Nouvelles perspectives en extraction d'information

Michel Dupont, Jean-Marc Vuillaume, Bernard Victorri, Patrice Enjalbert, Yann Mathet, Nicolas Malandain

Technique et Science Informatique, 21/1, 2002 – Version préliminaire

Nouvelles perspectives en extraction d'information

**Michel Dupont, Jean-Marc Vuillaume, Bernard Victorri,
Patrice Enjalbert, Yann Mathet, Nicolas Malandain**

1. Introduction

Les techniques dites d'extraction d'information (EI) ont connu un essor considérable ces dix dernières années. L'EI consiste à extraire de documents des informations précises et à les structurer sous une forme prédéfinie. Il s'agit en général de remplir des formulaires donnant certaines caractéristiques concernant des entités ou des événements évoqués dans les textes ainsi que des relations entre ces entités et ces événements. Le formulaire est constitué d'une liste d'attributs auxquels le système doit faire correspondre une liste de valeurs pour chaque texte analysé. Ainsi, l'EI permet de stocker dans une base de données factuelles l'information jugée pertinente en vue de traitements ultérieurs [PAZ 97][1].

Dans le vaste éventail des traitements automatiques de documents textuels, il sera commode de situer l'EI, tant par ses objectifs que par ses méthodes, comme un niveau intermédiaire entre la recherche documentaire d'une part, et la compréhension automatique, au sens de l'intelligence artificielle (IA), de l'autre :

- En recherche documentaire, l'objectif général consiste à faciliter la sélection d'un sous-ensemble de documents pertinents dans une base documentaire en réponse à la requête d'un utilisateur. Le résultat du traitement est le document lui-même, sans compréhension ou interprétation de son contenu. Les traitements se réduisent généralement à une analyse du contenu lexical du texte, et éventuellement de sa « structure matérielle » (titres, résumé…), sans prise en compte de la structure syntaxique et sémantique des phrases.

- A l'opposé, en compréhension automatique, le but est d'obtenir une représentation du sens du texte donné. Cet objectif réclame une analyse exhaustive, syntaxique et sémantique, de chaque phrase et des relations qu'elles entretiennent, et, ce qui est tout aussi difficile, la construction d'une base de connaissances et l'élaboration d'un formalisme de représentation du sens capables de couvrir des domaines très vastes de l'expérience humaine.

En EI, on produit également des représentations sémantiques externes au document. Mais on se donne la tâche de comprendre non pas le texte dans son ensemble mais des parties extrêmement ciblées quant à la structure de l'information recherchée et quant aux formes linguistiques qui la portent. L'EI apparaît ainsi comme un bon compromis, susceptible d'aboutir à la réalisation de systèmes opérationnels *d'analyse de contenu* de documents textuels, complémentaires de la recherche documentaire[2].

Cette « technologie » s'est notablement développée à la faveur de la série de conférences MUC (*Message Understanding Conferences*) qui ont permis de confronter divers systèmes d'EI sur des tâches et des corpus communs. Une large communauté scientifique s'est ainsi constituée, posant des bases méthodologiques solides et permettant une « accumulation primitive » de techniques et de systèmes logiciels. Nous pensons que ces acquis contribueront notablement à renouveler les objectifs et les méthodes de l'informatique documentaire dans son ensemble.

Le présent article présente les travaux menés dans notre équipe en extraction d'information et, à partir de cette expérience, propose différentes orientations tant pour améliorer la technologie elle-même que pour en élargir les applications. Dans la section 2 nous commencerons par une présentation rapide de l'expérience des conférences MUC et nous en discuterons les acquis et limites. La section 3 sera consacrée à nos propres travaux sur une tâche du même type concernant un corpus de constats

[1] Noter toutefois que le terme « extraction d'information » est aujourd'hui revendiqué par différentes approches textuelles, le point commun étant la recherche d'informations ciblée et prédéfinies dans un corpus de textes. Mentionnons notamment des approches à base essentiellement lexicale et statistique, visant à extraire « de l'information » non pas de chaque texte individuellement, mais d'un corpus pris dans son ensemble. Par exemple on essaiera de repérer les thématiques récurrentes et d'en analyser les variations [FER 00]. On tend à utiliser le terme *d'extraction d'information structurée* pour l'approche qui nous concerne.

[2] On le verra toutefois, la séparation entre recherche documentaire et EI est en fait bien moins tranchée.

d'accidents, occasion pour discuter quelques points de méthodologie. Nous aborderons alors la démarche que nous suivons pour surmonter les limites actuelles de l'EI. D'un côté, il s'agit de diversifier la tâche, en quelque sorte de l'affaiblir de manière à rendre les traitements suffisamment efficaces et fiables pour une utilisation « réelle ». Nous présenterons un système d'extraction de localisation spatiale et temporelle dans des documents de géographie humaine visant à un « encodage sémantique » du texte. Les applications visées vont de l'aide à la lecture à la structuration automatique de documents composites (section 4). La seconde orientation consiste à développer des méthodes sémantiques approfondies, en nous appuyant sur des connaissances linguistiques que nous adaptons à la tâche. Nous présenterons des travaux en cours sur la référence et sur la sémantique spatiale (section 5). Un bilan de l'ensemble de ces expériences pourra alors être tiré en conclusion.

2. L'expérience des conférences MUC

Les premières recherches en EI sont relativement anciennes. Parmi les précurseurs, on peut citer notamment un système d'extraction de noms propres et de titres à partir d'extraits de journaux [BOR 67] ou le système de Sager de traitement de rapports de radiologie [SAG 81]. Mais c'est à partir de la fin des années 1980 que la recherche en EI a pris véritablement son essor avec l'organisation des premières des sept conférences MUC organisées entre 1987 et 1998 [MUC 91, 92, 93, 95, 99]. Leur intitulé (*Message Understanding Conference*) est quelque peu trompeur puisqu'il ne s'agit pas à proprement parler de conférences et que le thème n'est pas la compréhension automatique mais bien l'extraction d'information. En fait, il s'agissait essentiellement lors de ces manifestations de confronter les systèmes réalisés par plusieurs équipes en comparant leurs performances grâce à des mesures précises et objectives. Comme toujours dans ce genre de « compétition » l'organisation de la confrontation comprend deux phases : une première phase d'entraînement pendant laquelle les systèmes sont mis au point sur un premier corpus ; et la phase d'évaluation proprement dite.

2.1 La tâche

Les textes traités sont des textes courts à caractère descriptif (dépêches d'agences, articles journalistiques) « ciblés » sur un type d'information très spécifique : opérations militaires navales (MUC 1 et 2), attentats terroristes (MUC 3 et 4), joint ventures et circuits électroniques (MUC 5), mouvements de personnels de direction de grandes entreprises (MUC 6) ou lancement d'engins balistiques (MUC 7). Il s'agit de remplir un ensemble de formulaires au format préétabli (*templates* dans la terminologie MUC), de la forme *entités typées – relations*, synthétisant les (principales) informations contenues dans le texte. La figure 2 de la section 3 présente un exemple, issu de nos propres travaux, de texte et le schéma des formulaires à instancier.

Au fil de l'expérience, certaines phases du traitement sont apparues comme suffisamment importantes et génériques pour faire l'objet d'une étude et d'une évaluation spécifiques, notamment :

- L'extraction des **entités nommées**, c'est-à-dire le repérage de toutes les formes linguistiques qui, à l'instar des noms propres, désignent de manière univoque une entité par leur pouvoir de sélectivité : dates, heures, valeurs monétaires, quantités…
- Le calcul de **coréférence**. Il s'agit de reconnaître toutes les formes linguistiques qui réfèrent à une même entité ou un même événement (*chaînes de coréférence*).

Plusieurs caractéristiques apparaissent déjà dans ce court descriptif : les textes traités relèvent d'un domaine d'une très grande spécificité. On connaît exactement le schéma des informations à extraire et on sait qu'elles sont présentes (à quelques détails près). Autrement dit, il s'agit *d'instancier* un formulaire au schéma préétabli en *reconnaissant* dans le texte les informations pertinentes. Il s'agit donc bien d'une forme de compréhension automatique — et nous verrons dans les prochaines sections que l'on en retrouve pratiquement tous les aspects « classiques » — mais d'une compréhension *limitée*, sur un ensemble *fermé* de textes et d'informations extraites. Ces caractéristiques sont des conditions essentielles de faisabilité avec un niveau suffisant de fiabilité et d'efficacité.

2.2 Architecture d'un système d'EI

L'architecture très générale présentée dans la figure 1 fait apparaître les principales composantes d'un système d'EI. La phase dite de prétraitements consiste en un ensemble d'opérations « de surface » sur le matériau linguistique, permettant d'entreprendre l'analyse linguistique avec un texte « nettoyé » et « préparé ».

