Sémantique et corpus spécialisés : Constitution de Bases de Connaissances Terminologiques

Anne Condamines

Sémantique et corpus spécialisés : Constitution de bases de connaissances terminologiques

Anne Condamines

Mémoire présenté en vue de l'obtention de l'Habilitation à Diriger les Recherches en Linguistique

Université Toulouse Le Mirail

Jury :
Andrée Borillo, professeur, Univ. Toulouse Le Mirail
Josiane Boutet, professeur, IUFM, Paris
Benoît Habert, professeur, Univ. Paris X, rapporteur
Jennifer Pearson, responsable de la traduction à l'Unesco, Paris
Marie-Paule Péry-Woodley, professeur, Univ. Toulouse Le Mirail, rapporteur
François Rastier, directeur de recherches, CNRS, Modyco, rapporteur

26 Juin 2003

Je vois de plus en plus à la fois l'immensité du travail qu'il faudrait pour montrer au linguiste ce qu'il fait … et en même temps l'assez grande vanité de tout ce qu'on peut faire finalement en linguistique.

Saussure, Lettre à Meillet, 4 janvier 1894, cité par (Benveniste, 1966, 37)

REMERCIEMENTS

Benoît Habert et François Rastier m'ont honorée en acceptant d'être rapporteurs de mon dossier d'Habilitation à diriger les recherches. Je les remercie chaleureusement.
Marie-Paule Péry-Woodley a joué parfaitement son rôle de « marraine », je lui en suis très reconnaissante ; bien avant cette habilitation, nos discussions scientifiques (et les autres aussi) avaient accompagné mon parcours de recherche.
Andrée Borillo est certainement celle qui est la mieux à même de mesurer l'évolution de ma réflexion; j'espère qu'elle n'est pas déçue d'une pensée qu'elle a certainement contribuée à former.
Josiane Boutet et Jennifer Pearson ont accepté de faire partie du jury ; j'apprécie l'intérêt que, ce faisant, elles témoignent à mon travail.

Même si elle s'élabore souvent dans la solitude, une réflexion n'est jamais le fruit d'une pensée isolée ; elle se nourrit de rencontres intellectuelles, de collaborations, de confrontations. De ce point de vue-là, je dois beaucoup à tous ceux qui ont accepté de faire un peu de route avec moi.

Je remercie mes amis de l'opération « Sémantique et corpus » de l'ERSS (en poste ou encore étudiants) qui m'ont permis d'assurer ma réflexion en l'ancrant dans une perspective fondamentalement linguistique. Nos discussions sur le rôle des corpus dans l'étude du sens, sur les possibilités d'utiliser des outils et sur des questions plus futiles aussi m'ont souvent donné du courage : A. Borillo, D. Bourigault, C. Fabre, N. Hathout, L. Tanguy, M. -P. Péry-Woodley, C. Pernet, C. Frérot, M. Hodac, M. -P. Jacques, M. Lecolle, H. Miguet, S. Osdowska, J. Rebeyrolle, P. Vergely. J'adresse un remerciement tout particulier à Josette Rebeyrolle qui m'accompagne dans mes « aventures » avec des entreprises depuis 1992 et qui a suivi de près l'élaboration de ce mémoire.

Je remercie mes collègues et amis du groupe TIA (Terminologie et Intelligence Artificielle) ; la pluridisciplinarité accompagnée de l'esprit d'ouverture qui règne dans ce groupe m'ont amenée à un point de vue distancé sur la linguistique qui m'a permis d'établir mes convictions, même si ce fut au prix de prises de consciences difficiles. Le groupe ASSTICCOT, bien que plus récent, a eu le même effet stimulant. Je ne peux citer tous les noms mais je tiens à mentionner Nathalie Aussenac-Gilles, informaticienne à l'Irit, ma fidèle complice, avec qui nous essayons, depuis 1986, de construire une pensée interdisciplinaire cohérente autour de l'acquisition de connaissances à partir de textes.

J'ai la chance de travailler dans un laboratoire de linguistique qui ne connaît pas d'antagonismes insupportables. J'ai plaisir à être chercheur dans un tel contexte et j'exprime ma gratitude à tous les membres de l'ERSS : les directeurs, l'ancien (M . Plénat) et le nouveau (J. Durand), les chercheurs anciens ou actuels (M. Aurnague, M. Bras, A. Le Draoulec, S. Lignon, L. Sarda, et tous les autres…). Pour des questions moins scientifiques mais tout aussi importantes, ma reconnaissance va à Laurence Lamy et Nicole Serna. Elles savent combien leur amitié m'est précieuse.

Grâce aux projets auxquels j'ai participé, j'ai rencontré beaucoup d'acteurs du monde « économique ». Je leur exprime ma gratitude : cette confrontation avec une certaine réalité m'a gardée d'une vision uniquement universitaire de la recherche.

SOMMAIRE

Liste des figures et des tableaux

Introduction

Les différents termes du titre du mémoire présenté ici annoncent bien, me semble-t-il, les principales questions qui vont y être abordées. Ces questions s'organisent en trois grands thèmes, l'un concerne les bases de connaissances terminologiques et l'évolution des perspectives en terminologie, examinées tant d'un point de vue théorique qu'applicatif ; l'autre s'intéresse à l'utilisation des corpus en sémantique et à ses conséquences épistémologiques; le troisième évalue les possibilités de relations entre linguistique et informatique, lorsqu'il s'agit de construire une représentation des connaissances à partir d'un corpus. Ces thèmes sont intimement liés. En effet, la création des bases de connaissances terminologiques a été possible du fait de la rencontre de la terminologie et de l'informatique, mais une des conséquences en a été que la terminologie s'est rapprochée de la linguistique et s'est ancrée dans la sémantique de corpus. Ainsi, si les problématiques couvertes semblent vastes, en réalité, elles s'étayent mutuellement tout en permettant un éclairage nouveau sur des phénomènes linguistiques traditionnellement observés sur des bases fondamentalement différentes. Citons des exemples aussi divers que la polysémie, la notion de métalangage, le genre textuel, le fonctionnement des nominalisations, les relations entre linguistique théorique et linguistique appliquée : autant de problèmes qui sont examinés dans ce mémoire mais d'une manière qui leur donne un relief particulier, en lien avec un point de vue qui consiste à pratiquer une analyse parfaitement située.

Les bases de connaissances terminologiques (BCT)

Officiellement apparu en 1992, le terme de base de connaissances terminologiques scelle le rapprochement entre l'intelligence artificielle et la terminologie sur la question de la représentation des connaissances. C'est plus particulièrement la mise en réseau des concepts qui symbolise ce rapprochement, ce mode de représentation étant utilisé à la fois par la psycholinguistique par exemple et par les langages de représentation informatiques. Depuis cette date, les relations IA/terminologie ont beaucoup évolué et les BCT ne symbolisent plus la fusion initialement envisagée entre ces deux disciplines. Impliquée dans l'interdisciplinarité, chaque discipline a dû préciser ses présupposés. La problématique qui anime actuellement l'ingénierie des connaissances est celle des ontologies : comment les constituer, les formaliser et les pérenniser ? Ce n'est pas de ce point de vue que je me placerai

principalement mais de celui de la sémantique textuelle. En effet, la terminologie a dû, quant à elle, s'interroger sur ses rapports avec la linguistique. Il a fallu d'une part s'interroger sur les rapports entre lexicologie et terminologie et définir le *concept*, qui est la notion de base utilisée par la terminologie. Il a fallu aussi engager une réflexion sur le corpus dans la constitution des données terminologiques et surtout sur la manière de le prendre en compte dans la construction de ces données, c'est-à-dire dans la construction d'un sens. La vision référentielle, traditionnelle en terminologie, est ainsi apparue comme insuffisante voire inadéquate pour rendre compte du fonctionnement des textes spécialisés, ce qui a accéléré le développement de la terminologie textuelle.
Curieusement ainsi, une des conséquences de la rencontre entre terminologie et IA a été d'amener chacune à revisiter ses postulats et la façon de poser sa problématique. Le mariage a donc été particulièrement fructueux et les liens initiaux – de symbiose, pour reprendre un qualificatif de Skuce et Meyer (Skuce et Meyer, 1991) – se sont transformés et enrichis.

Corpus, terminologie et sémantique

Initialement, l'utilisation des textes en terminologie est liée au fait que les linguistes-terminologues ne peuvent pas faire confiance à leur intuition linguistique puisqu'ils n'ont pas de compétence de locuteurs sur le domaine concerné. Mais, avec la prise en compte de textes réels, ces mêmes linguistes-terminologues sont aussi confrontés au fait que l'utilisation de corpus ne peut pas correspondre à une simple substitution des données introspectives par des données attestées. La conscience de cette mutation, particulièrement prégnante surtout lorsqu'on travaille dans le cas d'une demande réelle, donne un éclairage puissant sur les phénomènes à étudier. En effet, ne considérer les textes que comme un ensemble d'attestations relève d'une vision parcellaire qui les ampute de la dimension dynamique conférée par la situation dans laquelle ils ont été produits. C'est le premier constat, inévitable, que l'on fait lorsque l'on construit une BCT à partir d'un corpus. L'autre constat, beaucoup moins fréquent, vient de ce que la prise en compte de la situation de production des textes n'est pas suffisante pour expliquer certains choix de représentation. Il est évident que la construction d'une représentation relationnelle à partir d'un corpus relève d'une interprétation. Ce choix interprétatif peut-être parfois (ou en partie) justifié par des régularités de fonctionnement en corpus, en lien avec la situation de production des textes. Mais souvent, c'est l'objectif de la modélisation qui guide aussi les choix de représentation. Un des objectifs de ce mémoire est de donner à la situation d'interprétation une place comparable à celle qu'on a pu donner à la situation de production des textes. C'est en effet l'objectif interprétatif qui donne des critères de constitution du corpus, qui permet de justifier des choix de modélisation et qui, au bout du compte, joue un rôle majeur dans l'évaluation des résultats.
Ainsi examinée, l'analyse sémantique de corpus s'inscrit dans un contexte doublement situé ; finalement, il s'agit de comprendre comment cette double situation (de production des textes et d'objectif de l'étude) agit sur l'interprétation. Le cas de la construction de BCT constitue un exemple parfait de ce processus. En effet, l'objectif de la modélisation doit être particulièrement travaillé avant la constitution du corpus et le démarrage de l'étude proprement dite. Mais je soutiens qu'il s'agit d'une version grossie de ce qui est toujours à l'œuvre dans l'analyse d'un corpus, que ce soit dans une perspective théorique ou appliquée : l'objectif ou l'hypothèse de l'interprétation joue un rôle déterminant dès la mise en place de l'étude, avant même la constitution du corpus.
Réalisée dans ce cadre, la compréhension de beaucoup de phénomènes linguistiques gagne en dynamisme. Le revers de la médaille est sans doute que la variation apparaît alors comme omniprésente et difficile à contrôler et, encore plus, à prédire. Et cet élément-là est sans doute difficilement acceptable pour beaucoup de linguistes qui souhaiteraient inscrire leur réflexion dans un paradigme scientifique proche des sciences « dures ». Pourtant, en lien avec la

situation de production des textes et d'interprétation des corpus, on peut repérer des régularités, en particulier sans doute si l'on essaie de généraliser les résultats grâce à la notion de genre. Ces régularités sont beaucoup moins stables que celles que l'on pense repérer par introspection mais elles sont aussi certainement beaucoup plus proches de la réalité des fonctionnements langagiers.

Construction de BCT et informatique

Ainsi que je l'ai signalé, la notion de BCT est, dès l'origine, indissociable de l'informatique. Cette rencontre avec l'informatique se fait de deux manières. D'une part pour modéliser puis formaliser les données terminologiques construites à partir du corpus et, d'autre part, pour travailler le matériau textuel afin de construire ces données. D'un autre point de vue, les informaticiens, habitués à travailler sur le terrain, contribuent à faire émerger des demandes applicatives (par exemple besoin d'un index, d'un thésaurus ou d'une aide à la traduction…), pour lesquels la construction d'une BCT peut constituer, en tout cas en partie, une réponse.
En chacun de ces points, les possibilités de collaboration entre termino-linguistique et informatique doivent être examinées.
Les modèles de données qui ont été proposés pour les BCT ont été souvent le fait d'informaticiens, qui avaient pour perspective la formalisation et le raisonnement. Ce genre de perspective a des conséquences non négligeables sur la façon de considérer les phénomènes sémantiques. Il nécessite d'une part que le sens soit discrétisé, c'est-à-dire manifesté par des formes (des formes lexicales, le plus souvent des noms ou des groupes nominaux) et d'autre part qu'il soit maîtrisé ; si, par exemple, la polysémie est prise en compte, c'est avec la nécessité que tous les sens possibles aient été identifiés, avant la formalisation mais aussi souvent avant l'analyse de discours réels. Lorsqu'il est poussé jusqu'à son extrême, ce mode de représentation devient carrément incompatible avec la description des phénomènes sémantiques ; par exemple, la détermination d'une racine unique pour les taxinomies, indispensable pour la formalisation, n'a souvent pas de sens pour la sémantique, particulièrement lorsqu'elle est pratiquée à partir de corpus. On rencontre le même type de difficultés avec la représentation relationnelle. Il ne s'agit pas d'un mode de représentation qui va de soi et, contrairement à ce que pensent beaucoup d'informaticiens, il ne s'agit pas d'un mode de représentation qui permet d'introduire la transparence en nettoyant la langue de ses handicaps inhérents : la représentation sous forme de réseaux relationnels peut introduire de l'ambiguïté. Mais, au-delà de ces limites, et surtout, à condition que la situation d'élaboration de ces réseaux soit très claire, ce mode de modélisation peut aussi constituer un cadre et un guide pour l'interprétation, à la fois parce qu'il permet de focaliser sur certains phénomènes sémantiques et parce qu'il oblige le linguiste à faire des choix et à les rendre explicites, c'est-à-dire à stabiliser, voire maîtriser son interprétation.
Le mode d'utilisation des outils de traitement automatique de la langue (TAL) relève aussi d'un double point de vue. Parce qu'ils ne sont pas réalisés avec un point de vue neutre sur la langue et parce qu'ils visent un objectif souvent prédéterminé, ils peuvent être trop contraignants voire parasiter l'analyse des textes. En revanche, si on connaît parfaitement les choix qui ont présidé à leur conception, si on connaît leurs possibilités et leurs limites et si l'on sait parfaitement ce que l'on souhaite en faire, ces outils peuvent être utiles et, dans le cas d'analyse de corpus volumineux, ils sont indispensables. Ils peuvent mettre en évidence, par rapprochement de contextes, des régularités invisibles « à l'œil nu » et permettre ainsi de tester très rapidement des hypothèses linguistiques ; de ce point de vue-là, ils constituent, pour la sémantique de corpus, un élément de mutation majeure.
Il est évident ainsi que l'informatique, à la fois parce qu'elle s'inscrit dans des demandes « sociétales » identifiées, parce qu'elle permet la mise à disposition de textes sous format électronique et qu'elle fournit des outils, contribue de manière radicale à l'évolution de la

sémantique. Mais il revient à la sémantique de s'interroger sur la façon dont l'informatique vient éclairer ou au contraire biaiser ses observations. Les questions sur le sens ne sont pas fondamentalement différentes de ce qu'elles étaient il y a cent ans mais il est indéniable qu'elles sont radicalement renouvelées par les possibilités de l'informatique et que ces possibilités introduisent une dynamisation des réflexions. Il est ainsi capital que la réflexion sur le sens dans les textes, en lien avec l'utilisation d'outils, ne soit pas confiée aux seuls informaticiens. Cela suppose bien sûr un effort de la part des sémanticiens, qui doivent comprendre les présupposés, les objectifs et les méthodes des informaticiens mais cette confrontation est extrêmement enrichissante. C'est en tout cas, l'expérience que je tire de ma rencontre avec l'informatique dans mon projet de construire des BCT à partir de corpus.

S'il vise à affirmer des positions théoriques, ce mémoire est aussi une sorte de compte-rendu de mon expérience d'analyste de corpus, dans le cadre de situations d'interprétations particulières, pour construire des BCT. Ma conviction sur la nécessité de prendre en compte très tôt l'objectif de la modélisation est le fruit d'une réflexion nourrie de cette expérience. A l'origine, je croyais fermement en la possibilité de constituer un modèle à partir d'un corpus sans qu'il soit besoin de faire intervenir un objectif quelconque, espérant que le corpus et ses régularités « immanentes » seraient suffisants pour garantir la pertinence et la cohérence du modèle. Il doit rester des traces, dans ce mémoire, de cet espoir déçu, même si j'ai tenté d'effacer la dimension chronologique de ma réflexion pour ne rendre compte que de son aboutissement.

Le mémoire s'organise en cinq chapitres. Le premier présente le cadre sémantique dans lequel je me place. Après un panorama de la façon dont les courants de la linguistique prennent en compte les corpus, il argumente en faveur de la nécessité d'une sémantique textuelle qui, tout à la fois, prenne en compte la réalité des variations langagières et repère des régularités de fonctionnement. Ce mode d'analyse n'est possible que s'il est doublement contrôlé, par la situation de production des textes et par l'objectif de l'interprétation. Le deuxième chapitre concerne la problématique des bases de connaissances terminologiques, à l'intersection de l'ingénierie des connaissances et de la terminologie textuelle. Le troisième chapitre présente le modèle de données que nous avons constitué, dans une perspective de modélisation (plus que de formalisation) mais en situant cette étape dans un processus qui pourrait aller des textes aux modèles formels de l'intelligence artificielle. Les principales études qui sont nécessaires pour remplir les champs ainsi définis sont aussi présentées. Le quatrième chapitre s'intéresse à certains phénomènes et à leur analyse dans des corpus réels. Il met ainsi en évidence des fonctionnements qui confirment ou au contraire invalident des descriptions réalisées de manière intuitive sur des phénomènes tels que le « repérage » de termes, la polysémie et le fonctionnement des nominalisations en corpus technique. Le cinquième chapitre s'intéresse plus particulièrement à la question des relations conceptuelles et à la possibilité de baser leur construction sur des portions de textes auxquelles on attribue le rôle de marqueurs de ces relations.

Plusieurs projets menés avec des entreprises ou des organismes publics sont évoqués dans le mémoire, je les décrits rapidement ci-dessous. Ils sont présentés avec plus de détails en annexe (demande et réponse faites, interprétation en termes théoriques).

MMS (Matra Marconi Space), 1991-1993 : réalisé dans le laboratoire mixte ARAMIIHS : constitution d'une base de connaissances terminologiques pour l'aide à la rédaction et l'aide à la formation.

CNES 1 (Centre National d'Etudes Spatiales), 1995 : mise au jour de points de vue grâce à une réflexion sur l'étude de la polysémie en corpus.

EDF, 1997-1998 : étude d'un Manuel, MOUGLIS (Méthodes et Outils de Génie Logiciel pour l'Informatique Scientifique) et constitution d'une Base de connaissances terminologiques afin d'améliorer l'accès au contenu.

DDE de la Haute-Garonne, Projet SGGD (Système de Gestion Globale des Déplacements), 1998-1999 : constitution de 4 bases de connaissances terminologiques correspondant aux corpus de 4 organismes afin d'élaborer un référentiel terminologique.

CENA (Centre d'Etudes de la Navigation Aérienne), 1999-2003 : étude d'un corpus de dialogues pour identifier les modes d'expression du dysfonctionnement technique.

CNES 2, 2002-2004 : définition de méthodes linguistiques pour repérer l'évolution des connaissances dans le temps.

Chapitre I
La sémantique de corpus ou la confrontation avec la réalité langagière : pour une sémantique doublement située

La linguistique a connu un changement épistémologique majeur au début du XXè siècle, qui a été en grande partie motivé par la volonté de lui donner un statut de science. Dans cette évolution, la nécessité de construire un objet autonome, sans aucun lien avec le réel synchronique, ni avec la dimension diachronique, ni avec l'individu parlant, ni avec la variation sociale est apparu comme fondamental : un objet qui a été posé d'emblée comme idéal et, dans le même temps, observable d'une manière neutre car clairement séparé de celui qui l'analyse.

Beaucoup de linguistes pensent primordial de maintenir cette rupture avec la réalité des manifestations langagières, ce qui leur permet, en tout cas l'espèrent-ils, de faire accéder la linguistique au rang de science de la nature et la langue au rang d'objet digne d'étude. Ce choix épistémologique qui a eu des conséquences très fructueuses apparaît maintenant comme un cadre trop rigide qui ne tient pas compte de deux éléments. D'une part, l'idéalisation de l'objet n'a de pertinence que tant que cet idéal n'est pas confronté à la réalité des usages. Or, cette réalité fait maintenant irruption dans le champ de la linguistique, à la fois du point de vue de la demande sociale et de l'évolution de la réflexion théorique. D'autre part, même du temps où le structuralisme triomphait et influençait toutes les sciences humaines, les études sur la réalité des faits langagiers (texte, discours…) ont continué dans l'ombre et il apparaît aujourd'hui que les travaux qui se sont développés ne sont pas sans intérêt, voir par exemple la présentation de Rastier (2001).

La réflexion menée à partir de corpus revient à se demander comment la prise en compte du réel, de toutes façons incontournable, vient interroger la linguistique et tout particulièrement la sémantique. Il ne peut s'agir ni de revenir à une époque pré-structuraliste ni de nier une relation de l'objet langue avec d'autres manifestations humaines, sociales, psychologiques, historiques… La question est finalement toujours la même pour un linguiste : qu'est-ce qui fait que le discours a du sens, c'est-à-dire, s'inscrit dans une perspective collective tout en laissant des possibilités d'expression individuelle ? Cette question peut se décliner de plusieurs manières : la notion de système est-elle indispensable pour expliquer les régularités qui peuvent être mises au jour ? Par rapport à quels contours se définissent ces régularités ? Quelles stabilités les caractérisent (sociales, diachroniques) ?

Ce premier chapitre consistera en un panorama des points de vue qui existent à propos de la prise en compte des données textuelles en linguistique en général et tout particulièrement en sémantique[1]. Il s'agit d'essayer de montrer quels présupposés animent ces réflexions et les difficultés qu'ils soulèvent.
Dans une seconde partie, je présenterai comment je conçois l'analyse de corpus et, par voie de conséquence, ce que me semblent être les possibilités de la recherche en sémantique de corpus.

1. Modes de prise en compte des corpus en sémantique

Ce n'est sans doute pas par hasard si la sémantique est la discipline la plus jeune des sciences du langage tant le sens peut paraître labile et difficile à maîtriser : la notion de régularité est souvent mise à mal dès que l'on s'intéresse au sens. C'est peut-être pour s'éviter un trop grand vertige que beaucoup de sémanticiens refusent de baser leurs descriptions sur des études de productions réelles. Le mode de prise en compte des corpus peut être une façon d'éclairer les points de vue en sémantique et de mettre en évidence les principales questions qui s'y posent. Sans vouloir être exhaustive, je propose de balayer les différents courants de la sémantique en les organisant en trois grandes classes, depuis ceux pour qui l'utilisation des corpus est hors de propos jusqu'à ceux pour qui elle est le fondement de leur approche.

1.1. La notion de corpus n'a pas de pertinence

Pour une part importante des linguistes, le recours aux corpus est exclu du champ de la linguistique en général. C'est le cas de la linguistique structurale et de la linguistique générative qui justifient toutes deux leurs positions par la volonté d'inscrire la discipline linguistique dans une perspective scientifique.
Dans l'analyse générative, l'analyse de productions réelles n'est pas envisageable car elle est contraire à l'objectif de description de la « compétence » d'un « locuteur idéal », la description empirique est alors clairement présentée comme incompatible avec une vision théorique.

> « Ce travail sur l'anglais noir non standard ressemble à l'étude du coréen ou des langues amérindiennes. C'est un travail très utile. Mais ce qui me gêne, c'est sa prétention théorique. Nous avons affaire à de la bonne linguistique descriptive. » (Chomsky, 1977, 73).

Il s'agit de construire un modèle explicatif qui décrive « l'intuition linguistique – la compétence tacite – du sujet parlant » (Chomsky, 1971, 45). On peut s'étonner de trouver ce terme d'intuition dans une démarche qui revendique de s'inscrire dans une approche qui se veut aussi (rigidement) scientifique. On pourrait penser en effet que s'il est un élément variable chez un humain, c'est bien l'intuition. Mais pour Chomsky, d'une part, l'intuition ne concerne pas l'analyste mais le locuteur (même s'il s'agit d'une seule et même personne) et d'autre part, par hypothèse, cette intuition, cette compétence est considérée comme unique puisqu'elle est le fait d'un locuteur idéal.
Ainsi figée dans une idéalité, la langue perd son essence sociale et le sens son dynamisme créatif. La variation est hors de propos surtout si elle vient perturber le modèle explicatif qui a été élaboré.
Comme le montre Vandeloise (Vandeloise, 1991), c'est justement pour éviter le recours à l'intuition, considérée comme incompatible avec une démarche scientifique, que le

[1] En réalité, je crois que toute la linguistique est traversée par la question du sens, question que l'on peut sans doute momentanément suspendre pour des études ponctuelles en phonologie, voire en morphologie ou en syntaxe mais qui devrait être toujours posée à un moment ou l'autre de la réflexion.

structuralisme s'est construit un objet, la langue, distinct de réalisations réelles, relevant de la parole :

> «En faisant ainsi des significations l'objet de la sémantique, le structuralisme veut rendre celle-ci indépendante du locuteur et du contexte d'énonciation, la fondant sur des critères non-intuitifs et non introspectifs. » (Vandeloise, 1991, 71).

Structuralisme et générativisme justifient de manière différente l'autonomie de la langue. Pour Chomsky, les arguments sont d'ordre psychologique et s'appuient sur les notions d'innéité et d'universaux ; pour Saussure, parfaitement conscient de la dimension sociale de la langue, les arguments sont d'ordre méthodologique : il est indispensable de se construire un objet dégagé de ses attaches sociales pour établir la distance qui permettra l'analyse objective que requiert une approche théorique. Dans ce mouvement de détachement du réel, il semble que, pour Saussure, l'objet d'étude passe du domaine sociologique à celui du psychologique :

> « On peut donc concevoir une science qui étudie la vie des signes au sein de la vie sociale ; elle formerait une partie de la psychologie sociale, et par conséquent de la psychologie générale […]. La linguistique n'est qu'une partie de cette science générale.» (Saussure, 1982, 33).

Malgré des points de départ différents, on retrouve des constantes dans le structuralisme et le générativisme :

- Une désocialisation du langage : par hypothèse dans le générativisme, par définition méthodologique dans le structuralisme, l'objet d'étude n'a aucun lien avec les situations sociales dans lesquelles il apparaît. Le recours au corpus n'a pas de sens dans ce type d'approche.
- Une recherche de scientificité qui consiste à tout mettre en œuvre pour rapprocher la linguistique d'une science « dure », ce qui contribue à fixer un objet extérieur à l'analyste, consistant en des faits observables « objectivement », qui sont donc des données comme dans les sciences de la nature.
- Une méfiance par rapport au sens. Le sens est l'élément qu'il faut contrôler car c'est de lui que pourrait venir la menace de déstabilisation du système. Le générativisme est le plus radical sur ce point : le sens ne dépend que de l'agencement syntaxique et il n'a donc pas d'objet pour la linguistique. Le structuralisme s'intéresse au sens mais cherche à le contrôler par une analyse strictement interne, qui ne fait intervenir que l'aspect différentiel intra-langue.
- Une volonté de neutraliser l'intuition et la subjectivité de l'analyste.

Dans ce type d'approches, on peut voir se dessiner l'association suivante : l'appel aux productions réelles menace la dimension scientifique de la discipline, en particulier parce qu'elle suppose la prise en compte de la variation et, plus généralement, du sens, voire de l'interprétation.

On perçoit bien quel est l'objectif de ces approches : maîtriser le fonctionnement linguistique en en prévoyant toutes les manifestations possibles et, pour ce faire, maintenir par tous les moyens la variation hors du champ d'étude. Cet objectif est louable tant on comprend qu'il a pour ambition de hisser la linguistique au rang d'une science respectable et digne d'intérêt. Et il serait injuste de minimiser les apports du structuralisme et du générativisme. En particulier, le structuralisme a permis de dégager la langue d'une fonction référentielle qui biaisait l'analyse linguistique. Malheureusement, on ne peut faire éternellement abstraction du réel et il est patent que l'application des méthodes du structuralisme (approche différentielle du

système, unique, de la langue) ou du générationalisme est souvent incompatible avec l'étude de corpus réel[2].

1.2. Recours au corpus « introspectif »

Dans beaucoup d'approches sémantiques ou pragmatiques, la réflexion est menée à partir de données introspectives. D'une certaine façon alors, le corpus est constitué par les éléments langagiers que l'analyste s'autorise au moment même de l'élaboration de sa pensée. Il y a bien recours à des faits attestés (par le linguiste lui-même) mais ces faits sont élaborés dans une situation tout à fait particulière qui est celle du test linguistique. Or, d'une part, il s'agit d'une situation artificielle, qui ne correspond pas à une utilisation « spontanée » du langage, et, d'autre part, cette situation est empreinte de l'objectif de l'analyste qui vise à démontrer une hypothèse[3]. Il est évident que cette visée risque d'amener à filtrer les productions langagières et à ne conserver que celles qui répondent à l'hypothèse[4]. Ce problème a été souvent discuté, en particulier par Corbin (1980). Ce type d'approche est moins rigide que le courant générationaliste car il admet la notion de variation, en tout cas tant qu'elle ne remet pas en question la notion de système. La notion d'intuition est d'ailleurs aussi revendiquée comme un des éléments d'une science, elle-même souvent considérée comme empirique.

> « No empirical science can operate without human intuitive judgement intervening at some point.» (Cruse, 1986, 10).

On retrouve ce point de vue sur les corpus, qui n'admet qu'une variation contrôlée, dans bon nombre de travaux en sémantique lexicale ou cognitive mais aussi dans certains travaux sur l'énonciation.

1.2.1 Sémantique lexicale

Jusqu'à une date récente, la plupart des travaux en sémantique ne se sont intéressés qu'aux seuls mots, considérés comme les unités majeures. Ces travaux s'inscrivent certainement dans une perspective ancienne qui visait à dénommer, c'est-à-dire désigner, les objets du monde.
Dans la plupart de ces travaux, on trouve une double préoccupation, d'une part, comment expliquer les relations de la langue avec le monde et, d'autre part, comment expliquer les relations des mots entre eux. La première préoccupation consiste à expliquer les rapports de la langue avec le réel. Ce lien est considéré comme stable parce qu'il est de nature perceptive et s'appuie sur l'existence d'universaux ; il est aussi peu variable d'un individu à l'autre :

> « D'un point de vue sémantique, les expressions nominales sont des expressions référentielles. » (Lyons, 1980, 80).

---

[2] Il faut toutefois noter que le structuralisme n'est pas resté dans la vision, qui peut paraître figée, que lui avait donné Saussure (essentiellement pour des raisons méthodologiques qui ont d'ailleurs fait leurs preuves). Greimas, par exemple, n'est pas du tout opposé à l'utilisation de corpus : « Un certain nombre de textes individuels, à condition qu'ils soient choisis d'après des critères non-linguistiques garantissant leur homogénéité, peuvent être constitués en corpus et […] ce corpus pourra être considéré comme suffisamment isotope. » (Greimas, 1966, 93). Je ne crois pas que le générationalisme ait suivi la même évolution.

[3] Même lorsque cette approche recourt à des données attestées, c'est toujours dans un second temps, pour confirmer l'hypothèse et la plupart du temps, sans que les caractéristiques des textes d'où sont extraits les exemples soient mentionnées.

[4] L'expérience de l'analyse de textes réels permet de faire le constat très fréquent d'une inadéquation entre les résultats fournis par l'approche introspective et les faits textuels. Pour être précise, cette inadéquation est souvent en lien avec la fréquence des phénomènes : certains décrits comme marginaux sur des bases introspectives apparaissent comme massifs dans certains textes. Mais il n'est pas rare aussi de rencontrer des manifestations linguistiques pourtant affublées d'une astérisque dans des descriptions introspectives ; par exemple, les analyses de Manning (Manning, à paraître) sur les contextes syntaxiques d'apparition de verbes comme *consider* ou *regard* dans le New York Times font apparaître des différences importantes entre les jugements de grammaticalité *a priori* et la réalité des usages.

« Les items lexicaux présupposent l'existence (c'est-à-dire dans ce sens et uniquement dans ce sens réfèrent à) de concepts, c'est-à-dire d'entités générales plus connus sous le nom d'universaux lorsqu'elles se présentent sous une forme nominale. » (Kleiber, 1981, 24).

« Dans une vaste série de cas, nos conceptualisations ou notre modèle mental du monde est largement identique d'un individu à l'autre, ce qui forme une sorte de socle pour une intercompréhension réussie. » (Kleiber, 1999, 21-22).

C'est seulement une fois fixé le sens référentiel que l'on peut envisager de prendre en compte le fonctionnement des mots en contexte, ce qui peut laisser une certaine place à la variation, dans la mesure où on la contrôle en expliquant comment elle ne remet pas en question le sens préfixé. Mais il est bien entendu que ce contexte ne peut être que linguistique :

« There are good reasons for a principled limitation to linguistic context ; first, the relation between a lexical item and extra-linguistic contexts is often crucially mediated by the purely linguistic contexts… ; second, any aspect of extra-linguistic context can in principle be mirrored linguistically ; and, third, linguistic context is more easily controlled and manipulated. » (Cruse, 1986, 1).

