évalue pas à ce stade les problèmes qui peuvent s'ajouter du fait du traitement conjoint des mots-clés avec des critères calendaires.

Il s'agit d'évaluer un modèle de tri développé pour faire remonter les adverbiaux calendaires les plus pertinents d'un corpus par rapport à un adverbial considéré comme une requête.

Dans notre cas de figure, l'évaluation du modèle de sélection et de tri par pertinence a porté sur un corpus de 90 adverbiaux calendaires, scindé en trois ensembles. Chaque ensemble de 30 adverbiaux a été soumis à deux évaluateurs. Six évaluateurs différents au total ont participé à l'évaluation. Cinq adverbiaux calendaires considérés comme des requêtes étaient associés à chaque ensemble testé. Les évaluateurs devaient assigner une échelle graduée de pertinence allant de 0 à 4 à chaque adverbial de leur sous-corpus pour chacune des cinq requêtes (cf. tableau 13). On produit en annexe un exemple complet d'une fiche d'évaluation (cf. annexe 3). Pour associer une note finale de pertinence aux adverbiaux évalués, on fait la somme des deux mesures produites par les évaluateurs.

	Requête 1 avant 1905			Requête 2 au début du XIXe siècle		
Corpus	Eval 1	Eval 2	Total	Eval 1	Eval 2	Total
en 1905	1	0	1	0	0	0
en 1852	4	2	6	1	1	2
de 1933 à 1938	0	0	0	0	0	0
en septembre 1977	0	0	0	0	0	0
au matin du 21 janvier 1793	2	2	4	0	1	1
En août 1807	3	2	5	4	3	7
au XIXe siècle	4	2	6	2	1	3

Tableau 13 : Extrait d'une synthèse d'un corpus soumis à deux évaluateurs pour l'évaluation du module d'ordonnancement des adverbiaux calendaires

Afin d'obtenir un score de précision moyenne (MAP), où les documents ne sont séparés qu'en deux classes (pertinent/non pertinent), les scores inférieurs ou égal à 4 sont considérés comme non pertinents, ceux qui sont supérieurs étant considérés comme pertinents.

Le recours à deux évaluateurs pour chaque sous-ensemble du corpus soumis à l'évaluation a permis également de mesurer l'accord inter-évaluateurs. Pour cette expérience, prise telle quelle, la mesure classique d'accord dite *Kappa* ne convenait pas. Cette mesure (Cohen, 1960) consiste à comparer la proportion d'accord observée effectivement P_o pour la catégorisation des données et à celle correspondant à une catégorisation aléatoire P_a. Ces proportions sont comparées de la façon suivante :

$$K = P_o - P_a / (1 - P_a)$$

Cependant, dans le cadre de notre évaluation, les scores attribués à chaque adverbial cible forment une échelle graduée de 0 à 4 : il ne s'agit donc pas d'une classification dans des catégories disjointes et étanches, qui n'entretiendraient aucun rapport entre elles. Un score de pertinence de 2 est plus proche d'un score de 3 que de 4 et ainsi de suite. Aussi, si un évaluateur a cru bon d'attribuer une note de 4 à une adverbial cible pour une requête donnée et que l'autre évaluateur a pour sa part attribué la note de 3, la mesure de leur accord doit tenir compte de cette faible différence d'appréciation. L'accord doit donc être considéré comme étant moins grand si le second évaluateur a attribué la note de 2, par exemple.

Dans ce cas de figure, il existe une mesure dite du Kappa pondéré (*Weighted Kappa*) (Cohen, 1968). Les poids utilisés pour pondérer le calcul du Kappa ont été organisés de façon linéaire, dans la mesure où la différence entre deux catégories voisines était identique dans toute la graduation. En considérant k catégories, les poids sont calculés de la façon suivante :

$$w_i = 1 - \frac{i}{k - 1}$$

Le tableau 14 détaille la mesure d'accord inter-annotateur, la mesure de la précision moyenne (MAP) et la mesure nDCG :

mesures	scores
Kw	0.70
MAP	0.86
nDCG	0.93

Tab. 14 : Résultats de l'évaluation du système d'ordonnancement des adverbiaux calendaires

Dans le cadre de notre évaluation la mesure du Kappa (*Kw* = 0.70) tend à montrer que l'accord entre évaluateurs est plutôt bon, sachant qu'il y avait cinq catégories possibles. Les résultats de l'évaluation (MAP 0.86 et nDCG 0.93) montrent que le modèle d'ordonnancement des adverbiaux calendaires a un comportement très similaire à celui des évaluateurs humains.

8.3.3 Evaluation du système CaSE

Pour évaluer le système CaSE, nous partons du principe que pour industrialiser le prototype, plutôt que de soumettre les requêtes saisies par les utilisateurs au système d'annotation, il serait préférable d'en contraindre la saisie par une syntaxe constituée d'opérateurs simples. Ceci à la fois (i) parce que cette approche est plus conforme à la saisie sous forme de mots-clés reliés éventuellement par des opérateurs, (ii) parce que les contraintes des traitements en temps réel s'accommodent mal du temps de traitement de la chaîne d'annotation et enfin (iii) parce que cela permet d'éviter les éventuelles erreurs du système d'annotation lors de l'interprétation des requêtes. Ainsi, afin d'isoler l'évaluation du moteur de recherche de celle de l'annotation des requêtes, on s'est ainsi assuré que les requêtes testées ont été correctement interprétées par le moteur d'annotation.

Nous avons réuni un corpus de 28 requêtes en français qui combinent des mots-clés et un critère calendaire. On produit en annexe 4 le corpus des requêtes testées. On rappelle que le système ne traite que des requêtes dont le critère calendaire ne désigne qu'une zone unique sur le calendrier (cf. section 7.1.1). Ces requêtes ont été soumises au moteur de recherche expérimental ainsi qu'au moteur de Google. La pertinence des résultats a été analysée par deux évaluateurs.

Les requêtes ont été testées sur un ensemble d'articles de Wikipedia collecté de la façon suivante : pour chaque requête testée sur Google, un robot récupère les cent premiers résultats retournés à l'utilisateur ; ces cent résultats sont ensuite indexés par le moteur expérimental. En somme, ce

protocole permet d'observer la façon dont le moteur réordonne à sa manière les cent premiers résultats proposés par Google.

C'est là une limite et un biais dans l'évaluation, dans la mesure où si le moteur avait pu être testé sur l'ensemble des articles de Wikipedia en langue française, les résultats auraient parfois sans doute été différents : des articles que Google ne retourne pas dans les premiers résultats auraient pu apparaître et ainsi jouer sur les taux de rappel et de précision. En dépit de ce biais, l'évaluation rend sensible la différence entre les deux moteurs dans la façon dont ils ordonnancent les résultats et il est possible d'en tirer certaines conclusions.

Sont évalués les 10 premiers résultats fournis par chacun des moteurs. Deux évaluateurs se sont chargés en parallèle du dépouillement des résultats et ont dû attribuer un score de 0 à 3 à chauqe document retourné, 3 étant le meilleur score, 0 le moins bon. Sur la base de ces évaluations, la pertinence des résultats a été mesurée à l'aide de deux indicateurs : la précision moyenne (*Mean Average Precision* ou *MAP*) (Buckley et Voorhees, 2004) et la mesure des gains cumulés normalisés (*normalized Discounted Cumulative Gain* ou *nDCG*) (Järvelin et Kekäläinen, 2002). L'échelle de pertinence peut être décrite comme suit :

Score 3 = très pertinent
Score 2 = pertinent mais plutôt spécifique ou au contraire trop générique.
Score 1 = peu pertinent, très spécifique, anecdotique
Score 0 = non pertinent

Pour produire la mesure de la précision moyenne (MAP), qui ne permet que de différencier des documents pertinents par rapport à des documents non pertinents, l'échelle utilisée pour évaluer la pertinence d'un document a été modifiée comme suit : un document qui s'est vu attribué un score de 0 ou 1 est considéré comme un document non pertinent ; un document qui s'est vu attribué un score de 2 ou de 3 est considéré comme un document pertinent.

Cette mesure paraît moins adaptée pour deux raisons. La pertinence graduée (nDCG) produit une mesure plus fine, dans la mesure où il est possible qu'une partie seulement d'un document soit pertinente. En outre, la précision moyenne ne permet pas d'évaluer le taux de rappel : par exemple, une requête pour laquelle il n'y a que très peu de bons résultats, produira un score de précision très faible, alors même que le corpus ne contient pas nécessairement davantage de bons documents. De toute évidence, dans un corpus volumineux (tel que Wikipedia), mesurer le taux de rappel n'est pas possible. Pour autant, il est possible d'avoir une mesure qui en donne une sorte d'aperçu : en comparant avec un vecteur idéal pris dans le meilleur ensemble des résultats produits par les deux moteurs, on peut avoir une indication du taux de rappel : en effet, s'il y a peu de bons résultats dans le vecteur idéal, contrairement à la précision moyenne, la valeur de la mesure nDCG n'en sera pas affectée, puisqu'elle mesure l'écart avec le vecteur idéal.

Le vecteur idéal, dans notre cas de figure, est constitué des dix meilleurs documents pris dans les vecteurs de CaSE et de Google.

	CaSE	Google
MAP - Evaluateur 1	*0,56*	*0,54*
MAP - Evaluateur 2	*0,43*	*0,34*
MAP – moyenne	***0,50***	***0,44***
nDCG - Evaluateur 1	*0,75*	*0,62*
nDCG - Evaluateur 2	*0,70*	*0,60*
nDCG – moyenne	***0,72***	***0,61***

Tab. 15 : Résultats de l'évaluation comparative de CaSE et de Google

Si aussi bien la précision moyenne (MAP@10) que la mesure des gains cumulés normalisés (nDCG@10) présente une légère différence en faveur de CaSE par rapport à Google, l'écart est plus sensible pour l'indicateur nDCG. Ceci s'explique précisément par la plus grande finesse de cette mesure. La mesure nDCG est populaire pour la recherche sur le Web, où la qualité des pages varie beaucoup et où un grand nombre de pages plus ou moins pertinentes apparaît dans la collection (Najork *et al.*, 2007).

Aussi minces soient les différences, il reste possible d'en tirer quelques conclusions. La première d'entre elles, est que, bien qu'avec un traitement des mots-clés très sommaire (s'appuyant uniquement sur le TF-IDF et sans ressources linguistiques tels que des dictionnaires de synonymes), le système CaSE obtient des résultats comparables à ceux de Google, ce qui laisse à penser que la différence devrait s'accentuer si les deux systèmes disposaient de ressources équivalentes. Il faut ensuite regarder plus précisément les résultats en fonction de la nature du critère calendaire et du critère thématique exprimé dans les requêtes.