Dans la seconde phase, l'analyse morphologique consiste à étiqueter (*tagging*) les mots selon leur classe (nom, verbe, adjectif…) et à repérer leur genre, nombre, personne etc.[3] L'analyse syntaxique doit produire (de manière plus ou moins exhaustive) les relations grammaticales : relations sujet-verbe ou verbe-COD, rattachements prépositionnels (compléments de nom ou compléments indirects des verbes), etc. L'analyse sémantique vise à construire à partir de chaque proposition une « représentation conceptuelle » (au sens de l'IA [SAB 90]) sous forme d'expression logique ou de réseau sémantique. Enfin l'analyse du discours doit établir les liens entre les différentes phrases, typiquement repérer les coréférences ou l'ordre temporel des énoncés. L'organisation de ces traitements, et bien sûr les méthodes linguistiques utilisées constituent des caractéristiques importantes des différents systèmes d'EI.

Enfin, à partir de cette représentation conceptuelle « abstraite », il s'agit de remplir les champs des formulaires. Ce qui suppose notamment d'affiner le calcul de l'identification des entités et des événements. Du point de vue des méthodes, une caractéristique de cette étape est d'être complètement orientée par le but, c'est à dire par la structure des formulaires, et de recourir à des inférences mettant en jeu des connaissances du domaine (« le monde financier», « le monde de la route »…) alors que la phase d'analyse linguistique est, elle, plutôt orientée par la structure linguistique du texte. D'une certaine manière, il s'agit d'un pendant de la phase *d'interprétation en contexte* de l'architecture classique en compréhension automatique [ALL 95].

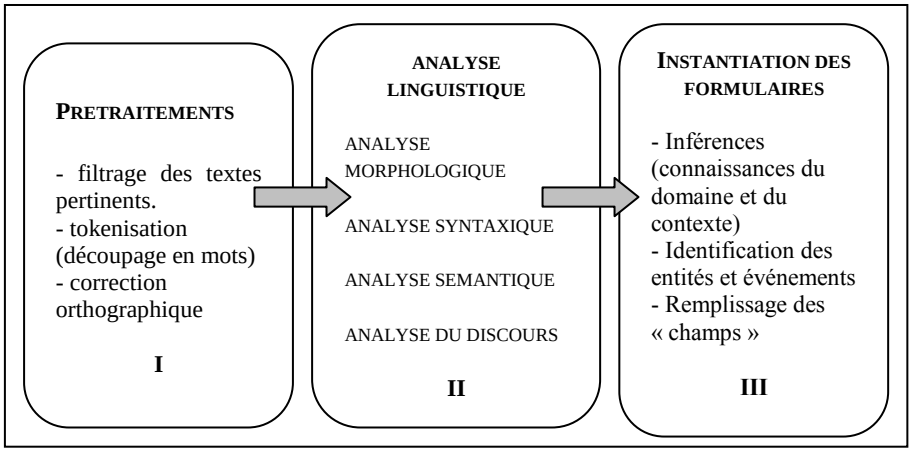


Figure 1 : architecture générale d'un système d'EI

2.3 Evaluation et résultats

La mise au point de méthodes d'évaluation des systèmes de Traitement Automatique du Langage (TAL) apparaît aujourd'hui comme une composante importante, quoique difficile, de la recherche dans ce domaine [RUB 00]. Même si cette évaluation présente bien sûr des limites [POI 99], la spécification des formats de sortie et des critères d'évaluation ainsi que la réalisation d'outils d'évaluation automatique constituent par elles-mêmes des avancées certaines pour la recherche.

[3] Cette tâche est parfois intégrée dans les « prétraitements ».

L'idée générale, inspirée de méthodes courantes en informatique documentaire, est de calculer deux mesures : le *rappel*, qui est le rapport du nombre de réponses correctes sur le nombre de réponses attendues ; et la *précision*, qui est le rapport du nombre de réponses correctes sur le nombre total de réponse fournies par le système. La *F-mesure* pondère rappel et précision en un indicateur unique permettant un classement général des systèmes. [4]

Les résultats des meilleurs systèmes vont de 42% à 62% en rappel et de 53% à 70% en précision pour les conférences MUC 4 à 7. Ces résultats doivent être mis en relation avec les performances humaines sur la même tâche, qui sont loin d'être parfaites, et que [HOB 96] estime à 65-80%. En fait, on peut avancer que les performances des meilleurs systèmes sont de l'ordre de 70% à 80% des performances humaines. Il faut ajouter à ce bilan que le portage d'un domaine à un autre peut être réalisé en quelques mois par un spécialiste de la technologie.

2.4 Les enjeux actuels

Le bilan d'ensemble de « l'expérience MUC » est, pensons-nous double :

- D'une part des progrès considérables ont été réalisés sur tous les plans : linguistique, méthodologique, informatique… Même si les résultats actuels ne sont pas encore à la hauteur des espérances des organisateurs de MUC, le domaine a acquis aujourd'hui une maturité suffisante pour que les recherches puissent se diversifier en fonction de besoins applicatifs précis [GRI 97].
- Néanmoins, les résultats mentionnés plus haut montrent que le niveau de fiabilité atteint et les temps de portage ne sont pas encore suffisants pour un large développement industriel sur des tâches du type MUC.[5] L'heure est donc à une appréciation critique de l'avenir de ces techniques.

En premier lieu il convient de resituer l'EI dans le champ plus général des traitements informatiques du document : Quelle est sa place vis-à-vis des tâches traditionnelles de l'informatique documentaire ? Quels nouveaux types de traitements sont envisageables à partir de l'expérience MUC ? A quelle condition peuvent-ils déboucher sur des applications industrielles à court terme ? L'idée est ici, comme l'argumente Th. Poibeau [POI 99], de ne pas hésiter à s'éloigner du modèle MUC pour mieux s'ajuster aux besoins industriels. Et de récupérer les acquis en EI, en complément d'autres techniques, pour des applications diversifiées réclamant des traitements moins élaborés que ceux qui étaient visés dans les conférences MUC. Nous présenterons dans la section 4 des travaux menés dans ce sens par notre équipe.

Faut-il pour autant abandonner totalement ce type de tâches ? Nous ne le pensons pas parce qu'il nous semble que l'objectif initial des MUC reste d'actualité pour des applications encore plus nombreuses et que des résultats acceptables peuvent être atteints à court terme sur certaines classes de documents judicieusement choisis pour minimiser la complexité de la tâche. D'autre part nous pensons qu'un travail à caractère plus fondamental est susceptible de conduire à des progrès significatifs. Nous présenterons dans la section 5 des travaux que nous menons en sémantique dans cette perspective.

3. Le projet TACIT.

Nous présentons maintenant le système d'EI développé par notre groupe[6] pour traiter des textes de constats d'accidents automobiles, le portage sur d'autres corpus devant faire l'objet d'une prochaine étape du projet. Le corpus sur lequel nous avons travaillé comprend 87 constats amiables fournis par une compagnie d'assurance. L'objectif est de remplir un formulaire pour chaque impact (*formulaire des événements*) ainsi que des sous-formulaires pour les entités impliquées : véhicules et conducteurs, parties de véhicules touchées, autres objets matériels (*formulaire des entités*). A titre d'exemple, la figure

[4] On pourra consulter les actes des différentes MUC pour suivre l'élaboration de la méthode d'évaluation, et notamment [CHL 93] [GRI95] sur l'adaptation des méthodes en vigueur en recherche documentaire. On pourra constater que cette adaptation n'est pas triviale, et particulièrement délicate en ce qui concerne le calcul de coréférence.

[5] Rappelons qu'en matière d'informatique linguistique, le « zero défaut » n'est pas accessible et qu'il s'agit bien d'obtenir des résultats de qualité *suffisante* pour être utiles dans une tâche donnée.

[6] Projet TACIT (Traitements Automatiques pour la Compréhension d'Informations Textuelles), dont certaines phases ont fait l'objet d'une collaboration avec le LIPN et d'un soutien du programme « Sciences de la Cognition » [VIC 98].

2 présente l'un de ces textes et les formulaires associés. Le système est réalisé en Prolog (à l'exception de l'analyse syntaxique).[7]