Le contexte extra-linguistique, quant à lui, est un risque pour la stabilité et il est donc au mieux relégué à un second plan :

« Cette existence [d'éléments « existants » réels ou fictifs] leur est garantie par cette modélisation intersubjective stable à apparence d'objectivité qui caractérise notre appréhension du monde. Modélisation qui se trouve appréhendée par deux sources : par notre expérience perceptuelle, mais aussi par notre expérience socio-culturelle incluant la dimension historique. La première étant donné notre commune condition humaine, a plus de chances d'apparaître universelle et donc stable, que la seconde liée aux groupes sociaux et à la dimension temporelle, donc au changement. » (Kleiber, 1999, 27).

Finalement, même lorsque le contexte est pris en compte, le sens, référentiel, est sauf :

« Le contextualisme n'arrive finalement pas – heureusement (?) – à se débarrasser d'un sens linguistique non construit. » (Kleiber, 1999, 45).

Dans les approches de sémantique référentielle, l'introspection est mise en œuvre aux deux niveaux : pour fixer le sens des mots, c'est-à-dire pour lui donner une référence (c'est-à-dire un lien avec le réel tel qu'il est humainement (linguistiquement ?) perçu) et pour évaluer comment ce sens s'adapte à différents contextes, ces contextes étant eux-mêmes construits sur des bases intuitives. L'hypothèse fondamentale est qu'il n'y a pas de différences majeures d'un individu à l'autre et pas de différence majeure de sens dans les différentes occurrences d'un mot.
La notion de corpus apparaît seulement à travers celle de contextes, qui sont les contextes que l'on construit pour expliquer le sens d'un mot et les éventuelles variations de sens que l'on tolère.
Du point de vue de la prise en compte des productions réelles, l'approche de la sémantique cognitive n'est pas très différente de celle de la sémantique lexicale référentielle.

1.2.2 Sémantique cognitive

La principale caractéristique de la sémantique cognitive vient de ce qu'elle revendique un lien étroit avec la psychologie. Là où, pour le générationalisme, le structuralisme ou la sémantique référentielle, ce lien est revendiqué comme présent mais posé *a priori* sans qu'il influence l'analyse strictement linguistique, pour la sémantique cognitive, ce lien est omniprésent et s'oppose à une analyse autonome du langage :

« La sémantique cognitive affirme que les concepts lexicaux ne peuvent être étudiés de manière adéquate qu'en relation avec les capacités cognitives générales de l'homme et, en particulier, qu'il

n'y a pas d'organisation spécifiquement linguistique ou sémantique de la connaissance, séparée de la mémoire conceptuelle au sens large. » (Geeraerts, 1991, 27).

Cette désautonomisation de la langue est d'ailleurs fortement interrogée par des sémanticiens de tous bords :

« Il nous semble que la recherche cognitive repose sur deux postulats d'ordre philosophique qu'elle transforme peut-être en thèses scientifiques. Le dualisme traditionnel entre l'esprit et le cerveau doit être restreint […]. L'homme peut simuler artificiellement les processus mentaux […]. Un troisième postulat, gnoséologique celui-ci, est généralement accepté, sans être discuté pour autant : la connaissance est une représentation symbolique du réel […]. » (Rastier, 1991, 35).

« Il y a cependant un danger, celui de perdre de vue le(s) fonctionnement(s) linguistique(s) au profit de principes cognitifs, dont la généralité est tellement puissante qu'elle ne peut être prise en défaut par les phénomènes linguistiques, ce qui n'est qu'une autre façon de dire que, linguistiquement, elle n'a plus réellement de vertus explicatives. » (Kleiber, 1990, 15).

Pourtant, les hypothèses mises en avant par la sémantique cognitive sont très proches de celles de la sémantique référentielle ; elles concernent les capacités perceptives humaines considérées comme quasi identiques d'un individu à l'autre :

« Il est clair que la qualification des couleurs, et celle des signifiés en général, est en grande partie motivée par notre système perceptuel et conceptuel. » (Vandeloise, 1991, 93).

Toutefois, mettant l'accent sur la polysémie et la non-discrétion des signifiés, la sémantique cognitive semble assez tolérante à l'idée de variation, diachronique ou sociologique, à condition qu'elle puisse être contrôlée par la notion de prototype qui est la notion fondatrice de cette sémantique. Pour autant, il ne s'agit pas de s'appuyer sur des manifestations linguistiques réelles. Comme pour la linguistique référentielle, les contextes qui sont convoqués sont construits sur des bases introspectives dans une visée explicative forte, la plupart du temps en lien avec des considérations extra-linguistiques comme pour le traitement de l'espace chez Vandeloise :

« J'ai montré ailleurs (Vandeloise, 1986) que les caractéristiques des contextes qui déterminent l'usage des termes spatiaux ne dépendent pas tant de concepts géométriques, topologiques ou logiques que de concepts fonctionnels liés à l'utilisation de l'espace. » (Vandeloise, 1991, 96).

Dans son livre de 1999, Werth montre que les cognitivistes américains (Fillmore, Minsky, Lakoff, Langaker) ont particulièrement travaillé sur la notion de *frame* (Werth, 1999, 42-43). Cette notion est aussi fortement basée sur la perception élaborée sous la forme d'une expérience :

« Fillmore comes up with many such examples, and from them we can perhaps arrive at some kind of intuitive understanding of what a frame is. It seems to be something like an 'area of experience' in a particular culture. In the terms developed in the present book, […] we might say that a frame is a cognitive space, mapping out an experiential category. » (Werth, 1999, 106).

Par ailleurs, du fait de la rencontre revendiquée de la linguistique cognitive avec la psychologie, le recours à des données linguistiques textuelles est souvent remplacé par des tests psycholinguistiques qui visent par exemple à montrer l'universalité de tel ou tel concept, tests la plupart du temps basés sur un calcul de temps de réponse. Comme en d'autres occasions, on retrouve le besoin de donner un statut scientifique à l'approche mise en œuvre :

« Sur le plan méthodologique, la psycholinguistique se distingue de la linguistique traditionnelle – linguistique descriptive, linguistique historique et linguistique générale – parce qu'elle est plus souvent de type quantitatif que de type descriptif et logique. De ce point de vue, la psycholinguistique se rapproche davantage de la psychologie expérimentale, de la biologie et de façon générale, des sciences exactes. » (Keller, 1985, 4).

Sémantique référentielle et sémantique cognitive, également centrées sur l'étude des mots, ne s'intéressent aux contextes que lorsque, convoqués dans un deuxième temps, ils viennent servir d'exemples à un sens considéré comme stable, c'est-à-dire qui prédit toutes les variations possibles. Dans un tout autre type d'approche, on retrouve le même recours au corpus introspectif : dans les approches énonciatives.

1.2.3 Linguistique de l'énonciation : Sémantique instructionnelle

Avec la linguistique de l'énonciation, la variation est au coeur de la problématique, y compris (en tout cas pour Benveniste) la variation individuelle :

> « L'énonciation est cette mise en fonctionnement de la langue par un acte individuel d'utilisation. » (Benveniste, 1974, 80).

Mais cette variation n'est pas le point de départ de l'étude. Le point de départ reste le système linguistique, qui doit prendre en compte la notion de variante. Pour ce faire, dans la perspective de Ducrot par exemple, il ne s'agit pas d'expliquer le sens littéral d'un énoncé mais de mettre au jour son sens instructionnel, c'est-à-dire d'expliquer les multiples lectures possibles :

> « La signification de la phrase ne doit pas être confondue avec son "sens littéral" qu'on retrouverait identique à lui-même dans le sens des énoncés. La phrase dit seulement ce qu'il faut faire pour découvrir le sens. » (Ducrot, 1980, 17).

Dans ce type d'approche, la signification doit pouvoir s'adapter à toutes les situations d'énonciation ; ces situations possibles doivent donc être imaginées au moment de la description afin qu'elle soit compatible avec toute nouvelle situation.

> « [La signification (du mot ou de la phrase)] contient surtout, selon nous, des instructions données à ceux qui devront interpréter un énoncé de la phrase, leur demandant de chercher dans la situation de discours tel ou tel type d'information et de l'utiliser de telle ou telle manière pour reconstruire le sens visé par le locuteur. » (*ibid*.,12)

> « [Le linguiste] peut demander au modèle d'envisager dès le niveau du système abstrait un maximum de valeurs possibles pour une structure donnée, le contexte fonctionnant alors surtout comme un filtre chargé de sélectionner dans ce vaste ensemble de valeurs virtuelles […]. » (Kerbrat-Orechhioni, 1996, 53).

Ainsi, si la variation est prise en compte, elle est entièrement prévue par le modèle. Pour cela, le linguiste doit reconstituer toutes les situations possibles d'énonciation. Les textes peuvent être utilisés, mais ils sont alors considérés comme des supports d'attestations, au même titre que le corpus introspectif :

> « Qu'il s'agisse de la sémantique des mots ou de celle des phrases, le linguiste est alors amené à prendre l'analyse de textes comme instruments nécessaires – que ces textes soient écrits ou oraux, qu'ils soient le texte de discours effectivement tenus ou attestés, ou de discours imaginaires (mais que l'on imaginera en même temps qu'un environnement les rendant possibles).» (Ducrot, 1980, 9).

Ainsi, lorsque la linguistique énonciative parle de situation extra-linguistique, elle ne l'étudie pas dans son rapport avec la production de textes mais comme un élément dont il faut tenir compte parce qu'il influence le sens que l'interlocuteur va donner à un énoncé. Ainsi, comme le dit Boutet :

> « […] dans la linguistique de l'énonciation telle que développée par Benveniste et Culioli, la notion de situation est rigoureusement distinguée des situations concrètes et empiriques dont parle, entre autres, Bakhtine. La situation d'énonciation est un construit théorique du chercheur, entièrement défini par l'acte même de l'énonciation . » (Boutet, 1995a, 54).

Les textes n'étant qu'un « instrument nécessaire », ils ne constituent pas le cœur de l'étude, c'est l'énoncé qui est l'unité maximale d'analyse :

> « […] la proposition ne peut entrer comme partie dans une totalité de rang plus élevé […]. Un groupe de propositions ne constitue pas une unité d'un ordre supérieur à la proposition […]. La phrase est l'unité du discours. » (Benveniste, 1966, 129-130).

Au mieux, le texte est une succession de phrases et c'est sans doute parce qu'il est nécessaire d'expliquer l'enchaînement des énoncés que bon nombre d'études de linguistique énonciative portent sur les connecteurs (par exemple sur 7 articles de l'ouvrage dirigé par Ducrot en 1980 (Ducrot, 1980), 4 portent sur un connecteur).

Au bout du compte, qu'il soit question de sens des mots ou des énoncés, une grande partie des travaux de sémantique s'est consacrée à construire des modèles et des théories sur des bases introspectives. Même lorsque la variation est modélisée, il est capital, pour ce type d'approche, qu'elle le soit une fois pour toutes et, en tout cas, avant la mise en situation réelle.

1.3. Recours à un corpus réel

Parallèlement, voire antérieurement aux approches linguistiques qui ont fait le choix de ne pas se fonder sur des manifestations langagières réelles, de nombreux courants, pour des raisons très diverses, se sont, eux, intéressés aux réalités linguistiques. Sans anticiper sur la suite de mon propos, il me semble que l'on peut dire que dans de nombreux cas (hormis sans doute l'analyse littéraire), l'utilisation de corpus s'est accompagnée d'un objectif que, si je n'avais peur de la dimension péjorative que peut contenir ce terme, je qualifierais « d'utilitaire ». Pas parce que l'objectif visé est prédominant dans ce type d'approche mais parce que la dimension textuelle suppose une telle plongée dans la réalité des usages que les textes sont non seulement situés par rapport à la situation où ils sont produits mais aussi très souvent accompagnés d'un objectif d'analyse qui, très certainement, influe sur l'étude réalisée.
Sans vouloir être exhaustive, je présenterai les recherches qui utilisent les corpus selon deux axes, l'un dans lequel le corpus est censé être représentatif de la compétence des locuteurs d'une langue, l'autre dans lequel il constitue l'unité d'analyse et, éventuellement, le point de départ d'une analyse plus générale.

1.3.1 Corpus comme représentatif de la compétence des locuteurs

A côté de travaux qui ne mettent en œuvre que le corpus introspectif pour décrire une langue, des travaux importants se sont développés, qui visent à une description basée sur des données réelles, c'est-à-dire un corpus.
L'utilisation d'un corpus « de langue » peut avoir trois objectifs : la description lexicologique, la description grammaticale, l'enseignement de la langue.
De manière très nette, cette approche s'est préférentiellement développée dans la communauté anglo-saxonne (Péry-woodley, 1995). La liste des corpus anglais présentée dans (Habert et al, 1997) est ainsi tout à fait parlante (Brown, LOB, Suzanne, London-Lund, …, BNC…). En France, on retrouve aussi cette approche mais de manière bien plus discrète.

1.3.1.1 Lexicologie et établissement de grammaires à partir de corpus

Lexicologie à partir de corpus

La lexicologie à partir de corpus consiste à essayer d'élaborer des définitions de dictionnaires à partir d'exemples attestés. Pearson présente comme pionnier le travail de Johnson qui publia en 1755 le premier dictionnaire dont les définitions étaient élaborées à partir d'usages réels :

« Johnson had begun his task of compiling a dictionary by reading the English writers which he deemed to be suitable sources for a dictionary of the English language. According to Reddick, Johnson's first step " was to mark passages in printed books to use as examples of usage " [..]. [These passages] were to be used not only for the selection of a word but also for the purposes of exemplification and as input for definitions. » (Pearson, 1998, 73-74).

Ce dictionnaire, enrichi et développé est devenu plus tard l'Oxford English Dictionary, publié en 1928. Mais le dictionnaire le plus fréquemment cité comme étant constitué à partir d'attestations réelles est le Cobuild. Son principal responsable, John Sinclair, a d'ailleurs mené une réflexion sur la lexicologie à partir de corpus qui inspire encore aujourd'hui beaucoup de travaux.

En France, ce n'est que depuis les années 60 qu'une démarche d'élaboration d'un dictionnaire à partir de corpus s'est mise en place avec la constitution de la base de données Frantext et du Trésor de la Langue Française. Il faut toutefois noter que cette entreprise n'a pas eu la même volonté de systématisation que le projet Cobuild ; le recours aux textes pour le TLF a surtout eu valeur d'attestation. Frantext est essentiellement composée de textes littéraires des XIXè et XXè siècles, ce qui lui a été souvent reproché. Il n'en reste pas moins que le projet TLF constitue une réelle avancée qui se manifeste par dix-sept volumes papier et par une version électronique désormais disponible sur le web (http//www.inalf.fr/tlfi/) qui présente de très réels intérêts.

Grammaires à partir de corpus

Pour Kennedy, les premières grammaires de l'anglais basées sur l'utilisation d'exemples informels remontent au début du XXè siècle (Jespersen, Kruisinga et Poutsma). (Kennedy, 1998, 17). Mais l'entreprise la plus aboutie de constitution systématique d'une grammaire de l'anglais à partir de corpus revient, plus récemment, à Randolph Quirk. Pour mener à bien son objectif, il constitua à partir de 1959 le Survey of English (SEU) Corpus contenant 200 extraits d'environ 5 000 mots chacun, censés être représentatifs de l'anglais écrit et oral. Ce corpus d'un million de mots fut étudié par de nombreux chercheurs et conduisit à la constitution d'une grammaire de l'anglais (Quirk et *al.*, 1985).

D'après Kennedy toujours, les études lexico-syntaxiques à partir de corpus se sont développées dans les domaines suivants (Kennedy, 1998, 89) : formes verbales (verbes transitifs, infinitifs, modaux, temps verbaux, particules pré et post-verbales), syntagmes nominaux (types de syntagmes nominaux, articles, quantifieurs, adjectifs, chaînes pronominales…), prépositions, adverbiaux, pré- et post-modification, subordination (conditionnelles, phrases nominales, temporelles…), complémentation (infinitives…), topicalisation.

Autant dire que peu de phénomènes grammaticaux de l'anglais semblent avoir échappé à une étude à partir de corpus. Mais ces études sont restées isolées ; ce n'est que plus récemment, et peut-être alors qu'on ne l'attendait pas sur ce terrain que Biber s'est lancé dans la prise en compte systématique des corpus pour construire une grammaire (Biber et *al.*, 2000)[5].

En France, des travaux d'analyse de phénomènes grammaticaux à partir de corpus existent ; mais il manque un recensement fin, qui permettrait de dresser un panorama afin de mieux évaluer les complémentarités entre ces travaux mais aussi les besoins en corpus. Comme je l'ai déjà souligné, je pense en effet qu'on a une bien meilleure vision des besoins en corpus lorsque l'on sait comment ces corpus vont être utilisés, tout au moins dans une première utilisation (ils peuvent être ensuite réutilisés pour d'autres objectifs).

[5] En effet, la mise au jour de types de textes, basés sur des régularités linguistiques, mais qui peuvent varier en fonction des phénomènes examinés semble incompatible avec la stabilisation que nécessite la définition d'une grammaire.

1.3.1.2 Enseignement à partir de corpus

Ce type d'objectif a ceci de particulier qu'il vise l'efficacité dans l'apprentissage d'une langue. On trouve toujours dans ces travaux l'idée qu'il y a un noyau de connaissances linguistiques à apprendre, qui constitue la base nécessaire pour se débrouiller dans une langue, que ce soit sa langue maternelle ou une langue étrangère. Ainsi, comme le signale Picoche, les travaux en lexicologie visent un ou l'autre objectif : une large couverture pour des résultats plutôt destinés aux natifs d'une langue afin de leur donner accès au sens de tous les mots, ou bien une couverture ciblée, pour des résultats plutôt destinés à l'apprentissage par des locuteurs étrangers :

> « L'étude du lexique d'une langue comporte deux pôles : la recherche du plus grand nombre possible de mots utilisables par un locuteur français d'une part : c'est la perspective du Trésor de la Langue française ; d'autre part, la recherche du petit nombre de mots très usuels, indispensables à la communication et communs à tous les locuteurs de langue française : c'est la perspective du Français Fondamental. » (Picoche, 1992, 48).

Comme pour la lexicologie ou la grammaire, les anglo-saxons ont été précurseurs dans la mise en place de l'apprentissage d'une langue à partir de corpus. Ainsi, dès le début du XXè siècle, Thorndike a mis en place un corpus de plus de 4,5 millions de mots et élaboré des listes de fréquences. Le point de vue de Thorndike n'était pas linguistique mais psychopédagogique[6] ; il fut un des théoriciens de l'apprentissage par essai/erreur et un des fondateurs du behaviorisme.

En France, le projet « français fondamental » s'est mis en place au milieu du XXè siècle (Gougenheim et *al.*, 1958). Clairement destiné à l'apprentissage du français par des étrangers, ce corpus d'un peu plus de 310 000 mots, essentiellement constitué d'enregistrements de conversations a permis de repérer les mots les plus fréquents du corpus, que l'on a considérés comme les mots les plus fréquents du français. L'objectif était de centrer l'apprentissage du français par des étrangers sur les premières centaines de ces mots.

Quel que soit l'objectif, la question commune à ces approches est celle de la représentativité du corpus. Pour toutes, le corpus doit permettre de rendre compte du fonctionnement d'une langue ; il est donc censé rendre compte de tous les usages de cette langue.

Pour réussir à atteindre ce but, deux méthodes de constitution de corpus ont pu être suivies (Péry-woodley, 1995), (Habert et *al.*, 1997, 146). L'une a consisté à constituer de très gros corpus (on espérait que la quantité permettrait une large couverture qui compenserait l'absence de choix raisonné des corpus sélectionnés), l'autre a consisté à constituer des corpus équilibrés, censés représenter tous les registres d'une langue (c'est le choix majoritaire des corpus anglo-saxons).

Les corpus équilibrés semblent bien mieux à même d'être représentatifs d'une langue. En réalité, leur constitution est basée sur un paradoxe. Il s'agit d'étudier les usages « réels » d'une langue, éventuellement en rendant compte des variations en fonction des registres ; mais ce corpus lui-même, que l'on a constitué pour qu'il soit représentatif, est basé sur une idée intuitive des registres et donc de ce qu'est l'usage réel de la langue.

C'est certainement le problème majeur qui traverse toute approche à partir de corpus : le corpus est censé venir pallier les problèmes liés à une approche seulement introspective mais la constitution de ce corpus est elle-même faite sur des bases intuitives. Diverses pistes ont été proposées pour pallier cette difficulté majeure (dont la définition de genres, *cf.* paragraphe 2), mais on doit reconnaître que ce problème reste la pierre d'achoppement de la linguistique de

[6] Il ne faut pas oublier que la constitution de corpus n'a pas toujours été inspirée par des objectifs linguistiques. Par exemple, Kubler (http:wall.jussieu.fr/tildenkubler/enseignement/Edcours.ppt) signale que, dès 1890, Käding avait constitué un corpus de 11 millions de mots pour étudier la fréquence d'apparition des séquences de lettres, pour améliorer les performances des sténographes !

corpus. Je reviendrai sur cette question de la représentativité du corpus dans le chapitre III, dans le cadre particulier de la constitution de bases de connaissances terminologiques, pour montrer comment ce problème peut être éclairé, sans être définitivement réglé.

1.3.2 Le corpus comme objet d'étude

Comme le montre Rastier par exemple (Rastier, 2001), l'analyse de textes est une tradition très ancienne et elle n'a jamais cessé. Depuis l'avènement d'une science linguistique, il est devenu nécessaire de situer l'analyse de textes par rapport aux préoccupations de cette discipline. Deux tendances semblent se dessiner. L'une considère que, dès lors qu'un corpus a été constitué, il devient l'objet d'analyse et il n'y a pas d'intérêt à s'interroger sur son statut par rapport à un éventuel système linguistique. L'autre considère que, tout en ayant une cohérence propre, le corpus n'en est pas moins une manifestation d'un usage de la langue et qu'il convient d'expliquer le lien entre langue et discours : il faut mettre alors en œuvre une linguistique variationniste qui prenne en compte les différences d'usages possibles. Par rapport aux approches pour lesquelles le corpus n'est pas indispensable et qui ont une orientation plutôt psychologique, mais aussi par rapport à celles pour qui, le corpus étant représentatif de la langue, la variation est peu prise en compte, les approches qui prennent le corpus comme objet d'étude ont une dimension sociologique importante. Dans tous les cas, les corpus étudiés sont situés socialement et/ou historiquement ; dès qu'il y a texte en effet, sa situation de production ne peut être ignorée et à un moment ou l'autre de l'étude elle est nécessairement prise en compte.

J'aborde ici des points de vue dont je me sens proche ; aussi j'évoquerai quelques courants qui prennent le ou les textes comme points de départ en soulignant ce qui me semble intéressant et ce qui me semble problématique : sociolinguistique ; langues spécialisées et terminologie ; sous-langage et TAL ; analyse de discours, linguistique textuelle et sémantique interprétative.

1.3.2.1 Sociolinguistique

A la suite de Malinowsli et de Firth – « the main point is that this study of meaning was a study of change » (Firth, 1957, 7) –, la sociolinguistique telle qu'elle a été établie par Labov est une linguistique de la variation. Pour Labov, il ne s'agit pas d'un choix méthodologique mais de la conviction que la linguistique ne peut être que sociolinguistique :

> « Pour Labov, la sociolinguistique n'est pas une des branches de la linguistique, et pas davantage une discipline interdisciplinaire : c'est d'abord la linguistique, toute la linguistique mais la linguistique remise sur ses pieds. » (Encrevé, 1976, 9).

Cette variation est nécessairement examinée à travers l'analyse de productions réelles, la sociolinguistique est ainsi clairement une linguistique de corpus. Enfin, cette variation est examinée dans ses relations avec des variations sociales. Il s'agit de prendre en compte la notion d'hétérogénéité des fonctionnements tout en essayant de repérer la systématicité qui se met en œuvre dans cette hétérogénéité :

> « Ces données ordinaires se présentent selon une forte hétérogénéité, généralement considérée comme aléatoire, mais au milieu de laquelle Labov va chercher à établir une systématicité. […]. Il s'avère que l'endroit où se lit une structuration de l'hétérogénéité, ce n'est pas le locuteur comme individu mais la communauté dans son ensemble, approchable par une prédictibilité statistique : il y a une stratification de l'usage de la langue dans la société, dont il a pu établir qu'elle était à la fois régulière et extrêmement fine ». (Gadet, 1992, 6).

Cette façon de voir a de quoi séduire parce qu'elle revendique de tenir compte de l'usage réel de la langue et d'en faire le point de départ de l'étude du système. Prendre en compte la variation et trouver des régularités dans cette variation : voilà un programme qui semble

réconcilier approche empirique et approche hypothético-déductive. Il me semble toutefois que pour aussi intéressante qu'elle soit, l'approche sociolinguistique (en tout cas celle qui se situe dans une perspective explicitement labovienne) n'aborde pas toutes les questions que pose la linguistique de corpus. Trois points en particulier sont problématiques.

Le rapport entre discours et système

L'étude de la variation étant au cœur même de la problématique, en principe l'étude des discours doit permettre de décrire le système linguistique. En réalité, les rapports entre discours et système ne sont pas aussi clairs que cela. Conein montre par exemple, qu'à propos du vernaculaire, Labov hésite entre deux positions :

> « – une définition interactionniste, où il est présenté comme un sous-système à l'intérieur d'un système linguistique […] – une définition communautaire, où il est un système qui s'apparente à un dialecte de l'anglais […] ». (Conein, 1992, 107).

On retrouvera ce même type de problème dans les prochains paragraphes, notamment avec les sous-langages ou les langues spécialisées : jusqu'à quel point s'agit-il d'un usage d'un système plus général et jusqu'à quel point ces systèmes sont-ils autonomes ? On a pu ainsi dire, dans une vision harissienne, qu'il y avait moins de différences entre des sous-langages de domaine dans deux langues différentes qu'entre les sous-langages de deux domaines dans la même langue.

Le problème du sens

La question du fonctionnement sémantique est à la fois fondamental en sociolinguistique et curieusement très peu évoqué en tant que tel. Toute la difficulté vient de ce que le sens semble souvent être l'élément qui permet de mesurer la variation. Ainsi, pour mesurer qu'il y a deux variantes phonologiques, il faut décider qu'il s'agit de deux formes qui renvoient à une même « chose » :

> « Le cadre labovien ne s'était pas embarrassé de nuances pour définir la variable syntaxique. Il se contentait de généraliser ce qui avait donné de bons résultats en phonologie et en morphologie : deux formes qui disent la même chose peuvent être considérées comme des variantes. » (Deulofeu, 1992, 69).

Dans une vision qui découpe aussi nettement les niveaux phonologique, morphologique et syntaxique, il ne reste plus que l'élément sémantique pour être constant. Ainsi, pour décider que deux structures syntaxiques sont des variantes d'un même élément, il faut pouvoir décider que cet élément est constant dans les deux formes ; c'est la rémanence sémantique qui permet au système de se maintenir. Mais on sait bien aussi que tout changement de forme s'accompagne (peut s'accompagner ?) d'un changement de sens. Comment décider que les structures syntaxiques utilisées par des populations différentes renvoient au même sens ? Mais s'il n'y a pas un sens commun, alors il n'y a pas de variantes, il y a des systèmes différents.

Le problème des classes sociales

On considère généralement que l'objectif de la sociolinguistique est de faire apparaître une co-variance entre des éléments sociaux et des éléments linguistiques. Concrètement, cela amène souvent (ne serait-ce que pour constituer les données) à définir des classes sociales considérées comme linguistiquement homogènes, entre lesquelles on va étudier des variantes éventuelles. Si bien que très souvent, les classes sont établies sur des bases idéologiques et la notion de groupe social est extrêmement prégnante : un individu est défini par son appartenance à un groupe social.

> « L'effet pour le locuteur de l'existence d'une variation structurée dans la communauté, c'est la possibilité que celle-ci devienne emblématique d'une identité, permette de se reconnaître et de reconnaître les autres comme ayant une certaine identité sociale. » (Gadet, 1992, 7).

Les variations linguistiques servent ainsi souvent à conforter l'existence de classes sociales et, corrélativement, chaque variation linguistique doit pouvoir être associée à un élément extra-linguistique social. La communauté linguistique se construit ainsi sur la base de la « densité des rapports de communication » (*ibid.*) ; dans les faits, cette communauté linguistique s'élabore bien souvent sur l'identification de groupes sociaux prédéfinis. Le point de vue est ainsi socialement orienté. Il n'est pas envisagé que de nouveaux groupes, linguistiquement homogènes apparaissent lors de l'étude. A ma connaissance, il n'a jamais été envisagé de corréler des variations linguistiques avec des communautés qui ne soient pas socialement préétablies. Par exemple, on n'a jamais étudié des variations dans une communauté professionnelle donnée entre un sous-groupe qui entretient des relations amicales très suivies en dehors du travail et un autre qui n'interragit que dans le cadre professionnel.
Malgré ces réserves, la sociolinguistique a pour mérite d'avoir pris en compte l'hétérogénéité de la langue et, peut-être plus encore, d'avoir mis en avant non seulement le rôle des données textuelles réelles dans l'analyse linguistique mais aussi, et par voie de conséquence, le rôle du linguiste dans la société :

> « J'ai tiré réconfort de la sereine conviction qu'avait Weinrich de ce que nous marchions dans la direction qu'une linguistique rationnelle et réaliste devait inévitablement emprunter. » (Labov, 1976, 39).

Boutet (Boutet, 1992) présente ainsi deux exemples de problèmes sociaux dans lesquels Labov et son équipe furent amenés à se prononcer. L'une a concerné l'évaluation de la lisibilité et de l'objectivité d'un document officiel envoyé à 600 ouvriers noirs engagés dans une action en justice pour discrimination (*ibid.*, 91). L'autre a concerné l'évaluation de la similitude entre deux enregistrements, l'un anonyme, annonçant une alerte à la bombe, l'autre étant l'enregistrement du présumé coupable (*ibid.*, 91).
Je reviendrai sur cette dimension du rôle social du linguiste.

1.3.2.2 Langues spécialisées et terminologie textuelle

Si j'accorde une place à l'utilisation des corpus en terminologie, c'est bien plus parce qu'il semble encore nécessaire de donner des arguments en faveur d'une terminologie textuelle que parce que les corpus sont considérés comme la base de la terminologie. En effet, il est tout à fait récent et encore peu répandu que la prise en compte des corpus en terminologie se fasse de manière systématique.
Je reviendrai plus longuement sur cette question dans le prochain chapitre. Il est cependant intéressant de noter que deux types de travaux se sont mis en place en parallèle sans parvenir à établir des liens très fructueux, alors qu'ils ont eu, en tout cas à certains moments, un même type de matériau (les textes spécialisés) et alors que leurs auteurs ont pu fréquenter les mêmes congrès ! Il s'agit des travaux qui ont concerné la terminologie d'une part et les langues spécialisées d'autre part.
Comme je le disais, les chercheurs qui ont considéré les textes comme étant leur matériau privilégié pour l'étude de la terminologie sont rares, à l'exception notable de Meyer à Ottawa, Pearson à Dublin et, en France les études socioterminologiques à Rouen (Gaudin, 1993) et les travaux du groupe TIA (Terminologie et Intelligence Artificielle) (Bourigault et Slodzian, 1999). L'histoire même de la terminologie a fait que non seulement les textes n'ont pas été utilisés mais qu'on a même pensé qu'ils pouvaient mettre en péril la discipline terminologique. Comme on peut le voir dans l'historique présenté par Foucault, dès le XIXè

siècle, on a voulu accorder un statut spécial au langage scientifique, comme pour le maintenir à l'écart des pollutions langagières :

> « Deux soucis ont été constants au XIXè siècle. L'un consiste à vouloir neutraliser et comme polir le langage scientifique, au point que, désarmé de toute singularité propre, purifié de ses accidents et de ses impropriétés – comme s'ils n'appartenaient point à son essence –, il puisse devenir le reflet exact, le double méticuleux, le miroir sans buée d'une connaissance qui, elle, n'est pas verbale. » (Foucault, 1966, 309).

S'il est en effet un domaine où on a voulu contrôler par le langage, c'est bien la terminologie ; au point qu'il a paru préférable de figer le fonctionnement des termes en en faisant des quasi étiquettes plutôt que de risquer d'en perdre le contrôle en acceptant d'examiner leur fonctionnement dans les textes. Là où le linguiste le plus référentialiste se refuse à se poser en normalisateur, pour beaucoup de terminologues, la normalisation est leur raison d'être. Ils pensent que c'est la seule façon d'améliorer la communication, en expurgeant la langue des scories qui en font un outil d'échange imparfait.
Les chercheurs qui s'intéressent aux langues spécialisées, en revanche, s'inscrivent, dans une démarche linguistique (Cabré, 1998), (Kocourek, 1991), (Lerat, 1995), (Sager, 1990). Mais ils ont bien des difficultés à définir ce qu'est une langue spécialisée. Ces difficultés sont de deux ordres. D'une part, il faut parvenir à définir ce que signifie « spécialisée », d'autre part, il faut situer la réflexion par rapport aux travaux sur la langue. Souvent, la langue spécialisée est considérée comme l'utilisation d'une langue dans un domaine particulier :

> « Une langue spécialisée est une langue naturelle considérée en tant que vecteur de connaissances particulières ». (Lerat, 1995, 20).

Mais, dans d'autres cas, on cherche à mettre au jour le système propre à la langue spécialisée :

> « La langue de spécialité, comme la langue tout entière au reste, ce sont d'abord les textes parlés et écrits […]. C'est principalement sur la base de ces textes que l'on cherche à saisir le système de la langue de spécialité » (Kocourek, 1982, 20).

On ne sait pas s'il y a un système par langue spécialisée ou bien un système unique, de la langue de spécialité. Mais, dans tous les cas, on ne sait pas comment on délimite cette ou ces langues de spécialité (Condamines, 1997). Qu'est-ce qui constitue un domaine de connaissance spécialisé ? A côté de découpages qui paraissent assez consensuels, il peut y avoir des points de vue qui organisent la connaissance de manières très différentes. Et, de fait, ces chercheurs n'ont pas beaucoup travaillé sur la question de la constitution du corpus.
En revanche, les linguistes des langues spécialisées ont parfaitement compris l'intérêt de prendre en compte les textes pour étudier la langue. En effet, leurs études passent nécessairement par l'utilisation de données réelles puisqu'ils n'ont pas de compétence sur le domaine étudié. Ils ont ainsi acquis une expérience d'analyse de textes qui leur a permis d'avancer sur l'étude des corpus en sémantique. Les travaux actuels en terminologie textuelle ont certainement bénéficié de ces avancées.

