Analyse sémantique des adverbiaux de localisation temporelle : application à la recherche d'information

Charles Teissèdre

STIH - EA 4509 – Université Paris-Sorbonne
28 rue Serpente, 75006 Paris, France
charles.teissedre@gmail.com

RÉSUMÉ. *Ces travaux abordent la question de l'accès aux textes numériques, et plus particulièrement de l'accès à leur « contenu informationnel » considéré sous l'angle de son ancrage temporel. Leur visée est double : il s'agit de concilier une démarche d'analyse linguistique et une approche applicative, dont un des objectifs est de participer à l'élaboration de nouveaux outils pour la fouille de textes et la recherche d'information. 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 formellement la sémantique d'une des unités textuelles qui contribue de façon saillante à ancrer dans le temps les situations décrites dans les textes: les* adverbiaux de localisation temporelle. *L'étude linguistique vise à montrer que l'ingénierie des langues peut gagner à envisager d'une façon renouvelée l'analyse des « expressions temporelles » dans les textes, en cherchant à formaliser les différentes opérations sémantiques à l'œuvre dans les adverbiaux de localisation temporelle. Nous montrons que cette démarche permet d'envisager l'élaboration de nouveaux systèmes de recherche d'information en mesure de traiter des requêtes associant un critère calendaire avec un ensemble de mots-clés, comme « les universités au début du XII^e^ siècle », par exemple.*

ABSTRACT. *This work addresses the issue of accessing the content of digital texts, in so as it is anchored in time. 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 and information retrieval. 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 "temporal expressions" in texts, by seeking to uncover the different semantic 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.*

MOTS-CLÉS : *analyse sémantique, adverbiaux de localisation temporelle, recherche d'information.*

KEYWORDS: *Semantic Analysis; Temporal Locating Adverbials; Information Retrieval.*

1. Introduction

L'accumulation documentaire conduit à s'interroger sur les différentes stratégies à mettre en place pour accéder à des documents pertinents, et, au sein de ces documents, aux informations pertinentes pour une recherche donnée. Les systèmes de recherche d'information élargissent ainsi progressivement les services d'interrogation et de filtrage de vastes collections documentaires qu'ils proposent et 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, lorsqu'une recherche semble porter sur un sujet d'actualité. Il y a toutefois peu de réalisations opérationnelles permettant de prendre en charge les informations temporelles exprimées dans le corps des documents (Alonso *et al.*, 2007).

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 XV^e^ siècle, coupe du monde 1998*, etc.). 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). C'est sans doute ainsi pour ce type d'usages spécialisés que l'intérêt des systèmes de recherche d'information permettant d'exprimer des critères temporels s'apprécie le mieux. Mais 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 les systèmes puissent inférer que les zones temporelles qu'ils dénotent sont incluses dans celle dénotée par le critère temporel exprimé dans la requête. De même, les moteurs ne sont pas en mesure de rapprocher un adverbial comme *en 1965* avec une requête qui porterait sur *les années 60*. Les utilisateurs, par expérience, n'ignorent pas que les moteurs de recherche n'interprètent pas la sémantique de 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 ces requêtes se répandrait davantage.

Le système de recherche d'information que nous présentons dans cet article, dédié à la recherche d'information temporelle, s'appuie sur une analyse sémantique des adverbiaux de localisation temporelle présents dans les textes. L'hypothèse qui sous-tend ces travaux est que cette analyse invite à réexaminer en partie la description des « expressions temporelles » qui s'est imposée en ingénierie des langues (section 2). Nous montrerons qu'il peut être utile de faire coexister deux représentations, l'une qui vise à modéliser la manière dont s'exprime en langue la localisation temporelle (section 3), l'autre associée aux standards normés de représentation des dates et de faire dépendre des besoins applicatifs la façon de transposer la première représentation vers la seconde (section 4)[1].

Un exemple simple devrait permettre d'illustrer l'intérêt de la démarche. L'étude par un archéologue ou un historien d'un objet ou d'un événement ancien ne permet pas toujours de le dater de façon très précise. Ainsi, les études sur le tableau *Judith décapitant Holopherne* du Caravage semblent estimer qu'il a été peint *à la toute fin du XVI^e^ siècle*, probablement *vers les années 1598-1599* (Schütze S., 2009). Ce type d'incertitude ne peut pas être exprimé avec les standards normés des dates (comme la norme ISO 8601), leur objectif étant précisément de chasser toute indétermination dans la représentation d'une référence datative. 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 ou de manipuler des repères temporels de granularité variable. Or, pour un archéologue, un historien ou un journaliste, qui s'intéresse à des périodes de temps plus ou moins étendues, aux contours parfois imprécis, pouvoir exprimer des repères temporels en conservant un certain degré d'indétermination est souvent une nécessité. Il peut donc être intéressant de faire coexister à côté des normes de représentation des dates un modèle formel découlant de l'analyse linguistique des expressions datatives présentes dans les textes.

Les travaux présentés ici abordent ainsi la question de l'accès aux textes numériques, et plus particulièrement de l'accès à leur « contenu informationnel » considéré sous l'angle de son ancrage temporel. Leur visée est double : il s'agit de concilier une démarche d'analyse linguistique et une approche applicative, dont un des objectifs est de participer à l'élaboration de nouveaux outils pour la fouille de textes et la recherche d'information. 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 formellement la sémantique d'une des unités textuelles qui contribue de façon saillante à ancrer dans le temps les situations décrites dans les textes : *les adverbiaux de localisation temporelle*. Ces adverbiaux ne recouvrent pas simplement les expressions datatives dans les textes (*15 juillet ; 2011*), au sens où ils intègrent des

1. Ces travaux ont été réalisés dans le cadre d'un doctorat en contrat CIFRE codirigé par Delphine Battistelli et Jean-Luc Minel. Ils s'articulent avec les travaux menés dans le cadre d'un atelier composé de Delphine Battistelli, Marcel Cori, Jean-Luc Minel et Charles Teissèdre. La formalisation du processus de transduction et l'heuristique permettant de mesurer la pertinence des adverbiaux calendaires pour la recherche d'information présentées dans cette section sont le fruit de cette collaboration.

prépositions ou locutions prépositionnelles (*avant fin 2011 ; à partir du 15 juillet*), mais aussi parce qu'ils réfèrent à des repères temporels non calendaires, ancrés sur l'énonciation (*depuis l'an dernier*) ou relativement à un procès, éventuellement nominalisé (*depuis les élections ; jusqu'à ce que le puits de forage de BP soit déclaré « mort »*).

Dans le sillage des travaux de Battistelli (2009), la représentation formelle que l'on propose consiste à décrire les adverbiaux de localisation temporelle sous la forme d'une succession d'opérations sémantiques portant sur une référence temporelle noyau. Cette représentation, que nous avons généralisée pour qu'elle recouvre non plus seulement les expressions dites calendaires, mais l'ensemble des adverbiaux de localisation temporelle, permet de conserver, si besoin est, l'indétermination sur les bornes de la zone temporelle à laquelle ils peuvent référer. En articulant cette représentation avec une représentation dite « référentielle » qui associe aux adverbiaux des valeurs calendaires, nous montrons qu'il devient possible d'élaborer de nouveaux systèmes de recherche d'information, susceptibles de traiter des requêtes contenant des critères calendaires et thématiques, telles que *la prohibition dans les années 30* (section 5).

2. Le problème de la représentation des « expressions temporelles » en ingénierie des langues

La caractérisation de l'« information temporelle » dans les textes 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 niveaux de représentation adopter 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 les projets d'annotation automatique des textes, 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 à la référence localisée dans le temps des valeurs calendaires, pose toutefois des difficultés.