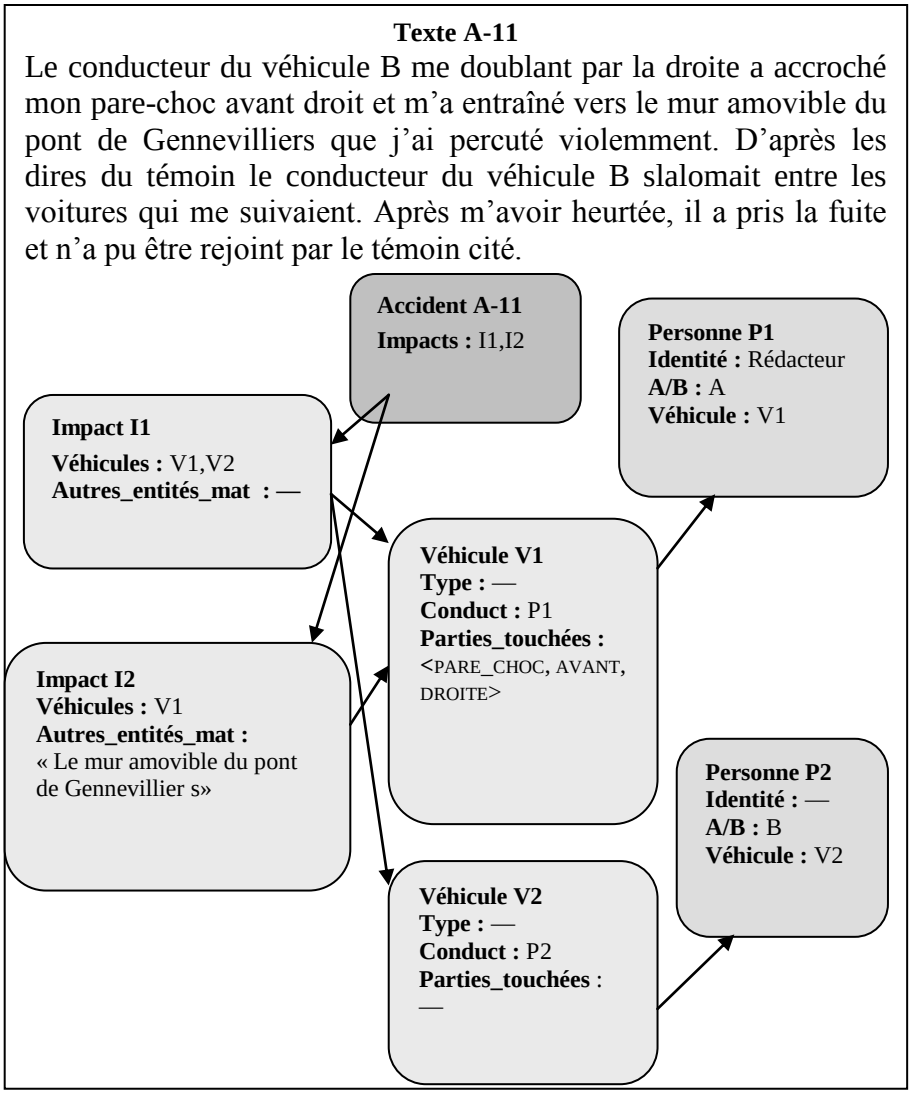


Figure 2 : Constat et formulaires associés

[7] Avec un codage de structures de traits fortement inspiré du format GULP de M. Covington [COV 94].

Des travaux en cours visent à extraire aussi des informations sur le lieu de l'accident et des précisions sur les mouvements des véhicules au moment du choc. Ces compléments nécessitent des traitements sémantiques plus complexes qui seront évoqués dans la dernière partie de l'article.

Nous allons maintenant présenter les méthodes employées dans TACIT. Notez que dans notre projet, la phase de prétraitements est limitée : tous les textes sont pertinents, la correction orthographique a été réalisée « à la main » et le repérage des entités nommées (Mr Dupont, le véhicule A, R21, Genevilliers…) est effectué de manière suffisante par l'analyseur morpho-syntagmatique. Dans le schéma général de la figure 1, nous nous concentrerons donc sur les phases II et III.

L'analyse morphologique et un certain niveau d'analyse syntaxique sont opérées dans la même passe. Puis nous procédons à une analyse sémantique, intégrant certains aspects « simples » du calcul de coréférence. Ce traitement est fortement *orienté* par l'identification des entités et des événements et leur mise en relation.

3.1. Analyse morpho-syntagmatique

Nous utilisons l'analyseur du GREYC, dû à J. Vergne et E. Giguet [VER 94, GIG 97], qui offre aujourd'hui d'excellents résultats adaptés à notre démarche. Cet analyseur met en évidence des segments minimaux (syntagmes « non récursifs » ou *chunks*) avec une très grande fiabilité, ainsi que certains liens reliant ces segments (notamment sujet-verbe, et, dans une moindre mesure, verbe-objet). Une sortie simplifiée de l'analyseur est illustré par la figure 3. C'est cette structure syntaxique « partielle » qui constitue la base de l'analyse sémantique.

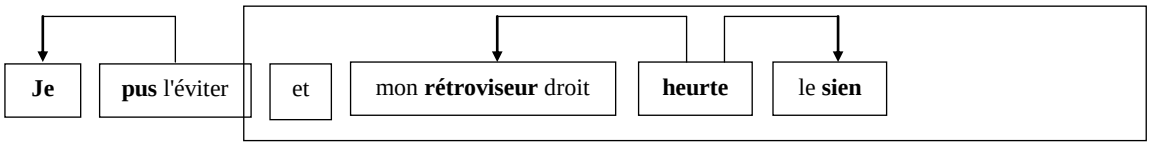


Figure 3 : Segmentation morpho-syntagmatique

Chaque segment est étiqueté (groupe verbal, nominal, etc.) de même que chaque lien entre segments (fonction sujet, objet, etc.). Les mots en gras sont les « têtes » des syntagmes.

3.2 Traitement des entités

L'extraction des entités consiste à repérer les groupes nominaux et pronoms pertinents et à leur associer des entités typées, en utilisant un lexique dans lequel des classes sémantiques sont associées aux unités linguistiques. Les types (véhicule, personne…) sont déterminés par les attributs des formulaires. Nous opérons en trois étapes :

1. *Filtrage* : repérage d'un segment nominal dont la tête est associée à une des classes sémantiques recherchées. Ce calcul est complété par la recherche d'autres termes (comme les possessifs) susceptibles d'introduire de nouvelles entités ;
2. *Mise en relation* : repérage de certains liens syntaxiques qui permettent de relier sémantiquement des entités filtrées ;
3. *Premier calcul local de coréférence*. Dans un premier temps, on crée une entité pour chacune des expressions linguistiques reconnues comme pertinentes. Un premier calcul de coréférence, fondé sur des critères très simples, est mis en œuvre dès cette étape et permet de fusionner certaines entités. Ce calcul sera complété en phase finale lors du remplissage des formulaires (voir *infra*, §3.4).

Illustrons cette phase de traitement à l'aide de l'extrait suivant :

*Le **conducteur** du **véhicule B me** doublant par la droite a accroché **mon pare-chocs avant droit** et **m**'a entraîné vers le **mur** amovible du **pont** de Gennevilliers **que j**'ai percuté violemment.*

En gras figurent les mots introduisant des entités de type 'personne' (*conducteur*, *me*, *j'*), 'véhicule' (*véhicule*), 'partie de véhicule' (*pare-chocs*), 'entité matérielle autre' (*mur*, *pont*), et des entités non typées (*que*). Un lien syntaxique liant *pare-chocs* et *avant droit* permet de repérer la localisation du choc sur

une partie de véhicule. Le possessif *mon* introduit un lien entre une entité de type 'personne' (l'énonciateur) et une entité de type 'partie de véhicule' (le pare-chocs). La résolution des anaphores permet d'identifier les diverses mentions du rédacteur du constat (pronoms et possessifs) et de relier *que* à *mur* ou *pont* (anaphore *partiellement* résolue à cette étape).

3.3 Traitement des événements

Les événements, ici les chocs, peuvent être évoqués de deux manières : par des verbes (*heurter, toucher,* mais aussi *abîmer*…) ou par des noms (*le choc, l'accident*…). Nous ne considérerons ici que le cas des verbes. Il faut donc déterminer :

- Les deux entités ayant subi le choc (véhicules, piétons, murs, etc.), que l'on appelle les *entités impliquées*. On distingue, parmi ces deux entités, celle qu'on appelle *l'acteur* et celle qu'on appelle *le patient* de l'événement : pour un verbe à la voix active, l'acteur du choc est lié plus ou moins directement au sujet grammatical du verbe et le patient au complément.
- les parties de véhicule touchées, lorsqu'elles sont mentionnées ; elles peuvent être exprimées par les compléments d'objet (*j'ai heurté son pare-chocs*) ou par des compléments circonstanciels (*je l'ai heurté à l'arrière*).

On commence par une phase de *filtrage* des groupes verbaux dont le verbe figure dans le lexique des verbes 'exprimant un choc'. À chacun de ces verbes est associé un ensemble d'actants sémantiques [MEL 95] : X (Véhicule) HEURTE Y (Entité) {EN Z (Partie de Véhicule)}, cette dernière information étant optionnelle. C'est cette *structure actancielle* qu'il faut mettre en correspondance avec une structure syntaxique ; par exemple pour une forme verbale en voie active : Sujet - Verbe - Complément Direct - {Complément Circonstanciel}.