Google présente de bons résultats si la période recherchée est très fréquemment associée aux mots-clés (par exemple « *prohibition au début des années 30* »), car alors le critère calendaire peut apparaître dans le texte et son traitement comme mot-clé renforce le score de la page. A l'inverse, si la période n'est pas fréquemment associée aux mots-clés, Google renvoie généralement des mauvais résultats (par exemple « *choc pétrolier de 1980 à 2000* »). En effet, le critère calendaire étant traité comme un mot-clé, sa présence peut perturber considérablement les résultats s'il n'apparaît pas tel quel dans les pages pertinentes pour une requête donnée : de fait, son traitement comme mot-clé peut favoriser des pages où tout ou partie de l'expression apparaît, même sans rapport avec la recherche thématique (cf. par exemple la requête « *langue française de 1520 à 1600* »).

CaSE produit de moins bons résultats lorsque les mots-clés de la requête sont polysémiques : ceci s'explique du fait que le modèle de traitement des mots-clés est très rudimentaire. En particulier, le moteur ne prend pas en charge la synonymie, les acronymes (PCF = Parti Communiste Français), ni la proximité entre termes (livre/lecture, littérature/écrivain).

Par ailleurs, si le critère calendaire exprimé dans la requête couvre une période de temps très étendue, alors ce critère devient moins discriminant (par exemple « *crise économique au XIXe siècle* »). Ainsi un document pertinent du point de vue du thème (mots-clés) a davantage de chances d'être pertinent également du point de vue du critère calendaire. Le modèle de pertinence des mots-clés devient alors plus discriminant.

Il est une autre donnée comparative intéressante qui ressort de l'évaluation, à savoir le temps passé à dépouiller les résultats : 709 minutes au total pour analyser la pertinence des 10 premiers résultats de Google pour les 28 requêtes testées (25 minutes en moyenne pour analyser les résultats d'une requête), contre 369 minutes avec le système CaSE (13 minutes en moyenne). Cette différence non négligeable s'explique par la façon dont les deux systèmes présentent les résultats : dans la liste des résultats, Google, comme tous les moteurs de recherche documentaire, ne fournit qu'un court extrait du document (*snippet*), qui ne permet pas toujours à l'utilisateur de se faire une idée de sa pertinence. Il faut parfois parcourir les documents dans le détail pour s'assurer qu'ils sont pertinents. Le système CaSE fournit pour sa part une liste des extraits des documents les plus pertinents (jusqu'à quatre phrases). En outre, il permet d'explorer chacun des documents individuellement en fonction de critères calendaires (au moyen de la frise chronologique ou de requêtes) : il est donc beaucoup plus simple de s'assurer qu'un document est pertinent. On touche là à un des aspects singuliers du système CaSE, à savoir sa capacité à permettre de faire des recherches intra-documentaires.

8.3.4 Limites de l'évaluation

Les mesures d'évaluation que l'on a fournies sont des indicateurs classiques utilisées pour l'évaluation de moteur de recherche documentaire. Elles laissent néanmoins dans l'ombre certaines facettes intéressantes du système CaSE, en particulier :

- la capacité du système à associer des périodes de temps à des mots-clés

Le système permet d'explorer temporellement un corpus en s'appuyant sur les adverbiaux calendaires associés à des mots-clés, notamment pour les requêtes ne contenant pas de critère calendaire. Ceci doit permettre par exemple de faciliter la constitution de biographies (cf. par exemple les résultats proposés pour des requêtes telles que *Balzac* ou *Clémenceau*), mais aussi de localiser certains événements, mouvements ou périodes (par exemple *croisades*, *commune de Paris*, *peste*, *gothique*).

- la capacité du système à circonscrire les parties les plus pertinentes d'un document long

Ce n'est pas toujours le choix d'un document, mais le fait de trouver des parties pertinentes d'un document qui peut avoir de la valeur pour l'utilisateur : parfois, seule une phrase, un paragraphe ou une section suffit à répondre aux besoins d'information de l'utilisateur. Or les moteurs de recherche grand public ne considèrent pas des unités inférieures au document. Le moteur CaSE considère lui les documents comme des ensembles de phrases. L'utilisateur peut ainsi faire des requêtes au sein d'un seul document, une fois qu'il l'estime pertinent (recherche intra-documentaire).

- la capacité à désambigüiser le critère calendaire de la requête à l'aide de la frise chronologique.

La frise chronologique est en effet une alternative à l'expression d'un critère calendaire : il est possible d'interagir avec cet outil pour affiner une requête et d'explorer un corpus ou un texte en balayant progressivement la ligne du temps, à différentes échelles.

En dépit de ces limites, l'évaluation montre ainsi que, même avec un traitement des mots-clés des plus simples, fourni de façon standard avec un moteur de recherche open-source, une représentation de la sémantique des adverbiaux calendaires issue d'une analyse linguistique permet d'améliorer sensiblement les résultats pour des requêtes contenant des critères calendaires. L'évaluation confirme ce qu'on souhaitait montrer, à savoir le bénéfice que les moteurs de recherche existant pourraient tirer de cette approche.

Chapitre 9 : Bilan et perspectives

9.1 Bilan

Les adverbiaux de localisation temporelle sont des unités textuelles privilégiées pour l'analyse de l'ancrage temporel des situations décrites dans les textes. En parcourant des textes à travers ces références, il devient possible d'observer sous l'angle chronologique les thèmes qui apparaissent dans leur contexte : on a vu qu'il était possible ainsi de suivre l'évolution dans le temps d'un thème, qu'il s'agisse d'un événement, d'une personne, d'un mouvement artistique, d'une période historique. La possibilité de parcourir des textes selon des spécifications nées de l'analyse linguistique des adverbiaux de localisation temporelle - plutôt que des « expressions temporelles » conçues comme des entités nommées - ouvre alors une piste originale pour la recherche d'information.

Les adverbiaux de localisation temporelle présentent ainsi un intérêt pour la recherche documentaire et la navigation dans les textes, mais également pour la gestion des connaissances : à la fois parce que, de leur analyse linguistique, il est possible de faire ressortir des opérations sémantiques qui peuvent contribuer à rendre plus expressifs les formats de représentations des données temporelles (données qui sont transverses et communes à de nombreux domaines d'application) et parce que la mise en œuvre de ressources pour les annoter automatiquement peut permettre de faciliter l'acquisition de connaissances à partir des textes.

Du côté applicatif, nous avons présentés plusieurs modules pour les traitements des adverbiaux de localisation temporelle. Nous avons ainsi présenté deux modules génériques, (1) un système d'annotation, de structuration et de transduction des adverbiaux de localisation temporelle présents

dans les textes et (2) un système de recherche d'information expérimental, le moteur de recherche CaSE, pour la recherche documentaire sous l'angle calendaire ; enfin, (3) nous avons présenté un système orienté sur un cas d'application industriel précis, le prototype TKA, pour assister et faciliter la saisie de dates et horaires d'ouverture dans une base de connaissances.

D'un point de vue formel, ces modules reposent sur une représentation *fonctionnelle* des adverbiaux de localisation temporelle donnée sous la forme d'une succession d'opérations sémantiques et sur une représentation dite *référentielle*, donnée sous la forme d'intervalles calendaires. Ils reposent également sur une heuristique de transduction, permettant de passer de la première représentation vers la seconde. Le système de recherche documentaire que nous avons présenté s'appuie à la fois sur ces deux modes de représentation et sur un algorithme de filtrage et de tri par pertinence d'une sous-partie des adverbiaux de localisation temporelle, les adverbiaux calendaires.

D'un point de vue théorique, nous avons essayé de montrer que l'analyse des adverbiaux de localisation temporelle présents dans les textes pouvait gagner à s'appuyer une représentation formelle découlant d'une analyse linguistique, parce que cette approche permet d'isoler des structures syntaxiques et sémantiques plus cohérentes d'un point de vue linguistique que ne le font généralement les modèles sous-jacents qui prédominent dans le traitement automatique des langues, dans la recherche d'information et dans le domaine de la gestion des connaissances, où les représentations temporelles, souvent ramenées aux représentations normées du calendrier, peinent à exprimer et prendre en charge des informations temporelles de granularité variée ou présentant un certain degré d'indétermination.

Dans ces domaines, en effet, le plus souvent on ne s'intéresse aux unités textuelles porteuses d'informations relatives à localisation temporelle qu'en tant qu'elles peuvent être ramenées à des valeurs calendaires, leur description sémantique paraissant être un problème intermédiaire de second plan. Notre objectif était de montrer qu'en renversant le problème – en considérant d'abord la question de la représentation de ces unités textuelles qui peut découler d'une analyse linguistique, avant de s'intéresser à leur transposition sous la forme de valeurs calendaires -, il devenait possible d'apporter des réponses aux difficultés rencontrées par les systèmes d'organisation des connaissances et les systèmes de recherche d'information lorsqu'il s'agit de manipuler certaines informations temporelles.

En renversant ainsi la perspective, on cherche moins à « normaliser » des représentations langagières qu'à les transposer vers des formats qui ne sont pas tous équivalents sémantiquement :

- une représentation formelle d'abord, qui découle d'une analyse linguistique, et des vues simplifiées de cette représentation qui s'expriment (i) dans un schéma d'annotation pour enrichir les textes de métadonnées sur les unités textuelles qui dénotent une expression de localisation temporelle, (ii) dans un schéma RDF pour le stockage sous la forme d'un réseau sémantique dans une base de connaissances et (iii) dans l'implémentation d'un modèle objet qui permet à des agents logiciels d'en manipuler les instances.
- une représentation calendaire qui peut s'exprimer dans différents formats standards et qui présente des propriétés calculatoires bien étudiées.

En posant le problème ainsi, il fait sens de faire dépendre des besoins applicatifs les mécanismes de transduction des représentations que l'analyse linguistique permet de faire émerger vers différents autres formats.

En outre, à côté du paradigme de comparaison entre intervalles de temps qui vise à catégoriser leurs relations (en termes d'inclusion, de précédence, de succession, etc.), nous avons montré qu'il peut être intéressant d'établir des mesures de similarité entre intervalles calendaires - mesures qui permettent de mettre en œuvre des systèmes de recherche d'information offrant la possibilité de fouiller des corpus et des documents sous l'angle chronologique. Cette approche, dont on a montré la faisabilité appliquée à des documents, est en outre reproductible dans le domaine de la recherche d'information appliquée à des données structurées : elle peut alors permettre d'explorer des données auxquelles sont associées des informations de localisation temporelle.

Cette approche répond ainsi en partie à la gestion, à la constitution et à la recherche dans des informations relatives à la localisation temporelle dans un fonds documentaire ou une base de connaissances. Ce sont des problématiques auxquelles sont confrontés les musées, les centres d'archives, les agences de presse, par exemple, lorsqu'ils doivent gérer des collections et des fonds documentaires où la perspective calendaire peut revêtir un intérêt majeur et où les informations relatives à la localisation temporelle ne se satisfont pas toujours des modèles de représentation calendaire très contraints.