Dans les applications d'annotation automatique des textes, les unités textuelles référant explicitement à un calendrier (le calendrier grégorien, par exemple) constituent un champ d'investigation exploré depuis de nombreuses années (Maurel, 1988). Le repérage des entités nommées dans les textes (Chinchor *et al.*, 1999), au nombre desquelles les dates sont très fréquemment 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 sixième 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 (date et heure), s'ajoutant aux autres types d'entités nommées. Lors de la septième conférence MUC, en 1998, la sous-tâche

dédiée aux expressions temporelles étend ses exigences aux expressions temporelles dites *relatives*, par opposition aux expressions dites *absolues*. Progressivement, cette classification s'affirme comme une distinction opératoire entre les expressions qui peuvent être placées sans ambiguïté sur un calendrier, les expressions « absolues » telles que *10 octobre 1999*, et celles, « relatives », qui requièrent d'autres informations pour pouvoir être situées sur un calendrier – en particulier, les expressions déictiques dont le repère temporel noyau est ancré sur le processus énonciatif (Setzer et Gaizauskas, 2000). Cette distinction étant posée, une tâche dite de *normalisation* se fait jour : elle consiste à résoudre la référence calendaire des expressions temporelles relatives et à leur assigner une date dans un format standard, généralement le format ISO 8601.

Parallèlement, lors de ces conférences, une nouvelle tâche apparaît qui consiste 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 par exemple). En plus des expressions datatives, aux informations temporelles visées dans les textes s'ajoutent ainsi les « événements ». De nombreux travaux s'intéressant aux corpus de presse (Filatova et Hovy, 2001 ; Schilder et Habel, 2001 ; Setzer, 2001 ; Setzer et Gaizauskas, 2002) décrivent alors des systèmes d'annotation permettant d'assigner des dates aux événements repérés. Pour la première fois en 2004, une tâche dévolue uniquement à la problématique du repérage et de la normalisation des expressions temporelles est proposée pour évaluer différents systèmes d'annotation (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 de ces expressions (Ferro *et al.*, 2003 ; Saquete *et al.*, 2004). C'est dans ce contexte que naît TimeML (Pustejovsky *et al.*, 2002 ; Saurí *et al.*, 2006 ; Saurí et Pustejovsky, 2009 ; Pustejovsky *et al.*, 2010), un langage d'annotation des informations temporelles dans les textes. Ce langage d'annotation est conçu à l'origine pour améliorer les systèmes de questions-réponses, afin qu'ils puissent prendre en charge des questions faisant intervenir des critères temporels, dont voici plusieurs exemples tirés de (Pustejovsky *et al.*, 2003a) :

- *Is Gates currently CEO of Microsoft?* (Gates est-il aujourd'hui PDG de Microsoft ?)
- *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 ?)
- *Did the Enron merger with Dynegy take place?* (La fusion d'Enron avec Dynegy a-t-elle eu lieu ?)

L'objectif qui anime cette démarche est ainsi 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.*, 2003a)). En l'occurrence donc, le contenu temporel des textes, qui trouve à s'exprimer essentiellement dans des « événements » et des « dates ».

Le langage d'annotation TimeML s'est aujourd'hui très largement imposé dans la communauté scientifique pour l'annotation des « expressions temporelles » (*TIMEX*) et des événements (*EVENT*). Des campagnes d'évaluation des systèmes d'annotation de ces expressions s'appuient sur cette spécification (TempEval-1 en 2007 (Verhagen *et al.*, 2007) et TempEval-2 en 2010 (Verhagen *et al.*, 2010)). Parallèlement, cet effort d'élaboration d'un langage d'annotation des expressions temporelles s'est accompagné, au fil du temps, de la diffusion de vastes corpus annotés tels que ACE et TimeBank (Pustejovsky *et al.*, 2003b) et, bien sûr aussi, d'un nombre croissant de systèmes automatiques pour annoter les expressions temporelles présentes dans les textes (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 d'applications opérationnelles qui sont en mesure de tirer parti de l'annotation des expressions temporelles pour la recherche d'information (Alonso *et al.*, 2007). Les systèmes actuels n'exploitent encore que très peu et avec difficulté les informations temporelles exprimées dans les textes. Fait significatif, l'unique réalisation grand public d'envergure, le service *view:timeline* de Google[2], a été abandonné en 2011.

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 tâches principales 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 de relations temporelles (tels que *avant*, *après*, *durant*), semble aisé à résoudre à l'aide d'automates à états finis (Maurel, 1990 ; Wilson *et al.*, 2001). La deuxième tâche, le repérage d'événements, pose, elle, des problèmes théoriques, la caractérisation de ce qu'est un événement dans un texte n'allant pas de soi. La troisième tâche pose des problèmes de deux ordres. D'abord, de nombreuses descriptions d'événements ne sont pas associées à des dates dans les textes. Par ailleurs, dater précisément des événements décrits dans les textes pose des difficultés comme le souligne Calabrese Steimberg (2006), parce que « *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*[3] *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) ;* Auschwitz *ne désigne pas un événement à proprement parler mais toute une famille d'événements, et les connotations*

2. Ce service associait aux résultats d'une recherche une frise chronologique présentant la fréquence de coapparition de dates et de mots-clés dans les textes : on pouvait ainsi visualiser sur ce graphique qu'un terme tel que *révolution* était associé fréquemment dans les textes aux années 1789, 1830, 1848 et 1917.

3. 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, p. 26).

dépassent largement la liste de faits que le désignant résume ; les guerres, par exemple, n'ont pas toujours des limites temporelles précises ».

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 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 Luc ; 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 du point de vue de l'analyse linguistique (il s'agit de considérer des adverbiaux plutôt que des dates et des marqueurs qui ne renvoient pas à des 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) demande à être reconsidéré : en effet, des adverbiaux temporels qu'il est possible de décrire d'une façon similaire d'un point de vue syntaxique et sémantique peuvent contenir des « dates » ou des « événements » (*depuis 2008 ; depuis le début de la campagne électorale*).

Le fait que les systèmes d'annotation ne prennent pas en compte des unités syntaxiques cohérentes explique sans doute en partie la difficulté qu'ont en aval les systèmes de recherche d'information pour exploiter ces annotations. Ceux-ci devraient en effet pouvoir différencier des expressions telles que *dans les années 80, depuis les années 80, au début des années 80, deux ans avant le début des années 80* au lieu de n'exploiter que la référence temporelle noyau *années 80*. Il nous semble ainsi qu'une analyse linguistiquement motivée des adverbiaux de localisation temporelle articulée au problème de leur transposition sous la forme de valeurs calendaires doit permettre de répondre à plusieurs difficultés rencontrées par les applications qui recourent à ces systèmes d'annotation (Arikan *et al.*, 2009 ; Berberich *et al.*, 2010 ; Matthews *et al.*, 2010).

3. Les adverbiaux de localisation temporelle

S'ils ne sont pas les seuls éléments contribuant à ancrer temporellement les situations décrites dans les textes (Gosselin, 1996 et 2005), les adverbiaux temporels, et plus particulièrement ceux que l'on nomme ici *adverbiaux de localisation temporelle*, 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), par rapport au référentiel calendaire (ex. 2 et 3) que par rapport à un procès, éventuellement nominalisé (ex. 4 et 5).

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 : 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. 3 : 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. 4 : 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. 5 : 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 (ex. 6 et 7), les adverbiaux de localisation temporelle contribuent également à la cohésion discursive.

Ex. 6 : 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. 7 : **Ce jour-là**, le Printemps proposera également une création autour des échanges musicaux entre la France et l'Afrique, baptisée "Yeke, Yeke".