Cette analyse utilise différents types de calculs et d'informations. Le fait de disposer, par l'analyse morpho-syntagmatique, d'une relation sujet-verbe très fiable est évidemment précieuse. Par contre la reconnaissance des compléments (directs ou circonstanciels) par des méthodes purement syntaxiques est beaucoup plus problématique, et le typage sémantique des entités, joint au typage des actants sémantiques dans le lexique des verbes est ici exploité. Par ailleurs, il faut noter que les actants ainsi identifiés ne correspondent pas toujours aux deux entités que nous recherchons. Reprenons l'exemple précédent pour illustrer ce point :

Le conducteur du véhicule B me doublant par la droite ***a accroché*** *mon pare-chocs avant droit et m'a entraîné vers le mur amovible du pont de Gennevilliers que j'****ai percuté*** *violemment.*

Deux verbes de choc sont repérés : *a accroché* et *ai percuté*. Pour le premier choc cité, les actants du procès sont deux entités créées au cours de la phase précédente, l'une de type 'personne' (le conducteur) et l'autre de type 'partie de véhicule' (le pare-chocs). Or les entités impliquées que l'on doit extraire sont deux véhicules. De même, l'agent du second procès est l'énonciateur, de type 'personne', alors que c'est son véhicule qui a subi le choc. Il n'y a que le patient du deuxième procès, une 'entité matérielle autre' (obtenue après résolution de l'anaphore qui relie *que* à *le mur*) qui est effectivement une entité impliquée dans le choc. Ces phénomènes de métonymie sont très fréquents dans notre corpus. C'est au cours de la dernière phase (instanciation des formulaires) que ces problèmes sont résolus.

3.4 Instanciation des formulaires (interprétation)

À ce stade, chaque phrase a été analysée isolément et il s'agit maintenant de rassembler ces informations parcellaires en un tout cohérent de manière à remplir les formulaires. C'est ce que nous appellerons ici l'étape *d'interprétation*.

Notre représentation contient encore des entités (véhicule, personne…) qui doivent être fusionnées. Ce problème de résolution de coréférences concerne également les événements puisque plusieurs verbes (ou noms : *le choc*…) peuvent référer au même impact. Ainsi dans *Le véhicule B m'a heurté ... m'abîmant tout l'avant gauche*, les verbes *heurter* et *abîmer* concernent le même choc.

Une autre question concerne l'établissement de liens entre les entités, pour savoir par exemple qui est conducteur de quel véhicule. Il est rare que le rédacteur dise explicitement *Je conduisais le véhicule A, une Peugeot 205*… Un ensemble d'indices linguistiques et de connaissances spécifiques au domaine

permettent alors d'établir ces liens, grâce à quelques raisonnements déclenchés systématiquement. Nous avons pour cela établi un ensemble de *règles*. Certaines mettent en œuvre des connaissances sur le type de texte traité : en l'absence d'indication contraire explicite, on peut supposer que le texte parle d'un accident, survenu entre deux véhicules, nommés A et B, dont l'un est conduit par le locuteur. D'autres traitent les phénomènes de métonymie évoqués précédemment. En voici quelques unes, données de manière informelle :

- **Si** il y a un lien entre l'actant d'un procès 'exprimant un choc' et une entité X de type 'véhicule', **alors** X est une entité impliquée du choc correspondant.
- **Si** l'actant d'un procès 'exprimant un choc' est de type 'personne' et que cette personne est à bord d'une entité X de type 'véhicule', **alors** X est une entité impliquée du choc correspondant.

A titre d'exemple, voici le résultat obtenu sur le texte suivant :

Etant arrêté momentanément sur la file de droite du Boulevard des Italiens j'avais mis mon clignotant j'étais à l'arrêt et m'apprêtant à changer de file. Le véhicule B arrivant sur ma gauche m'a serré de trop près et m'a abîmé tout le côté avant gauche.

L'analyse des entités et des événements donne les informations suivantes :

P1 : entité de type 'personne' (chaîne de référence : *j', mon, j', m', ma, m'*)
V1 : entité de type 'véhicule' (*le véhicule B*)
C1 : procès de type 'exprimant un choc' (*a abîmé*)
PV1 : entité de type 'partie de véhicule' (*côté avant gauche*)
V1 agent de C1. Spécification de V1 = véhicule B
P1 patient de C1. Spécification de P1 = énonciateur
PV1 partie de véhicule. Spécification de PV1 = partie touchée dans C1, liée à P1.

Les règles vont permettre de compléter ces informations de la manière suivante :

- « La personne acteur d'un procès 'choc' est à bord d'un véhicule » ; d'où la création d'une entité V2 de type 'véhicule', d'un lien 'conducteur' entre P1 et V2 ; or PV1 est liée à P1, V2 est donc le patient de C1 et PV1 est rattachée à V2 comme 'partie touchée'.
- « Il y a conducteur à bord d'un véhicule qui n'est pas explicitement déclaré à l'arrêt » ; d'où la création d'une entité P2 de type 'personne', liée à V1 comme 'conducteur' ;
- Enfin « dans un constat les deux véhicules impliqués sont nommés A et B » : d'où la spécification de V2 comme 'véhicule A'.

Il est remarquable que quelques règles simples donnent ici de très bons résultats pour un problème *a priori* complexe. Il s'agit là évidemment d'une conséquence de contraintes fortes sur le domaine de connaissances et les textes, et il serait intéressant de savoir si d'autres corpus présentent une situation similaire.

3.5 Discussion

Notons d'abord que les résultats obtenus sont tout à fait encourageants si on les compare à ceux des MUC présentés plus haut : précision : 84 %, rappel : 73 %, F-mesure : 78 %. Mais précisons qu'il ne s'agit là que d'une simple indication, l'évaluation n'ayant évidemment pas été conduite selon les règles d'objectivité qui prévalent dans une « compétition » telles que MUC. En outre la comparaison des performances devrait prendre en compte des facteurs de complexité de la tâche difficiles à évaluer [BAG 99]. Examinons maintenant quelques questions méthodologiques.

Analyse syntaxique

Nous opérons une analyse très partielle, reflétant en cela une démarche assez répandue dans la communauté de l'EI, connue sous le terme « d'analyse syntaxique légère » (*shallow parsing*) [COW

96] : voir par exemple les systèmes FASTUS [HOB 96] ou PROTEUS [GRI 97, YAN 99].[8] A l'inverse un système comme LaSIE [GAI 95, 99] reprend la démarche classique en compréhension automatique d'une analyse syntaxique autonome préalable la plus complète possible [ALL 95]. Notre choix est d'abord guidé par des considérants théoriques et une observation qui, pensons-nous, montrent l'impossibilité de parvenir à une analyse syntaxique complète sans information sémantique [POI 98]. Mais également par la remarque que seules certaines parties du texte, susceptibles de porter des informations pertinentes, sont à analyser « en profondeur ».

Analyse sémantique et interprétation

Tout système d'EI doit, d'une manière ou d'une autre, combiner deux types d'informations et de traitements :

- linguistiques, liés à « la langue » en général, donc génériques ;
- liés au domaine et à la tâche (remplissage de formulaires) et donc spécifiques.

Notre démarche peut être caractérisée par deux principes :

- ne pas hésiter à intégrer les deux types de traitements,
- découper éventuellement une tâche en plusieurs phases en fonction de l'information disponible pour obtenir des résultats sûrs.

Le calcul de la référence constitue une bonne illustration. Comme nous l'avons vu, il opère en deux passes et, pour une part importante, se trouve rejeté dans la phase III , en liaison étroite avec le remplissage des formulaires — dont la structure condense une information très « contrainte» et précieuse sur le contenu des textes. Cette organisation diffère du schéma séquentiel « standard » en compréhension automatique, illustré par la figure 2, et repris par un système comme LaSIE.

Sur le plan théorique, cette démarche s'appuie sur la notion de *langue de spécialité* [HAR 91]— dans une acception, il est vrai, quelque peu élargie — c'est-à-dire de régularités fortes liées à un domaine, tant en ce qui concerne la structure de l'information que son expression. Plutôt que des *outils linguistiques universels* nous recherchons donc des méthodes linguistiques générales *adaptables* ou *instanciables* dans un domaine particulier. Sans préjudice pour des techniques d'un plus grand degré de généralité, susceptibles de produire en amont certaines analyses partielles : voire l'idée déjà discutée d'une analyse syntaxique légère ou certains traitements de l'anaphore présentés dans la section 5.

4. Autres applications

4.1 L'EI dans le champ de l'informatique documentaire

Reprenons la question posée en fin de la section 2, en bilan de MUC, invitant à considérer avec recul la place de l'EI (structurée) dans la chaîne documentaire. Parmi diverses orientations possibles, l'idée développée ici (suivant [WIL 97]) est d'un décloisonnement entre la *recherche documentaire*, qui vise traditionnellement à sélectionner des items d'une base documentaire, et l'EI qui peut ensuite en fournir une analyse « en profondeur ».

En premier lieu, les techniques de l'EI peuvent parfaitement être mises à contribution dès la phase de recherche documentaire. Ainsi, dans le système FACILE [CIR 99] l'utilisateur peut formuler des requêtes telles que *Envoyez-moi les textes concernant des contrats passés par des institutions financières européennes dont le montant dépasse un million d'Euros.* Au lieu de la traditionnelle recherche thématique, dirigée d'une manière ou d'une autre par un ensemble de mots-clés, la requête porte ici sur le repérage d'une information structurée, ce qui réclame d'utiliser des techniques d'EI.

En second lieu, l'EI permet une recherche d'information *au sein même du texte*. Une bonne illustration en est donnée par l'idée d'un système d'*aide à la lecture* de documents volumineux. Imaginons par exemple un médecin recherchant dans des comptes-rendus d'hospitalisation des informations sur *les patients ayant présenté le symptôme S à la suite de l'intervention I (dans un délai D)*. Sans doute hésitera-t-il à se fier à des fiches produites par un système totalement automatisé de type

[8] En remarquant que la technique utilisée par ces auteurs (transducteurs à état finis) diffère de celle qui est mise en œuvre dans l'analyseur du GREYC (propagation de contraintes).

