Chapitre 2

Fondements de l'approche

Dans le chapitre précédent, nous avons présenté nos objectifs. Nous avons montré qu'il existe un véritable besoin d'aide personnalisée à l'accès au contenu des documents numériques. C'est pour cela que cette thèse a pour but la réalisation d'outils informatiques pour l'assistance personnalisée à l'accès aux documents numériques et à leur contenu. Ce chapitre présente et justifie les fondements de notre approche. Dans une première partie (2.1), nous débutons par un état de l'art des solutions existantes dans ce domaine et plus largement des modèles de la sémantique utilisés en Intelligence Artificielle (IA). Nous présentons en particulier les ontologies utilisées dans le cadre du Web Sémantique (2.1.1) héritières de travaux anciens en représentation de connaissances (2.1.2). Comme nous l'avons précisé dans le chapitre précédent, notre modèle s'appuie majoritairement sur le lexique, nous sommes donc amené à prendre position par rapport aux concepts d'exhaustivité et de normalisation inhérents aux approches terminologiques (2.1.3). Pour l'accès aux documents et à leur contenu, des théories linguistiques et sociales des phénomènes sous-jacents peuvent être proposées comme une alternative aux approches logiques et conceptuelles. Certaines solutions hypermédias les exploitent déjà. Nous les présentons (2.1.4) pour exposer en quoi nous nous en inspirons pour notre modèle et nos réalisations logicielles.

Les constats dressés (2.1.5), nous amènent, dans une deuxième partie, à mettre en place et à justifier les fondements théoriques et opératoires de notre modèle. Nous faisons l'hypothèse que certaines prises de positions pragmatiques d'un sujet interprétant sont fonction des contraintes prescrites par le matériau linguistique *et* de la situation de la tâche en cours. Elles peuvent être modélisées en interaction avec la machine. Cette modélisation s'effectue à travers des principes de catégorisation différentielle et de description componentielle du lexique inspirés de la valeur saussurienne (2.2.1), de la Sémantique Interprétative de Rastier (2.2.2) et d'un modèle de catégorisation différentielle initialement utilisé par Beust dans le cadre d'un modèle interactionniste du sens (2.2.3). Nous précisons enfin dans la partie 2.2.4 quels rôles sont alors alloués à la machine et quels sont ceux qui relèvent de l'utilisateur dans la construction des ressources et dans les processus.

2.1 Accès aux documents et à leur contenu

Dans cette partie, nous présentons les approches couramment utilisées en informatique pour faciliter l'accès aux documents à leur contenu. Nous constatons leurs limites et nous envisageons les techniques qui peuvent être retenues pour y remédier.

En 2.1.1, nous nous attarderons sur les ontologies et l'utilisation qui en est faite actuellement dans le cadre du Web Sémantique pour tenter de pallier les manques des systèmes classiques de recherche documentaire que nous avons présentés dans le premier chapitre. Ces solutions sont inspirées de propositions anciennes de représentation des connaissances encore utilisées en TAL pour « modéliser et formaliser le sens » d'énoncés ou de textes. Ces propositions seront présentées en 2.1.2. Certaines pratiques de ces approches sont en relation avec la terminologie, ce constat nous permettra en 2.1.3 de prendre position par rapport aux concepts d'exhaustivité et de normalisation lexicale. Enfin, nous présenterons en 2.1.4 des approches hors de la sphère logico-grammaticale. Il s'agit des systèmes hypermédias qui nous semblent plus à même d'apporter des solutions aux buts que nous poursuivons. Nous conclurons finalement sur cette partie en 2.1.5.

2.1.1 Les ontologies et le web sémantique

L'ontologie est originellement une branche de la philosophie qui a pour objet l'être en tant qu'être, c'est-à-dire l'étude des propriétés générales de ce qui existe par opposition à l'axiologique, science et théorie des valeurs (morales). Dans le cadre de la représentation des connaissances en IA, le terme désigne un modèle conceptuel de la représentation formelle de concepts, d'objets et d'entités et

de leurs relations. Le modèle conceptuel peut avoir pour vocation de décrire le monde [Lenat *et al.*, 1986]. Il peut avoir une valeur universelle et être réutilisable à volonté (voir par exemple le projet *Ontology Server Projects*[22] dont les prétentions en la matière ont été infléchies ces dernières années). Mais l'ontologie qui s'attache à décrire le monde, l'univers et le reste est de plus en plus considérée comme l'inaccessible étoile, non seulement pour des raisons terre à terre comme le coût économique et humain d'une telle entreprise, mais surtout pour des motifs qui sont davantage philosophiques : les connaissances peuvent-elles être représentées par des symboles et manipulées par des règles hors de tout point de vue [Rousselot et Frath, 2002] ? Les efforts sont dès lors concentrés sur des ontologies spécialisées (ou locales, régionales) consacrées à des domaines restreints comme dans [Assadi, 1998] par exemple.

Pour construire des ontologies, il faut définir des concepts et leurs relations. L'approche conceptuelle nécessite tout d'abord de décider des primitives constituantes des ontologies. Ces informations sont généralement collectées par des experts du ou des domaines et/ou à l'aide d'outils informatiques pour l'extraction de connaissances à partir de corpus. Ces outils utilisent des méthodes statistiques (comme le système ANA [Enguehard, 1992]) auxquelles s'ajoutent souvent des considérations morphologiques et syntaxiques (par exemple XTRACT [Smadja, 1992] et LEXTER [Bourigault, 1994] et son successeur UPERY [Bourigault, 2002]). Les résultats fournis sont alors principalement des syntagmes nominaux plus ou moins complexes (parfois rassemblés en regroupements significatifs à partir d'analyses distributionnelles comme pour MANTEX [Frath *et al.*, 1995]) qui sont ensuite placés dans des structures de données à héritage comme les *frames*, les logiques de descriptions, les ontologies formelles etc. Notons que les travaux présentés dans [Habert et Nazarenko, 1996] qui exposent une étude pour l'extraction de connaissances à partir de corpus à l'aide de méthodes linguistiques, montrent que le travail automatique doit être nécessairement complété par une interprétation humaine et que les résultats obtenus (en l'occurrence des regroupements sémantiques de termes) ne sont finalement valables que pour le corpus traité.

Les ontologies connaissent un nouveau succès depuis que l'idée du Web Sémantique (WS) a été émise par Tim-Berners Lee à la fin des années 1990 [Berners-Lee, 1998]. Le WS se veut une extension du web actuel pour les besoins des entreprises, visant à rendre les contenus, non plus seulement accessibles et affichables, mais également exploitables et interprétables par les machines. Dans le cadre du WS, les ontologies représentent dans un format opératoire les connaissances utiles à une méta-description. Elles peuvent être décrites entre autres à l'aide de RDF (*Resource Description Framework*). RDF permet de définir des triplets d'association de type (entité)/propriété/valeur. Ces triplets permettent d'exprimer, par exemple, qu'un département fait partie d'une région. Les ontologies sont alors des suites de triplets associées à des règles d'inférence du type « si un code département est asso-

[22] http://www.ksl.stanford.edu/

cié à un code de région et qu'une adresse utilise ce code département, alors cette adresse est associée au code région ». Les ontologies sont les supports des *méta-données* qui sont utilisées pour qualifier et décrire le contenu des documents. Ces descriptions facilitent leur indexation et leur accès direct via des logiciels d'interrogation. Les systèmes de calcul combinent les données issues de sources hétérogènes et produisent de nouvelles informations susceptibles d'être plus précises, moins redondantes, et enrichies par la présence éventuelle de descriptions ontologiques. Un certain nombre de langages informatiques, de modèles et de normes ont été créés à cet effet (RDF et RDF Schema pour la mise en place des relations, Topics Map pour l'annotation, DAML+OIL et OWL Lite, DL ou Full pour la définition de classes et de types de propriétés, WSDL et DAML-S pour la description des services possibles, DAML et SHOE pour la description des ontologies, etc.[23]) qui permettent entre autres d'associer au mot relié à un concept, un certain nombre de descriptions et de règles d'associations entre concepts.

Le rapport de l'Action Spécifique 32 CNRS/STIC de 2003 [Charlet *et al.*, 2003*] présente l'état actuel du WS. Il en donne une vue synthétique à ce jour et propose de nombreuses perspectives de recherche pour le domaine. Ces perspectives sont en forme de pré-requis techno-scientifiques à résoudre pour mener à bien un projet qui monopolise des moyens financiers et humains très importants mais dont la mise en place effective se heurte toujours à de nombreux problèmes. Ainsi, la conclusion du document parle d'obstacles *particulièrement cruciaux pour* les débuts même *du Web sémantique* [*ibid.* : 123]. Ces obstacles concernent entre autres l'absence de consensus sur les langages et les modèles informatiques à utiliser mais surtout, c'est la possibilité même de l'utilisation généralisée de méta-données qui est mise en doute : *la détermination et l'ajout, même de simples méta-données, n'est pas une activité naturelle pour la plupart des personnes. (…) Les expériences dans la construction d'ontologies sont (…) instructives et pourraient contribuer à lever quelques illusions.* [*ibid.* : 124]. Ce problème existait déjà il y a dix ans lorsque certains moteurs de recherche de l'Internet n'utilisaient que les balises HTML « `meta` » pour l'indexation (c.f. note 11 p.17).

Deux principales voies de recherche sont à ce jour envisagées pour l'extension du WS. La première consiste à créer des mécanismes inférentiels et de sémantique formelle toujours plus puissants. La seconde, plus semblable à notre approche (c.f. infra), met l'accent sur des représentations qualifiées de semi-formelles dont l'exploitation opérationnelle repose davantage sur l'utilisateur puisqu'il est désormais reconnu que *la volonté d'automatiser à outrance n'est certainement pas une voie réaliste* [*ibid.* : 107]. Ce constat est également proposé dans [Folch et Habert, 2004] où les auteurs proposent sur ce point une distinction entre les langages de méta-données du WS : RDF et Topic Maps. Selon ces auteurs, RDF est surtout valable dans le cadre *d'une démarche formalisante poussée,*

[23] Le manque certain de lisibilité sur les objectifs de ces langages, aggravé par leur multiplicité, est reconnu comme un problème majeur par la communauté du Web Sémantique [Charlet *et al.*, 2003 : 20].

contrôlée par une ontologie partagée, rendant possibles inférences et remodelages et limitant l'intervention humaine [*ibid.* : 70]. Tandis que Topic Maps[24] est plus adapté aux ressources du Web qui font coexister des points de vue différents et pour lesquelles la qualification humaine des informations reste centrale. Nous noterons alors que les espérances initiales d'exploitation et d'interprétabilité des contenus par les machines sont abandonnées en cours de route. Une partie importante de ce rapport est ainsi consacrée à l'adaptation et la personnalisation [*ibid.* : 71-91] considérées toutes deux comme des points clés pour *l'utilisation, la vente et l'accès aux services et documents par les entreprises et les particuliers*. Les approches préconisées dans le cadre de la création de services et de documents adaptatifs sont :

I. La mise au point de « profils d'utilisateurs », « profils comportements » ou de « profils d'intérêts ». Très demandée par l'industrie et en particulier par les e-commerçants, la mise au point de tels profils n'est pas sans poser des problèmes concernant la protection de la vie privée des usagers. Leur réalisation reste peu aisée en ce qui concerne la pertinence des informations à collecter et les services à mettre en œuvre en fonction de ces informations. La constitution de profils sous-entend parfois une uniformisation ou une standardisation des pratiques. De plus, il devient difficile de les modifier dès que de nouveaux usages ou de nouveaux besoins apparaissent. L'approche à base de profil pour la personnalisation ne semble donc pas adaptée à nos objectifs d'autant que ces propositions ne relèvent pas du contenu textuel proprement dit.

II. La mise en place de systèmes d'aide à la sélection d'ontologies par les utilisateurs : ce travail de grande ampleur suppose la compréhension et l'acceptation des ontologies par des non-spécialistes de la formalisation des connaissances dans le cadre des pratiques en cours.

III. La prise en considération des apports de la sociologie, de la psychologie cognitive et de l'ergonomie cognitive pour l'évaluation de l'utilisabilité, de l'utilité, de l'acceptabilité et de l'ergonomie des services de documents ainsi que pour la présentation de l'information, les formats et la compréhension.

Les problèmes sous-entendus par la proposition III ne sont pas inhérents au WS : nous avons pu montrer qu'ils étaient communs à de nombreuses études en linguistique de corpus et inhérents à la recherche d'information dans des documents numériques [Perlerin et Ferrari, 2003b]. Les manques actuels concernent entre autres l'accès à l'information pertinente (zones pertinentes dans un document, …), la navigation dans un grand espace de ressources (accès aux documents pertinents dans un ensemble,…) et l'assistance logicielle pour la compréhension d'une ressource complexe (l'interprétation et la compréhension d'un document ou d'une partie de document, la compréhension du fonctionne-

[24] http://www.topicmaps.org

ment du système informatique médiatisant l'information,…). Nous aurons l'occasion de préciser combien ces aspects sont en effet cruciaux pour les tâches documentaires informatisées.