1.3.2.3 TAL et sous-langages

Le traitement des corpus en TAL amène nécessairement un point de vue particulier. En effet, les outils ne peuvent traiter que des formes et toute la problématique du TAL vient de la nécessité d'associer un contenu (un sens) à ces formes. Le sens est ainsi à la fois inaccessible par les outils (seul, un humain peut donner du sens) et sans cesse recherché par eux.
Il peut sembler étonnant de présenter la théorie des sous-langages et la problématique du TAL dans le même paragraphe. Ce choix n'est pas seulement lié à la volonté de ne pas faire une présentation trop longue. En effet, d'une part, les informaticiens qui travaillent en TAL, à

partir de données textuelles réelles, se réclament souvent de l'approche distributionnelle[7]. D'autre part, ce mode de présentation permet de s'interroger sur le traitement du sens en informatique. Premier constat : la théorie harissienne est remise en question par les sémanticiens et même les analystes de discours de tous bords : (Adam, 1999), (Benveniste, 1966), (Lyons, 1980), (Rastier, 1991), (Vandeloise, 1991). Plusieurs raisons peuvent être avancées pour expliquer cette défiance. Tout d'abord, la vision de Harris est nettement béhavioriste :

> « It is empirically discoverable that in all languages which have been described we can find some part of one utterance which will be similar to a part of some utterance. « Similar » here means not physically identical but substituable without obtaining a change in response from native speakers who hear the utterance before and after the substitution… In accepting this criterion of hearer's response, we approach the reliance on « meaning » usually required by linguists » (Harris, 1966, 20).

Ensuite, et corrélativement, l'ambition de Harris consiste à comprendre comment des agencements de formes (les distributions de mots) peuvent créer du sens. En principe, il s'agit donc d'évacuer le sens de l'analyse puisqu'il doit émerger de la mise au jour de régularités de formes. Ainsi, au moins dans la vision initiale de Harris, il n'y a pas d'interprétation des résultats car les dépendances relationnelles mises au jour sont le reflet de la perception humaine :

> « …the dependence relation of words reflects what one might consider dependencies within man's perceivable world…word meanings and co-occurrence selections express a categorizing of the world. » (Harris, 1991, 347) cité par (Habert et Zweigenbaum, 2002a).

On comprend qu'une telle vision irrite quelque peu les sémanticiens… d'autant que la méthode décrite par Harris ne me semble pas correspondre à son application qui, elle, fait nécessairement intervenir une interprétation. C'est en effet sur la base d'une interprétation, d'un point de vue particulier, que l'on peut décider de donner (ou pas) un sens à une régularité distributionnelle. C'est aussi sur la base d'une interprétation que l'on peut décider qu'une phrase à l'actif et une autre au passif ont le même sens ! Dans les travaux d'auteurs qui mettent en œuvre la méthode harrissienne, on trouve souvent cette dimension interprétative, au détour d'un paragraphe :

> « What is needed to determine whether an underlying grammatical relation holds between words is a set of word classes in terms of which we can distinguish those occurrences that " make sense " in the domain in question (here clinical narrative) from those that do not. » (Sager, 1987, 15)

Cette dimension interprétative est maintenant revendiquée par beaucoup d'informaticiens :

> « Unsupervised learning of word classes according to distributional similarities provides clusters of words that require *a posteriori* human 'pruning', interpretation and labelling. » (Habert et Zweigenbaum, 2002a).

Mais dans le même temps, la pression reste forte sur les informaticiens pour créer du sens à partir de la forme et pour beaucoup, l'ambition demeure d'acquérir des connaissances sémantiques par simple repérage de régularités de formes.
Sur un autre plan, comme le montrent Habert et Zweingenbaum, les « meilleurs résultats » sont obtenus par la mise en œuvre d'une connaissance linguistique *a priori* (qu'elle soit d'ordre syntaxique ou sémantique), qui guide l'interprétation, ce qui semble contradictoire avec le fait de faire émerger du sens grâce à la mise au jour de régularités. Cette connaissance *a priori* n'est évidemment pas neutre et il serait peut-être intéressant d'évaluer, d'une part comment elle s'accorde ou non avec les régularités mises au jour et d'autre part, le degré de

[7] En tout cas, pour ce qui concerne les travaux qui s'intéressent au sens, en particulier dans les corpus spécialisés. Je ne parle pas dans cette partie des recherches menées sur l'analyse syntaxique.

dépendance entre cette connaissance et les résultats que l'on veut obtenir, c'est-à-dire avec le type d'interprétation que l'on souhaite faire des résultats. Dans cet objectif, sémanticiens et informaticiens ont intérêt à travailler main dans la main.
Malgré les réserves que l'on peut avoir envers la théorie harissienne (plutôt qu'envers sa mise en œuvre), le TAL basé sur l'approche distributionnelle est d'un apport majeur pour l'analyse de corpus : par la mise au jour de régularités imperceptibles « à l'œil nu », il ouvre des perspectives nouvelles, voir par exemple (Bourigault et Fabre, 2000). Il permet un point de vue que la sémantique n'avait encore jamais eu sur le matériau langagier et renouvelle de manière radicale les questions sur les rapports entre sens et forme. Il peut aider à systématiser l'approche distributionnelle qui, peu ou prou, est toujours celle qui est mise en œuvre lors de l'analyse de corpus. Il serait dommage que les sémanticiens de corpus se privent des outils proposés par le TAL. Ainsi, bien que, comme le note Blache :

> « on constate que ce terme même de linguistique de corpus n'est quasiment utilisé que par la communauté linguistique-informatique » (Blache, 2000, 83),

les linguistes ont tout intérêt à tenir compte, dans leur réflexion, de la possibilité d'utiliser des corpus électroniques et des outils d'analyse. Mais les résultats de ces outils doivent être examinés dans une perspective interprétative. Ces résultats sont obtenus grâce à des outils qui mettent en œuvre un point de vue sémantique et syntaxique, à la fois parce qu'ils utilisent des « connaissances linguistiques » et parce qu'ils recherchent des régularités dont on suppose qu'elles peuvent avoir un sens (pendant que d'autres types de régularités ne sont pas examinées) :

> « Un jeu d'étiquettes permet […] d'étudier certains phénomènes ou de développer des traitements ultérieurs déterminés, tandis qu'il laisse d'autres aspects linguistiques dans l'ombre et se révèle incompatible avec d'autres applications. » (Habert, 2000a, 107).

- Il est donc indispensable que les linguistes qui utilisent ces outils soient parfaitement conscients du point de vue qui a présidé à leur construction ; les utilisateurs peuvent ainsi intégrer (ou pas, en fonction de leurs objectifs) les résultats de ces outils dans leur réflexion ou pratique.
- Dans une perspective sémantique, les résultats obtenus constituent le point de départ de la réflexion ou de la construction d'un modèle ; ils ne sont pas un aboutissement. Je reviendrai sur cette question mais il est important de noter que, s'il s'agit par exemple de construire une base de connaissances terminologiques, le travail à effectuer à partir des résultats d'outils, par interprétation de ces résultats, reste considérable. Mais sans outils, le même travail ne serait même pas envisagé !

Il reste toutefois un problème majeur, récurrent quand on travaille à partir de corpus, qui est celui de la constitution du corpus. Comme le montre Pearson, (Pearson, 1998, 28-35), les propositions de Harris ne résolvent pas cette question. En effet, on retrouve toujours le même paradoxe, qui consiste à sélectionner des textes parce qu'on suppose qu'ils contiennent les mêmes régularités syntaxiques alors même que ce que l'on souhaite, c'est faire apparaître des régularités syntaxiques dans ces textes.
Hormis la définition de besoins en matière d'outils ou d'utilisation d'outils, la collaboration avec des informaticiens (en TAL ou en Ingénierie des Connaissances, dans les groupes TIA (Terminologie et Intelligence Artificielle) ou ASSTICCOT (Action Spécifique STIC Terminologie et COrpus), a eu pour moi un autre intérêt, majeur. Il concerne la prise en compte d'une demande extérieure. La plupart des informaticiens, particulièrement les ingénieurs de la connaissances mais aussi beaucoup de ceux qui s'intéressent au TAL en basant leur réflexion sur des corpus, s'inscrivent dans la perspective de répondre à un besoin réel. Cela leur pose beaucoup moins de problèmes qu'aux linguistes dans la même situation, qui se sentent obligés de justifier ces collaborations ; tout au plus, les chercheurs en

informatique s'interrogent-ils sur la différence entre leur approche et celle d'ingénieurs. Cette position, décomplexée par rapport à des besoins réels, m'a amenée à réfléchir d'une part au rôle social du linguiste et, d'autre part, au sens que pouvait avoir la prise en compte d'un besoin particulier dans l'analyse sémantique. Tout comme Léglise (Léglise, 2000), je préfère, au terme de linguistique appliquée, celui de linguistique impliquée qui me semble mieux rendre compte de la situation du linguiste lorsqu'il intervient dans un contexte professionnel. Mais je montrerai aussi que cette notion d'implication ne peut pas être totalement absente d'une analyse de corpus. Comme le souligne Boutet :

> « L'introduction des discours au travail comme objet d'étude des linguistes conduit donc à une réflexion méthodologique sur le statut des matériaux à analyser, sur les méthodes d'enquête, sur la place du chercheur. » (Boutet, 1995b, 248).

Je montrerai dans la partie 2-2 pourquoi je crois que cette situation d'analyse sur le terrain n'est pas fondamentalement différente d'une analyse théorique.

1.3.2.4 Linguistique textuelle, analyse de discours, sémantique interprétative

Si je présente dans la même rubrique linguistique textuelle et analyse de discours, c'est parce qu'elles partent toutes deux d'un postulat fort : l'unité d'analyse est le texte ; comme le rappelle Adam, citant Meyer :

> « Le texte est un tout, et non un simple assemblage de propositions indépendantes (et analysables comme telles) que l'on aurait mises bout à bout. En fait, le sens d'un texte se détermine par ses composantes mais ne s'y ramène pas : chaque phrase du texte renvoie à ce dernier comme à son sens profond. (Meyer1986 :238) » (Adam, 1999, 26).

Même si, comme le fait Adam (Adam, 1999), on considère que la linguistique textuelle est « un sous-domaine de l'analyse de discours », le contexte est toujours pris en compte, soit parce que, comme dans la linguistique textuelle, il a permis de délimiter la cohérence du texte à analyser, soit parce que, comme dans l'analyse de discours, il interagit en permanence avec les éléments textuels pour aider à construire un sens :

> « L'analyse de discours n'étudie pas de manière immanente les énoncés pour ensuite les rapporter à divers paramètres " extérieurs ", situationnels : elle s'efforce au contraire d'appréhender le discours comme une activité inséparable de ce « contexte » » (Maingueneau, 1996, 22).

Les auteurs s'accordent pour donner à l'analyse de discours (en tout cas dans l'analyse de discours « à la française ») des origines extrêmement diverses : psychanalyse, marxisme, philosophie, sciences humaines et sociales mais aussi herméneutique et philologie (Charaudeau et Maingueneau, 2002, 7).

L'école française d'analyse du discours, qui a été largement inspirée par les propositions de Michel Pécheux, s'est d'abord appuyée sur une analyse de type harrissien (dans la perspective d'une analyse automatique) et a d'abord porté sur des textes politiques. Aujourd'hui, les différents courants qui, dans les sciences du langage, se situent dans une optique discursive, entretiennent des parentés serrées, rarement identifiées, dans lesquelles on retrouve les influences de la sociolinguistique, de la théorie des sous-langages, des réflexions sur « langue et travail », du TAL… Et il est parfois bien difficile de savoir quelles problématiques sous-tendent les analyses.

Une constante de l'analyse de discours me semble être que le sens n'est pas un donné immédiat mais qu'il doit être soit mis au jour, soit construit. Comme le dit Maingueneau,

> « L'analyste du discours vient ainsi apporter sa contribution aux herméneutiques contemporaines. » (Maingueneau, 1987, 6).

Cette perspective herméneutique rejoint les préoccupations de la sémantique interprétative :

« Au lieu de partir d'une ontologie préfixée, dont le texte ne serait qu'une manifestation toujours partielle et imparfaite, [la conception rhétorico/herméneutique] cherche à faire émerger corrélativement des régularités et des singularités, et à leur faire correspondre, par construction interprétative, des fonds et des formes sémantiques. » (Rastier, 2001, 90).

S'il y a interprétation, c'est donc parce que le sens n'est pas un donné du texte ; cet élément s'oppose à une vision référentielle et même à une vision strictement lexicale. L'unité à étudier est le texte et il doit être pris comme élément en lien avec un contexte. Les questions majeures deviennent alors : peut-on avoir une approche systématique de l'interprétation d'un texte (cette question a-t-elle même un sens) ? Comment la notion de contexte peut ou non cadrer cette interprétation?

« S'il est souvent invoqué au lieu d'être défini, le contexte a un effet de problématisation […] Il témoigne d'une reconnaissance locale et partielle du problème de l'interprétation ». (Rastier, 1998, 97)

La dimension herméneutique peut faire peur tant elle peut ouvrir de possibilités : dans leurs diversités, ces possibilités sont très éloignées de la vision d'une approche scientifique censée mettre au jour la vérité dans une sorte de logicisation figée[8].
Aussi la réflexion sur l'interprétation doit se garder de deux écueils majeurs, parfaitement énoncés par Pécheux :

« L'analyste de discours ne prétend pas s'instituer en spécialiste de l'interprétation maîtrisant « le » sens des textes […]. L'enjeu crucial est de construire des interprétations sans jamais les neutraliser ni dans le " n'importe quoi " d'un discours sur le discours, ni dans un espace logique stabilisé à prétention universelle. (Pécheux, 1984) » (Maingueneau, 1987, 6).

Ce problème de l'interprétation est celui qui traverse toutes les questions qui se posent en analyse de discours et en sémantique textuelle. La deuxième partie de ce chapitre me permettra de préciser ma position sur ce sujet. Pour clore cette partie sur les analyses faisant appel à un corpus réel, il devient nécessaire que je m'interroge sur ce qu'est un corpus.

1.3.3 Qu'est-ce qu'un corpus ?

J'ai utilisé jusqu'à présent la notion de corpus sans la définir, en partie parce qu'elle n'est pas stable d'un point de vue à l'autre. Au terme de ce panorama sur les approches qui s'ancrent dans un corpus, il devient nécessaire de préciser ce que j'entends par corpus et, en particulier, comment je situe cette notion par rapport à celle de texte.
Pearson relève plusieurs définitions (Pearson, 1998, 42) qui ont été proposées par des lexicologues ou des grammairiens ; j'en retiendrai deux :

- celle de Sinclair (que l'on retrouve fréquemment citée) :

« a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language » (Sinclair, 1994, 2) (cité par Pearson, 1998) ;

[8] Il est intéressant de noter qu'à deux époques différentes, Milner et Rastier ont mis en parallèle (en opposition ?) « art » et « science » à propos de l'approche linguistique, Rastier, dans le titre même de son dernier livre (« Arts et science du texte »), Milner, dans le passage suivant qui est l'introduction d'un cours qu'il a donné en 1974 :
« Entre l'art et la science, la limite tient en un axiome que dénie le premier et dont la seconde se soutient : le réel de la langue est de l'ordre du calculable. Mais à l'axiome même, on ne parvient pas sans détours : il y faut : I) constituer la langue comme un réel …II) Constituer la langue comme un réel représentable par un calcul…III) Ne retenir de l'être parlant en général que ce qui le fait support d'un calculable, le penser comme un point sans division ni étendue, sans passé ni avenir, sans conscience et sans inconscient, sans corpus – et sans autre désir que celui d'énoncer… IV) Ne retenir de la multiplicité des êtres parlants que ce qui est nécessaire à constituer un réel calculable comme langue …. » (Milner, 1978, 7).
Les temps ont changé et le recours à des données réelles (les corpus) atténue le clivage énoncé par Milner. Sous l'influence de la sémantique de corpus, la linguistique est en passe de devenir une science moins rigide qu'elle n'était, sans doute colorée d'une dimension artistique (en tout cas artisanale) …

– celle de Atkins, Clear et Ostler :

« a subset of [Electronic Text Library] built according to explicit design criteria for a specific purpose […] » (Atkins et *al.*,1992,1) (*ibid*).

Le commentaire des principaux éléments de ces deux définitions va m'amener à définir ma position. Ces éléments sont au nombre de 5 : la taille, le support électronique, la notion d'échantillon de la langue, la notion de textes, la notion de visée spécifique.

– Taille : lorsqu'il s'agit de corpus qui sont censés être représentatifs d'une langue, il est entendu que l'on parle de corpus d'un volume important : au minimum 100 000 mots mais parfois plusieurs millions (100 millions pour le BNC). Comme le montre Pearson (*ibid*, 56-57), plusieurs auteurs pensent que la taille du corpus peut être liée à l'objectif poursuivi et que, dans certains cas, des corpus beaucoup plus petits peuvent être suffisants. Par ailleurs, comme le dit Pearson toujours, il est de nombreux cas où les données disponibles sont rares et ne peuvent constituer un corpus très volumineux ; c'est moins alors par choix que par nécessité que le corpus étudié est petit. J'ajouterai que le corpus à étudier n'est parfois pas le produit d'une construction par un linguiste mais est fourni par une entreprise et est constitué sur des bases qui lui sont propres et qui constitue une réelle cohérence. Souvent ces corpus ne sont pas très volumineux.
– Corpus électronique : c'est sous l'influence du traitement automatique des langues qu'on a pris l'habitude de ne concevoir de corpus qu'électronique. Mais la tradition d'analyse de textes (y compris dans un but lexicographique) existe depuis bien longtemps, bien avant en tout cas que le TAL se développe et de nombreux travaux réalisés sur des corpus non-électroniques conservent une valeur réelle. Par ailleurs, tout comme pour la taille, il n'est pas rare que la dimension électronique ne dépende pas du choix de l'analyste ; les corpus disponibles n'existent parfois que sous format papier et, sauf à les scanner, ce qui est long et pas toujours très efficace, il peut arriver que l'on décide de les conserver sous cette forme, surtout si les textes sont courts et que l'étude à réaliser ne justifie pas de passer du temps à scanner ces textes. Imaginons par exemple un texte de 10 pages dans lequel on veut étudier tel ou tel terme qui apparaît une vingtaine de fois ; il est facile de repérer les occurrences de ce terme manuellement et de travailler à leur distribution manuellement. De mon point de vue, ce type d'approche relève tout autant d'une linguistique de corpus qu'une approche qui se met en place sur des corpus très volumineux et sous format électronique. Pour autant, ce point de vue ne remet pas en cause l'apport du TAL, en particulier sur des corpus volumineux.
– Portions de textes. La définition de Sinclair rend compte d'un corpus comme étant composé d'extraits de textes. Cette position, qui prend son sens dans la perspective de constituer des échantillons de langue ou dans celle d'entraîner des modèles n-grammes, est fortement remise en question par les tenants de la linguistique textuelle. En effet, elle suppose une vision d'un texte comme étant un amas de données langagières dans lequel on peut prélever ce qui semble intéressant. Outre qu'il est bien difficile de décider ce qui est intéressant dans la perspective de représenter la langue, ce point de vue équivaut à faire éclater ce qui constitue une unité d'analyse qui a sa propre cohérence : le texte. Enfin, l'ordre dans un texte n'est pas aléatoire ; la linéarité de la lecture, qui permet de construire un sens, est complètement éclatée si on ne conserve que des extraits de textes.
– Sélection sur des critères linguistiques. Toujours dans la définition de Sinclair, le corpus est présenté comme étant constitué sur la base de critères linguistiques. Or, d'une part, ce point ne me semble pas exact et, d'autre part, il ne me semble pas toujours souhaitable. Pas exact car, pour la sociolinguistique par exemple, dont le

matériau d'étude est toujours un corpus, les critères de constitution du corpus sont d'abord sociaux (*cf.* ci-dessus). Pas souhaitable car on retrouve toujours ce même paradoxe qui consiste à partir d'une intuition linguistique pour constituer un corpus afin de mettre en évidence des régularités. Mais une fois le corpus constitué sur des « critères linguistiques » intuitifs, les textes non retenus sont définitivement exclus de l'étude et donc impossibles à rapprocher d'autres textes, sur la base de régularités linguistiques. Ce biais-là est probablement irrémédiable mais pour autant, il n'est pas souhaitable de le poser comme méthode d'élaboration de corpus.

– Objectif spécifique. La définition de Atkins et *al.* mentionne un objectif spécifique qui justifie la constitution du corpus. Pearson pense que cette notion n'est peut-être pas essentielle (Pearson, 1998, 42) parce que dans la plupart des cas, les informaticiens qui constituent un corpus ne savent pas comment il sera utilisé ; pourtant, elle-même reconnaît qu'elle utilise des « special purpose corpora » et que :

« We plan to use this term whenever the specific purpose for which the corpus has to be used (e.g. retrieval of definition statements, analysis of gender-related issues) is the reason for creating or selecting the corpus » (*ibid*, 48).

Mon avis est que ce cas de figure ne devrait pas être un cas à part de constitution de corpus mais bien le cas régulier. Je ne crois pas en effet que l'on puisse constituer un corpus sans avoir une idée précise de ce que l'on veut en faire (hypothèse linguistique ou application, j'y reviendrai). Même s'il s'agit de construire un corpus de « référence » (si l'on accepte que cela ait un sens), il est capital d'avoir mené une réflexion très poussée sur l'objectif (les objectifs) de cette constitution. En particulier parce que c'est la seule façon de caractériser les textes du corpus sur des bases non-linguistiques qui permettront ultérieurement des recoupements de genres. La constitution du corpus n'est pas indépendante de l'étude que l'on va mener et la réflexion sur les critères à mettre en œuvre doit être très élaborée. Dans cet ordre d'idée, un texte peut devenir un corpus, si on lui a donné le statut d'objet dans le cadre d'une étude particulière.

Malgré les réserves que j'ai énoncées par rapport aux définitions données par les lexicologues de corpus, il est important de retenir qu'un corpus est toujours le fruit d'une constitution. En tenant compte des éléments que j'ai développés ci-dessus, je propose la définition de corpus suivante :
« Collection de textes (éventuellement un seul texte) constitué à partir de critères linguistiques ou extra-linguistiques pour évaluer une hypothèse linguistique ou répondre à un besoin applicatif».

Pour conclure cette longue partie sur la présentation des modes de prise en compte des corpus, un dernier paragraphe va synthétiser l'ensemble des questions posées à la sémantique lorsqu'elle prend les corpus comme point de départ.

1.4. La langue est-elle un objet d'étude autonome ?

Une des questions récurrentes dans la réflexion en sémantique porte sur la possibilité d'une autonomie de la langue. La principale ambition du structuralisme a certainement été de créer une discipline parfaitement identifiée, dont le modèle a d'ailleurs été repris par bon nombre de disciplines des sciences humaines.
Cette autonomie de la langue est souvent perçue comme la seule façon de donner un statut scientifique à la linguistique : la langue, si elle ne se perd pas dans les méandres historico-

socio-psychologiques des autres sciences humaines devient un objet clair qui peut être étudié de manière objective. Dans l'approche générative et l'approche cognitive, un rapprochement est établi avec la psychologie, mais c'est pour mieux confirmer la permanence et la régularité de l'objet langue : parce qu'ils ont le même système perceptif et cognitif, les humains possèdent forcément des schémas langagiers réguliers et constants. L'approche sociolinguistique, quant à elle, donne une grande importance aux groupes de locuteurs mais ces groupes sont eux-mêmes assez figés et surtout, ils préexistent à l'analyse linguistique qui ne vient souvent que confirmer leur existence. Dans les deux cas, il me semble que l'on oublie que la langue s'inscrit aussi dans une histoire individuelle et une histoire collective. Cette histoire individuelle a un impact d'une part sur l'apprentissage de la langue (et donc sur la compétence de locuteur qui relève à la fois de l'expérience individuelle et de l'appartenance à une collectivité) et d'autre part sur les interactions d'un individu avec d'autres individus. C'est la rencontre entre histoire individuelle et histoire collective, qui ont également une dimension psychologique et une dimension sociale, qui crée le dynamisme linguistique, en particulier sémantique. Il se peut que ce dynamisme ne puisse être figé dans des règles définies et encore moins prédit et pourtant, là est peut-être l'essence même du sens. De ce point de vue-là, la linguistique ne peut être dissociée de la sociologie, de la psychologie, de l'ergonomie, de l'histoire… elle n'a peut-être même aucune raison d'être…
Et pourtant elle existe ! Elle existe parce que, qu'on le veuille ou non, la notion de régularité est tout à fait pertinente : dans tout texte, dans toute production langagière, il y a à la fois mise en œuvre de règles conscientes, celles du système de la langue utilisée, et mise en œuvre de règles moins conscientes voire inconscientes : celles d'autres systèmes sociaux ou psycho-sociaux ou même interindividuels. En tout état de cause, la linguistique n'existe que parce que le sens ne peut être le fait d'un individu isolé (le linguiste est d'ailleurs lui aussi un acteur éminemment social):

> « Seule donne sens la formation discursive qui englobe des énonciateurs isolables mais non pour autant singuliers, ensemble discursif dialectiquement construit et délimité par d'autres ensemble discursifs » (Béacco et Moirand, 1995, 46).

Un des nouveaux défis de la linguistique, en particulier de la linguistique de corpus, est de mettre au jour les liens entre les régularités qui se manifestent dans un corpus et la situation qui a présidé à la constitution de ce corpus. Evidemment, ce projet s'inscrit dans une perspective pluridisciplinaire, à la fois parce que les textes appartiennent à des genres (psycho-sociaux) et parce que l'interprétation est, elle-même, située (*cf* ci-dessous). Mais une fois le corpus constitué, alors me semble-t-il, l'analyse qui se met en place est essentiellement linguistique, au sens où, cadrée par une étude très fine de la situation, elle peut se concentrer à la mise au jour de régularités, régularités qu'on peut dire immanentes à ce moment-là, même si elles sont à rapporter au contexte d'énonciation et d'interprétation.

> « Méthodologiquement, nous prenons appui sur des régularités repérables de façon immanente et non sur un modèle applicable à tous les textes qui ne seraient que des occurrences de ce modèle universel. Ce principe d'immanence intervient à deux moments : lors du choix et de la délimitation des corpus, parce que les écritures institutionnelles sont des écritures produites dans des cadres historiquement situés qui leur donne sens. […]. De même, lorsque nous passons du matériau collecté aux formes linguistiques retenues pour l'analyse, nous sélectionnons des régularités immanentes aux texes. » (Branca-Rosoff et *al.*, 1995, 58).

Je crois même tout à fait légitime, une fois le corpus constitué, de mettre en place une analyse structurale et différentielle « à la Saussure ». D'une certaine façon, lorsque le cadre de l'interprétation est clairement établi, ce qui demande une analyse qui peut être longue et qui peut faire intervenir différents partenaires, on peut penser que l'interprétation fait place à la

description, à entendre ici comme la volonté d'une interprétation qui s'inscrit dans une perspective uniquement linguistique :

> « J'ai essayé de montrer ici qu'il existe un point de vue de linguiste sur les discours et que la spécificité du linguiste tient à sa capacité à prendre seulement en compte la matérialité même des organisations linguistiques. Non seulement lexicale mais aussi morphologique ou syntaxique. La suspension de l'interprétation au profit de la description minutieuse de la mise en mots caractérise une démarche linguistique ». (Boutet, 1995a, 213).

Outre les analyses que j'ai pu mener (présentées dans la suite du mémoire), qui me semblent être essentiellement linguistiques même si les régularités mises au jour sont en relation avec des éléments extra-linguistiques, mon choix d'une autonomisation possible (voire souhaitable) de l'objet d'étude et, plus encore peut-être, d'une identité propre de la linguistique de corpus, s'appuie sur des expériences que j'ai pu avoir en entreprises, dans lesquelles d'autres chercheurs en sciences humaines, interviennent, par exemple des ergonomes[9]. Souvent amenés à étudier des textes produits dans ces entreprises, ces ergonomes sont parfois bien dépourvus de méthodes et même de recul pour remplir la tâche qu'on leur propose. La syntaxe et la sémantique (y compris dans leur dimension introspective) ont accumulé des connaissances qui, très souvent, ont une pertinence pour (à défaut d'une adéquation avec) les questions posées par l'étude de ces textes dans un contexte de travail. Evidemment, en tant que locuteurs d'une langue, les ergonomes, tout comme les ingénieurs de la connaissance, par exemple, peuvent faire de ces textes des interprétations intéressantes (je reviendrai sur cette question dans le chapitre V). Mais les linguistes me semblent mieux armés pour repérer systématiquement des régularités, pour les mettre en relation avec des éléments extra-linguistiques et pour les situer comme connaissance nouvelle dans une tradition disciplinaire (c'est-à-dire pour thésauriser cette connaissance). Si bien d'ailleurs qu'il me semble pertinent que le travail de laboratoire alterne avec le travail de terrain ; il y a un temps pour comprendre le terrain, ses besoins, ses traditions, ses attentes et puis, une fois intégrée la situation, il y a un temps pour interpréter le corpus constitué, pour en repérer les régularités « immanentes » puis à nouveau un temps d'évaluation des résultats par les acteurs du terrain mais aussi par les pairs linguistes.

On a pu appeler « ergonomie linguistique » la compétence exigée par les besoins d'analyse de textes en entreprise (Rastier et *al.*, 1994), (Péry-Woodley, 1995). A la vérité, je crois qu'il faut, certes, travailler à la mise au jour de complémentarités entre approche ergonomique et approche linguistique mais qu'il faut surtout envisager de mettre en place des formations en sémantique orientées vers une description située des textes : la linguistique de corpus a maintenant acquis une somme de connaissances qui lui est propre. Même si elle n'a pas encore constitué ces connaissances en discipline, il me semble légitime qu'elle revendique son autonomie.

J'ai souvent évoqué la situation (ou le contexte au sens large) comme l'élément fondamental qui permet de justifier les possibilités interprétatives. La deuxième partie de ce chapitre va me permettre d'aborder ce que peut recouvrir cette notion et la manière dont je l'intègre dans ma réflexion.

[9] Il faut d'ailleurs noter que si les ergonomes sont très souvent sollicités pour faire des analyses de textes, c'est souvent parce qu'ils sont déjà sur place et qu'on considère que cela fait partie de leur compétence. La psychologie, en ne rechignant pas à étudier des productions réelles d'entreprises, dès le milieu du XXè siècle, a ainsi pris une longueur d'avance sur la linguistique pour ce qui concerne l'analyse des textes d'entreprises.

2. Pour une sémantique de corpus doublement située

La situation dans laquelle s'est élaboré un texte est à peu près toujours prise en compte par les analystes de discours, au sens large. La situation qui permet d'interpréter ce texte est beaucoup moins fréquemment évoquée. Or, elle joue un rôle tout aussi important dans la construction d'un sens. Dans la constitution de terminologies à partir de textes, je crois même que la situation d'interprétation joue un rôle plus important que la situation de production de ces textes, j'essayerai de le montrer. La question qui va traverser ma présentation de ce que j'appelle « sémantique doublement située » (par la situation de production et la situation d'interprétation) est celle de la possibilité (ou non) de stabiliser les variations situationnelles, de les regrouper pour mieux les prédéfinir et les poser comme cadre d'étude aussi systématisé que possible. Pour la situation de production, c'est à travers la notion de genre textuel que la question se pose ; je souhaiterais poser la même question pour ce que j'appelle les « genres interprétatifs ».

2.1. La situation de production des textes : la question du genre textuel

Que le sens d'un texte ne puisse être mis au jour ou construit sans faire intervenir le contexte à un moment ou l'autre de l'analyse est un constat très largement partagé par tous ceux qui ont les discours comme objet d'étude, quelle que soit d'ailleurs la discipline concernée. Mais, alors que pour la sociologie, la psychologie, l'histoire… la nécessité de prendre en compte l'élément situationnel ne pose pas de problème majeur, pour la linguistique, cette dimension s'accompagne d'interrogations sur l'autonomie du langagier, sur les possibilités de systématiser les approches et les résultats, sur le mode d'interaction, ou de fusion entre situation et discours. Avec la prise en compte de la situation, c'est la variation qui fait irruption dans l'étude syntaxico-sémantique et, il faut le reconnaître, une variation qui peut être extrêmement labile et difficile à maîtriser. Comparée à une sémantique cognitive ou référentielle, dont les résultats sont contrôlés par ce que l'on suppose être un réel tangible et relativement immuable, la sémantique textuelle est nécessairement une sémantique de l'instable, d'ailleurs revendiquée comme telle. Dans le même temps, pour essayer de comprendre ce que peut être l'interprétation d'un texte, c'est-à-dire ce que peuvent être ses sens, il faut bien tenter de se donner des cadres fixes d'interprétation, de fonder les interprétations sur des éléments que l'on peut contrôler. Cela signifie en particulier qu'il n'est pas possible d'étudier un texte comme s'il était une production langagière totalement nouvelle, dans une situation totalement nouvelle. D'une part, ce texte, écrit dans une langue donnée, obéit aux règles de cette langue ; d'autre part, en tant que texte socialement situé, il obéit aux règles du genre auquel il appartient. Cette notion de « genre textuel », que l'on rencontre aussi dans pratiquement tous les travaux d'analyse de discours est ainsi un des piliers qui fonde la théorie interprétative en linguistique textuelle : si l'on peut mettre au jour des régularités sémantico-syntaxiques, c'est parce que préexistent des situations sociales codées qui imposent aux locuteurs, souvent à leur insu, des règles linguistiques et qui permettent de regrouper des textes.

