Prendre en compte la dimension globale d'un corpus dans la contextualisation du sens : expérimentations en informatique linguistique

Pierre Beust, Thibault Roy

PRENDRE EN COMPTE LA DIMENSION GLOBALE D'UN CORPUS DANS LA CONTEXTUALISATION DU SENS : EXPÉRIMENTATIONS EN INFORMATIQUE LINGUISTIQUE

Pierre BEUST, Thibault ROY

GREYC CNRS UMR 6072 & pôle ModeSCoS
Université de Caen Basse-Normandie

Cet article s'inscrit dans le cadre de recherches en cours dans le domaine du Traitement Automatiques des Langues (TAL). Plus précisément, nous cherchons à mettre en œuvre des traitements sémantiques adaptés à certaines tâches informatisées où les spécificités socio-linguistiques des utilisateurs (par exemple leurs centres d'intérêt, leurs habitudes terminologiques) et leurs interprétations sont au centre de l'interaction homme-machine. C'est par exemple le cas dans bon nombre de tâches dans le domaine de la Gestion Electronique de Documents (GED) tels que le classement, l'archivage ou la recherche de documents. C'est plus encore le cas dans le domaine de la veille documentaire ou de la recherche d'information sur l'Internet.

La question du sens (sa construction et sa nature) est bien sûr très liée aux rapports entre des documents (majoritairement textuels) et des sujets interprétants travaillant avec ces documents et en produisant par ailleurs. Elle est centrale dans plusieurs phases de travail fréquentes que constituent l'élaboration et les analyses de corpus ou encore la construction et la gestion de ressources terminologiques. Nous allons montrer dans cet article comment nous envisageons et expérimentons dans nos travaux en informatique linguistique les rapports entre ces notions complexes et fortement connexes de sens, de contexte, de corpus et de ressources.

Après avoir posé notre cadre d'étude et plus précisément notre démarche centrée sur les besoins d'un utilisateur, nous décrirons comment nous abordons à travers nos recherches la question de la construction du sens. A partir de notre point de vue sur la nature du sens nous aborderons les notions de contexte, co-texte et d'intertexte. Nous préciserons par là ce que nous entendons par les rapports entre le local et le global. Nous présenterons alors nos travaux et nos expériences en cours dans le domaine de la cartographie thématique de corpus. Enfin nous ferons état des perspectives de recherche qui s'ouvrent à nous à ce moment de nos travaux.

Cadre d'étude

Les recherches que nous menons au sein de l'équipe ISLanD (Interactions, Sémiotique, Langues et Diagrammes) du laboratoire GREYC CNRS UMR 6072 à l'Université de Caen –

Basse Normandie s'inscrivent et trouvent leurs principales applications dans le cadre d'étude de la veille documentaire et de la recherche d'information, le plus généralement sur Internet. Afin de mettre en évidence les enjeux scientifiques, nous allons ici dresser un bref « état des lieux » des rapports entre ce cadre d'étude et de la problématique de la construction du sens.

Les technologies de l'information sur l'Internet forment un domaine d'application direct de l'ingénierie linguistique et plus précisément de l'accès au contenu des documents, d'où le rapport avec la problématique du sens. La taille des données textuelles à traiter ainsi que le nombre et la variété des traitements à réaliser rendent incontournable le développement de méthodes d'analyses automatiques fiables et rapides. A titre d'exemple, on peut se rappeler que le fameux moteur de recherche Google indexe aujourd'hui environ 8 milliards de documents et estimait en février 2003 traiter environ 250 millions de requêtes par jour.

Plusieurs types d'outils de TAL sont spécifiquement dédiés à la problématique du document. Ils constituent une évolution majeure du TAL aujourd'hui. Dans certains cas y sont réinvestis des travaux sur la compréhension des textes provenant de la tradition logico-grammaticale (c'est par exemple le cas des systèmes mis en compétition dans le cadre des conférences MUC[1]). Dans d'autres cas, on observe des démarches plus pragmatiques qui tentent de tirer de profit de larges corpus et de méthodes d'apprentissage automatiques (Claveau, 2003).

Adeline Nazarenko dans (Condamines & al., 2005, Chap. 6) établit quatre familles de méthodes automatiques d'accès au contenu des documents : l'extraction d'information, les méthodes de question/réponse, le résumé automatique et l'aide à la navigation. On entend par extraction d'information les méthodes qui consistent à rechercher dans un corpus très homogène (par exemple des dépêches d'actualité dans ou encore des articles scientifiques) des informations dont on sait qu'elles s'y trouvent. Ainsi on cherche par exemple dans un corpus d'actualité boursière à extraire les transactions de rachats et de fusions de sociétés ce qui revient à chercher à remplir des sortes de formulaires électroniques indiquant notamment qui a acheté qui, à quel prix et quand. Il est donc souvent visé ici d'alimenter de manière entièrement automatique des bases de données préexistantes à partir de corpus soigneusement sélectionnés. Les méthodes dites de Questions/Réponses n'ont pas le même objectif. Elles consistent à chercher un fragment de texte extrait d'un corpus volontairement assez généraliste dans lequel un sujet interprétant a de bonnes chances de trouver la réponse à une question qu'il aura formulée en langue naturelle. Par exemple extraire une séquence du style « *(…) la vie de Baudelaire, auteur des Fleurs du mal, fut (…)* » à la question « *Qui a écrit les Fleurs du mal ?* ». La bonne construction linguistique de la réponse n'est pas ici visée car il ne s'agit que de fournir une « fenêtre » dans une chaîne de caractères, éventuellement en essayant tout de même de ne pas couper des mots en leur milieu. Lors des conférences d'évaluation TREC9[2], les systèmes de questions/réponses avaient pour consigne de rendre des réponses de moins de 250 caractères à partir de 980 000 documents et de 700 questions. A la différence des méthodes d'extraction d'information, la construction de sens dans le cours de l'analyse est moins primordiale dans la mesure où finalement on s'en remet à l'interprétation d'un sujet humain. Les méthodes de résumé automatique s'appuient aussi largement sur l'interprétation de celui à qui est destiné le résumé. Bien souvent il est plus juste de parler de condensation ou de réduction de textes plutôt que de résumé (dans le sens de ce qu'est un résumé quand il est rédigé par un sujet humain). L'enjeu technique est de rechercher des phrases dont on pense qu'elles ont un statut assez significatif (par exemple une phrase qui commencerait par « *en somme, on constate que (…)* » a de bonnes chances de synthétiser ce qui est dit avant) et de les juxtaposer dans un « résumé » où l'on compte que celui qui le lira

[1] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/proceedings_index.html (consultée le 29-06-2006)

[2] http://trec.nist.gov/pubs/trec9/t9_proceedings.html (consultée le 29-06-2006)

pourra rétablir une certaine cohérence textuelle, par exemple relativement aux rattachements anaphoriques.

Les méthodes d'extraction d'information, de Question/Réponse et de résumé automatique s'adressent principalement à la dimension rhématique des documents en cherchant d'une certaine façon à savoir ce qui est dit, où et comment. En général, les méthodes d'aide à la navigation s'adressent plus spécifiquement à la dimension thématique des documents (dans le sens où l'on cherche de manière plus globale à savoir de quoi traite un document ou un ensemble de documents). Les applications les plus courantes de ces méthodes sont l'indexation de document, l'extraction de terminologies, l'aide à la lecture (visualisation de documents ou encore création d'index par exemple), le groupement en classes de documents, la cartographie de corpus (qui, comme nous le verrons par la suite, est l'objet de nos recherches et nos développements).

Les quatre familles de méthodes d'accès au contenu présentées ci-dessus regroupent des projets de recherche où sont mis en œuvre beaucoup d'intelligence du point de vue des collaborations interdisciplinaires, notamment entre la linguistique et l'informatique. Cependant force est de constater que la majeure partie de ces projets de recherche sont toujours à l'état de prototypes de laboratoire et sont jusqu'à présent très peu mis en application et évalués dans des outils sur Internet à destination du plus grand nombre. Cela a des conséquences comme le montrent (Lavenus & al., 2002) à propos des méthodes de Question/Réponse en mettant en évidence la différence entre les corpus de référence utilisés dans les conférences TREC par rapport à des vraies questions d'utilisateur en recherche documentaire. Les auteurs notent que les questions du corpus de référence sont toutes des interrogatives canoniques courtes (par exemple *« What does a defibrillator do ? »*) alors que la majorité des demandes de « vrais » utilisateurs sont couramment des affirmatives complexes du style « *je voudrais savoir (…)* ».

Paradoxalement, si d'un point de vue informatique et algorithmique les méthodes couramment utilisées notamment par les moteurs de recherche sont très fines et fiables, on constate effectivement qu'elles restent linguistiquement très pauvres, à la fois du point de vue de leur fonctionnement propre mais également du point de vue de l'interaction avec leurs utilisateurs.