Le WS vise à un abandon (partiel) des méthodes de calculs d'occurrences et d'indexation à partir de simples mots-clefs et préconise finalement un retour aux propositions anciennes de représentation de connaissances pour permettre un accès au contenu même des documents. La partie suivante revient succinctement sur les principes mis en jeu dans les modèles de représentation des connaissances, initialement envisagés comme des modèles de la mémoire des individus pour l'IA.

2.1.2 Représentation des connaissances

En IA, des modèles de la mémoire sont utilisés, le plus souvent implicitement, dans la manière de définir les bases de connaissances ou les bases de cas. Nous avons pu en démontrer l'importance pour l'ingénierie du logiciel et en présenter un historique dans [Nicolle *et al.*, 2002]. À la fin des années 1970, les capacités des machines ont permis l'émergence de travaux qui avaient pour but de modéliser la mémoire sémantique, i.e. celle de la structure des choses, de leurs relations, de leurs fonctions et de leur genèse. La mémoire sémantique contient le langage, les codes, les règles et les connaissances et ses tentatives de modélisation les plus reconnues sont les systèmes de productions et les réseaux sémantiques. Ces propositions sont contemporaines des *frames* et des *scripts* (ou scénarii) qui tentent de modéliser l'organisation de la mémoire épisodique : cette dernière est constituée des événements et des sensations, faits, dates et lieux qui leur sont associés – il s'agit d'une mémoire peu fiable chez l'homme, comme le montrent les contradictions dans les témoignages de personnes de bonne foi.

Les réseaux sémantiques ont eu d'abord pour objet de représenter la mémoire sémantique des êtres humains sous forme de graphes dont les nœuds représentent des concepts et où les arcs figurent les relations entre ces concepts. La relation avec les ontologies actuelles est explicite. Dans les réseaux sémantiques, les principales relations sont les relations de généralisation-spécialisation entre les concepts, les relations entre un objet complexe et ses parties, et les relations structurelles entre certains concepts (comme les jours de la semaine ou les couleurs de l'arc-en-ciel). Les algorithmes de parcours de ces graphes à partir d'instances de concepts seraient l'analogue des activations de la mémoire [Falhman, 1979]. Les réseaux sémantiques ont été utilisés pour améliorer les systèmes experts en associant des règles aux concepts au lieu de les présenter en vrac. Avec de tels modèles, il est en effet possible de ne se concentrer que sur les règles relatives au problème en cours. Les hiérarchies de généralisation entre les concepts, qui représentent des connaissances ontologiques, sont le mode d'organisation le plus souvent utilisé mais on utilise également la logique formelle. Elles permettent de factoriser les connaissances, en les associant au concept le plus général pour lesquelles elles sont pertinentes et en les évoquant par héritage pour les concepts plus spécialisés. Dans les systèmes de

productions purs où la base de connaissances n'est pas structurée, on ne peut pas présélectionner les propriétés pertinentes, il faut toutes les examiner à chaque cycle d'inférence. Les réseaux sémantiques (généralisés par Sowa [Sowa, 1984] en graphes conceptuels) sont parfois utilisés pour « représenter la sémantique » d'une phrase ou d'une scène [Nazarenko, 1994], dans les faits, pour produire un jeu de formules logiques censées formuler le sens d'un énoncé. Ces modèles restent très en deçà des subtilités des langues en particulier parce qu'ils conçoivent le signe de manière isolé et ne prennent donc pas en considération les co-adaptations des significations des constituants d'un énoncé ou d'un texte. Ils établissent des représentations propositionnelles canoniques du sens qui ne peuvent rendre compte des diverses interprétations possibles d'un énoncé.

Les *frames* [Minsky, 1975] (ou schémas) et les scripts [Schanck et Abelson, 1977] sont inspirés de travaux menés en psychologie cognitive sur la mémoire de l'Homme. Ils ont pour objet la représentation synthétique de la mémoire épisodique. L'hypothèse sous-jacente est que les situations ne sont pas stockées de manière complète et successive mais qu'elles sont organisées en une partie générique : on trouve les schémas de situations fréquemment rencontrées, et une partie spécifique à une situation donnée. Une situation particulière est donc appariée à un ou plusieurs schémas connus et mémorisée en terme de schémas et particularités. Il existe des schémas plus ou moins généraux organisés les uns par rapport aux autres. Les schémas mémorisés peuvent être évoqués pour résoudre des problèmes, pour « comprendre » des histoires, en complétant les descriptions données explicitement par référence aux situations les plus courantes. Dans la conception des schémas, la principale difficulté est le choix des informations pertinentes à mémoriser. Chaque situation peut donner lieu à une infinité de descriptions dont la pertinence et le niveau de détail dépendent du contexte : il faut choisir celles qui seront considérées comme assez générales pour faire partie d'un schéma. Avec l'utilisation des *frames*, le principal problème est l'évocation du ou des schémas pertinents. Contrairement à l'appariement entre les règles de production et la base de faits qui est toujours strict, la comparaison entre un schéma et une situation doit accepter des différences si elles sont moins significatives que les ressemblances. Utilisés en TAL, les schémas sont des ensembles structurés de connaissances autour d'un mot donné. Lors d'analyses à partir de schémas, on fait correspondre chaque mot avec son schéma et de ce point de vue, les problèmes sont les mêmes que ceux évoqués pour les graphes conceptuels.

Minsky fut l'un des consultants de Stanley Kubrick pour le film *2001, l'Odyssée de l'espace*. En 1968, il avait affirmé au réalisateur britannique que les machines de 2001, grâce aux représentations conceptuelles du type *frame* (qui ne s'appelaient pas encore ainsi), seraient capables de parler couramment avec les humains à l'horizon d'une trentaine d'années. Comme souvent lorsqu'il s'agit d'anticipation, les faits ont prouvé qu'il n'en était rien. D'une manière générale, les *frames* et les autres propositions du type permettent une représentation de connaissances conceptuelles sur le monde et

son organisation. En les utilisant dans le cadre du TAL, il est possible de résoudre par exemple des problèmes d'indexation ou de raisonnement automatique ciblé mais ces structures ne s'intéressent pas directement à la langue dans le discours. Leur portée s'en trouve donc limitée puisqu'en particulier, elles tendent à isoler le signe et à faire abstraction des sujets interprétants. Sowa lui-même avoue désormais les faiblesses de cette approche dans [Sowa, 2001]. Il rapporte par exemple que Winograd, le père du système de dialogue SHRDLU, se tourne à l'heure actuelle vers le domaine de l'Interaction Homme-Machine (IHM) pour tenter enfin de prendre en considération le contexte, le cadre de référence et les exceptions dans ce type de problématique[25].

La modélisation des connaissances, qu'elle soit effectuée dans le cadre de travaux sur la mémoire ou pour le TAL, a des points communs avec la terminologie. Nous allons en examiner certains dans la partie suivante.

2.1.3 Terminologie et linguistique

La terminologie trouve son origine dans l'antiquité grecque mais son développement moderne se situe au début du siècle dernier, lorsqu'on éprouve le besoin de normaliser le vocabulaire de la commission électrotechnique internationale (CEI). La terminologie en tant que science a pour origine les travaux d'Eugen Wüster (Théorie Générale de Terminologie, 1931). À l'époque, elle se fondait sur les travaux du Cercle de Vienne, dont un des objectifs fut de séparer la philosophie et la science de la métaphysique [Soulez, 1985]. Des savants tels que Carnap, Schlick et Neurath préconisèrent alors *la recherche d'un système formulaire neutre, d'un symbolisme purifié des scories des langues historiques* [*ibid.*]. L'approche est encore ici conceptuelle : le terme désigne un concept, lui-même relié à d'autres concepts dans une taxinomie. La relation entre mot et notion est considérée comme univoque ou peut faire l'objet d'une normalisation par le terminologue. Dans ces conditions, la terminologie est donc prescriptive au contraire de la linguistique qui se veut plus descriptive. On n'est pas étonné de constater que Wüster est l'un des pères spirituels de l'ISO (*International Organization of Standardization*).

La terminologie se place parfois dans l'optique logique où l'on assigne au terme un statut le faisant échapper aux contraintes linguistiques d'ambiguïté et de signification contextuelle. Condamines [Condamines, 1994] affirme ainsi que *la connaissance scientifique procédant du raisonnement logique, il est possible de bâtir un système sémiotique optimal entièrement fondé sur la logique. L'unité minimale est le terme, 'pur' de toute connotation, univoque, monoréférentiel et précis*. Ces positions radicales ont été discutées par des acteurs impliqués dans le TAL. Dans [Bourigault et Slodzian, 2000], la terminologisation est ainsi présentée comme un processus parallèle et non postérieur à l'élaboration conceptuelle. L'antériorité des notions sur les mots est réfutée. Le terme n'est plus *hors-*

[25] C.f. la page personnelle de Terry Winograd : http://hci.stanford.edu/winograd/

langue mais bien plongé dans le champ des contraintes liées à son statut *en langue* (on parle alors de linguistes-terminologues). Dans ces conditions, la construction d'une terminologie ne peut se soustraire à une analyse linguistique. Ainsi, en terminologie, le signe peut être linguistique *et* conceptuel : les ontologies (ou les autres propositions analogues de représentation) donnent le primat à l'aspect conceptuel, tandis que les travaux en acquisition de connaissances à partir de corpus prennent aussi en considération l'aspect linguistique.

Les contraintes de la terminologie sont essentiellement l'exhaustivité et la normalisation. Un domaine doit être abordé le plus précisément possible pour que la terminologie qui lui est associée soit efficiente et pérenne. L'ambiguïté des termes doit être prise en considération et décrite en tant que telle dans les structures fournies. Les contraintes d'exhaustivité et de normalisation ne relèvent pas de notre problématique puisque nous constatons la très grande variabilité contextuelle des significations des termes en contexte. En outre, les applications de notre modèle visent une *sémantique légère* (c.f. chapitre 1 - 1.4). Cependant, la terminologie assistée par l'ordinateur présente des aspects communs à notre étude. Il nous faut, en particulier, considérer le terme, non pas comme élément isolé, mais bien comme élément d'un texte ou d'un corpus. La nécessité actuelle d'élaborer des outils d'assistance aux utilisateurs pour l'acquisition de termes depuis un corpus [Nazarenko et Hamon, 2002b*] est également une phase nécessaire à la mise en œuvre de nos propositions. Dans ce domaine, des outils mettant en œuvre des moyens d'interaction dédiés ont déjà vu le jour. Il s'agit par exemple de LEXTER [Bourigault, 1994] (présentation en réseaux), ASIUM [Faure, 2000], PROMETHE [Morin, 1999] ou encore SYNOTERM [Hamon et Nazarenko, 2001] (où l'utilisateur est sollicité pour valider les résultats des calculs). De même, la démarche itérative de certains systèmes [Barrière et Copeck, 2001] avec lesquels il est possible d'augmenter le lexique terminologique initial à la suite de résultats d'analyses de textes, est reprise dans le cadre de nos propositions moyennant une semi-automatisation des analyses et une possibilité d'augmenter le lexique non pas forcément en quantité mais également en qualité (c.f. chapitres 4 et 5).

En informatique, il existe de nombreux algorithmes pour classer automatiquement des mots en fonction de grains contextuels variables (documents, phrases, syntagmes…). Le but est alors de construire des classes de mots cernant un contexte donné. Comme il a été souligné dans [Pincemin, 1999b : 79], les classes sont le plus souvent constituées en partition (un élément ne peut se retrouver que dans une seule classe) et sont considérées sur le modèle des classes d'équivalence, avec les propriétés de réflexivité, de symétrie et de transitivité. On postule donc ici des phénomènes de synonymie et d'homonymie strictes qui n'entrent pas dans les propriétés structurelles de la langue. Lorsqu'à l'inverse, on tente de classer les contextes en fonction des mots pour trouver les termes les plus significatifs de ces contextes (comme dans [Meunier *et al.*, 1997] cité dans [Pincemin, 1999b*]), les classes de contextes ont les même propriétés que celles précitées : l'association univoque classe / contexte in-

duit une distorsion linguistique qui n'existe pas dans la langue ; les résultats obtenus avec ce type de ressource sont intéressants mais ne peuvent pas entreprendre par exemple, des classifications automatiques de textes véritablement en rapport avec leur contenu tel que nous l'envisageons.

