Chapitre 1

Vers une *sémantique légère* pour le TAL

Cette thèse a pour but la réalisation d'outils informatiques pour l'assistance personnalisée à l'accès aux documents numériques textuels et l'exploration de leur contenu. Notre étude participe au courant de recherche qui tend à transformer l'ordinateur en un média à valeur ajoutée pour la manipulation de textes. D'une manière abusive, l'utilisation du terme *document* fera référence dans l'ensemble de ce tapuscrit au document numérique textuel ou majoritairement textuel.

Ce chapitre introductif présente un aperçu général de nos travaux. Afin de provoquer l'attention du lecteur, nous commençons par répondre brièvement aux trois questions fondamentales qui président habituellement à ce type de présentation : quoi (Un modèle de ressources lexicales personnalisées - 1.1) ? pourquoi (Définition des objectifs - 1.2) ? et comment (Démarche - 1.3) ? Nous justifions en particulier notre approche interactionniste de la sémantique computationnelle qui a comme conséquence la minimisation des données, la simplification des traitements et la personnalisation des résultats obtenus de la machine. Ces caractéristiques placent nos travaux dans l'optique d'une *sémantique légère* pour le TAL (1.4) qui nous permet de tendre vers des propositions multilingues. Ce chapitre n'a bien sûr pas le but de lever entièrement le voile sur l'ensemble de l'étude mais de fournir les informations nécessaires à la compréhension de l'ensemble de notre démarche et par la même occasion, de présenter le plan de ce mémoire (1.5).

1.1 Un modèle de ressources sémantiques personnalisées

Dans le chapitre *Le Language Analytique de John Wilkins*[1] de ses *Autres Enquêtes :1937-1952*[2], J.L. Borges (1899-1986) fait référence à une encyclopédie chinoise intitulée *Marché Céleste des Connaissances Charitables* qui présenterait la classification des animaux suivante :

a. ceux qui appartiennent à l'empereur ;
b. les embaumés ;
c. ceux qui sont entraînés ;
d. les porcs ;
e. les sirènes ;
f. les fabuleux ;
g. les chiens perdus ;
h. ceux qui sont inclus dans cette classification ;
i. ceux qui tremblent comme s'ils étaient fous ;
j. les innombrables ;
k. ceux dessinés avec un très fin pinceau en poils de chameau ;
l. d'autres ;
m. ceux qui viennent de casser une cruche ;
n. ceux qui ressemblent de loin à des mouches.

Cette classification peut paraître choquante à bien des égards et a suscité un nombre très important de commentaires depuis sa publication ; les plus fameux sont certainement ceux de Michel Foucault [Foucault, 1966] qui y voit le lieu de naissance de son ouvrage *Les mots et les choses*. Outre le fait que notre culture scientifique nous amène spontanément à rejeter toute ambiguïté et que certaines catégories ne sont pas définies en fonction de propriétés mais simplement par leur dénomination courante (comme les sirènes ou les porcs par exemple), le plus remarquable est que cette classification ignore résolument la filiation aristotélicienne de l'*Organon* [Aristote], où l'ordre proposé se fait par genres et différences exclusives. L'arbre de Porphyre n'est ici d'aucune aide car les attributs de différenciation (lorsqu'ils existent) ne sont pas exclusifs. Le choix des instances de chaque catégorie est un facteur de trouble supplémentaire : où classer le porc Brille Babil qui vient de casser une cruche et qui vraisemblablement peut ressembler à une mouche si on l'observe d'assez loin ? La classification de Borges ne correspond à aucun modèle connu car elle ne propose non plus aucun prototype et ne présente aucune condition nécessaire et suffisante pour les classes proposées. Cette classification est cer-

[1] Le texte intégral en anglais et en espagnol est présent à l'adresse suivante : http://www.crockford.com/wrrrld/wilkins.html

[2] Borges, J.L., 1964, *Other Inquisitions: 1937-1952*, 223 p., Austin, Texas, University of Texas Press, ré-édition de 1975, 0292760027.

tes incongrue, mais elle pourrait trouver une justification pratique. Nous nous abstiendrons d'en faire la démonstration ici, mais il est fort probable que quelques oulipiens éclairés pourraient sans peine s'y exercer (Perec en fait la preuve dans *Penser / Classer* [3]). Au-delà de l'anecdote, les propos de Borges ou de Perec révèlent une propriété inhérente à nos pratiques : il est tout à fait envisageable de concevoir des catégorisations éminemment dépendantes d'un point de vue particulier. Ce point de vue dépend de la pratique dans laquelle s'inscrit l'acte langagier ; tout est lié à la manière dont on parle d'une tâche ou d'un sujet. C'est l'utilisation particulière de la langue dans une situation donnée qui prévaut et non le monde tel qu'il est, ou qu'il nous semble être ; la langue est ici conçue en action dans le discours d'un locuteur, d'un écrivant et dépend, en outre, de l'interprétation du lecteur ou de l'interlocuteur. Certains exemples analogues dépassent le cadre de la littérature. Dans son discours sur l'état de l'Union de janvier 2002[4], le Président des États-Unis d'Amérique G.W. Bush distinguait deux types de pays dans le monde : ceux de « l'axe maléfique » (*an axis of evil* i.e. la Corée du Nord, l'Iran et l'Irak – il leur adjoindra plus tard la Syrie) et les autres. Ces propos n'ayant fait l'objet d'aucun consensus international, ils ont suscité à l'époque un déluge de commentaires. L'expression a été maintes fois réutilisée depuis[5] ; elle a naturellement induit des interprétations différentes selon le lectorat ou l'auditoire et selon les contextes de ses apparitions. La catégorisation proposée par le président G.W. Bush ne s'appuie pas sur une référence au monde universellement partagée[6] mais trouve cependant une légitimité pour son auteur (sa fonction n'y étant pas étrangère). Elle s'est inscrite durablement dans son discours et ceux de ses partisans dans l'affaire en question. L'intertexte de ces discours peut être interprété à la lumière du premier, non seulement en fonction des références explicites vers les pays désignés, mais également en fonction du manichéisme et de la charge morale que l'on a pu déceler dans les propos. C'est ainsi que certains opposants au régime étasunien se sont appropriés l'expression, n'en gardant que l'aspect moral et dual et en détournant les références explicites au monde[7].

Les champs d'application des disciplines du TAL (Traitement Automatique des Langues) se sont considérablement élargis avec la multiplication sans cesse croissante des besoins informatiques

[3] Perec, G., 1985, *Penser/Classer*, Col. La Libraire du XXI siècle, 175 p., Paris, Seuil, 2020587254.

[4] « *President focuses on War on Terrorism, Homeland Security, Jobs – President Bush's State of the Union Address* » 29 janvier 2002 - http://usinfo.state.gov/topical/pol/terror/02012914.htm. L'expression « axe maléfique » provient de la traduction officielle du document par le département d'état des programmes d'information internationale des États-Unis d'Amérique (http://usinfo.state.gov/francais/f0213001.htm).

[5] Voir par exemple l'article de M. Reynolds « "*Axis of Evil*" *Rhetoric Said to Heighten Dangers* " » dans le *Los Angeles Times* du 21 janvier 2003 - http://www.commondreams.org/headlines03/0121-03.htm.

[6] « "It was harmful both conceptually and operationally," said Graham Allison, government professor and former dean of the Kennedy School of Government at Harvard University. "Conceptually, the 'axis' suggested a relationship among the entities that doesn't exist. " ».- *ibid*.

[7] Voir par exemple l'article de A. de Borchgrave *Fissures in the House of Saud* dans *The Washington Times* du 19 janvier 2004 - http://www.washtimes.com/commentary/20040120-085115-5893r.htm

pour l'accès aux documents numériques et à leur contenu. Les discours de G.W. Bush ou les textes de Borges ne sont que de minuscules îlots dans l'océan documentaire électronique. Les textes à traiter ne sont plus uniquement les documents techniques et scientifiques dont l'univocité est souvent la principale caractéristique. Les contenus de ce type de document, lorsqu'ils ne sont pas formatés (au sens informatique du terme), ne sont d'ailleurs jamais triviaux à traiter : des pratiques diverses peuvent amener à les interpréter différemment, les besoins informationnels ne sont pas identiques selon la profession, la pratique ou l'utilisateur. Nous avons pu apprécier les exigences des industriels dans ce domaine, en particulier au cours de nos rencontres avec les dirigeants du CRITT BNC[8]. Dans un cadre de veille technologique tel que le propose cette entreprise, les tâches consistent à réaliser des dossiers techniques présentant l'état de l'art d'un sujet donné à un instant précis. On utilise pour cela des bases de données bibliographiques à partir desquelles il est aisé de produire automatiquement des listes d'articles intéressants ou des réseaux de collaboration puisque les documents de ces bases sont annotés, dûment indexés et formatés pour leur exploitation automatique. Lorsqu'il s'agit de traiter des documents non préparés à l'avance pour ce type de tâche comme les pages de sites de l'Internet par exemple, le travail est assisté par la machine à l'aide des techniques classiques de recherche d'information. Cependant, l'exploitation des documents jugés pertinents par ces systèmes doit se faire le plus souvent à la main faute de ressources et de traitements adaptés aux besoins précis de la tâche en cours. Un des partenaires du CRITT BNC est une entreprise spécialisée dans la biocorrosion et la biodétérioration des matériaux. Accéder à tous les documents qui traitent de ce sujet comme veut le voir aborder cette entreprise n'est pas une tâche aisée : l'entreprise n'est intéressée que par certaines applications du domaine et n'a, par exemple, que faire de sites de vulgarisation ou ne faisant qu'aborder succinctement le sujet dans un but promotionnel pour une solution déjà largement connue. Les documents rapatriés sont parfois très longs et les moyens d'en accélérer le parcours, d'accéder aux parties réellement intéressantes sont difficiles à mettre en œuvre. Il y a bien une dimension personnelle - ici au client de l'entreprise - pour la recherche documentaire à effectuer.