Polysémiques et parfois difficiles à catégoriser sur un plan sémantique (Van Raemdonck, 2007), les adverbiaux temporels au sens large 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 des difficultés d'analyse parce qu'elles concentrent dans leur dénomination même un aspect syntaxique et un aspect sémantique. 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 (Blumenthal, 1990). S'ils présentent une grande diversité du point de vue morphosyntaxique, 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.

Dans sa proposition de description 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 (ex. 1 et 2), des syntagmes prépositionnels (ex. 3 et 4), des subordonnées (ex. 5 et 6), des syntagmes nominaux (ex. 7 et 8) ou même des noms communs sans déterminant en position d'adverbial (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 […] 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.

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.

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, 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 apparente des adverbiaux de localisation temporelle (et géographique), il est possible de dégager une structure morphosyntaxique régulière, dont l'analyse est formalisée sous la forme d'un arbre syntaxique qui peut s'enrichir de façon récursive (*deux jours avant la réunion ; deux jours avant la réunion d'après les vacances*). Les adverbiaux de localisation prennent ainsi généralement la forme d'un syntagme prépositionnel (d'autres formes sont possibles comme les syntagmes nominaux tels que *le 23 mars*

2002) qui peut connaître deux types de complémentation : un complément classique de la préposition à droite (*avant les vacances ; depuis ce jour-là*) et un complément en position dite de « spécifieur » (*peu avant ; quelques jours après*). Ces différents 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*). 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*).

Plusieurs travaux se sont attachés à décrire formellement la sémantique des adverbiaux de localisation temporelle en les représentant sous la forme d'une combinaison d'opérateurs sémantiques en petit nombre permettant d'exprimer des relations entre des repères temporels (Gagnon et Lapalme, 1996 ; Battistelli, 2009). Nos travaux s'inscrivent dans cette veine, en étendant la formalisation proposée par Battistelli (2009) qui représente les expressions dites calendaires comme une succession d'opérations s'appliquant sur une référence calendaire nommée *base*, qui correspond à la zone initialement pointée sur le calendrier (l'année 1976 dans l'expression *avant 1976* par exemple). Cette modélisation s'attache à mettre en lumière la façon dont ces expressions déterminent un ancrage temporel à partir d'un repère calendaire noyau qui est l'opérande d'une succession d'opérations sémantiques (opération de régionalisation, de déplacement et de focalisation). Étendant cette proposition (Teissèdre, 2012), on peut distinguer, selon la nature du repère temporel impliqué dans l'opération de localisation, trois types d'adverbiaux de localisation temporelle : les adverbiaux calendaires (*dès les années 30*), les adverbiaux relatifs à un procès – éventuellement nominalisé – (*dès son retour*) ou à un nom de temps (*avant la Belle Époque*) et les adverbiaux déictiques (*depuis le siècle dernier*). Conformément à la proposition de Battistelli *et al.* (2008), on distingue trois grands types d'opérateurs portant sur une référence temporelle noyau : les opérateurs de *régionalisation*, les opérateurs de *focalisation* et les opérateurs de *déplacement*.

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 sur une référence temporelle noyau et permettent de déterminer la région pointée par l'adverbial sur un axe du temps (ex. 1 et 2 où les marqueurs de l'opération mis en italique).

Ex. 1 : 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. 2 : 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* permet d'encoder les changements de « zoom » (focal) opérés par rapport à la granularité du repère temporel noyau (ex. 3 et 4).

Ex. 3 : 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. 4 : À Montpellier dimanche, l'ancien Toulousain a collé à l'encéphalogramme de l'équipe: totalement plat **jusqu'*au milieu de*** **la seconde période** (…).

Les opérations de *déplacement* décrivent les décalages opérés par rapport à la zone temporelle pointée initialement par la base (ex. 5 et 6).

Ex. 5 : « 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. 6 : 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.

La succession d'opérations sémantiques à l'aide de laquelle on représente les adverbiaux de localisation temporelle peut être notée sous la forme d'*expressions fonctionnelles* (*EF*) encodant les valeurs de chacune des opérations sémantiques[4] :

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

Ex. 2 : jusque trois mois avant le début des années 30
```
Régionalisation Jusque(
        Déplacement(mois,-3)(
                Focalisation Début(
                        Base Calendaire(
                                décennie : 1930))))
```

Cette analyse sémantique vise à rendre compte de la façon dont les adverbiaux de localisation temporelle déterminent un ancrage temporel, indépendamment de toute visée applicative. Elle présente cependant un intérêt pour l'ingénierie des

4. La valeur ID signifie qu'un opérateur laisse son opérande inchangé.

connaissances, puisqu'elle permet de décrire ces adverbiaux sans nécessairement lever l'éventuelle indétermination quant aux bornes de la zone temporelle qu'ils dénotent (*vers la mi-mars, au début des années 80, peu après la fin des élections*). Nous avons proposé un langage d'annotation qui reflète cette analyse et développé des ressources pour annoter une partie des adverbiaux de localisation temporelle (Teissèdre *et al.*, 2010). Si ce langage d'annotation est transposable vers le standard TimeML, qui est beaucoup plus englobant, au sens où les expressions temporelles qu'il vise peuvent recouvrir d'autres unités textuelles que les seuls adverbiaux de localisation temporelle, il découle néanmoins plus directement d'une analyse linguistique, en s'attachant à isoler et à décrire des adverbiaux que le standard ne décrit pas en tant que tels, mais sous une forme éclatée.

Dans la section suivante, nous décrivons un processus permettant d'articuler cette représentation qui découle d'une analyse linguistique avec une notation exprimée sous la forme d'intervalles calendaires, afin de pouvoir interagir avec des formats de représentation standard et mettre en œuvre des systèmes de recherche d'information en mesure de traiter des requêtes mêlant à la fois des critères thématiques et des critères calendaires. Nous décrivons également une heuristique pour mesurer la pertinence d'un document ou d'une portion de document pour ce type de requêtes.

4. Recherche d'information selon des critères calendaires

4.1. *Transposer la description sémantique des adverbiaux calendaires sous la forme d'intervalles calendaires*

L'heuristique décrite dans (Battistelli *et al.*, 2011 et 2012) 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 (*laïcité en France avant 1905 ; peinture italienne au début du XV^e^ siècle*, par exemple). Pour pouvoir traiter le critère calendaire dans ces requêtes (critère qui peut lui-même s'exprimer sous la forme d'un adverbial : *avant 1905* ; *au début du XV^e^ siècle*), il faut pouvoir évaluer la similarité entre ce critère et les unités textuelles dénotant un repérage temporel dans les textes.

La mesure de similarité que nous proposons articule deux représentations des adverbiaux de localisation temporelle, l'une *fonctionnelle* (décrite dans la section précédente), l'autre dite *référentielle*, qui s'exprime sous la forme d'intervalles calendaires. Nous définissons un processus de transduction de la première vers la seconde, qui est utilisée dans le calcul de similarité. 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 est une heuristique qui implique parfois une surdétermination dans la représentation résultante. C'est pourquoi il est utile de conserver la représentation sous la forme d'une succession d'opérations sémantiques et de faire dépendre des

besoins applicatifs la façon de transposer cette représentation vers une représentation sous la forme d'intervalles calendaires. Pour des expressions telles que *fin juin*, *vers la mi-mars*, *l'hiver prochain*, il n'y a pas de transposition sémantiquement équivalente sous la forme d'intervalles calendaires. En effet, s'il est possible de dire avec vraisemblance que *le 29 juin* est bien inclus dans la zone temporelle dénotée par une expression telle que *la fin du mois de juin*, en revanche, selon le contexte, il se peut que *le 19 juin* le soit également ou ne le soit pas : il n'y a donc pas de transposition univoque de cette expression vers la représentation calendaire.