S'intéresser au contenu des documents et vouloir en faciliter l'accès vise à rendre compte des thèmes présents dans les textes. Les approches terminologiques tendent ainsi à construire des ressources permettant de détecter les thèmes des documents pour par exemple les classer automatiquement. Dans [Pichon et Sébillot, 1999], on appelle *thèmes*, des listes de mots qui ont trait aux sujets abordés dans un texte ou dans un corpus. L'approche proposée permet la création des regroupements non exclusifs de termes qui figurent dans les mêmes cotextes à partir de la recherche de leurs cooccurrences[26]. Si l'expérience descriptive est très encourageante, les possibilités de personnalisation des données proposées sont limitées puisqu'elles sont obtenues automatiquement et validées par un spécialiste. Ainsi, toutes les méthodes que nous venons d'aborder permettent effectivement d'obtenir des résultats logiciels intéressants mais l'absence d'interaction avec l'utilisateur ne permet pas de répondre aux deux exigences que nous avons formulées pour notre modèle : la possibilité d'être assisté de façon personnalisée pour l'accès au document *et* l'exploration de son contenu.

2.1.4 Subjectivité, hypermédias et interprétation

Les communautés de la recherche d'information et de la modélisation des connaissances ne sont pas les seules à s'intéresser à l'accès aux documents numériques. Les notions d'intertexualité et d'intersubjectivité autour de l'interprétation de textes ont mené depuis de nombreuses années à l'élaboration de modèles hypermédia qui permettent des parcours de lecture personnalisés et assistés dans des ensembles documentaires. Ils permettent aussi l'annotation, la conservation et le partage de ces parcours. Ces études participent à la transformation de la machine en un média à valeur ajoutée pour l'accès personnalisé aux documents et à leur contenu ; nous en proposons donc un rapide exposé, inspiré entre autres de Benel [Benel, 2003 : 41-50]. Nous distinguerons deux familles de solutions dans ce domaine : celles inspirées de l'hypertexte, qui consistent à associer des fragments ou « blocs d'informations » par l'intermédiaire de liens ou de relations, et celles qui sont plus ancrées dans des considérations linguistiques, documentaires et herméneutiques.

Le *World Wide Web* (ou l'Internet), par le truchement des hyperliens, permet des parcours de lectures libres, ou tout du moins, non linéaires. La mise en place effective des pages dynamiques, en particulier avec le développement du PHP, des CGI et de l'ASP, apporte certes des solutions de haut niveau pour la personnalisation des parcours mais leur subjectivité est contrainte par les choix de

[26] Ces regroupement figurent les isotopies de la SI puisqu'ils évoquent un thème grâce à des éléments de sens qu'ils partagent. Ces éléments de sens sont figurés par la dénomination de ce thème.

l'auteur du document en cours de consultation[27]. La conservation de ces parcours n'est pas aisée [Delepine, 2003] et l'accès au contenu des documents reste dépendant des systèmes cités dans les parties précédentes. Si l'exploration des documents peut être facilitée par ces parcours, elle n'en est pas davantage directement assistée par les logiciels.

L'Internet a depuis ses débuts été fortement influencé par les travaux de Ted Nelson et son projet Xanadu[28] qui date des années 1960. La principale originalité du système provient de la présence de liens bidirectionnels inamovibles qui permettent de concevoir un document comme un ensemble de fragments d'autres documents et de liens réutilisables. Ce système concrétise donc un intertexte explicite pour un document. Du point de vue de la subjectivité, l'utilisateur a la possibilité de créer des documents à partir de fragments de documents déjà existants et d'expliciter et de concrétiser ainsi ses références. Ces fonctionnalités se retrouvent dans Hyper-G et son avatar commercial Hyperwave [Maurer, 1996] qui rend, de plus, possible la création de *collections* dans lesquelles on peut placer des documents et d'autres collections. Ce principe est maintenant exploité dans les annuaires de recherche des moteurs de l'Internet. Il s'agit là encore d'une assistance au parcours de lecture et à sa mémorisation.

L'ATLAS.ti [Muhr, 1994] a la particularité de proposer au lecteur la définition de citations, qui sont des fragments sur un *document primaire*. Les citations sont reliées à l'aide d'hyperliens et décrites par des *codes*. Des codes communs à plusieurs citations sont reliés à d'autres par des liens de relation de cause, d'équivalence ou de généralisation par exemple. Les codes, les citations ou les documents primaires peuvent eux-aussi être commentés à l'aide de mémos. Les codes, les documents primaires et les mémos peuvent être regroupés en *familles*. Enfin, un *supercode* permet une définition en intention des citations qu'il décrit. L'utilisateur a ainsi la possibilité de créer ses propres commentaires et ses propres familles pour faciliter le parcours et la conservation de documents. On trouvait également une partie de ces fonctionnalités dans le logiciel Ideliance de Jean Rohmer et Sylvie Le Bars de la société Dallas mais celui-ci semble à ce jour abandonné et n'a pas fait l'objet, à notre connaissance, de publications.

Parfois désignés sous le terme d'*hermeneutics softwares*, ces solutions informatiques permettent essentiellement d'organiser et de commenter de façon personnelle des ensembles de documents. Elles mettent en place des principes de conservation, d'annotation et d'organisation pour aider la réalisation de projets ou la prise de décision. C'est aussi le cas pour TheBrain[29] de Hugh qui est fondé ~~sur les trois notions de *contenu*,~~ de *pensée* et de *relation*. À chaque pensée, il est possible d'associer un

[27] Les Smart Tags, Smart Links ou Ezula Top Text pourraient bien modifier cette donne puisqu'il est désormais possible de transformer automatiquement n'importe quel mot ou groupe de mots sur une page Web en un hyperlien vers un site ou un service commercial d'une entreprise, et cela sans l'accord préalable de l'auteur de la page en question.

[28] http://xanadu.com

[29] http://www.thebrain.com

les trois notions de *contenu*, de *pensée* et de *relation*. À chaque pensée, il est possible d'associer un contenu et les pensées sont associées par des relations de paternité ou de saut. Ce système permet de mettre en place des réseaux de notes et de documents pour créer des répertoires personnels ou des réseaux de façon collaborative. Il est lui-même inspiré du Memex de Bush [Bush, 1945] qui envisageait, dès les années 1940, de « mécaniser » les dispositifs de stockages des livres, disques et autres supports pour les rendre plus rapidement consultables et les transformer en un *supplétif agrandi de la mémoire*.

Le système Porphyry[30] [Benel, 2003*] s'adresse à des communautés d'experts appelées à travailler sur des corpus numérisés de documents. Il est fondé sur l'enrichissement itératif des corpus par des structures hypermédias. Ces structures sont construites par les experts en fonction de leurs problématiques et de leurs spécialisations. Ces derniers expriment leur point de vue sur le corpus à l'aide d'objets documentaires (fragments, graphiques, notes…), de parcours de lecture (permettant d'ordonner le corpus selon un ordre temporel, spatial…) et des réseaux de description (permettant d'organiser le corpus en le décrivant de manière semi-formelle). Le but principal de l'application est d'assister le chercheur dans la lecture de publications savantes en le soulageant des aspects répétitifs de son activité et en le laissant se concentrer sur les aspects créatifs, intuitifs et à haut niveau d'abstraction. Porphyry ne s'appuie pas sur la logique ou une approche conceptuelle mais sur la Sémantique Interprétative (SI) de Rastier [Rastier, 1987*], en tant qu'il met en place les moyens de tenir compte de traces d'une interprétation dans le processus de lecture assistée (ces traces sont exploitées pour l'indexation des documents). Nous aurons l'occasion de revenir sur la SI en tant que source d'inspiration de nos travaux. Pour l'heure constatons simplement que Porphyry s'adresse essentiellement aux chercheurs et demande aux utilisateurs des efforts importants pour la mise en place des principes de personnalisation.

PASTEL de Tanguy [Tanguy, 1997b] s'inspire aussi de la SI pour l'analyse de textes (PASTEL correspond à « Programme d'Aide à l'Analyse Sémantique de TExtes, même Littéraires »). De l'aveu même de l'auteur, PASTEL ne sert en fait qu'à *remplacer (…) un support classique du type papier-crayon* en prenant en charge *le côté calculatoire de l'interprétation*. Le côté calculatoire s'appuie sur les mécanismes décrits par la SI (voir 2.2.2). Dans les faits, le logiciel permet de partager avec la machine les contraintes d'une interprétation formalisées à partir des concepts de la SI pour mieux les objectiver et les réutiliser. Le lecteur/utilisateur partage une première vision du texte avec le logiciel, qui en retour, peut l'amener à la découverte de nouveaux aspects sémantiques relatifs au matériau analysé à l'aide d'informations textuelles et d'une interface rudimentaire. PASTEL et le modèle sous-jacent sont d'une qualité rare car ils permettent la prise en considération des contraintes du matériau textuel pour l'assistance à l'interprétation, tout en posant clairement les limites de ce qui est accessible à la machine et ce qui relève de l'humain. Cependant, PASTEL présente deux défauts majeurs par

[30] http://www.porphyry.org

rapport aux buts que nous poursuivons : il s'adresse essentiellement à des spécialistes de l'interprétation de textes (il manipule des concepts complexes peu accessibles à des novices et les résultats des analyses valent principalement pour service rendu – la découverte d'autres aspects que ceux repérés est possible), il est particulièrement approprié pour l'analyse d'*un* texte (court) et son utilisation par exemple à l'échelle d'un corpus n'est pas envisagée.

La plupart des systèmes que nous venons de présenter permettent aux usagers de personnaliser leurs interactions avec la machine dans des tâches documentaires. Loin de rendre les services des moteurs de recherche classiques, ils permettent de tenir compte d'un certain point de vue sur l'exploitation du matériau textuel, et pour certains, de prendre en considération la dimension intertextuelle d'une lecture. Lorsqu'ils sont dédiés au travail collaboratif, ils mettent alors en place des principes d'intersubjectivité. La voie empruntée par Porphyry et surtout PASTEL, nous semble la plus en adéquation avec nos objectifs. La Sémantique Interprétative semble une théorie valable à exploiter pour les mener à bien. En effet, elle cherche à caractériser le sens des mots en restant au niveau linguistique, sur la base des contributions rencontrées au sein de textes ou d'énoncés et cela, à partir des *différences* d'emplois. Les travaux de Tyavert présentés dans le premier chapitre rejoignent d'ailleurs ces principes.

2.1.5 Conclusion

Le fait que les travaux du WS ou que des chercheurs comme Winograd se tournent maintenant vers l'utilisateur en tant que *comprenant des ressources* est symptomatique de manques évidents des systèmes qui font abstraction des usagers pour ce qui touche au sens. Ceci révèle dans le même temps une prise de conscience des limites de l'automatisme (et de l'universalisme) pour l'accès au contenu des textes. Les ressources conceptuelles sont pour la plupart construites *a priori* ou tout du moins indépendamment de la situation concrète de l'usager ; à l'heure actuelle, les moyens de personnalisation proposés à partir de ces techniques sont de ce fait limités. Mais un autre point a de plus amples incidences sur l'utilisation de ces techniques en TAL : comme nous allons le montrer, ces dernières sont en effet en contradiction avec la nature structurale de la langue.

Dans le premier chapitre, nous avons succinctement abordé les travaux de recherche documentaire qui se fondent sur la simple présence de mots isolés. L'absence de prise en considération des contextes d'apparition des mots est une limite avouée de ces systèmes. Pour pallier ce manque, on envisage très couramment de compléter une requête en ajoutant des synonymes, des hyperonymes ou des hyponymes de chaque mot-clef à partir d'un dictionnaire ou d'un thésaurus [Fluhr, 2000] et [Hiyakumoto et Veloso, 2002]. L'idée est de pallier les manques des approches du type de celles qui utilisent des ontologies générales et qui sous-entendent qu'il y a une correspondance entre les mots et les concepts ; ici on tente de cerner une idée, un concept, à l'aide de plusieurs mots en relation dans un

support lexical structuré. Dans les faits, il s'avère que de telles extensions de requête ne permettent pas le repérage *du* concept associé au mot d'une requête dans tous les textes. Ces travaux reposent sur des phénomènes qui ne se retrouvent pas dans la langue : des synonymies et des homonymies exactes, aucun glissement de sens possible, etc. Les résultats obtenus ne sont d'ailleurs pas à la hauteur de la quantité de données à fournir au système. Ces approches sont exemplaires car elles sous-entendent deux propriétés de la sémantique des langues couramment envisagées dans les travaux de TAL (même avec l'utilisation des ontologies ou de la logique formelle) : la compositionnalité et la référentialité.