> « L'analyse du discours – pour moi analyse des pratiques discursives qui renoncent à traiter comme identiques les discours judiciaire, religieux, politique, publicitaire, journalistique, universitaire, etc. – s'attarde quant à elle prioritairement sur la description des régulations descendantes que les situations d'interaction, les langues et les genres imposent aux composantes de la textualité. » (Adam, 1999, 35).

> « Entre en jeu la notion (ou pour certains, le concept) de genre. Ces désignations, par lesquelles les membres d'une communauté de communication s'accordent à classifier les textes, constituent une source de connaissance des pratiques langagières en usage dans une communauté de communication. Mais, de plus, en tant que représentations qui informent la production langagière, elles fournissent à

la recherche des cadres au moins institutionnels et/ou sémiotiques qui autorisent le rapprochement de textes. » (Béacco et Moirand, 1995, 47).

Les genres textuels

Comme le rappelle Branca-Rosoff, la notion de genre est utilisée depuis longtemps dans une perspective à la fois descriptive et prescriptive : si l'on voulait être reconnu comme un bon « écrivain », il fallait obéir aux règles de la bonne rédaction. Mais les siècles passant, la dimension prescriptive est débordée par la réalité des usages et il devient de plus en plus difficile de contrôler (ou décrire) l'évolution des genres :

« La terminologie des genres, construite en vue de l'acquisition pratique des modèles a été descriptivement adéquate tant que l'institution scolaire a travaillé sur le corpus fermé des textes de la tradition. Mais une fois encore les classements se sont périmés à partir du XVIIè siècle parce que les pédagogues ont figé les catégories et ont exclu du champ littéraire les textes qui ne correspondaient pas à leur grille d'analyse alors que leur importance sociale allait croissant […]. A partir du XIXè siècle, la crise s'accroît car la modernité revendique la déstabilisation des genres… » (Branca-Rosoff, 1999, 18).

A l'époque contemporaine, deux auteurs sont considérés comme les théoriciens du genre, Wittgenstein d'une part, à travers l'hypothèse des jeux de langage et surtout, Bakhtine.

« En fait, les formulations de Bakhtine quant aux " genres de discours " peuvent être tenus pour parents des énoncés de Wittgenstein sur les jeux de langage, notamment en ce qu'ils thématisent l'instabilité mais aussi le lien à la situation d'apprentissage.» (Bouquet, 1998, 118).

« Tout énoncé pris isolément est, bien entendu, individuel, mais chaque sphère d'utilisation de la langue élabore ses types relativement stables d'énoncés, et c'est ce que nous appelons les genres de discours » (Bakhtine, 1984, 265).

« Les genres du discours, comparés aux formes de langue, sont beaucoup plus changeants, souples, mais, pour l'individu parlant, ils n'en ont pas moins une valeur normative : ils lui sont donnés, ce n'est pas lui qui les crée. » (*ibid*., 287).

C'est entendu, il existe, entre le système de la langue et le discours, un palier intermédiaire où s'organisent des régularités langagières auxquelles obéissent les locuteurs. La question qui se pose alors est celle de la possibilité d'établir une liste des genres de discours. A cette question, la plupart des auteurs répondent par la négative :

« La richesse et la variété des genres du discours sont infinies car la variété virtuelle de l'activité humaine est inépuisable et chaque sphère de cette activité comporte un répertoire des genres du discours qui va se différentiant et s'amplifiant à mesure que se développe et se complexifie la sphère donnée. Il faut souligner tout particulièrement l'hétérogénéité des genres du discours (oraux et écrits). » (Bakhtine, 1984, 265).

Différentes raisons sont avancées pour justifier la difficulté à dresser une typologie des genres. Voyons-les.

- La principale raison évoquée a trait à « la variété de l'activité humaine ». Il est évident qu'il est impossible de prévoir toutes les situations de communication, surtout si l'on considère que toute communication est nouvelle, en particulier si l'on prend en compte l'évolution dans le temps. Inversement, il existe des situations où la création d'un genre ne se fait pas spontanément mais est imposée. C'est le cas dans la plupart des entreprises, qui proposent une liste de « types de documents » sur le contenu desquels le linguiste a peu d'intuitions et qu'il a souvent du mal à rattacher à un genre plus connu. Ainsi, le CNES (Centre National d'Etudes Spatiales), sur le projet Rosetta, propose une liste de 24 types de documents (liste « non-exhaustive et qui peut être modifiée en cas de besoins ») comme par exemple : Appel d'offre, Cahier des charges fonctionnel, Décisions du

Comité Directeur, Procédures, Spécifications… Il serait sans doute nécessaire d'examiner la notion de genre dans ce type de situation qui combine protocole imposé et rédaction spontanée.

- Une autre de ces raisons concerne le fait que, dans un même texte, plusieurs genres peuvent s'entrecroiser. Il est ainsi plutôt rare qu'un texte relève tout entier d'un seul et même genre ; par exemple, un article peut contenir des citations de lettres ou des définitions de dictionnaire.
- Une autre raison se base sur le décalage qui peut exister entre genres repérés *a priori*, sur des bases linguistiques intuitives et régularités linguistiques réelles identifiées grâce à l'étude croisée de différents paramètres. Plusieurs auteurs ont proposé de distinguer ces deux types de régularités ; c'est le cas par exemple de Achard :

> « P.Achard mettait l'accent sur l'impossibilité de proposer une typologie générale […] il proposait de s'intéresser avant tout au problème posé par le non-recouvrement entre [registre (rôle langagier attendu dans les différentes situations sociales et genre (fonctionnements linguistiques devant y correspondre)] : " la dynamique des écarts est le critère central par lequel s'articulent le linguistique et le social " » (Branca-Rosoff, 1999, 19).

C'est aussi le cas de Biber :

> « I use the term " genre " to refer to text categorizations made on the basis of external criteria relating to author/speaker purpose » (Biber, 1988, 68).
>
> « I use the term " text type " on the other hand, to refer to groupings of texts that are similar with respect to their linguistic form, irrespective of genre categories » (*ibid.,* 70).

Or, l'étude des régularités sémantico-syntaxiques peut permettre, selon les phénomènes étudiés, de rapprocher certains genres et d'en éloigner d'autres, qui pourtant paraissaient proches, intuitivement.

- Enfin, une raison est liée à la diversité des critères qui peuvent être utilisés :

> « Les genres de textes demeurent cependant des entités foncièrement vagues. Les multiples classements existants aujourd'hui restent divergents et partiels, et aucun d'entre eux ne peut prétendre constituer un modèle de référence stabilisé et cohérent […]. Cette difficulté de classement tient d'abord à la diversité des critères qui peuvent légitimement être utilisés pour définir un genre… » (Bronckart, 1996, 76).

Toutes les raisons invoquées sont pertinentes. C'est bien à cause de ce type de problèmes qu'il est aussi difficile de constituer un corpus représentatif : parce qu'il est difficile de délimiter des genres et, encore plus, d'évaluer leur part de représentation dans la somme totale des genres. Au final, la notion de genre permet de mieux expliquer et contrôler la variation linguistique mais certainement pas de la maîtriser. Il faut voir alors comment, tout en renonçant à établir des typologies définitives, le linguiste peut travailler (avec) cette notion pour mieux la cerner et lui donner un fondement, ce qui pourrait être une façon de comprendre le projet de Rastier :

> « Alors même que le nombre de langues décroît rapidement, un second chantier – moins grandiose mais aussi difficile – s'ouvre à la linguistique : décrire la diversité des discours et des genres ». (Rastier, 2001, 228).

Il ne faut pas oublier tout d'abord que certains genres sont, à l'évidence, repérés par la langue qui leur a donné un nom : le roman, la presse, la petite annonce, le manuel, le sermon, le cours… Il ne faut pas minimiser ce phénomène de dénomination qui manifeste la conscience d'une régularité, en tout cas dans la situation de communication (dont on suppose qu'elle s'accompagne de régularités linguistiques). Et il n'est pas illégitime d'utiliser ces genres dénommés comme points de départ, quitte à ce que ce soit pour les remettre en question (tout

comme il n'est pas illégitime d'utiliser une notion intuitive de domaines). On sait bien que ces dénominations peuvent être elles-mêmes déclinées en sous-genres, eux-mêmes difficilement cernables : roman historique, roman d'aventure, roman épistolaire, presse quotidienne, presse hebdomadaire, presse féminine, sportive… manuel de logiciel, d'appareil ménager… Mais cette possibilité n'empêche pas que certains genres, sans doute les plus saillants, sont marqués linguistiquement et qu'il faut tenir compte de cette caractéristique.
Un deuxième élément important concerne ce qui me semble une piste très prometteuse, qui est la nécessité de prendre en compte non seulement la situation de production des textes, dont presque tout le monde est d'accord pour admettre qu'elle permet de générer des genres textuels mais aussi, exactement au même titre, la situation d'interprétation.

2.2. La situation d'interprétation des textes : vers la définition de genres interprétatifs ?

2.2.1 La situation d'interprétation

Bien qu'évoquée dans les travaux des analystes de discours, la situation d'interprétation est rarement considérée comme pouvant faire l'objet d'une réflexion approfondie alors qu'elle intervient tout autant dans la construction du sens :

> « On doit reconnaître que le sens n'est ni dans l'objet (texte), ni dans le sujet (interprète) mais "dans " leur couplage au sein d'une pratique sociale. Pour l'interprète comme pour l'énonciateur s'imposent deux contraintes *in praesentia*, la situation et le contexte, et deux contraintes *in absentia*, le genre et l'intertexte. » (Rastier, 2001, 125).

Ce qui me semble un manque dans la réflexion sur le genre vient aussi souvent du fait que la situation qui est la plus couramment prise en considération est la situation d'interlocution, qui fait intervenir l'énonciateur et le destinataire initialement pressenti. La notion d'interprétation vient alors interroger celle de compréhension d'un texte. Or, il est courant pour le linguiste de se trouver dans une situation d'interprétation tout à fait différente d'abord parce qu'il est rarement le destinataire des textes qu'il étudie (hormis s'il mène un entretien), ensuite parce que l'interprétation qu'il propose ne se fait pas en direct :

> « Dans le cas des discours naturels : l'analyste est un récepteur non destinataire, c'est-à-dire qu'il traite un objet qui ne lui est en rien destiné (sauf s'il a lui-même participé à l'interaction qu'il étudie, ce qui pose d'autres problèmes) ; s'immisçant en intrus dans l'échange communicatif, il est généralement incapable de reconstituer la totalité des informations contextuelles pertinentes.» (Kerbrat-Orecchioni, 1996, 47-48).

> « …Interpréter dans l'interaction c'est ce que fait, par exemple, un formateur engagé dans une interaction avec un groupe… Ce n'est pas là, on le voit, le niveau où peut se situer le travail d'interprétation d'un linguiste. Pour lui, l'interprétation est une activité différée dans le temps : il prend comme objet une interaction déjà construite… » (Boutet, 1995a, 27).

Pour autant (et peut-être d'autant plus) l'interprétation du linguiste n'est pas neutre, elle est située socialement et culturellement :

> « …il convient de distinguer entre la situation d'interprétation et la situation d'énonciation. Elles sont relatives à des pratiques sociales et à des statuts individuels ; et les rôles énonciatifs et interprétatifs font alors la médiation entre les pratiques et les statuts sociaux qu'elles mettent en jeu. Les interprètes comme les énonciateurs doivent être socialement habilités. Qu'il y ait ou non identité spatio-temporelle des deux situations fondamentales, chacune suppose son univers de référence et ses univers d'assomption ». (Rastier, 1998, 105).

Il me semble tout à fait nécessaire de s'interroger sur cette situation d'interprétation propre au linguiste, en tant qu'il est lui-même socialement situé.

On peut essayer de dresser un premier panorama des situations possibles d'interprétation en fonction de ces deux éléments :

- interprétation par un destinataire/par un non-destinataire,
- interprétation in praesentia/in absentia.

L'interprétation par un destinataire correspond au cas classique du lecteur, soit que le rédacteur ait une idée précise du type de lecteur auquel il s'adresse, (par exemple, documents rédigés en interne, dans une entreprise), soit qu'il n'en ait pas d'idée précise (romans, par exemple).

L'interprétation *in praesentia* correspond à la situation dialogique mais aussi par exemple à la situation où un interprète non destinataire assiste à un échange entre protagonistes.

Dans le cas d'une interprétation par un non-destinataire, on peut identifier des types d'interprètes, en fonction de leur rôle social :

- ergonomes,
- critiques littéraires,
- traducteurs,
- correcteurs,
- ingénieurs de la connaissance,
- …
- linguistes.

Pour ce qui est de l'interprétation par des linguistes, elle peut obéir à au moins deux visées, que l'on a pu appeler théorique et appliquée. Dans la mesure où ces deux visées correspondent à deux modes de situations à mon avis nullement incompatibles, je me refuse à considérer la première comme digne de l'approche scientifique et la deuxième comme la seule mise en œuvre des résultats de la première, comme c'est généralement le cas :

> « Par linguistique appliquée, on désigne l'ensemble des recherches qui utilisent les démarches de la linguistique proprement dite pour certains problèmes de la vie courante et professionnelle…Partie utilitaire et pratique de la linguistique, elle est nécessaire mais ne peut évidemment constituer la fin unique des recherches en matière de langage.» (Dubois et *al.*, 1973, 43).

> « Participer à un travail de terrain et appréhender des données de langage socialement situées constitue un projet intellectuel qui pour de nombreux linguistes ne fait pas sens. » (Boutet, 1995a, 2).

Dans ma perspective, il y a entre ce que l'on appelle visée théorique et visée appliquée une différence de point de vue correspondant à deux situations différentes, l'un et l'autre de ces points de vue étant également utiles pour éclairer un objet extrêmement difficile à cerner : le sens textuel. Avec le point de vue théorique, on cherche à situer les résultats de l'étude appliquée, par rapport à des courants, par rapport à des présupposés théoriques, bref, par rapport à l'état des connaissances en linguistique (par exemple, on cherche à justifier l'existence de genres interprétatifs !).

Avec un point de vue appliqué, on cherche à construire un sens qui prenne en compte un besoin généralement exprimé par un tiers[10], souvent issu du monde économique. Mais cette demande peut parfaitement être elle-même interprétée en des termes qui la rapproche de problématiques généralement considérées comme plus théoriques[11]. Ainsi, il est clair que, quelle que soit la demande de la part des entreprises, elle peut toujours se résumer à un problème de « compréhension », le plus souvent entre deux populations d'acteurs ; ce que le chercheur en linguistique peut interpréter comme un problème de construction du sens et de divergences dans cette construction.

[10] Mais ce tiers existe aussi dans le cas de l'interprétation théorique, il est constitué par la communauté scientifique qui demande de trouver des régularités.

[11] *Cf.* en annexe la présentation des différents projets menés avec des entreprises et la façon dont nous avons interprété les demandes « appliquées » en questionnements théoriques.

Ainsi, si l'on accepte que l'interprétation est elle-même située, c'est-à-dire pas laissée à l'appréciation d'un individu isolé mais d'un individu en tant qu'il répond à une demande extérieure, on peut considérer que, quelle que soit cette demande (sociétale ou académique), il s'agit de mener une interprétation située.
Evidemment, il ne s'agit pas de considérer qu'il n'y a pas de différence entre interprétation dans un contexte appliqué et interprétation dans un contexte théorique mais il est possible de préciser ces différences sans opposer ces situations d'interprétations.

Validation des résultats

Avec un point de vue appliqué, les résultats sont validés par ceux-là même qui ont formulé la demande ; avec un point de vue académique, les résultats sont validés par les pairs universitaires. Et il n'est pas du tout assuré que les résultats de validation soient similaires. Dans le cas où il n'y a pas concordance, plusieurs possibilités peuvent s'ouvrir. Il se peut que, le point de vue étant très différent, les résultats obtenus soient très différents. Par exemple, il arrive que l'on fasse une réponse très *ad hoc* à un besoin applicatif, réponse qui, du point de vue de la puissance explicative n'est pas très satisfaisante. Il se peut aussi que la non-adéquation des validations ouvre des réflexions extrêmement fructueuses, soit qu'il faille revoir l'interprétation que l'on a faite du besoin applicatif et qu'on arrive à le repenser en termes plus théoriques, soit que le besoin applicatif permette sur les données un point de vue tout à fait original qui conduit à revoir la position théorique et à la formuler en des termes qui tiennent mieux compte d'un fonctionnement réel des phénomènes. Par exemple travailler sur la question de la mise au jour de points de vue dans l'étude de corpus du CNES permet de donner un éclairage nouveau sur la question de la polysémie (*cf*.chapitre IV). Dans ces cas-là, loin de relever de la seule mise en œuvre de connaissances, la réflexion sur le terrain vient au contraire alimenter la réflexion théorique et lui sert donc de point de départ.

Réutilisation des résultats

Dans un contexte théorique, la réutilisation des résultats est primordiale. Le fondement de la science linguistique (et de toute science sans doute) consiste à décrire des régularités de fonctionnement qui, autant que faire se peut, fassent système. Décrire signifie donc, en l'occurrence, prédire. Il faut pouvoir prédire que, dans la même configuration linguistique et/ou extralinguistique, c'est tel phénomène qui se produira (ou tel type de phénomène).
Dans un contexte appliqué, deux situations peuvent se présenter. Dans certains cas, on doit traiter une demande locale, en lien avec un corpus particulier (par exemple, étudier tel document pour en constituer un index). Dans ce cas-là, les résultats n'ont pas à être réutilisés. Dans d'autres cas, essentiellement pour des raisons financières, la demande concerne la constitution d'une ressource censée pouvoir être réutilisée dans différentes applications : par exemple, construire une BCT pour aider à la traduction, aider à la formation des nouveaux arrivés et aider à l'indexation de la documentation. Or, cette demande n'est pas toujours évidente à satisfaire. Elle suppose ou bien qu'on considère que tous ces types d'objectifs sont proches, ou bien que l'on réalise une étude pour évaluer la proximité de ces objectifs et leur compatibilité. Une troisième possibilité consiste à convaincre le demandeur qu'il faut concevoir la situation différemment et constituer trois types de modélisations, en fonction des trois objectifs (applications) identifié(e)s. Cette position n'est pas toujours facile à tenir face à une entreprise qui doit dégager des crédits pour faire réaliser trois types de produits là où elle espérait n'en avoir qu'un à financer. Une des propositions du groupe TIA (Terminologie et Intelligence Artificielle, *cf*. présentation dans le chapitre II, 1.3) consiste à dire que ce qui va être réutilisable n'est pas tant les résultats que les méthodes pour y parvenir. L'important ne serait pas de constituer des ressources mais de définir les méthodes pour les constituer rapidement (constitution de corpus, utilisation d'outils, utilisation de connaissances linguistiques).

Aussi bien dans une perspective théorique que dans une perspective appliquée, l'expérience d'analyse de corpus amène à revoir le niveau voire le type d'établissement de régularités. Dans les deux cas, elle suppose de travailler à définir des cadres d'interprétation aussi précis que possible.

Interprétation/méta-interprétation

C'est peut-être la différence la plus sûre, encore qu'elle ne soit pas toujours aussi tranchée que l'on pourrait croire (il y a certainement de la méta-interprétation dans certaines applications).
Il n'en reste pas moins que, de mon point de vue, on peut penser que, dans le cas de l'analyse de corpus, l'interprétation de type théorique est plutôt de la méta-interprétation parce qu'elle vise à généraliser, à stabiliser et expliquer les phénomènes, et à étudier les dépendances qui existent, d'une part entre phénomènes décrits et genre du corpus et d'autre part entre phénomènes décrits et objectif de la description. L'interprétation « applicative », quant à elle, est liée à un genre textuel (voire un texte) et/ou à une demande particulière. Il me semble ainsi que la méta-interprétation (donc l'approche théorique) consiste à se situer dans une perspective essentiellement langagière en essayant de repérer les relations de dépendance avec les éléments extra-linguistiques.
Interprétation théorique et interprétation applicative relèvent donc pour moi de deux situations, certes différentes mais qui ne sont pas incompatibles. Et, de mon point de vue, si elles ne sont pas incompatibles, c'est aussi parce qu'elles s'inscrivent dans des paradigmes, certes différents mais pas opposés que l'on peut peut-être constituer en genres.

2.2.2 Comment justifier l'existence de « genres interprétatifs »

Il me semble qu'il peut être très fructueux d'établir un parallèle entre situation de production et situation d'interprétation et de pousser ce parallélisme jusqu'à définir des genres interprétatifs sur le modèle des genres textuels. Deux éléments majeurs me paraissent justifier ce parallélisme, l'un repose sur des réflexions théoriques, l'autre sur la réalité des fonctionnements linguistiques.

2.2.2.1 Arguments théoriques pour la définition de « genres interprétatifs »

Plusieurs arguments théoriques me semblent plaider en faveur de l'existence de genres interprétatifs. Fondamentalement toutefois, le principal argument repose sur le constat que la construction du sens est bien trop souvent considérée du seul point de vue du locuteur ou du rédacteur, simplement parce que c'est lui qui a la parole. Mais celui qui reçoit ou utilise la parole, interlocuteur reconnu comme tel ou autre type d'interprète, me semble également impliqué dans un processus de construction du sens[12]. La situation d'interprétation, tout comme la situation de production, relève donc d'un processus d'élaboration sémantique. Cela signifie que comprendre relève d'un processus sémantique mais aussi qu'interpréter, même lorsqu'on n'est pas l'interlocuteur pressenti, relève aussi d'une élaboration de ce type.
S'il en va ainsi, le même type de questionnement se pose que dans la situation de production : comment le processus sémantique peut-il être à la fois individuel et collectif ? Tout comme on a proposé la notion de genre pour expliquer l'inscription collective d'une production textuelle, il me semble intéressant et justifié de proposer cette même notion dans le cadre de la situation d'interprétation. Tout comme la situation de production donc, la situation d'interprétation pourrait relever d'un genre, c'est-à-dire d'un paradigme préexistant, voire « normatif » et en partie inconscient dans lequel s'inscrirait toute activité interprétative, ces trois éléments (inscription collective, caractère « normatif » et phénomène en partie inconscient) me

[12] Au point qu'on a pu dire que locuteur et interlocuteur étaient engagés conjointement dans le processus de construction du sens, malgré l'apparente passivité de l'interlocuteur.

semblant correspondre aux éléments principaux qui décrivent le genre textuel. Cette hypothèse signifie à la fois que, à cause de sa nature essentiellement sémantique, l'interprétation d'un texte ne peut être strictement individuelle et que les points de vue interprétatifs peuvent être regroupés. Mais tout comme pour les genres textuels, il paraît difficile de dresser une liste des genres interprétatifs. Toutefois, tout comme pour les genres textuels, certains genres interprétatifs semblent reconnus comme autant de points de vue interprétatifs possibles ; la liste, préliminaire, que j'ai établie ci-dessus (ergonomes, critiques,…) me semble pouvoir constituer une première liste de genres interprétatifs. La constitution de Bases de Connaissances Terminologiques relèverait ainsi d'un sous-genre interprétatif, qui pourrait être lui-même réorganisé en sous-genres : constitution d'index, de thésaurus, de modèles de connaissances pour des systèmes formels. Ce sous-genre interprétatif (les BCT) a comme particularité majeure de viser la mise sous forme relationnelle du contenu d'un corpus. Comme je le montrerai dans le chapitre V, cet objectif n'a rien de naturel et il suppose une adhésion collective à la possibilité de modéliser le sens sous cette forme.
Enfin, ce qui semble assez acceptable pour un objectif applicatif pourrait être examiné aussi pour des objectifs théoriques qui, certainement aussi, s'inscrivent dans un ou des paradigme(s) collectif(s) (présupposés scientifiques, approche théorique, mode d'étude…). On ne saurait mieux dire qu'Auroux :

> « Comme phénomènes, [les sciences] sont complexes et manifestent en général au moins trois types de composants : un *composant théorique* (concepts, procédures, observables), un *composant sociologique* (formation et organisation de la main-d'œuvre scientifique, sociétés savantes, organes de diffusion, etc.) et un *composant pratique* (les types de finalités que l'on se propose en construisant des connaissances scientifiques). […] Comme toutes les activités humaines elles n'existent pas indépendamment d'une représentation que l'on en a lorsqu'on les pratique. Autrement dit, elles donnent lieu à des concepts qui peuvent être construits en fonction de différents prototypes, selon le modèle dominant que l'on se donne ». (Auroux, 1998, 9).

Ce qui pourrait constituer la plus grande résistance à cette notion de genres interprétatifs vient de ce qu'il peut sembler difficile de l'ancrer dans un substrat langagier. Il est bien évident que cette dimension langagière est intimement liée à la définition du genre textuel ; c'est toujours par la supposée existence de régularités lexico-syntaxiques que se justifie un genre textuel[13]. Pourtant, il est possible de montrer que l'établissement de certaines régularités linguistiques n'a de sens que par rapport à un objectif bien précis. Le prochain paragraphe présentera quelques unes de ces régularités qui seront largement reprises par la suite.

2.2.2.2 Arguments empiriques

La suite de ce mémoire va m'amener à montrer que certains résultats d'analyse sémantique confirment non seulement l'existence de genres textuels mais aussi, et c'est sans doute plus nouveau, que d'autres confirment le rôle de la situation d'interprétation dans la construction du sens, voire même l'existence de genres interprétatifs. Plus exactement, je soutiens que la prise en compte de l'objectif de l'interprétation peut amener à mettre au jour des régularités originales qui dépendent plus de ce point de vue interprétatif que du point de vue textuel. Par ailleurs, certains résultats nécessitent de faire intervenir à la fois le genre textuel et le genre interprétatif.

[13] Mais il faut rappeler aussi que la mise au jour de régularités linguistiques peut amener à revoir certains genres établis sur des bases intuitives, comme l'ont montré par exemple les travaux de Biber dont j'aurai l'occasion de reparler.

Je me place dans une situation d'interprétation qui vise à construire des bases de connaissances à partir de corpus, c'est-à-dire dans une situation qui concerne à la fois une visée applicative (les BCT sont généralement construites en réponse à une demande) et une visée théorique : comment systématiser les méthodes de construction permettant d'élaborer, à partir d'un corpus, un modèle consistant en des entités reliées par des relations ?
Comme en une saisie synthétique des études qui suivent, je peux ainsi montrer comment cette corrélation genre/régularités linguistiques se met en place :

- Concernant la réflexion sur les nominalisations déverbales, on peut dire que leur sur-utilisation dans les textes techniques permet de considérer qu'il s'agit d'une caractéristique propre à ce genre textuel (chapitre IV).
- La réflexion sur la polysémie dans un corpus spécialisé, quant à elle, n'a de sens que par rapport à la notion de points de vue à l'œuvre dans les corpus. Ainsi, il ne semble pas suffisant de prendre en compte la nature spécialisée d'un corpus pour décider de la présence ou non du phénomène de polysémie (les corpus spécialisés sont réputés comme comportant peu de polysémie) ; cela dépend du point de vue interprétatif que l'on adopte. Dans une situation interprétative qui vise à mettre en évidence des points de vue de différents acteurs, il est tout à fait possible de mettre au jour de la polysémie (chapitre IV). La recherche de polysémie relève ainsi du point de vue que l'on adopte, point de vue qui s'inscrit lui-même dans un genre.
- Avec la question des marqueurs de relations conceptuelles (chapitre V), le problème devient particulièrement complexe. Tout d'abord, rechercher des marqueurs de relations conceptuelles constitue déjà un point de vue, une situation interprétative qui permet de donner un relief particulier à certains éléments linguistiques. Trois résultats seront particulièrement décrits. L'un permet de montrer que *chez* peut marquer la méronymie dans certains textes : s'ils sont du domaine des sciences naturelles et didactiques. L'explication du fonctionnement « méronymique » de *chez* ne peut donc s'expliquer que par l'existence d'un genre textuel et d'un genre interprétatif (rechercher des marqueurs de méronymie). Le second résultat concerne l'étude de *avec* lorsqu'il marque la méronymie. Il se trouve que c'est le cas en particulier dans certains textes : petites annonces immobilières, description d'itinéraires, catalogues de jouets. L'objectif d'interprétation (le genre interprétatif) vient ici prendre le pas sur le genre textuel. En effet, il semble que le marquage de la méronymie par *avec*, amène à remettre en question l'existence d'un genre « petites annonces » les petites annonces immobilières se montrant plus proches des catalogues de jouets que des petites annonces de ventes de voiture. La troisième étude met elle aussi au jour, peut-être de manière encore plus évidente, l'importance du genre interprétatif. Elle concerne les cas de reprise anaphorique par un hyperonyme (*Un chat entra. Cet animal au pelage soigné…*). Or il s'avère que ce « marqueur » permet de repérer des hyperonymes tout à fait différents de ceux que l'on repère avec des marqueurs définitoires par exemple. En effet, les noms repérés avec le marqueur anaphorique semblent être de plus haut niveau que ceux repérés avec d'autres marqueurs. Ainsi, selon le niveau de hiérarchie où l'on veut situer sa construction taxinomique, il est possible qu'il ne faille pas utiliser les mêmes marqueurs. Si cette hypothèse se vérifie, cela signifie que le fonctionnement des marqueurs est à mettre en relation avec le type de résultats que l'on souhaite obtenir.

La nature du travail que l'on souhaite mener joue un rôle déterminant sur la nature du corpus que l'on va constituer. Par exemple, comme on le verra dans la suite du mémoire, si l'on souhaite constituer une base de connaissances terminologiques en utilisant des marqueurs de

relation, ce qui est important est moins que les textes utilisés relèvent de tel ou tel genre textuel mais plutôt que ces textes soient riches en marqueurs. On peut alors envisager de faire des tests rapides sur certains textes afin de sélectionner les plus adaptés à l'étude. Certes, on peut penser que ces textes, riches en marqueurs, relèvent d'un genre « didactique » (genre étudié par Béacco et Moirand en particulier, (Béacco et Moirand, 1995)). Mais cette caractérisation est intuitive ; or, d'une part, on n'a parfois pas d'intuition sur la nature didactique ou non des textes que l'on a à disposition et, d'autre part, certains textes, pourtant réputés comme didactiques contiennent peu de marqueurs de relations.
L'application d'une première série de marqueurs dont on sait qu'ils sont pertinents pour construire une BCT (certains marqueurs d'hyperonymie par exemple) pourrait permettre de sélectionner des textes pertinents au regard de l'objectif de l'étude et de la méthode choisie.
Toutes ces questions seront abordées dans le détail dans la suite du mémoire.

3. Conclusion

Ce premier chapitre a une importance particulière qui a permis de préciser ma position. Il m'a amenée à poser mon cadre d'étude, entre théorie et application, l'application étant entendue comme un lieu d'interprétation qui vient nourrir une méta-interprétation plus propre à l'analyse théorique. La prise en compte de textes réels oblige à la confrontation avec une réalité qui semble souvent échapper à toute possibilité de catégorisation : il y a variation parce qu'il y a création. La sémantique textuelle est toujours une linguistique de l'interprétation, ce qui ne veut pas dire que cette interprétation n'a aucun fondement ou qu'elle est entièrement libre. Tout comme la production textuelle, il est probable que l'interprétation obéit à des régularités qui peuvent se constituer en genres, qui restent largement à étudier.
La question qui me préoccupera dans la suite de ce mémoire est celle de comprendre ce que signifie construire une Base de Connaissances Terminologiques à partir de corpus, ce que cela implique comme type d'interprétation et comme choix de modélisation. Finalement, il s'agit d'expliquer selon quelles modalités il est possible de passer d'un texte linéaire à une modélisation spatiale qui privilégie les formes nominales pour les nœuds et verbales pour les relations.
Si la recherche en sémantique ne relève plus d'une maîtrise du fonctionnement du sens mais plutôt d'une tentative d'explication de phénomènes socialement et culturellement situés, elle gagne sans doute en perspective d'études ce qu'elle perd en (ce que d'aucuns voudraient appeler) scientificité. La sémantique est ainsi fondamentalement une science humaine et sociale :

> « Toutes les sciences, même les sciences de la Nature, sont des sciences de l'homme, elles ont l'homme pour auteur et aussi pour instrument et destinataire. L'homme est à la fois le sujet et l'objet du savoir, non pas un homme abstrait, un sujet idéal et universel, mais un homme situé dans une communauté humaine au sein d'un milieu culturel qui lui fournit son vocabulaire, ses moyens de connaissance, et sanctionne en fin de compte les résultats obtenus ». (Gusdorf, 1988,212).