Plus largement, le document numérique et les techniques qui le médiatisent font émerger de nouveaux besoins. Ces derniers concernent en particulier l'usager et la place qui lui est réservée dans les systèmes et cela, à tous les stades de manipulation et pour tous les publics intéressés par les documents et leur contenu. La numérisation a non seulement bouleversé les habitudes documentaires des entreprises et des particuliers, mais aussi certaines pratiques scientifiques : parmi ceux-ci, les linguistes sont particulièrement concernés par le renouveau de la linguistique de corpus informatisée. Au passage, l'informatisation de la science a également changé l'échelle de ses données empiriques. Qu'il s'agisse des tâches traditionnelles de la linguistique – terminologie, analyse de faits de langue, etc. –

[8] Centre Régional d'Innovation et de Transfert de Technologie de Basse-Normandie Cotentin.

* Les citations suivies d'une astérisque apparaissent plusieurs fois au sein du tapuscrit.

ou que soient désormais concernées de nouvelles utilisations de l'accès à l'information dans les entreprises ou chez les particuliers, on constate la nécessité croissante de fournir des moyens d'assistance personnalisée aux utilisateurs (voir par exemple [Nazarenko et Hamon, 2002 : 15] en terminologie, [Bourigault et Fabre, 2000 : 135] en syntaxe ou encore [Polity, 1999] pour une approche plus générale autour des « sciences de l'information »). Comme il est souligné dans [Valette, 2004 : 229], les situations d'utilisation des documents numériques par les utilisateurs se sont à ce point multipliées et complexifiées qu'à l'heure actuelle, ce ne sont plus d'outils de recherche dont l'usager a besoin, ce sont d'outils d'aide à l'interprétation. Depuis quelques années, certains travaux ont amené à ne plus considérer le sens des documents comme une donnée mais comme un processus d'interprétation et d'appropriation (parfois qualifiée de cognitive comme dans [Polity, 1999*]) propre à un individu ou un groupe donné. Dans ce nouveau paradigme, l'utilisateur est envisagé comme un acteur social qui, dans le cadre d'une activité, fait appel à des systèmes informatiques pour lui faciliter, non seulement l'accès au document, mais aussi l'accès à son contenu. Nos travaux se placent résolument dans cette optique qui place au cœur des préoccupations la personnalisation des traitements et les modalités d'interactions qui leur sont nécessaires.

À la question quoi (que veut-on faire), nous répondrons donc : un modèle informatique permettant la réalisation d'applications logicielles qui allouent à un utilisateur ou un groupe d'utilisateurs, la possibilité d'exprimer un point de vue sur une tâche pour obtenir en retour des services personnalisés. Ces services relèvent de l'assistance à l'accès à des documents numériques textuels et à l'exploration de leur contenu.

1.2 Définition des objectifs

1.2.1 Accès aux documents

Les techniques d'accès aux documents numériques sont encore à l'heure actuelle dominées par l'indexation dont on estime qu'elle a été permise par la disparition du *volumen* au profit du *codex* introduit en occident vers les premiers siècles de notre ère. La forme moderne des index remonterait à un ouvrage de 1682 pour l'édition du catalogue de la bibliothèque de Lamoignon [Fayet-Scribe et Canet, 1999]. En machine, l'indexation subit certaines des contraintes du support papier. Dans l'index, le document doit être réduit à un nombre restreint d'informations structurées. Ces informations doivent être les plus pertinentes possibles pour retrouver rapidement un document en fonction d'un ensemble de critères. Dans les bibliothèques, les Centres de Documentation et d'Information ou les autres bases

de données bibliographiques (marché dominé actuellement par la société Sirsi Corporation[9]), ces informations sont généralement le titre, le nom du ou des auteurs, la date de publication du document, le domaine auquel on peut le rattacher, etc. ; l'indexation s'effectue principalement à la main. Le principal apport de l'informatique dans le domaine est l'automatisation des tâches de gestion des ensembles documentaires et la possibilité d'interroger rapidement et plus ou moins facilement des index de plusieurs milliers de documents. Notons au passage, que certaines études dans des domaines spécifiques montrent que près de 30% des utilisateurs de tels systèmes les trouvent compliqués ou mal faits et ne savent pas définir correctement leur recherche ni utiliser les fonctionnalités de bases de ces logiciels[10]. Parfois, les informations contenues dans l'index sont des ensembles de mots-clefs censés représenter « le contenu » des documents, et les discriminer les uns par rapport aux autres dans l'index. Des outils informatiques existent pour automatiser l'indexation à l'aide de mots-clefs. Ces systèmes ne s'avèrent véritablement satisfaisants qu'agrémentés de techniques spécifiques ou lorsqu'ils sont destinés à des domaines restreints réservés à des spécialistes comme le montrent par exemple le succès de Citeseer.IST la bibliothèque numérique spécialisée dans les publications scientifiques en informatique. Les résultats obtenus au GREYC dans des systèmes de gestion de documents géographiques [Turbout, 2002] montrent que les solutions qui semblent le mieux répondre aux attentes des utilisateurs sont celles qui sont destinées à des usagers producteurs de mêmes types de documents que ceux auxquels ils désirent avoir accès.

Dans les systèmes majoritairement automatiques, les documents sont explorés par des programmes qui en extraient des mots-clefs et qui, à l'aide d'opérations statistiques ou probabilistes, en évaluent la pertinence pour l'index. C'est par exemple le cas des systèmes classiques de recherche documentaire (ou de recherche d'information) comme les moteurs de recherche de l'Internet. L'accès au document est alors vu comme la mise en correspondance d'une représentation d'un document avec une représentation d'une requête par l'intermédiaire d'un index. Il existe plusieurs modèles d'appariement entre la requête et l'index, tous plus ou moins inspirés des modèles suivants :

- Le modèle booléen basé sur l'algèbre de Boole [Frakes et Baeza-Yates, 1992] se fonde uniquement sur la présence des mots considérés comme isolés. Aisé à mettre en place, il est difficile d'accès pour les utilisateurs tant au niveau de la formulation des requêtes que pour l'interprétation des résultats fournis par les systèmes.
- Le modèle vectoriel : chaque document est représenté par un vecteur de mots à partir d'une approche *bag of words* (ou sacs de mots). Combiné avec l'approche booléenne [Salton et

[9] Les produits informatiques de la société Sirsi Corporation sont utilisés pour la gestion et l'accès à plus de 10 000 bibliothèques dans le monde (dont celle de l'Université de Caen) – http://www.sirsi.com.

[10] http://www.educagri.fr/renadoc/rapport02/sommaire.htm - « Evaluation de l'usage des bases documentaires par les utilisateurs des CDI des établissement d'enseignement agricole » Comité National d'Orientation 2001-2002.

MacGill, 1983], de tels systèmes permettent des interrogations plus aisées mais la linéarité des documents n'est pas davantage prise en considération et les moyens d'expression des besoins des utilisateurs restent restreints.

- Le modèle probabiliste [Robertson et Spark Jones, 1976] s'appuie sur le modèle vectoriel (et souffre donc des même manques) ; il utilise en outre des notions probabilistes plutôt que statistiques pour calculer la pertinence de chaque index pour un document donné. Le but du modèle est de fournir à l'utilisateur des documents ayant la plus grande probabilité d'être en rapport avec la requête de l'utilisateur.