Comme nous l'avons déjà signalé dans le premier chapitre, la vision compositionnelle amène à considérer que la signification de toute phrase est fonction des significations de ces parties, ce qui signifie que le sens de chaque unité lexicale 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. De nombreux arguments ont été avancés contre cette vision compositionnelle du sens mais celle-ci demeure sujette à débat[31]. Il apparaît donc important d'y revenir dans cette partie. La compositionnalité sémantique s'oppose à une vision de co-détermination des mots par les textes et des textes par les mots. [Tyvaert, 2003* : 49] dit que « *chaque mention d'un mot du lexique dans un texte est (…) l'occasion d'une élaboration particulière de son sémantisme en une signification qui dépend des autres mots du texte, cette signification dépend du texte où la présence des mots est observée* ». Un exemple simple suffit pour illustrer cette position. Dans le fameux énoncé *le chat mange la souris*, le chat n'est pas nécessairement celui que l'on croit en fonction des phrases précédentes ou suivantes dans un véritable contexte. Si l'on considère que le chat est un mammifère carnivore digitigrade domestique, on aura du mal à interpréter une phrase atroce et tristement célèbre de notre histoire récente : *Mais celui qui détruit la vie s'expose lui-même à mourir (…). Qui doit-on blâmer, le chat ou la souris, si le chat mange la souris ?*[32] Les comparaisons ou les métaphores ne sont pas les seules formes à mettre à mal la compositionnalité (voir par exemple [Thlivitis, 1998 : 32-33]) : c'est bien la contextualisation qui permet d'envisager la signification d'un terme, vu l'influence entre les termes d'un même texte sur l'interprétation que l'on peut en faire.

Les réseaux sémantiques comme les *frames* ou encore les logiques de descriptions participent à un courant logico-informatique ayant pour but la modélisation des connaissances nécessaires au TAL par la logique formelle. Les ontologies ne sont finalement que les représentants les plus modernes de ce type d'approche. Celles-ci profitent souvent d'une assise mathématique et logique leur assu-

[31] L'introduction de l'appel à soumissions pour un numéro de la revue TAL en 1997 consacré à au principe de compositionnalité sémantique débutait par ces mots : « *Le principe de compositionnalité constitue une référence omniprésente dans les travaux de sémantique dans le domaine du TAL sans pour autant être réellement discuté, à l'exception des travaux qui s'en démarquent et le critiquent comme ceux de F. Rastier, par exemple ».* Adeline Nazarenko, http://www.atala.org/tal/appel-compositionnalite.html

[32] http://www.phdn.org/histgen/hitler/declarations.html - site « *Pratique de l'histoire et dévoiements négationnistes* » de Gilles Karmasyn.

rant une certaine respectabilité académique. Si l'on s'intéresse à l'interprétation des textes en tant qu'activité humaine, ces approches compositionnelles ne sont pas envisageables puisqu'elles délaissent le contexte et la situation de lecture au seul profit de l'ordre référentiel. La vision référentielle de la sémantique des langues correspond à un appariement de la langue avec des concepts extra-linguistiques. La référence pose ainsi le problème du rapport de la langue à la réalité, de la langue au monde. Les textes renvoient à quelque chose, évoque quelque chose chez le lecteur ; ils permettent de décrire des expériences de pensées ou de perception. Mais les théories linguistiques divergent pour rendre compte de cette réalité. Par exemple, Fodor [Fodor, 1975] considère que les *mots de la pensée* sont en relation directe biunivoque avec des objets du monde réel. Au contraire, Rastier considère les représentations induites par la langue comme des épiphénomènes d'une lecture [Rastier, 2001b] et nomme *ordre référentiel*, cet ordre de description des rapports entre les signes d'un texte et d'autres types d'expérience perceptive ou mentale. Chez Rastier, il existe quatre ordres de la sémantique des textes : les ordres syntagmatiques, paradigmatiques, herméneutiques et référentiels. L'ordre herméneutique est celui des conditions de production et d'interprétation des textes (nous présenterons plus loin les autres ordres). Il comprend par exemple les phénomènes d'interprétation en fonction de la situation, que l'usage fait relever de la pragmatique.

L'une des premières critiques modernes envers l'approche référentielle a été proposée à partir de certaines avancées de la théorie des actes de langage. Fondée dans les années soixante par Austin [Austin, 1962], cette conception fut principalement développée par Searle [Searle, 1969]. Le postulat de la théorie affirme que la production de certains énoncés s'assimile à l'accomplissement d'actions : si agir c'est transformer l'état des choses, parler c'est, également, transformer l'état mental des interlocuteurs. Austin conçoit le langage comme une activité sociale permettant par exemple de faire aussi bien des promesses ou des déclarations, que de baptiser, marier, parier ou faire des assertions. Pour Searle, un texte peut être interprétable de maintes façons différentes. Il admet qu'il y ait une ambiguïté littérale des textes et que la signification littérale d'une énonciation puisse différer avec le temps dans la mesure où ce qu'il appelle le 'réseau' et 'l'arrière-plan' sont sujets à des variations. Il rejette ainsi les théories référentielles en faisant dépendre le sens et l'interprétation au moins d'un cadre temporel. Ces travaux ont montré que l'activité langagière vise autant à agir sur le monde qu'à le décrire ; elle ne met pas en place un rapport de la langue à un réel préexistant par la référence mais un rapport de la langue au monde, qu'elle construit en le décrivant ou en permettant de l'inventer. À cet égard, les textes de Bush et de Borges que nous avons cités dans le premier chapitre, sont singulièrement éloquents : les auteurs de fiction ne sont pas les seuls à suspendre les hypothétiques règles de référence. Dans la tradition logique, le sens repose sur la relation de représentation entre des symboles formels, logiques et des objets du monde. Il s'appuie ainsi sur la triade Mot / Concept / Chose. Dans une telle perspective, l'interprétation d'un mot ou d'un syntagme par un usager est conçue comme l'action qui consiste à associer une occurrence à un type puis à l'identification d'une signification. Cette approche suppose la

définition et le figement des types *a priori* indépendamment d'une pratique. Les approches logiques président à l'institution des types tandis que d'autres orientées vers l'herméneutique président à l'interprétation des occurrences[33]. Le lien entre les deux problématiques n'est à ce jour pas résolu. [Rastier, 1999 : 226] ou encore [Habert, 2000] ont montré que l'approche conceptuelle n'établit aucun lien entre langue et conception du monde. Il n'y a dès lors aucune garantie linguistique sur les données. Dans un cadre informatique, la validité linguistique des données ou des processus n'est bien entendu pas nécessaire pour obtenir des résultats applicatifs. Cependant, les conceptions sous-jacentes de la langue qui permettent ces approches s'interdisent de traiter certains aspects que nous nous proposons au contraire d'aborder ici. Les différents travaux exposés dans les parties 2.1.1 et 2.1.2 ainsi que leurs différents avatars, permettent de mettre en œuvre des systèmes complexes utiles à l'indexation automatique et à la recherche d'information dans des cadres précis d'utilisation, i.e. le plus souvent dans des domaines techniques et pour des langues de spécialité. Ces cadres sont d'autant plus limités pour le WS que les solutions proposées sollicitent toujours, du moins pour la plupart, les producteurs dont la bonne fois et la rigueur sont les premiers garants de la fiabilité de l'ensemble de la démarche[34]. Elles s'avèrent cependant très étroites lorsqu'il s'agit d'offrir des outils d'interprétation ou d'aide à l'interprétation, en bref, lorsqu'il faut faciliter l'exploration *du contenu* des documents à un usager. C'est là une différence fondamentale entre les approches qui visent à extraire et formaliser *le sens* d'un texte et celle qui tendent à en saisir *le contenu*, c'est-à-dire évaluer des critères pour rendre plausible une lecture relativement à un utilisateur et une situation et évaluer les informations pertinentes dans ces circonstances.

La perspective conceptuelle et référentielle, qui vise les entités du monde telles que « nous les distinguons » et dont nous envisageons « les relations », est opposée à la perspective sémantique linguistique qui se contente de décrire la signification des mots [Habert, 2000*]. La principale critique qui est adressée à l'approche linguistique est qu'elle s'enferme dans les distinctions d'une langue et d'une culture déterminées. C'est cependant celle que nous adoptons : nous revendiquons cette limitation puisque nos visées ne sont pas générales, ni exhaustives quant aux représentations manipulées. Notre ambition se limite à favoriser l'assistance à l'interprétation, la prise en considération de dimensions propres à la situation de lecture par un interprétant ou un groupe d'interprétants. Ainsi, même si les positions radicales de Greimas, Rastier et leurs affiliés informaticiens comme Tanguy sont la résultante d'un certain militantisme contre les excès des théories référentielles et mentalistes (ce qui leur est

[33] Considérer la hiérarchie des types comme évidente et donnée *a priori* corrobore d'ailleurs l'utilisation du terme « ontologie » (du grec ontos : « ce qui est »).

[34] Nous avons cité de nombreux exemples de détournements de méta-données pour augmenter le nombre de visiteurs de sites sur l'Internet dans [Perlerin, 2000].

parfois reproché [Frath, 1997 : 102]), leurs propositions semblent plus à même d'apporter des solutions adaptables à notre problématique et à notre façon d'en fixer les contours[35].

Contrairement aux systèmes Porphyry ou PASTEL, il nous faut exploiter les principes de la SI pour accéder à des documents non encore connus de l'utilisateur et cela sans nécessairement mettre en place un index ou le solliciter tout au long de ses lectures. Nous essaierons ainsi d'obtenir des solutions personnalisées pour l'analyse du contenu des documents en tentant de soulager l'usager au niveau de la maintenance des ressources et du temps consacré à leur mise en place.

Il s'agit désormais d'exposer les principes théoriques et opératoires que nous avons retenus de ces systèmes pour nos travaux.

2.2 Fondements

Dans cette partie, nous dévoilons les fondements théoriques de nos travaux que nous justifions par rapport aux applications et aux systèmes déjà présentés. Nous commençons par un constat : un même terme peut avoir trait à divers concepts ou choses du monde réel ou imaginaire en fonction de son utilisation dans un texte. Il peut également apporter différentes évaluations et sa participation à un texte peut présenter des intérêts divers en fonction de la tâche située de l'utilisateur. Pour pouvoir tenir compte de ce phénomène, il nous faut tout d'abord expliquer comment nous traiterons du rapport entre sens et référence. Pour cela, nous nous appuyons sur Saussure et le concept de valeur des signes (2.2.1). Ces considérations nous permettent au passage d'obtenir un principe d'organisation de nos ressources : un principe différentiel.

Notre approche est interprétative dans le sens où elle se focalise sur l'interprétant (humain) dans son activité d'interprétation de textes, c'est donc tout naturellement que nous nous tournons vers la Sémantique Interprétative comme une source d'inspiration pour à la fois, la construction des ressources du modèle et les fondements des analyses à produire (2.2.2). Le principe de structuration computationnelle n'en est pas pour autant réglé et c'est vers les travaux de Beust et le modèle ANADIA que nous nous sommes tournés pour obtenir des solutions attestées dans ce domaine (2.2.3). Enfin, il nous a fallu prévoir les rôles de la machine et de l'utilisateur dans le système (2.2.4).

[35] Ceci est d'autant plus pertinent que notre approche opportuniste de la SI nous permet de nous soustraire aux critiques couramment émises à son encontre [Frath, 1997* : 107-111] : en particulier, nous ne définissons pas nos attributs (pendants des sèmes de la Sémantique Interprétative) en tant qu'objets linguistiques dont l'existence pourrait par exemple être prouvée expérimentalement (c.f. chapitre 3 partie 3.2.1, p.77).

2.2.1 Valeur saussurienne

2.2.1.1 Variabilité contextuelle des significations

La variabilité contextuelle des significations des termes lexicaux est tout à fait remarquable dans les textes. La question de la pertinence des critères ontologiques pour l'analyse et la représentation du matériau lexical est donc posée. Des critères sociaux, culturels et praxéologiques sont souvent plus pertinents que des critères ontologiques. Rastier [Rastier, 1987*] rapporte ainsi le résultat d'une étude sur le contenu du mot *caviar* suite à une enquête au sein d'une population de collégiens. Il apparaît que le trait *luxueux* est le plus fréquemment cité tandis que d'autres traits ontologiques tels que *texture granuleuse* ou encore *salé* ne sont jamais évoqués.