Chapitre II
Les Bases de Connaissances Terminologiques

Les bases de connaissances terminologiques (BCT) sont un concept récent (1992). Leur apparition manifeste deux évolutions importantes, d'ailleurs reliées : l'une concerne l'affirmation d'une relation entre problématique de la terminologie et problématique de l'intelligence artificielle (dans le terme de bases de connaissances terminologiques, on retrouve en effet, celui de base de connaissances, qui provient de l'intelligence artificielle), l'autre concerne la nécessité de donner une représentation relationnelle aux définitions terminologiques jusqu'alors existant sous une forme uniquement discursive. La réflexion qui est menée sur les BCT est ainsi d'emblée pluridisciplinaire et la création du terme de BCT ne vient en fait que consommer un lien qui s'est mis en place plusieurs années auparavant. Bien que très récent, ce concept de BCT a déjà beaucoup évolué et on peut penser que cette évolution vient de ce que la confrontation interdisciplinaire a conduit chacune des disciplines à éclaircir ses postulats et à développer la réflexion. Ainsi, pour la terminologie, c'est la réflexion sur le recours aux corpus qui a bénéficié de la rencontre avec l'informatique ; et pour l'intelligence artificielle, la réflexion sur les ontologies et leur éventuelle généricité a certainement été enrichie par l'apport de la vision de la terminologie textuelle. Si bien que, si on les replace dans leur chronologie, les BCT peuvent être vues à la fois comme un point de jonction qui concrétise une réflexion pluridisciplinaire et comme une ouverture vers des analyses mono-disciplinaires enrichies. Ce chapitre va faire le point sur ce double mouvement, tout en me permettant de préciser mon point de vue de linguiste. En effet, dans cette évolution de la terminologie vers les textes d'une part et vers la modélisation d'autre part, la rencontre avec la lexicologie textuelle était urgente et inévitable. Au-delà des seules questions méthodologiques à propos de la constitution des données, cette rencontre a eu des conséquences qui pourraient être majeures (en tout cas qui l'ont été pour moi) sur le rôle des corpus et de la prise en compte d'un contexte appliqué en linguistique, mais aussi sur les possibilités de collaboration entre sémantique textuelle et traitement automatique des langues.
La constitution de Bases de Connaissances Terminologiques (BCT) permet ainsi d'aborder un grand nombre de problèmes sémantiques en lien avec un fonctionnement lexical, tout en imposant un questionnement sur deux points fondamentaux : comment passe-t-on d'un corpus à un modèle ? Et en quoi l'informatique vient-elle assister ou contraindre ce passage ?

Ces deux questions constituent le fondement de ma réflexion depuis plusieurs années et balisent ma conviction que la linguistique, et particulièrement la sémantique, est à un tournant majeur de son histoire du fait de la mise à sa disposition de corpus sous une forme électronique et d'outils pour traiter ces corpus. Il s'agit pour la linguistique de prendre la mesure de cette possibilité pour permettre son évolution. La collaboration avec le TAL est nécessaire mais il est nécessaire aussi que l'évolution incontournable de la sémantique ne soit pas confiée aux seuls informaticiens qui n'ont ni les mêmes objectifs, ni la même histoire ni les mêmes méthodes que les linguistes. Ce chapitre m'amènera à préciser mon cadre d'analyse, par rapport à la terminologie « classique » et par rapport à l'informatique.

1. Origine du concept de BCT

Très clairement, les Bases de Connaissances Terminologiques s'inscrivent d'emblée dans une vision interdisciplinaire. Si le terme de BCT est apparu chez les Canadiens Meyer et Skuce en 1992 (Meyer et *al.*,1992 a et b), les collaborations possibles et nécessaires entre terminologie et intelligence artificielle étaient déjà mentionnées dans des articles antérieurs, comme nous le soulignions dans notre article de 1995 (Bourigault et Condamines, 1995). Quelle que soit leur conception du fonctionnement de la terminologie, de nombreux auteurs, travaillant sur la terminologie et bien informés des travaux sur les systèmes experts et des besoins des entreprises avaient montré que cette rencontre était inévitable, que ce soit Parent (Parent, 1989), Wijnands (Wijnands, 1989, 1993) ou encore Felber :

> « Etant donné que les études sur l'intelligence artificielle et la mise au point de systèmes experts sont amenés à traiter des systèmes de notions, des combinaisons de notions, de la représentation conceptuelle de la réalité, etc., il faut s'attendre à un renforcement des affinités entre la théorie générale de la terminologie et l'informatique dans un avenir proche ». (Felber, 1987, 91).

A peu près à la même période, les informaticiens chargés de la modélisation des connaissances ont pris conscience de la nécessité de s'appuyer sur des textes pour « extraire et modéliser » la connaissance ; le projet Esprit KADS (Knowledge Analysis and Documentation System, 1983-1993) a joué un très grand rôle dans cette communauté. Mais si chacune des communautés (terminologie et IA) voyait un intérêt à la collaboration avec l'autre, la rencontre n'était pas encore consommée. Ainsi, si, à la fin des années 80, linguistes et informaticiens collaboraient depuis déjà longtemps, sur des problèmes de Traitement automatique de la langue (voir par exemple Sabah, 1988) ou des problèmes de formalisation de connaissances, un autre type de collaboration était encore en gestation, qui s'appuyait sur l'analyse de corpus réels et qui allait amener des mutations importantes, aussi bien en terminologie, qu'en intelligence artificielle et même en linguistique. Outre la dimension interdisciplinaire, cette collaboration a eu pour caractéristique de prendre en compte d'emblée, la réalité des applications. La première étape de cette mutation s'est faite par la création du terme de Base de Connaissances Terminologiques (plus exactement Terminological Knowledge Base) par les Canadiens Meyer, linguiste terminologue et Skuce, informaticien, tous deux à l'Université d'Ottawa.

1.1. Premiers projets de constitution de BCT : CODE et QUIRK

Au début des années 90, plusieurs projets ont manifesté la rencontre de la terminologie et de l'informatique pour la constitution de modèles de connaissances.

1.1.1 Le projet COGNITERM

Conscients de la grande parenté existant entre la constitution de terminologies et l'acquisition de connaissances à partir de textes – dans (Skuce et Meyer, 1991), ils parlent de « symbiotic

relationship » –, ces deux chercheurs ont mis sur pied le projet COGNITERM dont l'un des objectifs était de montrer comment un outil d'aide à l'acquisition et à la modélisation des connaissances pouvait être utilisé dans un contexte terminologique (Meyer et *al.*, 1992). Le projet a ainsi été d'abord conçu comme un système qui aide un cogniticien à construire une base de connaissances décrivant les concepts d'un domaine. Fonctionnant grâce à un langage de frames classique, incluant un calcul d'héritage et permettant une interface de visualisation des taxinomies construites, cet outil a ensuite été testé pour aider un terminologue à construire un produit terminologique d'un nouveau type : une base de connaissances terminologiques. Dans un premier temps, la collaboration terminologie/informatique s'est plutôt faite dans le sens de l'informatique vers la terminologie puisque l'outil CODE (Conceptually Oriented Description Environment), conçu dans une visée informatique était mis à disposition des terminologues, avec toutes les possibilités mais aussi les contraintes que cela entraînait pour les utilisateurs.

1.1.2 Le system QUIRK

Egalement conçu dans un contexte interdisciplinaire, Ahmad étant informaticien et Roggers, linguiste à l'Université de Surrey, le system QUIRK est un outil d'aide à l'analyse de texte et au développement de ressources lexicales. Ce système a été pensé d'emblée pour différents types de constructions : bases de données terminologiques, dictionnaires, bases de connaissances. Par rapport à CODE, l'outil QUIRK propose une aide à l'exploration de textes ; cette aide, sous la forme de l'outil TEXT ANALYSER (Kavanagh, 1996) a d'ailleurs été intégrée plus tard dans CODE.
QUIRK apparaît ainsi, dans un premier temps, plus complet et plus autonome que CODE en tout cas d'après ce qu'en dit le créateur de QUIRK :

> « It appears that code users have to interpret the cogniterm methodology, whereas System Quirk guides the user through details of text typology and tries to autonomously interpret the data it has access to » (Ahmad, 1993, 67).

La principale différence entre les deux systèmes est peut-être à noter dans le type d'utilisations visées. Les Canadiens se limitent à une utilisation restreinte à la terminologie et l'ingénierie des connaissances, et plutôt à la mise à disposition pour des terminologues d'un outil créé par et pour des informaticiens, alors que les Anglais envisagent une utilisation qui concerne aussi bien les terminologues ou les ingénieurs de la connaissance que les lexicologues. Si cette plus grande couverture peut paraître intéressante, elle présente le risque de dissimuler des questions sur la réalité des besoins en confondant les problématiques : la spécificité de l'utilisation des corpus pour chaque type de besoin (dictionnaire « général », terminologie, ontologies formelles…), et des corpus spécialisés, n'est pas ainsi clairement analysée ce qui peut conduire à un outil trop généraliste qui ne répond vraiment à aucun besoin particulier.

1.2. Le système toulousain « ARAMIIHS »[14]

De manière assez parallèle aux travaux canadiens et anglais[15], une réflexion sur les relations entre terminologie, linguistique et informatique a démarré dans le laboratoire ARAMIIHS. En stage post-doctoral cofinancé par Matra Marconi Space (MMS) et le CNRS, j'ai eu en grande partie la charge d'animer cette réflexion qui a donné lieu à la constitution d'un premier

[14] Action, Recherche et Application Matra/Irit en Interface Homme Système, Laboratoire mixte Matra Marconi Space/CNRS.
[15] Il est d'ailleurs intéressant de noter que lors du même colloque TKE (Terminology and Knowledge Engineering), en 1993, les équipes de ces trois pays ont présenté des travaux sur des thèmes similaires.

modèle de données et d'une première méthode d'analyse. Ce contexte de travail était particulièrement intéressant car il était interdisciplinaire de deux façons : il était constitué à la fois d'informaticiens et de linguistes et à la fois de chercheurs et de membres de l'entreprise. J'étais responsable de la partie recherche en terminologie et des relations avec les informaticiens et avec l'entreprise[16]. Nous avons été jusqu'à une dizaine de personnes, travaillant sur le projet européen EUROLANG qui visait avant tout l'aide à la traduction. Pour moi qui avais commencé mon travail de recherche en linguistique par une thèse sur la subordination en français, thèse somme toute assez classique même si elle avait pour originalité d'avoir été réalisée dans un laboratoire d'informatique (Institut de Recherche en Informatique de Toulouse) et même si la description visait une possible utilisation en informatique, la confrontation était triple, avec les travaux en terminologie, avec la réalité d'une demande en entreprise et ses conséquences aussi bien sur la réflexion en intelligence artificielle qu'en linguistique.

– Confrontation avec les travaux en terminologie.

Ma surprise a été grande de constater que les travaux existant en terminologie entretenaient pour la plupart une parenté très éloignée avec les travaux en lexicologie. Leur mérite était de s'appuyer souvent sur des données réelles, par exemple sur des données textuelles, mais les méthodes préconisées par les auteurs faisaient preuve d'une grande ignorance de la réalité linguistique et étaient fortement motivées par une volonté de normalisation. Il m'a semblé nécessaire et urgent de situer la terminologie comme une discipline linguistique, tout en intégrant ses caractéristiques propres comme le nécessaire recours à un corpus et le lien avec des besoins réels.

– Confrontation avec une demande réelle, du point de vue de l'informatique.

Dans le cadre de projets en entreprise, les travaux en informatique ne peuvent pas concerner de simples modélisation ou formalisation introspectives sans lien avec une demande précise ; la demande relève au contraire de besoins clairement identifiés et concerne des outils qui doivent être rapidement opérationnels. Les besoins identifiés à MMS étaient divers, d'une part la traduction assistée par ordinateur (en lien avec le projet EUROLANG) et d'autre part la formation des nouveaux arrivés (besoin émis par l'organe de formation interne : l'Ecole de l'Espace), enfin l'aide à la rédaction de documents. Besoins multiples qu'il n'était pas facile de prendre en compte, le premier demandait que soient particulièrement travaillées l'acquisition et la modélisation des termes en discours pour qu'elles puissent s'intégrer au système de Traitement Automatique de la Langue d'EUROLANG, le deuxième demandait une attention accrue à la modélisation des termes en tant que vecteurs de connaissances, le troisième participait des deux préoccupations. Ces différents besoins nous ont conduits à un premier choix de représentation qui consistait à créer un modèle dans lequel étaient nettement distingués un champ linguistique et un champ conceptuel (la présentation du modèle de données est détaillée dans le chapitre III).

Il est rapidement apparu que l'élaboration des données devait prendre appui sur une analyse de corpus, ce qui supposait de mettre au point une méthode d'analyse mais aussi un outil d'aide. En effet, aux débuts des années 90, nous n'avons pu trouver un outil d'analyse de textes français et nous avons été contraints de construire un petit outil de ce type, qui nous permettait d'accéder à des occurrences en corpus : Amsili a constitué un outil à partir des fonctionnalités Lex et Yak de Unix.

Pour ce qui est du stockage des données, le projet était assez innovant puisque les données terminologiques ont été stockées en SGML.

[16] Pour la première fois, j'étais confrontée à la réalité du terrain et assez mal préparée aux questions du type « quel est le retour sur investissement ? » qu'on ne manquait pas de me poser dans les réunions d'avancement !

– Confrontation avec une demande réelle, du point de vue de la linguistique.

Le contexte même du projet faisait qu'il était impossible de mener une linguistique « de bureau », pour reprendre le terme de Corbin (Corbin, 1980). La demande de l'entreprise n'était d'ailleurs pas claire dans un premier temps et il fallait surtout montrer ce que la linguistique pouvait lui apporter tant, pour la plupart des ingénieurs, il pouvait s'agir d'une discipline étrangère. A la fois pour sensibiliser les ingénieurs et pour évaluer leur intérêt pour des problèmes de langue, nous avons lancé une enquête sur les besoins en terminologie à l'intérieur de MMS. Les résultats ont été très encourageants, d'abord par le nombre de réponses obtenues (plus de 30% de réponses alors que ce genre de demande « spontanée » ne recueille en général que 10 % de réponses), mais aussi par la conscience de difficultés liées à la langue qu'ils révélaient. En même temps, ces réponses témoignaient d'une attente importante qui ne pouvait pas être comblée dans le cadre d'un projet de quelques mois ou de quelques années.

Confrontations finalement très riches, qui nous ont conduits à définir un modèle de données (Condamines et Amsili, 1993) mais aussi à démarrer les premières réflexions sur l'analyse de corpus (Condamines, 1995), les liens entre lexicologie et terminologie, les liens entre linguistique et informatique et l'inévitable rencontre, en matière de terminologie, avec une demande réelle.

Le modèle de données que nous avons constitué était très proche de ceux proposés par d'autres groupes ; il s'est élaboré dans un type de problématique très similaire, (*cf.* partie 2. dans ce chapitre). Ce modèle a d'abord servi de base, à MMS, à un outil d'aide à la rédaction puis il a été utilisé par l'équipe de Nathalie Aussenac, de l'IRIT (Institut de recherche en Informatique de Toulouse), pour constituer un outil de gestion de terminologie (GEDITERM), (Aussenac-Gilles et Séguéla, 1999). En 1995, nous avons inauguré avec Nathalie Aussenac-Gilles une réflexion commune sur les BCT, grâce à un projet du GIS Science de la Cognition[17], cette réflexion commune n'a jamais cessé depuis.

1.3. Le groupe TIA

Avec une même conscience de la nécessité d'organiser une réflexion sur les relations entre terminologie, linguistique et informatique, Didier Bourigault[18] et moi-même avons décidé de constituer un groupe de travail pluridisciplinaire, en 1993, groupe de travail qui fonctionne encore aujourd'hui[19]. Rassemblant une quinzaine de chercheurs[20], ce groupe s'est donné à l'origine un double objectif :

– Réfléchir sur le concept de Base de Connaissances Terminologiques et son lien avec les ontologies.

  Depuis son origine, TIA constitue un lieu de réflexion original qui a largement contribué à faire évoluer la problématique des BCT et des ontologies, en créant un

[17] « Terminologie, Modélisation des connaissances et Systèmes Hypertextuels de consultation de documentation technique », Projet du GIS Sciences de la Cognition, 1995-1997.

[18] A cette époque à la DER d'EDF, Didier Bourigault a été recruté au CNRS, d'abord au LLI (Laboratoire de Linguistique Informatique) puis à l'ERSS.

[19] Il est actuellement animé par Sylvie Szulman (LIPN) et moi-même.

[20] Participent ou ont participé à ce groupe : M. Amar (Paris X), N. Aussenac-Gilles (IRIT, Toulouse), J. Bouaud (Diams, AP-HP, Paris), B. Biebow (LIPN, Villetaneuse), D. Bourigault (actuellement ERSS, Toulouse), F. Cerbah (Dassault), J. Charlet (Diams, AP-HP, Paris), A. Condamines (ERSS, Toulouse), R. Dieng (INRIA, Nice), P. Frath (Univ. De Strasbourg), C. Enguehard (IRIN, Nantes), B. Habert (actuellement LIMSI, Orsay), C. Jacquemin (actuellement, LIMSI, Orsay), G. Otman (à l'époque CTN, Villetaneuse), A. Nazarenko (LIPN, Villetaneuse), F. Rousselot (ERIC, Strasbourg), J. Royauté (INIST, Nancy), M. Slodzian (Inalco, Paris), S. Szulman (LIPN, Villetaneuse), Y. Toussaint (CRIN, Nancy), P. Zweigenbaum (Diams, AP-HP, Paris).
http://www.biomath.jussieu.fr/TIA/

courant nouveau qui met l'accent sur le rôle des corpus (voir ci-dessous). Ce courant est maintenant très vivant en France et commence à être reconnu de manière internationale : la plupart des colloques sur l'Ingénierie des Connaissances mettent à leur programme un thème sur la constitution d'ontologies à partir de textes (EKAW (European Knowledge Acquisition Workshop, workshop de l'ECAI (European Conference on Artificial Intelligence) et de l'IJCAI (International Joint Conference on Artificial Intelligence), par exemple.

– Animer la recherche sur ces thèmes en particulier en France.
  On peut reconnaître que ce groupe a particulièrement bien réussi dans cette tâche puisque chacune des journées organisées (TIA'95 à Paris, TIA'97 à Toulouse, TIA'99 à Nantes et TIA'2001 à Nancy, Journée ATALA « Terminologie et Traitement automatique des langues », 1995, workshop « Ontologies and Texts », EKAW 2001) a permis à plus de cent personnes de se rencontrer (hormis le workshop EKAW, ouvert aux seuls auteurs d'une communication).

Ainsi, différents groupes, dans différents pays, et toujours avec la caractéristique d'être interdisciplinaires ont travaillé de manière concomitante sur le concept de BCT. Avec près de 10 ans de recul, on voit mieux à présent quels ont été les enjeux, les attentes et les impasses des BCT.

2. Le concept de BCT : caractéristiques du modèle de données

Depuis sa création, le concept de BCT a été beaucoup travaillé et remanié, on peut désormais examiner ce qui a fondé l'homogénéité de ce concept et comment les problématiques soulevées ont évolué.
Le premier élément qui fonde le concept de BCT concerne les différences qu'il manifeste par rapport aux anciennes bases de données terminologiques. Trois éléments peuvent être repérés : la mise en réseau des concepts, la distinction terme/concept, la prise en compte de l'usage.

2.1. Mise en réseau des concepts

La mise en réseau des concepts se retrouve dans tous les modèles de BCT ; la représentation relationnelle est l'élément majeur qui distingue les BCT des Bases de Données Terminologiques (BDT), les deux autres éléments n'étant pas toujours pris en compte dans les modèles de BCT. En effet, dans les BDT classiques, l'information conceptuelle se manifeste essentiellement sous la forme de définitions en langue naturelle, inexploitables de manière automatique. On peut d'ailleurs s'étonner de cet état de faits car l'idée de « systèmes notionnels » est très présente dans les écrits « traditionnels » sur la terminologie. Si la rencontre terminologie/intelligence artificielle s'est faite principalement sur la question des relations, c'est bien parce qu'il y avait là un besoin reconnu comme très proche dans les deux disciplines : besoin pour la terminologie de valider l'idée des réseaux notionnels, besoin pour l'IA de trouver des méthodes pour constituer des réseaux conceptuels qui, tout à la fois, s'ancreraient dans une réalité linguistique, et seraient prêts à l'emploi pour des utilisations en TAL puisqu'ils obéiraient aux exigences de cohérence et de complétude. L'idée de système constituait un vecteur de convergence très fort qui a permis la rencontre de ces disciplines. Ajouté à cela, il y avait, pour chaque discipline, une espérance moins scientifique mais tout aussi légitime : pour la terminologie, obtenir une reconnaissance et une valorisation de ses travaux, longtemps cantonnés aux services annexes dans les entreprises ; pour l'informatique, trouver des moyens systématiques et donc automatisables d'acquérir des connaissances dans

les textes et donc de gagner en efficacité. La rencontre s'est faite avec un espoir, clairement affiché, de « symbiose », pour reprendre le terme de Meyer et Skuce. Elle fut à la fois très fructueuse et porteuse de désillusions, et ce double résultat se manifeste particulièrement autour de la question des relations qui sera longuement évoquée par la suite.

2.2. Le problème du concept

La notion de concept est utilisée par un grand nombre de disciplines : linguistique, psychologie, terminologie, pédagogie, représentation des connaissances… pour n'en citer que quelques-unes. Mais chaque utilisation prend un sens différent selon le point de vue que l'on adopte. Si cette notion a pu servir de point de contact entre l'IA et la terminologie, cela s'est souvent fait au prix d'une occultation de leurs points de vue. Or, si ce même terme peut être utilisé à la fois en terminologie et en IA, il est nécessaire de mettre au jour ce qui constitue les différences dans les deux disciplines. Alors seulement peut-on caractériser la complémentarité des disciplines et leurs apports mutuels.

2.2.1 Concept et terminologie

La notion de concept est à la base de la plupart des travaux en terminologie, en tout cas dans une vision traditionnelle. Dans ce type d'approche, un terme sert à désigner un concept, également appelé notion et considéré comme préexistant :

> « Le terme se définit comme unité signifiante constituée d'un mot (terme isolé) ou de plusieurs mots (termes complexes) qui désigne un concept, de façon univoque à l'intérieur d'un domaine … » (OLF, 1985).

> « Terme : désignation au moyen d'une unité linguistique d'une notion définie dans une langue de spécialité. » (ISO 1087, 1990).

Il faut noter que ce point de vue s'intègre dans une vision d'ensemble de la terminologie qui s'inscrit dans l'approche initiée par Wuster dans les années 30. Comme l'a montré Monique Slodzian (Slodzian, 1994), la doctrine wustérienne se fonde sur le postulat d'une langue universelle permettant l'accès à la connaissance. Avec les langues « spécialisées », associées à des domaines parfaitement maîtrisés par des experts, ce type de doctrine a cru trouver confirmation de son hypothèse de la possibilité d'une langue qui, à défaut d'être pure, pourrait être purifiée (normalisée) :

> «[…] jusqu'à une date récente, la linguistique n'a fait valoir que l'évolution libre, non dirigée, de la langue. C'est l'usage effectif de cette dernière qui, dans la langue commune, sert de norme. On peut appeler cette norme la norme descriptive. En revanche, en terminologie, fertile en notions et en termes, cette évolution libre de la langue mène à une confusion inacceptable… » (Wuster, 1981, 65).

Un tel point de vue suppose des éléments parfaitement définis et cernés : les concepts, et un parti-pris exclusif pour la monosémie. Même si le dogme s'est parfois assoupli, la notion de concept reste très présente en terminologie et il est nécessaire de s'interroger sur sa pertinence linguistique.

En linguistique, le terme de concept est peu utilisé. Lorsqu'il l'est, il me semble que deux opinions essentielles prévalent : l'une fait du concept l'équivalent du signifié, position que l'on retrouve chez Saussure par exemple, l'autre donne au concept un statut particulier, premier :

> « Nous pensons que les items lexicaux présupposent l'existence de concepts, c'est-à-dire d'entités générales plus connues sous le nom d'universaux lorsqu'elles se présentent sous forme nominale. » (Kleiber, 1981, 24).

Cette deuxième position n'est pas très éloignée de celle des terminologues « classiques », en tout cas pour ce qui concerne la préexistence des concepts.

Une position souvent adoptée consiste à présenter le concept comme l'élément qui va permettre de rendre compte de certains phénomènes sémantiques comme la polysémie ou l'homonymie ou l'équivalence d'une langue à l'autre :

> « Le concept ne peut se confondre avec un signifié interlinguistique, à la fois parce que plusieurs signes linguistiques peuvent être synonymes (ou équivalents de langue à langue) et parce qu'un concept peut être propre à un groupe social ou universel et également lexicalisé dans une langue. » (Lerat, 1989, 57-58).

> « la distinction entre signe et concept peut contribuer à mettre en valeur ce genre de phénomène [la possibilité pour un signe d'être polysémique en corpus spécialisé], en prenant en considération la nature foncièrement polysémique des signes de langues » (Depecker, 2000, 107).

Dans ce type d'approche, si l'on a besoin du concept c'est pour permettre une généralisation, une reconnaissance de similitudes à travers des formes différentes. Cette façon de voir pourrait plaider pour une vision qui maintient un lien fort avec le linguistique, les usages ; ce n'est pas toujours le cas et beaucoup d'auteurs tiennent à maintenir le concept comme premier et universel. Pourtant, il est possible de retenir la notion de concept comme un élément qui rassemble, qui permet de définir, sans qu'il soit nécessaire de le poser comme premier. On peut au contraire considérer qu'il se constitue à partir d'usages de la langue d'une part et de construction de terminologues (ou de n'importe quel type d'interprétant) d'autre part. Je reprendrai la définition de Rastier du concept comme « signifié normé » (Rastier et *al.*, 1994) mais en attribuant à *normé* deux sens. L'un qui serait lié à celui de « normaison », l'autre à celui de « normalisation » :

> « L'analyse tirerait profit à opposer deux procès normatifs : la normaison, relevant de l'activité spontanée à l'œuvre dans tout échange, et la normalisation, domaine des interventions conscientes et planifiées » (Gaudin, 1993, 173).

Je retirerais simplement le terme de *planifié* de la définition de Gaudin qui me semble renvoyer à une obligation d'utilisation par une instance investie d'un pouvoir particulier. En revanche, dans une vision qui va des usages à la construction d'un modèle (*cf.* chapitre III), l'élaboration d'un signifié en concept se fait bien, me semble-t-il, d'abord par le repérage de régularités d'usages (normaison) puis par le choix de conférer à certains signes linguistiques un statut particulier (normalisation)[21].

De mon point de vue, il n'y a donc concept que si :

- il existe des conditions d'énonciation communes à un ensemble de locuteurs qui permettent de neutraliser les éléments propres à ce locuteur ; on a pu parler d'un locuteur collectif dans le courant de l'analyse de discours,[22] ou de communautés de locuteurs (Gaudin, 1995);
- un interprétant (terminologue, linguiste, documentaliste, expert…), ayant un objectif précis, et qui, à partir du constat de régularités « immanentes »

[21] Je monterai dans le chapitre IV que ce processus de « normalisation » concerne en réalité l'interprétation de l'ensemble du texte. La normalisation terminologique n'est ainsi qu'un aspect d'un processus beaucoup plus complexe qui peut concerner la totalité des phénomènes textuels et qui vise à construire une cohésion globale.

[22] « Le concept de communauté discursive, en tant qu'institution qui reçoit sa cohérence de ses pratiques discursives, quelle que soit la nature de son organisation sociale et technique est probablement de nature à fonder des analyses de discours autres que monographiques, puisqu'il assure la constitution d'espaces discursifs structurés par des instances de production et de diffusion repérables » (Beacco et Moirand, 1995, 49).
« Le locuteur collectif désigne le groupe social (groupes politiques, religieux, syndicaux, etc.) partageant un certain type de culture et produisant un discours qui apparaît comme celui de toute la comunauté » (Dubois et al., 1994, 289).

(normaison), attribue à certains des signifiés d'un texte, le statut de concept (normalisation) ; il y a donc passage d'un système sémiotique à un autre, relevant tous les deux du linguistique.

Ainsi, si l'idée de concept est à retenir dans le cadre de l'analyse de corpus spécialisés, c'est, me semble-t-il, sous deux conditions :

- les concepts ne sont pas préexistants mais construits par un interprétant,
- les manifestations linguistiques à partir desquelles sont établis les concepts ont en commun des caractéristiques liées à la situation d'énonciation.

La notion de concept est donc à corréler avec celle d'interprétation qui se reconnaît comme telle. La particularité de l'interprétation du sémanticien de corpus est qu'elle doit être la plus consciente possible ; elle doit tenir compte de tous les éléments en sa connaissance, y compris et peut-être surtout de son objectif propre d'interprétation. En effet, le linguiste de corpus analyse du matériau textuel qui, le plus souvent, ne lui était pas destiné ; il y a donc entrecroisement d'intentions diverses : celle des rédacteurs des textes, celle des lecteurs à qui ils étaient destinés et enfin celle du linguiste qui est guidé par un objectif précis. D'une certaine façon, il ne peut y avoir concept pour l'interprétant linguiste que lorsqu'il y a *conscience* du passage d'un système sémiotique à un autre. Le rôle du linguiste est alors d'expliquer comment se fait ce passage : à partir de quels éléments, avec quelle élaboration, pour quels objectifs… C'est en ce sens que l'interprétant linguiste se distingue d'un autre type d'interprétant : terminologue, ingénieur de la connaissance… : il cherche à justifier ses choix par les différents éléments qui les influencent (connaissances linguistiques, régularités internes, objectif de la modélisation), avec l'objectif de pouvoir dégager des régularités. Ainsi s'amorce une véritable linguistique de corpus[23].

Enfin, on peut dire que, d'une certaine façon, le sens, tellement labile dans son dynamisme, ne se laisse approcher que par la conceptualisation, c'est-à-dire par la création de concepts, au sens où je les ai définis ci-dessus. Cette création s'accompagne d'un acte de définition, c'est-à-dire de maîtrise du sens (d'un sens).

La non-préexistance des concepts est clairement un postulat (tout comme l'est leur préexistance). Ce choix théorique est le fruit d'une réflexion menée autour d'un ensemble de questions :

- Si les concepts préexistent, quelle est leur nature ? On a pu parler d'une nature perceptive, qui permettait de justifier l'idée d'universaux communs à tous les hommes (tous les hommes ayant le même type de fonctionnement sensoriel) ; il me semble que c'est faire peu de cas de la dimension culturelle et expériencielle, certainement au moins aussi présente dans la langue que la dimension perceptive ;
- Si les concepts préexistent, comment peut-on y accéder, comment vérifier leur existence ? Les chercheurs en sciences cognitives en particulier pensent prouver l'existence de ces universaux en montrant que le même type de lexicalisation existent dans des langues différentes. Ce type de fonctionnement est certainement avéré pour certains éléments, par exemple, il est certainement question de temps et d'espace dans toutes les langues. Mais de telles preuves sont-elles généralisables à

[23] Ainsi, contrairement à ce que pensent parfois les informaticiens qui construisent leur modèles à partir de corpus, il ne suffit pas de travailler à partir de textes pour avoir une approche linguistique ; il me semble qu'il n'y a linguistique de corpus que lorsque l'analyse de corpus se fait dans la perspective d'une insertion des questionnements dans l'histoire et les problématiques qui ont constitué la discipline linguistique. Cependant, la linguistique de corpus n'a pas encore défini un (des) modèle(s) et une théorie qui permettent de lui donner toute sa place dans la linguistique

l'ensemble des éléments linguistiques, pour toutes les langues et pour toutes les variations d'usages de ces langues ?

– Tout postulat s'accompagne d'une idéologie plus ou moins consciente. Il me semble que derrière le postulat des universaux (des concepts préexistants) se cache l'idée de la langue/moyen de communication idéal et moyen de reconnaissance entre êtres humains. C'est méconnaître :
    - que la langue n'est pas le seul moyen de communication, la dimension interactionnelle intervient de bien d'autres manières,
    - que les discours sont chargés d'histoire et de culture, collectives ou individuelles, qui interviennent dans la construction du sens,
    - que la volonté de communication ne doit pas être toujours associée à l'idée d'une transparence totale entre les interlocuteurs.

Les réflexions que m'ont inspirées ces questions m'ont amenée au choix d'une approche qui étudie la terminologie à partir d'usages réels et qui élabore des concepts à partir de ces usages, par une interprétation qui tient compte à la fois des régularités spontanément à l'oeuvre dans ces usages et des besoins qui ont conduit à la mise en place de l'analyse (besoin d'utilisateurs ou hypothèses linguistiques).

2.2.2 Concept et formalisation de la connaissance

Ce même terme de concept est utilisé en intelligence artificielle pour tous les langages de représentation qui utilisent des réseaux sémantiques (graphes conceptuels, logiques de description…). Il est donc clairement associé à une mise en relation ; or, la représentation relationnelle vient se substituer, en terminologie, à la définition sous forme discursive. La parenté des deux approches est donc avérée. Cependant, le point de vue formel de l'IA amène un certain nombre de contraintes qui ont une grande influence sur la façon de concevoir le concept. Deux éléments caractérisent le concept en IA : le fait qu'il soit associé à la perception et le fait qu'il soit associé à des éléments logiques.

Concept et perception

La théorie des graphes conceptuels est fortement reliée, à l'origine, à la psychologie de la vision :

> « Bien qu'ayant de nombreuses sources… l'origine des idées de Sowa peut être située dans la psychologie de la perception » (Sabah, 1988, 228).

N'oublions pas aussi que les réseaux sémantiques eux-mêmes ont été constitués par un psycholinguiste, Quillian, pour rendre compte de la mémoire sémantique.
Cette justification de la formalisation en réseaux par la psychologie n'est pas sans rappeler la vision des terminologues classiques, qui considèrent que les concepts préexistent, puisque la perception préexiste. Comme je l'ai déjà mentionné, c'est en partie sur cette base que s'est faite la rencontre originelle entre terminologie et IA.
Il faut noter que, depuis la création du concept de BCT (peut-être grâce à cette création), une évolution très nette, parallèle à celle de la terminologie, s'est faite en IA qui a conduit à ce que cette vision soit elle aussi remise en question et remplacée par une approche qui prend les corpus pour référence.

Concept et logique

Etant donné que l'objectif de l'IA est de faire raisonner à des machines, la présence d'une vision logique des concepts est omniprésente chez les informaticiens :

« Many researchers…have chosen to identify the notion of a concept with the notion of a predicate in first-order logic » (Woods, 1991, 48).