Les méthodes d'indexation utilisées par les moteurs de recherche pour associer des documents à des mots clés potentiels en sont un exemple. Cette indexation est dite « *Full Text* » dans le sens où tous les mots figurant dans un document sont gardés comme entrée d'index pour ce document. Pas étonnant dans ces conditions que les mots grammaticaux indexent une multitude de documents (expérience faite sur Google le 25/8/05 : une recherche stupide avec le mot clé unique « de » donne 873 000 000 réponses[3] et encore il est clair qu'en ce qui concerne les mots grammaticaux et le nombre de réponses possibles potentiellement, seules les pages considérées comme relativement importantes sont rendues). Ceci a des inconvénients, notamment la taille énorme des bases d'index que le moteur de recherche doit archiver et doit être capable d'interroger rapidement. En fait, l'intérêt de cette indexation un peu brutale réside dans le fait de pouvoir garder facilement comme index tout ce qui dans un texte ne peut être retrouvé dans un dictionnaire. C'est surtout le cas des entités nommées telles que des noms propres, des expressions temporelles ou encore des noms de quantité qui restent importants par rapport aux documents (on imagine mal par exemple que le nom d'une

[3] Il ne faut rester assez prudent sur l'interprétation du nombre de réponses rendu par un moteur de recherche car ce nombre est en fait souvent estimé plutôt que réel. Cette estimation en parfois fausse de manière évidente comme en témoigne une expérience menée par Jean Véronis (cf. http://aixtal.blogspot.com/2005/01/web-google-perd-la-boole.html (consultée le 29-06-2006)) en février 2005 où la requête Chirac rendait 3,2 millions de documents alors que la requête Chirac OR Sarkozy en rendait un peu moins de 2 millions, ce qui n'est pas logiquement cohérent.

société ne puisse pas être gardé comme entrée d'index pour son site web) mais dont il est difficile de dresser un catalogue fiable et durable[4]. La question du repérage et même de l'étiquetage (en tant que nom d'organisation ou de nom de lieu par exemple) des entités nommées est un enjeu important du TAL aujourd'hui et de nombreux projets de recherche abordent cette question avec des résultats intéressants mais leurs avancées n'ont pas encore eu de retombées sur les méthodes d'indexation utilisées sur Internet.

Dans leurs interactions avec les utilisateurs, les moteurs de recherche sont souvent assez rudimentaires d'un point de vue linguistique. Il faut bien souligner que l'utilisateur et son objectif de recherche sont uniquement considérés sous la forme d'une liste de mots clés (dont la casse et l'accentuation et même l'ordre sont d'ailleurs rarement pris en compte[5]) considérés pour une seule recherche dans la mesure où toutes les requêtes sont traitées indépendamment les unes des autres. Dans la pratique on s'aperçoit que pour mener à bien une recherche sur le web, il convient en fait d'interroger successivement plusieurs fois le (ou les) moteur(s) en ajoutant ou en précisant certains mots clés en fonction des résultats rendus à chaque étape. C'est donc le plus souvent à l'utilisateur seul qu'il convient de développer des stratégies efficaces pour trouver des mots clés adaptés à sa recherche. Certaines tentatives sont mises en place par certains moteurs pour aller un peu plus loin que la simple prise en compte de mots clés. Par exemple, Google permet de rechercher un mots clé ou un de ses synonymes avec l'opérateur tilde ~ (par exemple une recherche sur `powerpoint ~help` effectuera une recherche sur `powerpoint ET help ou tips, faq, tutorial`). Cependant, c'est le moteur lui-même qui établit ses listes de synonymes et il serait peut être plus judicieux que celles-ci soit validées par les utilisateurs quand ils les utilisent. Il convient donc, en tant qu'utilisateur, de rester très prudent quant aux compétences linguistiques des moteurs. Toujours à propos de Google, on trouve un exemple de résultat assez malheureux de l'opérateur `define` sur le *blog* de Jean Véronis[6]. L'opérateur `define` (disponible pour les pages en français depuis avril 2005) sert à rechercher à propos d'un mot des pages Web où ce mot ferait visiblement l'objet d'une définition. L'expérience relatée consiste à rechercher ainsi sur Google une définition du mot *femme* avec la requête `define:femme`. Les résultats donnés sont pour le moins plus que contestables. On aurait donc bien tort de croire à la fiabilité de l'opérateur `define` (qui pourtant est présenté par Google comme un outil de recherche de définition sans plus de détails) comme on aurait tort aussi de considérer le Web dans son ensemble comme une encyclopédie dans lequel on puisse rechercher des définitions attestées, notamment d'un point de vue moral.

En matière d'ingénierie documentaire la tendance actuelle est pourtant de renforcer ce genre d'utilisation du Web en cherchant à en faire une vaste base de connaissances, ce qu'évidemment il n'est pas. C'est la démarche considérée dans le projet du Web Sémantique où l'objectif annoncé par Tim Berners-Lee (Berners-Lee, 1998), initiateur du projet et directeur du W3C, est d'enrichir (notamment au moyen des technologies développées autour du langage XML) les documents (à l'aide d'ontologies normalisées, soit automatiquement, soit en assistant leur auteurs) avec des informations sur leur propre sémantique qui soit directement interprétables par des agents logiciels sans la supervision d'une interprétation

[4] Bien qu'elles soient délicates à construire et maintenir dans le temps, de telles ressources existent. C'est le cas de CELEX (CELEX, 1998), un lexique de 160 595 mots fléchis (avec leur lemme et leur catégorie syntaxique), une liste de 8 070 prénoms et de 211 587 noms de familles, une liste de 22 095 entreprises et 649 noms d'organisations, une liste de 7 813 villes et une autre de 1 144 pays et une liste sur les unités physiques et monétaires. Certaines de ces listes proviennent de sources déjà connues (c'est le cas des 22 095 entreprises issues du « *Wall Street Research Network* »), d'autres (les noms d'organisations) sont issues d'une acquisition lexicale sur Internet (Jacquemin & al., 2000) et d'autres sont construites manuellement (par exemple les unités physique et les monnaies)

[5] C'est par exemple le cas du moteur Exalead : http://www.exalead.com (consultée le 29-06-2006)

[6] http://aixtal.blogspot.com/2005/04/web-la-femme-selon-google.html (consultée le 29-06-2006)

humaine. Ceci fait l'hypothèse que la valeur sémantique d'un passage de document est le fait de son auteur alors que c'est finalement bien plus celui de son lecteur. L'expérience sur la définition de *femme* nous apprend bien qu'une définition considérée comme telle par quelqu'un n'a pas pour autant cette valeur pour d'autres et qu'au final c'est celle de l'utilisateur du moteur qu'il faudrait considérer.

Notre approche de la veille documentaire sur l'Internet se situe à l'opposé de celles défendues dans le cadre du Web Sémantique. Elle s'en distingue essentiellement par le fait que nous mettons l'accent sur des traitements et des ressources termino-ontologiques (bases de données terminologiques, représentations du contenu lexical etc.) avant tout centrés sur leur utilisateur, de sa tâche, de ses besoins et de ses centres d'intérêt. Là où le Web Sémantique cherche à rendre le plus possible partagées de vastes ontologies qui synthétisent une connaissance pensée comme objective et devant convenir à tous les utilisateurs, nous préférons manipuler des ressources propres à un utilisateur ou un petit groupe d'utilisateurs. Il en découle une certaine *légèreté* de ces ressources, au sens de (Perlerin, 2004), dans la mesure où elles ne représentent que ce qui est important du point de vue de l'utilisateur et restent ainsi de taille raisonnable (par exemple une centaine de termes) ce qui les rend moins complexes à construire, à maintenir et à enrichir.

Cette approche centrée utilisateur conduit à opérer un certain renversement scientifique relativement aux ressources qu'utilisent les modèles de TAL. Premièrement, d'un point de vue très pratique, force est de constater que des ressources très généralistes, valables pour tout type de traitement envisagé ainsi qu'à destination de tout utilisateur potentiel, ne sont pas facilement disponibles (sous forme électronique pour des traitements automatiques) et encore moins gratuites. Deuxièmement, nous soutenons que l'idée même d'une ressource généraliste est illusoire car elle dépend inévitablement du contexte qui lui préexiste (le but recherché par le ou les auteurs ainsi que leurs spécificités socioculturelles). Le rapport de l'Action Spécifique 32 du CNRS/STIC en 2003 (Charlet & al., 2003) va également dans ce sens en précisant un obstacle au projet du Web Sémantique : la détermination et l'ajout, même de simples méta-données, n'est pas une activité naturelle pour la plupart des personnes.

La tradition logico-grammaticale et plus précisément la sémantique formelle et computationnelle cherchent à représenter et à produire, automatiquement ou pas, des formes le plus possible objectivées des significations et du sens. Dans la démarche centrée utilisateur, on part d'une position duale où l'on considère que les traitements sémantiques appliqués à l'accès au contenus des documents ont tout à y gagner à être le plus possible subjectivés, tant du point de vue des ressources que du point de vue des résultats opératoires. Cette démarche nous paraît être une réponse au constat que dressent Didier Bourigault et Nathalie Aussenac-Gilles à propos de la variabilité des terminologies :

> *(...) le constat de la variabilité des terminologies s'impose : étant donné un domaine d'activité, il n'y a pas* une *terminologie, qui représenterait le savoir du domaine, mais autant de ressources termino-ontologiques que d'applications dans lesquelles ces ressources sont utilisées (Bourigault & al., 2003).*

Les ressources utilisées dans nos expérimentations sont produites de manière endogène dans une boucle d'interaction entre un outil logiciel, un utilisateur et des corpus où chaque pôle est déterminant. Il en découle une importance significative des corpus utilisés qui du coup ne peuvent plus être considérés uniquement comme un réservoir de formes attestées sur lequel on tenterait de mettre en œuvre un calcul à base de ressources exogènes. Le corpus utilisé dans le cadre de nos outils de TAL est à l'origine des ressources lexicales construites et constitue en même temps le matériau d'expérimentation de nos propositions. Ainsi notre démarche s'inscrit dans un processus de recherche et de développement en aller-retour entre des outils (des logiciels d'étude), des corpus (des corpus d'étude) et des utilisateurs, les uns étant conditionnés par les autres.

Comme on le voit bien, bon nombre de démarches diffèrent grandement dans les méthodes d'accès au contenu. Il y a un point incontournable sur lequel il convient également de préciser les ancrages épistémologiques, c'est la notion même de contenu et plus largement la question du sens. C'est ce que nous allons faire très succinctement dans la partie suivante.