Le système d'annotation des adverbiaux de localisation temporelle est un composant qui peut ainsi s'insérer dans une chaîne d'analyse documentaire pour l'enrichissement d'une base de connaissances. Le système CaSE apporte, lui, un mode de recherche complémentaire à la recherche à l'aide de mots-clés, en permettant d'ajouter à des mots-clés l'expression de critères calendaires. Les interactions avec la frise chronologique mises en œuvre dans le système offrent également une manière parallèle de parcourir un corpus de documents ou de ressources sous l'angle calendaire.

En croisant les perspectives abordées successivement dans ce mémoire – la perspective linguistique, la perspective de l'ingénierie des langues, celle de la gestion des connaissances et enfin celle de la recherche d'information -, nous avons donc souhaité montrer qu'il était possible de faire émerger de nouvelles manières de consulter des documents et de contribuer à l'enrichissement de bases de connaissances.

L'évaluation des différentes ressources mises en œuvre pour manipuler des adverbiaux de localisation temporelle laisse à penser que les prototypes que l'on a présentés ne sont pas hors de portée d'une opérationnalisation plus avancée : on espère que la preuve de leur « faisabilité » a été apportée.

9.2 Perspectives

De toute évidence, ce projet de recherche n'épuise pas la question de l'expression de la localisation temporelle dans les textes, ni sous l'angle de leur modélisation, ni sous l'angle de leur exploitation

pour la recherche d'information ou l'acquisition de connaissances à partir des textes. On espère toutefois qu'il a contribué à sa manière à renouveler la façon d'aborder ces problèmes. Plusieurs des pistes qui nous ont retenus un temps n'ont pas pu donner lieu à des développements approfondis, comme la navigation textuelle conduite sous l'angle chronologique, la mise en œuvre d'interactions graphiques avec des informations temporelles ou l'assistance à la production de chronologies. Les trajectoires dessinées dans nos travaux peuvent ainsi être prolongées dans plusieurs directions : sur le plan linguistique, sur le plan de l'ingénierie des langues et sur ceux de la gestion des connaissances et de la recherche d'information.

Sur le plan de l'analyse linguistique, on a vu que l'on pouvait généraliser la catégorie des adverbiaux calendaires pour constituer la catégorie des adverbiaux de localisation temporelle. On a montré que si le type de la référence au cœur de ces unités textuelles peut varier (base calendaire, base déictique, base relative à un « événement »), en revanche, sur un plan formel, l'approche consistant à décrire ces adverbiaux sous la forme d'une succession d'opérations sémantiques demeure opératoire. On a vu néanmoins que cette représentation, que l'on s'est attaché à enrichir par rapport au modèle initial proposé par (Battistelli, 2009), demanderait à être encore affinée : il nous semble ainsi que l'effort de formalisation pourrait être poursuivi dans deux directions au moins.

L'une d'elle consisterait à affiner le modèle et sa sémantique, en associant des transformations spécifiques à chacune des valeurs possibles des opérations utilisées pour décrire les adverbiaux de localisation temporelle. Il s'agit de pouvoir capter, par exemple, au sein des opérations de focalisation, la différence entre « au début de » et « au tout début de », différence qui doit se répercuter sur la façon dont cette opération modifie son opérande, sachant que cet opérande peut être représenté sous la forme d'un intervalle de temps. Les valeurs possibles des opérations à l'aide desquelles on décrit les adverbiaux de localisation temporelle traduisent ainsi des transformations sur des intervalles de temps : l'intervalle obtenu à l'issue de ces transformations correspond à celui que (Gosselin, 2006) propose de nommer « intervalle circonstanciel ».

Une autre direction possible pour poursuivre ce travail consisterait à faire dépendre les valeurs possibles des opérations non plus seulement des marqueurs lexicaux qui composent les adverbiaux, mais aussi d'autres unités discursives qui contribuent à ancrer dans le temps un procès : en effet, afin de pouvoir décrire la différence entre les prépositions « dès » et « depuis » dans deux adverbiaux tels que « dès la mi-mars » et « depuis la mi-mars », il devient nécessaire de considérer non plus seulement des intervalles circonstanciels (ceux que l'on peut associer aux adverbiaux), mais également le point de vue aspectuel à travers lequel est considéré le procès qu'ils contribuent à ancrer dans le temps : sur un plan formel, il ne s'agit plus donc d'associer une certaine transformation sur un intervalle de temps permettant de déterminer l'intervalle circonstanciel à l'aide duquel on peut représenter un adverbial, mais de préciser en quoi ces marqueurs contribuent à déterminer les liens entre cet intervalle et celui qui représente le procès considéré. Ceci rejoint la question de la portée des adverbiaux et de la façon dont ils participent à l'ancrage d'un procès dans le temps.

Ces travaux feraient d'autant plus sens qu'ils pourraient ouvrir des pistes pour améliorer le système de recherche d'information expérimental que l'on a présenté : en effet, on a vu que l'absence d'information sur la portée des adverbiaux de localisation temporelle nous conduisait à rechercher

les termes formant le critère thématique d'une requête dans toute une phrase et non pas seulement sur l'unité discursive que complémente un adverbial.

Sur le plan de l'ingénierie des langues, on a vu qu'il pouvait être intéressant de faire converger le système d'annotation que l'on a décrit avec ceux qui s'appuient le langage TimeML, afin de progresser à la fois sur le terrain de l'annotation automatique des adverbiaux qui n'opèrent par pas une localisation relative au calendrier (comme les adverbiaux qui font intervenir des *événements* au sens de TimeML), mais aussi sur celui du repérage de la portée des adverbiaux de localisation temporelle. Nous avons esquissé une façon d'entrecroiser le langage d'annotation que l'on a présenté, ChronolocationML, avec TimeML.

Ces différentes pistes de travail, qui se situent à la fois au niveau de l'analyse linguistique (la portée des adverbiaux), au niveau de la mise en œuvre de systèmes d'annotation et au niveau de la spécification d'un langage d'annotation permettant de dresser des ponts entre l'approche de TimeML et l'approche complémentaire que l'on propose, sont à l'étude dans le projet ANR Chronolines[59]. Nos prochains travaux en ce sens viseront à établir et diffuser un corpus de référence annoté avec ChronolocationML accompagné d'un guide d'annotation, de sorte à ce qu'il soit possible à la fois d'établir des comparaisons avec TimeML, mais aussi éventuellement de reproduire de telles annotations sur de nouveaux textes ou encore de tester et d'entraîner d'autres systèmes d'annotation, tels que des systèmes d'apprentissage automatique.

Pour ce qui est de la recherche d'information, il pourrait être intéressant d'affiner le processus de transduction de la représentation fonctionnelle associée aux valeurs des adverbiaux calendaires vers une représentation référentielle donnée sous la forme d'intervalles calendaires. Il s'agirait ainsi d'améliorer la qualité du système CaSE pour la recherche d'information, en affinant les critères de filtrage et de tri par pertinence des adverbiaux calendaires. Pour l'heure, le processus consiste à associer un intervalle calendaire ou un ensemble d'intervalles à un adverbial calendaire. Cette représentation pourrait être enrichie pour lui associer non plus un intervalle calendaire, mais plusieurs intervalles (éventuellement des ensembles d'ensembles), afin, par exemple, de conserver dans toute la chaîne des traitements les informations sur la granularité propre à un adverbial : ainsi une adverbial calendaire pourrait être transcrit par exemple sous la forme de deux intervalles, dont l'un représenterait l'étendue temporelle qu'il couvre sur le calendrier et l'autre la granularité initiale de sa base calendaire. Cette information pourrait ensuite être exploitée dans la mesure de la similarité entre deux adverbiaux.

Dans le cadre d'une intégration dans un système plus vaste de recherche d'information, au sein d'une base documentaire ou d'une base de connaissances, il serait intéressant de croiser la perspective chronologique avec d'autres modes de recherche, tels que la navigation par facettes, afin de pouvoir diversifier les parcours possibles au sein d'un document ou d'un corpus. Dans cette optique, il faudrait sans doute également formaliser un langage constitué de quelques opérateurs simples pour faciliter l'expression de critères calendaires dans une requête. On a vu que cette approche serait sans doute plus conforme à l'habitude forgée par les contraintes des moteurs de recherche, qui conduit les utilisateurs à saisir leurs requêtes sous la forme de mots-clés reliés

[59] http://chronolines.fr/

éventuellement par des opérateurs, plutôt que sous la forme d'expressions en langage naturel. Ce langage permettrait aussi de s'accommoder mieux des exigences liées au traitement en temps réel des données : le processus d'analyse du système d'annotation que l'on a décrit pèse en effet sur le temps de restitution des résultats. Enfin, comme tout système d'annotation automatique, il peut produire des erreurs d'analyse qui peuvent ensuite affecter la qualité des résultats.

Une autre piste possible pourrait consister à définir des microformats pour l'ajout de métadonnées dans les pages Web où des informations de localisation temporelle sont présentes. Ces métadonnées pourraient alors être exploitées par les moteurs de recherche dont on voit qu'ils s'approprient progressivement, tout en contribuant à les redéfinir, les technologies du Web Sémantique, en particulier du fait de la puissance d'interrogation qu'offrent les représentations structurées (navigation par facettes, désambiguïsation des requêtes, etc.). En outre, cela permettrait de découpler les problématiques de l'annotation des textes avec celles de la recherche d'information.

9.3 D'enjeux scientifiques à des enjeux industriels

Aujourd'hui encore, l'acquisition des connaissances à partir de textes est surtout pensée comme un processus d'extraction d'information destiné à nourrir des bases de connaissances ou à améliorer la qualité des systèmes de recherche d'information. Parallèlement, la lecture sur support numérique, qui se développe et reproduit encore pour l'heure en grande partie la lecture sur support papier, va vraisemblablement s'accompagner de services autour des textes, comme les services de navigation textuelle permettant d'effectuer différents types de parcours dans les textes. Si les moteurs de recherche enrichissent leurs fonctionnalités pour faire émerger des documents et des ressources pertinentes, ils n'ont pas encore investi complètement le champ de la lecture et de la recherche au sein d'un document précis. Le système CaSE, qui permet de basculer d'une recherche documentaire vers une recherche intra-documentaire, montre qu'une continuité est possible entre les problématiques de la recherche d'information et de la lecture sur support numérique. De nombreuses voies restent à explorer dans cette direction, non pas seulement pour poser automatiquement des métadonnées sur les textes et faciliter l'accès aux documents, mais pour s'approprier des textes, en les parcourant et en déposant dessus sa propre lecture. Dans la redéfinition des activités éditoriales, des pratiques de lecture et de production documentaire, liée à la numérisation des documents, à la variété des supports de lecture, aux outils de recherche et, plus généralement, aux systèmes offrant des interactions avec les textes, le développement de services autour du texte est amené à jouer un rôle important. Il pourrait par exemple être intéressant de proposer des services permettant de naviguer parallèlement dans un texte et sur le paratexte dont on peut l'enrichir. La navigation temporelle dans les textes pourrait ainsi s'intégrer dans un ensemble plus vaste de tels services.