À chacune des opérations sémantiques décrites dans une expression fonctionnelle à l'aide de laquelle on décrit les adverbiaux calendaires, on fait correspondre des transformations appliquées à l'intervalle calendaire associé au repère temporel noyau de l'adverbial. Ce repère est l'année *1905* dans l'adverbial *avant 1905*, par exemple. Ce processus de transduction nous permet d'associer un intervalle calendaire à chaque expression fonctionnelle. Nous décrivons à travers quelques exemples une partie de ces règles de transduction[5]. Notons que cette heuristique ne traite que les adverbiaux calendaires, c'est-à-dire des adverbiaux dont la référence temporelle noyau dénote une (ou plusieurs) zone(s) du calendrier[6].

Nous considérons un ensemble fini d'unités calendaires U, par exemple : $\{millénaire, siècle, décennie, année, mois, jour\}$. À chaque unité u de l'ensemble U est associée une séquence infinie de dates : $S(u) = \langle -\infty, \dots, u_{-n}, \dots, u_{-1}, u_0, u_1, \dots, u_m, \dots, +\infty \rangle$. $S(u)$ 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 : $S(u) = \langle \dots, 2010/11, 2010/12, 2011/01, 2011/02, \dots$.

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 u_i précède une date u_j, on notera $u_i < u_j$. Par exemple, on a $jour < année$ et $2010/11 < 2010/12$.

Afin de pouvoir comparer entre elles les interprétations associées à un ensemble d'adverbiaux calendaires (adverbiaux qui sont susceptibles de s'exprimer à différentes unités), on définit des applications permettant de les représenter à la même unité calendaire, celle la plus petite de l'ensemble considéré. On définit ainsi deux applications $début_{u \to v}$ et $fin_{u \to v}$ (où $v < u$), de telle sorte que pour une date u_i d'unité u :

5. Se reporter à (Battistelli *et al.*, 2012 ; Teissèdre, 2012) pour une présentation détaillée des règles de transduction.

6. Pour pouvoir traiter des adverbiaux dont le repère temporel noyau est formé par un déictique (*hier*, *jeudi dernier*) ou ancré relativement à un procès (*deux jours avant les élections*), il faudrait disposer d'un système en mesure de leur associer un intervalle calendaire. De nombreux travaux décrits plus haut traitent ce problème (« normalisation » des références déictiques et datation des événements décrits dans un texte).

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

Pour deux dates u_i et u_j, si $u_i < u_j$, on a :

$$fin_{u \to v}(u_i) < début_{u \to v}(u_j)$$

Si on a $début_{u \to v}(u_i) = v_j$ et $fin_{u \to v}(u_i) = v_k$, v_j correspond au début de u_i et v_k à la fin de u_i conformément à l'unité v. Ainsi, à toute date u_i, on peut associer un intervalle de dates d'unité v (où $v < u$) : $\langle v_j, v_k \rangle$. Par exemple, pour la date correspondant à l'*année 1997*, on a : $début_{année \to mois}(année\ 1997) = 1997/01$ et $fin_{année \to mois}(année\ 1997) =$ *1997/12*. On peut ainsi représenter cette date sous la forme d'un intervalle de dates d'unité *mois* $\langle 1997/01, 1997/12 \rangle$.

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 $u_i \leq u_j$) : il s'agit donc d'un ensemble compris entre deux dates u_i et u_j. Une autre notation possible est : $\langle i, j, u \rangle$. où i représente la date de début de l'IC, j la date de fin de l'IC et u l'unité calendaire considérée. Cette notation destinée à faciliter la lecture ne signifie pas qu'une date puisse être définie indépendamment d'une unité calendaire.

À chaque IC $\langle i, j, u \rangle$ 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 $\langle 1995/03/01, 1996/05/31, jour \rangle$.

4.2. *Propriétés des IC*

Considérons deux IC A et B, dont les unités sont respectivement u et v et où $v < u$, on a : $A = \langle g, h, u \rangle$, $t_{u \to v}(A) = \langle i, j, v \rangle$ et $B = \langle k, l, v \rangle$.

L'intersection de A et de B forme un IC qui se définit de la façon suivante : $A \cap B = \langle max(i, k), min(j, l), v \rangle$, 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 contient B (ou de B qu'il est inclus dans A) si et seulement si $i \leq k$ et $j \geq l$. Par exemple, on dira de l'IC $A = \langle 1980, 1989, année \rangle$ (pour lequel $t_{année \to mois}(A) = \langle 1980/01, 1989/12, mois \rangle$) qu'il contient l'$IC$ $B = \langle 1987/01, 1987/12, mois \rangle$.

Soit deux IC C et D. La longueur relative de C et de D (avec $C \neq \emptyset$) correspond à la valeur suivante :

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

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

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

On associe un intervalle infini aux adverbiaux tels que « *après 1989* » par exemple ($\langle 1989, +\infty, année\rangle$). Si D est un intervalle infini et C un intervalle fini non vide, alors on aura $\mathrm{rl}(\mathrm{C}/\mathrm{D}) = \varepsilon$, où ε est le plus petit des nombres strictement positifs. Si D et C sont des intervalles infinis, alors on a $\mathrm{rl}(\mathrm{C}/\mathrm{D}) = 1$.

À chaque *base calendaire* (la référence temporelle noyau d'un adverbial calendaire) 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 :

Ex. 1 : 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 : $\langle 198\_, 198\_, décennie\rangle$

4.3. *Interprétation calendaire des expressions fonctionnelles*

Considérons une expression fonctionnelle α à laquelle un IC $\langle i, j, u\rangle$ est associé. Si Ω est un opérateur de *focalisation*, on associe un IC à Ω (α) pour chaque unité v strictement inférieure à u de la façon suivante : on définit un coefficient τ compris entre 0 et 1/2 qui 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 19 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.

4.3.1. *Exemple de la* focalisation début

À 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 :

$$\langle début_{u \to v}(i), début_{u \to v}(i) + \lfloor \tau(fin_{u \to v}(j) - début_{u \to v}(i) + 1)\rfloor, v\rangle$$[7]

Le résultat de l'opération de *focalisation début* appliquée à l'IC $A = \langle 198\_, 198\_, décennie\rangle$ représentant les *années 80* dans l'adverbial *au début des années 80* peut par exemple être exprimé à l'unité $année$ ou à l'unité $mois$. Par exemple, l'image de A conformément à l'unité $année$ est $\langle 1980, 1989, année\rangle$.

7. La notation $\lfloor x\rfloor$correspond à la partie entière par défaut : le résultat est donc toujours un entier.

Telle que définie, à l'unité $année$, la transformation associée à l'opération de *focalisation début* produit le résultat suivant :

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

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 :

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

4.3.2. *Exemple du déplacement après*

Soit le *déplacement après* $(v, +n)$. À l'IC $\langle i, j, u \rangle$ associé à une 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 :

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

4.3.3. *Exemple de la régionalisation avant*

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

À l'adverbial *avant août 1974*, on associe ainsi, à l'unité mois, l'IC suivant :

$$\langle -\infty, 1974/07, mois \rangle$$

Ces règles de transduction permettent ainsi d'associer un intervalle de temps à un adverbial calendaire. Dans la section suivante, nous présentons une heuristique permettant de comparer entre eux ces intervalles et d'obtenir une mesure de similarité. L'implémentation de cette heuristique est exploitée par le système de recherche d'information que nous présentons dans la section 5.