Or, ce nécessaire lien avec la logique, qui suppose que soient respectés des critères de complétude et de cohérence, amène à une normalisation supplémentaire. Prenons le cas de la constitution de taxinomies. Lorsque ces taxinomies sont construites à partir d'un corpus qui sert de référence, le linguiste s'en tient en principe aux éléments qui justifient qu'il identifie des relations d'hyperonymie (en fait, des marqueurs de ces relations). Or, cette approche ne garantit ni la cohérence (il se peut, surtout sur des corpus volumineux, que des points de vue différents fassent apparaître des hiérarchisations différentes et pas nécessairement compatibles) ni la complétude. Il est ainsi très fréquent que les hiérarchies construites ne s'organisent pas en une seule et unique taxinomie avec une racine (un « top »), élément constituant l'origine de la hiérarchie, parfaitement identifié (*cf.* par exemple (Bowker, 1997)). Si bien que les critères de constitution d'une taxinomie généralement préconisés (principe de communauté avec le père, principe de différence avec le père, principe de différence avec les frères, principe de communauté avec les frères (Bachimont, 1995, 78)) ne peuvent pas et ne doivent pas même être suivis, en tout cas dans un premier temps, par l'analyste de corpus. C'est seulement dans une étape de normalisation informatique que le respect de ces critères logiques peut être envisagé. Notons d'ailleurs que, comme le souligne Bachimont, bien que logiques, ces choix de formalisation entretiennent, eux aussi, une parenté forte avec l'objectif de la formalisation.
Ainsi, après la normaison et la normalisation linguistique, la normalisation informatique permettrait de passer d'un corpus à un système formel.

Finalement du corpus à une BCT formelle, deux types d'interprétation sont à l'œuvre, l'une qui permet de modéliser, c'est-à-dire de décontextualiser les fonctionnements linguistiques (ce qui constitue un premier type de normalisation), l'autre qui permet de formaliser, c'est-à-dire de donner un statut logique à cette représentation formelle (et qui constitue un second type de normalisation). Dans la plupart des cas, pour plus d'efficacité, lorsque les ontologies sont constituées par des informaticiens, les deux étapes sont confondues en une seule. Il revient aux chercheurs en linguistique de mener des études approfondies sur la première étape, à la fois parce qu'elle permet d'éclairer des fonctionnements sémantiques et parce qu'elle permet de justifier (ou non) des raccourcis effectués par les informaticiens.

Concept et terme

Parce qu'il est un signifié normé, le concept est d'abord un signifié, le terme étant, lui, un *signe* normé. Par signe normé, on peut entendre un signe discursif auquel on a donné un statut particulier. Cela signifie que n'importe quel signe discursif peut devenir un terme, dans la mesure où il peut avoir un sens dans une interprétation contextualisée[24].
Dans le modèle de BCT qui est décrit dans le chapitre III, la distinction terme/concept est mise en place d'une autre façon : comme dans la conception terminologique la plus courante, elle sert alors à rendre compte de phénomènes comme la polysémie ou la synonymie, ce qui ne va pas sans poser de questions.

[24] C'est parce que l'on est dans un contexte d'interprétation particulier (qui concerne surtout la dénomination), qui s'effectue sur des corpus particuliers (des corpus spécialisés), que les termes les plus fréquemment repérés sont des noms ou des groupes nominaux. Mais on pourrait imaginer des situations d'interprétation de ce même type de corpus, pour lesquelles les termes seraient des éléments morphologiques (par exemple dans les cas de recherche d'information).

2.3. Prise en compte de l'usage

Le concept de BCT ne s'accompagne pas nécessairement d'une prise en compte de l'usage réel des termes. Comme on vient de le voir les deux éléments principaux qui caractérisent les BCT sont d'une part la mise en réseau des concepts et d'autre part la distinction terme/concept. D'une certaine façon, les BCT sont une évolution des BDT (Bases de Données terminologiques) qui étaient, elles, souvent élaborées par des entretiens avec des experts. Toutefois, il est évident que cette évolution s'est accompagnée d'une réflexion sur les méthodes de constitution et que les grands projets que j'ai évoqués (QUIRK et CODE) proposent tous de constituer les BCT à partir de données textuelles. Des outils pour assister cette opération sont d'ailleurs souvent constitués en même temps que les modèles de données et les outils de stockage et d'interfaçage de ces données. Ce passage d'une méthode basée sur des entretiens avec des experts à une méthode basée sur l'analyse de données textuelles n'a pas toujours été bien problématisé. Les textes utilisés, de toute nature, sont ainsi souvent considérés comme un amas de données réelles dont on parvient toujours à obtenir des éléments pertinents. Le lien entre le corpus et les données que l'on en obtient n'est pas clairement établi, la question de la constitution du corpus n'est pas évoquée. Enfin, la technique des marqueurs, souvent mise en œuvre pour repérer les relations, n'est pas examinée par rapport aux spécificités des corpus étudiés.
L'originalité des travaux qui sont menés à Toulouse se situe justement dans cette problématique qui consiste à s'interroger sur le rôle du corpus dans l'élaboration des données (Condamines, 2000). Dans les prochains chapitres, je montrerai comment nous avons abordé cette question pour ce qui concerne les données terminologiques et quelles propositions nous avons pu formuler pour élaborer une méthode complète d'analyse qui va de la constitution du corpus jusqu'au repérage des données pertinentes.

On voit que la principale innovation des BCT a consisté à proposer un modèle alternatif à celui des BDT ; cette évolution s'est accompagnée d'un rapprochement avec l'informatique et particulièrement la représentation des connaissances, au point que les BCT sont parfois considérées comme des ontologies (au sens de l'ingénierie des connaissances). Cette confusion, même si elle peut avoir une certaine pertinence, suppose que les points de vue et les problématiques respectifs qui président à la constitution de l'un ou l'autre produit soient éclaircis.

3. BCT et ontologies

3.1. BCT et ontologies : un problème similaire ?

Le débat sur les liens entre BCT et ontologies (c'est-à-dire entre modélisation d'un point de vue lexicologique et formalisation d'un point de vue « ingénierie des connaissances ») est apparu de manière assez récente. En ingénierie des connaissances[25], la constitution des ontologies est devenue l'un des thèmes qui suscitent le plus de travaux ; citons par exemple les travaux de Guarino (Guarino, 1995) ou ceux de Gomez-Perez (Gomez-Perez, 1999). Notons que le choix du terme « ontologie », emprunté à la philosophie, n'est pas des plus heureux ; il laisse supposer que les modèles que traite l'informatique ont à voir avec l'étant de

[25] Ce terme est venu se substituer à celui d'intelligence artificielle, en tout cas sur les questions de gestion de la connaissance ; en effet, à la suite d'un certain nombre d'échecs de systèmes visant à substituer le raisonnement de la machine à celui d'un humain, l'IA a évolué vers des systèmes d'aide au raisonnement, qui prennent en compte très tôt le besoin de l'utilisateur (*cf.* Charlet et al., 2000) et qui, ce faisant, s'apparentent à un processus d'ingénierie tout en ayant une approche scientifique beaucoup plus pragmatique face aux types de problèmes à résoudre.

la philosophie, il est d'ailleurs probable que ce mythe constitue encore le moteur de beaucoup d'informaticiens qui travaillent sur la représentation des connaissances. Si la dimension relationnelle est constante dans la définition d'une ontologie, certains éléments peuvent apparaître ou non dans les définitions.

La dimension formelle

Dans la plupart des cas, cette dimension est présente puisque le terme d'ontologie est associé à l'informatique. Les modélisations utilisées sont le plus souvent des langages inspirés des réseaux sémantiques « à la Quillan », comme les graphes conceptuels ou les logiques terminologiques. Mais chez certains auteurs, proches de la linguistique, cette dimension ne semble pas fondamentale (« par ontologie, on entend ici un ensemble de concepts d'un domaine structuré par des relations » (Frath et *al.*, 2000, 292)). La distinction BCT/ontologie n'existe plus du tout dans ce type d'approche.

Le lien avec un domaine

Cette question fait actuellement l'objet d'un enjeu majeur. De nombreux chercheurs sont à la recherche d'une ontologie générale, qui serait indépendante d'un domaine et donc, en principe, réutilisable pour n'importe quelle application, cela a été le cas par exemple de Cyc, « the Intelligent Encyclopaedia Project 1983-1993. » (Lenat et *al.*, 1990). Guidés par leur désir de constituer des données une fois pour toutes, ce qui présente aussi un avantage économique majeur pour les financeurs, les chercheurs qui participent à ces projets ne mènent pas une véritable réflexion sur la réutilisabilité. Ils considèrent souvent comme réutilisation le fait de conserver pour une nouvelle application des éléments de très haut niveau dans l'ontologie : les principales classes d'éléments. On peut se demander alors si le gain de temps et d'argent est réel. Il est probable que le fait d'utiliser une ressource déjà constituée, par d'autres chercheurs, a un rôle de sécurisation et sa mise en œuvre donne l'impression d'une valeur ajoutée et cumulative.
Or, même par domaine, la constitution d'ontologies semble difficile : des chercheurs ont pu montrer que la réutilisation d'une ontologie sur le domaine médical pour une application particulière s'avérait très peu utile voire impossible (Charlet et *al.*, 1996).

Une tendance plus récente, que le groupe TIA a contribué à élaborer, vise à considérer qu'une ontologie n'a de pertinence que si elle est associée non seulement à un domaine mais aussi à une application :

> « Une ontologie régionale reflète une structure seulement valable pour accomplir des actions dans un domaine donné… » (Bachimont, 2000, 316).

Se pose alors la question de la réutilisation des modèles construits, réutilisabilité qui, on vient de le voir, représente la principale motivation des travaux sur les ontologies générales. S'ils ne sont construits que pour une application, ces modèles risquent de ne plus avoir de pertinence pour une nouvelle application et il peut être plus efficace de construire un autre modèle pour une nouvelle application, la capitalisation des connaissances ne se fait plus alors à travers des modèles de connaissances mais à travers des méthodes de constitution de ces modèles, le plus souvent à partir de corpus (ce qui rend cruciale la question de la constitution des corpus). Cette position est clairement défendue par les membres du groupe TIA engagés dans la constitution d'ontologies ou de BCT :

> « We assert that text analysis could significantly improve the efficiency of the process of ontology building, as well as the quality and the relevance of the resulting ontologies ». (AussenacGilles et *al.*, 2000, 3);

cf. aussi (Slodzian, 2000).

Méthode de construction : descendante *vs* ascendante

La méthode de construction des ontologies constitue un autre élément distinctif dans les travaux existants ; élément qui est d'ailleurs en continuité avec la question du lien des ontologies avec un domaine et une application. Grossièrement, deux méthodes sont envisagées, l'une qui permet de construire une ontologie par introspection, méthode descendante (« top-down), l'autre qui permet de la construire à partir de données réelles, méthode ascendante (« bottom-up »). La méthode « tout introspection » est revendiquée pour construire des ontologies « générales » qui permettent de rendre compte de connaissances « naïves » sur le monde, c'est-à-dire partagées par l'ensemble des humains, censés avoir la même perception du monde. Dès qu'il s'agit de traiter de domaines particuliers de connaissances, cette méthode ne peut plus même être envisagée puisque les ingénieurs de la connaissance n'ont le plus souvent pas de compétence sur le domaine et donc pas d'intuition pour le modéliser. Il faut alors faire appel à des experts du domaine (qui vont expliciter les concepts qu'ils manipulent) ou/et à des textes rédigés par eux, supposés contenir cette connaissance experte. L'interrogation directe des experts ayant montré ses limites, la tendance est maintenant, pour les tenants de l'approche ascendante de focaliser les travaux sur l'analyse de corpus. C'est précisément sur ce thème-là que se fait la rencontre avec la linguistique et la terminologie de corpus. On peut dire ainsi que le principal élément de comparaison entre ontologie et BCT consiste en une mise en réseau des relations ; les méthodes de constitution des données à partir de corpus permettent, elles, la rencontre entre ingénierie des connaissances et terminologie textuelle. Les ontologies se situent du côté informatique au sens où elles prennent en compte la dimension formelle ; les BCT, en tout cas lorsqu'elles sont élaborées à partir de corpus, se situent plutôt du côté de la terminologie et de la linguistique.

3.2. Ontologies, BCT et TAL

Il est particulièrement fructueux de s'interroger sur les méthodes et les objectifs de sa discipline lorsqu'on les examine à partir de problématiques traitées par des disciplines connexes, avec des objectifs et des méthodes différentes.

On l'aura remarqué, de nombreuses questions qui se posent à propos des ontologies rappellent des problèmes qui pourraient concerner non seulement la constitution des BCT, mais aussi la constitution des données lexicales en TAL et, plus généralement l'analyse lexicale à partir de corpus.

Rôle du domaine

Tout comme en ingénierie des connaissances, il y a, en TAL, des tentatives de constitution de grands réseaux sémantiques, censés représenter le sens d'une langue voire le sens en général. C'est le cas du projet EUROWORDNET, qui se définit comme une suite du projet WORDNET. Il faut reconnaître que les résultats du projet WORDNET (Fellbaum, 1999), à l'origine mené dans une perspective psycholinguistique, ont été souvent utilisés comme ressource dans le traitement automatique de la langue ; mais l'évaluation du supposé gain de temps réalisé par rapport à une ressource qui serait construite directement pour un nouveau projet n'a pas été faite, à ma connaissance. En tout cas, les informaticiens ont manifesté un tel intérêt pour ce type de ressource que la version européenne de ce réseau sémantique, EUROWORDNET, a été, elle, d'emblée orientée vers le TAL, l'anglais jouant pratiquement le rôle de langue pivot, par rapport à laquelle se structurent les autres langues européennes.

Du point de vue de la méthode de constitution, EUROWORDNET n'est pas très éloigné d'un dictionnaire de langue classique avec une modélisation des données (en particulier des relations) beaucoup plus systématique, en fait surtout plus systématiquement utilisable d'un

point de vue informatique. Comme avec les ontologies générales, on trouve, à la base de la construction de ces réseaux sémantiques généraux la volonté d'investir une fois pour toutes dans des travaux dont les résultats seront réutilisables pour toutes sortes d'applications (c'est par millions d'euros que ces projets sont financés). Cette position est d'autant plus étonnante que beaucoup d'expériences en traitement automatique de la langue (TAL) ont montré, dans les années 80, que les systèmes de TAL (traduction, dialogue…) ne fonctionnaient bien que sur des domaines restreints, parfaitement circonscrits du point de vue linguistique et pour lesquels une caractérisation lexico-syntaxique préalable s'avérait indispensable par exemple les projets de dialogue Homme/machine du CRIN (Centre de Recherche en Informatique de Nancy).

Introspection *vs* analyse de données

Les possibilités offertes par la grande quantité de corpus désormais accessibles viennent ébranler un certain nombre de certitudes qui fondent la constitution des données, aussi bien en ingénierie des connaissances qu'en TAL ou en sémantique. Fondamentalement, ce qui est interrogé est le rôle de l'intuition et de l'introspection. Tout comme en ingénierie des connaissances, le fait de travailler sur des domaines spécialisés en TAL ou en sémantique rend ce problème incontournable : le recours à la seule introspection est impossible. Et tout comme les ingénieurs de la connaissance, les linguistes doivent alors se tourner vers les données réelles, des corpus de documentation d'un domaine. Sur ce terrain de rencontre, linguistes et informaticiens peuvent collaborer et discuter de méthodes et d'outils. Pour le linguiste cependant, le recours à des corpus amène un ensemble d'interrogations et finalement une véritable mise en question de ce qui fonde sa discipline, tout particulièrement dans le domaine sémantique. En effet, si pour l'informaticien, il y a un enjeu méthodologique à défendre une approche ascendante plutôt qu'une approche descendante, l'enjeu pour le linguiste est essentiel car il s'agit de prendre position sur la question du sens et des possibilités qu'ouvre l'utilisation de corpus réels. J'ai déjà évoqué cette problématique dans le chapitre I.

Ainsi, si BCT et ontologies entretiennent une parenté très nette, chacune s'inscrit dans un paradigme disciplinaire différent, qui fait intervenir des histoires, des méthodes et des projets différents. Nourris par les apports de l'interdisciplinarité, les chercheurs qui, travaillant sur l'élaboration de données lexicales, acceptent l'ouverture vers des disciplines proches peuvent contribuer à renouveler la manière de poser les questions dans chaque discipline d'origine… A condition peut-être que chaque point de vue maintienne son originalité et sache situer son apport par rapport à sa discipline d'origine. C'est une des raisons qui nous ont amenés à redéfinir le concept de BCT pour rendre compte de son double rôle, à l'interface de la linguistique de corpus et de l'ingénierie des connaissances.

3.3. BCTCorpus et BCTApplicative

L'étape la plus récente de l'évolution des BCT a consisté, en tout cas dans notre vision toulousaine, à redécouper le concept de BCT. En effet, les travaux que nous avons menés sur des corpus et les discussions qui en ont découlé nous ont amenées, Nathalie Aussenac-Gilles et moi-même, à l'idée que la notion de BCT devait être réorganisée en deux types de modélisation, correspondant à deux étapes, l'une plus proche du corpus, ui se manifeste par une modélisation appelée BCTcorpus, l'autre plus proche de l'application informatique, c'est-à-dire aussi plus formelle, ui se manifeste par une modélisation appeléé BCTapplicative, et qui se situe clairement dans la perspective des ontologies de l'ingénierie des connaissances. En effet,

« L'organisation formelle dans l'ontologie met l'accent sur la classification et la différenciation des concepts entre eux, au sein d'une hiérarchie de types ou de classes. Or, ce type de critère n'est pas présent pour décider de l'organisation des concepts par les linguistes, tout simplement parce que le texte lui-même ne reflète pas systématiquement une telle hiérarchie ». (Aussenac Gilles et Condamines, 2001, 159).

En revanche, le maintien du terme BCT dans chacune des dénominations rappelle la très forte parenté qui unit les deux concepts, l'un prenant le relais de l'autre dans une élaboration qui va d'un corpus à une ontologie (donc à un modèle formel).
Cette distinction permet aussi de caractériser les deux types de concepts que j'ai présentés en 2, l'un proche du fonctionnement en corpus, qui s'élabore par normalisation linguistique, l'autre qui permet une utilisation dans les représentations informatiques (normalisation informatique)[26].
Enfin, cette répartition permet d'examiner les questions du point de vue de chacune des disciplines. Il est souhaitable en effet que chaque discipline s'interroge sur la façon dont sa collaboration avec l'autre amène un éclairage nouveau sur des questions jusque là traitées d'un seul point de vue. Ce retour à une distinction des résultats et des méthodes peut être considéré comme la fin d'une relation « symbiotique ». Mais chaque discipline a su tirer parti de cette période fusionnelle et la collaboration peut se faire maintenant sur des bases plus claires.
Il semble évident que la constitution des BCT interroge la linguistique sur deux points majeurs :

- le rôle de l'analyse de corpus dans l'étude du sens, et dans le même courant le rôle des outils informatiques dans l'analyse de corpus,
- le passage d'un système discursif à un système sémantique abstrait (asyntaxique). La réflexion sur ce type d'élaboration complexe constitue comme un fil conducteur dans ma réflexion puisque dès 1988, je me suis interrogée sur ces questions de passage du linguistique au conceptuel et sur les complémentarités entre linguistique, TAL et IA (Condamines, 1988).

3.4. BCT et linguistique

Pour l'ingénieur de la connaissance, il n'est pas trop difficile de se situer dans une des approches : ontologie générale ou ontologie régionale (même si la première correspond à la vision encore largement dominante) dans la mesure où l'évaluation de la pertinence de ces concepts se fait le plus souvent (ou devrait se faire) en termes d'efficacité.
Pour le linguiste, la question se pose de manière différente. La notion de système est fondamentale en linguistique : le repérage de régularités qui mettent de l'ordre dans le flux des phénomènes linguistiques est à l'évidence ce qui fonde la linguistique comme science et ce qui motive profondément le chercheur en linguistique. Mise en ordre va ici avec maîtrise : il s'agit de maîtriser des phénomènes qui menacent sans cesse d'échapper à notre contrôle. L'informatique vient parfaitement cadrer avec ce projet de maîtrise puisqu'il faut lui fournir des données parfaitement calibrées, parfaitement ordonnées et donc maîtrisées. L'informatisation vient parfois ainsi comme la confirmation du fantasme d'un contrôle possible du sens. La rencontre avec la terminologie « classique » s'est certainement faite, à l'origine, sur le même espoir de contrôle du sens. Comme le note Rey d'ailleurs (Rey, 1979), *définition* et *terme* comportent le même trait de clôture, englobement que l'on peut, selon les cas, interpréter comme limitation ou comme possibilité d'échange.
La rencontre entre IA et terminologie sur le concept de BCT menaçait de se transformer en justification réciproque d'illusions, la prise en compte des corpus, incontournable dans des

[26] Dans la suite de ce mémoire, c'est par abus de langage que je parlerai de BCT au lieu de BCT Corpus.

domaines spécialisés, est venue comme une réalité frustrante mais certainement salutaire qui a conduit à des interrogations majeures à la fois en ingénierie des connaissances et en linguistique.
Inévitable en terminologie, la question des corpus atteint maintenant la linguistique dans ce qu'elle a de plus essentiel : le sens. En effet, si l'on accepte de ne plus considérer la terminologie au mieux comme relevant de la linguistique appliquée, au pire comme une simple technique, alors, les questions qu'elle soulève viennent ébranler la linguistique théorique à tous les niveaux qu'elle aborde et particulièrement au niveau sémantique. En charriant avec elle une longue pratique des corpus (à défaut d'une théorisation), indispensable dans les domaines qu'elle a à traiter, la terminologie vient interpeller la linguistique sur un problème longtemps volontairement marginalisé : l'utilisation des corpus, c'est-à-dire d'usages réels de la langue. Corbin, dès 1980 (et sans doute avant lui, à leur manière, les sociolinguistes) avait pointé certains des abus d'une linguistique uniquement introspective. Ses inquiétudes sont maintenant d'une brûlante actualité.
Bien sûr, de nombreuses disciplines, et depuis longtemps, prennent les corpus pour matériau d'étude. On peut citer par exemple :

- la linguistique historique et comparative,
- l'analyse littéraire,
- la sociolinguistique,
- l'analyse de discours,
- le traitement automatique de la langue (en tout cas une partie des travaux).

Dans ces approches, le sens apparaît nécessairement en filigrane soit parce que, considéré comme stable, il soutient l'analyse syntaxique et morphologique et permet d'étudier la variation, soit parce que, insaisissable par hypothèse, il est toujours le fruit d'une construction et que seules les modalités de cette construction sont accessibles ; on ne perçoit donc du sens que ses effets, psychologiques ou sociaux.
Mais il est rare que le sens soit mis au cœur de la problématique et qu'une réflexion soit menée pour mettre en lumière la place du sens dans l'élaboration des résultats d'analyse à partir de corpus.
Il me semble que la **compétence linguistique** (j'entends, de locuteur) est construite à partir des expériences que peut avoir un individu en tant que locuteur/auditeur d'une langue ; cette compétence est sans doute hétérogène[27] :

> « …pour [Labov], nous sommes des sujets entendants, tout autant et peut-être plus que parlants : nous nous constituerions notre " compétence hétérogène " par les traces de nos confrontations constantes à des productions elles-même non unifiées » (Gadet, 1992, 10).

La **compétence de linguiste** me semble, quant à elle, relever d'une tentative pour essayer de comprendre quelle part de connaissance sémantique est convoquée *a priori* dans les études de corpus et quelle part est construite par ces études. Au fond, il ne s'agit pas d'opposer introspection (qui met en œuvre une connaissance pré-existante, relevant de la compétence de locuteur et de la compétence de linguiste) et construction par analyse de corpus mais d'examiner comment, pour chaque nouvelle étude, l'une et l'autre se mettent en œuvre.
La difficulté majeure dans l'analyse de corpus consiste à repérer les phénomènes qui fonctionnent sur un mode attendu et ceux qui sont propres à un corpus ou pertinents pour une certaine utilisation. L'analyse de corpus a ceci de particulier qu'elle nécessite à la fois de faire appel à son intuition, et à sa compétence de linguiste et à s'en méfier en permanence pour arriver à rendre compte au plus juste du fonctionnement linguistique propre à un corpus ou à un besoin en lien avec ce corpus.

[27] Au sens où je l'entends, cela ne signifie pas qu'on ne peut pas faire l'hypothèse qu'elle est proche, d'un locuteur à l'autre, *cf.* chapitre IV, 1-3.

Finalement, le rôle du linguiste qui fait de l'analyse sémantique à partir de corpus me semble être de diverses natures :

- trouver des régularités pertinentes pour ce corpus et/ou une application donnée, en faisant appel à sa compétence de locuteur et de linguiste,
- essayer d'ouvrir les résultats et de voir comment le contexte, au sens large, influence la mise au jour de ces régularités,
- définir des méthodes permettant de repérer les fonctionnements propres à un corpus.

Je reviendrai sur ces questions à propos de la constitution effective de bases de connaissances terminologiques. En effet, la constitution de BCT met le linguiste au pied du mur puisqu'elle lui impose de se confronter avec la réalité des corpus et des utilisations réelles qui vont être faites des résultats qu'il produit, ce qui l'oblige à pratiquer une linguistique située. En retour, cette confrontation est riche d'enseignements et elle permet d'éclairer d'un jour nouveau nombre de phénomènes sémantiques.

4. Conclusion

Après la présentation du contexte théorique (sémantique) dans lequel je me place (chapitre 1), ce chapitre m'a permis de situer plus précisément ma réflexion, qui s'élabore dans la perspective de la constitution des bases de connaissances terminologiques à partir de corpus. Ce contexte suppose non seulement une réflexion sur l'élaboration d'une représentation à partir d'un corpus mais aussi sur les rapports entre terminologie textuelle et outils d'analyse et de représentation. Cette pluridisciplinarité est en effet incontournable lorsque l'on évoque les BCT mais il est aussi nécessaire d'étudier comment cette problématique commune prend sens du point de vue de chacune des disciplines concernées. Un autre mode d'interdisciplinarité dans la constitution de BCT se réalise dans la prise en compte des outils dans l'analyse de corpus. Ce thème constituera une des principales parties du prochain chapitre qui commencera à baliser le processus d'élaboration d'une BCT à partir d'un corpus.

Chapitre III
Constitution de bases de connaissances terminologiques : Expérimentations, théorisation

De nombreuses questions se posent lorsqu'on se donne pour but de passer d'un corpus à une modélisation de ce corpus. Il faut déterminer quel type de modélisation on vise, quel corpus servira de point de départ et avec quelles méthodes s'élaborera la modélisation. Lorsqu'on se place du point de vue de la linguistique, il est évident que c'est ce troisième élément qui focalise l'attention. Pour le chercheur, l'objectif est alors de repérer des régularités de fonctionnement dont il faut mesurer la dépendance avec le corpus étudié et avec l'objectif visé. La recherche de régularités permet à la fois de généraliser les résultats afin de les rendre utilisables pour d'autres types d'études et afin de les transmettre (à des étudiants par exemple). Pourtant, la focalisation sur le processus de modélisation qui paraît, seule, relever des préoccupations du linguiste ne doit pas faire complètement délaisser les questions qui peuvent se poser à propos du matériau étudié (le corpus) et du résultat recherché (en lien avec un modèle).

En effet, la mise en œuvre d'une approche doublement située, telle que je l'ai présentée en I, prend corps de deux manières dans la constitution de BCT. Il s'agit d'une part de construire un modèle de données qui prenne en compte l'objectif (ou les objectifs) de l'étude et, d'autre part, d'établir des critères d'élaboration du corpus qui prennent en compte tout à la fois l'homogénéité de genre textuel et de pertinence par rapport à l'objectif d'étude.

Ce chapitre III, appuyé sur des exemples réels, montre ainsi comment on peut jalonner la réflexion qui permet d'élaborer une représentation à partir d'un corpus. Dans la première partie, la présentation du modèle de données montrera les choix qui ont été faits et évoquera la manière dont ces choix peuvent contraindre l'interprétation. Une réflexion sur la constitution du corpus m'amènera à donner une définition affinée de la notion de représentativité. Dans la deuxième partie, un premier balayage des questions que pose l'élaboration des données d'une BCT à partir de corpus sera proposé. Enfin, la troisième partie présentera les outils de TAL qui peuvent exister pour assister cette construction et évaluera les possibilités et les limites de leur utilisation.

1. Du corpus à la BCT : mise en place et problématiques

L'élaboration d'une BCT à partir d'un corpus demande que des questions d'objectif et de méthodes soient parfaitement clarifiées, c'est-à-dire des questions qui concernent la mise en place de l'étude, qui joue un rôle de balisage fondamental pour la suite de l'analyse. Ces questions sont abordées ici dans trois paragraphes qui concernent la constitution du modèle de données, la question de la représentation des relations sémantiques et la constitution du corpus.

1.1. Constitution d'un modèle de données

Le modèle de données va servir de guide à l'analyse puisqu'il va déterminer *a priori* le type des éléments que l'on va chercher en corpus et la façon de les représenter. Il ne s'agit pas à ce niveau de définir un modèle spécifique à une application mais plutôt un modèle général qui permette de rendre compte des phénomènes lexicaux les plus courants : ellipse, polysémie, synonymie, homonymie, et bien sûr, relations conceptuelles. Comme dans tout modèle, les fonctionnements sont figés, voire contraints par la représentation choisie. Mais ce modèle sert aussi à stabiliser les résultats et à leur donner une cohérence, à un moment donné. Le modèle que nous[28] avions constitué pour le projet EUROLANG à Aramiihs (Condamines et Amsili, 1993) est encore d'actualité même s'il a été quelque peu modifié pour s'adapter aux différents projets et outils, par exemple la constitution de l'outil support par une équipe de l'Irit (Aussenac-Gilles et Séguéla, 1999), et aussi aux différentes difficultés d'utilisation rencontrées par les linguistes utilisateurs. En voici une schématisation.

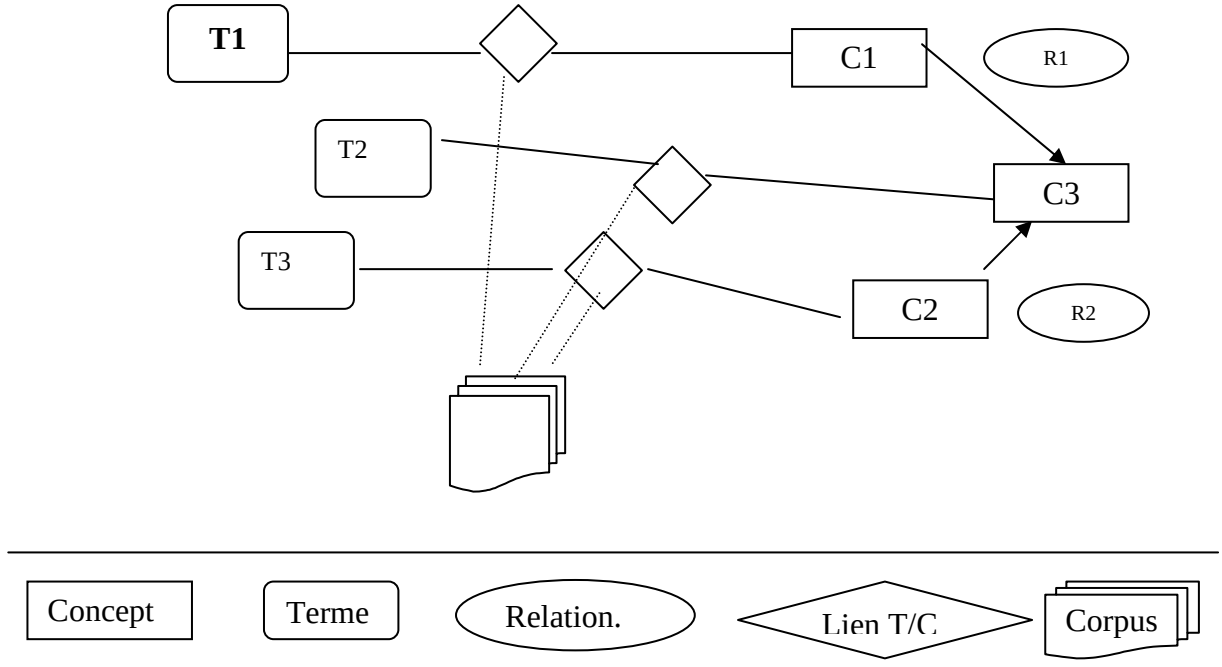


Figure 1 : Modèle de BCT

[28] Le "nous" que j'utilise dans ce chapitre me permet d'associer dans cette présentation tous les collègues et collaborateurs qui ont travaillé sur différents projets de constitution de BCT : P. Amsili, J. Rebeyrolle, A.-M. Soubeille, M.-P. Jacques, N. Aussenac-Gilles, P.Séguéla, J. Feliu, I. Orlac, P. Monnier, E. Lecorgne, D. Fournier, J. Mell, F. Zerguini

Ce modèle est organisé en trois champs principaux : le concept, le terme et le lien terme/concept.
En réalité, ce type de modélisation n'est pas extrêmement rigoureux du point de vue linguistique. Il sert surtout d'une part à rendre compte de phénomènes sémantiques comme la synonymie, l'homonymie et la polysémie et d'autre part à maintenir un lien important entre la modélisation et le texte. Mais ce choix de modélisation laisse entière la question de la nature **linguistique** du lien entre le terme et le concept. En effet, les relations sémantiques s'établissent en principe de signe à signe, sans intervention d'un autre élément. L'utilisation du concept pour rendre compte de ces phénomènes constitue une facilité de représentation qui a le mérite, en tout cas, de mettre en évidence que l'instauration de relations lexicales et conceptuelles relève d'un choix d'interprétation et de représentation qui se manifeste par la création d'un concept.
Dans les faits, les informations sur le fonctionnement en discours sont plutôt dans le champ « terme » et celles sur le fonctionnement hors discours, fruit d'une interprétation, dans le champ « concept » ; l'établissement des liens terme/concept est également le fruit d'une interprétation même si ce choix d'interprétation se base sur une analyse des occurrences de termes (ce qu'indique le lien qui existe avec le corpus dans le modèle).