La question de la construction du sens

Dans une communication sur l'histoire des traitements sémantiques en TAL, Gérard Sabah[7] précise ce que peut être le sens du point du résultat visé par telle ou telle sémantique :

- Préciser les conditions de vérité de l'expression traitée (dans le cas d'une sémantique vériconditionnelle) ;
- Décrire une expression comme l'ensemble des propriétés théoriques que possèdent les concepts correspondants (comme en sémantique intensionnelle) ;
- Décrire une expression comme l'ensemble des objets ou des situations du monde de référence que cette expression peut désigner (on parlera alors de sémantique extensionnelle ou également de sémantique dénotationnelle ou référentielle) ;
- Chercher à décomposer le contenu des mots en éléments de sens plus primitifs pour étudier les possibilités de combinaison de ces éléments (on est ici dans le cadre d'une sémantique componentielle) ;
- Décrire une expression comme l'ensemble des actions à effectuer pour trouver l'objet désigné (on parle ici de sémantique procédurale) ;
- Mettre en évidence les marqueurs et les constructions utilisées pour qu'un énoncé puisse servir comme un argument en faveur d'un autre énoncé (il s'agit alors d'une sémantique argumentative).

Cette catégorisation, qui ne se veut pas exhaustive, montre la grande variété des points de vue sur le sens. Aucune de ces sémantiques n'a complètement tort ou complètement raison. Par exemple, il existe bon nombre d'énoncés en langue naturelle dont le sens a un certain rapport avec la vérité ou encore la référence, mais cela ne veut pas dire que *tout* énoncé est interprétable en terme de vérité de ou de référence. Il faut donc en adoptant l'une ou l'autre de ces sémantiques garder à l'esprit que l'on ne cherchera avec à rendre compte uniquement que d'une partie du sens qui de fait ne le couvre pas dans son ensemble.

Selon la catégorisation de Gérard Sabah, nous situons notre approche de la notion de sens dans le cadre d'une sémantique componentielle et plus précisément dans un héritage scientifique de la Sémantique Interprétative (SI) de François Rastier (Rastier, 1987), elle-même en filiation avec les travaux en sémantique structurale dont l'origine remonte à Hjelmslev. De la SI on retient principalement deux aspects. Premièrement on en tire (parfois avec quelques adaptations au cadre d'une instrumentation informatique[8]) un « appareillage » théorique fin pour la description d'effets de sens. Par exemple, les notions de sèmes, d'isotopies (i.e. récurrences d'un même sème dans un texte), etc. Deuxièmement, on s'approprie de la SI un positionnement épistémologique relativement à la question du sens. Celui-ci consiste à préférer la tradition rhétorique et herméneutique à la tradition logico-grammaticale. Ainsi on défend, d'une part, le principe selon lequel le global détermine le local (ce qui marque une rupture avec le principe de compositionnalité) et, d'autre part, que le sens ne peut pas être intégralement objectivé (Rastier, 1998).

Si le sens ne peut pas être objectivé, il peut encore moins l'être de manière formelle (comme cela est souvent le cas dans la tradition logico-grammaticale). On rejoint ici l'avis de (Nicolle, 2005) pour qui le sens n'est jamais capturé par ses représentations. Toute

[7] http://www.limsi.fr/Individu/gs/textes/ATALA-14.12.96/LePointSurLeSens.html (consultée le 29-06-2006)
[8] C'est par exemple le cas concernant la notion de sème que nous avons redéfini dans (Beust, 1998).

représentation du sens est forcément incomplète et il n'y a donc pas de langage formel qui puisse reproduire fidèlement le sens d'un énoncé en langue naturelle alors que tout énoncé formel peut être reformulé dans une langue. Anne Nicolle en tire la conséquence que la langue est un langage terminal. L'interprétation langagière est en cela bien distincte de l'interprétation logique qui se résume à la traduction dans un autre langage. Dans le cas de l'interprétation langagière, il n'y a pas d'autre langage.

Ainsi, à la manière de Jacques Courcil qui définit les principes de non consignation et de non préméditation de la chaîne parlée (Courcil, 2000), nous défendons un principe de non transformation du sens en langue naturelle dans la mesure où il n'y a pas de pensée construite possible qui ne soit pas déjà sous forme langagière. Le sens est donc une réalité concrète intralinguistique et subjective. Dans le cadre de la SI, l'interprétation est considérée comme une perception sémantique, perception forcément individuelle, dont toute tentative d'objectivation est une sommation incomplète de points de vue. Ainsi le sens d'un texte est une interprétation à un moment donné et dans une tâche donnée d'un sujet interprétant, ce qui est à notre avis un argument fort pour une instrumentation de la sémantique des langues individu-centrée.

Beaucoup de travaux en sémantique formelle (logique, DRT[9], SDRT[10] etc.) ont depuis des années déployé beaucoup d'intelligence pour obtenir de façon compositionnelle un « calcul du sens » acceptable. Force est de constater qu'un tel résultat n'est toujours pas atteint à l'heure actuelle. Il ne s'agit pas ici que d'un problème d'évaluation dont on n'aurait pas encore bien mis en place la méthodologie mais d'un problème beaucoup plus profond. Dès lors qu'on parle de « vrais » textes, de « vrais » corpus, et pas simplement de phrases d'exemples artificiellement construites en dehors d'un contexte linguistique et pragmatique, il convient de se rendre compte que la dimension interprétative personnelle fait qu'il n'y a pas de consensus évident sur ce qu'est ou n'est pas le sens d'un texte. Il en résulte, à notre avis, que le sens ne peut être modélisé à la façon d'un résultat calculatoire qui serait plus ou moins complété ou dégradé d'un interprétant à un autre. Le sens n'est pas de nature symbolique ; c'est un processus sémiotique au centre de l'activité de l'interprétant qui est complexe, notamment parce qu'il est réflexif.

Dès lors, la construction du sens n'est pas une question d'extraction à partir du matériau linguistique, ni même de calculs sur une extraction d'informations à partir du matériau linguistique. Ce ne sont pas tant les caractéristiques propres des mots, des phrases ou des paragraphes qui priment dans le sens des textes mais c'est ce que les interprétants en attendent ou y projettent. Des critères externes aux textes peuvent être tout aussi importants. D'une certaine manière, le succès du moteur de recherche Google nous en donne un exemple à propos du rapport entre le contenu d'une page et l'importance de cette page du point de vue des utilisateurs du moteur de recherche. Comparativement à d'autres moteurs de recherche, le classement par importance des pages Web répondant à une requête ne dépend pas avant tout de leur contenu. Pour AltaVista par exemple, la pertinence d'une page dépend de critères liés à son contenu (présence répétée d'un mot clé de la requête dans le contenu, dans le titre ou dans les méta-données). Pour Google, c'est l'algorithme de *PageRank*[11] qui conditionne la pertinence d'une page avec, en plus des techniques classiques donnant une importance particulière à certaines zones (par exemple les titres), le principe suivant[12] : plus il existe de pages qui ont un lien vers la page P, plus P est pertinente (quelle que soit son contenu et quels que soient les mots clés de la requête). Les travaux menés dans le cadre du projet PRINCIP (Valette & Grabar, 2004) visant la détection automatique de sites Internet au contenu illicite

[9] Discourse Representation Theory (Kamp, 1981)
[10] Segmented Discourse Representation Theory (Asher, 1993)
[11] http://www.google.com/technology/ (consultée le 29-06-2006)
[12] Par rapport à d'autres moteurs de recherche, ce principe donne à Google un caractère résolument socio-centré.

(principalement des propos racistes ou antisémites) montrent également que le sens est de nature pluri-sémiotique. Il provient de la présence conjointe de plusieurs facteurs dont certains sont extérieurs aux textes. Ainsi, du point de vue de la thématique des textes, des propos racistes ou anti-racistes sont parfois très proches à tel point qu'une détection automatique fiable uniquement basée sur la recherche de certains mots clés du texte n'est pas facile à obtenir. Si en plus on veut qu'elle soit fiable dans la durée, cela devient très difficile étant donné qu'on ne peut pas prévoir à l'avance l'usage de certains mots dans certains contextes. Il faut alors exploiter d'autres critères pour fiabiliser cette détection : la ponctuation qui traditionnellement n'est pas prise en compte (on remarque de façon statistiquement significative que les sites racistes utilisent fortement le point d'exclamation et que les sites anti-racistes utilise plutôt des points de suspension), le type de police de caractère utilisé (la police Arial semble significativement caractéristique des sites racistes), les couleurs de fond et de police de caractères utilisées (le rouge et le noir sont aussi significativement caractéristiques des sites racistes), les contenus des images entourant le texte (la thématique de l'animal dans les textes racistes est souvent corrélée avec des dessins montrant des animaux souvent connotés de façon péjorative, le rat par exemple).

Dans la logique de nos partis pris épistémologiques, la question de la construction du sens se trouve déplacée. Nous ne l'abordons pas avec l'idée d'un traitement automatique qui produirait un résultat (quand bien même ce résultat serait un processus). Nous l'abordons d'une autre manière sous l'angle de l'instrumentation informatique dédiée à *l'assistance* à l'interprétation. Sous cet angle de vue il convient de développer des outils logiciels pour mettre en place des interactions homme-machine où l'utilisateur, les textes et ses interprétations sont l'objet de la boucle interactive. L'adaptation des interfaces à cette boucle interactive, notamment par l'utilisation de méthodes de visualisation adéquates tient une place importante (c'est notamment ce qu'ont bien compris les terminologues qui ont travaillé à la réalisation de concordanciers). On va donc chercher à développer des méthodes d'accès au contenu qui plutôt que d'extraire du sens vont chercher à alimenter des interactions (notamment des techniques de visualisation telles que des cartes ou des diagrammes) avec des signes qui ont pour objectif de faire sens du point de vue de l'utilisateur. La mise en œuvre de ces méthodes d'accès au contenu confère, comme nous allons le voir dans la suite, un statut important à la dimension globale des corpus.