4.4. *Une heuristique pour mesurer la similarité entre des intervalles calendaires*

Pour comprendre la démarche que nous proposons, considérons par exemple un corpus de documents relatifs à l'histoire des États-Unis. Un utilisateur pourrait s'intéresser, au sein de ce corpus, à des informations relatives à « *la prohibition au début des années 30* ». Une des réponses les plus pertinentes pourrait par exemple être dans cet extrait : « 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 dénotée 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. L'approche que l'on présente ne vise ainsi 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 similarité entre différents adverbiaux calendaires. L'inclusion n'est qu'un de ces critères, important certes, mais pas le seul.

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 plus haut. 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.

Considérons un ensemble A d'intervalles calendaires associés à un ensemble d'adverbiaux calendaires présents dans un corpus, conformément au processus de transduction présenté plus haut : $A = \{A_1, A_2, \dots, A_n\}$. Le critère calendaire exprimé dans une requête est également traduit sous la forme d'un intervalle calendaire, nommé Q. L'objectif est d'extraire un sous-ensemble $A(Q) = \left\{A_{i_1}, A_{i_2}, \dots, A_{i_p}\right\}$ dans l'ensemble A et de l'ordonner du plus pertinent au moins pertinent. Pour évaluer la pertinence des intervalles calendaires, on considère en premier lieu un critère d'adéquation avec la requête, et en cas d'égalité au niveau du critère d'adéquation, un critère d'ordre.

4.4.1. *Critère d'adéquation*

On établit trois critères d'adéquation entre un intervalle pris dans A et l'intervalle Q, du meilleur au moins bon : (1) l'*égalité*, (2) l'*inclusion* (de A_i dans Q) et (3) l'*inclusion inverse* (inclusion de Q dans A_i) ou l'*intersection* (entre A_i et Q). Dans le cas d'une *intersection vide*, on dira ainsi qu'aucun critère d'adéquation n'est satisfait. Ces critères peuvent être décrits sous la forme de relations d'Allen (Allen, 1983).

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

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*). Par exemple, si

Q représente l'adverbial *en 1980* et A_1 l'adverbial *de mars à mai 1980*, A_1 est inclus dans Q.

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*. Par 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. Si A_i et Q se chevauchent partiellement (*overlaps*), on parle d'*intersection*. Par 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. L'*intersection* n'est pas considérée comme étant un meilleur ou un moins bon critère d'adéquation que l'*inclusion inverse*. 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 qui font intervenir la longueur relative de deux intervalles calendaires :

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

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

Pour toute réponse satisfaisant le critère d'*égalité*, 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* (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* (par exemple la réponse A_2 *de 1978 à 1982*). Le rappel tend ainsi à privilégier des réponses de granularité supérieure à la requête, alors qu'un évaluateur humain tend à les considérer comme moins pertinentes. Dès lors, on introduit un coefficient α inférieur à 1, afin de 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 :

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

Le tableau 1 présente les scores attribués à différents adverbiaux susceptibles de répondre à la requête *en 1980*, pour une valeur d' α fixée à 0,4, permettant de privilégier fortement la précision sur le rappel. Des intervalles calendaires infinis sont également des réponses possibles : les scores qu'ils sont susceptibles d'obtenir peuvent être comparés à ceux des intervalles calendaires finis (ex. A'_3 et A''_3).

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 les requêtes auxquelles on associe des intervalles infinis (par exemple, une requête telle que *depuis 1980*). Pour cela, on introduit une mesure de distance, qui dépend des pôles des intervalles considérés.

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

Tableau 1. *Liste ordonnée de réponses pour la requête «* en 1980 *»*

On associe un pôle à chaque intervalle calendaire. 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 opération *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 31 décembre 1964* si l'intervalle calendaire associé à l'adverbial est exprimé à l'unité *jour*. Les expressions *mars 2009* et *mi-mars 2009* ont ainsi le même pôle.

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.

Réponses à la requête *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. *Liste ordonnée 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 leur a associés. L'ordonnancement des réponses est bien sûr discutable et soumis à interprétation (qu'est-ce qui permet de considérer dans l'absolu que la réponse A'_1 est meilleure que la réponse A''_1 ?). La mesure du Kappa (cf. section 5.2.) montre néanmoins qu'il y a un accord plutôt bon entre différents évaluateurs et que l'heuristique adoptée s'approche de leurs choix d'ordonnancement déduits à partir de la mesure de pertinence qu'ils ont associée à chacune des réponses cibles possibles proposées dans l'expérience d'évaluation.

4.5. *Traiter conjointement un critère thématique et un critère calendaire dans une requête*

Pour la mise en œuvre effective d'une application de recherche d'information temporelle, le critère calendaire d'une requête doit être traité conjointement au critère thématique exprimé sous la forme de mots-clés. Afin de limiter la recherche des mots-clés à un contexte voisin de celui où apparaît un adverbial calendaire,

l'index est formé avec des phrases contenant des adverbiaux calendaires et non avec des textes entiers. Ainsi, au lieu de courts extraits de textes segmentés autour des mots-clés retrouvés (*snippets*) que présentent généralement les moteurs de recherche, le démonstrateur présente une ou plusieurs phrases pour chaque document de la liste de résultats. Cela renvoie à deux problèmes auxquels doivent faire face les systèmes de recherche d'information, l'un générique, l'autre spécifique au traitement des adverbiaux 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 : comment s'assurer que l'adverbial calendaire présent 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 le problème de la résolution de la relation entre les adverbes et les segments textuels sur lesquels ils portent – problème qui appelle une analyse linguistique profonde. Cependant, même en restreignant la recherche des mots-clés et des adverbiaux dans le cadre d'une phrase, cette approche empirique ne garantit pas pour autant que l'adverbial et le segment textuel qui répond au critère thématique soient liés. En outre, elle ne permet pas de renvoyer des segments textuels de plus d'une phrase, même s'ils pourraient correspondre au critère thématique, alors que des adverbes de localisation temporelle, en particulier les adverbes dits cadratifs, sont susceptibles d'avoir une portée textuelle débordant une phrase (Le Draoulec et Pery-Woodley, 2005). Nous envisageons par la suite d'ajouter à la chaîne de traitements une phase d'analyse syntaxique pour repérer le segment textuel sur lequel porte les adverbiaux.

Pour mesurer la pertinence d'un document par rapport à une requête exprimée sous la forme de mots-clés, les systèmes de recherche d'information recourent à différentes formules de pondération, dont une des plus répandues est la formule du TF-IDF. Cette mesure statistique permet d'évaluer l'importance d'un terme contenu dans un document par rapport à sa distribution dans la collection indexée. Si cette approche est efficace pour rechercher des documents, elle l'est moins pour rechercher des phrases, le nombre d'occurrences d'un terme n'étant pas, à cette échelle, un bon critère de discrimination. Ainsi, plutôt que de recourir à ce type de mesure, nous distinguons plusieurs cas de figure du plus pertinent au moins pertinent pour évaluer l'intérêt d'une phrase du point de vue du critère thématique exprimé dans une requête : soit l'ensemble des mots-clés y est présent, soit un des mots-clés n'y est pas présent, soit deux des mots-clés n'y sont pas présents, etc. À l'intérieur de ces classes disjointes et ordonnées qui permettent d'obtenir un premier classement, le système trie ensuite par pertinence les phrases que chacune de ces classes regroupent en fonction du critère calendaire de la requête, au moyen de l'heuristique décrite précédemment. On obtient ainsi un ensemble ordonné de phrases. À l'échelle supérieure, pour trier les documents par pertinence, le système les ordonne dans un premier temps d'après la phrase la plus pertinente de chacun d'eux. En cas d'égalité entre deux documents, le système compare alors les

deuxièmes phrases les plus pertinentes extraites de ces documents. Ces étapes sont répétées jusqu'à ce que l'ensemble des documents soit ordonné.