MUC. Mais un système d'EI pourrait analyser les comptes-rendus et simplement repérer les passages correspondant à la requête de manière à les présenter en sur-brillance, aidant ainsi le praticien à prendre connaissance rapidement des informations pertinentes. Des systèmes de ce type font l'objet de recherche en informatique médicale [ZWE 97]. Notre équipe travaille sur une application assez voisine, concernant des documents de géographie humaine, intéressant par exemple des « décideurs » chargés d'administrer des territoires. C'est ce que nous allons maintenant détailler.

4.2 Encodage sémantique de documents géographiques

Objectifs et corpus

Le corpus considéré dans ce projet[9] est constitué d'atlas géographiques, c'est-à-dire de documents de taille importante, intégrant à la fois du texte et des planches graphiques (cartes et autres représentations de données statistiques). Ce sont soit des ouvrages imprimés numérisés (tel que *l'Atlas de la France scolaire* de Robert Herin [HER 94]), soit des hyperdocuments électroniques rédigés de manière coopérative (tel que l'Atlas Trans-Manche [TUR 99]). L'analyse montre que l'information y est consignée sous une forme très « canonique » présentant un *phénomène*, dont les *valeurs ou variations* sont quantifiées, relativement à un ensemble de *zones géographiques* et de *périodes temporelles*. En voici un exemple typique [HER 94]:

Jusqu'au milieu des années 1980, les taux de retard scolaire ont fortement varié selon les configurations géographiques… Ainsi dans l'Aveyron, à Paris ou dans les Pyrénées-Atlantiques, seulement un enfant de 6° sur trois est en retard scolaire…

De tels documents constituent en quelque sorte des « réserves d'information » que le lecteur souhaitera consulter de manière sélective, plutôt que par un parcours exhaustif et linéaire. Comment l'aider à prendre connaissance des informations pertinentes ? La réponse classique consisterait à effectuer une analyse thématique, qui pourrait porter ici sur les notions de *retard*, de *taux de scolarisation*, différentiées selon les *niveaux scolaires*, etc. Mais on peut avoir une autre approche. Connaissant la thématique générale de l'ouvrage, le lecteur peut être intéressé par des données concernant certaines zones géographiques (tels départements ou régions) et/ou certaines périodes temporelles.

Le but du projet sera donc de repérer dans le texte les expressions spatiales (*au nord de Paris*, *dans l'est du Massif Central*, etc.) et temporelles (*en 1980*, *dans les années 1980*, etc.) et d'en faire une analyse sémantique produisant une représentation formelle. Cette représentation est associée aux expressions sélectionnées par un encodage XML. On obtient ainsi un texte enrichi permettant une sélection rapide de passages au gré du lecteur, sur critères spatiaux ou temporels[10]. Noter que l'accès « par la localisation » est totalement pertinent pour des ouvrages « géographiques », dont une caractéristique est de porter des informations *géoréférencées*, c'est à dire intrinsèquement liées à des territoires.

Une autre application de cette analyse concerne la structuration même du document, en tant qu' « hybride » intégrant texte et illustrations. Nous visons en effet à établir des liens de type hypertextuel entre le corps du texte et les planches graphiques (cartes, schémas), grâce à une interprétation conjointe de ces dernières portant sur leur légende (texte) mais aussi sur les cartes et graphiques eux-mêmes en tant qu'image [MAL 99]. On peut ainsi établir des liens que nous appelons « globaux ») entre sections textuelles et planches[11], mais aussi des liens de granularité plus fine (dits « minimaux ») entre des passages de l'ordre du syntagme (« unités d'information textuelles ») et des

[9] Mené en collaboration avec le groupe « InfoDoc » du GREYC, particulièrement M. Gaio, et M. ould Ahmed Liman.

[10] Sans préjudice pour un croisement des deux méthodes, permettant de chercher les passages concernant, par exemple, « le retard scolaire dans l'Ouest depuis 1990 ».

[11] On pourrait penser que la référence aux planches est totalement explicite dans le texte, par des indications du type « voir figure tant ». Or ce n'est pas la pratique courante des géographes qui laissent au lecteur le soin d'établir ces liens, dans une certaine zone de proximité de quelques pages avant ou après le passage textuel concerné. Noter également que la numérisation des ouvrages imprimés rompt cette structure « spatiale » de l'ouvrage ce qui complique d'autant la tâche du lecteur.

composantes de l'illustration (« unités d'information graphiques ou cartographiques »). La figure 4 illustre les liens que nous souhaitons établir.

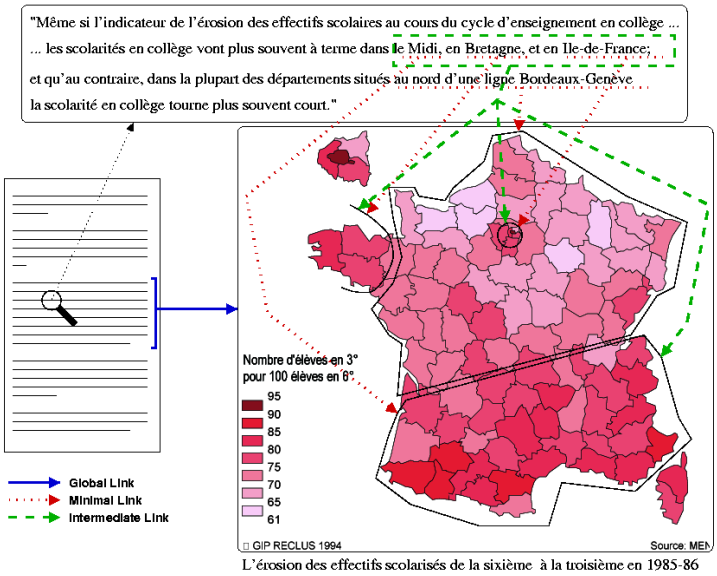


Figure 4 : Carte et texte associé

Méthodes

Nous ne présenterons ici que les principes des traitements linguistiques, en montrant comment ceux-ci se rapprochent des méthodes, présentées *supra*, utilisées en EI ; et uniquement l'analyse spatiale, l'analyse temporelle se révélant similaire et sensiblement plus simple[12].

Une étude de corpus montre que la localisation est exprimée par des groupes nominaux ou prépositionnels, que l'on peut répartir en classes de complexité (syntaxique et sémantique) croissante selon l'entité géoréférencée (EG) dénotée :

- EG directe et précise, exprimées notamment par une entité nommée (*en Bretagne*, *dans le Massif Central*) ;
- EG indirectes définies par une orientation en référence à une EG directe, explicite (*au nord / dans le nord du Massif Central*.) ou non (*dans le nord*).
- D'autres constructions permettent de définir des EG par une forme géométrique (*une ligne Bordeaux-Genève*)

Une représentation formelle de type attributs-valeurs peut être définie pour coder ces différents types d'EG, et représentée en XML (cf. figure 5). Le traitement linguistique se décompose alors comme suit.

- Recherche de mots « déclencheurs » indiquant la présence d'une expression spatiale. Il peut s'agir de mots « pleins » (*département*, *région*, *nord*…) ou de prépositions à valeur spatiale (*dans*, *à*…) ;
- Analyse syntaxique permettant de reconstruire, autour de ces termes, des *chunks* et des syntagmes complexes. Cette opération peut être réalisée par des méthodes traditionnelles (grammaires), soit à partir du texte brut, soit (et c'est pour nous la voie d'avenir) en partant du texte traité par un analyseur morpho-syntagmatique du type de celui de J. Vergne qui donne le découpage en chunks et diverses informations morphologiques.
- Construction de la structure sémantique. Il est remarquable qu'ici une démarche compositionnelle fonctionne fort bien[13]. (cf. Figure 5).

[12] Elle permet d'associer à une expression temporelle une période (approximative) représentée par un certain intervalle de dates, par exemple [1980,1985] pour « le début des années 1980 ».

[13] Ceci est lié à la très forte structure syntactique et sémantique, et au caractère « local » des expressions spatiales. Contraster cette situation avec celle qui prévaut au niveau de la proposition et de la phrase [POI 98]. Il s'agit-là d'un bon exemple des simplifications que permet une analyse *limitée* de type EI.

- Interprétation. L'établissement d'un hyper-lien avec une carte suppose une interprétation en quelque sorte plus « concrète » relativement à la carte et aux données (de type SIG) qu'elle représente. Nous renvoyons à [MAL 99, MAL 00] pour plus de détails sur ce traitement.

4.3 Conclusion

La tâche qui vient d'être présentée a bien les caractéristiques de l'EI : il s'agit de compréhension automatique, mais orientée par une tâche précise et limitée produisant une information formalisable, et reposant sur une analyse linguistique locale. Un certain nombre d'outils génériques peuvent être utilisés : repérage des entités nommées, analyse morphosyntagmatique « légère », etc. Bien évidemment, elle en constitue une *variation*, certaines étapes devant être menées de manière spécifique : il en va ainsi notablement de l'analyse sémantique compositionnelle qui repose sur des caractéristiques spécifiques des groupes prépositionnels spatiaux.