Le rapport local/global dans la contextualisation du sens

Le principe de détermination du local par le global propre à l'approche herméneutique est un principe de contextualisation du sens (au passage, principe alternatif à la compositionnalité). La question de la contextualisation du sens de telle ou telle unité linguistique du texte (mot, syntagme, paragraphe par exemple) est d'une grande importance dans la perspective de la sémantique interprétative car elle est la base de l'établissement de parcours interprétatifs, c'est-à-dire des suites d'opérations permettant d'assigner un ou plusieurs contenus à des expressions (ce qui explique comment les langues peuvent s'acquérir réflexivement par leur pratique).

Dans la contextualisation entrent en compte, selon nous, 3 notions : le co-texte, le contexte extralinguistique et l'intertexte :

- On entend par co-texte d'une unité linguistique son « entourage » dans le texte, c'est-à-dire un passage de texte : une zone de localité sémantique pertinente autour d'une unité. Cette zone est appelée Période (Rastier & al. 1994, p. 116) et elle est délimitée par l'étendue des relations d'isotopies, de prédication et d'anaphore ;

- Le contexte extralinguistique regroupe les conditions pragmatiques liées à l'interprétation du texte. Dans le cadre de nos expérimentations en recherche d'information, on limitera ce contexte à l'utilisateur et sa tâche (voire au groupe d'utilisateurs et leur tâche) ;
- L'intertexte rassemble tous les documents que l'utilisateur estime liés à un texte du point de vue de son interprétation. Tout texte mis en relation avec d'autres textes en reçoit des déterminations sémantiques et modifie potentiellement le sens de chacun des autres textes (c'est le principe d'architextualité définit dans (Rastier, 2001). Un document peut appartenir à plusieurs intertextes comme un texte peut s'interpréter dans plusieurs intertextes en fonction des relations sémantiques établies, c'est-à-dire des objectifs interprétatifs. On peut penser par exemple qu'un même article de presse n'indique pas le même point de vue quand il y est fait référence dans une revue de presse des plus sérieuses ou dans les colonnes d'un journal satirique.

Ainsi contextualiser, c'est établir au sein du co-texte des parcours interprétatifs qui tiennent compte du contexte extralinguistique et de l'intertexte. Ce que l'on peut analyser avec les moyens d'une sémantique componentielle comme mettre en évidence dans certains passages du texte des sèmes particulièrement importants. Ainsi l'analyse de la détermination du local par le global consiste à identifier localement des sèmes pertinents issus du global (le contexte ou l'intertexte). Le contexte et l'intertexte restent deux notions aux contours assez flous qu'il convient de circonscrire ici. Comme nous le verrons par la suite, le contexte est vu dans le cadre de nos expérimentations via les ressources personnelles construites par le ou les utilisateurs et l'intertexte via le corpus d'étude[13] sur lequel les outils logiciels développés permettent de travailler. A l'égal des ressources exploitées, le corpus « matérialisant » l'intertexte en tire un rôle central par rapport à la construction du sens en contexte, ce qui lui confère bien plus d'importance qu'un simple réservoir de formes attestées.

Identifier et caractériser l'importance d'un sème dans la contextualisation du contenu d'une unité est le résultat d'opérations interprétatives d'actualisation et de virtualisation de sèmes. Dans l'énoncé *Le facteur m'a donné une lettre* le sème /courrier/ est actualisé dans le contenu de *lettre* parce qu'il se répète dans le contenu de *facteur* formant ainsi une isotopie. Cette actualisation permet de retenir la signification pertinente de *lettre* dans l'énoncé (on ne retient donc pas, par exemple, la signification de *lettre* en tant que caractère de l'alphabet) et précise une sélection du co-texte sur une partie du signifié de *lettre*. Ainsi, dans cet exemple, le sème /courrier/ est renforcé par le co-texte alors que ce n'est notamment pas le cas du sème /en papier/ appartenant également au contenu de *lettre* ; à l'inverse, ce sème serait probablement actualisé dans *Il a chiffonné sa lettre* et pas /courrier/. La virtualisation est l'opération interprétative duale de l'actualisation. Elle décrit une neutralisation d'un sème en contexte. Par exemple, dans le syntagme *Le chat immortel*[14], on dira que le sème /mortel/ appartenant au contenu de *chat* est virtualisé car non seulement il n'est pas répété dans l'énoncé mais, de plus, il est invalidé par le contenu sémique de *immortel*. La virtualisation ne doit pas être considérée comme un simple retrait d'un sème. L'idée d'une neutralisation temporaire est plus juste car, si dans la suite du texte et/ou de l'intertexte le sème virtualisé venait à réapparaître dans d'autres unités, il serait alors ré-actualisé.

L'actualisation et la virtualisation jouent un rôle important sur la mise en co-texte de contenus sémiques définis en langue, c'est-à-dire sur les sèmes dits inhérents. Cette notion de sème inhérent est à opposer à celle de sème afférent. Dans des contextes particuliers, on peut

[13] Le corpus n'est pas pour autant une simplification de l'intertexte, c'en est selon (Rastier 1998, note de bas de page n°17) une objectivation : « *Le corpus est la seule objectivation possible (philologique) de l'intertexte, qui sinon demeure une notion des plus vagues* ».

[14] Extrait de la phrase « *Le chat immortel a fui sur les toits comme s'il avait un démon à ses trousses.* » sur la page web http://rouscaille.tripod.com/rouscaill/id163.html (consultée le 29-06-2006).

actualiser des sèmes qui ne font pas partie des contenus des unités du texte. Ces sèmes sont dits afférents et l'opération interprétative qui consiste à les actualiser porte le nom d'afférence. L'afférence consiste en la production d'un sème qui vient créer ou renforcer une isotopie. Selon Thierry Mézaille[15], il y a une telle production d'un sème isotopant dès lors qu'est établie une relation sémantique d'assimilation ou de dissimilation. L'assimilation est une afférence co-textuelle (c'est-à-dire que le sème afférent est inhérent dans d'autres unité du co-texte) d'un sème générique par présomption d'une récurrence sémique. Par exemple, dans l'énoncé *Voici des choux, des concombres et des scoubidous* l'assimilation consiste en une afférence d'un sème à *scoubidou* où le but est de renforcer une répétition déjà initiée dans le co-texte. En l'occurrence, il s'agit d'enrichir le contenu sémique de *scoubidou* avec le sème afférent générique[16] /légume/ pour rendre le thème de l'énoncé uniforme[17]. L'opération interprétative inverse de l'assimilation est la dissimilation. Alors que l'assimilation diminue, par afférence, les contrastes forts, la dissimilation, quant à elle, augmente les contrastes faibles. La dynamique des sèmes en cause dans la dissimilation n'est plus une afférence de sème générique, comme c'est le cas pour l'assimilation, mais une afférence de sèmes spécifiques pour différencier, dans un co-texte, les contenus sémantiquement proches. Par exemple, dans l'énumération *routes et autoroutes*, le contenu de *route* doit décrire une signification spécifique qui exclut la signification de *autoroute* et non une signification générique qui inclut cette signification. La dissimilation est encore plus flagrante dans l'exemple suivant où il y a répétition de la même lexie dans *Il y a musique et musique*. Ici, la dissimilation consiste, à distinguer par des sèmes spécifiques les deux significations de *musique*. Ainsi, on peut aférer à la première le sème spécifique /agréable/ et à la seconde /désagréable/.

L'assimilation et la dissimilation sont des formes d'afférences co-textuelles mais la sémantique interprétative décrit également une autre forme d'afférence : l'afférence socialement normée. Dans de tels cas d'afférence, il y a bien enrichissement contextuel du contenu d'une lexie dans un énoncé par un (ou plusieurs) sème(s) (c'est bien pour cela qu'il s'agit toujours d'une afférence) mais cette fois, le ou les sèmes afférents ne sont pas inhérents dans d'autres contenus d'unités linguistique du co-texte. L'afférence est alors le fait d'une norme sociale partagée au sein d'une communauté linguistique. C'est, par exemple, le cas du sème /tristesse/ afférent au contenu de *noir* dans *il broie du noir* ou encore le cas du sème /bonheur/ dans *rose* dans *la vie en rose*. Là où l'afférence co-textuelle (par assimilation ou dissimilation) est le résultat d'un parcours interprétatif local, l'afférence socialement normée résulte de parcours interprétatifs beaucoup plus globaux (à titre d'exemple d'afférences globales sur l'ensemble d'un corpus, on peut citer l'étude de (Rastier, 1987) sur le roman de Stendhal *Le rouge et le noir* où il y a une afférence jusque dans le titre du roman des sèmes /armée/ et /Église/).

Le concept d'afférence est un outil théorique très fin pour la description de la dynamique des sèmes dans les corpus et pour la caractérisation des effets du global sur le local. Le problème de la modélisation de l'afférence (surtout en ce qui concerne l'afférence socialement normée) c'est qu'il faudrait avoir des ressources très larges pour pouvoir l'expérimenter de façon opératoire dans un traitement automatique. Comme on l'a dit plus

---

[15] cf. « Quels mécanismes pour (r)établir la cohésion sémantique textuelle ? Sur la prééminence des processus d'assimilation et de dissimilation dans l'interprétation des énoncés contradictoires et métaphoriques » disponible en ligne à l'adresse : http://www.chez.com/mezaille/contraphore.htm (consultée le 29-06-2006).

[16] L'assimilation concerne bien des sèmes génériques car c'est le sème /légume/ qui est afférent et pas le sème /vert/ inhérent à *choux* et *concombre* mais spécifique, ce qui a pour conséquence qu'il n'y ait pas de défaut d'assimilation dans *Voici des choux, des concombres et des carottes*).