Le champ concept

Ce champ joue un rôle fondamental dans le modèle, les autres champs s'organisant autour de lui. C'est en effet dans ce champ que sont déclarées les relations d'un groupe terme/concept à un ou plusieurs autre(s). Les concepts et plus exactement les relations conceptuelles jouent un rôle majeur pour la modélisation mais aussi pour la méthode d'analyse du corpus (ces questions seront longuement évoquées dans le chapitre V). Dans la mesure du possible, les relations viennent remplacer la définition en langue naturelle, classiquement donnée dans les bases de données terminologiques mais qui n'a d'intérêt que pour une lecture humaine. La représentation sous forme relationnelle oblige, elle, à s'interroger sur la sémantique de ces relations et constitue, comme on l'a vu dans le chapitre précédent une ouverture très nette sur la problématique des ontologies. Dans certains cas, lorsque, par exemple, elle est donnée dans le corpus, il est possible de maintenir une définition en langue naturelle, une des possibilités d'utilisation des résultats étant la consultation directe des données.
Le concept est étiqueté par le (ou un des) terme(s) qui lui correspond(ent). Le fait qu'il s'agit d'une étiquette est signalé par un # qui précède le terme choisi. Les étiquettes pourraient être de nature numérique mais le concept perdrait alors tout lien avec sa dimension linguistique, ce qui poserait un problème à la fois du point de vue de la réflexion théorique qui a amené à la constitution du modèle et du point de vue de la lecture même des données.

Le champ terme

Le champ « terme » contient des informations de nature lexicale et discursive :

- Nombre et genre : le nombre n'est précisé que lorsque le terme est toujours utilisé soit au singulier soit au pluriel.
- Catégorie grammaticale : cette information sera surtout utilisée pour calculer les cas de synonymie,
- Existence de variantes : certains termes apparaissent dans le corpus sous différentes formes ; le cas le plus fréquent concerne l'ellipse d'un déterminant et/ou d'une préposition, par exemple, dans un corpus d'EDF : *phase de spécification produit/ phase de spécification du produit*,
- Existence d'un sigle (qui est une autre forme de variable). De nombreux termes sont utilisés sous leur forme siglée (*SCAO* pour *Système de Contrôle d'Attitude et d'Orbite*, dans le corpus MMS).

Le terme est entré sous la forme la plus étendue mais la forme la plus utilisée dans le corpus (forme étendue, variante, sigle) est précisée. Dans le cas d'une BCT multilingue, comme celle que nous avons construite pour MMS, la langue à laquelle appartient le terme est précisée, ceci afin de permettre un calcul d'équivalence d'une langue à l'autre.
Dans la perspective d'une utilisation en TAL, ce champ « terme » pourrait être enrichi de données sur le type de collocations dans lequel apparaît le terme : par exemple, type d'arguments pour un verbe ou une nominalisation comme dans (L'Homme et Gemme, 1997) par exemple. Hormis dans le cas de la BCT MMS, où les données devaient être intégrées à un outil d'aide à la traduction, les autres BCT construites étaient beaucoup plus en lien avec la modélisation de connaissances ou de fonctionnements sémantiques, si bien qu'il n'a pas été nécessaire de développer ce champ, ce qui nous a certainement évité bien des difficultés !

Le champ lien terme/concept

Ce champ contient des informations sur l'usage en général, c'est-à-dire sur des éléments extra-linguistiques qui justifient l'établissement d'un lien entre terme et concept. Il s'agit en fait surtout de préciser quel « locuteur collectif » utilise tel terme pour renvoyer à tel concept. Le plus souvent, dans une entreprise, les différents « locuteurs collectifs » possibles correspondent à des Divisions ou Départements internes. Au moment de la constitution du corpus, chacune des Divisions concernées par le projet a été représentée sous la forme d'un sous-corpus, identifié comme tel. Ainsi, nous verrons en IV que le traitement de la polysémie de *satellite* dans un corpus du CNES tient compte d'usages langagiers dans trois Divisions : Mathématiques spatiales, Systèmes électriques et automatiques et Observation de la terre.
Ce champ déroge à la règle qui veut que l'on ne rende compte que de l'usage dans le corpus puisqu'il permet aussi de rendre compte de « conseils d'utilisation » d'une instance normalisatrice ; par exemple, le terme *senseur*, pourtant largement utilisé à MMS est déconseillé par le Journal Officiel car il s'agit de la francisation de *sensor* ; les termes conseillés sont *capteur* et *détecteur*, seul le premier est utilisé à MMS. La BCT permet de rendre compte du fonctionnement de l'usage de ces trois termes.

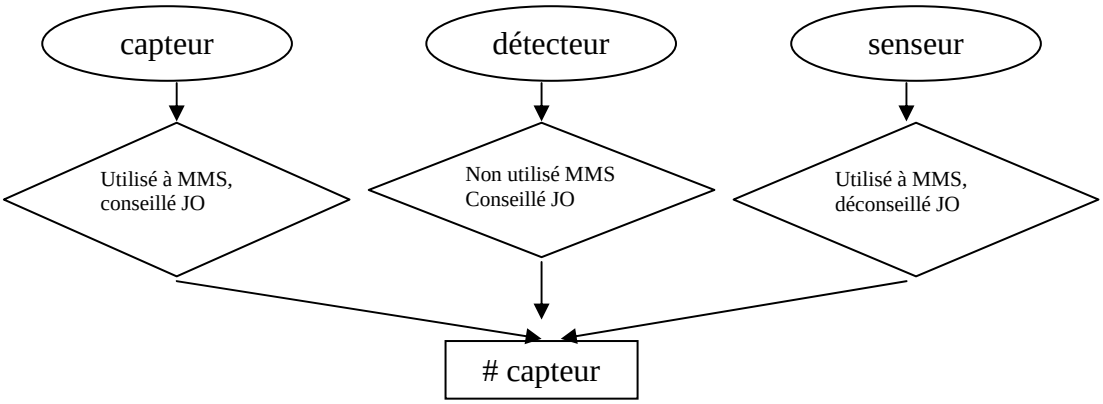
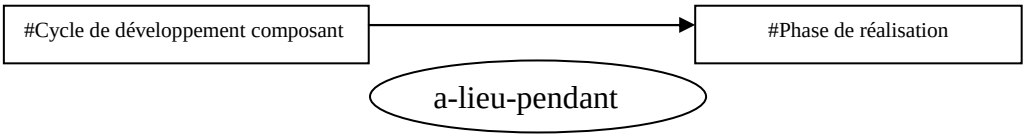


Figure 2 : Exemple de traitement de recommandations par un organisme officiel

Autre exemple, le service qualité, commanditaire du projet EDF (voir présentation en annexe) ne souhaitait pas que le sigle *PVC* soit utilisé pour renvoyer à *Plan de Validation Composant*. Cette recommandation a été indiquée dans le champ terme/concept. On voit bien ici combien la nécessité de constituer un modèle qui soit aussi économique que possible peut amener à ce qui peut apparaître comme des contradictions du point de vue linguistique. En effet, ce choix de modélisation nous a amenés à rendre compte dans le même champ de ce qui relève de l'usage attesté et de ce qui relève de l'usage souhaité (par les instances de normalisation). L'important est alors de maîtriser parfaitement le mode de représentation choisi et de contrôler comment se décide l'instanciation des champs définis.

Dans la perspective de maintenir un lien fort entre le corpus et la modélisation qui en est proposée, des éléments du corpus peuvent être utilisés pour justifier tel ou tel choix de modélisation, tout particulièrement pour justifier les relations conceptuelles comme dans l'exemple suivant, issu du corpus fourni par EDF :



Le cycle de développement composant se déroule pendant la phase de réalisation

Figure 3 : Exemple de relation entre concepts justifiée par un extrait de corpus

Ainsi, lorsque le modèle est mis en oeuvre dans un outil (comme dans l'outil GEDITERM), le corpus est stocké en même temps que les données du modèle de façon à ce que des liens puissent être établis entre certaines données du modèle et des passages du corpus qui justifient la modélisation choisie. Ces liens données/corpus permettent à un utilisateur humain de comprendre comment s'est faite l'élaboration du modèle. La lecture des seuls passages sélectionnés peut aussi donner un bon aperçu du contenu du texte puisque ces passages ont été choisis en fonction de leur pertinence sémantique, c'est-à-dire, le plus souvent, de leur capacité à exprimer une relation sémantique. Du point de vue de la construction de la BCT, la nécessité d'établir ces liens peut paraître fastidieuse (et elle l'est souvent). En revanche, elle amène les linguistes à avoir une réflexion sur leur manière d'analyser les données textuelles et à essayer de systématiser cette analyse.

1.2. Représentation et calcul des relations sémantiques : synonymie, homonymie, polysémie

Les questions les plus intéressantes du point de vue linguistique, lorsque l'on construit une BCT se posent à propos des phénomènes bien connus en sémantique que sont les relations sémantiques. Les deux prochains paragraphes décrivent la façon dont ces phénomènes sont représentés dans le modèle de données. Les chapitres IV et V décriront comment ils sont identifiés en corpus et les nombreuses questions que posent cette identification.

1.2.1 Représentation des relations sémantiques

Le modèle de données défini permet de trouver des conventions de représentation efficaces de certains phénomènes sémantiques. C'est le cas de l'homonymie, de la polysémie et de la synonymie.

Polysémie

Si l'on considère, conformément à la définition habituelle retenue, que la polysémie concerne des dénominations d'éléments qui entretiennent une parenté sémantique, la représentation la plus adaptée aurait été un seul terme qui renvoie à deux concepts entretenant une relation sémantique. Par exemple, dans le corpus d'EDF, nous avons rencontré un cas classique de polysémie qui s'établit pour une nominalisation déverbale entre l'action et le résultat de l'action : *documentation logiciel* correspondait ainsi soit à la rédaction, soit au document produit. La relation qui s'instaure entre les deux concepts est alors `a-pour-conséquence`.

Le problème vient de ce qu'il n'est pas toujours aisé d'identifier en corpus une relation qui distingue clairement deux concepts et ce pour trois raisons :

- il peut être difficile de représenter la différence entre deux concepts sous une forme relationnelle,
- la forme relationnelle peut être adaptée mais pas sous la forme d'une relation directe : des concepts peuvent être reliés mais par le biais d'un autre concept ; si on admet ce type de représentation pour la polysémie, il est impossible alors de contrôler ce phénomène puisque presque tous les concepts sont reliés, soit directement soit indirectement entre eux,
- il peut exister une relation entre les concepts mais elle n'apparaît pas clairement en corpus, c'est par une connaissance extérieure que l'on peut l'établir. Nous ne nous interdisons pas de faire appel à une connaissance extérieure mais lorsque c'est possible, nous essayons de justifier nos choix par des passages du corpus et nous hésitons toujours à représenter une relation qui n'est pas exprimée dans le corpus.[29]

Pour pallier cette difficulté, nous avons décidé de conserver la représentation un terme/plusieurs concepts sans imposer qu'une relation conceptuelle explicite soit indiquée entre les concepts (*cf* figure 4). Ce choix a eu pour conséquence d'imposer une modélisation particulière pour l'homonymie[30].

Homonymie

Du point de vue représentationnel, l'homonymie pourrait être proche de la polysémie : un seul terme renvoyant à plusieurs concepts. Ce mode de représentation ayant été retenu pour la polysémie, nous avons opté pour une autre solution qui consiste à dupliquer le terme. C'est donc par une similitude de la forme du terme que l'on repère l'homonymie.

Synonymie

La représentation de la synonymie est moins problématique : deux termes renvoient au même concept.

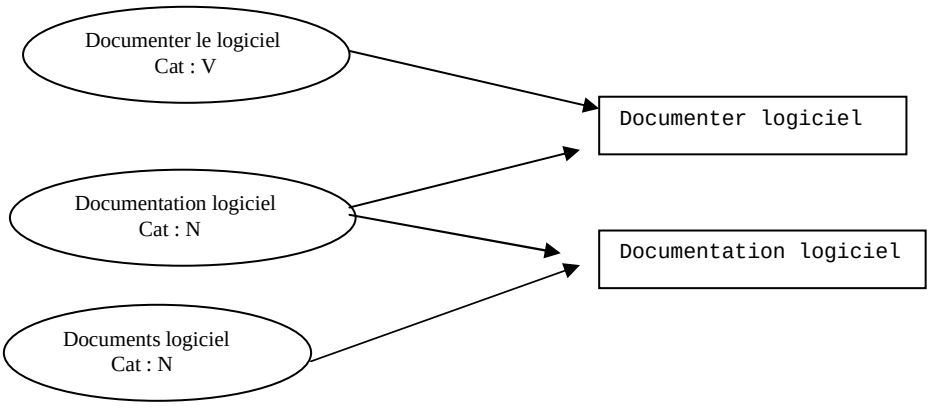


Figure 4 : Exemple de représentation de polysémie et de synonymie

[29] C'est le cas très souvent avec des termes qui sont composés d'une tête qui, à l'évidence, joue un rôle d'hyperonyme par rapport à l'ensemble du terme sans que cette relation soit explicitement mentionnée dans le corpus, par exemple, dans le corpus Mouglis : une *phase d'architecture produit* est bien une *phase* ; en revanche, un *espace de développement* n'est pas un *espace*…

[30] Il est bien évident que ce choix de modélisation n'est pas très élaboré, beaucoup moins par exemple que celui présenté dans (Victorri et Fuchs, 1996). Mais la question fondamentale dans notre perspective est moins celle de la représentation des phénomènes que celle de leur repérage en corpus.

Dans la figure 4, *document logiciel* et *documentation logiciel* sont synonymes, *documentation logiciel* est polysémique.

1.2.2 Calcul des relations sémantiques

La modélisation présentée ci-dessus concerne un choix de représentation des relations sémantiques dans laquelle ces relations ne sont pas déclarées comme telles contrairement aux relations conceptuelles ; lorsque le modèle est mis en œuvre, ces relations sont donc calculées selon les modalités suivantes :

Synonymie

Deux termes sont synonymes s'ils pointent sur le même concept et s'ils ont la même catégorie grammaticale. Par exemple, dans la figure 4, *documenter le logiciel* et *documentation logiciel*, ne sont pas synonymes puisqu'ils n'ont pas la même catégorie grammaticale.
Si on le souhaite, on peut rajouter une contrainte concernant l'équivalence des conditions d'usage (lien terme/concept) mais on peut aussi considérer que l'on souhaite connaître les alternatives possibles d'un terme dans une autre Division par exemple, et on peut alors relâcher la contrainte d'équivalence d'usage. On obtient ainsi un panorama de tous les termes utilisés dans une entreprise pour renvoyer à un même concept.
Un mode de calcul très proche peut servir à rendre compte du fonctionnement dans deux langues différentes, à travers l'information concernant la langue qui est mentionnée dans le champ « terme ». Sont équivalents alors deux termes de langue différente pointant sur le même concept.
En dehors de l'expérience à MMS, où nous avons constitué une BCT bilingue, à partir de deux corpus, l'un en français l'autre en anglais, avec d'une part une équipe francophone (linguistes et experts) et d'autre part une équipe anglophone (linguistes et experts), nous n'avons pas d'expérience d'utilisation de ce modèle dans des projets plurilingues. Nul doute que des questions nouvelles se poseraient dans ce cas, en particulier sur la compatibilité des réseaux conceptuels d'une langue à l'autre. Nous avons d'ailleurs rencontré un problème identique dans le projet SGGD, où nous avons constitué quatre BCT correspondant aux quatre sous-corpus des quatre partenaires (principaux), que nous avons ensuite fusionnées, ce qui oblige à faire une fois encore des choix de représentation (Jacques et Soubeille, 2000).

Homonymie

Deux termes sont homonymes lorsqu'ils ont la même forme.
Ce mode de repérage est facile puisque nous avons décidé de dupliquer les termes homonymes.
Comme pour la synonymie, on peut tenir compte ou non de variations dans le champ « terme » et dans le champ « lien terme/concept ». Pour qu'ils soient homonymes, on peut choisir par exemple que deux termes aient le même genre (par exemple, *greffe* peut être féminin ou masculin et on peut décider qu'il n'est donc pas homonyme). On peut aussi décider qu'il n'y a homonymie qu'à l'intérieur d'une langue (équivalence de la donnée langue dans chacun des champs « terme »), ou à l'intérieur d'une même entreprise (équivalence de cet élément dans le champ « lien terme/concept »).

Polysémie

Un terme est polysème lorsqu'il renvoie à au moins deux concepts. Comme pour la synonymie et l'homonymie, on peut rajouter des contraintes sur les données du « lien terme/concept ».

Ces exemples de calculs possibles montrent bien que ce genre de modélisation et l'usage qu'on peut en faire viennent interroger le linguiste sur sa connaissance et peuvent introduire une perturbation par rapport aux fonctionnements généralement admis en lexicologie.
Cette perturbation peut être de deux ordres.
Le premier type de perturbation relève d'une nécessaire recherche d'efficacité et il est alors indispensable que le linguiste soit conscient de cette nécessité et de la simplification théorique à laquelle elle conduit. Que, par exemple, on arrive à retrouver, par le même type de calcul, à la fois la synonymie intralangue et l'équivalence de langue à langue ne va pas sans poser de question. Cette possibilité vient de l'utilisation sans doute abusive du champ « concept » qui, en permettant des calculs de différentes natures, joue un rôle métalinguistique qu'il n'a (peut-être) pas dans la vision théorique que j'ai présentée en II. Le deuxième type de perturbation est lui, tout à fait assumé par la méthode d'analyse des données que nous souhaitons élaborer, qui va du corpus au modèle et qui tient compte de l'objectif de la modélisation. C'est ce contexte de travail qui justifie par exemple que les liens « terme/concept », qui concernent des informations d'usage, soient pris en compte pour calculer des relations sémantiques. Cela signifie en effet que ces relations sémantiques existent relativement à des groupes de locuteurs, préalablement identifiés, les phénomènes sémantiques étant observés à l'intérieur d'un corpus constitué de sous-corpus émanant de ces groupes de locuteurs (*cf.* traitement de la « polysémie » de *satellite*, chapitre IV). Il est probable qu'une vision plus classique des phénomènes linguistiques ne verrait pas là une véritable polysémie. En effet, dans ce type de vision, la polysémie n'est envisagée que dans une perspective descendante, qui va du système à son utilisation, comme l'indique les termes mêmes utilisés : *décomposition, dérivation :*

> « La polysémie des unités lexicales ‘ pleines ‘ a été décrite à l'aide de deux grands types de techniques, qui participent de deux cadres théoriques distincts : *la décomposition en traits sémantiques* et la *dérivation à partir d'un sens* " *premier* "» (Victorri et Fuchs, 1996 , 46).

Confronté à la nécessité de modéliser le fonctionnement lexical qu'il étudie en corpus, le linguiste oscille souvent ainsi entre un sentiment de simplification abusive et un autre d'ouverture permise par la recherche de stabilisation des phénomènes. Dans tous les cas, il est indispensable qu'il ait une conscience très claire des choix qui sont faits et de la raison pour laquelle ils sont faits (conformité à une théorie, recherche d'efficacité, adaptation à un besoin particulier). Alors seulement, la réflexion et le travail sur corpus peuvent trouver à s'épanouir.

1.3. Constitution du corpus

Dès que l'on se situe dans un domaine spécialisé, le recours à un corpus est la seule possibilité pour le linguiste. En effet, il n'a pas la compétence linguistique qui lui permettrait éventuellement de substituer à l'analyse des données réelles une analyse introspective. On pourrait dire d'une certaine façon que le linguiste ne peut être juge et parti dans ce cas de figure. Les réflexions que l'on peut mener sur la constitution du corpus, obligatoire lors de l'analyse sémantique dans un domaine particulier, peuvent ainsi servir de cadre plus général à la problématique de l'analyse linguistique de corpus. En effet, la dimension « spécialisée » du corpus est un élément extra-linguistique à prendre en considération qui n'est pas fondamentalement différent d'autres éléments situationnels. Le problème de la constitution du corpus est traversé par deux questions, liées entre elles, celle de la clôture et celle de la représentativité.

1.3.1 Le problème de la clôture

Dans la perspective de la terminologie traditionnelle, le corpus et son statut dans l'étude est très rarement problématisé. On considère qu'il y a des langues de spécialité par domaines et la difficulté est tout entière dans la nécessité de définir ces domaines ; il s'agit moins d'un

problème linguistique que d'un problème cognitif. A l'intérieur du système autonome garanti par la notion de domaine, tout ce qui permet d'accéder au fonctionnement jugé propre à ce domaine est pris en considération et retenu comme attestation : textes de toutes sortes, entretiens avec des experts, glossaires déjà existants…
On retrouve le même type de perception du domaine comme garant de la cohérence dans la théorie des sous-langages où la notion de clôture est fondamentale. Le corpus constitué pour étudier le sous-langage à l'œuvre dans ce domaine joue un rôle de référence ; il doit donc être très représentatif du système que l'on veut décrire.
Dans ces deux approches, la notion de clôture est donc clairement associée à celle de système autonome :

> « Les composantes du texte spécialisé sont étudiées dans le cadre de la linguistique de spécialité qui tient pour acquis que la langue de spécialité (LSP) peut être étudiée comme un système linguistique et opposée à des ensembles mieux connus, comme la langue commune. » (Auger et L'homme, 1995).

> « Une question que l'on peut se poser, selon Harris, est celle de savoir si l'on n'aurait pas intérêt à considérer que les sous-langages sont des systèmes " linguistiques " spécifiques plutôt que des sous-langues d'une langue naturelle. » (Dachelet, 1994, 111).

Ainsi, s'il peut y avoir clôture d'un corpus, c'est parce que cette idée est sous-tendue par la notion de système. Cela permet aussi de ne pas considérer le corpus comme un ensemble d'attestations mais plutôt comme des manifestations linguistiques d'un système sous-jacent que l'on peut mettre au jour. Une telle conception permet donc de décrire une langue (que l'on dit spécialisée) et pas seulement un discours. Dans (Condamines, 1997), j'évoque cette opposition langue/discours en essayant de montrer qu'une des conditions de passage du discours au système est celle d'un locuteur collectif qui permet de stabiliser en partie (et de donner un sens à) des régularités ; mais il est peu probable que ce locuteur collectif soit systématiquement assimilable à un locuteur compétent dans un domaine spécialisé. Ce locuteur collectif peut même n'avoir une pertinence que temporaire : le temps de la réalisation d'un projet ou de la rédaction d'un manuel, la situation de co-locution et l'objectif commun de cette collaboration créant en eux-mêmes des conditions suffisantes pour que s'instaurent des régularités d'usage constatables en corpus.
Cette idée de clôture permettant de délimiter un système autonome est évidemment séduisante et, d'une certaine façon, elle est à la base de notre point de vue sur les BCT, qui prennent le corpus pour référence. En effet, nous cherchons à décrire au plus près les régularités qui apparaissent en corpus avec l'hypothèse que ces régularités permettront de travailler sur l'ensemble du corpus en faisant appel le moins possible à des connaissances extérieures. Dans le même temps, les expériences que nous avons menées sur des corpus nous montrent que le seul recours au corpus est impossible. Même si la plupart des choix de modélisation sont inspirés par des éléments du corpus, il faut reconnaître que des fonctionnements sont décrits dans la BCT qui n'apparaissent pas dans le corpus comme tels mais qui font appel à des connaissances que nous avons sur le fonctionnement linguistique, c'est-à-dire, d'une certaine façon, des connaissances que nous avons rencontrées dans d'autres corpus, d'autres productions qui ont précédé notre analyse du corpus à l'étude, des régularités de différente nature que nous avons intégrées inconsciemment. Dans ce cas-là, la nécessité de faire des choix est encore plus cruciale car ils ne peuvent être étayés par des passages du corpus. Prenons un exemple. Dans un corpus sur les maladies coronariennes[31], plusieurs noms qui dénomment des affections potentielles des artères ont été repérés : *lésion, obstruction,*

[31] Il s'agit du corpus MENELAS constitué par les membres du groupe DIAM de l'AP-HP et qui a été mis à la disposition du groupe TIA afin de servir de matériau d'étude commun.

sténose, occlusion, réocclusion. Pour la plupart de ces noms, on trouve un dérivé de type adjectival :

Lésion d'une artère	*artère lésée*
Sténose d'une artère	*artère sténosée*
Occlusion d'une artère	*artère occluse*

En revanche, on ne trouve ni *artère obstruée* ni *artère réoccluse.* Ces deux termes ont un statut différent puisque *obstrué* fait partie des éléments connus de la langue alors que *réocclus* n'en fait pas partie. Que décider alors ? Qu'*artère obstruée* peut être intégré comme élément terminologique mais pas *artère réoccluse* ? Que la régularité de la dérivation est suffisante pour que l'on puisse retenir également *artère réoccluse* ? Que l'on décide de demander à un expert lesquels de ces deux termes il accepte (et il n'est pas rare que les experts, eux-mêmes influencés par leur compétence de locuteurs de la langue et des règles de dérivation aient des jugements bien plus tolérants que les productions réelles que l'on trouve en corpus)? Quel que soit le choix, doit-on conclure que le corpus n'était pas assez important et donc avait été clos trop tôt, i.e, qu'il ne comportait pas assez de données ?
Même si des éléments quantitatifs interviennent dans la constitution d'un corpus (Habert et *al.*, 1998), aucun corpus, aussi volumineux soit-il ne peut prétendre être parfaitement clos : cela signifie qu'en réalité, même constitué avec le plus grand soin, un corpus ne se suffit pas lui-même pour son interprétation. Outre des connaissances sur les éléments extra-linguistiques qui ont présidé à sa constitution, le contexte au sens large, son interprétation fait nécessairement appel à des connaissances linguistiques, c'est-à-dire à des connaissances acquises lors de l'interprétation, au sens large de compréhension, d'autres textes, c'est-à-dire de manifestations linguistiques contextualisées. Bien entendu, le corpus à l'étude doit entretenir un lien étroit avec la situation extra-linguistique dans laquelle s'inscrit son interprétation ; il doit donc être constitué avec cet objectif. A un moment donné, on va le déclarer clos mais il faudra alors savoir que cela ne signifie pas qu'il aura un fonctionnement parfaitement autonome. Il aura un statut spécial qui lui confèrera un statut de référence fort mais pas d'autonomie complète ; comme le dit François Gaudin :

> « …Cette relative clôture existe dans les faits, dans les objets étudiés… mais la pensée, les mots ne connaissent pas de frontières… » (Gaudin, 1995, 229).

Par ailleurs, même, si l'on peut mettre au jour des régularités dans un tel corpus, ces régularités ne sont pas en lien seulement avec un domaine défini *a priori* mais avec un ensemble d'éléments situationnels dont le domaine (par exemple, si l'on travaille sur un corpus provenant de Matra Espace, le domaine spatial sera à prendre en considération). Ce critère du domaine est d'ailleurs souvent difficile à délimiter : lorsque, dans le projet SGGD, nous avons à travailler sur un corpus traitant de la circulation dans l'agglomération toulousaine, quel domaine est concerné ? Est-ce que la « circulation automobile » constitue un domaine qu'on aurait pu déterminer *a priori* ?
Parfois, les contraintes extra-linguistiques qui ont permis de stabiliser les régularités d'usage sont si fortes que l'on peut se demander comment on pourra dégager les résultats de cette gangue contextuelle pour pouvoir les réutiliser tels quels dans d'autres analyses. Dans le chapitre 5, sur les relations conceptuelles, je donne des pistes pour cette généralisation. La difficulté vient donc de concilier à la fois le dynamisme de la langue en lien avec des situations extra-linguistiques toujours mouvantes et la nécessaire stabilisation que requiert une description « scientifique ».

1.3.2 La question de la représentativité

La question de la représentativité est liée d'une part à celle de la clôture et d'autre part à celle de la généralisation des résultats. Si un corpus est considéré comme représentatif, alors il sera également considéré comme clos. Mais pour qu'un corpus soit représentatif, il faut que les données qu'il contient puissent être considérées comme un échantillon fiable de l'ensemble des données qui pourraient apparaître dans le même contexte extra-linguistique.

En fait, on a souvent parlé de représentativité à propos de la possibilité d'utiliser un corpus pour élaborer le système d'une langue ; par exemple, à la fin des années 50, le français fondamental s'est élaboré à partir d'un corpus supposé représentatif des usages du français pour en permettre un apprentissage (Gougenheim et *al.*, 1958). La notion de contexte extra-linguistique est alors pratiquement inexistante, l'objectif n'est pas d'apprendre des usages particuliers d'une langue en fonction des situations mais de constituer des données censées représenter tous les usages possibles de la langue, considérés comme des exemples du système général. En réalité, dans ce type d'approche, la notion de compétence linguistique s'accompagne pour la plupart des auteurs de la suppression de la notion de contexte d'utilisation.

On a pu retrouver le même type de dérive en TAL, comme l'a montré Péry-Woodley (Péry-Woodley, 1995), où on a eu tendance à substituer la quantité des données à leur qualité.

La notion de genre textuel, parce qu'elle a pour objectif de regrouper des textes, s'inscrit dans la perspective de la représentativité. En effet, si un texte est reconnu comme participant d'un genre, alors, on peut faire l'hypothèse qu'il va avoir les mêmes caractéristiques linguistiques que tous les textes du même genre. Malheureusement, pour plusieurs raisons (présentées dans le chapitre 1, 2-1-1), l'intuition que nous pouvons avoir *a priori* sur l'existence de genres est souvent démentie par l'étude des phénomènes langagiers attestés dans les textes et de leur supposée régularité.

A l'opposé de cette vision descendante des situations extra-linguistiques, qui oblige à définir des « genres » *a priori*, l'approche ascendante permet une définition de la représentativité plus située.

Ainsi, dans mon expérience d'utilisation des corpus, que ce soit dans une visée applicative ou dans une visée plus théorique, la question de la représentativité me semble se poser à trois niveaux :

- Au niveau de la situation d'énonciation dans lesquels les textes du corpus ont été élaborés. Si, par exemple, on veut travailler sur la terminologie propre à MMS, on évitera de travailler sur un corpus de dialogues entre un ingénieur de MMS et un du CNES. En effet, par un phénomène d'adaptation, il se peut que l'un des deux adopte la terminologie de l'autre et le texte ne sera pas représentatif de la terminologie propre à MMS.
- Au niveau de l'objectif de l'étude : selon l'objectif que l'on s'est fixé, on ne sélectionnera pas les mêmes corpus. Prenons deux exemples ; si l'on veut constituer un thésaurus pour un Département d'une entreprise, on veillera à sélectionner un ensemble de textes concernant une connaissance partagée par les acteurs de ce Département : des présentations générales plutôt que des notes personnelles ou des mails. Dans le cas d'une étude plus théorique, si l'hypothèse est d'identifier des corrélations entre tel genre et tel phénomène, on veillera à sélectionner des textes très représentatifs de ce genre. Si, par exemple, on veut tester une hypothèse linguistique sur un corpus de genre didactique, on sélectionnera des textes propres à ce genre : des cours par correspondance plutôt que des articles de quotidiens par exemple. On voit bien ici que la constitution du corpus oblige à définir très précisément l'hypothèse et à donner des critères très fins de distinction de contextes extra-linguistiques.

– Au niveau de la méthode mise en œuvre : on se situe ici entre la notion de représentativité et celle de pertinence. La méthode que l'on va mettre en œuvre intervient dans la constitution du corpus. Dans notre perspective, comme le montre le modèle décrit ci-dessus, les relations conceptuelles jouent un rôle fondamental et la méthode pour les mettre au jour est celle des marqueurs ; cette problématique sera longuement discutée dans le chapitre IV. Dans une première appréciation de la notion de marqueur (« élément linguistique qui renvoie à un élément de contenu »), il est évident que les corpus les mieux adaptés seront les corpus riches en marqueurs, c'est-à-dire en éléments linguistiques renvoyant explicitement à une relation conceptuelle. Explicitement pour l'interlocuteur pressenti et aussi pour le linguiste qui va interpréter le texte. Meyer parle de « knowledge rich contexts » (Meyer, 2000); on pourrait parler ici de « knowledge rich corpus », notion qui est à mettre en lien avec la méthode utilisée et avec l'objectif de l'étude plus qu'avec le contenu du texte. Ce type de corpus correspond souvent à une situation où le locuteur est un peu plus expert que l'interlocuteur et se place donc dans une visée didactique au sens large. Il est évident qu'alors, la constitution du corpus privilégiera ce type de textes plutôt que des textes écrits par des experts pour d'autres experts.

Ainsi, au moment de la constitution du corpus, une certaine représentativité est recherchée. Elle fait intervenir ces trois points de vue, qui entretiennent des parentés : représentativité par rapport aux locuteurs que l'on pense concernés par l'objectif de l'analyse, représentativité par rapport à l'objectif d'analyse, lui-même en lien avec la méthode d'élaboration. La généralisation possible des résultats ne peut s'envisager que si l'on peut considérer ces points de vue comme relevant d'un paradigme : celui des genres textuels pour caractériser la situation de production des textes, celui des genres interprétatifs pour caractériser la situation d'interprétation ; c'est en tout cas l'hypothèse que je fais.

2. Présentation générale de la méthode d'analyse de corpus pour construire une BCT

Etant donné un corpus construit pour un objectif déterminé, et étant donné un modèle de données, il reste à présent à définir le type d'études qu'il faut mener pour aller du corpus à la modélisation.

Si l'on considère le modèle présenté ci-dessus, on peut voir que quatre tâches devront être effectuées :

- Repérage des termes,
- Repérage des variantes de termes,
- Repérage des liens terme/concept,
- Repérage des relations concept/concept.

Les méthodes permettant de réaliser certaines de ces tâches sont présentées dans ce paragraphe.