Wordnet 2.0[36], le système de références lexicales de l'université de Princeton, propose les informations suivantes pour le mot *caviar* :

Sense 1 caviar, caviare - (*salted* roe of sturgeon or other large fish; usually served as an hors d'oeuvre) => roe, hard roe – (fish eggs or egg-filled ovary; having a *grainy texture*)	Sens 1 caviar – (œufs *salés* d'esturgeon ou d'autres grands poissons ; habituellement servis en hors d'œuvre). => oeufs de poissons – (ovaires pleins ; ayant une *texture granuleuse*)

Les informations proposées par EuroWordNet[37] sont absolument identiques. Dans une des ontologies produites dans le cadre d'un projet pour le WS[38], le « concept de caviar », est associé successivement à *Fish Roe* (oeuf de poisson), *Fish Product* (produit ichtyologique), *Animal Agricultural Product* (produit agricole animal), *Food Fish* (espèce de poisson comestible), *Sea Food* (fruits de mer) etc. Or, en soumettant au moteur de recherche Google l'unique mot-clef *caviar* sans préciser la langue de recherche, les informations obtenues par l'exploration des cinquante premières pages Internet[39] relativisent l'utilisation de *caviar* comme ayant trait aux œufs d'esturgeon ou d'autres poissons (figure 4, p.53) : 30% des pages ne présente aucune information sur des productions animalières ou comestibles et 4% évoque des aliments salés sans rapport avec les poissons. En revanche, 60% des pages est consacré à l'information, l'achat ou la préparation culinaire des œufs d'esturgeon, 6% concerne les œufs d'autres poissons. Si l'on réitère l'opération en précisant que la langue attendue est le français, les résultats sont très différents (figure 5, p.53). Outre le fait qu'une des pages est en fait en espagnol, on peut voir que seules 36% d'entre elles abordent les œufs d'esturgeon, 31% proposent des recettes à base d'aliments salés (essentiellement des légumes – une seule aborde le hareng mais il ne s'agit pas

[36] Word Net 1.5 : http://nipadio.lsi.upc.es/cgi-bin/wei/public/wei.consult.perl

[37] Web EuroWordNet Interface 0.2 : http://www.cogsci.princeton.edu/cgi-bin/webwn

[38] Projet DAML Agent Semantic Communication Service - ASCS http://reliant.teknowledge.com/DAML/

[39] Les pages explorées correspondent à des sites différents i.e. aux 56 premières URL retournée - 45 documents sont en anglais, 3 en français, 1 en portugais, 1 en tchèque, 1 en espagnol et 1 en iranien.

de ses œufs et le hareng ne semble pas être considéré comme un « grand poisson »), 16% sont des sites de vente de matériel informatique (un constructeur de disque dur ayant une série nommée « Caviar ») et 16% n'évoquent ni la nourriture, ni les productions animalières.

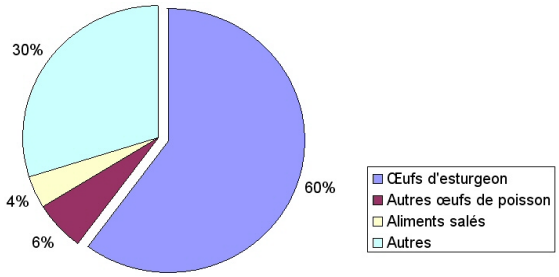

Figure 4 –Requête *caviar* sous Google sans précision de langue (56 premiers résultats).

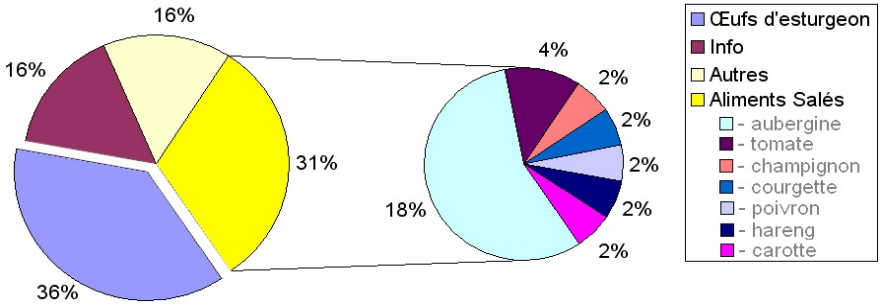

Figure 5 – Requête *caviar* sous Google en français.

Le classement proposé par le moteur Google à la suite d'une requête est principalement fondé sur l'algorithme du *PageRank*. Puisque le classement proposé s'appuie essentiellement sur la popularité des sites, ces chiffres ne sont pas comparables aux résultats obtenus par exemple à partir d'un concordancier, d'une analyse linguistique avancée sur corpus ou d'une expérience de Psycholinguistique. Cependant, ils tendent à montrer que les informations proposées par WordNet et par l'ontologie que nous avons utilisée en exemple (et dans une moindre mesure le TLF[40]) ne suffisent pas pour traiter du terme dans les pratiques langagières courantes : les œufs d'esturgeons côtoient ici la purée de tomate, le disque dur et la coupe Davis de tennis[41]. Dans un bon nombre de pages Internet explorées n'abordant pas la nourriture, il est possible d'inférer le caractère luxueux par l'utilisation du terme

[40] En plus de l'œuf d'esturgeon, le Trésor de la Langue Française Informatisé - http://atilf.inalf.fr/tlfv3.htm - propose comme définition associée à « caviar », : « *N'importe quel mets considéré comme délicat et symbolisant le luxe suprême* ».

[41] Ce zeugme s'actualise par l'analyse des sites en français : http://www.aufeminin.com/__r247_Caviar_de_tomates.html (rang 5 dans la liste de résultats - il s'agit d'une recette à base de tomate), no.kelkoo.com/b/a/ss_Western_digital_caviar.html (rang 8 - où l'on peut acheter un disque dur « caviar ») et sports.fr/fr/cmc/ tennis/200249/cmc_5286.html (rang 23, où l'article intitulé « *Du caviar dans le Saladier* » célèbre la victoire de l'équipe Russe de tennis lors de la Coupe Davis de 2002).

mais ce n'est pas systématique. Dans les pages en français de recettes de cuisine ou de vente de préparations culinaires, *caviar* est principalement employé pour désigner une préparation, le plus souvent salée, qui s'étale sur un toast et peut se servir en hors d'œuvre. Pour les sites en anglais, les seuls aspects *salés* et *granuleux* apparaissent pas de façon concomitante dans environ 34% des pages obtenues. Notons au passage, que les travaux visant à traduire strictement une ontologie d'une langue à l'autre dans le cadre du web sémantique pourront se heurter à des différences flagrantes d'utilisation malgré l'orthographe et certains éléments de signification communs. Pour en finir avec le caviar, notons qu'une étude espagnole sur la catégorisation des produits alimentaires en fonction de la perceptions des consommateurs le place parmi les productions qui ont une « composante sociale » (*social component*) dénotant un certain prestige au même titre que les vins renommés, les liqueurs ou le jambon fumé traditionnel [Colomer et Clotet, 2000]. Cette étude tend à confirmer les résultats obtenus par Rastier pour l'expérience précédemment citée.

Cet exemple est éloquent : il semble matériellement impossible de faire une liste exhaustive des emplois d'un terme donné et d'en déterminer *a priori* les usages et les effets de leur utilisation pour la lecture d'un texte. De même, le sens d'un texte semble dépendre des situations de lecture et la signification d'un mot dépend des causes et des conditions de son énonciation, ces paramètres pouvant varier selon l'imagination de l'homme (dans les limites d'une certaine sociabilité, c.f. les propos de Meillet rapportés p.59), i.e. dans l'absolu, à l'infini dans le temps et dans l'espace. Même dans une approche synchronique, qui supposerait de définir d'emblée des bornes temporelles correspondant à une certaine stabilité des usages, la tâche paraît impossible si elle se veut exhaustive. Dans ces circonstances, comment prendre en considération la relation qui existe entre le sens et la référence puisque cela paraît utile à l'accès au contenu des documents (le cuisinier peut être intéressé par l'aspect salé du caviar tandis que le sociologue pourra être concerné par l'aspect luxueux du produit) ? Comment prendre en considération le fait que l'utilisation de *caviar* puisse faire référence aussi bien aux œufs de poissons qu'aux disques durs ou au luxe en fonction de ce qui est intéressant pour l'utilisateur ? En tentant de répondre à ces questions, nous verrons dans la suite de cette partie, comment nous gagnons au passage un principe de structuration de nos données en nous appuyant sur la *valeur* des signes définie par Saussure.

2.2.1.2 Valeur des signes

La référence n'est pas un simple rapport entre la langue et le monde. C'est l'essence des interactions langagières ; c'est l'opération de construction d'un monde, créée dans et par le discours. Les relations entretenues avec la « réalité » ne sont pas déterministes, prévisibles ou normées. La référence n'est donc pas fonction de choses préexistantes. Un agencement de mots peut avoir trait à des objets existants ou non, à des situations réelles ou imaginaires, à des évènements passés, présents, futurs, fic-

tifs ou avérés, à des idées, à des états mentaux et même à un agencement de mots comme dans les phrases auto-référentielles de Hofstadter ([Hofstadter, 1988 : 6], par exemple *Cette phrase contient cinq mots*.). Il y a donc une distinction entre le domaine des signifiés, qui construisent les contraintes sémantiques de l'interprétation, et le domaine des référents des signes, construits par ceux qui lisent ou écoutent. Ceci nous place dans les distinctions saussuriennes et en particulier dans celle opposant langue et parole.

Selon Saussure, la langue est un système de signes qui permet d'unir un concept et une image acoustique, i.e. unir un signifié et un signifiant. En d'autres termes, il s'agit de faire apparaître ce qui est purement subjectif ou plutôt intersubjectif. Toute langue est facteur d'intersubjectivité ; par exemple, nommer une chose institue cette chose en objet de connaissance et implique un certain découpage de la réalité qui doit être partagé par les utilisateurs de la langue. Dans cette mesure, une langue implique une certaine conception du monde ; l'exemple fameux des 35 termes Inuits pour distinguer l'aspect de la neige est éloquent à cet égard[42]. La langue correspond donc selon Saussure au social (en tant que partagé) et à l'essentiel (en opposition à l'accidentel). En revanche, la parole correspond à l'utilisation d'une langue dans une situation donnée. Elle est signe de liberté, de créativité. Elle correspond aux actes langagiers, aux textes réels qui contrairement au système de la langue, sont directement observables en tant que procès (au sens de Hjelmslev) ou dans leur dimension rhétique (au sens d'Austin). Les concepts langue / parole, signifiant / signifié et système / procès s'opèrent ainsi sous la forme de couples d'oppositions.

La signification d'un signe linguistique n'est pas aussi transparente qu'il pourrait apparaître à première vue. Un signe n'est jamais assimilable parfaitement à un référent, comme nous venons de le voir avec l'exemple du caviar. Dans un texte, un signe linguistique est doté d'une signification particulière en fonction du rôle qu'il joue dans ce système actuel. La signification d'un signe s'échafaude en s'appuyant sur deux axes: l'axe paradigmatique et l'axe syntagmatique. Le paradigme, c'est la classe sémantique à laquelle on rattache un signe donné. Les éléments d'une même classe peuvent se substituer les uns aux autres. On peut associer un signe donné à un nombre indéfini de paradigmes. Aussi, c'est l'enchaînement syntagmatique dans lequel s'inscrit ce signe qui permet la détermination du paradigme actuel. Barthes [Barthes, 1964] écrit : *Il y a donc, devant tout syntagme, un problème analyti-*

[42] Il semblerait d'ailleurs que ce fait, très souvent rapporté dans les cours de linguistique, ne soit finalement erroné et qu'il relèverait plus du mythe urbain que d'une observation scientifique. Geoffrey Pullum en fait la démonstration dans son livre polémique *The Great Eskimo Vocabulary Hoax and Other Irreverent Essays on the Study of Language* (1991, University of Chicago Press, ISBN 0-226-68534-9). Dans les faits, il existe cinq familles de langues eskimo. La plus répandue est l'Inuit (ou *Yup'ik*) dont les différents dialectes sont utilisés du nord de l'Alaska jusqu'au côtes du Groenland. Dans son dictionnaire *Yup'ik Eskimo dictionary* (Alaska Native Language Center, University of Alaska, Fairbanks, USA), Steve A. Jacobson ne relèvent qu'une dizaine de termes (ou racines) pouvant être utilisée pour désigner la neige (neige fine : *kanevvluk*, croûte sur la neige qui est tombée : *qetrar-*, neige fraîche sur le sol : *nutaryuk*, neige tombée flottant sur l'eau : *qanisqineq*, etc.). Ces termes sont recensés par Anthony C. Woodbury à l'adresse suivante : http://www.princeton.edu/~browning/snow.html.

que : le syntagme est à la fois continu (fluent, enchaîné) et cependant il ne peut véhiculer du sens que s'il est articulé. On peut représenter cette articulation entre le paradigmatique et le syntagmatique sous la forme d'axes perpendiculaires :

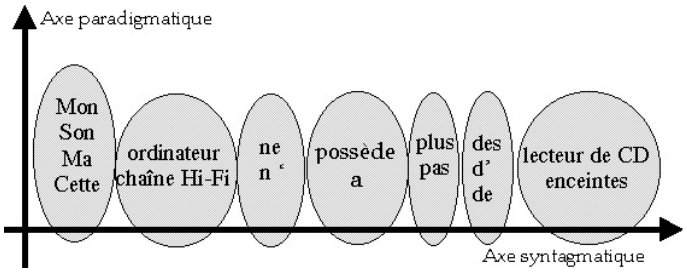

Figure 6 – Articulation paradigmatique / syntagmatique.