5. Implémentation et évaluation

5.1. *Le système CaSE*

Le système CaSE (*Calendar Search Engine*) implémente l'algorithme de filtrage et d'ordonnancement d'intervalles calendaires présentés dans la section précédente. Le moteur de recherche permet aux utilisateurs d'exprimer des requêtes associant des critères thématiques et des critères calendaires (figure 1).

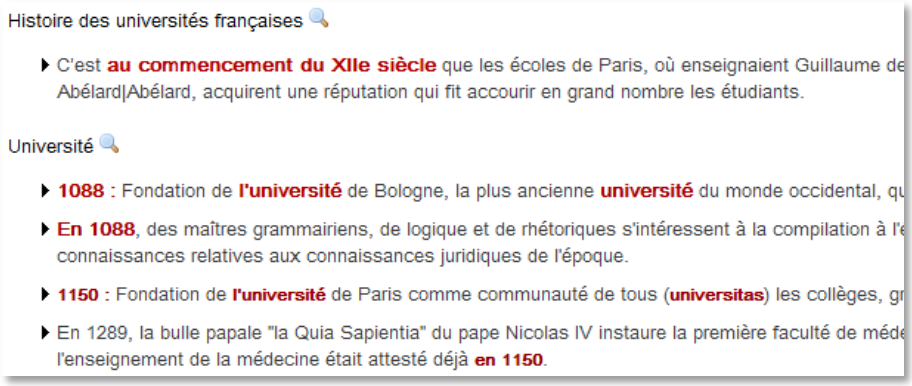

Figure 1. *Copie d'écran de la liste des résultats proposés pour la requête* « université au début du XII[e] siècle »

Les documents indexés et les requêtes sont annotés par un système d'annotation des adverbiaux calendaires présenté dans (Teissèdre *et al.*, 2010 et Teissèdre, 2012). S'appuyant sur un ensemble de lexiques et de patrons (ou grammaires locales), implémentés manuellement sous la forme de transducteurs Unitex (Paumier, 2002), ce module d'annotation exploite un langage d'annotation découlant de l'analyse linguistique des adverbiaux de localisation temporelle. Ses performances ont été évaluées sur le corpus FR-Timebank (Bittar, 2010) et sont comparables aux systèmes existants qui visent à la résolution de tâches similaires (précision : 0,87, rappel : 0,82, F-Mesure : 0,84). Les adverbiaux sont ensuite transposés sous la forme d'intervalles calendaires pour pouvoir effectuer les mesures de similarité. Le système exploite le moteur de recherche Open Source Apache Solr[8] pour l'indexation et la recherche des mots-clés. Le système CaSE est ainsi une surcouche logicielle du moteur de recherche Solr. L'ensemble des ressources développées a

8. http://lucene.apache.org/solr/

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 de recherche d'information qui procèderaient d'une telle démarche.

Les expérimentations menées avec le système CaSE illustrent 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 permet de traiter des requêtes exprimant des critères calendaires « absolus » (*en 1920, depuis la fin du XIX^e^ siècle, dans les années 80*).

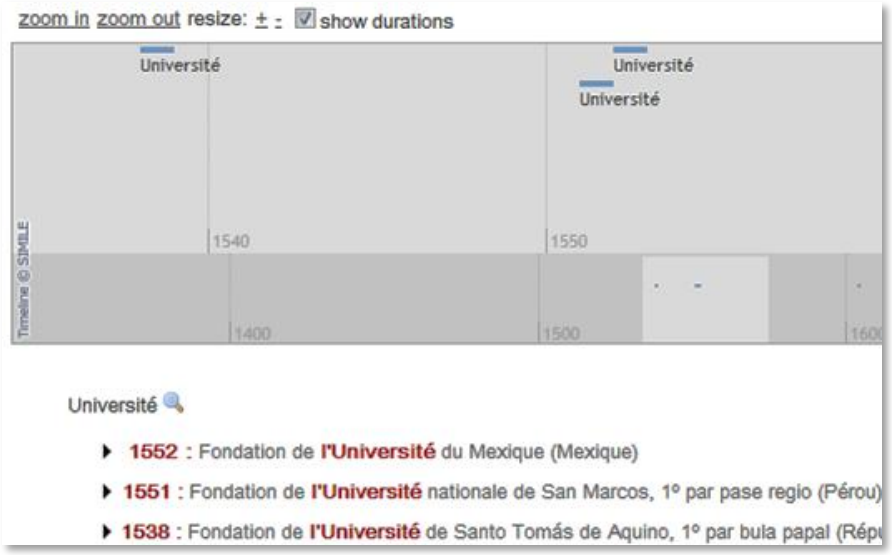

Figure 2. *Copie d'écran de la liste des résultats proposés suite à un déplacement sur la frise chronologique*

En plus de la recherche documentaire, le système permet également à l'utilisateur d'effectuer des recherches à l'intérieur d'un document donné. Une frise chronologique[9] est également présentée au-dessus de la liste des résultats d'une recherche (figure 2). Les interactions avec la frise chronologique (déplacement-zoom) génèrent de nouvelles requêtes sur l'index, ce qui permet à l'utilisateur d'affiner sa recherche et de parcourir temporellement un document donné ou l'ensemble de la collection documentaire.

9. Cette représentation graphique exploite le composant SIMILE Timeline (http://www.simile-widgets.org/timeline/)

5.2. *Évaluation de la mesure de pertinence des adverbiaux calendaires*

L'algorithme de sélection et de tri par pertinence des adverbiaux calendaires utilisé dans le système CaSE a été évalué à l'aide de la mesure de *précision moyenne* (*Mean Average Precision*). Cette évaluation visait à mesurer la capacité du modèle de tri à faire remonter les adverbiaux calendaires les plus pertinents dans un ensemble par rapport à un adverbial considéré comme une requête : par exemple, pour une « requête » telle que *dans les années 80*, est-ce que les adverbiaux *en 1981* et *à la fin des années 70* sont pertinents ? L'évaluation a consisté à comparer les sorties du système avec un filtrage et une mesure de pertinence effectuée à la main. Le corpus d'évaluation consistait en un ensemble de quatre-vingt-dix adverbiaux calendaires. Cet ensemble a été scindé en trois sous-ensembles de trente adverbiaux, soumis chacun à deux évaluateurs. Cinq adverbiaux calendaires considérés comme des requêtes étaient associés à chaque ensemble testé. Les évaluateurs ont assigné une mesure de pertinence à chaque adverbial de leur sous-corpus pour chacune des cinq requêtes. La mesure de pertinence était comprise dans une échelle graduée allant de 0 à 4. Afin d'obtenir un score de précision moyenne (MAP), où les adverbiaux ne sont séparés qu'en deux classes (pertinents ou non pertinents), les scores inférieurs ou égal à 2 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 interévaluateurs. Dans la mesure où les classes de pertinence ne sont pas disjointes (il s'agit d'une graduation), cet accord est évalué à l'aide de la mesure dite Kappa pondéré (Weighted Kappa) (Cohen, 1968).

Dans le cadre de notre évaluation la mesure du Kappa tend à montrer que l'accord entre évaluateurs est plutôt bon (Kw = 0,70), sachant qu'il y avait cinq catégories possibles de pertinence. Les résultats de l'évaluation (MAP 0,86) montrent de son côté que le modèle d'ordonnancement des adverbiaux calendaires produit des résultats très proches de ceux des évaluations produites manuellement.