On pourra se reporter à [Turbout, 2002*] pour un aperçu très complet de toutes ces techniques. À l'heure actuelle, celles-ci apportent des solutions valables pour l'accès aux documents dans de très grands ensembles. Leur efficacité est accrue de façon importante lorsqu'elles sont associées à d'autres techniques, c'est par exemple le cas avec le *PageRank* [Page *et al.*, 1998]. Cette méthode de classement se base sur le comptage des liens entrant et sortant pour les pages Internet et fait actuellement le succès planétaire du moteur Google. Mais les détournements de *PageRank* et les dommages dus au *spam indexing*[11] sont symptomatiques de l'inaptitude de ces systèmes à accéder au contenu réel des documents, à prendre en considération les informations utiles pour la tâche de l'utilisateur. Le document n'est conçu que comme un sac de mots auquel on a parfois ajouté des méta-données et son contenu réduit à un index structuré[12]. Dans l'article intitulé *Anti-war slogan coined, repurposed and Googlewashed… in 42 days* paru dans *The Register* (SF, CA) du 3 avril 2003, le journaliste Andrew Orlowski rapporte un détournement avéré de *PageRank* autour de l'expression *the second superpower*. Celle-ci désignait initialement l'opinion publique mondiale contre la seconde guerre en Irak et a été détournée et placée en haut des résultats de Google par le biais d'une campagne de mise en ligne de liens hypertextes vers un article où *second superpower* réapparaissait dans une version édulcorée[13].

[11] Le *Spam Indexing* est l'acte qui consiste à insérer des noms de marque ou de personnes dans les méta-données d'un site pour en augmenter artificiellement le nombre de visiteur. De façon plus générale, l'insertion de mots très populaires dans les requêtes de l'Internet comme *météo*, *humour*, *jeux* et bien sûr *sexe* et *porno* dans les méta-données d'une page sans que cela ne corresponde véritablement à son contenu relève du *spam indexing*. Par exemple, en juin 2000, les mots-clefs utilisés dans les méta-données du site de l'une des candidates au poste de Maire de Paris étaient les suivants : *francoise*, *gaulliste*, *séguin*, *seguin*, *chirac*, *panafieu*, *rpr*, *rassemblement pour la république*, *2001*, *député*, *pamela anderson*, *municipales*, *paris*, *élections*, *arrondissements*, *candidat*, *politique*, *bertrand delanoë*, *affaires*, *tiberi*… Il n 'est pas nécessaire de désigner l'intruse.

[12] Certains algorithmes pour le calcul de *PageRanks* « sensibles au contexte » sont à l'étude depuis plusieurs années. Le « *Topic Sensitive Pagerank* » consiste par exemple à donner plus d'importance à des pages dont on connaît la thématique de départ. Le contenu des documents par rapport à la lecture d'un usager n'en est pas moins négligé [Haveliwala, 2002].

[13] Le détournement du principe de *PageRank* est même devenu l'enjeu de concours. Au cours du premier semestre 2004, il s'agissait d'être classé premier dans la liste de résultats fournie à partir de la requête « *mangeur de Cigogne* » (http://concours.promo-web.org/).

Les deux textes avec l'expression *second superpower*, comme tous les documents répertoriés par le moteur, ont été explorés par des robots qui ont repéré les chaînes de caractères des parties remarquables de leur structure (entête, adresse, titre, corps voire images avec les balises type `alt` en HTML et méta-données avec les balises type `meta` en HTML). La figure 1 et la figure 2 présentent de manière simplifiée le principe d'indexation utilisé.

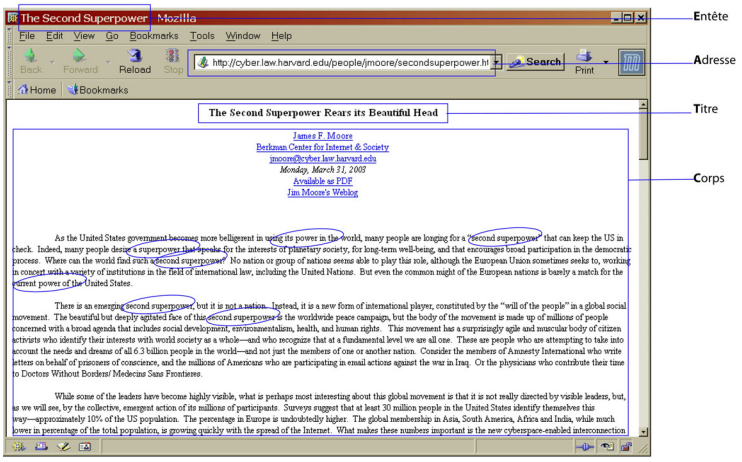

Figure 1 – Repérage des termes pour la création d'un index.

Index de la page http://cyber.law.harvard.etu/...

	Entête	Adresse	Titre	Corps
second	1	1	1	49
superpower	1	1	1	58
war	0	0	0	8
...				

Index inversé pour "superpower"

	Entête	Adresse	Titre	Corps	estimation du PageRank normalisé
http://cyber.law.harvard.etu/...	1	1	1	58	7/10
http://www.theregister.co.uk/...	0	0	0	10	6/10
http://www.worldchanging.com/	1	1	1	25	5/10
...					

Figure 2 – Index et index inversé simplifiés[14].

Dans cet exemple, nous voyons que la linéarité du texte et sa structure sont abandonnées au profit de l'index et de l'index inversé qui ne permettent pas d'anticiper pleinement le contenu des documents puisque le sens des termes en contexte n'est pas pris en considération. L'approche *bag of words* néglige la linéarité du texte et l'interprétation que l'on peut en faire en fonction des besoins de la tâche documentaire en cours. Les systèmes qui la mettent en œuvre ne permettent pas une personnalisation des services d'accès aux documents à proprement parler : l'utilisateur n'est considéré que comme le producteur d'une requête, la plupart du temps limitée à une suite de mots-clefs. Les services rendus par ces systèmes sont appréciables car ils permettent de discriminer un ensemble de résultats parmi les très nombreux documents connus du système mais l'absence d'analyse du contenu véritable des documents amène de nombreux problèmes. En particulier, les listes de résultats sont soit très longues à explorer (de l'ordre de la centaine, voire du millier de documents pour une seule requête) et

[14] Le PageRank des pages a été obtenu à partir des indications de la *Google ToolBar*.

présentent de nombreux documents inadéquats pour l'usager (c'est ce qu'on appelle le bruit), soit très courtes voire vides et dans ce cas l'utilisateur ne peut satisfaire les besoins documentaires en rapport avec sa tâche (c'est qu'on appelle le silence).

En réaction, certains chercheurs reviennent actuellement vers des principes anciens de l'IA et du TAL : la constitution de bases de connaissances conceptuelles et la réalisation de terminologies. C'est le cas pour le projet du « Web Sémantique » (WS) qui consiste à annoter les documents avec des méta-données structurées manipulables par la machine. Le WS vise à un abandon (partiel) des méthodes de calculs d'occurrences et d'indexation à partir de mots-clefs et préconise un retour à la « représentation des connaissances ». Ces solutions nécessitent des données en quantité phénoménale et des processus d'inférence toujours plus complexes. Cette complexité n'est pas sans poser problème aux utilisateurs, qui se voient souvent réduits à faire des suppositions quant aux fonctionnements de tels systèmes pour tenter de les exploiter au mieux. Nous verrons plus loin d'autres limites de ces techniques qui montrent que la problématique du traitement des documents et de leur contenu par les machines semble appelée à connaître quelques inflexions théoriques. Nos travaux ont pour but d'aider à l'émergence de ces inflexions en particulier en plaçant l'utilisateur au cœur du système. Il s'agit d'allouer à l'usager des moyens d'interagir avec la machine pour obtenir des services personnalisés pour non seulement, accéder aux documents mais aussi, être assisté pour l'exploration de leur contenu en fonction de préoccupations propres. Ces mécanismes nécessitent de concevoir les textes dans leur globalité et leur linéarité. Une de nos hypothèses de travail est que l'analyse du contenu des documents est un préalable indispensable à la mise en oeuvre de moyens d'accès à des derniers. Le contenu d'un document semble alors n'être véritablement intéressant dans cette optique que sous le couvert du point de vue de l'utilisateur sur sa tâche.

Nous reviendrons plus en détail sur les techniques utilisées et sur certaines solutions plus inhabituelles dans le domaine de l'accès aux documents et à leur contenu dans le chapitre 2. Retenons pour le moment que l'analyse du contenu d'un document apparaît comme un préalable à la mise en place de techniques d'assistance à son accès et que ces analyses doivent s'appuyer sur l'expression de besoins et d'un point de vue particulier à un utilisateur ou un groupe d'utilisateurs. À la question *pourquoi*, nous répondrons donc, à ce stade de la présentation, que l'accès aux documents et leur gestion doivent passer par l'analyse de leur contenu et que cette analyse ne peut être efficiente que si elle prend en considération les buts documentaires de l'utilisateur.