Enfin, il est à noter que la tâche est ici plus simple que celle de MUC, du fait d'une plus grande simplicité des expressions linguistiques (pas de niveau « prédicatif») et d'une meilleure tolérance aux erreurs (l'aide à la lecture est moins critique que l'alimentation d'une base de données puisque le document demeure disponible pour l'utilisateur). De ce fait ce travail illustre bien, pensons-nous, les potentialités ouvertes par une adaptation adéquate des techniques de l'EI, avec des résultats opérationnels à court terme.

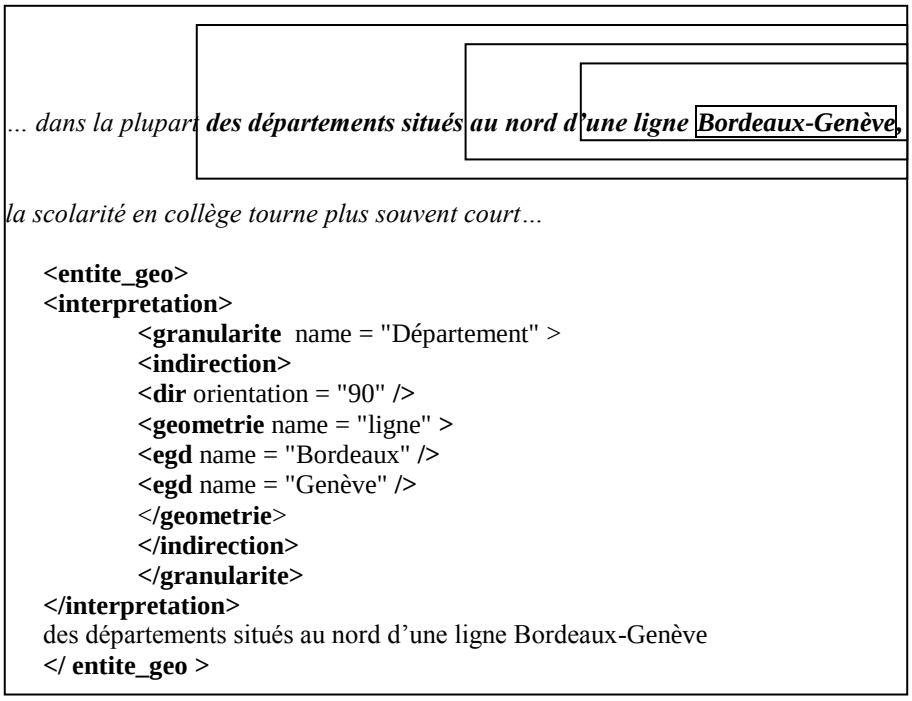


Figure 5 : expression spatiale et encodage XML

5. Traitements sémantiques approfondis

Nous venons de présenter quelques applications possibles à relativement court terme, avec un degré de fiabilité acceptable, sur la base des techniques actuelles. A plus long terme, nous pensons que des progrès notables peuvent être réalisés, permettant de traiter des tâches plus complexes, à partir d'un travail linguistique plus approfondi. Nous pensons notamment à un travail en sémantique, exploitant — en les adaptant et, à l'occasion, en les développant — un certain nombre de théories linguistiques qui font l'objet de recherches récentes ou en cours. Notre équipe est fortement engagée dans cette démarche que nous illustrerons par deux exemples ; le traitement de la référence d'une part, de la spatialité de l'autre.

5.1 Calcul des chaînes de coréférence

Les conférences MUC l'ont montré (cf. section 2), le calcul de coréférence constitue un problème clé pour la qualité des systèmes d'EI. Pour chaque occurrence de groupe nominal ou de pronom, il faut déterminer si cette forme linguistique réfère à une entité en excluant notamment les pronoms impersonnels (comme *il* dans *il pleuvait*, les nominalisations d'adjectifs (*l'ampleur* dans *l'ampleur des dégâts*). Il faut ensuite déterminer à quelle *chaîne de coréférence* appartient cette forme linguistique en initialisant au besoin une nouvelle chaîne. Notons que la reprise d'une entité (une voiture, par exemple) peut se faire de différentes manières : par un pronom (*elle, celle-ci* etc. : on parle alors d'anaphore pronominale) ou par un groupe nominal (*cette voiture, la Peugeot 205,* etc. : anaphore nominale).

Pour traiter ce problème de manière efficace et sûre, nous avons élaboré une méthode, inspirée pour une part des travaux d'Alshawi en compréhension automatique [ALS 97] et de la théorie psycholinguistique de l'accessibilité de Mira Ariel [ARI 90], mais originale par nombre de ses aspects [DUP 96a, 96b, 97]. Nous allons en résumer ici les caractéristiques essentielles.

A l'instar d'Alshawi qui préconise la gestion d'un « Context Model », nous proposons de gérer un « modèle des attentes » qui synthétise les anticipations que le lecteur fait sur la suite du texte. Ce modèle des attentes contient une liste d'entités avec pour chacune d'elles un attribut numérique, la « saillance ». La saillance est d'autant plus grande que l'entité est présente à l'esprit du lecteur et qu'en conséquence il s'attend à rencontrer une réalisation de cette entité dans la suite du texte. Une entité de forte saillance se trouve en quelque sorte en avant-plan sur la scène évoquée par le texte alors qu'une entité de faible saillance se trouve plus en arrière-plan, moins focalisée par le lecteur.

Le modèle des attentes est initialisé à partir de la nature particulière du document : au début de la lecture d'un texte, on s'attend à ce que l'on parle de certaines entités, que l'on appelle les entités pré-construites (par exemple dans un constat d'accident, on s'attend à ce que le locuteur soit le conducteur d'un véhicule, qu'il soit question d'un autre véhicule, d'un accident...). Après cette initialisation, ce sont les marques linguistiques qui vont, au fur et à mesure du déroulement du texte, confirmer ou infirmer ces éléments présents par défaut, changer leur saillance, introduire de nouvelles entités, etc. Pour chaque phrase, le modèle des attentes est mis à jour et ses données sont utilisées pour le traitement de la phrase suivante.

Voyons quels sont les facteurs à prendre en considération pour mettre à jour les saillances. Nous considérons que deux seuils permettent d'établir trois classes: faible, moyenne et forte saillance. Toute mention d'une entité lui confère une forte saillance. Les entités liées de manière prototypique à des entités fortement saillantes (ainsi, pour poursuivre notre exemple, les principaux organes ou parties d'un véhicule) sont *ipso facto* moyennement saillantes même si elles n'ont pas été elles-mêmes mentionnées. La saillance des entités se dégrade lorsqu'elles ne sont plus évoquées et, à l'inverse, se trouve renforcée par toute nouvelle réalisation linguistique. Dégradation et renforcement prennent en compte divers facteurs : proximité de l'occurrence précédente (autrement dit « récence » de la dernière évocation) ; marques linguistiques de focalisation; « profondeur syntaxique », fonction grammaticale de l'occurrence dans sa proposition, etc.

Cette notion de saillance joue un rôle important dans la résolution des anaphores. En effet, on peut énoncer un « principe de concordance » qui relie la saillance de l'entité et les marques linguistiques susceptibles de l'évoquer à nouveau. Ainsi un pronom personnel (comme *elle*) ne peut reprendre qu'une entité fortement saillante. A l'opposé, un groupe nominal indéfini comme « *une Renault 21* » ne peut servir qu'à introduire une nouvelle entité (faible saillance). Entre les deux, un groupe nominal défini court comme « *la Peugeot 205* » reprend une entité de saillance forte ou moyenne. Un groupe nominal défini long comme « *la voiture me précédant* » peut servir à introduire une nouvelle entité (faible saillance) ou à reprendre une entité moyennement saillante. On associe donc à chaque marque linguistique une « plage des saillances admissibles », qui dépend en premier lieu de la nature de cette marque linguistique.

La saillance n'est cependant pas le seul facteur à prendre en compte. D'une manière générale, la résolution des anaphores se présente comme un problème de résolution de contraintes. Certaines

marques linguistiques imposent des conditions très spécifiques. Ainsi, pour *ce dernier*, le critère de concordance s'efface devant celui de récence, *je* indique le locuteur sans que d'autres critères interviennent, etc. D'autres facteurs plus généraux peuvent intervenir :

- compatibilité des accords en genre nombre et personne ;
- synonymie ou hyperonymie (par exemple *le camion* repris par *le véhicule*) ;
- catégories sémantiques des arguments d'un prédicat (par exemple dans « *il m'a dit que...* » le pronom « *il* » réfère à un humain) ;
- contraintes syntaxiques imposant dans certains cas de rechercher l'antécédent d'un pronom dans un segment bien délimité de la phrase, ou au contraire excluant ce même segment (« règles du liage ») ;
- cohérence avec les connaissances du domaine.