[17] On retrouve ce type d'assimilation dans certaines formes d'humour lorsque le sème afférent est particulièrement inattendu dans le contenu sémique de la lexie en cause. Le titre du livre de G. Lakoff : *Women, Fire and Dangerous Things : What Categories Reveal about the Mind* (1987) en témoigne en provocant l'afférence du sème /dangereux/ au lexème *Women*.

haut, un texte n'est pas lié à un unique intertexte. Du point de vue du sujet interprétant et de son histoire propre, tous les intertextes forment un univers de textes construit individuellement de manière le plus souvent inconsciente. Théodore Thlivitis l'appelle l'*anagnose* (Thlivitis 1998, p. 41). Dans la perspective centrée utilisateur qui est la nôtre (ainsi que celle Théodore Thlivitis) l'anagnose serait idéalement ce qu'il faudrait formaliser pour représenter au mieux l'utilisateur et modéliser avec satisfaction ses afférences. Seulement, il n'est pas du tout évident, au contraire, que cette anagnose rassemblant l'histoire d'un individu (l'histoire de ses interprétations comme son histoire propre), sa culture, sa société, les données de son époque soit formalisable. On est ici face aux mêmes problèmes que ceux de la constitution des ontologies généralistes. Dés lors, les différentes ressources personnelles que nous serons amenés à construire et à manipuler ne toucheront que des infimes parties de cette anagnose. Nous ne prétendons donc pas rendre compte de l'ensemble des mécanismes d'affférence dans nos expérimentations informatiques. Ce n'est pas pour autant un problème majeur car il est possible de personnaliser efficacement (comme l'a montré Vincent Perlerin) des tâches de recherche d'information et d'accès au contenus de documents avec des ressources légères car d'emblée non exhaustives. C'est ce que nous allons montrer dans la suite en détaillant nos expérimentations à propos de la cartographie thématique.

La cartographie thématique : expérimentation sur corpus

La plate-forme ProxiDocs[18] (Roy et Beust, 2004) permet de construire différentes représentations globales d'un corpus de textes à partir de ressources terminologiques construites par l'utilisateur. Ces représentations sont appelées des cartes. Ce sont des visualisations topologiques interactives, personnalisées en fonction d'un utilisateur ou d'un petit groupe d'utilisateurs.

Les ressources utilisées pour produire les cartes vont représenter les thèmes ou les domaines choisis par l'utilisateur pour intervenir dans ses analyses. Deux types de représentations sont possibles selon les besoins de l'utilisateur : la représentation en « sacs de mots » où chaque thème est représenté par un ensemble de lexies s'y rapportant selon le point de vue de l'utilisateur ; et la représentation selon le modèle LUCIA de sémantique différentielle (Perlerin, 2004) où chaque domaine est représenté par un ensemble (appelé un dispositif) de catégories de lexies dont la signification est représentée par des différences de sèmes. Afin de construire de telles ressources terminologiques, nous proposons un ensemble de logiciels d'étude[19] complémentaires apportant une aide à l'utilisateur. Par exemple, les outils MemLabor (Perlerin, 2002) et FlexiConcord permettent une première analyse du corpus d'étude en réalisant respectivement une extraction des graphies répétées et une mise en contexte de termes et de leurs flexions. Après avoir isolé des termes pouvant intervenir dans ses analyses, l'utilisateur peut organiser ces termes en classes thématiques (qu'on appelle thèmes) à l'aide de l'outil ThemeEditor (Beust, 2002). Un principe de surlignage avec différentes couleurs (une couleur correspondant à un thème) permet de mettre en évidence la répartition, l'alternance et les enchaînements au long d'un texte des thèmes ainsi construits. Selon les besoins de sa tâche, l'utilisateur peut choisir de construire des représentations sémantiques plus fines selon le modèle LUCIA. Pour cela, l'outil LuciaBuilder (Perlerin, 2004, pp. 151-160) est mis à sa disposition afin de l'assister dans les différentes étapes de création des dispositifs représentant les domaines de son choix.

[18] http://www.info.unicaen.fr/~troy/proxidocs (consultée le 29-06-2006).

[19] Ces logiciels d'étude sont tous disponibles sur le site de l'équipe ISLanD du laboratoire GREYC de l'université de Caen : http://www.greyc.unicaen.fr/island/logiciel/ (consultée le 29-06-2006).

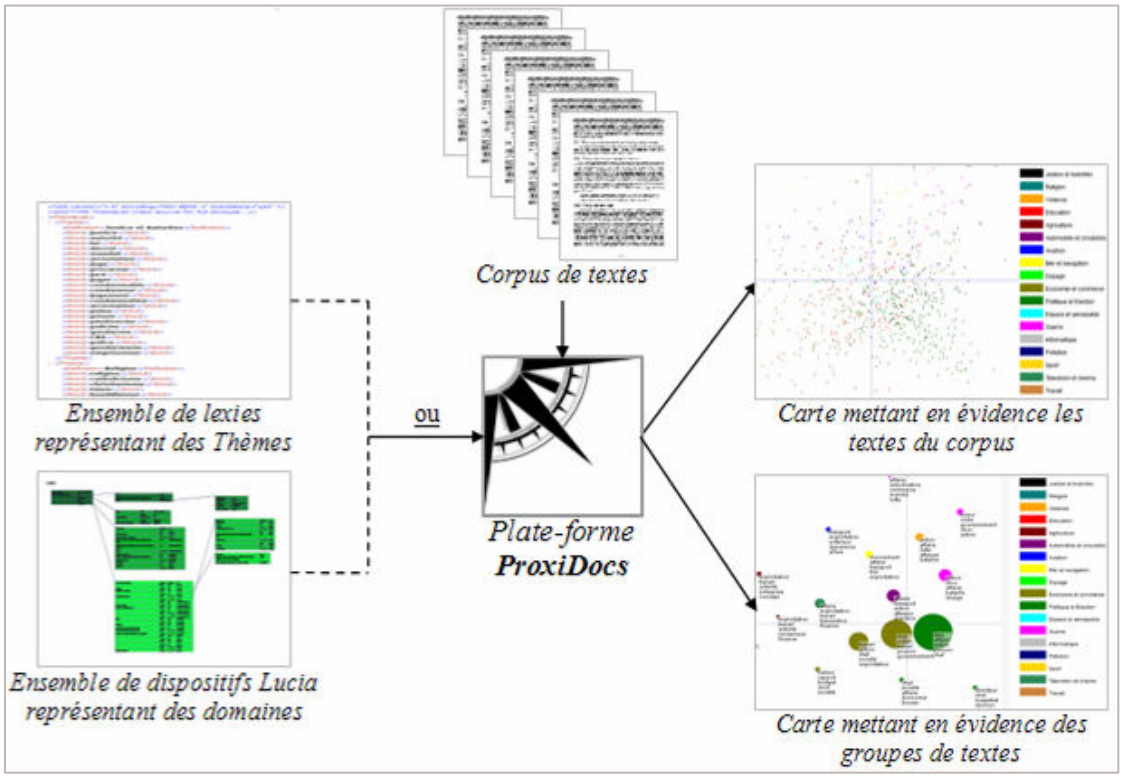

Figure 1 : Utilisation de la plate-forme ProxiDocs.

La plate-forme ProxiDocs permet d'opérer différentes visualisations d'une analyse globale du corpus à partir des ressources de l'utilisateur (la figure 1 illustre les possibilités de la plate-forme exploitées dans cet article) :

- des cartes en 2 ou 3 dimensions représentant chaque texte du corpus analysé par un point. Chacun de ces points est un lien hypertexte vers le texte où les lexies des thèmes ou des dispositifs sont mises en valeur à l'aide d'une technique de coloriage. La couleur d'un point correspond au thème majoritaire repéré dans le document représenté. De telles cartes sont utiles afin d'étudier les liens et les différences de thématiques abordées entre les documents du corpus ;
- des cartes en 2 ou 3 dimensions mettant en évidence des groupes de textes abordant des thèmes proches. Chaque groupe est représenté sur la carte par un disque ou une sphère de diamètre proportionnelle au nombre de textes qu'il contient. La couleur attribuée au disque ou à la sphère correspond au thème majoritaire repéré dans les textes du groupe. Les disques représentant les groupes portent un jeu d'étiquettes (au plus 5 étiquettes) indiquant les lexies localement les plus fréquentes dans l'ensemble des documents du groupe. Chacun de ces groupes est un lien hypertexte vers un rapport sur le contenu du groupe représenté. Ce rapport indique le texte le plus « représentatif » du groupe (celui situé le plus près du centre de gravité du groupe), présente la répartition des thèmes ou des catégories abordés, met en évidence les lexies répétées et propose un classement des isotopies les plus « importantes » des documents du groupe dans le cas d'une cartographie réalisée à partir de dispositifs LUCIA. Ces cartes permettent d'avoir un regard sur les principaux sujets abordés ainsi que sur la répartition des thèmes dans les textes du corpus ;
- des cartes en 2 dimensions animées mettant en évidence l'évolution des thèmes abordés dans les textes lorsque les documents qui constituent le corpus sont datés (c'est par exemple le cas d'un corpus de dépêches d'agence de presse). Ce type de cartes permet

de mettre en évidence les différentes thématiques abordées sur certaines périodes ainsi que leur enchaînement.

Expériences sur corpus réalisées avec ProxiDocs

Dans le but de caractériser la part de la dimension globale intertextuelle dans l'accès aux contenus de documents, nous relatons ici les principes et les résultats de quatre expérimentations logicielles avec la plate-forme ProxiDocs :

- la première sur corpus généraliste avec des ressources elles aussi assez généralistes sous forme de listes de thèmes créées par nos soins ;
- la deuxième sur corpus très spécialisé avec des ressources très spécifiques développées sous formes de dispositifs LUCIA (représentations terminologiques différentielles) également créés par nous même. ;
- la troisième sur un flux documentaire avec des ressources créées par nous même et visant les différentes thématiques abordées à différents moments du flux ;
- la quatrième sur un corpus et des ressources lexicales utilisées dans une recherche en cours sur la terminologie médicale par une chercheuse de l'université de Rouen.