1.2.2 Accès au contenu des documents

Les parties précédentes nous ont permis d'introduire la notion de contenu de document relativement à une tâche documentaire : il s'agit des informations utiles qu'un lecteur peut en retirer dans le cadre d'une tâche donnée. La justification de nos travaux, ainsi que leur fondements, se trouvent en

particulier dans un article de Tyvaert [Tyvaert, 2003] qui présente une critique des approches formelles pour représenter le contenu des documents. Ces approches, très courantes en TAL, tentent de « formaliser les langues naturelles » ; il s'agit de produire des représentations (logiques) du « contenu » de documents (plus généralement de simples énoncés) et les rendre manipulables par la machine. Les représentations produites font abstraction de l'usager dans son activité de lecture : elles tendent vers une objectivation *du* sens d'un document.

Il y a une centaine d'années, la formalisation de la logique a permis de décrire avec précision certains raisonnements. Il était désormais possible de produire des démonstrations constituées en dérivations réglées à partir d'axiomes. Le décalage qui existe toujours entre la nouveauté technique ou scientifique et la langue a mené à l'émergence d'une métaphore particulière : on parla alors de « langages formels ». Cette dénomination était parfaitement recevable mais l'usage la fit glisser vers la notion beaucoup plus problématique de « langues formelles ». Tyavert [*ibid.* : 37] écrit à ce propos : *Il était certes tentant de comparer les langues naturelles avec les artefacts nouvellement introduits qui pouvaient avoir hérité de leur préhistoire scientifique une parenté avec elles. Tout y invitait : il semblait qu'en combinant symboles de constantes et symboles de fonction et de relation, on reprenait sinon ce qui était noms et verbes, au moins arguments et prédicats. (…) On parla de syntaxe (formelle) et, plus tard, de sémantique (formelle), et on commença à penser (…) qu'on avait inventé les principes d'une langue universelle et parfaite.* La fascinante réussite de la formalisation dans son propre champ permit de croire que l'exactitude était enfin à la portée de l'homme. L'être humain avait dépassé la parole, outil décidément bien imparfait, beaucoup trop diversifié de par les idiomes existants, beaucoup trop souple sémantiquement, beaucoup trop souvent source d'erreurs et de confusions. Mais c'était oublier que l'exactitude des langages formels était d'un type particulier *qu'elle n'avait de sens que dans une tâche limitée, celle du calcul et du raisonnement déductif* [*ibid.*] [15]

Nous aurons l'occasion dans le chapitre 2 de nous attarder sur l'inadéquation des approches critiquées par Tyvaert avec le « quoi » qui nous préoccupe. Notons simplement la principale critique

[15] Il n'est pas étonnant de constater que le phantasme décrit par Tyvaert est tellement partagé et ancien, qu'à l'heure des débuts de l'utilisation de la logique en TAL, Asimov le reprenait déjà à son compte en inversant chronologiquement le processus et en y voyant une dégénérescence des capacités de communication humaines et la cause des misères de l'homme : *À l'origine, le langage articulé constituait le moyen grâce auquel l'homme avait appris à transmettre, quoique imparfaitement, les émotions et les idées issues de son esprit. (…) il avait mis au point une méthode de communication, (…) dont le caractère sommaire et grossier avait provoqué la dégénérescence d'un intellect rompu à toutes les subtilités (…). Décadence sans cesse accentuée, dont on ne peut mesurer les résultats : toutes les souffrances dont l'humanité a été la victime peuvent être imputées au seul fait que, dans toute l'histoire de la Galaxie, nul homme, (…), ne fut véritablement capable de comprendre son semblable. Chaque être humain vivait derrière un mur impénétrable, un brouillard étouffant, en dehors duquel nul autre que lui n'existait. (…) parce qu'ils ne pouvaient se comprendre, parce qu'ils n'osaient pas se faire mutuellement confiance et nourrissaient depuis leur enfance les terreurs et l'insécurité nées de cet ultime isolement, ils éprouvaient cette crainte de l'homme traqué par l'homme, cette sauvage rapacité de l'homme pour l'homme*. I. Asimov. *Seconde Fondation*, Le cycle de Fondation III, Paris : Folio, Denoël, 1966, p.112-113. traduit de l'états-unien par Pierre Billon.

de l'enseignant chercheur qui regrette l'absence de prise en considération dans ce cadre, de la différence qui existe entre l'expression en langue et la formulation en langue d'une relation purement symbolique au sens combinatoire. L'assimilation entre la formule logique et ses exemples langagiers pose des problèmes sur l'analyse des langues en général : la démarche implique, dans une tradition fregéenne, que la signification de toute phrase est fonction des significations de ses parties, ce qui signifie que le sens de chaque unité lexicale non ambiguë est fixé dans le lexique, et que la combinaison de ces sens est guidée par la structure syntaxique de la phrase, et aboutit à l'interprétation de la phrase complète. Si ce principe, dit de compositionnalité, semble effectivement valable pour les formules logiques ou mathématiques, il ne semble pas être transférable à la langue naturelle. Par exemple, la construction du sens de la formule arithmétique 1+2=3 relève d'une procédure compositionnelle. Nous sommes en présence de symboles de constantes (1, 2 et 3), d'un symbole de fonction (+) et d'un symbole de relation (=) dont la signification est bien définie et sans ambiguïté ; leur univocité est leur nature d'être. Les pratiques langagières courantes ne procèdent pas de la même manière car les corpus, les textes, les phrases, sont autant d'occasions de contre-sens qu'il y a de force d'indécision sémantique dans les mots lexicaux qui y sont mentionnés. Distinguer le chat qui rapporte la souris en tant qu'animal de la famille des félidés ou en tant que petit grattement au fond de la gorge qui démange, relève justement de la cotextualisation du mot « chat ». De même que si l'on lit qu'il « rapporte », le fait de retenir que l'action de « rapporter » a trait au déplacement d'une chose dans la gueule d'un animal vers son maître et non au comportement du premier de la classe qui dénonce son petit camarade, dépend de la prise en considération du cotexte d'apparition du terme. Les contre-sens sont finalement rares puisque les indices intrinsèques et extrinsèques aux textes, qui sont accessibles au sujet interprétant, permettent dans la plupart des cas de les éviter. Ces indices ne sont pas nécessairement conscientisés, à tel point que ce sont les traitements informatiques de la langue qui en ont finalement révélé l'ampleur.

Tyvaert apporte alors un nouvel argument à la question qui nous préoccupe dans cette partie. À la question pourquoi, nous répondrons, en nous appuyant sur ses propos, que si les approches compositionnelles d'analyse de la langue sont utiles dans des systèmes de déduction, leur non-conformité avec le fonctionnement des pratiques langagières réelles est un obstacle à leur utilisation dans le cadre de nos travaux ; en particulier, elles supposent une unicité des significations des mots ou au moins l'énumérabilité de toutes leurs significations possibles, ce qui semble être difficilement réalisable lorsqu'on s'intéresse à la langue naturelle, même dans une approche computationnelle. De plus, l'utilisateur est absent des systèmes logiques : il n'est que l'interprétant des résultats automatiques. Il n'est pas possible dans ces circonstances de lui proposer des moyens d'assistance personnalisés.

1.3 Démarche

1.3.1 Décrire des significations

C'est une nouvelle fois sur l'article de Tyvaert que nous nous fondons pour présenter un aperçu de nos propositions. En réaction à ses critiques des modèles formels, le linguiste y présente un modèle de compréhension de textes qui permet de *rendre compte, en respectant les acquis de la linguistique (en particulier ceux de la sémantique lexicale), de l'élaboration, provisoire certes, mais sûre et bien conduite, du sens de tel ou tel propos par celui qui en est le destinataire* [*ibid.*]. Il s'inspire pour cela de la Sémantique Interprétative (SI) de Rastier [Rastier, 1987].

Reprenons une phrase proposée en exemple dans l'article pour expliquer la démarche :
Les coquelicots flamboient dans le couchant.