Des contraintes de préférence peuvent intervenir lorsque plusieurs candidats restent en compétition. Ainsi une solution qui maintient la même fonction syntaxique entre les réalisations successives d'une entité est préférable à une solution où la fonction syntaxique change. De même, la position anaphorique (où l'antécédent précède le pronom) est préférée à une position cataphorique (où le pronom précède l'antécédent). Un calcul de score ou un système de priorité sur des règles peuvent être implantés pour exprimer ces critères de préférence.

Comment une telle théorie peut-elle être prise en compte dans un système d'EI ? Peut-elle donner lieu à des traitements suffisamment « légers » ? Pour quel gain ?

Nous remarquerons d'abord que la plupart des facteurs analysés ci-dessus sont totalement indépendants du domaine, et peuvent donc faire l'objet de traitements totalement génériques, voire de modules spécifiques intégrés très en amont du processus d'EI. De fait, nous avons montré que l'on pouvait, en se restreignant aux critères indépendants du domaine, obtenir des résultats très encourageants sur des textes tout-venant [DUP 97]. D'autre part il est possible d'exploiter certaines informations et connaissances spécifiques fournies au système pour d'autres raisons : par exemple une taxonomie du domaine et le typage du lexique associé (synonymie, liens d'hypo- et hypéronymie) pourront permettre de prendre en compte des contraintes de concordance sémantique.

Enfin, si l'analyse syntaxique *légère* ne permet pas de prendre pleinement en compte les contraintes de liage, elle fournira néanmoins des informations exploitables et précieuses. Ainsi, avec un découpage en propositions on peut exploiter le fait qu'un pronom non réfléchi ou réciproque ne peut pas être coréférentiel avec le sujet dans une même proposition, et qu'un pronom situé dans une proposition principale ne peut pas être coréférentiel avec un groupe nominal qui se trouve dans une subordonnée postposée. Si on ne dispose pas de ce niveau d'analyse syntaxique on peut utiliser des contraintes moins fortes, et dire par exemple qu'un pronom qui se trouve en tête d'une phrase ne peut pas, en général, avoir son antécédent dans la phrase. Remarquons également que ces traitements peuvent se faire en différentes étapes, comme nous l'avons vu dans la section 3.

Ainsi l'utilisation, même partielle, d'une théorie linguistique « solide » de l'anaphore peut être intégrée dans des traitements d'EI et doit permettre d'améliorer sensiblement les performances sur la tâche clé du calcul des chaînes de coréférence.

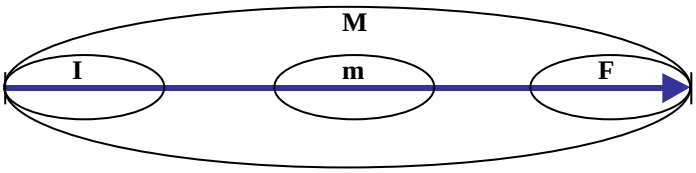
5.2. Traitements sémantiques pour l'extraction d'informations spatiales

Nous abordons à présent une étude qui montre l'intérêt de recourir à des traitements sémantiques approfondis pour l'extraction d'informations non explicitement formulées. Exposons sommairement le problème : en reprenant les textes de constats dont nous avons fait état en section 3, nous nous fixons l'objectif de déterminer le lieu géographique (*Paris, la RN13*) et/ou le type de lieu (*en ville, à un carrefour*) où s'est produit l'accident. La difficulté provient du fait que les lieux en question ne sont pas explicitement corrélés à la description de l'impact. Nous ne trouvons jamais dans le corpus de mention directe telle que : *l'accident s'est produit au carrefour* ou *à la sortie de Mours* etc ; Nous voyons plutôt apparaître une série de 'lieux' généralement issus de l'expression de déplacements (eg. *me rendant à Beaumont-sur-Oise depuis Cergy…*), parmi lesquels figurent diverses caractérisations du lieu de l'impact (cf. texte A1 ci-dessous). Décrivons le traitement sémantique proposé pour réaliser cette tâche.

Texte A1. Me rendant à **Beaumont sur Oise** depuis **Cergy**, je me suis retrouvée à **un carrefour** juste après **la sortie Beaumont sur Oise**. J'étais à **un stop** avec 2 voitures devant moi tournant à droite vers **Mours**. Alors que la première voiture passait **ce stop** je fis mon contrôle à gauche et je démarrais mais je percutais la deuxième voiture qui n'avait pas encore passé **le stop**.

Le premier problème consiste à détecter dans le texte les expressions susceptibles d'exprimer la localisation d'une phase du récit de l'accident. Cette tâche est *a priori* complexe, car il n'est pas possible « en général » d'attribuer de manière certaine l'étiquette 'lieu' à des items lexicaux (on a bien affaire à un lieu dans *traverser un carrefour* mais pas dans *rénover un carrefour*). Toutefois, le fait de se focaliser sur un ensemble de textes homogène permet de le faire avec un très grand degré de confiance : on parvient à 100% de pertinence dans notre corpus où toutes les entités urbaines ou routières mentionnées le sont en tant qu'elles permettent de renseigner spatialement l'action.

Il faut ensuite établir un lien entre les 'lieux ' ainsi repérés et le déroulement de l'action. Nous nous appuyons pour cela sur les travaux linguistiques de [LAU 91] qui fournissent un traitement compositionnel très simple du complexe *verbe de déplacement* / *préposition spatiale* dans la phrase simple, c'est-à-dire dans les phrases ne faisant intervenir qu'un seul complément. Le principe consiste à attribuer à tout verbe une *catégorie aspectuelle spatiale* qui peut être *initiale*, *médiane* ou *finale*, visant à caractériser le fait qu'un verbe dénote intrinsèquement respectivement un début de déplacement (*quitter*), son milieu (*passer*), ou sa fin (*aller*). Il en va de même pour les prépositions (respectivement : *de*, *par*, *vers*), à ceci près qu'une quatrième catégorie regroupe les prépositions, dites *positionnelles*, n'exprimant rien de tel *a priori* (*sur*, *dans*, *à*). Toute combinaison donne lieu, par un calcul simple, à une valeur sémantique globale du complexe *verbe / complément prépositionnel* caractérisé comme initial, médian ou final ; par exemple, *se rendre à*, qui est *final / positionnel* donne lieu à une interprétation *finale* (**F**) de sorte que *me rendant à Beaumont-sur-Oise* indique que Beaumont est le lieu final du déplacement exprimé ; de même, *se rendre […] depuis*, qui est *final / initial* donne lieu à une interprétation *initiale* (**I**) et *me rendant […] depuis Cergy* indique donc que *Cergy* est le lieu initial. Notons, sans détailler, que nous avons quelque peu aménagé la théorie de Laur pour autoriser le traitement des compléments multiples, et pour affiner le traitement de l'interprétation médiane (nous distinguons entre la catégorie **m** correspondant par exemple à *passer par* et **M** correspondant à *circuler dans*). Voici ci-dessous une illustration des parties de procès auxquelles réfèrent ces catégories :



Nous doublons ce premier traitement d'un calcul simplifié de l'aspectuo-temporalité, afin de déterminer si le procès dénoté est présenté comme accompli ou inaccompli. Si en toute rigueur un calcul complexe est nécessaire [GOS 96] une attribution de cette valeur par la seule considération du temps du verbe conjugué donne déjà un premier niveau de résultats intéressant.

Il reste à croiser les deux informations sémantiques pour savoir si un lieu est *passé*, c'est-à-dire qu'au moment de référence verbal, le sujet l'a déjà quitté, *présent*, dans le cas où le sujet s'y trouve au moment de référence, ou *futur*, c'est-à-dire non encore atteint. On obtient le tableau de combinaisons suivant :

Aspect temporel	Aspect spatial	Temporalité du lieu
Non-Accompli	I M m F	Passé Présent Indéterminé Futur
Accompli	I, M, m F	Passé Présent

Les lieux marqués "présent" peuvent être considérés comme décrivant effectivement le lieu de l'accident. [14]

A titre d'exemple, examinons le traitement du texte A1 de notre corpus, repris ci-dessous. Nous nous trouvons, après la première étape de filtrage lexical, en présence d'un certain nombre de 'lieux' fournis par le texte : *Beaumont-sur-Oise, Cergy, carrefour, stop, Mours*. Il est intéressant de constater que la première phrase du texte (*Me rendant à Beaumont-sur-Oise depuis Cergy*) introduit deux lieux qui ne sont pas pertinents quant à la localisation de l'accident. En effet, elle exprime que *Beaumont-sur-Oise* est F (final), que *Cergy* est I (initial), et se trouvent dans un procès Non-Accompli. De la sorte, les deux premiers lieux introduits, *Beaumont* et *Cergy*, sont respectivement Futur (Non-Accompli + F) et Passé (Non- Accompli + I), donc bien rejetés par le système. Il en est de même pour *Mours*, au contraire de *stop* et *carrefour* qui sont bien retenus. Le traitement est donc pertinent sur ce texte.