La première expérience réalisée consiste à cartographier un corpus thématiquement hétérogène constitué d'environ 800 articles issus du journal *Le Monde* de 1987 à 1989. Cette expérience (détaillée dans (Roy, 2005)) prend place dans le domaine de la veille d'informations : elle a pour objectif de découvrir les principaux sujets abordés dans cet ensemble d'articles. Les cartes obtenues à l'issue de cette expérience (présentées en figure 2) ont été réalisées avec un ensemble de 18 thèmes généralistes que nous avons construits, tels la justice, la télévision, l'éducation, etc. La carte des articles met en évidence un nombre très important d'articles de thèmes majoritaires Politique et élection (quadrant inférieur droit de la carte) et Économie et commerce (quadrant inférieur gauche). Ces observations sont confirmées par la carte des groupes d'articles, la couleur, la taille et la disposition des groupes sur cette carte donnent une information sur les thèmes abordés dans les textes du corpus ainsi que sur leur répartition. En visualisant les rapports des groupes, l'utilisateur peut avoir une idée plus précise des thèmes abordés dans les articles de chaque groupe : il est ainsi facilement observable que le groupe de thème majoritaire Politique et élection contient principalement des articles traitant des futures élections européennes. Les différentes cartes construites durant cette expérience nous ont alors permis de mettre rapidement en évidence les grandes tendances du corpus (principaux sujets abordés dans les textes et groupes de textes abordant des thèmes proches), ce qui est un premier résultat satisfaisant pour une interface de lecture rapide particulièrement utile dans une tâche de veille d'informations.

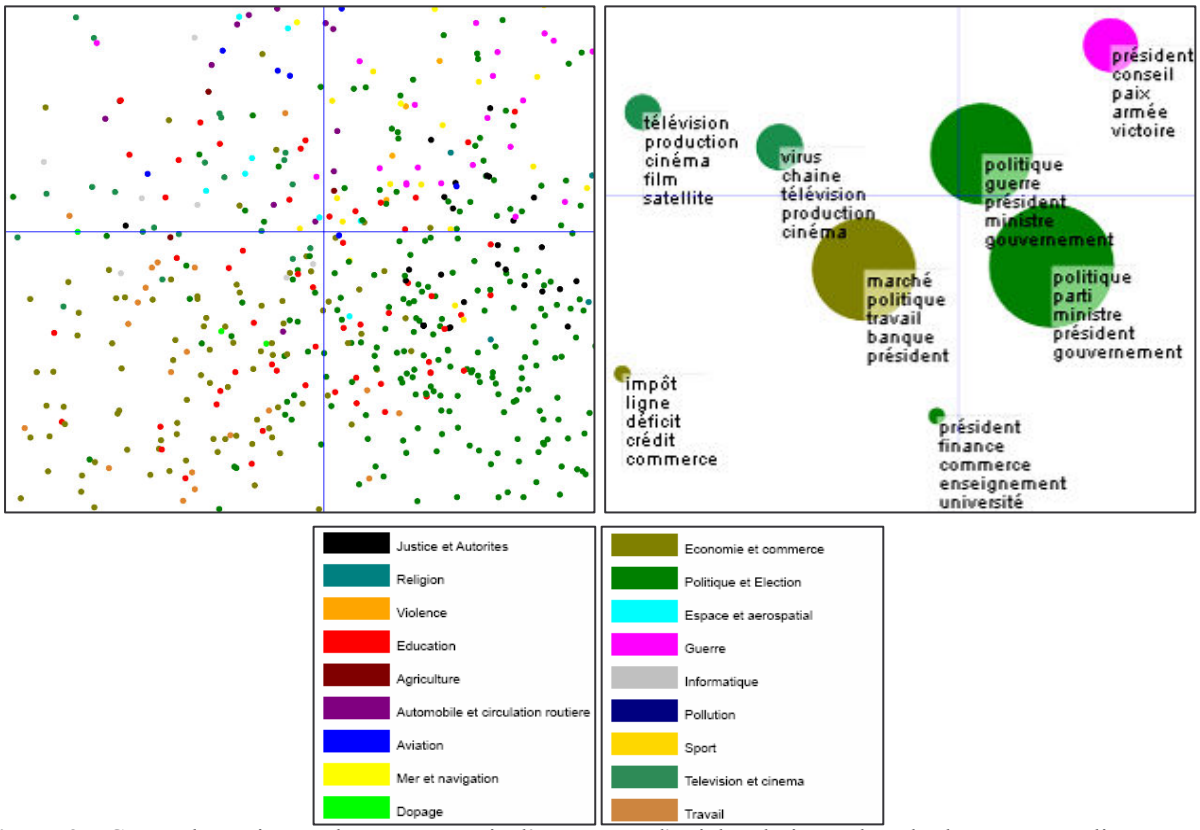

Figure 2 : Cartes thématiques obtenues à partir d'un corpus d'articles de journal et de thèmes généralistes. La carte mettant en évidence les articles du corpus est située en haut et à gauche de la figure, la carte mettant en évidence des groupes de textes est située en haut et à droite de la figure, la légende de couleur de la carte est indiquée sur la partie inférieure de la figure.

La deuxième expérience présentée ici (et détaillée dans (Roy & al., 2005)) consiste à observer un fait de langue : la métaphore conceptuelle au sens de Lakoff et Johnson (Lakoff & al., 1980). Cette observation est faite sur un corpus d'environ 300 articles boursiers issues du journal *Le Monde* entre 1987 et 1989. Les analyses ont porté sur trois métaphores conceptuelles : la *météorologie boursière*, la *guerre économique* et la *santé financière*, un nombre important de ces trois métaphores ayant été observé dans notre corpus d'étude[20]. Les cartes obtenues à l'issue de cette expérience (présentées en figure 3) ont été réalisées avec un ensemble de 3 dispositifs LUCIA représentant les domaines de la météo, la santé et la guerre. Les cartes obtenues nous ont notamment montré une proximité entre des documents contenant des emplois métaphoriques d'un même lexique. Il a aussi été possible d'observer un lien entre le type d'articles (bilans, dépêches, etc.) et les métaphores conceptuelles qui y apparaissaient. De cette manière, nous avons pu déterminer que les métaphores de la *guerre économique* se situaient plutôt dans des dépêches détaillant des événements boursiers ponctuels alors que les métaphores de la *météorologie boursière* et de la *santé boursière* se retrouvaient de manière simultanées dans les bilans boursiers hebdomadaires et mensuels.

[20] Des extraits de notre corpus d'étude illustrant respectivement ces trois métaphores conceptuelles :
- *« Une véritable **tempête** de hausses, alimentée par une marée de capitaux, étrangers pour partie, en quête de placement. »* Le Monde 03/08/87
- *« Le dénouement dans la **bataille** autour de la première banque commerciale privée du pays a eu peu d'effet sur les cours. »* Le Monde 27/02/89
- *« La pente fut longue à remonter, et il fallut bien douze mois pour **panser** les **plaies** du sinistre et à commencer à croire à de nouveaux records d'altitude pour le CAC. »* Le Monde 01/08/89

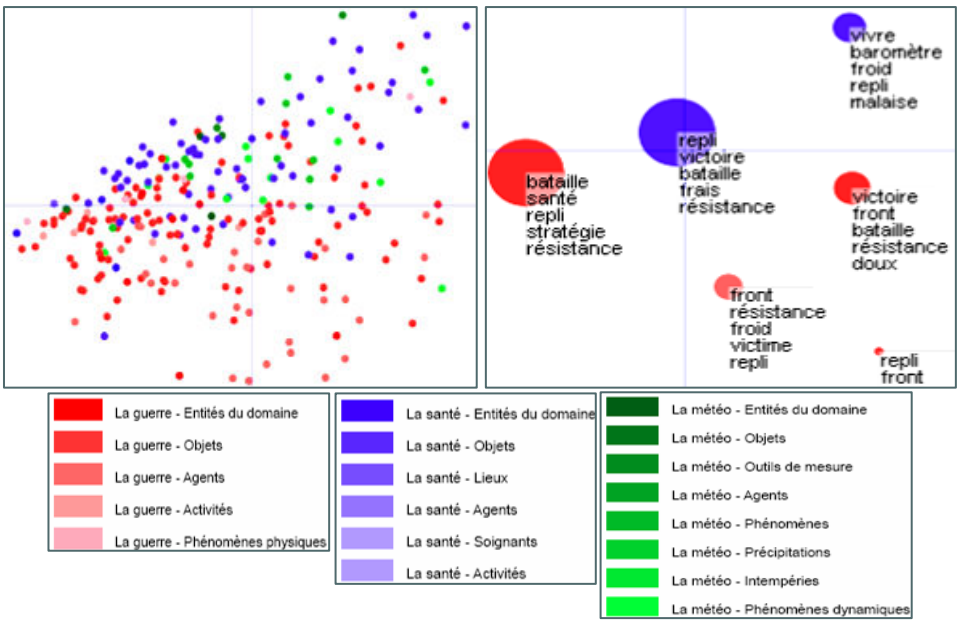

Figure 3 : Cartes thématiques obtenues à partir de notre corpus constitué d'articles boursiers et de dispositifs LUCIA représentant les domaines des métaphores conceptuelles étudiées.