Ainsi, initiés à l'origine par des problématiques industrielles liées à la gestion d'informations temporelles, ces travaux de recherche qui ont notamment visé à affiner un objet d'un point de vue linguistique, en retour, conduisent aussi à renouveler la question des usages possibles des systèmes de recherche d'information et de gestion des connaissances : les problématiques industrielles s'en trouvent ainsi elles-mêmes redéfinies.

Références bibliographiques

Adam, J-M. (2001). Genres de la presse écrite et analyse de discours. In *Semen* n°13, P. U. de Franche-Comté, pp. 7-14.

Ahn D., van Rantwijk J., de Rijke M. (2007). A cascaded machine learning approach to interpreting temporal expressions. In *Proceedings of NAACL-HLT'07* Rochester, NY, USA, April, pp. 284-291.

Allen J. F. (1983). Maintaining knowledge about temporal intervals. In *Communications of the ACM*, 26 (11), pp. 832-843.

Allen J. F., Swift M. et De Beaumont W. (2008). Deep semantic analysis of text. In *Symposium on Semantics in Systems for Text Processing (STEP), Venice, Italy,* pp. 343-354.

Alonso O., Gertz M. et Baeza-Yates R. (2007). On the value of temporal information in information retrieval. In Proceedings of *ACM SIGIR Forum*, 41 (2), pp. 35-41.

Alonso O., Gertz M. et Baeza-Yates R. (2009). Clustering and exploring search results using timeline constructions. In *Proceeding of the 18th ACM conference on Information and knowledge management*, pp. 97–106.

Alonso O., Berberich K., Bedathur S. et Weikum G. (2010). Time-Based Exploration of News Archives. In *HCIR 2010*, New Brunswick, pp. 12-15.

Amardeilh F. (2007). *Web Sémantique et Informatique Linguistique: propositions méthodologiques et réalisation d'une plateforme logicielle*. Thèse de Doctorat. Université de Paris X.

Arikan I., Bedathur S. et Berberich K. (2009). Time Will Tell: Leveraging Temporal Expressions in IR. In Proceedings of *WSDM'09*, Springer, pp. 13-25.

Asher N. et Lascarides A. (2003). *Logics of Conversation*. Cambridge University Press.

Asur S. et Buehrer G. (2009). Temporal analysis of web search query-click data. In *Proceedings of SNA-KDD'09*, ACM, Paris, France.

Auer S., Bizer C., Kobilarov G., Lehmann J., Cyganiak R. et Ives Z. (2007). Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, pp. 722–735.

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

Bach E. (1986). The algebra of events. In *Linguistics and philosophy*, vol. 9 (1), pp. 5-16.

Bacot P., Douzou L., Honoré J. (2008). Chrononymes. La politisation du temps. In *Mots. Les langages du politique*, 87 (2), pp. 5–12.

Baeza-Yates R. (2005). Searching the future. In S. Dominich, I. Ounis, & J.-Y. Nie (Ed.)*, MFIR2005 Proceedings of the Mathematical/Formal Methods in Information Retrieval Workshop associated to SIGIR 2005, 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Salvador, Brazil. August 15-19, ACM Press.

Battistelli D., Schwer S. R. et Minel J.-L. (2006). Représentation des expressions calendaires dans les textes : vers une application à la lecture assistée de biographies. In *TAL*, vol. 47, pp. 11-37.

Battistelli D. et Chagnoux M. (2007). Représenter la dynamique énonciative et modale de textes. In *Actes de TALN'07*, pp. 13-23.

Battistelli D., Couto J., Minel J.-L. et Schwer S. R. (2008). Représentation algébrique des expressions calendaires et vue calendaire d'un texte. In *Actes de TALN'08* (1), pp. 9-13.

Battistelli D. (2009). La Temporalité Linguistique : Circonscrire un objet d'analyse ainsi que des finalités à cette analyse. *Habilitation à Diriger des Recherches*. Université Paris-Ouest Nanterre La Défense (Paris 10), novembre 2009.

Battistelli D. (2011a). Linguistique et recherche d'information : la problématique du temps. In *Hermes*, coll. *Traitement de l'Information*, avril 2011, 250 p.

Battistelli D., Cori M., Minel J.-L. et Teissèdre C. (2011b). Semantics of Calendar Adverbials for Information Retrieval. In *Proceedings of ISMIS 2011*, June 28-30, Warsaw, Poland, pp. 622-631.

Battistelli D., Cori M., Minel J.-L. et Teissèdre C. (2012). Information Retrieval: Ranking Results according to Calendar Criteria. *IPMU 2012*, July 9-13, Catania, Italy (à paraître).

Bécher G., Enjalbert P., Fievé E., Gosselin L., Lévy F. et Ligozat G. (2005). *Rapport du Projet OGRE - Ordres de Grandeur et REpétition*. Rapport technique, CTAN, 93p.

Bécher G., Clerin-Debard F. et Enjalbert P. (2000). A qualitative Model for Time Granularity. In *Computational Intelligence*, Vol. 16 (2), pp. 137-175.

Bejan C. A. et Harabagiu S. (2010). Unsupervised Event Coreference Resolution with Rich Linguistic Features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1412-1422.

Benveniste E. (1970). L'appareil formel de l'énonciation. In *Langages*, vol. 5 (17), pp. 12-18.

Benveniste E. (1974). Problèmes de linguistique générale. Paris *Gallimard*, vol. II.

Berberich K., Bedathur S., Neumann T. et Weikum G. (2007). A Time Machine for Text Search. In *Proceedings of ACM SIGIR'2007*, pp. 519-526.

Berberich K., Bedathur S., Alonso O. et Weikum G. (2010). A language modeling approach for temporal information needs. In *Advances in Information Retrieval (ECIR 2010)*, Springer, pp. 13–25.

Bärenfänger M., Lobin H. Lüngen H. et Mirco H. (2008). OWL ontologies as a resource for discourse parsing. In *LDV-Forum. GLDV-Journal for Computational Linguistics and Language Technology*, vol. 23 (1), pp.17-26.

Berners-Lee T., Hendler J. et Lasilla O. (2001). *The Semantic Web*, Scientific American, vol. 284 (5), May 2001, pp. 34-43.

Bettini C., Jajodia S. et Wang S. (Eds). (2000). *Time granularities in Databases, Datamining, and Temporal Reasoning*. Springer-Verlag New York Inc.

Bevort E., Bonvoisin S., Frémont P. et Savino J. (éds). (1999). *Historiens et géographes face à la médiatisation de l'événement*. Paris, Centre National de Documentation Pédagogique, collection *Documents, actes et rapports pour l'éducation*. 196 p.

Bilhaut F., Ho-Dac M., Borillo A., Charnois T., Enjalbert P., Le Draoulec A., Mathet H., Pery-Woodley M.-P. et Sarda L. (2003). Indexation discursive pour la navigation intradocumentaire : cadres temporels et spatiaux dans l'information géographique. In *Actes de TALN'03*, Batz-sur-Mer, pp. 315-320.

Bittar A. (2008). Annotation des informations temporelles dans des textes en français. In *Actes de RECITAL 2008*, 18 (1), pp. 139-157.

Bittar A. et Danlos L. (2009). Intégration des constructions à verbe support dans TimeML. In *Actes de TALN 2009.* Session posters, Senlis, France.

Bittar A. (2009). Annotation of events and temporal expressions in French texts. In *Proceedings of the 19th Meeting of Computational Linguistics in the Netherlands*. Morristown, NJ, USA, Association for Computational Linguistics, pp. 25-38.

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.

Bizer C., Heath T. et Berners-Lee T. (2009). Linked data-the story so far. In *Heath T., Hepp M., and Bizer C. (eds.). International Journal on Semantic Web and Information Systems (IJSWIS)* (Special Issue on Linked Data), vol. 5 (3), pp. 1-22.

Blumenthal P. (1990). Classement des adverbes : Pas la couleur, rien que la nuance ?. In *Langue française*, 88, pp. 41-50.

Bontcheva, K. et Cunningham, H. (2003). The semantic web: A new opportunity and challenge for human language technology. In *Proceedings of the Workshop on Human Language Technology for the Semantic Web and Web Services.* Held in conjunction with the 2nd International Semantic Web Conference (ISWC'03). H. Cunningham, Y. Ding, A. Kiryakov (eds). Florida, USA. [en ligne] http://www.gate.ac.uk/sale/iswc03/iswc03.pdf

Borillo A. (1998). Les Adverbes de référence temporelle comme connecteurs temporels de discours. In *Temps et discours*. Vogeleer S., Borillo A., Vetters C. et Vuillaume M. (eds.), pp 131-145.

Bourigault D., Aussenac-Gilles N. et Charlet, J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes : Un cadre unificateur pour trois études de cas. In *Revue d'Intelligence Artificielle*, vol. 18, n° 1, pp. 87-110.

Büchi E. (1996). Les structures du « Französiches Etymologisches Wöterbuch » : Recherches métalexicographiques et métalexicologiques. Tübingen, Niemeyer.

Buckley C. et Voorhees E. M. (2004). Retrieval Evaluation with Incomplete Information. In *Proceedings of ACM SIGIR'04 conference on Research and development in information retrieval*, July 25-29, Sheffield, South Yorkshire, UK, pp. 25-32.

Büttcher S., Clarke C. et Gordon V. C. (2010). Information Retrieval: Implementing and Evaluating Search Engines. *The MIT Press*, 632p.

Calabrese Steimberg L. (2006). La construction de la mémoire historico-médiatique à travers les désignations d'événements. *Travaux du Cercle belge de linguistique, n° 1, p. 1-16.* [en ligne] http://webh01.ua.ac.be/linguist/online/paps2006/cal2006.pdf

Calabrese Steimberg, L. (2008). Les héméronymes. Ces évènements qui font date, ces dates qui deviennent évènements. In *Mots. Les langages du politique*, vol. 3 (n° 88), pp. 115-128.

Calabrese Steimberg, L. (2009). *Nommer un événement ou les marges du sens dans la désignation médiatique : l'exemple de la canicule*. In *Le sens en marge : Représentations linguistiques et observables discursifs*. Sous la direction de Evrard I., Pierrard M., Rosier L. et van Raemdonck D. L'Harmattan, pp. 15-28.

Campos R., Dias G. et Jorge A. M. (2009). Disambiguating Web Search Results by Topic and Temporal Clustering: A Proposal. In *Proceedings of International Conference on Knowledge Discovery and Information Retrieval (KDIR'09)*, pp. 292-296.

Chagnoux M. (2006). *Temporalité et aspectualité dans les textes français : modélisation sémantico-cognitive et traitement informatique*. Thèse de Doctorat. Paris IV Sorbonne.

Charolles M. (1997). L'encadrement du discours : univers, champs, domaines et espaces. In *Cahier de recherche linguistique*, vol. 148 (4), pp. 9-30.

Charolles M. (2003). De la topicalité des adverbiaux détachés en tête de phrase. In *Adverbiaux et topiques*. M. Charolles et S. Prévost éds, Louvain la Neuve, Travaux de Linguistique, 47, pp. 11-51.