L'axe syntagmatique est l'axe de la chaîne linguistique. C'est l'axe du temps dans l'apparition des constituants de la production langagière (énoncé, phrase, texte, récit …). L'axe paradigmatique représente les significations que l'on peut commuter dans une chaîne linguistique en garantissant qu'il y aura toujours un sens accessible, une interprétation possible. C'est l'axe des classes décrivant les différentes terminologies et les domaines thématiques. Les langues dites naturelles, par opposition aux langages formels, ne fonctionnent donc pas sur un mode strictement compositionnel : le sens d'un énoncé n'est pas uniquement construit à partir des significations de ces constituants mais aussi à partir de l'absence de ce qui pourrait être à leur place. Ce sont les artistes qui en ont apporté la plus belle démonstration. Nous pensons en particulier à Perec et à sa Production Automatique de Littérature Française (PALF) [Bénabou, 1989]. La PALF consiste à substituer plusieurs fois des enchaînements syntagmatiques dans un énoncé de départ par d'autres de significations équivalentes (considérés comme tels dans les dictionnaires ou dans d'autres textes). Les effets de sens produits par l'énoncé de départ et d'arrivé sont bien sûr très différents[43]. Cet exemple nous permet de distinguer le sens de la signification. Le sens est une propriété liée aux énoncés, aux textes et la signification (ou l'acceptation) est liée au signe. La signification est alors un artefact. Les effets de sens proviennent de la co-adaptation des significations de plusieurs constituants d'un énoncé, d'un texte, d'un corpus...

Parce que les langues naturelles procèdent par répétitions et différenciations, l'interprétation d'une chaîne linguistique, d'un texte, ne peut pas se réduire à une structure syntaxique ornée par la sémantique ; les significations qui composent la chaîne ne sont pas premières par rapport au sens de la chaîne. Le sens d'un énoncé et les significations des mots de la chaîne sont identifiés en même temps dans l'articulation entre le paradigmatique et le syntagmatique. Ainsi le rapport syntagmatique entre

[43] Dans l'exercice n°2 intitulé « Presbytère et prolétaire » [*ibid.* : 50-61], l'auteur part de l'énoncé « *Le presbytère n'a rien perdu de son charme ni le jardin de son éclat* » (repris par Gaston Leroux dans *Le mystère de la chambre jaune*, la phrase est à l'origine de George Sand) pour aboutir par un jeu de substitutions à un texte d'une quarantaine de ligne dont voici un extrait « *Les pieds bottés excitèrent l'odorat des chiens de garde sur les passantes.*».

les termes concerne le fait de parole (le procès) ; pour Saussure, c'est le rapport *in præsentia*. Le rapport paradigmatique entre les termes concerne la langue en tant que système. Selon Saussure, c'est le rapport *in absentia*[44]. La notion de « valeur » du signe est celle qui permet d'instituer véritablement la langue en tant que système. Pour Saussure, le concept est la contre-partie de l'image auditive dans l'intérieur du signe mais ce signe lui-même, c'est-à-dire le rapport qui relie ces deux éléments, est aussi, et tout autant la contre-partie des autres signes de la langue. La valeur d'un signe résulte donc de sa place dans un réseau de relations binaires. Le signifié (la valeur) d'un signe est donc déterminé par ce qui l'entoure. *Le tout vaut par ses parties, les parties valent aussi en vertu de leur place dans le tout, et voila pourquoi le rapport syntagmatique de la partie au tout est aussi important que celui des parties entre elles* [Saussure, 1916* : 177]. Le signifié de *caviar* n'est pas défini par les propriétés ontologiques du caviar en général mais représente tout ce que « caviar » n'est pas. Le signifié, contrairement à la notion de concept ontologique, n'entretient aucun lien avec une classification des espèces animales ou des productions animalières. Il est défini de façon purement différentielle. Tous les signes sont solidaires et la valeur de chaque signe est un point de contact avec l'ensemble du système de la langue organisé en réseau d'oppositions.

Le système de valeurs n'est pas figé, il se modifie dans le temps. Selon Saussure, cette évolution est l'effet d'une *force sociale*, la langue *partie sociale du langage (…) n'existe qu'en vertu d'une sorte de contrat passé entre les membres de la communauté* [*ibid.*][45] et ce « contrat social » se construit entre autres par la parole[46]. Ce sont les relations qu'entretiennent les éléments d'un paradigme qui organisent les significations en un système de valeurs. Les paradigmes peuvent être liés entre eux puisque les significations différentes d'un même signifiant peuvent se placer dans des paradigmes différents. Ce que nous rechercherons donc à faire du point de vue de la structuration des signifiants, c'est de les organiser selon leurs différences pour rendre compte de leur valeur. Il ne s'agira pas pour autant de modéliser le système de la langue proposé par Saussure : nos aspirations sont bien plus modestes. Nous nous inspirons du système de valeurs pour organiser les signifiés jugés intéressants pour l'utilisateur dans un cadre donné. Nous utiliserons donc un principe de langue pour structurer des éléments repérés et utilisables pour l'analyse de la parole (ce qui est légitime si l'on conçoit que l'on ne

[44] Saussure pose que le médium véritable de la communication n'est alors pas la parole comme donnée immédiate considérée dans ses effets et sa matérialité observable mais la langue comme système de relations objectives. C'est ce système qui rend possible la production du discours et son interprétation. Mais la langue ne peut être appréhendée en dehors de la parole. On apprend une langue par la parole et la parole permet de transformer une langue. Cependant ces deux processus ne sont pas chronologiquement hiérarchisés puisque l'un aura le primat sur l'autre selon que l'on se situe sur le terrain de l'histoire individuelle et collective ou sur les conditions de l'interprétation.

[45] Voir l'analyse de Robert Marty - http://www.univ-perp.fr/see/rch/lts/marty/s018.htm

[46] Nous ne faisons aucunement référence ici à Jean-Jacques Rousseau, il s'agit du contrat social saussurien.

peut représenter en langue que ce qui a été décrit en contexte). Il s'agit là d'un premier choix théorique quant à la modélisation de la signification des mots qui seront utilisés par notre système. Nous n'en avons pas pour autant terminé avec la référence.

2.2.1.3 Sens et référence

La pragmatique a entre autre pour objet la relation observable entre les mots et les choses et pour revenir à la distinction saussurienne langue/parole, on considère souvent que la pragmatique a pour objet l'étude de l'usage de la langue (la parole) par opposition à l'étude du système de la langue qui concerne la linguistique. On peut relativiser cette distinction et considérer la linguistique et la pragmatique comme deux plans d'analyse orthogonaux (comme dans [Bricon-Souf, 1994]). La référence est alors conçue comme l'activation du potentiel d'un terme, d'un énoncé, ou d'un texte à être porteur d'indications pour la mise en contexte de son sens. Beust, [Beust, 1998] p.40, cite par exemple, un extrait de dialogue où l'enchaînement conversationnel permet tour à tour de concevoir l'expression *le bureau de l'association* comme ayant trait au lieu et au groupe de personnes par des actualisations différentes de son potentiel de sens. Il montre ainsi que la référence ne provient pas seulement de l'interprétation mais de l'actualisation des contraintes dans le contexte[47].

Rappelons les deux questions que nous nous posions à propos du caviar : comment prendre en considération la relation qui existe entre le sens et la référence puisque cela est utile à l'accès au contenu des documents ? Comment prendre en considération le fait que l'utilisation de *caviar* puisse faire référence aussi bien aux œufs de poissons qu'aux disques durs ou au luxe ? Dans notre système, nous partons des signifiants présents dans les textes. Nous savons désormais qu'ils seront organisés de façon différentielle en fonction de leurs signifiés en référence à la valeur saussurienne. Nous nous intéressons aux textes, aux interactions langagières, à la parole. Les signifiés de signifiants observés dans un texte ou un ensemble de textes, seront organisés par l'utilisateur en terme de ce qui est utile à sa tâche. Cela pourra donc se faire en fonction de leur potentiel à faire référence à des objets du monde (« aux œufs de poisson » si c'est utile) aussi bien qu'à des aspects plus personnels à l'utilisateur ou culturels (« au luxe » si la tâche le nécessite). D'une manière générale, ils seront organisés en fonction de leurs différences avec les autres signifiants utiles à l'utilisateur. Les points communs entre ces signifiés permettront de créer des classes paradigmatiques dont l'expertise de la validité d'un texte à un autre dans un contexte identique (celui d'une tâche documentaire définie) sera l'objet central du travail d'assistance de la machine. Nous nous appuierons pour cela sur l'étude de textes dans leur ensemble et non sur de simples phrases ou énoncés.

[47] Les maximes de Grice expliquent pourquoi, en général, les contraintes issues de l'interprétation du matériau linguistique donnent lieu à des références non ambiguës [Grice, 1975].

Il nous reste à régler quelques questions importantes, en particulier celle qui nous apportera une solution valable pour décrire différentiellement les signifiés. C'est ce que nous allons aborder dans les parties suivantes.

2.2.2 Approche interprétative, Sémantique Interprétative

Une approche interprétative consiste à se focaliser sur l'interprétant (humain) dans son activité d'interprétation de textes. Le sens est considéré comme étant le fait de l'interprétant, comme un processus qui se fonde à la fois sur l'observation du matériau textuel dans un contexte donné et sur les connaissances du lecteur en tant que sujet interprétant. Pour nous, le contexte sera celui d'une tâche documentaire définie dans laquelle la variabilité contextuelle des significations est limitée. Elle est limitée dans le sens où les cotextes et les intentions du lecteur partagent des caractères redondants d'un texte à l'autre, d'une activité particulière à l'autre dans le cadre d'une tâche documentaire générique.

La Sémantique Interprétative de F. Rastier donne un cadre scientifique à l'étude de l'interprétation des textes : ce cadre établit un ensemble de considérations influençant l'interprétation que l'on peut en faire. Ces considérations relèvent de la situation sociale et historique et de l'intention de l'interprétant dans le cadre de sa lecture ; elles sont projetées sur les éléments qui le constituent pour y trouver une justification et peuvent être modifiées ou servir de base à la découverte d'autres aspects. Au cœur de la théorie se trouve donc la notion d'interprétation définie comme *l'assignation d'une signification à une séquence linguistique* [Rastier, 1987*]. Selon Rastier, l'interprétation est une opération guidée par une stratégie qui dépend d'indices intrinsèques et extrinsèques au texte. Les indices intrinsèques relèvent du principe herméneutique de détermination du local par le global : c'est le texte qui détermine l'interprétation que l'on peut en faire. Les indices extrinsèques relèvent à la fois des connaissances, de la situation et de la pratique de lecture de l'interprétant. Ils concernent également les informations connues de l'interprétant sur les conditions de production du texte et la place qu'il peut lui octroyer dans un intertexte. F. Rastier refuse l'immanentisme du sens tout autant que l'impossibilité d'accéder à celui-ci : il considère l'interprétation comme une construction et non comme une redécouverte[48]. Cette construction est un processus qui met en jeu certains facteurs qui vont bien au-delà du domaine purement linguistique. Ces facteurs sont appelés par Rastier des normes

[48] Certaines prises de position de la Sémantique Structurale peuvent paraître radicales et certains n'ont pas manqué de tenir à leur sujet des propos alarmants. C'est le cas par exemple pour le psychologue Hillman [Hillman, 1989 : 28] (cité dans [Dufresne, 1992]) qui accuse cette linguistique d'être la cause d'une *« crise de confiance, puisque nous n'osons plus nous abandonner aux mots en tant que porteurs de sens ! Le langage est frappé d'une véritable phobie du sens.* ». Il s'agit d'une véritable caricature. Dans sa définition de la langue, Meillet disait au contraire, dès 1948 : « *Une langue est un système rigoureusement lié de moyens d'expressions communs à un ensemble de sujets parlants : il n'a pas d'existence hors des individus qui parlent (ou qui écrivent) la langue ; néanmoins, il a une existence indépendante de chacun d'eux ; car il s'impose à eux ; sa réalité est celle d'une institution sociale, immanente aux individus, mais en même temps indépendante de chacun d'eux, (…)* ».

ou des degrés de systématicité. Il s'agit de l'idiolecte, du sociolecte et du dialecte qui correspondent respectivement à l'*usage d'une langue et d'autres normes sociales propres à un énonciateur*, à l'*usage d'une langue fonctionnelle, propre à une pratique sociale déterminée* et à la *langue fonctionnelle — ou langue considérée en synchronie, par opposition à la langue historique* [Rastier, 2001a]. L'idiolecte est donc lié au locuteur en tant qu'individu avec ses singularités langagières personnelles. Le sociolecte découle de considérations culturelles voire communautaires et donc sociales. Le dialecte quant à lui, pose les normes de la langue. Les systèmes classiques du TAL destinés à l'analyse du contenu de documents ne peuvent intégrer toutes ces normes si l'utilisateur n'y occupe pas une place centrale. Lorsqu'on utilise exclusivement des dictionnaires, des ontologies ou encore des thésaurus, seuls le dialecte et le sociolecte sont concernés : le dialecte est en effet l'instrument qui permet de s'adapter à la langue des documents manipulés ; quant au sociolecte son usage permet de mettre en place des systèmes spécialisés touchant à des domaines restreints. Pour ce qui concerne l'idiolecte, tout traitement devient impossible sans l'intervention de l'utilisateur car il s'agit de prendre en considération ses particularités personnelles. On se prive ainsi d'une possibilité d'assistance personnalisée, d'une manipulation des textes électroniques qui permettrait à l'utilisateur d'adapter les processus informatiques mis en jeu à sa tâche et sa pratique. Nous proposons à l'utilisateur la prise en considération non seulement du dialecte de son intérêt, mais également des dimensions sociolectales et idiolectales en rapport avec sa tâche dont nous obtiendrons des traces à travers les ressources qu'il aura lui-même construites. À travers ces constructions, nous proposons l'expression d'un point de vue particulier sur le contenu de textes médiatisés par des logiciels. Les exemples exposés dans les prochains chapitres permettront d'illustrer les moyens mis en œuvre pour mettre en lumière ces aspects.