Chacun des termes de cette phrase est polysémique dans le sens où, par exemple, *coquelicot* peut avoir trait à la 'fleur', à la 'plante', 'au parasite (des cultures céréalières)', à l' 'opiacé' et donc au 'somnifère' qu'il peut être dans certains cas. Pour rendre compte de la compréhension que l'on peut faire de cette phrase, Tyvaert se fonde sur les *indications [sémantiques] associées par l'usage à chaque terme considéré artificiellement hors contexte* pour donner à chaque terme sa signification dans la phrase. Selon lui, l'usage tend à associer au mot *coquelicot* les indications 'plante', 'fleur', 'champ', 'parasite', 'rouge', 'tige (barbue)' etc. tandis qu'au mot *flamboyer* sont associés 'action', 'rayonnement', 'chaleur', 'lumière', 'rouge (orangé)' et à *couchant* 'lieu (de l'horizon)', 'soir', '(fin de la) lumière (du jour)', '(ciel) rouge (orangé)', '(coucher du) soleil', etc. À partir de ces associations, la cooccurrence des trois termes *coquelicot*, *flamboyer* et *couchant* classe ces indications diverses – dans leur rapport à chacun des termes – en fonction de l'importance de leur présence dans la phrase. Nous voyons que 'rouge (orangé)' revient trois fois et 'lumière' deux fois. Il semble alors que dans cette phrase la signification de « coquelicot » ordonne les indications relevées hors texte en plaçant au premier rang l'indication 'rouge (orangé)' et au deuxième 'lumière'. Cette démonstration exhibe une procédure très simple, mobilisable par tout locuteur même peu expérimenté, et néanmoins rigoureuse, pour rendre compte de l'interprétation de cette phrase (l'auteur la généralise d'ailleurs à des textes – des poèmes – autrement plus complexes). Dans les faits, les mots semblent véhiculer des significations en nombre incomparablement plus grand que le leur et même que celui des indications que l'on peut légitimement associer à chacun d'entre eux, en particulier du fait du grand nombre de combinaisons qui peuvent être envisagées (au sein de phrases, de textes, de corpus…). À partir de ce principe très simple, Tyvaert prône pour le TAL la conception de *langages formels de deuxième génération*. À tous les signes, on pourrait attacher des collections d'indications (des traits qu'on pourrait lister) et qu'on utiliserait, dès qu'il y a texte (et seulement quand il y a texte) pour constituer selon la procédure de

textualisation, la signification de chaque signe dans ce texte. Ces significations ainsi définies pourraient ensuite être utilisées pour *calculer un sens* en fonction des récurrences de traits.

L'approche de Tyvaert est très intéressante à plusieurs titres. Tout d'abord, et cela est loin d'être négligeable, qui plus est dans une optique computationnelle, elle est très simple. Ensuite, elle s'absout des travers des méthodes classiques logiques et conceptuelles d'appréhension du sens et permet de s'intéresser à un sujet interprétant (l'auteur prône pour l'utilisation *d'un lexique propre à chaque sujet*). Cependant, une mise en place informatique est loin d'être triviale et encore moins, dans le cadre de l'élaboration d'outils applicatifs hors de la sphère de la recherche. Le premier problème qui se pose est la limitation des « indications sémantiques » associables à un mot pour un utilisateur. La tâche semble irréalisable si elle tend vers l'exhaustivité (souvent inhérente à l'implantation de telles mécanismes). En nous fondant sur ces propositions, nous prenons position pour une démarche praxéologique et personnelle à un utilisateur ou à un groupe restreint d'utilisateurs. Ne seront stockées en machine que les indications sémantiques, les traits, utiles à la tâche, découverts et jugés adéquats en fonction de l'interprétation de textes ou de connaissances de l'usager convocables dans la situation. Ces traits pourront être distingués en fonction de leur nature au sein de classes de traits qui nous serviront à des analyses automatiques. Il pourra s'agir de traits que l'usage permet d'associer au terme mais également, de traits plus personnels à un utilisateur si ceux-ci s'avèrent utiles pour sa tâche. En cela, et contrairement à Tyvaert, nous ne postulerons rien sur l'éventuelle « naturalité » de ces associations. Du point de vue des données informatiques manipulées, nous assisterons l'usager dans cette description à plusieurs niveaux :

- dans le choix des mots utiles à décrire (à travers des aides à l'exploration d'un corpus d'observation en rapport avec sa tâche) ;
- dans le choix des descriptions à produire (à travers un principe de catégorisation permettant à la fois de structurer les traits et les mots et de limiter le nombre de traits associés).

L'approche de Tyvaert sous-entend la possibilité de *calculer un sens* ce qui revient non seulement à savoir ce qu'est le sens mais aussi à le considérer comme figé, immanent à un texte, à une phrase. Dans ce cadre, le lecteur n'est vu que comme un déchiffreur qui n'aurait qu'à convoquer les *bonnes indications sémantiques* pour *bien* accéder au sens d'un texte, *pour montrer sa vraie interprétation* (sic). Nous aurons l'occasion de défendre une position transversale qui consiste à considérer plutôt le sens comme un processus et d'en fixer ainsi les limites de calculabilité comme si c'était un résultat. Par « calcul », on entend généralement une opération qui a pour but l'obtention d'un résultat à partir d'une combinaison de nombres ou d'autres symboles. L'ordinateur est en ce sens un très bon calculateur. En informatique comme en linguistique, par « calculer le sens », on entend souvent une traduction de la langue (dite naturelle) en une représentation formelle et par abus de langage, cette représentation est souvent entreprise comme « le sens » de l'énoncé, du texte, qui a permis son élabora-

tion. Dans les faits, il est très difficile d'évaluer la relation entre un énoncé et une représentation formelle, entre un texte et, par exemple, une suite de formules logiques, ou comme le propose Tyvaert, une succession de traits à retenir. Si l'on considère que le sens est dans le texte, on doit obligatoirement avouer une perte de sens entre le texte et sa représentation qui n'est autre qu'une reformulation. Nous refusons le statut de représentation du contenu des textes à ce type de formalisation et considérons ce contenu comme incalculable puisqu'il dépend du lecteur et de l'interprétation qu'il fait d'un texte dans une situation donnée. Comme il a été par exemple souligné dans [Prié, 1995], l'interprétation d'un texte n'est pas déterministe : il y a pluralité de sens pour un texte (multiplicité qui n'est ni unicité, ni infinité). Comment dès lors prétendre calculer le sens ? Pour notre part, à la manière de capteurs qui permettent d'effectuer des calculs indiciaires pour la prédiction de certains phénomènes sans les cerner pleinement (comme en météorologie par exemple), nos processus automatiques produiront des indications et des moyens d'exploiter ces indications. Au final, le dernier mot sera toujours celui du lecteur/utilisateur du système. Si les résultats obtenus automatiquement répondent à ses besoins, l'usager pourra s'en contenter. Mais en aucun cas, nous n'aurons la prétention de proposer une *vraie* interprétation, ou de calculer *le* sens. Néanmoins, selon les propositions de Tyvaert, nous nous fonderons sur le principe de récurrence des traits sémantiques pour rendre les services attendus par l'utilisateur.

Nous aurons l'occasion d'aborder dans le prochain chapitre : un certain nombre de réalisations informatiques exploitent déjà ces principes. Celles-ci s'inspirent non pas de Tyavert mais de la Sémantique Interprétative de François Rastier qui donne entre autres un cadre rigoureux pour l'organisation des traits sémantiques. L'utilisation des « indications sémantiques » ressemble en effet aux *sèmes* (traits sémantiques) de la sémantique componentielle et la découverte de la redondance de ces traits au sein des textes est le pendant de l'*isotopie*. Ces concepts sont utilisés par la Sémantique Interprétative (SI) dont nous aurons maintes fois l'occasion de montrer l'influence sur nos travaux. Cependant, d'aucuns pourront signaler que les travaux de Tyvaert correspondent à une vision erronée des sèmes de la SI de Rastier puisque dans ce cadre, les sèmes sont le résultat de l'interprétation d'un texte et non des primitives utilisables pour décrire le sens des mots. Les sèmes en SI ne sont pas des composants sémantiques préexistants dont la composition permettrait de décrire le sens des mots. C'est entre autres pour cela que les indications sémantiques du système seront de préférence associées à des mots par l'intermédiaire d'études de textes. Comme nous le verrons plus en détail dans le chapitre 2, les traits sémantiques ne se définissent que dans la confrontation de plusieurs unités linguistiques.

Notre position, orientée vers la tâche et l'utilisateur, nous amène à simplifier la plupart des concepts de la SI. Ces simplifications nous permettent à la fois de proposer des services avérés, et de pouvoir proposer ces services à des non-spécialistes[16] de la langue et de l'interprétation.

À la question comment, nous pouvons répondre pour l'instant qu'il s'agira de solliciter un usager pour décrire des mots intéressants pour sa tâche à l'aide de traits sémantiques. La machine se chargera alors d'analyser des textes en se fondant sur les récurrences de ces traits.

1.3.2 Organiser les descriptions

Pour faciliter la compréhension de l'ensemble de ce tapuscrit, nous proposons dès maintenant de lever quelque peu le voile sur la partie de notre modèle qui concerne les ressources et leur structuration. Loin de contenir la description complète du modèle, cette partie a simplement pour but d'en présenter la terminologie.