Charolles M. et Vigier D. (2005). "Les adverbiaux en position préverbale : portée cadrative et organisation des discours." In *Langue Française*, 2 (148), pp. 9-30.

Chen Z., Yang H., Ma J., Lei J. et Gao H. (2011). "Time-based Query Classification and its Application for Page Rank.". In *Journal Of Computational Information Systems*, 9, pp. 3149-3156.

Chinchor N., Brown E., Ferro L. et Robinson P. (1999). *1999 Named Entity Recognition Task Definition, version 1.4*. Technical Report. MITRE and SAIC. [en ligne]
ftp://jaguar.ncsl.nist.gov/ace/phase1/ne99_taskdef_v1_4.pdf

Christin O. (2008). *Ancien Régime. Pour une approche comparatiste du vocabulaire historiographique*. In Mots. Les langages du politique, 87 (2), pp. 5-12.

Cohen J. (1960). "A coefficient of agreement for nominal scales". In *Educational and Psychological Measurement*, vol. 20 (1), pp. 37–46.

Cohen J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. In *Psychological Bulletin*, vol. 70 (4), Oct 1968, pp. 213-220.

Couto J. (2006). *Modélisation des connaissances pour une navigation textuelle assistée. La plate-forme logicielle NaviTexte*. Thèse de Doctorat. Université Paris-Sorbonne.

Couto J. et Minel J.-L. (2006a), *Navigation textuelle : représentation des textes et des connaissances*. In *Revue TAL*, vol. 47, n°2, pp. 225-254.

Couto J. et Minel J.-L. (2006b). SEXTANT, un langage de modélisation des connaissances pour la navigation textuelle. In *Actes de ISDD'06, Colloque International Discours et Document*, Schedae, Caen, p. 80-90.

Cunningham H., Maynard D., Bontcheva K., Tablan V. et Ursu V. (2002). The GATE user guide. http://gate.ac.uk/

Denis P. et Sagot B. (2010). *Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français*. In *Actes de TALN 2010*, Montréal, 19–23 juillet 2010.

Desclés J.-P. (1994). *Quelques concepts relatifs au temps et à l'aspect pour l'analyse des textes*. in Studia Kognitywne (*Etudes cognitive*), Semantyka kategorii Aspektu i czasu, n°1, Polska Akademia Nauk, Institut Slawistiki, pp. 57-88.

Desclés J.-P. (1995). Les référentiels temporels pour le temps linguistique. In *Modèles linguistiques*, 16, pp. 9-36.

Desclés, J.-P. et Guentcheva Z. (2000). *Enonciateur, locuteur, médiateur*. In Erikson P. et Monod-Becquelin A. (éds), Les rituels du dialogue, Société d'ethnologie, Nanterre, pp. 79-112.

Desclés, J.-P. (2000). Imparfait narratif et imparfait de nouvel état en français. In *Etudes linguistiques romano-slaves offertes à Stanislas Karolak*, Eds. Oficyna Wydawnicza "Edukacia" (*Colloque de Cracovie, Pologne, septembre 2000*).

Desclés, J.-P. La Grammaire Applicative et Cognitive construit-elle des représentations universelles ? In *Linx*, 48, Revue des linguistes de l'Université de Paris X Nanterre, pp. 139-160.

Diaz F. et Jones R. (2004). Using Temporal Profiles of Queries for Precision Prediction. In M. Sanderson, *et al.* (eds.), *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Sheffield, UK, pp. 18–24.

Doddington G., Mitchell A., Przybocki M., Ramshaw L., Strassel S., Weischedel R. (2004). The automatic content extraction (ACE) Program – Tasks, Data, and Evaluation. In *Proceedings of LREC 2004*, vol. 4, pp. 837–840.

Doerr M. (2003). The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata. In *AI Magazine*, vol. 24 (3), pp. 75–92.

Dowty D. R. (1986). *The effects of aspectual class on the temporal structure of discourse: semantics or pragmatics?* In Linguistics and Philosophy, vol. 9 (1), pp. 37-61.

Ehrmann M. et Hagège C. (2009). Proposition de caractérisation et de typage des expressions temporelles en contexte. In *Actes TALN'09*, pp. 24-26.

Elsas J. L. et Dumais S. T. (2010). Leveraging temporal dynamics of document content in relevance ranking. In *Proceedings of the third ACM international conference on Web search and data mining WSDM'10*, ACM, pp. 1-10.

Etcheverry Méndes M. (2010). *Reconocimiento e Interpretacion de Expresiones Temporales en Español*. Tesis Doctoral. Universidad de la República – Uruguay.

Ferro L., Mani I., Sundheim B. et Wilson G. (2001). *TIDES Temporal Annotation Guidelines*. Version 1.0.2 MITRE Technical Report, MTR 01W0000041, McLean, Virginia: The MITRE Corporation.

Ferro L., Gerber L., Mani I., Sundheim B., Wilson G. (2003). *TIDES Standard for the Annotation of Temporal Expressions*. [en ligne] http://www.mitre.org/work/tech-papers/tech-papers-04/ferro-tides/ .

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

Fortin J., Carloni O., Leclère M. et Weiser S. (2009). Extraction et exploitation de données temporelles pour un portail d'e-tourisme. In Actes de *EGC'09 - Atelier "Fouille de Données Temporelles - Analyse de Flux de Données",* Strasbourg*,* France.

Freitag D. (1998). Machine Learning for Information Extraction in Informal Domains. Thèse de Doctorat. Université Carnegie Mellon.

Gagnon M. et Lapalme G. (1996). "From conceptual time to linguistic time." In *Computational Linguistics*, vol. 22 (1), pp. 91-127.

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

Gosselin L. (2005a). *Temporalité et modalité*. De Boeck-Duculot, Bruxelles.

Gosselin L. (2005b). L'itération dans le modèle SdT. *Article paru dans le Rapport du Projet OGRE - Ordres de Grandeur et REpétition,* Rapport technique, CTAN, pp. 17-33.

Gosselin L. (2006). De la distinction entre la dimension temporelle de la modalité et la dimension modale de la temporalité. In *Cahiers de praxématique*, vol. 47, Presse Universitaire de la Méditerranée, pp. 21-52.

Grishman R. et Sundheim B. (1996). Message understanding conference-6: A brief history. In *Proceedings of the 16th conference on Computational linguistics (COLING) - Vol. 1*. Association for Computational Linguistics, pp. 466-471.

Hagège C. et Tannier X. (2008). XTM: A robust temporal text processor. In *Proceedings of the 9th international conference on Computational linguistics and intelligent text processing*, Springer-Verlag, pp. 231–240.

Han B., Gates D. et Levin L. (2006). From language to time: A temporal expression anchorer. In *Proceedings* of *TIME'06*, IEEE Computer Society, June, pp. 196–203.

Hitzeman J., Moens M. et Grover C. (1995). Algorithms for analysing the temporal structure of discourse. In *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*, Morgan Kaufmann Publishers Inc, pp. 253–260.

Hobbs J. R. (1985). *On the Coherence and Structure of Discourse*. Report No. CSLI-85-37, Stanford University, Center for the Study of Language and Information, October, 1985.

Hobbs J. R. et Pustejovsky J. (2003). Annotating and reasoning about time and events. In *Proceedings of AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, AAAI Press, Menlo Park, California, pp 74-82.

Hobbs J. R. et Pan F. (2004). An ontology of time for the semantic web. In *ACM Transactions on Asian Language*, vol. 3 (1), pp. 66-85.

Hobbs J. R. et Pan F. (2006). *Time Ontology in OWL*. W3C Working Draft. http://www.w3.org/TR/owl-time.

Hovy E. (1990). Parsimonious and Profligate Approaches to the Question of Discourse Structure Relations. In *Proceedings of the Fifth International Workshop on Natural Language Generation,* pp. 128-136.

Jackiewicz A. et Minel J.-L. (2003). L'identification des structures discursives engendrées par les cadres organisationnels. In *Actes de TALN'03,* pp.11-14.

Jackiewicz A. (2002). Repérage et délimitation des cadres organisationnels pour la segmentation automatique des textes. In *Actes de CIFT'02, Colloque International sur la Fouille de Textes*, Hammamet, Tunisie, 20-23 octobre 2002, pp. 95-105.

Jacques M.-P. et Aussenac-Gilles N. (2006). Variabilité des performances des outils de TAL et genre textuel. In *Traitement automatique des langues*, 47, pp. 11-32.

Järvelin K. et Kekäläinen J. (2002). Cumulated Gain-based Evaluation of IR Techniques. In *ACM Transactions on Information Systems*, Vol. 20, No. 4, October 2002, pp. 422-446.

de Jong F., Rode H. et Hiemstra D. (2005). Temporal language models for the disclosure of historical text. In *Humanities, computers and cultural heritage: Proceedings of the XVIth International Conference of the Association for History and Computing (AHC 2005)*, 14-17 Sept., Amsterdam, The Netherlands, pp. 161-168.

Jones R. et Diaz F. (2007). Temporal profiles of queries. In *Proceedings of ACM Transactions on Information Systems*, vol. 25 (3), 14p.

Kanhabua N. et Nørvag K. (2010). Determining time of queries for re-ranking search results. *In Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL, Glasgow, UK, 2010, pp. 261–272.

Klein W. (1994). *Time in Language*. Psychology Press, Routledge, Londres.

Lagoze C. et Hunter J. (2001). The ABC Ontology and Model. In *Proceedings of the International Conference on Dublin Core and Metadata Applications 2001,* Tokyo, 2001*, pp. 160-176*.

Lascarides A. et Asher N. (1993). Temporal interpretation, discourse relations and commonsense entailment. In *Linguistics and philosophy,*vol. 16 (5), pp. 437-493.

Laublet P., Reynaud C. et Charlet, J. (2002). Sur Quelques Aspects du Web Sémantique. In *Assises du GDR I*, Editions Cépadues, Nancy, Décembre 2002, pp. 59-78.

Lebranchu J. et Mathet Y. (2011). Vers une prise en charge approfondie des phénomènes itératifs par TimeML. In *Actes de TALN 2011*, Montpellier, France, 2011. [en ligne] http://jlebranc.perso.info.unicaen.fr/publication/taln2011-timeml-phenomenes-iteratifs.pdf

Lebranchu J. (2011). *Etude des phénomènes itératifs en langue. Inscription discursive et calcul aspectuo-temporel. Vers un traitement automatisé.* Thèse de doctorat. Université de Caen.

Le Draoulec A., Péry-Woodley M. (2005). Encadrement temporel et relations de discours. In *Langue française*, vol. 148, pp. 45-60.

Le Goffic P. (1995). *La double incomplétude de l'imparfait*. In Modèles Linguistiques, 31, Temps et Langage. XVI-1, pp. 133 – 148.

Leeman D. (2008). Prépositions du français : état des lieux. In *Langue française*, vol. 157 (1), pp. 5-19.

Leroy S. (2004). De l'identification à la catégorisation : l'antonomase du nom propre en français. Peeters Eds., Bibliothèque de l'Information Grammaticale, vol. 57, Louvain, 225p.