La mise en œuvre comporte pour l'essentiel deux aspects : le lexique et les règles de combinaison lexicale. Le lexique nominal consiste à donner aux éléments concernés la catégorie 'lieu', celui des prépositions et des verbes consiste à renseigner la 'catégorie aspectuelle spatiale' (I, M ou F) ; quant aux règles compositionnelles, elles sont, à quelques aménagements près, celles de Laur.

L'expérience a porté sur un nombre trop réduit de textes pour permettre une évaluation probante, mais nous pouvons néanmoins porter un jugement qualitatif sur les sorties fournies par le système. Le bruit apparaît faible : sur 17 textes d'essai, seul un cas est litigieux (cas qui reste ambigu même « manuellement »). Le rappel est par contre encore médiocre, mais cela semble dû pour beaucoup à une analyse encore insuffisante des groupes nominaux complexes (par exemple le système ne donne que *mur* au lieu de *mur du pont de Gennevilliers*), plus qu'au principe d'extraction proposé.

5. Conclusion

Ainsi, de nouvelles perspectives s'ouvrent dans la recherche en EI. Après une première phase de développement, qui a imposé ce champ de recherches en lui donnant des bases méthodologiques solides, l'heure est à l'approfondissement des techniques et à l'extension de leur champ d'application.

Il faut souligner tout l'intérêt de ce domaine de recherche. En premier lieu, les retombées industrielles dans le domaine de l'informatique documentaire, sous les formes variées, sont probablement proches. Mais son impact pour la recherche fondamentale ne doit pas être sous-estimé. La méthodologie utilisée constitue un grand pas en avant pour tous ceux qui s'intéressent aux traitements automatiques des textes. L'utilisation systématique de corpus, la définition de tâches précises et la mise au point de méthodes chiffrées d'évaluation comptent certainement pour beaucoup dans la nouvelle dynamique de l'ingénierie linguistique, et il est patent que l'expérience MUC a fortement contribué à cette évolution. Bien des idées théoriques issues des travaux en compréhension automatique peuvent trouver là un terrain expérimental solide, indispensable à leur validation ; inversement la tâche définie par l'EI peut constituer un stimulant précieux pour envisager de nouvelles orientations théoriques et de nouvelles méthodes de traitement.

6. Bibliographie

[ALL 95] ALLEN, J., *Natural Language Understanding*, 2nd edition, Benjamin/Cummings, 1995.

[ALS 87] ALSHAWI H., *Memory and context for language interpretation*, Cambridge University Press, 1987

[ARI 90] ARIEL M., *Accessing Noun Phrases Antecedents*, London, Routledge, 1990.

[BAG 99] BAGGA A., « Analysing the Complexity of a Domain With Respect To An Information Extraction Task », in [MUC 99].

[14] D'autres heuristiques plus subtiles peuvent être envisagées.

[BOR 67] BORKOWSKI C., « An experimental system for automatic identification of personal names and personal titles in newspaper texts », *American Documentation*, 18(3), p. 131–138, 1967.

[CHL 93] CHINCHOR N., HIRSCHMANN L., LEWIS D., « Evaluating Message Understanding Systemes : An Analysis of the Third Message Understanding Conference MUC-3 », *Computational Linguistics*, 19(3) p. 409-449.

[CIR 99], CIRAVEGNA, F. *et al*., « FACILE : Classifying Texts Integrating Pattern matching and Information Extraction », *Proceedings of. IJCAI'99*, p. 890-895, 1999.

[COV 94] COVINGTON, M.A., *Natural language Processing for Prolog Programmers*, Prentice Hall, 1994.

[COW 96] COWLE,J., LENHNERT, W., Information extraction, *Comm. of the ACM*, 39(1), p. 80-91, 1996.

[DUP 96a] DUPONT M., « Le Modèle des Attentes du Lecteur », *Actes de ILN'96*, Nantes, 1996, p. 219-229.

[DUP 96b] DUPONT M., « Le Modèle des Attentes du Lecteur dans le calcul de la référence », *Actes de RECITAL'96*, Courcelles, 1996, p. 155-160.

[DUP 97] DUPONT M., CLAVIER S., « Calcul des Chaînes de Référence dans des Textes Tout Venant », *Actes de JST'96*, 1997, p.345-351.

[FER 00] Ferrari, S. et al., « Projet LINGUIX, recherche d'informations et traitements linguistiques: le cas des métaphores », Conférence internationale sur le document électronique (CIDE 2000), Lyon, 2000 .

[GAI 95] GAIZAUSKAS R. et *al*, « Description of the LaSIE System as Used for MUC-6 », in [MUC 95].

[GAI 99] GAIZAUSKAS R. et *al*, « Description of the LaSIE System as Used for MUC-7 », in [MUC 99].

[GIG 97] GIGUET E., VERGNE J., « From part of speech tagging to memory-based deep syntactic analysis », *Proceedings of International Workshop of Parsing Technologies*, Massachussetts, MIT, 1997.

[GOS 96] GOSSELIN L., *Sémantique de la temporalité en français*, Louvain-la-Neuve, Duculot, 1996.

[GRI 95] GRISHMAN R., B. SUNDHEIM, « Design of the MUC-6 Evaluation », in [MUC 95].

[GRI 97] GRISHMAN R., « Information extraction : Techniques and challenges », in Pazienza M.T. (éd.), *Information Extraction*, Springer Verlag, p. 10–27, 1997.

[HER 94] HERIN R. *et al*, « Atlas de la France scolaire de la maternelle au lycée », La documentation française, Reclus, 1994.

[HOB 96] HOBBS J. R. et *al*, « FASTUS: Extracting Information from Natural-Language Texts », http://www.ai.sri.com/~appelt/fastus.html, 1996.

[LAU 91] LAUR D., « Sémantique du déplacement et de la localisation en français : une étude des verbes, des prépositions et de leur relation dans la phrase simple », Thèse de doctorat, Université de Toulouse le Mirail. 1991

[MAL 99] MALANDAIN N., GAIO M., « Extraction d'unités d'informations géographiques dans des documents composites », *Actes de la Conférence CIDE'99*, Damas, 1999.

[MAL 00] MALANDAIN N., « La Relation Texte/Image. Essai de Modélisation dans un Corpus Géographique », Doctorat de L'Université de Caen – Basse-Normandie 2000.

[MEL 95] MEL'CUK I, CLAS A, POLGUERE A, « Introduction à la lexicologie explicative et combinatoire », Éditions Duculot, 1995.

[MUC 91] *Proceedings of the Third Message Understanding Conference (MUC-3)*, Morgan Kaufmann Publisher, 1991.

[MUC 92] *Proceedings of the Fourth Message Understanding Conference (MUC-4)*, Morgan Kaufmann Publisher, 1992.

[MUC 93] *Proceedings of the Fifth Message Understanding Conference (MUC-5)*, Morgan Kaufmann Publisher, 1993.

[MUC 95] *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, Morgan Kaufmann Publisher, 1995.

[MUC 99] *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, http://www.muc.saic.com/proceedings/muc_7_toc.html, 1999.

[PAZ 97] PAZIENZA M.T. (éd.), *Information Extraction*, Springer Verlag, 1997.

[POI 98] POIRIER C., MATHET Y., ENJALBERT P. « La compositionnalité à l'épreuve des faits, dans un projet de compréhension automatique de constats d'accidents », *TAL, Traitement automatique des langues*, 39(1), p. 99-130, 1998.

[POI 99] POIBEAU T., « Evaluation des systèmes d'extraction d'information : une expérience sur le français », *Langues*, 2(2), p. 110-118, 1999.

[RUB 00] RUBIO A. et al Eds, *Second International Conference on Language Resources and Evaluation, LREC 2000*, Athènes, Grèce, 31 Mai - 2 Juin 2000, Publié par European Language Resource Association, http://www.icp.inpg.fr/ELRA/

[SAB 90] SABAH G., *L'intelligence artificielle et le langage*, 2ème édition, Hermes, 1990.

[SAG 81] SAGER N., *Natural Language Information Processing*, Addison-Wesley, 1981.

[TUR 99] TURBOUT C.., « *Systèmes d'informations documentaires : spécificités du document géographique* », O1'Design pages 67-77, Saint Férréol, France, Décembre 1999, http://infodoc.unicaen.fr/OhRAGE.

[VER 94] VERGNE J., « A non-recursive sentence segmentation, applied to parsing of linear complexity in time » *Proceedings of International Conference on New Methods in Language Processing*, Manchester, 1994.

[VIC 98] VICTORRI B. *et al*, « Le projet TACIT : Traitements Automatiques pour la Compréhension d'Informations Textuelles », GIS Sciences de la Cognition, http://elsap1.unicaen.fr/operations.html, 1998.

[WIL 97] WILKS Y., « Information Extraction as a Core Language Technology », in [PAZ 97].

[YAN 99] YANGARBER R., GRISHMAN R. : Description of the Proteus/PET System as used for MUC-7, in [MUC 99].

[ZWE 97] ZWEIGENBAUM P., *Traitement automatique de la langue médicale*, Mémoire d'Habilitation à Diriger des Recherches, Université de Paris-Nord, 1997.