5.3. *Évaluation du système CaSE*

Comparer les sorties du système CaSE avec celles d'autres systèmes similaires n'a pas été possible, parce que à notre connaissance, soit l'approche du problème retenue dans ces systèmes était trop éloignée (approche par filtrage sur des plages de temps plutôt que par évaluation de la pertinence selon des critères temporels et thématiques, comme dans (Matthews *et al.*, 2010)), soit les corpus utilisés dans leurs évaluations n'étaient pas disponibles (Arikan *et al.*, 2009), soit encore les requêtes testées étaient limitées à des dates simples (*juin 2009*, par exemple) ne permettant pas d'exprimer un repérage temporel plus complexe (*fin juin 2009* ou *avant juin 2009*, par exemple), ce qui, à l'issue d'une évaluation complexe à mettre en œuvre, n'aurait permis une comparaison que sur une portion du problème traité (Berberich

et al., 2010). L'évaluation expérimentale présentée ici pourrait néanmoins être complétée par la suite par une évaluation plus parcellaire avec d'autres systèmes.

Un corpus de vingt-huit requêtes combinant des mots-clés et un critère calendaire a été constitué pour évaluer le système CaSE (*peine de mort depuis les années 70 ; De Gaulle après mai 68 ; vote des femmes depuis 1900 ; journaux à la fin du XVIII^e siècle*, etc.). Ces requêtes ont été testées sur un corpus de plus de 3 500 articles de Wikipédia collectés à l'aide de l'API Google Search. Afin de constituer l'index, pour chaque requête du corpus, l'API de Google a été successivement appelée une fois avec, puis une fois sans le critère calendaire (par exemple, *vote des femmes depuis 1900*, puis *vote des femmes*). Les cent premiers documents retournés par l'API Google Search pour chacun de ces appels (hors les doublons) ont été reversés dans la collection documentaire indexée.

L'évaluation a consisté à comparer les résultats proposés par le système CaSE avec ceux d'un moteur de recherche Solr paramétré de façon standard pour une recherche plein texte. Pour chaque requête, les dix premiers résultats fournis par chacun des moteurs ont été évalués manuellement, en évaluant (1) le système CaSE (MAP 0,43), (2) le système Solr avec des requêtes associant des mots-clés et un critère calendaire (MAP 0,27) et (3) le système Solr avec des requêtes ne contenant que les mots-clés, sans le critère calendaire (MAP 0,23). L'évaluation donne des indicateurs de comparaison entre les différents systèmes testés plutôt qu'elle ne mesure la performance des systèmes en eux-mêmes, car il faudrait parcourir manuellement toute la collection des documents pour établir un taux de rappel et de précision juste. Les résultats obtenus montrent néanmoins l'intérêt de la démarche, les performances du système CaSE étant sensiblement supérieures à celles des deux autres systèmes testés.

Le critère calendaire étant traité comme un mot-clé par le moteur Solr paramétré de façon standard pour une recherche plein texte, sa présence peut perturber considérablement les résultats. De fait, son traitement comme mots-clés dénués de sémantique peut favoriser des pages où tout ou partie de l'expression apparaît, même sans rapport avec la recherche thématique (par exemple la requête *langue française de 1520 à 1600*). À l'inverse, si le critère calendaire exprimé dans la requête couvre une période de temps très étendue ou mieux s'il est fréquemment associé aux mots-clés, alors ce critère devient moins discriminant (par exemple *crise économique au XIX^e siècle* ou *ségrégation depuis les années 50*). 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. Néanmoins, les résultats obtenus par le moteur Solr testé avec les requêtes combinant à la fois des mots-clés et un critère calendaire sont légèrement supérieurs à ceux obtenus par le même système testé en ne conservant que le critère thématique : en effet, si l'adverbial exprimant le critère calendaire apparaît tel quel dans des documents, ceux-ci sont avantagés par rapport aux autres documents où il n'apparaît pas (par exemple pour une requête telle que *université vers le XIII^e siècle* où la présence d'une référence au *XIII^e siècle* écarte l'ensemble des documents où aucune référence à ce siècle n'est présente).

6. Perspectives

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, afin, par exemple, de suivre l'évolution d'un événement, de retracer la biographie d'une personne, d'observer un mouvement artistique. 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 ainsi une piste originale pour la recherche d'information.

L'analyse des adverbiaux de localisation temporelle présente ainsi un intérêt pour la recherche documentaire et la navigation dans les textes, mais également pour la gestion des connaissances, 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 représentation des données temporelles dans des bases de connaissances – données qui sont présentes dans de nombreux domaines d'application. Nos travaux ultérieurs s'attacheront ainsi à proposer des modèles formels plus expressifs pour constituer des bases de connaissances contenant des informations temporelles.

Nous avons essayé de montrer que l'analyse linguistique des adverbiaux de localisation temporelle présents dans les textes invite à réexaminer en partie les modèles qui prédominent dans le champ du traitement automatique des langues. En effet, le plus souvent ces modèles ne s'intéressent aux unités textuelles permettant d'ancrer dans le temps les situations décrites dans les textes qu'en tant qu'elles peuvent être ramenées à des valeurs calendaires (à des dates), 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 d'un point de vue linguistique, avant de s'intéresser à leur transposition sous la forme de valeurs calendaires –, il devenait possible d'apporter des réponses à une partie des difficultés rencontrées par les systèmes de recherche d'information temporelle, parce que l'on replace ainsi au cœur du problème la question de l'articulation de deux représentations non équivalentes.

7. Bibliographie

Ahn D., van Rantwijk J., de Rijke M., A cascaded machine learning approach to interpreting temporal expressions, *Proceedings of NAACL-HLT'07*, Rochester, NY, USA, April, 2007, p. 284-291.

Allen J. F., Maintaining knowledge about temporal intervals, *Communications of the ACM*, 26 (11), 1983, p. 832-843.

Alonso O., Gertz M., Baeza-Yates R., On the value of temporal information in information retrieval, *Proceedings of ACM SIGIR Forum*, 41 (2), 2007, p. 35-41.

Arikan I., Bedathur S., Berberich K., Time Will Tell: Leveraging Temporal Expressions in IR, *Proceedings of WSDM'09*, Springer, 2009, p. 13-25.

Aurnague M., Bras M., Vieu L., Asher N., The syntax and semantics of locating adverbials, *Cahiers de Grammaire*, 26, 2001, p. 11-35.

Battistelli D., Couto J., Minel J. L., Schwer S. R., Représentation algébrique des expressions calendaires et vue calendaire d'un texte, *Actes de TALN'08 (1)*, 2008, p. 9-13.

Battistelli D., *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, novembre 2009.

Battistelli D., Cori M., Minel J. L., Teissèdre C., Semantics of Calendar Adverbials for Information Retrieval, *Proceedings of ISMIS 2011*, Warsaw, Poland, 2011, p. 622-631.

Battistelli D., Cori M., Minel J. L., Teissèdre C., Information Retrieval: Ranking Results according to Calendar Criteria, *Proceedings of IPMU 2012*, July 9-13, Catania, Italy, volume 297 of CCIS, Springer, 2012, p. 460-470.

Berberich K., Bedathur S., Alonso O., Weikum G., A language modeling approach for temporal information needs, *Proceedings of Advances in Information Retrieval (ECIR 2010)*, Springer, 2010, p. 13-25.

Bevort É., Bonvoisin S., Frémont P., Savino J., *Historiens et géographes face à la médiatisation de l'événement*, Centre national de documentation pédagogique, 1999, p. 197.