Lettieri C. (2008). *L'Italie et ses Années de plomb. Usages sociaux et significations politiques d'une dénomination temporelle*. In *Mots. Les langages du politique*, ENS Editions, vol. 87, pp. 43-55.

Llorens H., Saquete E., Navarro B. (2010). TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010*, July, pp. 284-291.

Mani I., Wilson G., Ferro L. et Sundheim B. (2001). Guidelines for annotating temporal information. In *Proceedings of HLT 2001, First International Conference on Human Language Technology Research*. Morristown, NJ, USA, Association for Computational Linguistics, pp. 299-302.

Mani I. et Pustejovsky J. (2004). Temporal discourse models for narrative structure. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pp. 57-64.

Mani I. et Wilson G. (2000). Robust temporal processing of news. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL2000)*. Hong-Kong, pp. 69-76.

Mann W. C. et Thompson S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. In *Text*, vol. 8 (3), pp. 243-281.

Marcu D. (2000). The Theory and Practice of Discourse Parsing and Summarization. The MIT Press, 250p.

Marcu D. et Echihabi A. (2002). An Unsupervised Approach to Recognizing Discourse Relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, pp. 368-375.

Mathet Y. (2007). Une approche cognitive de l'itération et sa modélisation objet. In *Actes de Représentation et raisonnement sur le Temps et l'Espace (RTE 2007)*, pp 53-62.

Matthews M., Tolchinsky P., Mika P., Blanco R. et Zaragoza H. (2010). Searching through time in the New York Times Categories and Subject Descriptors. In *Proceedings of HCIR 2010 - Challenge Report*, New Brunswick, pp. 41-44.

Maynard D., Cunningham H., Bontcheva K. et Dimitrov M. (2002). Adapting a Robust Multi-Genre NE System for Automatic Content Extraction. In *Proceedings of AIMSA 2002, The Tenth International Conference on Artificial Intelligence: Methodology, Systems, Applications*, Varna, Bulgaria, pp. 47-63.

Maurel D. (1988). Grammaire des dates, Etude préliminaire à leur traitement automatique. In *Linguisticae Investigationes* XII, pp. 101-128.

Maurel D. (1992). Reconnaissance automatique d'un groupe nominal prépositionnel. Exemple des adverbes de date. In *Lexique*, 11, P.U.L., pp. 147-161.

Metzler D., Rosie J. Fuchun P. et Ruiqiang Z. (2009). Improving search relevance for implicitly temporal queries. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 700-701.

Minel J.-L. (2003). Filtrage sémantique, du résumé automatique à la fouille textuelle. Hermes Sciences, vol 67, Paris, 202p.

Moirand S. (2001). Du traitement différent de l'intertexte selon les genres convoqués dans les événements scientifiques à caractère politique. In *Semen*, vol. 13, *Genres de la presse écrite et analyse de discours*, Presses universitaires de Franche-Comté, pp. 97-117.

Moirand S. (2004). L'impossible clôture des corpus médiatiques. In *TRANEL (Travaux neuchâtelois de linguistique)*, *Approche critique des discours,* vol. 40, pp. 71-92.

Moreau-Moquay C. (2012). *A la recherche des adverbiaux de temps: extraction automatique d'expressions calendaires en allemand à l'aide d'Unitex*. Mémoire de Master II Recherche. Université Paris-Sorbonne.

Muller P. et Tannier X. (2004). Annotating and measuring temporal relations in texts. In *Proceedings of the 20th international conference on Computational Linguistics - COLING '04*. Morristown, NJ, USA, Association for Computational Linguistics, pp. 50-56.

Najork M., Zaragoza H. et Taylor M. (2007). HITS on the Web: How does it Compare?. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in Information Retrieval*, pp. 471-478.

Noël L. et Azémard G. (2008). From semantic web data to inform-action: a means to an end. In *Proceedings of CHI2008 Workshop on Semantic Web User Interaction*, Florence, Italie, 5-10 avril. [en ligne] http://swui.semanticweb.org/SWUI2008CHI/Noel.pdf

Nørvag K. (2004). Supporting temporal text-containment queries in temporal document databases. In *Journal of Data & Knowledge Engineering*, vol. 49 (1), pp. 105–125.

Nunes S., Ribeiro C. et David G. (2008). Use of Temporal Expressions in Web Search. In *Proceedings of European Conference on IR Research (ECIR 2008),* 30th March-3rd April, Glasgow, Scotland. Springer Berlin/Heidelberg, vol. 4956/2008, pp. 580-584.

Padró L., Collado M., Reese S., Lloberes M. et Castellón I. (2010). FreeLing 2.1: Five Years of Open-Source Language Processing Tools. In *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)*, ELRA. La Valletta, Malta. May, 2010. [en ligne] http://nlp.lsi.upc.edu/freeling

Palmer M., Kingsbury P. et Gildea D. (2005). The proposition bank: An annotated corpus of semantic roles. In *Computational Linguistics*, vol 31 (1), pp. 71–106.

Pan F. et Hobbs J. R. (2005). Temporal aggregates in OWL-Time. In *Proceedings of the 18th International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pp. 560-565.

Parent G., Gagnon M. et Muller P. (2008). Annotation d'expressions temporelles et d'événements en français. In *Actes de TALN 2008*, Avignon.

Paumier S. (2002). Manuel d'utilisation du logiciel Unitex. IGM, Université de Marne-La-Vallée. http://www-igm.univ.mlv.fr/~unitex/manuelunitex.pdf

Prévost S. (2005). Adverbiaux temporels et structuration textuelle au 15ème siècle. In *Actes du XIIIe colloque sur le moyen Français*. Anvers, pp. 95-108.

Pustejovsky J., Belanger L., Castano J., Gaizauskas R., Hanks P., Ingria R., Katz G., Radev D., Rumshisky A., Sanfilippo A., Sauri R., Sundheim B. et Verhagen M. (2002). TERQAS Final Report. [en ligne] http://www.timeml.org/site/terqas

Pustejovsky J., Castaño J., Ingria R., Saurí R., Gaizauskas R., Setzer A. et Katz G. (2003a). TimeML: Robust Specification of Event and Temporal Expressions in Text. In *Proceedings of IWCS-5, Fifth International Workshop on Computational Semantics,* pp. 28-34.

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 *Corpus Linguistics*, pp. 647–656.

Pustejovsky J., Ingria R., Saurí R, et Castaño J., Littman J., Gaizauskas R., Setzer A., Katz G. et Mani I. (2004). The specification language TimeML. In *The Language of Time: A Reader.* Oxford University Press, pp. 545-557.

Pustejovsky J., Littman J. et Saurí R. (2006). Argument structure in TimeML. In *Dagstuhl Seminar Proceedings Annotating, Extracting and Reasoning about Time and Events*. [en ligne] http://pages.cs.brandeis.edu/~roser/pubs/lrec06_arglink.pdf

Pustejovsky J., Lee K., Bunt H. et Romary L. (2010). ISO-TimeML: An International Standard for Semantic Annotation. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. [en ligne]
http://lexitron.nectec.or.th/public/LREC-2010_Malta/pdf/55_Paper.pdf

Raimond Y. et Abdallah S. A. (2006). "The event ontology," OWL-DL ontology, 2006. [en ligne] http://purl.org/NET/c4dm/event.owl

Raimond Y., Sutton C., Sandler M. (2008). Automatic Interlinking of Music Datasets on the Semantic Web. In *Proceedings of the 1st Workshop about Linked Data on the Web (LDOW2008)*. [en ligne] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.123.9753&rep=rep1&type=pdf

Reichenbach H. (1947). *Elements of Symbolic Logic*. London, Macmillan, 444 p.

Saïto M., Yamamoto K. et Seline S. (2006). Using Phrasal Patterns to Identify Discourse Relations. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, New York, pp. 133-136.

Saquete E., Martinez-Barco P., Muñoz R., Vicedo J.L. (2004). Splitting Complex Temporal Questions for Question Answering systems. In *Proceedings of Association for Computational Linguistics (ACL),* Barcelona, Spain, July 2004, pp. 566-573.

Saurí R., Knippen R., Verhagen M., Pustejovsky J. (2005). Evita: A Robust Event Recognizer For QA Systems. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05*. October, pp. 700-707.

Saurí R., Littman J., Knippen B., Gaizauskas R., Setzer A. et Pustejovsky J. (2006). TimeML annotation guidelines. [en ligne] http://www.timeml.org/timeMLdocs/AnnGuide14.pdf

Saurí R. et Pustejovsky J. (2009). TimeML in a Nutshell. [en ligne]
http://www.timeml.org/tempeval2/tempeval2-trial/guidelines/introToTimeML-052809.pdf

de Saussure F. (1916). *Cours de Linguistique générale*. Payot, 1980, 509 p.

de Saussure L. (2003). *Temps et pertinence*. Bruxelles, De Boeck-Duculot, 321p.

Scherp A., Franz T., Saathoff C. et Staab. S. (2009). F-A Model of Events based on the Foundational Ontology DOLCE+ Ultra Light. In *Proceedings ot the 5th International Conference on Knowledge Capture (K-CAP'09)*, Redondo Beach, California, USA, pp. 137-144.

Schilder F. et Habel C. (2001). From temporal expressions to temporal information: Semantic tagging of news messages. In *Proceedings of the ACL-2001 Workshop on Temporal and Spatial Information Processing, Toulouse, France*, pp. 65-72.

Schwer R. S. (2007). Traitement de la temporalité des discours : une analysis situs. In Information temporelle, procédure et ordre discursif. In *Cahier Chronos*, vol. 18, pp. 7-22.

Setzer A. et Gaizauskas R. (2000). Annotating events and temporal information in newswire texts. In *Proceedings of the Second International Conference On Language Resources and Evaluation*, pp. 1287-1294.

Setzer A. (2001). Temporal Information in Newswire Articles: an Annotation Scheme and Corpus Study. Ph.D. thesis. University of Sheffield, UK.

Setzer A. et Gaizauskas R. (2002). On the importance of annotating event-event temporal relations in text. In *LREC 2002 Workshop on Temporal Annotation,* pp. 52-60.

Shaw R., Troncy R. et Hardman L. (2009). Lode: Linking open descriptions of events. In *The Semantic Web*, pp. 153–167.

Teissèdre C., Battistelli D. et Minel J.-L. (2011). Recherche d'information et temps linguistique : une heuristique pour calculer la pertinence des expressions calendaires. In *Actes de TALN 2011*, 27 juin-1er juillet 2011, Montpellier, pp. 161-172.

Vandenbussche P.-Y. et Teissèdre C. (2011). Events Retrieval Using Enhanced Semantic Web Knowledge. In *Proceedings of ISWC Workshop DeRiVE2011*, October 23-27 2011, Bonn, Germany, pp. 112-116.

UzZaman N. et Allen J. F. (2010). TRIPS and TRIOS System for TempEval-2: Extracting Temporal Information from Text. In *Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010*, pp. 276-283.