Tout d'abord, la tradition scientifique de notre discipline nous obligeant à nommer notre modèle (et si possible à l'aide d'un acronyme en anglais), nous avons choisi **LUCIA** pour *Located User-Centred Interpretative Analyser*. *Located* nous sert à préciser qu'il s'agit de prendre en considération un utilisateur situé dans une histoire, dans une culture, dans une langue et dans une tâche précise. *User-centred* affirme notre volonté d'élaborer des propositions autour des désirs et des besoins des utilisateurs, pour leur rendre des services personnalisés et leur faciliter au maximum toutes les tâches pour lesquelles ils seront sollicités. Enfin, *Interpretative Analyser* exprime notre intérêt pour l'interprétation, c'est-à-dire l'activité de lecture d'un texte par un usager.

Nous venons de voir que les ressources de notre système seront des mots associés à des traits. Ces ressources seront le reflet d'un point de vue de leur auteur à un moment donné sur les choses qui y sont décrites. Ce point de vue sera exprimé à travers les choix effectués lors des étapes de constitution et de structuration des ressources. Pour structurer les descriptions, nos « indications sémantiques » ou nos traits auront la forme d'**attributs**. Ils seront constitués d'un mot ou d'une paraphrase permettant de les dénommer et d'un ensemble de valeurs opposées également représentées par des mots. L'association d'un attribut et d'une entité lexicale permettra de distinguer cette entité d'autres qui partagent des traits identiques selon le point de vue de l'auteur des associations. Par exemple, les signifiés des entités *anticyclone* et *dépression* pourront être décrits à l'aide de l'attribut dénommé « Pression »

[16] Jacques Coursil dans un entretien accordé à Antoine Perraud dans l'émission « *Tire ta langue* » de France Culture en septembre 2002, déniait le statut de « spécialiste de la langue » aux linguistes arguant qu'ils n'étaient pas les auxiliaires les plus recherchés pour la rédaction d'une lettre d'amour – tâche où l'enjeu semble plus primordial, du moins pour les intéressés, qu'une analyse littéraire d'un poème romantique par exemple. Dans nos propos, le spécialiste de la langue doit être entendu comme celui ou celle qui est capable de manipuler des concepts linguistiques plus ou moins ardus et qui possède une culture assez vaste dans cette discipline ou tout du moins une culture académique relativement développée en linguistique, telle qu'on l'enseigne à l'université.

ayant comme valeurs opposées [haute vs. basse], *anticyclone* actualisant la valeur « haute » et *dépression* la valeur « basse ». Nous noterons cet attribut [Pression : haute vs. basse]. On représentera cette description à l'aide d'une **table** que l'on pourra nommer pour notre exemple « Phénomènes atmosphériques » (Figure 3).

Phénomènes atmosphériques	**Pression**
anticyclone	haute
dépression	basse

Figure 3 – Table LUCIA des « Phénomènes atmosphériques ».

La description présentée dans la figure 3 sera corrélée à une observation dans un corpus donné et/ou aux connaissances convoquées pour la tâche en cours et au point de vue du lecteur/utilisateur. Ainsi, les propositions exprimées dans la figure 3 pourront être remises en question si, par exemple, un corpus d'observation fait place à une utilisation métaphorique ou relevant d'un autre domaine pour l'une ou l'autre entité. Il pourra en être également autrement si l'usager est plus intéressé par le fait qu'un anticyclone amène le beau temps, etc. Le caractère subjectif des représentations nous permet de pointer ici la notion centrale de notre système : l'interprétation. Cette dernière est considérée relativement au lecteur/utilisateur dans la situation de sa tâche. Nous rejoignons ici Cavazza [Cavazza, 1996 : 62] en considérant toute description sémantique de la langue comme de l'interprétation figée. Ceci relativise la portée des descriptions à une situation donnée : celle de l'utilisateur et de sa tâche.

Plusieurs attributs pourront être utilisés pour décrire un ensemble de mots partageant des éléments de significations. Plusieurs tables croisant des mots (en fait leurs signifiés) et des attributs/valeurs pourront être créées pour décrire un même domaine ou un même sujet d'intérêt, elles formeront alors un **dispositif**.

À la question comment, nous pouvons répondre qu'il s'agit de s'inspirer de théories linguistiques pour la mise en place d'un modèle de description et de catégorisation lexicale permettant l'expression du point de vue d'un utilisateur ou d'un groupe d'utilisateurs sur une tâche documentaire. Ce modèle permet d'effectuer des analyses de documents numériques textuels qui auront pour but d'en faciliter l'accès et l'exploration du contenu pour une tâche donnée. Plus précisément, il s'agit d'organiser et de décrire en termes de traits des mots à partir de leur observation dans un contexte précis. Les mots sont décrits à l'aide d'attributs, agencés au sein de tables elles-mêmes regroupées au sein de dispositifs. Placés dans les structures, ces traits sont considérés comme des éléments de significations potentiellement valables dans d'autres textes que ceux qui ont permis de les élaborer. Ces ressources sont exploitées pour présenter le matériau textuel de façon personnalisée. Elles sont également l'objet d'analyses automatiques de textes basées sur le repérage des récurrences d'attributs et valeurs d'attributs. Ces récurrences sont répertoriées et quantifiées pour mettre en place les moyens

d'assistance attendus. Les tâches que nous avons choisies pour valider nos propositions, et qui servent donc d'exemple tout au long de ce tapuscrit, sont la veille documentaire et l'analyse de fait de langue (l'analyse d'une métaphore conceptuelle).

1.3.3 L'interaction comme alternative aux approches classiques

À ce stade de la présentation, nous voyons que l'interaction est une orientation majeure de nos travaux. Nos préoccupations opératoires quant à la manipulation de mots et de textes sont placées au même niveau d'importance que celles relatives aux interactions entre l'utilisateur et le système.

Notre approche se concentre sur l'utilisateur dans le cadre d'une tâche documentaire donnée, sur sa façon d'envisager cette tâche et sur ses besoins spécifiques. Dans notre système, le contenu des textes ne pourra pas être analysé entièrement automatiquement car, comme nous le verrons dans les prochains chapitres, les aspects intéressants de ce contenu pour une tâche documentaire ne sont pas tous accessibles à partir de dépendances sémantiques prévisibles entre les termes contextualisés ou à travers leurs seules cooccurrences. L'indéterminisme des opérations de construction d'un sens pour un utilisateur implique notre incapacité à automatiser entièrement toutes les opérations. Le processus de construction du sens est envisagé comme accessible seulement à l'utilisateur. Le rôle de la machine est de l'assister le plus efficacement possible. C'est en exploitant les traces d'une interprétation de l'usager qu'elle pourra réaliser des tâches classiques comme filtrer et classer des documents inconnus. Dans l'optique applicative que nous avons choisie, nous sommes tout autant soucieux de l'élaboration d'un modèle opératoire d'analyse de textes et de représentation de significations, que des moyens proposés aux usagers pour utiliser effectivement les réalisations logicielles mettant en œuvre ces principes. Cet intérêt se concrétise par la réalisation d'interfaces d'interaction entre l'utilisateur et la machine pour l'assistance personnalisée à des tâches qu'il est impossible d'automatiser. Nous avons ainsi été amené à nous interroger sur la généricité et la spécificité de certaines étapes des tâches documentaires pour proposer des principes communs ou non à nos logiciels en fonction du cadre de leur utilisation et du type d'utilisateurs concernés. Nous exploitons au maximum les capacités actuelles des machines en ce domaine, en particulier par la création de représentations graphiques, l'utilisation de la couleur, voire de la 3D. Nos propositions se limitent cependant aux appareillages technologiques les plus classiques (écran, souris, clavier) et aux logiciels les plus courants (logiciels multi-plateformes et navigateurs pour l'Internet) pour en assurer une utilisabilité maximale.

Dans nos travaux, l'interaction est vue comme un moyen de personnaliser et d'optimiser qualitativement les résultats. Elle relève de la personnalisation des traitements. À la question comment ? Nous pouvons donc aussi répondre que c'est à travers l'interaction, envisagée dans sa mise en place effective dans des logiciels, que nous parvenons à personnaliser autant les traitements que la présentation des résultats automatiques.

1.4 Sémantique légère

Les buts de cette thèse sont d'évaluer les possibilités de traitements du contenu des documents à partir de ressources minimales en quantité et de processus minimaux en complexité[17].