Dans la lignée des Bréal et Saussure, Greimas, Coseriu et Pottier, la Sémantique Interprétative s'inscrit dans la continuité de la Sémantique Structurale européenne. La SI propose l'utilisation des sèmes (ou traits sémantiques élémentaires) pour la description des signifiés linguistiques (appelés sémèmes)[49]. Les sémèmes sont attribués à des signifiants comme les morphèmes ou les lexies. Au niveau local, un sémème correspondra à une collection de sèmes de statuts différents. Au niveau global, ces sèmes sont rassemblés en classes de sémèmes ou à des formes de structuration de ces classes. Au cœur de la théorie, se trouve le principe de justification de l'attribution d'un sème à un sémème. Cette justification se fait au niveau global par la mise en relation d'un sémème à décrire avec un ou plusieurs autres sémèmes. Les normes évoquées plus haut constituent des niveaux de validité et de pertinence des caractérisations sémantiques d'une entité lexicale. Ainsi, les sèmes relevant du système fonction-

[49] Les sèmes de la Sémantique Interprétative se distinguent néanmoins de ceux de ses aînés : ils ne sont plus des qualités d'un référent ou des parties d'un concept, ils n'ont plus de caractère universel et leur nombre cesse d'être indéfini. Cependant, comme pour Pottier, le sème est ici « *un trait distinctif sémantique d'un sémème, relativement à un petit ensemble de termes réellement disponibles et vraisemblablement utilisables, chez le locuteur dans une circonstance donnée de communication* » [Pottier, 1980]. Les aspects contextuels et pragmatiques priment désormais sur les considérations sémiques.

nel de la langue sont définis comme *inhérents*. Ceux des autres normes sont *afférents*, donc plus liés au contexte que les inhérents. Les sèmes afférents sont non-définitoires et relèvent donc des normes sociolectales et idiolectales. Ils sont hors du système fonctionnel de langue au contraire des sèmes *inhérents* qui doivent être « reconnus » (et non « construits ») au cours du processus interprétatif puisqu'ils sont donnés *a priori*. Cette première nuance entre sèmes ne se situe finalement qu'*en langue*. En contexte, les deux types de sèmes peuvent apparaître ou disparaître de la même manière. Par exemple, le sème /force/ pour le sémème 'homme' peut être actualisé par la perception d'instructions contextuelles, c'est un sème afférent. Le sème /noir/ pour 'corbeau' est un sème inhérent : c'est un sème dont l'occurrence de 'corbeau' hérite du type (le sème en langue), par défaut. Les signifiés sont organisés selon plusieurs formes. On constitue une classe sémantique à partir d'un sème commun entre tous ces éléments. Un tel sème est appelé *générique* : il organise les sémèmes au sein d'une même classe. Certains sèmes pourront au contraire distinguer les sémèmes d'une même classe, ils seront appelés *spécifiques*.

Dans les textes, les sèmes supportés par les sémèmes forment des récurrences le long de l'axe syntagmatique. On appelle ces récurrences des *isotopies*. Divers statuts attribués aux sèmes peuvent se combiner au sein d'une même isotopie, et donc un même sème peut jouer divers rôles organisateurs en langue. L'identification de la classe d'un sème et de son statut sont propres à une interprétation. L'isotopie permet de repérer ce que l'on peut appeler un thème dans une phrase ou un texte. Elle peut également servir d'indice au repérage des sèmes et être utiles en cela (comme c'est le cas avec PASTEL) pour capter dans un cadre rationnel les « présomptions d'isotopie ». Dans [Tanguy et Thlivitis, 1996], il est ainsi proposé l'exemple suivant : si l'on aborde une recette de cuisine, nous nous attendons à y trouver des expressions ayant trait à l'alimentation, donc à repérer une isotopie du sème /alimentation/. En abordant une poésie de la période romantique, on s'attend à trouver les thèmes classiques comme l'amour ou la nature mais ces thèmes peuvent être difficilement repérables dans des termes pris isolément. PASTEL guide le lecteur de la présomption d'isotopie à l'isotopie proprement dite et l'aide à la justifier et rationaliser son interprétation. On voit ici que les préoccupations de la SI et du logiciel PASTEL ne sont pas identiques aux nôtres puisque nous tentons non pas de rationaliser une interprétation mais de l'exploiter pour rendre des services documentaires à l'utilisateur. Les présomptions d'isotopie pourront cependant nous servir pour l'aide à l'analyse du contenu de documents en terme d'appartenance à un domaine précis. Dans ce cas, elles correspondront plus à des « isotopies réclamées » dans le sens où, si l'utilisateur est intéressé par un thème particulier, il sera demandeur de textes présentant (au moins) une isotopie en rapport avec ce thème.

La SI propose les opérations interprétatives suivantes : l'actualisation, la virtualisation ainsi que l'afférence et la dissimilation. L'actualisation consiste à identifier un sème dans un contexte. Par exemple, dans l'énoncé *le volailler m'a vendu un canard*, le sème /viande/ est actualisé dans le sé-

mème ‘canard’ parce-qu’il se répète dans ‘volailler’. L’actualisation permet entre autres de lever les ambiguïtés lexicales et de ne pas prendre en considération dans l’exemple (et sans autre forme de contexte) le canard qui aurait trait à la presse. La virtualisation consiste au contraire à neutraliser un sème en contexte. Par exemple, dans …*le canard qui aurait trait à la presse*… le sème /volaille/ est virtualisé car non seulement il n’est pas répété dans l’énoncé mais, de plus, il ne trouve pas d’autres justification dans le présent cotexte. Comme nous venons de le voir, l’actualisation et la virtualisation concernent les sèmes inhérents ; ces deux opérations jouent un rôle important dans la mise en cotexte du contenu sémique dans l’ordre paradigmatique. Certains sèmes ne sont actualisables qu’en fonction de certaines instructions contextuelles, ces sèmes sont dits afférents. L’actualisation d’un sème afférent est appelé afférence. Par exemple, /non alcoolisé/ est un sème afférent pour « apéritif » dans *Mister Mint, apéritif sans alcool*[50]. Enfin, la dissimilation consiste à actualiser des sèmes afférents opposés dans deux occurrences du même sémème (ou dans deux sémèmes parasynonymes). L’exemple des énoncés interprétés comme des tautologies par les systèmes logiques est le plus explicite à cet égard. Par exemple, *Il y a chasseur et chasseur* amène à produire une dissimilation des sèmes /agile/ et /maladroit/ ou encore /respectueux/ et /cruel/ pour les sémèmes de l’une ou l’autre occurrence de chasseur. Les opérations interprétatives précisent donc la dynamique des sèmes en contexte.

Pour la Sémantique Interprétative, la langue propose et le texte dispose. Nous entendons par-là que les sèmes « en langue » (les types), doivent être actualisés, transformés en occurrences au cours de l’interprétation – en cela, on peut dire que le texte propose aussi. Pour actualiser un sème, il faut produire ou trouver un contexte qui permette son activation par défaut ou par propagation du fait des relations de dépendance sémantique entre les composants syntagmatiques. Certaines de ces relations de dépendances sémantiques tendent à renforcer les propriétés communes des composants et à effacer celles qui ne donnent pas lieu à une répétition dans le co-texte : c’est cela que nous appelons la dynamique des sèmes en contexte. Les opérations interprétatives permettent de rendre compte de façon rationnelle d’une interprétation. Cependant, elles ne sont pas entièrement prédictibles et la plupart de ces opérations ne peut donc relever que de l’humain. Nous remarquerons la possible mise au jour ou formalisation de ces opérations relève du spécialiste. Étant donnés nos objectifs, il nous faudra ne les réserver qu’à ce type d’utilisateur et proposer aux autres des moyens performants plus axés sur leurs propres préoccupations.

Dans cette partie, nous avons vu que la SI représente une base théorique importante pour expliquer les principes d’une interprétation en langue naturelle. Cependant, dans le cadre d’un travail qui a pour but de fournir des outils informatiques utiles à des tâches documentaires dépassant l’objectivation ou l’explication d’une interprétation, elles apparaissent complexes à exploiter telles quelles. Dans notre système, nous garderons le principe componentiel qui permet l’utilisation des sè-

[50] http://www.aperosansalcool.com/M_MINT.html

mes pour décrire des signifiés en contexte. Nous exploiterons également l'organisation des signifiés en classes et le principe d'isotopie pour rendre compte de certains aspects de l'interprétation et pouvoir exploiter ces aspects pour l'analyse automatique de textes inconnus de l'utilisateur. Cependant, les exigences d'une implémentation informatique nous amèneront à utiliser une autre définition du sème, plus effective dans une modélisation computationnelle. Nous verrons par ailleurs tout au long de ce travail que nous nous éloignons des notions de la SI en les transposant dans nos applications informatiques.

Dans les parties suivantes, nous présentons le modèle de catégorisation ANADIA, que nous exploitons pour organiser différentiellement et de façon componentielle les signifiés utiles à l'utilisateur dans le cadre d'une tâche. Certains psychologues envisagent la catégorisation comme une activité primitive des êtres vivants [Dubois, 1991]. Parmi tout ce qui peut être observé du monde qui nous entoure et de notre activité, seulement ce qui *fait différence* pour les résultats de l'activité de l'individu est construit et conservé en mémoire. L'activité humaine de catégorisation a une dimension sociale : les choses inconnues abordées par un interlocuteur dans une conversation ou par un auteur dans un texte, sont interprétées par différenciation avec les choses connues et/ou admises socialement. Selon ces principes, nous proposons d'organiser les entités lexicales du domaine d'intérêt d'une tâche documentaire comme une représentation praxéologique, en vue d'une pratique ; il s'agit d'un modèle de terrain commun des interactions entre l'utilisateur et les logiciels fondé sur la différence. Le modèle de catégorisation ANADIA nous permet ainsi de rejoindre les propositions de Saussure ; l'utilisation d'une description componentielle des signifiés des entités lexicales nous permet d'exploiter certains principes de la SI.

2.2.3 Modèle de catégorisation différentielle et modèle oppositionnel du sème

Notre modèle se nomme LUCIA en référence à ANADIA. ANADIA est une méthode de catégorisation initialement proposée par Coursil [Coursil *et al.*, 2000] pour rendre compte de la dimension défective de la langue, le fait que la langue se constitue essentiellement par des différences suivant en cela les propositions de Saussure. ANADIA met en place des grilles de catégorisation qui permettent de montrer les différences pertinentes entre des entités. En fonction des remarques que nous avons proposées précédemment, le modèle ANADIA nous est apparu pleinement adapté à la modélisation des signifiés telle que nous la proposons. Nous aurons l'occasion dans le chapitre suivant de revenir sur les détails de ce modèle puisqu'il est à la base de LUCIA.