Van De Velde D. (2000). Existe-t-il des noms propres de temps ? In *Lexique*, n° 15, p. 35-45.

Van Raemdonck D. (2001). *Est-il pertinent de parler d'une classe d'adverbes de temps ?* In *Circulo de lingüística aplicada a la comunicación, Número 7*.

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

Verhagen M., Gaizauskas R., Schilder F., Hepple M., Katz G. et Pustejovsky J. (2007). “Semeval-2007 task 15: Tempeval temporal relation identification.” In *Proceedings of the 4th International Workshop on Semantic Evaluations* (June), pp. 75-80.

Verhagen M., Saurí R., Caselli T., Pustejovsky J. (2010). SemEval-2010 Task 13: TempEval-2. *Computational Linguistics*. (July), pp. 57-62.

W3C, (2006), *Time Ontology in OWL*, W3C Working Draft, Septembre 2006, http://www.w3.org/TR/owl-time/

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

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.

Wilmet M. (2003). *Grammaire critique du français*. Editions Duculot.

Wilson G., Mani I, Sundheim B. et Ferro L. (2001). A multilingual approach to annotating and extracting temporal information. In *Proceeding of the workshop on Temporal and spatial information processing TASIP ’01*, pp. 81-87.

Annexe 1 : exemples de représentation des adverbiaux sous la forme d'une succession d'opérations sur un repère temporel noyau

On présente dans cette annexe différents exemples de représentation des adverbiaux de localisation temporelle sous la forme d'une succession d'opérations sémantiques présentés dans des diagrammes d'objet. Les exemples d'analyse sont présentés pour les adverbiaux suivants : *hier*, *ce jour-là* et *durant les deux dernières semaines de la campagne électorale*.

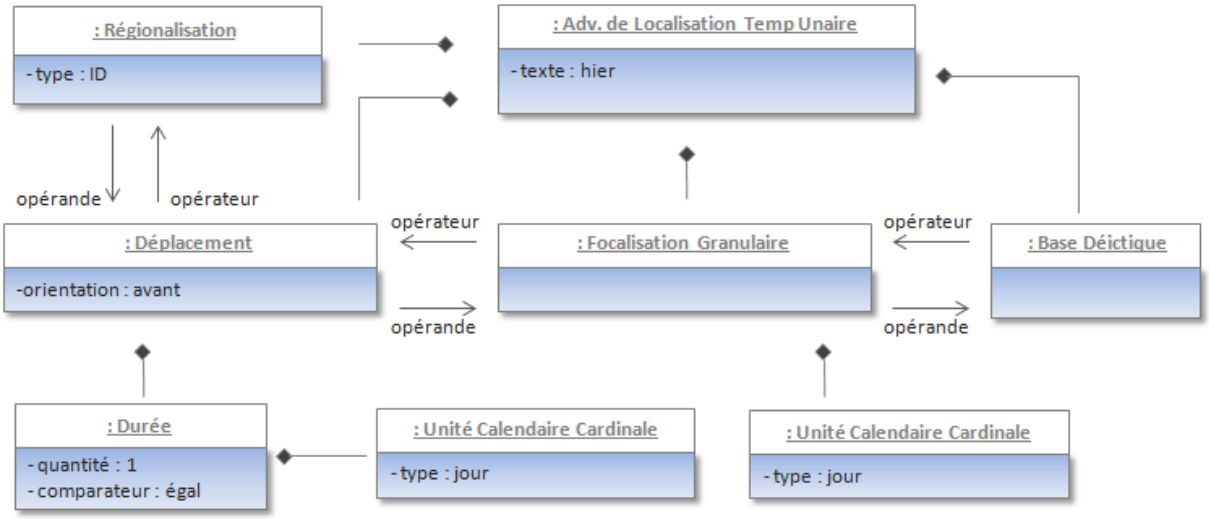


Fig. 76 : diagramme d'objet associé à l'adverbial hier

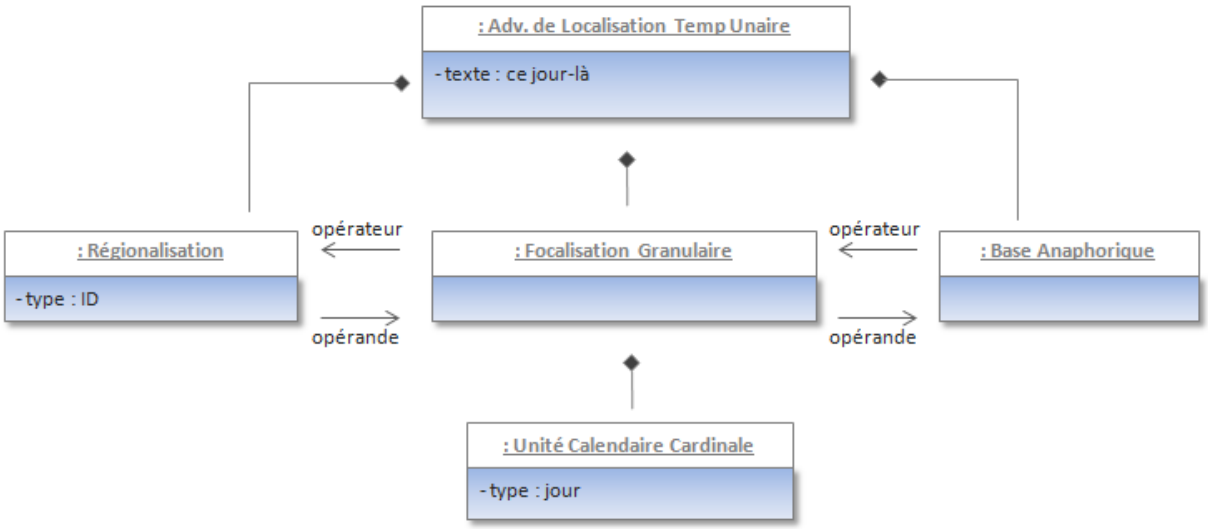


Fig. 77 : diagramme d'objet associé à l'adverbial ce jour-là

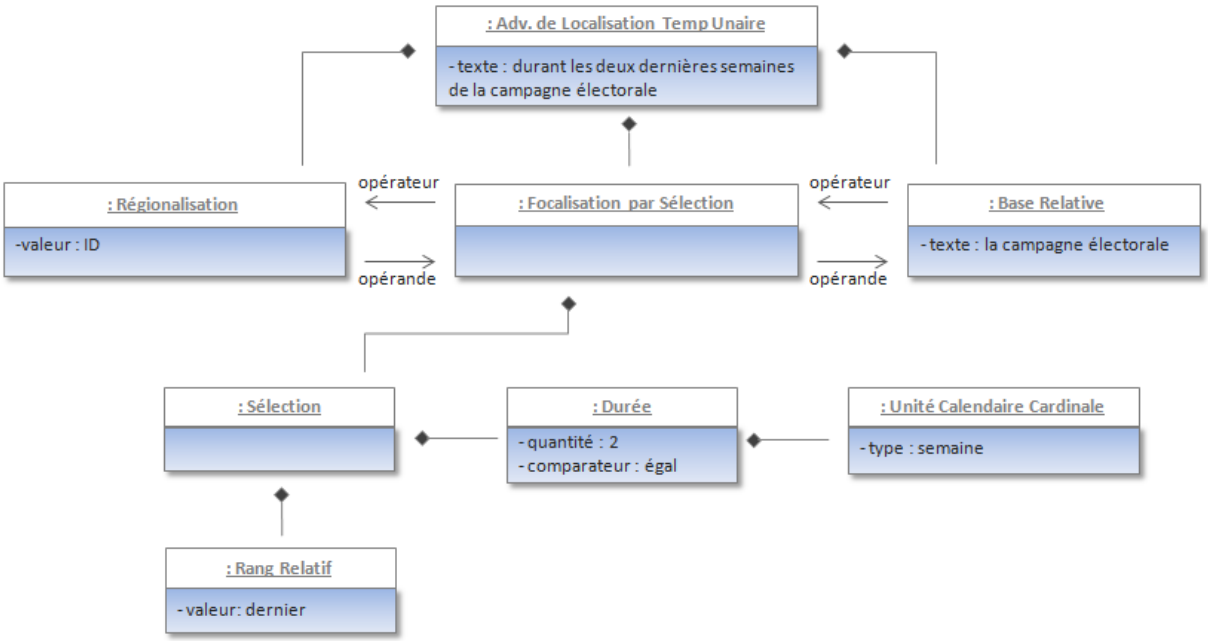


Fig. 78 : diagramme d'objet associé à l'adverbial durant les deux dernières semaines de la campagne électorale

Annexe 2 : le schema d'annotation ChronolocationML

Cette annexe reproduit la DTD du langage *ChronolocationML* qui définit un jeu de balises XML pour annoter des expressions de localisation temporelle présentes dans les textes.

```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT ChronolocationML ( TEXT  | INTERVAL_LINK | COMPOSITION_LINK | CHRONOLOC_LINK )* >

<!ELEMENT TEXT ( #PCDATA | CHRONEX  | INTERVAL | EVENT )* >

<!ELEMENT EVENT ( #PCDATA ) >
<!ATTLIST EVENT eiid ID #REQUIRED >

<!ELEMENT CHRONEX ( #PCDATA | ZONING | SHIFTING | ZOOMING | BASE  )* >
<!ATTLIST CHRONEX cid ID #REQUIRED >

<!ELEMENT ZONING ( #PCDATA)* >
<!ATTLIST ZONING zonid ID #REQUIRED >
<!ATTLIST ZONING type ( before | after | until | since | around | ID ) #REQUIRED >
<!ATTLIST ZONING operand CDATA #IMPLIED >

<!ELEMENT ZOOMING ( #PCDATA | CALENDAR_UNIT)* >
<!ATTLIST ZOOMING zmid ID #REQUIRED >
<!ATTLIST ZOOMING subdivision ( quarter | third | half | beginning | mid | end ) #IMPLIED >
<!ATTLIST ZOOMING rank CDATA #IMPLIED >
<!ATTLIST ZOOMING operand CDATA #IMPLIED >

<!ELEMENT SHIFTING ( #PCDATA | CALENDAR_UNIT )* >
<!ATTLIST SHIFTING sid ID #REQUIRED >
<!ATTLIST SHIFTING orientation ( before | after ) #IMPLIED >
<!ATTLIST SHIFTING quantity CDATA #IMPLIED >
<!ATTLIST SHIFTING comparisonOperator CDATA #IMPLIED >
<!ATTLIST SHIFTING operand CDATA #IMPLIED >

<!ELEMENT BASE ( #PCDATA | EVENT | CALENDAR_UNIT )* >
<!ATTLIST BASE bid ID #REQUIRED >
<!ATTLIST BASE isIterative CDATA #IMPLIED >
<!ATTLIST BASE type ( calendar | deictic | relative | anaphoric | unknown ) #REQUIRED >

<!ELEMENT CALENDAR_UNIT ( #PCDATA ) >
<!ATTLIST CALENDAR_UNIT type CDATA #REQUIRED >
<!ATTLIST CALENDAR_UNIT value CDATA #IMPLIED >
```

```
<!ELEMENT INTERVAL (#PCDATA | CHRONEX )* >
<!ATTLIST INTERVAL iid ID #REQUIRED >
<!ATTLIST INTERVAL type ( between | fromTo ) #REQUIRED >
<!ATTLIST INTERVAL start CDATA #REQUIRED >
<!ATTLIST INTERVAL end CDATA #REQUIRED >


<!ELEMENT COMPOSITION_SIGNAL (#PCDATA) >
<!ATTLIST COMPOSITION_SIGNAL type ( concatenation | exception ) #REQUIRED >


<!ELEMENT COMPOSITION_LINK EMPTY >
<!ATTLIST COMPOSITION_LINK clid ID #REQUIRED >
<!ATTLIST COMPOSITION_LINK type ( concatenation | specification | exception ) #REQUIRED >
<!ATTLIST COMPOSITION_LINK operator CDATA #IMPLIED >
<!ATTLIST COMPOSITION_LINK operand CDATA #IMPLIED >
<!ATTLIST COMPOSITION_LINK list CDATA #IMPLIED >


<!ELEMENT CHRONOLOC_LINK EMPTY >
<!ATTLIST CHRONOLOC_LINK chronlid ID #REQUIRED >
<!ATTLIST CHRONOLOC_LINK operator CDATA #REQUIRED >
<!ATTLIST CHRONOLOC_LINK operand CDATA #REQUIRED >
```