1.4.1 Les ressources

En TAL, lorsqu'on effectue des traitements ayant trait au sens, il est courant d'utiliser une quantité très importante de données. Cette quantité est d'ailleurs érigée en objet de fierté tant ce type de ressources est généralement complexe et long à élaborer. Les ontologies, les thésaurus ou les dictionnaires utilisées contiennent couramment plusieurs milliers de mots classés, annotés voire enrichies de définitions, de relations et d'informations d'ordre syntaxique. Pour créer ces données, on a recours soit à des spécialistes, soit à des traitements « d'extraction de connaissances » qui, comme l'a souligné [Pincemin, 1999a : 113], mobilisent nécessairement une collaboration humaine. Les données obtenues automatiquement appellent forcément des corrections et des ajustements. La multiplicité des calculs possibles et des corpus pour créer ces données est telle que chaque résultat obtenu de la machine a une valeur d'inachèvement et d'ouverture. Si ces données ont pour objectif une visée générale sur un domaine, leur validité dépend néanmoins à la fois du point de vue du concepteur du système (ou du spécialiste qui a permis de les mettre au jour) et des théories sous-jacentes. De plus, leur pérennité est limitée du fait de l'évolution constante des usages en langue. Nous avons pour preuve les dictionnaires et les thésaurus des grandes maisons d'édition qui se vantent chaque année d'avoir plusieurs dizaines de mots nouveaux, de nouvelles définitions et de nouvelles relations entre termes. Des néologismes apparaissent régulièrement et leurs formes sont créées non seulement par dérivation et composition, mais également pas siglaison, abréviation, emprunt, etc. voire *ex novo* dans le cas de noms de marques commerciales par exemple. L'évolution sociale de la langue amène des changements significatifs d'emploi de certains termes et des emprunts nombreux d'une langue à l'autre. Il y a moins de 30 ans des mots comme *arobase* (ou *arrobe*) et *bogue* (ou *bug*) étaient absents des dictionnaires courants alors qu'il paraît impossible de nos jours de ne pas les voir y figurer[18]. La recherche de néologismes est d'ailleurs un sujet de recherche à part entière inhérent aux approches fondées sur des données préconstruites que l'on désire toujours plus exhaustives [Tanguy et Hathout, 2003]. Les dictionnaires re-

[17] Si la complexité algorithmique sera prise en considération pour la mise au point de nos propositions, nous parlons ici avant tout de la complexité en tant que complication, en tant que facteur qui entraîne beaucoup d'efforts et de connaissances pour rendre compréhensibles les mécanismes.

[18] Si le signe @ remonte au moyen-âge, les termes utilisés actuellement pour le désigner n'ont pas plus d'une quarantaine d'années. En France, la commission spécialisée de terminologie et de néologie de l'informatique et des composants électroniques a choisi de préconiser les termes *arrobe* et *arobase*. Tandis que le bogue est la version francisé de *bug* qui sert à désigner une erreur dans un programme informatique – elle n'a rien à voir avec la *bogue* (de la noix par exemple) dont l'apparition dans les dictionnaires est bien antérieure.

cèlent bien des connaissances sur la langue, sur la parenté des mots, sur leurs significations d'usage mais leur utilisation en TAL pose encore d'autres problèmes que ceux que nous venons d'évoquer. D'une part, ils ne comportent pas nécessairement de marques des domaines qui pourraient situer les secteurs ou les pratiques sociales d'usage des termes, ou encore de précisions quant au genre des textes correspondant aux significations décrites. D'autre part, ce qu'ils enregistrent n'est ni jamais complet, ni toujours valide. Les significations en contexte sont innombrables et des contextualisations et des situations amènent à rendre les descriptions obsolètes ou non-avenues. Pourtant les approches computationnelles qui utilisent ce type de ressources postulent souvent ces qualités pour asseoir la validité des traitements proposés. L'utilité des dictionnaires ou des thésaurus dans la vie courante n'est plus à démontrer, cependant, même sous forme électronique, ces objets sont par nature statiques. Lorsqu'on s'intéresse à l'activité de lecture, i.e. à l'interprétation, le sujet d'étude est dynamique. Il y a là une contradiction manifeste. Le dictionnaire peut être envisagé comme un objet linguistique dans lequel un contenu sémantique isolable ne s'identifie pas à une définition mais se construit à partir d'un contexte donné par le dictionnaire. Ce contexte est composé d'une définition mais aussi des citations, des illustrations d'usage, etc. Mais la formalisation et donc l'accessibilité de ce contexte par la machine ne peut suffire pour déterminer une signification contextualisée puisque celle-ci *se construit* dans l'acte d'interprétation.

C'est pour toutes ces raisons, que nous tenons dans nos travaux de limiter les ressources lexicales. La personnalisation des données à travers les principes d'interaction mis en place nous permet d'en limiter la quantité nécessaire et la complexité pour arriver à des résultats satisfaisants pour l'usager. Sur le modèle des traitements syntaxiques sans dictionnaire proposés par Vergne de l'équipe ISLanD du GREYC [Vergne, 2000], nous essayons d'être dépendants au minimum de données exogènes aux textes et aux corpus analysés. Les données du système ne doivent pas être inconnues ou incompréhensibles pour l'utilisateur. En limitant les données, nous tentons d'éviter les écueils que nous venons d'exposer, et de prendre en considération le point de vue de l'utilisateur et non uniquement celui du concepteur, qu'il soit linguiste, informaticien ou les deux.

1.4.2 Les processus

La plupart des travaux de TAL qui s'intéressent aux documents et à leur contenu mettent en place des processus très complexes. Cette complexité est principalement due à la quantité de problèmes qu'ils se proposent de résoudre sans solliciter l'utilisateur. Elle est parfois un frein à l'utilisation des logiciels puisque les usagers en sont réduits à faire des prospectives hasardeuses sur le fonctionnement réel des outils pour les exploiter au mieux. Ceci est aggravé par le peu d'intérêt suscité par les documentations proposées avec les logiciels ; la démocratisation de l'informatique a porté la formation

autodidacte au rang de seul apprentissage admissible pour nombre d'usagers[19]. L'utilisation des opérateurs booléens pour la construction de requêtes pour les moteurs de recherche de l'Internet s'avère, par exemple, inaccessible à la plupart des utilisateurs et l'on constate que la majorité des requêtes soumises contiennent au plus deux mots sans connecteur[20]. Ainsi, un moteur comme HotBot propose à l'heure actuelle un formulaire avec plus d'une vingtaine de champs pour constituer une requête complexe sans en connaître la syntaxe[21].

La complexité des traitements couplée à des grandes quantités de ressources amènent, en outre, à des solutions logicielles difficiles à mettre à jour et à adapter en fonction des tâches et des utilisateurs. Ce problème est parfois évité lorsque les traitements sont clairement distingués des données comme pour les moteurs à bases de règles d'inférence, mais l'explosion combinatoire n'en est pas pour autant résolue. Lorsque les solutions informatiques s'inspirent de théories linguistiques qu'ils tentent de suivre le plus rigoureusement possible, une des conclusions généralement admises est que ces théories n'ont finalement que rarement des buts communs avec les problématiques du TAL. Le travail de l'informaticien consiste alors à adopter une approche opportuniste et de distinguer ce qui peut être véritablement exploité en tant que tel de ce qui peut relève d'une orientation théorique. Les travaux de Bommier-Pincemin [Bommier-Pincemin, 1999] qui s'appuient à la fois sur la SI et sur la technique vectorielle de représentation du contenu de documents sont à ce titre exemplaires. La distinction entre ce qui relève d'une volonté d'implanter une théorie et ce qui relève de l'intérêt pour une tâche véritable pour laquelle on peut solliciter la machine n'est pas simple à effectuer.

1.4.3 Vers une *sémantique légère* pour le TAL

Nos travaux tentent de rendre des services à l'utilisateur à l'aide de programmes informatiques fondés sur des principes simples. Nous évitons ainsi les mises à jour trop complexes et l'impossibilité d'adapter les propositions à des tâches diverses. Les principes d'interaction mis en place nous permettent d'impliquer l'utilisateur dans toutes les phases de fonctionnement du système où son point de vue et les enjeux personnels de sa tâche peuvent être exploités pour mieux lui rendre service. Les données sont limitées à celles qui lui sont utiles et qu'il peut fournir à la machine. Les traitements sont simples car ils reposent principalement sur le calcul à partir des différences et de points communs entre certaines informations. Les traitements sémantiques légers montrent ainsi tous les services qu'ils peuvent rendre.

[19] Ceci a par exemple été montré en particulier par Beguin pour les progiciels : *« ... on constate que les manuels d'utilisation sont peu utilisés : les usagers préfèrent apprendre en faisant. »* [Beguin, 1991]

[20] Voir par exemple http://www.yooda.com/info/article/lycos_voyeur/lycosvoyeur.php et [Loupy, 2000].