La troisième expérience présentée ici consiste à cartographier un forum de discussion spécialisé portant sur l'apprentissage d'un langage de programmation. Le forum étudié est issu de la plate-forme INES[21]. Il permet à des étudiants de DEUST Technicien des Systèmes d'Information et de Communication d'échanger des messages en rapport avec leur formation. Ce forum est constitué d'environ 200 messages échangés entre le 18/02/2003 et le 27/04/2005 par 27 intervenants différents (enseignants et étudiants). Les cartes de ce forum ont été construites à partir d'un ensemble de 5 thèmes spécialisés que nous avons construits, ces thèmes portant sur l'enseignement, son déroulement, la recherche d'informations, etc. De telles cartes (présentées en figure 4) mettent en évidence les sujets principalement abordés dans les messages. Il est ainsi possible d'observer que la thématique liée au déroulement des enseignements est très majoritaire dans le forum, alors que la thématique liée au contenu même des enseignements est très peu abordée. Ces résultats globaux peuvent alors être des signes pour les enseignants de certaines attentes de leurs étudiants.

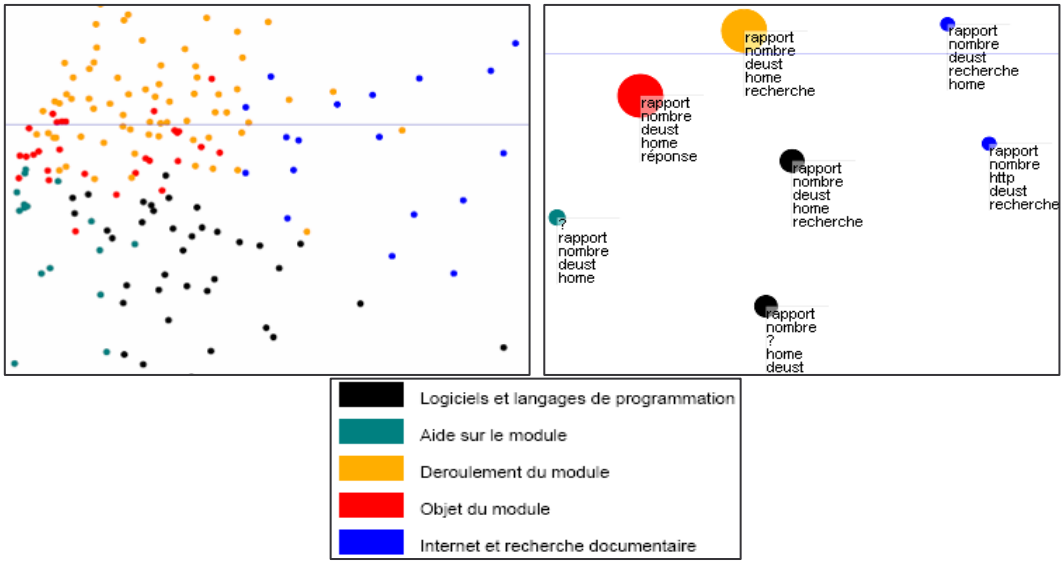

Figure 4: Cartes thématiques obtenues à partir de notre forum de discussion et de thèmes spécialisés.

[21] http://www.dep.u-picardie.fr/ (consultée le 29-06-2006).

La quatrième et dernière expérience présentée ici consiste à cartographier un corpus d'articles scientifiques médicaux à partir de ressources lexicales portant sur ce domaine. Cette expérience est en cours de réalisation et s'inscrit dans une recherche sur la terminologie médicale menée par une chercheuse en TAL de l'université de Rouen (Aurélie Néveol (2005)). La carte présentée en figure 5 illustre les premiers résultats obtenus. Elle met ainsi en évidence des sous-ensembles d'articles et caractérise chacun de ces sous-ensembles par ses « métatermes[22] » les plus occurrents. Lors du passage de la souris sur l'un de ces métatermes, les métatermes identiques caractérisant les autres groupes se mettent en relief permettant ainsi de mettre en évidence les métatermes partagés ou non entre sous-ensembles de documents. Cette mise en relief peut alors être particulièrement utile dans une tâche d'indexation de ces groupes d'articles. L'intérêt ici est que l'indexation peut être guidée principalement par la dimension intertextuelle des documents plutôt qu'uniquement par leur contenu.

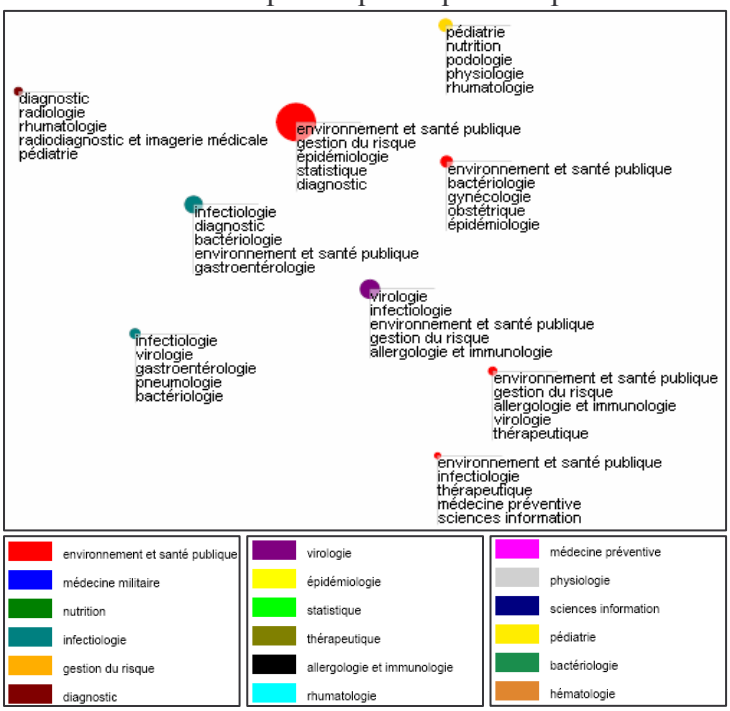

Figure 5: Carte thématique obtenue à partir d'un corpus d'articles médicaux et de ressources terminologiques spécialisées dans ce même domaine.

Perspectives de recherche

Les expériences présentées ci-dessus constituent des premières tentatives de prise en compte de la dimension globale d'un corpus ou d'un flux documentaire dans la contextualisation du sens pour les méthodes d'accès au contenu des documents. Il est clair qu'il faut encore chercher à aller plus loin dans la modélisation de ces méthodes d'accès. L'objectif est la mise au point de modèles opératoires (utilisant des ressources légères) pour les opérations interprétatives et plus particulièrement pour l'afférence. Pour cela il faut continuer à étudier finement des corpus afin d'en dégager différentes formes contextuelles de distribution sémique.

[22] Les métatermes sont des regroupements de mots-clés réalisés en fonction de spécialités médicales, par exemple le métaterme « *ophtalmologie* » regroupe notamment les termes « *œil* », « *myopie* », « *hypermétropie* » (Soualmia & al., 2002).

C'est en partie le travail que nous menons dans le cadre du projet ISOMETA visant la caractérisation d'emplois métaphoriques dans des corpus d'actualité boursière. Dans (Perlerin & al. 2005), auquel nous renvoyons le lecteur pour de plus amples détails, nous avons donné quelques descriptions de telles formes de dynamique sémique contextuelle. Par exemple, dans l'extrait de corpus suivant

> Extrait de l'article n°126
>
> *Le Dow Jones par exemple, le thermomètre de la Bourse de New York, qui avait chuté de 508 points (...).*

la lexie *thermomètre* est source d'une métaphore *in præsentia* dont la cible est exprimée par la lexie *Dow Jones*. Entre ces 2 deux lexies, nous avons montré une isotopie d'un sème indiquant la fonction d'un objet qui sert à l'étude et l'analyse (au même titre que d'autres lexies définies dans nos ressources, *graphique* ou *courbe*, partageant également ce sème). Une telle interprétation rend compte de la nature analogique du lien métaphorique. Une autre forme de dynamique sémique nous a permis de nous prononcer sur une autre interprétation d'une métaphore dans l'exemple suivant :

> Extrait de l'article n°153
>
> *Ce krach était dû (...) à la chute vertigineuse et incontrôlée du dollar, signe que la tempête affecte dorénavant les marchés financiers.*

Dans cet exemple, la lexie *tempête* est source d'une métaphore *in absentia* dont la cible n'est donc pas dans le cotexte. Entre *tempête* et *krach* nous avons montré une isotopie indiquant quelque chose évalué comme *mauvais*. De plus le contenu sémique de *tempête* porte un sème potentiellement partageable par plusieurs domaines thématiques qui indique un phénomène *violent*. Cette nature partageable du sème permet de l'actualiser dans le cotexte. Nous en avons déduit cette fois une caractérisation de l'aspect créatif du lien métaphorique.

D'autres pistes de recherche peuvent aussi être explorées pour tenter de rendre compte de mécanismes d'afférence sans pour autant nécessiter le recours à de vastes ressources. Notamment, nous cherchons à l'heure actuelle à exploiter l'importance relative d'une isotopie par rapport à une autre au sein d'un document. L'idée principale est de « positionner » les textes et les groupes de textes (déterminés automatiquement par la plate-forme ProxiDocs par exemple) par rapport aux éléments de plus haut niveau les englobant (respectivement, groupes de textes et corpus) en tenant compte de domaines représentés par l'utilisateur selon le modèle LUCIA. Ce positionnement d'une entité textuelle par rapport à une entité textuelle plus globale la contenant permet d'obtenir des informations pertinentes sur la répartition et la localisation des domaines représentés en corpus. Pour aller plus loin dans de telles analyses, nous tenons compte des particularités du niveau global (que l'on pourrait appeler « signaux forts ») à un niveau plus local (où prennent place ce que l'on pourrait appeler des « signaux faibles »). Pour mettre en œuvre cette prise en considération du niveau global, nous pondérons les classements des isotopies que nous retrouvons d'un document à l'autre selon les deux critères suivants :

- si dans le groupe et dans le corpus, une même isotopie est très présente, alors on diminue son importance dans le groupe (atténuation du signal fort) ;
- si au contraire, dans le groupe, une isotopie est présente et qu'elle l'est moins dans le corpus, alors on augmente son importance (amplification du signal faible).