ANADIA a déjà été utilisé dans le cadre du TAL et cela entre autres par Pierre Beust pour l'élaboration d'un modèle interactionniste du sens [Beust, 1998*] fondé en partie sur la Sémantique Interprétative. Pour parvenir à ses fins, Pierre Beust a imaginé un modèle oppositionnel du sème : ce-

lui-ci est rapporté à un jeu d'oppositions de valeurs au sein d'un domaine d'interprétation. Rastier [Rastier, 2001b] propose la définition du sème suivante : un sème est un *élément d'un sémème, défini comme l'extrémité d'une relation fonctionnelle binaire entre sémèmes*. Par sa définition oppositionnelle du sème, Beust ne considère plus le sème comme l'extrémité de cette relation mais cette relation entre signifiés. Selon lui, une approche computationnelle (et donc oppositionnelle) du sème, amène à refuser au sème le simple rôle d'une entité autosuffisante issue de l'interprétation d'un texte ; il le considère au contraire comme une entité relationnelle permettant à un agent logiciel[51] de fonder une combinatoire de représentations. Ce qui est alors pertinent, ce n'est plus seulement le trait sémantique en lui-même mais également ce à quoi on peut l'opposer. Le sème oppositionnel est donc rapporté à un jeu d'opposition de signes structurant au niveau même des éléments de base de la signification. Ces propositions sont issues entre autres de l'étude d'un corpus oral de conversations dans lequel Beust a pu apprécier que l'évocation d'une opposition dans un contexte ne se limite pas à la négociation de termes opposés. Les oppositions significatives sont rattachées à un domaine qui représente le cadre de son interprétation, qui n'est autre que son domaine d'interprétation tout court. Pour ne prendre que ce seul exemple, on peut être amené dans le cadre d'une description linguistique du contenu en sémantique componentielle classique à affecter à 'livre' et à 'billet' le sème /papier/ [Nicolle *et al.*, 2002*]. Cependant, on peut s'interroger sur l'ubiquité de ce sème : est-ce véritablement le même dans les deux cas ? En ce qui concerne le 'livre' le trait /papier/ peut être pertinent dans les oppositions /papier/ *versus* /logiciel/ ou encore /papier/ *versus* /oral/. Si ces oppositions concerneraient le domaine d'interprétation des supports de l'information, nous utiliserions alors le sème oppositionnel suivant :

Domaine d'interprétation :	*supports de l'information*
Oppositions :	*papier* vs. *logiciel*
	papier vs. *oral*

Avec *billet* l'opposition pertinente concernerait les types de monnaie : papier *versus* métallique *versus* électronique[52]. Nous utiliserions donc pour 'billet' le sème :

Domaine d'interprétation :	*type de monnaie*
Opposition :	*papier* vs. *métallique* vs. *électronique*

Les sèmes différentiels sont justifiés par leur potentiel de différenciation et les sémèmes sont dans ce cadre, inter-définis de façon relationnelle par leur contenu sémique. Le modèle de Beust est rendu opérationnel en machine par l'utilisation du modèle de catégorisation ANADIA dont les attributs de catégorisation sont des sèmes oppositionnels. Nous reprenons également à notre compte le modèle oppositionnel de Beust mais nos objectifs n'étant pas là non plus identiques, nous serons amené à apporter quelques changements du point de vue de leur utilisation et de la formalisation. En parti-

[51] Les travaux de Beust ont pour vocation de permettre aux agents logiciels, i.e. des programmes informatiques communicants, d'interagir plus efficacement entre eux ou avec des agents humains.

[52] Ces oppositions pourraient dans une tâche spécialisée dans le domaine monétaire avoir la forme : scripturale *vs.* métallique *vs.* fiduciaire *vs.* électronique.

culier, nous nous distinguerons par une approche non aristotélicienne (dans l'esprit de l'*Organon*) des catégories du modèle. En effet, Beust se place dans l'optique de l'élaboration d'un modèle de langue alors que nous nous restreignons à un modèle de représentation lexicale et d'analyses du contenu de documents. Dans les faits, nous verrons alors que les différences entre ANADIA et LUCIA relèvent plus de la nature et du statut des concepts utilisés que des représentations manipulées par les deux modèles.

À ce stade de la présentation des fondements de notre approche, nous avons pris position pour :

- une représentation différentielle des signifiés utiles à la tâche d'un utilisateur ;
- une appropriation de principes de la SI pour effectuer les analyses à partir de ces signifiés (le principe d'isotopie) et les organiser (les classes sémantiques) ;
- une utilisation d'un modèle de catégorisation différentielle et d'un modèle oppositionnel du sème pour une mise en place computationnelle de ces principes.

Avant de conclure ce chapitre, il nous reste à préciser les places allouées à la machine et à l'utilisateur dans un telle cadre.

2.2.4 Interaction dans le système

Pour préciser les rôles de la machine et de l'utilisateur dans notre système, nous nous appuierons sur trois types d'approches : celles technocentrées, celles anthropocentrées et enfin, celles centrées sur l'utilisateur.

Dans les systèmes documentaires de conception logique, les processus automatiques sont centraux. Si les résultats sont limités (ils sont fonction des constats présentés dans les premières parties de ce chapitre), l'effort demandé à l'utilisateur se situe principalement au niveau de l'interprétation des résultats fournis par la machine[53]. L'usager peut ainsi obtenir satisfaction en utilisant ces systèmes mais il ne peut les adapter précisément ni à ses besoins, ni à ses compétences dans le cadre d'une situation donnée. L'approche est alors dite *technocentrée* : c'est à l'utilisateur de s'adapter au système. Il n'y a aucune place pour la subjectivité, pour un autre point de vue que celui qui a été prévu. Les incidences contextuelles et l'indéterminisme de l'interprétation ne sont pas pris en considération. Dans la SI, l'interprétation est conçue comme un parcours, un processus sur lequel influent quatre facteurs : (i) l'interprète situé, (ii) la pratique sociale et donc (iii) l'action et (iv) la temporalité. Cette conception permet donc de s'intéresser à l'interprétation comme activité humaine et hors des capacités de la ma-

[53] Ce qui n'est pas toujours évident : nous reviendrons ultérieurement sur ce point, en particulier dans le chapitre 5 (partie 5.2 p.190).

chine[54]. Envisagée en informatique, la SI est souvent plus ou moins déformée à l'implantation – ceci est reconnu par les informaticiens travaillant sur la question puisqu'il est reconnu que les objectifs de la théorie linguistique ne sont pas les mêmes que ceux d'applications informatiques. Les applications informatiques sont essentiellement destinées à l'assistance à l'analyse de textes littéraires ou savants. Dans ces systèmes destinés aux linguistes ou aux chercheurs, l'utilisateur est central. Les services sont limités parce qu'entre autres les ressources proviennent presque exclusivement de l'usager et les services rendus ne sont pas investis pour d'autres types de tâches. En retour des efforts demandés, les logiciels peuvent s'adapter aux besoins de l'utilisateur. L'interprétation des résultats fournis est un effort moins important pour l'usager parce qu'en principe celui-ci est impliqué dans toutes les étapes des processus et n'a pas besoin de faire des prospectives hasardeuses sur le fonctionnement des programmes pour en faire bon usage. L'approche est alors dite *anthropocentrée* : la machine s'adapte à l'utilisateur. D'après [Thlivitis, 1998*] ; le rôle de l'utilisateur est alors d'*expliciter* (une interprétation) et de *décider* de la validité des propositions de la machine dont le rôle se limite à *contrôler*, *comparer*, *organiser* et *conserver* les propositions de l'usager, pour lui *suggérer* en retour des aides qui lui seront utiles. La distinction entre ces deux conceptions ne correspond pas nécessairement à la distinction entre les approches logiques (ou logico-grammaticales) et herméneutiques : certaines applications de l'approche logique tentent de s'intéresser à l'utilisateur en le modélisant ou en effectuant des adaptations au niveau des ressources [Razmerita, 2003], [Ranwez, 2000], [Brewster *et al.*, 2002], [Charlet *et al.*, 2003* : 79-91] et certaines applications de l'approche rhétorique/herméneutique ne prennent pas directement en considération l'activité humaine d'interprétation en s'abstenant de faire intervenir explicitement les utilisateurs humains [Beust, 1998*] et [Valette, 2003].

Dans nos travaux, nous envisageons, nous aussi, l'interprétation comme une activité humaine que nous nous proposons d'assister de façon logicielle. La machine ne peut, en définitive, que calculer (rapidement), stocker (beaucoup) et afficher (bien). Les capacités de stockage sont exploitées ici pour conserver aussi bien les ressources fournies par l'utilisateur, que celles permettant de les créer (dans les faits, des corpus d'observation). Le calcul est utilisé pour assister l'utilisateur dans la création des ressources en automatisant toutes les phases qui relèvent de manipulations symboliques accessibles aux logiciels. Il permet également, à partir du repérage des différences et des points communs entres les symboles proposés par l'utilisateur (les entités lexicales décrites réifiées en machine) et ceux trouvés dans des textes, de proposer des résultats chiffrés (comptages, statistiques…), parfois retranscrits sous la forme de rapports d'analyse textuels, de graphes, de coloriages, d'interfaces de lecture… Les résultats chiffrés seront exploités pour d'autres calculs (classement, filtrage…) pour mieux assister l'utilisateur dans la tâche pour laquelle il aura choisi le système (veille documentaire, analyse de fait

[54] Les travaux issus de la SI considèrent l'interprétation comme hors des capacités de la machine, puisqu'elle « dépendrait d'un ensemble de facteurs subjectifs (et intersubjectifs), et de nombreuses incidences contextuelles, qui ne sauraient se décider à l'avance. Ni suivre toujours des chemins déterministes. » [Kanellos *et al.*, 1999].

de langue…). Du point de vue de notre approche, nous nous situons donc dans un principe anthropocentré puisque les rôles de la machine et de l'utilisateur seront ceux précisés par Thlivitis. Cependant, l'approche interactionniste nous amène à ne pas considérer uniquement les principes théoriques de mise en place informatique de notre système et à apporter le plus grand intérêt à la réalisation d'interfaces, de vues et de lieux d'interactions les plus adaptées possibles à l'utilisateur et à sa tâche. Dans ce cadre, notre approche est également *centrée utilisateur* puisque les processus de conception sont principalement guidés par la prise en considération des besoins et des objectifs potentiels des utilisateurs. L'utilisabilité aussi bien par des novices que des spécialistes est une priorité de notre démarche et la présentation des moyens alloués aux utilisateurs prend donc en compte cet aspect tout au long de notre présentation.

2.3 Conclusion

Nous avons montré dans la première partie de ce chapitre en quoi les approches classiques pour l'accès aux documents et à leur contenu ne pouvaient satisfaire pleinement les besoins de personnalisation de ces services. Nous avons pu voir que certaines approches inspirées de la SI de Rastier répondaient déjà à certains de ces besoins. Les travaux qui s'en réclament permettent de rendre compte de l'activité d'interprétation sans pour autant réinvestir les résultats d'*une* interprétation, par *un* interprète, dans *une* situation donnée pour d'autres tâches courantes accessibles à des non-spécialistes. Ces constats nous ont amené dans une seconde partie à établir les fondements théoriques d'une approche interactionniste des tâches documentaires. Nous avons ainsi choisi d'utiliser de la valeur saussurienne pour organiser les ressources de notre système qui seront construites par l'utilisateur. De la SI, nous avons gardé l'approche componentielle et différentielle et les analyses interprétatives proposées pour rendre compte d'une interprétation. Enfin, nous avons vu que le modèle ANADIA et le modèle oppositionnel de Pierre Beust apportait un cadre computationnel attesté à ces principes.

Nos travaux traitent de la parole, de l'utilisation de la langue dans les actes langagiers, de l'utilisation de la langue en vue de fins pratiques. Nous ne considérons pas le langage dans son autonomie et son autosuffisance, c'est-à-dire comme finalité sans fin, sans autre fin, en tous cas, que d'être interprété à la façon de l'œuvre d'art. Nous nous attachons à concevoir « le comprendre pour agir » et dans le domaine informatique qui est le nôtre, à concevoir des outils permettant aux utilisateurs d'agir et d'être assistés en fonction de leur interprétation et leur compréhension du matériau textuel. Notre hypothèse principale est que certaines prises de positions pragmatiques d'un sujet interprétant en fonction des contraintes prescrites par le matériau linguistique *et* de la situation de la tâche en cours, peuvent être modélisées en interaction avec les logiciels et être exploitées pour l'assistance à des tâches documentaires. Les fondements théoriques de notre travail étant posés, il nous faut maintenant rentrer dans les détails de nos propositions. Le chapitre suivant est consacré au modèle LUCIA en tant que

modèle lexical de catégorisation et de représentation. Nous y montrons qu'il permet d'organiser des signifiés selon un point de vue particulier de l'interprétant utilisateur du système.