[21] http://www.hotbot.fr/?command=adv

Nous tentons ainsi de mettre en place les bases d'une *sémantique légère* en informatique, qui ne nécessiterait pas beaucoup de ressources ni de traitements très complexes mais qui permettrait cependant de rendre des services touchant au sens utiles aux utilisateurs. Pour arriver à nos fins, nous avons adopté une approche opportuniste de théories linguistiques comme la SI de Rastier qui représente néanmoins notre principale source d'inspiration. Nous avons questionné cette théorie à la lumière d'applications informatiques réelles : l'articulation entre un modèle linguistique descriptif et les services informatiques possibles est interrogée et évaluée à partir du modèle LUCIA et de ses réalisations logicielles dédiées à des tâches documentaires.

Ces principes nous permettent, en outre, de proposer finalement des solutions logiciels multilingues, utilisables pour plusieurs langues ou au moins, des fonctionnements les moins dépendants possibles d'une langue donnée. Le multilinguisme apparaît en effet comme une nécessité dans les travaux qui touchent aux tâches documentaires, vu les facilités et les nécessités actuelles d'accéder à des documents écrits dans de nombreuses langues.

1.5 Plan de la thèse

Dans ce chapitre, nous avons présenté succinctement l'orientation générale de nos travaux. Nous avons vu en particulier que nous nous intéressons principalement à l'utilisateur dans la situation d'une tâche documentaire précise. Il ne faut pas s'y tromper, les catégories de Borges proposées en début de ce chapitre ne constituent pas un exemple du type de ressources personnalisées que nous entendons construire. Les catégories très personnelles de Borges ne sauraient être la base d'une structuration de ressources telle que nous l'entendons car c'est leur singularité qui a justement amené à tant de commentaires. Dans notre système, l'utilisateur n'est pas tant un individu donné, qu'un groupe socialement marqué dont les membres partagent une tâche commune et une façon identique de l'envisager et pour qui, des mots ont un emploi commun attesté permettant une intercompréhension optimale (ce qui dans l'utilisation du modèle peut amener la création de catégories communes ou proches). À propos de l'immutabilité et la mutabilité du signe, Saussure écrivait «*… s'il l'on veut démontrer que la loi admise dans une collectivité est une chose que l'on subit, et non une règle librement consentie, c'est bien la langue qui en offre la preuve la plus éclatante.* » [Saussure, 1916 : 104]. Mais si le signe échappe à notre volonté, le sens se construit et trouve sa source dans le consensus. C'est ce consensus au sein d'une communauté donnée (fut-elle très restreinte) qui devra servir de critère pour les analyses effectuées à partir de notre modèle. Ainsi la distinction entre utilisateur et groupe d'utilisateurs pourra dans certaines configurations être considérablement réduite si ce n'est devenir totalement obsolète et Borges, iconoclaste dans sa façon de parler du monde, risquera de se retrouver fort isolé. C'est le caractère social et donc partagé de la langue qui constituera le frein majeur au « dire

n'importe quoi » à travers la construction et la constitution des ressources personnelles selon le modèle LUCIA. En outre, les contraintes du modèle limiteront les fantaisies possibles.

Notre étude se focalise sur un autre aspect, crucial à nos yeux : l'interaction entre l'utilisateur et la machine. L'informatique en général et le TAL en particulier n'ont pas pour unique tâche la réalisation de modèles opératoires. Les productions supportées par l'ordinateur (sons, images, vidéos, mondes virtuels, langages informatiques, etc.) et les interactions nécessaires à leur constitution ou leur utilisation (pointer avec une souris, frapper sur un clavier, etc.) sont autant d'*artefacts sémiotiques* dont les modalités d'étude scientifique au regard de l'informatique sont en cours de construction [Nicolle, 2001]. Il incombe ainsi à l'informaticien non seulement de créer ces artefacts mais également de les ériger en objet d'étude à part entière pour mieux cerner leur utilité et leur faisabilité. Développer des programmes informatiques qui permettent des actions réciproques en mode dialogué avec des utilisateurs (en d'autres termes : des logiciels interactifs) dont une tâche au moins requiert l'analyse de documents textuels, transporte l'informaticien taliste dans des cadres épistémologiques qui dépassent souvent les limites académiques de son domaine :

- le TAL entretient des relations fortes avec la linguistique, la philosophie du langage, etc. ;
- l'Interaction Homme-Machine (IHM) relève aussi de la sociologie, de l'ergonomie, etc.

D'une manière générale, le TAL ne peut plus aujourd'hui se soustraire à l'influence de ces sciences qui, depuis le milieu des années 1970, se sont vues attribuer l'épithète *cognitive*. Mais si nos travaux côtoient parfois l'une ou l'autre de ces disciplines, nous nous bornerons à n'exposer ici que leurs aspects informatiques et linguistiques.

Cette thèse comporte quatre chapitres après celui-ci ainsi qu'une conclusion :

Le second chapitre présente et justifie les fondements de notre approche en s'appuyant sur l'examen des solutions classiques utilisées en TAL pour faciliter l'accès aux documents et à leur contenu. Nous nous attardons en particulier sur les solutions qui tentent de pallier les problèmes inhérents aux approches en termes de recherche documentaire telles que nous les avons exposées en 1.2. Nous précisons nos objectifs et fixons nos choix théoriques et opératoires en nous appuyant sur certaines théories existantes.

Le troisième chapitre présente en détail le modèle de catégorisation et de description de significations créé. Nous y fixons également les critères de sélection et les moyens de manipulation choisis pour les entrées du système. Les résultats d'une expérience avec des usagers potentiels nous permettent une première évaluation de la mise en place effective de ces principes.

Le quatrième chapitre est relatif à la définition et la structuration des ressources en fonction de différentes tâches documentaires envisagées. Nous présentons entre autres les deux applications choisies pour l'évaluation de nos propositions : une tâche de veille documentaire et une tâche d'analyse de fait de langue (une métaphore conceptuelle).

Le cinquième chapitre présente en détails les analyses effectuées et les processus mis en œuvre pour rendre les services attendus par les usagers. Il apporte des éléments de validation du modèle et des outils logiciels proposés dans les deux champs d'applications choisis. Les moyens d'interaction et de visualisation prennent une part importante de cette présentation. Nous précisons également les limites comme base de réflexion aux voies de recherche envisageables pour y remédier.

Enfin, nous proposons dans une conclusion générale de replacer l'ensemble de ce travail dans les perspectives actuelles des STIC (Sciences et Technologies de l'Information et des communications).

Le travail présenté a donné lieu à plusieurs publications :

Chapitre d'ouvrage :

- **Perlerin**, V. et Beust, P. (2003). Pour une instrumentation informatique du sens. *Variation, construction et instrumentation du sens*. Siksou, M. (ed.). chap. 8. pp. 197-228. Hermès, Lavoisier. Paris.

Revue internationale avec comité de lecture :

- Nicolle, A., Beust, P., et **Perlerin**, V. (2002). *Un analogue de la mémoire pour un agent logiciel interactif*. In Cognito. n°21. pp. 37-66.

Conférences internationales avec comité de lecture et publication des actes :

- **Perlerin**, V. (2001). *La recherche documentaire : une activité langagière*. Actes de TALN/RECITAL 2001. tome 1. pp. 469-479. Tours, France.
- **Perlerin**, V. (2002). Memlabor, un environnement de création, de gestion et de manipulation de corpus de textes. Actes de TALN/RECITAL 2002. tome 1. pp. 507-516. Nancy, France.
- **Perlerin**, V., Beust, P., et Ferrari, S. (2003). *Computer-assisted interpretation in domain-specific corpora: the case of the metaphor*. Proceedings the 14th Nordic Conference on Computational Linguistics NODALIDA'03. [abstract] Reykjavik, Iceland.
- **Perlerin**, V., Beust, P., et Ferrari, S. (2002). *Métaphores et dynamique sémique.* Actes des 3e Journées de Linguistique de Corpus. Presses universitaires de Lorient, Lorient. (en cours de publication)
- **Perlerin**, V. et Ferrari, S. (2003). *Les besoins d'interaction en TAL et en Linguistique de Corpus : étude de cas.* Actes des 4e Journées de Linguistique de Corpus. Presses universitaires de Lorient, Lorient. (en cours de publication)
- Beust, P., Ferrari S., et **Perlerin**, V. (2003). *NLP model and tools for detecting and interpreting metaphors in domain-specific corpora*. Proceedings of the Corpus Linguistics 2003 conference. Archer, D., Rayson, P., Wilson, A. and McEnery, T. (eds.). Lancaster, UK, 114-123. UCREL technical paper n°16. UCREL, Lancaster University.

  (Presented both at the Main Conference and the Interdisciplinary Workshop on Corpus- Based Approaches to Figurative Language. UK, Lancaster, 27 March - 1 April 2003)
- Ferrari, S. et **Perlerin**, V. (2004). *Modèle sémantique et interactions pour l'analyse de documents*. Approches Sémantique du Document Electronique. Actes du septième Colloque International sur le Document Électronique CIDE.7. pp. 231-251. La Rochelle, France.