Annexe 3 : une fiche de synthèse des évaluations manuelles de la pertinence d'un ensemble d'adverbiaux-cibles par rapport à des adverbiaux-requêtes

Cette annexe présente une fiche de synthèse regroupant les scores attribués par deux évaluateurs chargés de mesurer la pertinence d'un corpus d'adverbiaux de localisation temporelle cibles par rapport à un ensemble d'adverbiaux requêtes.

Ces évaluations manuelles ont permis de mesurer un accord inter-annotateur et de comparer l'ordre qui découle de ces mesures avec celui qui découle de l'application de l'heuristique visant à mesurer la pertinence temporelle d'un adverbial calendaire candidat par rapport à un adverbial requête (cf. chapitre 8, section 8.3.2). Ces fiches ont ainsi servi à évaluer l'heuristique et de montrer qu'elle produit des résultats proches de ceux des évaluateurs : dit autrement, elle ordonne les adverbiaux-candidats d'une façon sensiblement similaire à ce que fait un humain. Cette méthode de calcul de la pertinence d'un adverbial calendaire est celle implémentée dans le moteur de recherche expérimental (cf. chapitre 7).

	requête 1			**requête 2**			**requête 3**		
	à partir de 1715			en février 1871			au XVIIIe siècle		
	évaluateur 1	**évaluateur 2**	**somme**	**évaluateur 1**	**éval. 2**	**somme**	**évaluateur 1**	**éval. 2**	**somme**
au début du XXe siècle	2	3	5	0	0	0	0	0	0
en février 1871	2	3	5	4	4	8	0	0	0
À partir de 1747	3	3	6	0	2	2	3	2	5
vers 1850	2	3	5	0	1	1	0	0	0
le 26 janvier 1871	2	3	5	1	1	2	0	0	0
avant 1851	2	2	4	0	0	0	0	0	0
en 1716	3	3	6	0	0	0	3	3	6
à l'automne 1924	2	3	5	0	0	0	0	0	0
le 21 janvier 1916	2	3	5	0	0	0	0	0	0
de 1849 à 1866	2	3	5	0	0	0	0	0	0
En octobre 1937	2	3	5	0	0	0	0	0	0
Du 26 mars au 22 mai 1871	2	3	5	1	0	1	0	0	0
le 6 juillet 1918	2	3	5	0	0	0	0	0	0
entre 1916 et 1923	2	3	5	0	0	0	0	0	0
entre 1931 et 1935	2	3	5	0	0	0	0	0	0
en mai 1938	2	3	5	0	0	0	0	0	0
À partir du XVIIIe siècle	3	2	5	0	3	3	3	3	6
À partir de 1925	2	3	5	0	0	0	0	0	0
le 10 octobre 1714	0	0	0	0	0	0	3	3	6
durant l'hiver 1756-1757	3	3	6	0	0	0	3	3	6
en 1799	3	3	6	0	0	0	3	3	6
Fin mars 1801	2	3	5	0	0	0	0	0	0
à la fin du XVIIIe siècle	3	1	4	0	0	0	3	3	6
entre 1815 et 1853	2	3	5	0	0	0	0	0	0
depuis 1906	2	3	5	0	0	0	0	0	0
en août 1795	3	3	6	0	0	0	3	3	6
mai 1871	2	3	5	1	0	1	0	0	0
durant les années 1840	2	3	5	0	0	0	0	0	0
à partir de 1838	2	3	5	0	2	2	0	0	0
le 13 janvier 1940	2	3	5	0	0	0	0	0	0

	requête 4			**requête 5**		
	jusque dans les années 30			après 1848		
	évaluateur 1	**évaluateur 2**	**somme**	**évaluateur 1**	**évaluateur 2**	**somme**
au début du XXe siècle	3	2	5	2	3	5
en février 1871	2	3	5	3	3	6
À partir de 1747	2	2	4	0	1	1
vers 1850	2	3	5	3	3	6
le 26 janvier 1871	2	3	5	3	3	6
avant 1851	2	3	5	3	2	5
en 1716	2	3	5	0	0	0
à l'automne 1924	3	3	6	2	3	5
le 21 janvier 1916	3	3	6	2	3	5
de 1849 à 1866	2	3	5	3	3	6
En octobre 1937	3	3	6	2	3	5
Du 26 mars au 22 mai 1871	2	3	5	3	3	6
le 6 juillet 1918	3	3	6	2	3	5
entre 1916 et 1923	3	3	6	2	3	5
entre 1931 et 1935	3	3	6	2	3	5
en mai 1938	3	3	6	2	3	5
À partir du XVIIIe siècle	2	2	4		2	2
À partir de 1925	3	2	5	2	2	4
le 10 octobre 1714	2	3	5	0	0	0
durant l'hiver 1756-1757	2	3	5	0	0	0
en 1799	2	3	5	0	0	0
Fin mars 1801	2	3	5	0	0	0
à la fin du XVIIIe siècle	2	3	5	0	0	0
entre 1815 et 1853	2	3	5	2	1	3
depuis 1906	3	2	5	2	3	5
en août 1795	2	3	5	0	0	0
mai 1871	2	3	5	2	3	5
durant les années 1840	2	3	5	2	2	4
à partir de 1838	2	2	4		2	2
le 13 janvier 1940	0	0	0	2	3	5

Annexe 4 : le corpus de requêtes pour l'évaluation du système CaSE

On produit ci-dessous la liste des requêtes soumises au système pour évaluer le moteur de recherche expérimental CaSE. On rappelle que l'ensemble des requêtes testées associent des critères thématiques (un ou plusieurs mot(s)-clé(s)) et des critères calendaires. Les résultats de l'évaluation sont présentés dans le chapitre 8 (cf. section 8.3.3).

- Nicolas Sarkozy dans les années 90
- linguistique dans les années 60
- université vers le XIIIe siècle
- agriculture biologique fin du XXe siècle
- ségrégation depuis les années 50
- prohibition au début des années 30
- peine de mort depuis les années 70
- langue française de 1520 à 1600
- monuments historiques jusqu'en 1851
- croisade début du XIIIe siècle
- industrie nucléaire depuis mars 2011
- de Gaulle après mai 1968
- lecture numérique depuis les années 60
- grippe aviaire en 2004
- décolonisation de 1945 à 1970
- vote des femmes depuis 1900
- poètes français du XVIe siècle
- crises économiques au XIXe siècle
- journaux à la fin du XVIIIe siècle
- franc-maçonnerie depuis 2000
- choc pétrolier de 1980 à 2000
- hygiène jusqu'au XVIIIe siècle
- peste au XVe siècle
- syndicalisme vers 1880
- chômage en France dans les années 70
- christianisation de la fin du Ve siècle au VIIIe siècle
- écologie de 1980 à 1990
- énergies renouvelables depuis 2010

Résumé

Cette thèse aborde la question de l'accès aux textes numériques, en particulier de l'accès à leur « contenu informationnel » ancré dans le temps. La visée de ces travaux est double : il s'agit de concilier une approche linguistique et une approche applicative, afin de contribuer à l'élaboration de nouveaux outils pour la fouille de textes, la recherche d'information et la gestion des connaissances - nouveaux outils en mesure de tirer parti de la sémantique des informations temporelles exprimées dans les textes. Il s'agit ainsi à la fois de mettre en œuvre des systèmes d'interaction avec les utilisateurs et de parvenir à modéliser la sémantique des unités textuelles qui contribuent de façon saillante à l'ancrage dans le temps des situations décrites dans les textes : les *adverbiaux de localisation temporelle*. L'étude linguistique vise ici à montrer que l'ingénierie des langues peut gagner à envisager d'une façon renouvelée les phénomènes de localisation temporelle, en cherchant à mettre au jour les différentes opérations à l'œuvre dans les adverbiaux de localisation temporelle. En faisant émerger les valeurs sémantiques de ces opérations, nous montrons qu'il devient possible d'élaborer de nouveaux systèmes de recherche d'information, susceptibles de traiter des requêtes associant un critère calendaire avec un ensemble de mots-clés, telles que « *les universités au début du XIIe siècle* », par exemple. S'appuyant sur les outils développés en ce sens, on montre qu'il devient également possible d'interagir avec des données structurées décrivant des informations temporelles, à la fois pour les interroger et pour les enrichir de façon semi-automatique.

Mots-clés : Extraction d'informations temporelles ; Annotation sémantique des adverbiaux de localisation temporelle ; Recherche d'information ; Acquisition de connaissances

Abstract

This Ph. D. thesis addresses the issue of accessing the content of digital texts, in so as it is linked to the expression of temporal location. The aim of this work is twofold: it consists in conciliating a linguistic approach and an applied approach, to participate in the development of new tools for text mining, information retrieval and knowledge management - new tools that would able to take advantage of the semantics of temporal information expressed in texts. It is thus both about implementing interaction systems and modeling the semantics of one the most salient textual units that contributes to anchor in time the situations described in texts: *temporal locating adverbials*. The linguistic analysis aims to show that language engineering can benefit from considering in a new way the phenomena of temporal location, by uncovering the different operations at work in temporal locating adverbials. Pinpointing the semantic values of these operations, we show that it becomes possible to develop new Information Retrieval systems, able of processing queries involving both a calendar criterion and a set of keywords, such as "*universities in the early twelfth century*", for instance. We show that it also becomes possible to interact with structured data describing temporal information, both for search process and for data semi-automatic enrichment.

Keywords: Temporal Information Extraction; Semantic Annotation of Temporal Locating Adverbials; Information Retrieval; Knowledge Acquisition.