Bittar A., Construction d'un TimeBank du français : un corpus de référence annoté selon la norme ISO-TimeML, *Thèse de doctorat*, Université Paris-Diderot, 2010.

Blumenthal P., Classement des adverbes : pas la couleur, rien que la nuance ?, *Langue française*, 88, 1990, p. 41-50.

Borillo A., Les adverbes de référence temporelle comme connecteurs temporels de discours, *Temps et discours*, 1998, p 131-145.

Calabrese Steimberg L., La construction de la mémoire historico-médiatique à travers les désignations d'événements, *Travaux du Cercle belge de linguistique*, n° 1, 2006, p. 1-16.

Chinchor N., Brown E., Ferro L., Robinson P., Named Entity Recognition Task Definition, version 1.4. Technical Report, MITRE and SAIC. 1999 [en ligne] ftp://jaguar.ncsl.nist.gov/ace/phase1/ne99_taskdef_v1_4.pdf

Cohen J., Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit, *Psychological Bulletin*, vol. 70 (4), 1968, p. 213-220.

Ferro L., Gerber L., Mani I., Sundheim B., Wilson G., TIDES Standard for the Annotation of Temporal Expressions. 2003 [en ligne] http://www.mitre.org/sites/default/files/pdf/ferro_tides.pdf

Filatova E., Hovy E., Assigning time-stamps to event-clauses, *Proceedings of ACL Workshop on Temporal and Spatial Information Processing*, 2001, p. 88-95.

Gagnon M., Lapalme G., From conceptual time to linguistic time, *Computational Linguistics*, vol. 22 (1), 1996, p. 91-127.

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

Gosselin L., *Temporalité et modalité*, De Boeck-Duculot, Bruxelles, 2005.

Han B., Gates D., Levin L., From language to time: A temporal expression anchorer, *Proceedings of TIME'06*, IEEE Computer Society, June 2006, p. 196-203.

Hobbs J. R., Pustejovsky J., Annotating and reasoning about time and events. *Proceedings of AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, AAAI Press, Menlo Park, California, 2003, p. 74-82.

Järvelin K., Kekäläinen J., Cumulated Gain-based Evaluation of IR Techniques, *Proceedings of ACM Transactions on Information Systems*, vol. 20, n° 4, October 2002, p. 422-446.

Le Draoulec A., Péry-Woodley M.n Encadrement temporel et relations de discours, *Langue française*, vol. 148, 2005, p. 45-60.

Mani I.,Wilson G., Robust temporal processing of news, *Proceedings of the Association for Computational Linguistics (ACL2000)*, Hong-Kong, 2000, p. 69-76.

Matthews M., Tolchinsky P., Mika P., Blanco R., Zaragoza H., Searching through time in the New York Times Categories and Subject Descriptors,. *Proceedings of HCIR 2010* – Challenge Report, New Brunswick, 2010, p. 41-44.

Maurel D., Grammaire des dates, étude préliminaire à leur traitement automatique, *Linguisticae Investigationes*, vol. 12, n° 1, 1988, p. 101-128.

Maurel D., Adverbes de date : étude préliminaire à leur traitement automatique, *Linguisticae Investigationes*, vol. 14, n° 1, 1990, p. 31-63.

Muller P., Tannier X., Annotating and measuring temporal relations in texts, *Proceedings of COLING '04*. Morristown, NJ, USA, 2004, p. 50-56.

Nunes S., Ribeiro C., David G., Use of Temporal Expressions in Web Search. *Proceedings of European Conference on IR Research (ECIR 2008)*, 30th March-3rd April, Glasgow, Scotland, Springer Berlin/Heidelberg, vol. 4956, 2008, p. 580-584.

Paumier S., Manuel d'utilisation du logiciel Unitex. IGM, université de Marne-la-Vallée, 2014, http://www-igm.univ-mlv.fr/~unitex/ManuelUnitex3.1.pdf

Pustejovsky J., Belanger L., Castano J., Gaizauskas R., Hanks P., Ingria R., Katz G., Radev D., Rumshisky A., Sanfilippo A., Sauri R., Sundheim B., Verhagen M., TERQAS Final Report, 2002, [en ligne] http://www.timeml.org/site/terqas

Pustejovsky J., Castaño J., Ingria R., Saurí R., Gaizauskas R., Setzer A., Katz G., TimeML: Robust Specification of Event and Temporal Expressions in Text. *Proceedings of IWCS-5, Fifth International Workshop on Computational Semantics*, 2003a , p. 28-34.

Pustejovsky J., Hanks P., Sauri R., See A., Gaizauskas R., Setzer A., Radev D., Sundheim B., Day D., Ferro L., Lazo M., The TimeBank Corpus, *Corpus Linguistics*, 2003b, p. 647-656.

Pustejovsky J., Lee K., Bunt H., Romary L., ISO-TimeML: An International Standard for Semantic Annotation, *Proceedings of LREC 2010*, 2010.

Saquete E., Martinez-Barco P., Muñoz R., Vicedo J. L., Splitting Complex Temporal Questions for Question Answering systems, *Proceedings of the Association for Computational Linguistics (ACL)*, Barcelona, Spain, July 2004, p. 566-573.

Saurí R., Littman J., Knippen B., Gaizauskas R., Setzer A., Pustejovsky J., TimeML annotation guidelines, 2006, http://www.timeml.org/timeMLdocs/AnnGuide14.pdf

Saurí R., Pustejovsky J., TimeML in a Nutshell. 2009 [en ligne] http://www.timeml.org/tempeval2/tempeval2-trial/guidelines/introToTimeML-052809.pdf

Schilder F., Habel C., From temporal expressions to temporal information: Semantic tagging of news messages, *Proceedings of the ACL-2001 Workshop on Temporal and Spatial Information Processing*, Toulouse, France, 2001, p. 65-72.

Schütze S., Caravage, L'œuvre complet. *Taschen*, Cologne, 2009.

Setzer A., Gaizauskas R., Annotating events and temporal information in newswire texts, *Proceedings LREC 2000*, 2000, p. 1287-1294.

Setzer A., *Temporal Information in Newswire Articles: an Annotation Scheme and Corpus Study*. Ph.D. thesis University of Sheffield, UK, 2001.

Setzer A., Gaizauskas R., On the importance of annotating event-event temporal relations in text, *Proceedings of the LREC 2002 Workshop on Temporal Annotation*, 2002, p. 52-60.

Teissèdre C., Battistelli D., Minel J. L., Resources for Calendar Expressions Semantic Tagging and Temporal Navigation through Texts, *Proceedings of LREC 2010*, Malta, 2010, p. 3572-3577.

Teissèdre C., *Analyse sémantique automatique des adverbiaux de localisation temporelle : application à la recherche d'information et à l'acquisition des connaissances*, Thèse de doctorat, université Paris-Ouest Nanterre La Défense, 2012.

Van Raemdonck D., Est-il pertinent de parler d'une classe d'adverbes de temps ? *Circulo de lingüística aplicada a la comunicación*, vol. 7, 2001.

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

Verhagen M., Gaizauskas R., Schilder F., Hepple M., Katz G., Pustejovsky J., Semeval-2007 task 15: Tempeval temporal relation identification, *Proceedings of the 4th International Workshop on Semantic Evaluations*, 2007, p. 75-80.

Verhagen M., Saurí R., Caselli T., Pustejovsky J., SemEval-2010 Task 13: TempEval-2, *Computational Linguistics*, 2010, p. 57-62.

Wilson G., Mani I, Sundheim B., Ferro L., A multilingual approach to annotating and extracting temporal information, *Proceeding of the workshop on Temporal and spatial information processing TASIP '01*, 2001, p. 81-87.