Ces deux conditions permettent de faire ressortir des propriétés des groupes masquées par les propriétés globales du corpus dont chaque texte hérite. Ce positionnement des groupes par rapport aux corpus aide ainsi à identifier comment les ressources sont projetées sur les différents paliers textuels (texte et groupe) du corpus analysé et en quoi elles permettent de différencier et d'isoler des groupes et des textes.

Ces quelques heuristiques de prise en compte de la dimension globale sont encore au stade de l'expérimentation au sein de l'outil ProxiDocs. L'enjeu est maintenant de les évaluer. Cette évaluation ne pourra être limitée à une quantification technique des résultats produits car, en

tant qu'outil individu-centré, ProxiDocs doit faire l'objet d'évaluations extrinsèques au sens de (Spark Jones & al., 1995), c'est-à-dire des évaluations construites sur le recueil et l'analyse des avis des utilisateurs. Il s'agira donc d'une certaine façon de caractériser si l'apport de la dimension globale d'un corpus sur l'accès au contenu des documents est consensuel ou au contraire fait l'objet de variations interpersonnelles.

Conclusion

Plus que jamais, notre objectif reste de poursuivre nos travaux sur la cartographie personnalisée de corpus pour aller toujours plus loin dans la modélisation et l'amélioration des Interactions Homme-Machine (IHM) où un rapport sémiotique au langage et aux textes est central. Dans ce genre d'IHM, nous préférons de loin l'idée d'une instrumentation du sens à celle de la construction du sens. D'une manière imagée, il nous semble qu'un traitement sémantique informatisé a plus de points communs avec un outil tel un microscope par exemple, c'est-à-dire quelque chose qui nous montre ce qui est déjà là mais qu'on ne voyait pas de cette façon, qu'avec un outil tel un transformateur électrique qui produirait à partir de quelque chose une autre chose qui n'existait pas avant. Il s'agit donc de considérer que le sens tel que le produit une interprétation humaine n'est pas à la portée d'un seul traitement informatique. Cela ne veut pas dire pour autant que les machines ne puissent pas interpréter des textes. Elles le font à leur manière comme les sujets humains le font aussi à leur manière. Nous opposons ici l'idée d'une Interprétation Calculatoire (IC) à celle d'une Interprétation Humaine (IH). Ces deux formes d'interprétation ne sont pas en concurrence car l'IC n'a en aucun cas le but de supplanter l'IH. Au contraire, nous les pensons comme complémentaires dans le sens où une interprétation calculatoire a pour objectif de produire dans l'interaction des traces qui vont participer aux interprétations humaines du ou des utilisateurs.

Comme le calcul rapide des opérations numériques complexes ou encore le traitement immédiat de vastes corpus sont inaccessibles aux capacités cognitives humaines, les finesses de sens (par exemple celles que l'on trouve dans un bon nombre de formes d'humour ou d'interprétations littéraires) ainsi que les mises en relation diverses et variées bien spécifiques à l'interprétation humaine ne sont pas, pour la majeure partie, modélisables dans le cadre d'une interprétation calculatoire. La recherche de ce qui est à la limite des deux formes d'interprétation et qui peut se prêter à une modélisation dans le cadre d'une interprétation calculatoire suscite à l'évidence bien plus de questions que de réponses. La question des rapports entre le global et le local dans la contextualisation du sens nous paraît notamment se situer exactement sur cette limite. L'amélioration des compétences sémiotiques et interactionnelles des machines constitue donc en tout cas un domaine de recherche à part entière au croisement de plusieurs disciplines et pas uniquement un domaine d'ingénierie. En cela, à l'encontre de Rastier (Rastier, 2005, p. 41), nous militons pour une scientificité propre des traitements sémantiques et plus largement des traitements automatiques des langues.

Bibliographie

ASHER N., 1993, *Reference to Abstract Objects in Discourse*, Dordrecht, Kluwer.

BERNERS-LEE T., 1998, What the Semantic Web can represent ?, W3C, http://www.w3.org/designissues/rdfnot.html (consultée le 29-06-2006), MANN, W.C., & THOMPSON, S.A.

BEUST P., 1998, *Contribution à un modèle interactionniste du sens*, Thèse de doctorat en Informatique, Université de Caen Basse-Normandie.

BEUST P., 2002, Un outil de coloriage de corpus pour la représentation de thèmes : 6èmes *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*.

BOURIGAULT D., AUSSENAC-GILLES N., 2003, Construction d'ontologies à partir de textes, Actes de *Traitement Automatique des Langues Naturelles (TALN)*, Tome 2, pp. 27-47.

CELEX, 1998, http://www.ldc.upenn.edu/readme_files/celex.readme.html (consultée le 29-06-2006), UPenns, Eds., Actes de *Consortium for Lexical Resources.*

CHARLET, J., LAUBLET, P., REYNAUD, C., 2003, *Web Sémantique*. Rapport de l'Action Spécifique 32 CNRS / STIC. V3. http://rtp-doc.enssib.fr/IMG/pdf/ASWebSemantique2003.pdf (consultée le 29-06-2006).

CLAVEAU V., 2003, *Acquisition automatique de lexiques sémantiques pour la recherche d'information*, Thèse de doctorat en Informatique, Université de Rennes 1.

CONDAMINES A. (dir.), 2005*, Sémantique et corpus,* Hermès, Paris, ISBN : 2-7462-1055-X.

COURSIL J., 2000, *La fonction muette du langage*, Ibis Rouge Editions, Petit-Bourg (Guadeloupe), ISBN 2-84450-090-0.

JACQUEMIN C., BUSH C., 2000, Fouille du Web pour la collecte d'entités nommées, Actes de *Traitement Automatique des Langues Naturelles (TALN)*, Lausanne 187-196.

KAMP H., 1981, A theory of truth and semantics representation. In *Formal Methods in the Study of Language* sous la direction de GROENENDIJK, JANSEN & STOKHOF, Amsterdam, Mathematical Centre Tracts.

LAKOFF G., JOHNSHON M., 1980, *Metaphors we live by*. University of Chicago Press, Chicago, U.S.A.

LAVENUS K., LAPALME G., 2002, Évaluation des systèmes de question réponse, revue *Traitement Automatique des Langues*, vol. 43, n°3/2002, p. 181-208.

NÉVÉOL A., 2005, *Automatisation des tâches documentaires dans un catalogue de santé en ligne*, Thèse de doctorat en Informatique, INSA de Rouen.

NICOLLE A., 2005, Comparaison entre les comportements réflexifs du langage humain et la réflexivité des langages informatiques, Actes des 12e journées de Rochebrune « *Réflexivité et auto-référence dans les systèmes complexes* », pp. 137-148.

PERLERIN V., 2002, Memlabor, un environnement de création, de gestion et de manipulation de corpus de textes. Actes de *TALN-RECITAL* dans le cadre des *Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, Tome 1, pp. 507-516.

PERLERIN V., 2004, *Sémantique légère pour le document. Assistance personnalisée pour l'accès au document et l'exploration de son contenu*, Thèse de doctorat en Informatique, Université de Caen Basse-Normandie.

PERLERIN, V., BEUST, P., FERRARI, S., 2005, Métaphores et dynamique sémique, In *La Linguistique de Corpus*, sous la direction de G. WILLIAMS, Rennes, Presses universitaires de Rennes, ISBN 2-7535-0046-0, pp. 323-336.

RASTIER F., 1987, *Sémantique interprétative*, Paris, Presses Universitaires de France.

RASTIER F., 1998, *Le problème épistémologique du contexte et le problème de l'interprétation dans les sciences du langage*, *Langages*, n°129, pp.97-111.

RASTIER F., 2001, *Arts et Sciences du texte*, Paris, Presses Universitaires de France.

RASTIER F., 2005, Enjeux épistémologiques de la linguistique de corpus, In *La Linguistique de Corpus*, sous la direction de G. WILLIAMS, Rennes, Presses universitaires de Rennes, ISBN 2-7535-0046-0, pp. 31-45.

RASTIER F., CAVAZZA M., ABEILLE A., 1994, *Sémantique pour l'Analyse*, Paris, Masson.

ROY T., 2005, Une plate-forme logicielle dédiée à la cartographie thématique de corpus, Actes de TALN-RECITAL 2005 dans le cadre des *Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pp. 545-554.

ROY T., BEUST P., 2004, ProxiDocs, un outil de cartographie et de catégorisation thématique de corpus, Actes des 7èmes *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, pp. 978-987.

ROY T., FERRARI S., BEUST P., 2005, Cartographie thématique des corpus pour l'étude des métaphores, Actes des *Journées de Linguistique de Corpus (JLC)*, WILLIAMS G. Ed., à paraître.

SOUALMIA L.F., BARRY-GREBOVAL C. ABDULRAB H & DARMONI S.J., 2002, Modélisation et représentation des connaissances dans un catalogue de santé, Actes de *Ingénierie des Connaissances (IC)*, pp. 139-149.

SPARK-JONES K. & GALLIERS J. R., 1995, *Evaluating Natural Language Processing Systems: An Analysis and Review*. Number 1083 in Lecture Notes in Artificial Intelligence, Springer.

THLIVITIS T., 1998, *Sémantique Interprétative Intertextuelle*. Thèse de doctorat en Informatique, Université de Rennes I.

VALETTE M. & GRABAR N., 2004 Caractérisation de textes à contenus idéologiques : statistique textuelle ou extraction de syntagme ? l'exemple du projet PRINCIP, Actes des 7èmes *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, pp 1107-1117.