Puisqu'il s'agit de faire une étude à partir de corpus, seule une approche « distributionnelle », au sens large (et non au sens de l'application stricte de la méthode harissienne, trop coûteuse et pas toujours possible) est envisageable, c'est-à-dire une analyse des contextes d'apparition de certains éléments linguistiques, analyse qui se décline différemment selon la tâche à laquelle on s'intéresse. Notons enfin qu'en terminologie, certains chercheurs ont essayé d'expliquer les caractérisations possibles des fonctionnements à partir de l'analyse de corpus, *cf.* par exemple (Kocourek, 1982) ou (Meyer et Mackintosh, 1996).

2.1. Approche sémasiologique vs onomasiologique

Contrairement à l'approche préconisée dans la terminologie traditionnelle, qui est une approche onomasiologique, qui va de l'extra-linguistique au linguistique :

> « La terminologie retient surtout la relation onomasiologique, c'est-à-dire qu'à partir du référent, elle cherche l'étiquette ou signe désignatif » (Dubuc, 1985, 36),

l'approche qui est présentée ici est clairement sémasiologique puisqu'elle va du texte aux concepts. Cette approche pourrait paraître ainsi centrée sur le linguistique et d'une certaine façon, coupée de la réalité (ce lien avec la réalité étant fortement revendiqué par les terminologues qui l'utilisent pour justifier une approche onomasiologique). Ce serait oublier qu'antérieurement à l'analyse linguistique, le réel n'a cessé de guider les préparatifs : définition du besoin précis, définition d'un modèle de données, et tout particulièrement, constitution du corpus. Lorsque l'analyse sémantique se met en place, l'extra-linguistique a été défini et, d'une certaine façon, maîtrisé. Même si l'objectif de la modélisation est toujours présent dans l'analyse, on peut dire qu'au moment où débute l'étude, son cadre a été parfaitement défini ce qui, loin de la contraindre va au contraire lui permettre de prendre toute son ampleur[32]. Les éléments extra-linguistiques qui concourent à l'établissement du sens sont, à ce moment de l'étude, à peu près contrôlés ; l'étude peut alors se concentrer sur le contexte linguistique (le co-texte) et son rôle dans l'établissement du sens.

La mise en place d'une approche sémasiologique rapproche l'analyse terminologique de l'analyse lexicologique et inscrit clairement la terminologie dans une parenté avec la linguistique, rapprochement dont Wuster se méfiait :

> « Pour Eugen Wuster, la démarche sémasiologique était typique de la linguistique tandis que la démarche onomasiologique était la position caractéristique de sa théorie générale de la terminologie » (Kocourek, 1982, 45).

Ainsi, les premiers théoriciens de la terminologie se sont plus ou moins volontairement privés de l'apport possible de la linguistique dont ils connaissaient les hypothèses comme l'ont montré divers auteurs, par exemple (Van Campenhoudt, 1994).

Et pourtant, force est de constater que l'approche onomasiologique n'est pas opératoire. Chez la plupart des auteurs « classiques » – et également dans les normes officielles fortement inspirées par les travaux wustériens, *confert* par exemple la définition de l'ISO : « terme : désignation au moyen d'une unité linguistique d'une notion définie dans une langue de spécialité » (ISO 1087, 1990) –, le principal indice d'un fonctionnement terminologique consiste en un lien de désignation avec, soit une notion (ou un concept), soit un référent. Cette façon de voir confère au référent ou à la notion un statut premier. Or, peu de choses sont dites sur la façon d'accéder à cet élément premier, pourtant considéré comme la clé pour identifier les termes. Souvent, les experts d'un domaine sont censés être les dépositaires de ces notions. L'expérience nous a montré que les experts en question sont souvent bien embarrassés pour donner la liste des notions d'un domaine (la question récurrente est alors : dans quel objectif ?) et ils ont conscience de faire des choix tout à fait arbitraires. Par ailleurs, ce qu'ils croient être leur façon de conceptualiser et de parler n'est souvent pas confirmé par l'analyse des usages réels dans des textes.

La réalité de la pratique terminologique montre d'ailleurs que les productions écrites sont massivement utilisées pour le repérage des termes :

[32] Notons aussi que les résultats de l'analyse linguistique sont eux aussi confrontés à la réalité extra-linguistique puisqu'ils sont systématiquement proposés au jugement d'experts du domaine à qui l'on soumet simultanément les résultats et les passages du corpus qui permettent de les étayer.

« La terminologie puise sa matière première, constituée d'unités terminologiques et de contextes définitoires, dans des corpus de langue naturelle, surtout écrits mais aussi oraux. » (Otman, 1995, 109).

De fait, la démarche régulièrement utilisée en terminologie est de nature sémasiologique plus qu'onomasiologique mais il s'agit d'une démarche intuitive, qui ne vise pas à s'inscrire dans une théorisation. Ce type d'approche n'est pas très éloignée de celle utilisée pour la constitution du Trésor de la Langue Française qui, certes, met en oeuvre des textes pour attester des usages mais sans que cette utilisation soit faite de manière méthodique.
L'approche sémasiologique que nous avons mise en place pour la constitution de BCT recherche, elle, à systématiser les différentes étapes afin d'évaluer le type de connaissances qui est mis en œuvre et définir des éléments de méthode.

2.2. Sens et contexte

La difficulté de l'approche sémasiologique est qu'elle se base essentiellement sur l'analyse de co-textes pour décider du sens. Même si cette analyse tient compte aussi de tout ce que j'ai appelé contexte (de production des textes, de constitution du corpus, d'objectif de l'étude), le matériau à traiter est d'abord et avant tout linguistique. Il faut donc mener une réflexion sur les liens ente co-texte et sens. Dans cette perspective, l'analyse du co-texte consiste à identifier des parentés sémantiques manifestées sous des formes différentes, c'est-à-dire à constituer des classes d'équivalences contextuelles sur la base d'une interprétation. Une approche comparable est présentée dans (Descamps et *al.*, 1992).
Si, pour une forme donnée, on arrive à constituer une classe de contextes unique, alors on décidera en faveur d'un sens unique associé à cette forme et qui sera définissable par interprétation des contextes. Dans le cas contraire, on décidera en faveur d'une homonymie ou d'une polysémie. Dans l'élaboration de ces classes, faite le plus souvent de manière non consciente, il y a donc une forte dimension interprétative qui se fait pour l'essentiel d'après la compétence du linguiste et aussi en fonction de la modélisation recherchée. Cette interprétation n'est donc pas complètement neutre puisqu'elle fait appel à une compétence qui peut être variable selon les analystes. Toutefois, lorsque nous étions plusieurs linguistes à travailler de manière séparée, nous sommes très souvent arrivés à une interprétation similaire[33], ce qui semble plaider pour une compétence assez similaire d'un interprète à l'autre à condition qu'elle s'élabore avec une perception consensuelle de l'objectif de la modélisation. Il semble bien que la compétence linguistique soit ici à l'œuvre et que, à condition que l'analyse soit cadrée par une situation bien maîtrisée, elle puisse se mettre en place de manière assez régulière. Il est d'autres cas aussi où les interprétations possibles, les choix de modélisation sont moins liées à une compétence linguistique qu'à l'objectif de la modélisation ou de l'application voire à la nécessité de rendre compte des fonctionnements langagiers à l'aide d'un modèle prédéfini.
Le travail sur le sens à partir de corpus se fait donc clairement dans le cadre d'une interprétation. Cette interprétation fait appel à trois types d'éléments :

- la compétence de l'analyste, entendue comme la somme de ses expériences de locuteur/auditeur (compétence de locuteur) et de ses expériences antérieures d'analyste de cette langue, étude de données réelles, introspection, lectures de résultats de travaux, compétence de linguiste, (compétence de linguiste) ; cette compétence n'est pas figée mais toujours réactualisée ;
- les données du corpus, qui constituent un cadre de contrôle des phénomènes,

[33] Bien qu'il ne soit pas le résultat d'une expérience rigoureusement menée, ce constat est très différent de celui de J.Véronis, qui, faisant travailler 6 étudiants sur la polysémie « en général », c'est-à-dire en dehors de tout contexte linguistique ou situationnel a obtenu des résultats très hétérogènes (Véronis, 1998).

– l'objectif de l'analyse, qui permet de justifier *in fine* le choix d'une interprétation.

L'hypothèse d'un travail sur le sens en corpus spécialisé est donc que cohabitent un fonctionnement attendu, au sens où il est conforme à la (double) compétence de l'analyste et qui se met en œuvre dans l'analyse des contextes, et un fonctionnement inattendu qu'il s'agit de décrire. Toute la difficulté consiste à repérer ce qui relève d'un fonctionnement attendu, sur lequel va s'appuyer l'analyse et ce qui relève d'un fonctionnement inédit qui va être décrit[34]. Lors d'une même étude, la progression dans la description se fait, en principe, par la connaissance d'un nombre de plus en plus important de fonctionnements. Cette progression passe aussi par des périodes où ce que l'on tenait pour du connu se révèle de l'inédit, *cf.* par exemple, dans le projet SGGD, le cas de *carrefour* qui, pour un groupe de locuteurs (la mairie de Toulouse) s'est avéré renvoyer à un croisement d'itinéraires plutôt qu'à un croisement de voies, ces locuteurs appelaient ainsi *carrefour* un *passage piétons* (Jacques,2000). Il est donc nécessaire tout à la fois de mettre en œuvre sa compétence tout en restant très vigilant sur la réalité des fonctionnements à étudier. Il faut le dire donc, l'analyse de corpus est faite de tâtonnements et de voies sans issues mais aussi d'intuition. La mise au jour de la nature de cette intuition et de ses modes de mise en œuvre reste un des grands chantiers de l'analyse linguistique ; nul doute que les recherches sur corpus, qui demandent de s'interroger sur l'objet d'étude, devraient permettre de grandes avancées dans ce domaine

2.3. Première approche des études à effectuer

Cette partie présente les premières questions que soulève la mise en place d'une approche sémasiologique pour effectuer les quatre tâches qui ont été présentées ci-dessus : repérage de termes, de variantes de termes, des liens terme/concept, des relations conceptuelles. Certaines de ces tâches feront l'objet d'une description poussée dans les prochains chapitres.

Repérage de termes

Le repérage des termes, qui semble à la base de l'analyse terminologique, est un des problèmes les plus importants auquel se heurte la discipline, à la fois dans sa dimension théorique et opérationnelle. En réalité, aucune définition de « terme » n'est réellement satisfaisante ni opérationnelle. On retrouve, amplifié par la présence des corpus et de besoins applicatifs, le problème de la définition du signe linguistique, ou du mot et les nombreux débats qu'il a soulevés. La réponse, une fois encore, pourrait venir de la prise en compte de l'objectif de l'étude ; il n'y a peut-être pas des termes dans l'absolu mais des éléments linguistiques qui, compte tenu de ce qu'on veut en faire, du point de vue que l'on adopte, prennent une certaine pertinence. Les objectifs de l'étude peuvent être variés : traduction, modélisation d'un domaine, indexation, résumé… Chaque objectif lui-même met en œuvre des méthodes d'analyse différentes au point que la définition d'un terme semble plutôt liée à la méthode par laquelle on le repère qu'à une définition intrinsèque. Cette question sera traitée plus longuement dans le chapitre suivant.

Variantes de termes

Pour repérer qu'un terme est une variante d'un autre, il faut repérer d'une part que ces deux éléments entretiennent une parenté morphologique forte[35] et d'autre part qu'ils renvoient au

[34] Je montrerai dans le chapitre IV comment, tout en étant conforme à l'intuition de ce qu'est un travail d'analyse de corpus, le repérage d'une « déviance » est difficile à mettre en œuvre systématiquement parce qu'il est impossible de définir une norme qui pourrait servir de référence ; en effet, cette norme est elle-même réactualisée en permanence parce qu'elle correspond à la compétence du linguiste (de locuteur et d'analyste de la langue) qui est, elle-même, toujours variable.

[35] Les variantes sémantiques existent aussi mais, d'une part, elles sont plus difficilse à repérer automatiquement et, d'autre part, la difficulté est alors de décider si l'on a affaire à une variante ou à un autre terme, *cf* ci-dessous.

même concept, c'est-à-dire que les contextes dans lesquels ils apparaissent sont similaires. C'est le cas de *IVA moyenne* et de *segment moyen de l'IVA* dans le corpus MENELAS :

L'IVA moyenne présente une sténose

On a détecté une sténose du segment moyen de l'IVA

qui apparaissent dans des contextes quasi identiques[36].

Pour ce qui est des sigles, nous avons mis au point une petite méthode de repérage des sigles avec Judit Feliu, doctorante de l'IULA[37], en stage pour quatre mois dans l'ERSS (septembre-décembre 1999) :

- Recherche des mots inconnus : les logiciels d'analyse de textes qui comportent une base de données lexicales, comme SATO[38], fournissent cette liste,
- Parmi cette liste, repérage et élimination des noms propres (par exemple, AGDE qui, dans le corpus SGGD correspond au nom d'une voie (route d'AGDE) mais qui pouvait avoir une apparence de sigle.
- Parmi les mots restants, repérage des contextes où le mot est suivi ou précédé d'une parenthèse ou d'un trait, elle même suivi de la première lettre du mot étudié ; par exemple :

  SCAO (Système de Contrôle d'Attitude et d'Orbite).

Cette méthode est assez proche de méthodes appliquées de manière plus automatique comme dans (Bowden, 1998) et elle donne des résultats satisfaisants.

Dans certains autres cas, le repérage de variantes est beaucoup plus difficile ; plus exactement, ce qui est difficile est le choix entre variante de termes et nouveau terme. En principe, c'est la notion de concept qui permet d'opter pour un choix ou l'autre mais concrètement, ce qu'il faut décider, c'est si l'élément variant fait sens ou non. Même des adjectifs apparemment anodins peuvent suffire pour amener à la création d'un terme nouveau. Par exemple, dans le projet EDF, nous avons été amenés à créer deux termes *petit projet* et *grand projet* (et pas un seul terme *projet* avec deux variantes) qui renvoient à des concepts parfaitement définis en termes financiers : la barre des 4 MF permettant de distinguer un petit d'un grand projet.

Jacquemin a axé sa réflexion sur le repérage des variantes de manière automatique (Jacquemin, 2001). Il est très utile, en effet, que l'analyste se voie proposer des ensembles lexicaux, suffisamment proches morphologiquement ou syntaxiquement pour que des éléments soient des variantes d'autres éléments. Il reste qu'il faut décider si on donne à ces éléments un statut de variante ou un statut de terme à part entière, ce qui n'est pas un mince problème. (*cf.* (Jacques, 2000) pour un questionnement à ce sujet).

Lien termes/concepts

Le repérage des liens termes/concepts concerne en fait l'identification de phénomènes sémantiques bien connus : synonymie, polysémie, homonymie. Les choix de modélisation de ces phénomènes ont été présentés dans le chapitre II.

Reste à définir une méthode de repérage de ces phénomènes. Comme je l'ai déjà dit, l'analyse se fait sur la base du classement de contextes et également sur la base de la forme elle-même des termes étudiés.

---

[36] La difficulté ici est que les deux termes n'ont pas le même genre (l'un est féminin, l'aute masculin) ce qui, du point de vue du modèle défini devrait conduire à la création de deux termes plutôt qu'un seul avec une variante.

[37] Institut Universitaire de Linguistique Appliquée, dirigé par Teresa Cabre, Barcelone.

[38] SATO (Système d'analyse de Textes par Ordinateur) a été conçu à l'UQAM par Jean-Guy Meunier et développé par François Daoust. L'accès au texte s'effectue au moyen de concordances, c'est-à-dire de recherche de l'ensemble des occurrences d'un mot dans chacun de ces environnements contextuels. La grande originalité de SATO, par rapport aux autres concordanciers, réside dans le fait qu'il permet d'ajouter des propriétés aux mots ou aux segments textuels.

Dans une première approche, on peut caractériser ainsi le repérage de ces phénomènes.
SYNONYMIE : sont synonymes deux termes qui ont des formes complètement différentes mais qui apparaissent dans des classes de contextes similaires.
HOMONYMIE : il y a homonymie lorsque un terme candidat apparaît dans des classes de contextes complètement différentes.
POLYSEMIE : il y a polysémie lorsque un terme candidat apparaît dans au moins trois classes de contextes, deux (ou plus) n'entretenant aucune parenté et une troisième qui neutralise ces différences. Un exemple de repérage d'un fonctionnement polysémique est traité plus longuement dans le chapitre IV.
On le voit, au-delà de l'objectif de « remplir » une base de données, se posent des questions de sémantique, qui, dans ce type d'étude, ont pour particularité de s'appuyer sur des données réelles.

Relations conceptuelles

Les relations conceptuelles jouent un rôle majeur dans la constitution de BCT. On l'a vu, c'est la mise en réseau des concepts, de manière aussi systématique que possible, qui a permis la rencontre avec l'informatique. Au-delà de cet aspect, la constitution d'un réseau relationnel permet au bout du compte de définir les concepts présents dans le corpus et finalement de sélectionner les éléments linguistiques qui seront considérés comme termes. En effet, dans une approche basée sur l'analyse de corpus, on attribuera le statut de terme aux éléments linguistiques qui pourront être élaborés en concept, c'est-à-dire aux éléments qui, en corpus, sont reliés par des marqueurs de relations. Je reviendrai longuement sur ces questions dans les chapitres IV et V.

Cette rapide présentation a permis de poser une vision globale de la façon de mettre en œuvre l'analyse de textes pour repérer certains phénomènes sémantiques. Les prochains chapitres donneront des exemples précis de traitement.

3. D'un corpus à une BCT : les outils

Du strict point de vue de la théorie linguistique, l'analyse de corpus n'est pas nécessairement mise en œuvre à l'aide d'outils, comme je l'ai montré dans le chapitre I. L'histoire de l'analyse de textes (à défaut de corpus) a d'ailleurs commencé bien avant la mise à disposition d'outils. Par ailleurs, la caractérisation de certains phénomènes ne peut se faire à l'aide d'outils.
En revanche, lorsqu'il est possible, le recours à des outils facilite et enrichit l'analyse de deux points de vue :

- Ils permettent d'assister le travail d'analyse de textes, par exemple en proposant rapidement l'ensemble des contextes où apparaît telle ou telle structure;
- Ils donnent un point de vue particulier sur certains phénomènes lexico-syntaxiques en rapprochant des éléments éloignés dans le corpus mais proches par la forme ou par le contexte distributionnel dans lequel ils apparaissent.

Comme le note Blanche Benveniste :

> « Les concordanciers établis sur des corpus informatisés, en permettant de totaliser les emplois des mots sur des millions de cas, changent totalement la nature de l'analyse distributionnelle. Pour Halliday (1991), les changements sont si considérables qu'on ne peut plus maintenir les grandes oppositions méthodologiques héritées de Saussure. » (Blanche-Benveniste, 1996, 32).

L'élaboration de terminologies semble particulièrement inspirer les chercheurs en TAL : la constitution d'outils d'aide à cette élaboration constitue d'ailleurs une thématique importante

du TAL (Traitement Automatique des Langues) ; des workshops – *cf.* Computerm, lors de Coling 1998 –, des revues – par exemple le volume 43 (n°1/2002) de la revue TAL –, des livres – par exemple (Bourigault et *al.*, 2000) – sont consacrés à ce thème. Notons qu'aucun des outils construits dans cette thématique ne prétend construire seul une BCT ; il s'agit toujours d'outils d'assistance, qui laissent une part de choix importante et donc d'interprétation à l'utilisateur. Pour cette raison, on parle plus volontiers d'outils d'aide au repérage de termes candidats ou de relations candidates.

La présentation qui va être faite de ces outils ne se veut pas exhaustive. Je me place plutôt dans la perspective de l'utilisation de ces outils pour la constitution de BCT et des problèmes méthodologiques qu'ils posent du point de vue linguistique[39]. Pour comprendre leur pertinence (ou leur non-pertinence) dans la position théorique que j'adopte, il est nécessaire de présenter les principes qui président au fonctionnement de ces outils. En 1997, avec J. Rebeyrolle (Condamines et Rebeyrolle, 1997), nous avions identifié deux grands principes, toujours valables pour décrire les outils existants : l'un correspondant à une vision que nous avions caractérisée d'ascendante, l'autre à une vision descendante. Les principes de ces outils sont décrits dans les trois prochains paragraphes (certains de ces outils étant considérés comme mixtes), un quatrième paragraphe s'intéresse aux outils « généraux » d'analyse de textes.

3.1. Outils de type descendant

Certains outils se placent dans la perspective d'une connaissance linguistique définie *a priori* et projetée sur le corpus d'étude afin de repérer des phénomènes. Ces outils sont basés sur l'idée que les corpus spécialisés, comme tous les corpus, sont des actualisations d'un système linguistique unique et stable auquel on peut accéder en grande partie par introspection. On retrouve ce principe dans les deux grandes classes d'outils dédiés à la terminologie : les outils de repérage de candidats termes et les outils de repérage de relations conceptuelles candidates.

Outils descendants de repérage de candidats termes

Les outils de repérage de candidats termes, lorsqu'ils utilisent un principe descendant, partent du constat que dans toutes les terminologies existantes, la grande majorité des termes sont des syntagmes nominaux. On peut s'interroger sur une telle abondance de formes nominales. En effet, étant donné que la définition même de ce qu'est un terme n'est pas claire, on peut se demander pourquoi la forme nominale est privilégiée. On peut tenter deux types d'explications. La première concerne le fait que la terminologie contemporaine est l'héritière des travaux de classement qui ont été menés dans le domaine des sciences naturelles et des techniques aux XVIIè et XVIIIè siècles : il faut nommer et ordonner les éléments pour leur donner une réalité scientifique. Le lien de nomination s'établit alors avec les objets du monde avec cette perspective très prégnante de constitution de taxinomies. Les noms, seuls, sont alors pris en considération parce qu'ils constituent la forme la plus adaptée pour nommer des objets du monde. C'est d'ailleurs dans une perspective assez proche que se situe la terminologie « classique ».
L'autre explication, plus en lien avec la réalité des phénomènes discursifs, relève du fait que les noms sont perçus comme les éléments qui ont atteint le degré maximum d'abstraction, de dégagement par rapport au contexte et donc susceptibles de transporter le maximum de sens « intrinsèque ». Pour cette raison aussi, les actions apparaissent bien plus volontiers sous la

[39] Pour une description plus exhaustive et plus informatique, on peut se référer à (Bourigault et Jacquemin, 2000).

forme d'une nominalisation que sous celle d'un verbe (*cf.* chapitre IV), comme le note Rastier :

> « La nominalisation est fort utilisée pour créer un effet d'objectivation : c'est pourquoi elle est massivement attestée dans les textes scientifiques (notamment positivistes) et dans les discours qui les imitent (langue de bois) » (Rastier, 1995, 51).

Cette deuxième raison explique le fait que dans tous les types de modélisation à partir de corpus (terminologies mais aussi thésaurus, bases de connaissances…), on retrouve une prédominance nette des noms : d'une part, on a le sentiment que la forme nominale est d'un bon rapport forme/contenu (mais sans toujours avoir conscience que cette situation peut créer de l'ambiguïté), d'autre part, la mise en réseau relationnel semble plus facile si l'on n'utilise que des noms.

Ces outils de repérage de termes candidats sont basés sur l'étude des terminologies existantes et sur la structure des termes (des syntagmes nominaux) qu'ils contiennent, par exemple, *Nadj, N de N, N prép N…*Les outils construits sur ce principe supposent que soit d'abord effectuée un étiquetage grammatical puis que soient repérées les suites d'éléments qui correspondent à ces structures. L'outil NOMINO[40], construit par David et Plante fonctionne sur cette base (David et Plante, 1990). Il s'agit d'un outil commandé par le RINT (Réseau International de Terminologie), d'un coût modique et qui est très utilisé par la communauté des terminologues.

Outils descendants de repérage de relations candidates

Le fonctionnement de ce type d'outils suppose, en principe, que soient définis les éléments suivants :

- l'ensemble des relations conceptuelles propres à une langue,
- l'ensemble des marqueurs de ces relations.

En France, les outils constitués par l'équipe LALIC (Langage, Logique, Informatique), Cognition du CAMS (Centre d'Analyse et de Mathématique Sociale)), fonctionnent sur ce principe, qu'ils mettent en œuvre dans le cadre de la Grammaire Applicative Contextuelle. Cette grammaire introduit, en plus de la reconnaissance de structures, des contraintes sur le contexte dans lequel apparaissent ces structures. Dans sa mise en œuvre, cette « exploration contextuelle » (Desclés et *al.*, 1997) n'est pas très différente de l'approche que je propose dans le chapitre V. L'élément fondamentalement différent vient de ce que, dans la perspective du CAMS, les relations sont définies *a priori* et l'analyse de textes réels sert à identifier les marqueurs qui correspondent à ces relations, indépendamment du genre textuel dans lequel ils apparaissent.
En réalité, pour l'instant, ces outils s'intéressent à des relations très connues comme l'hyperonymie avec des outils comme : SEEK, réalisé au CAMS (Jouis, 1995), TEXT ANALYSER, réalisé à l'Université d'Ottawa ((Kavanagh, 1996), (Davidson, 1998)) ou la cause, avec des outils comme : COATIS, également développé au CAMS (Garcia, 1998).
L'intérêt de ces outils est qu'ils fonctionnent sans étiquetage grammatical, par simple repérage de formes ce qui permet un gain de temps et ne semble pas nuire à la qualité des résultats.

3.2. Outils de type ascendant

Les outils de type ascendant visent à faire émerger les fonctionnements propres aux corpus. Ces outils sont basés sur l'idée que le corpus est suffisamment clos pour qu'y soit en œuvre

[40] http://www.ling.uqam.ca/nomino/

un fonctionnement cohérent et autonome ; en fait ces outils reprennent l'idée harrissienne de sous-langage.

Outils ascendants de repérage de termes candidats

Les outils de repérage de candidats termes dans ce type d'approche consiste à utiliser la méthode classique du repérage des segments répétés. Cette méthode est à la base du fonctionnement d'ANA par exemple (Enguehard et Pantera, 1995). Pour limiter le nombre de segments répétés non pertinents, cet outil part d'une liste de termes de départ proposée par les experts du domaine. L'idée sous-jacente ici est que plus un groupe de mots est répété, plus il a de chances d'être un terme.

Outils ascendants de repérage de relations candidates

L'objectif de ce type d'outils est d'aider à trouver en corpus des passages susceptibles d'exprimer une relation conceptuelle. Cette relation peut être préalablement connue ou non.
PROMETHEE (Morin, 1999) se situe dans cette perspective. Cet outil s'inspire des travaux de Hearst (Hearst, 1992) qui propose une méthode dans laquelle les résultats positifs obtenus sont immédiatement utilisés pour faire une nouvelle recherche. A partir de couples de termes reliés par une certaine relation connue, il s'agit de rechercher dans le corpus tous les endroits où ces termes coocurrent afin de faire apparaître éventuellement des marqueurs de relation nouveaux. PROMETHEE essaie de calculer, sur la base de récurrences de formes, quels éléments peuvent jouer le rôle de marqueurs dans les contextes où coocurrent les termes.
Cet outil ne peut donc être utilisé que si l'on connaît une relation pertinente et un ensemble de couples unis par cette relation. Ainsi, il s'agit plutôt d'un outil de repérage de marqueurs candidats puisque la relation est connue *a priori.*
Cette connaissance n'est pas nécessaire dans STARTEX (Rousselot et al, 1996), qui propose à l'utilisateur un ensemble de contextes dans lesquels coocurrent des termes, préalablement validés. Il s'agit pour l'utilisateur de repérer les contextes pertinents, c'est-à-dire ceux où s'expriment une seule et même relation conceptuelle comme la relation causale dans :

Un infarctus du myocarde par sténose de l'IVA
La sténose de l'IVA est responsable de l'infarctus du myocarde
Un IDM en relation avec une sténose de l'IVA.

Autres outils de type ascendant

D'autres outils de type ascendant proposent des résultats qui s'inspirent directement de la théorie distributionnelle pour essayer de constituer des classes sémantiques sur la base de parentés distributionnelles. C'est le cas de ZELLIG (Habert et al, 1996) et de LEXICLASS (Assadi, 1998). L'interprétation des classes, comme la mise au jour de relations à l'intérieur des classes est à la charge de l'utilisateur comme l'est l'interprétation des résultats de tous les outils présentés. Il faut d'ailleurs reconnaître que les classes proposées ne sont pas toujours interprétables d'un point de vue sémantique.
Qu'ils soient descendants ou ascendants, la plupart de ces outils d'aide à la terminologie génèrent à la fois du bruit (les candidats proposés ne correspondent ni à des termes ni à des marqueurs de relations conceptuelles), et du silence (des éléments pouvant être des termes ou des marqueurs de relations ne sont pas proposés), ce qui est beaucoup plus ennuyeux ; ces inconvénients sont dus à des raisons que l'on peut résumer dans le tableau suivant.

	BRUIT	SILENCE
Extracteurs de termes descendants	Tous les SN correspondant à certaines structures ne sont pas des termes	Les termes ne sont pas toujours des SN
Extracteurs de relations descendants	Les contextes ne sont souvent pas assez contraints	Seules les relations prédéfinies, marquées par des éléments prédéfinis sont recherchées
Extracteurs de termes ascendants	Les segments répétés ne sont pas toujours des termes	Certains termes ne sont pas des segments répétés
Extracteurs de relations ascendants	Startex : la cooccurrence de termes en corpus correspond rarement à l'expression d'une relation	Prométhée : seules les relations identifiées *a priori* peuvent être repérées
Autres outils ascendants	La cooccurrence de mots n'a pas toujours un sens, pas plus que la constitution de classes sur cette base	

Tableau 1 : Bruits et silences générés par les outils d'aide à l'extraction terminologique

Pour pallier ces difficultés, une nouvelle génération d'outils essaie de mettre en œuvre les deux types de méthode (ascendante et descendante) afin de limiter les problèmes engendrés par l'application d'une seule méthode. Il s'agit des outils « mixtes ».

3.3. Outils mixtes

Les outils mixtes, dont il est question dans ce paragraphe, semblent être une spécificité française. A l'origine de ces outils, il y a certainement la conscience des insuffisances des outils seulement ascendants ou seulement descendants. D'une manière plus générale, l'évolution s'est faite dans le sens d'une recherche d'efficacité, qui emprunte à chaque approche ce qui semble le plus adapté, à tel ou tel moment de l'analyse.

Outils mixtes de repérage de candidats termes

Les outils mixtes de repérage de candidats termes mettent en œuvre une approche qui s'appuie à la fois sur une connaissance *a priori* et sur un fonctionnement propre au corpus. ACABIT (Daille, 1994) et XTRACT (Smadja, 1993) combinent la recherche de patrons et la méthode des segments répétés. LEXTER (Bourigault, 1996)[41] met en œuvre une approche linguistique originale qui consiste à rechercher les candidats-termes en creux. Il s'agit de repérer les éléments qui apparaissent parmi d'autres éléments pouvant jouer le rôle de frontières de termes, c'est-à-dire de frontières de syntagmes nominaux (par exemple, un verbe et une préposition). Dans un second temps, l'outil met en œuvre une méthode d'apprentissage endogène qui tient compte de la combinaison des mots dans le corpus étudié.

Outils mixtes de repérage de relations candidates

CAMELEON (Séguéla, 2001) est un outil de ce type. Il utilise des marqueurs connus de méronymie et d'hyperonymie qu'il projette sur le corpus pour évaluer leur productivité ; dans un second temps, à partir des résultats obtenus, il projette les couples de termes repérés pour

[41] L'approche endogène mise en œuvre dans Lexter est à la base d'un nouvel outil (Syntex) qui est en train d'être développé par Didier Bourigault et Cécile Fabre (Bourigault et Fabre, 2000). Cet outil ne concerne plus seulement les syntagmes nominaux mais aussi les verbes et leur structure argumentale. Ce changement de perspective correspond à une vision du corpus plus approfondie qui se rapproche de l'objet d'étude des linguistes : il ne s'agit plus seulement de repérer, en surface, des éléments que l'on va retenir comme termes. Dans le même temps, ce type d'outils devient sans doute moins utilisable par les terminologues car il nécessite de bonnes connaissances en linguistique. En revanche, il pourrait être d'une aide précieuse pour la linguistique de corpus en mettant au jour des régularités de toutes natures, propres au corpus à l'étude.

identifier des marqueurs propres au corpus. Il y a donc une recherche d'un fonctionnement propre au corpus à l'étude. L'outil et la méthode assistent l'utilisateur durant toute la session. Ainsi, contrairement à PROMETHEE, qui utilise une hypothèse assez proche, il y a, avec CAMELEON, recherche d'adaptation à des besoins et des utilisateurs réels, ce qui amène à poser les questions différemment. De ce point de vue-là, CAMELEON est un véritable outil d'ingénierie linguistique.

D'autres types d'approches visent à augmenter la liste des termes ou des relations conceptuelles entre ces termes sur la base de connaissances sémantiques ou morphologiques. Ces connaissances ne sont pas mises en œuvre sur des corpus mais sur des terminologies existantes. Ainsi (Hamon, 2000) utilise des connaissances sur la synonymie en langue générale pour l'étendre à la terminologie d'un domaine afin d'essayer d'augmenter le nombre de termes. De la même façon, Grabar et Zweigenbaum (Zweigenbaum et Grabar, 2000) utilisent des règles de dérivation morphologique pour augmenter la couverture des relations existantes dans une terminologie médicale (la SNOMED).
Ces outils ne concernent pas l'analyse de corpus mais ils pourraient être mis en œuvre dans un second temps, après une première élaboration de termes et de relations conceptuelles à partir de corpus.

Presque tous réalisés dans le cadre d'une thèse, ces outils ont le plus souvent servi à approfondir une hypothèse et exploiter les possibilités de sa mise en œuvre en TAL. La préoccupation majeure de leurs concepteurs a rarement été de prendre en compte au plus juste le besoin des utilisateurs (hormis peut-être SEEK, qui a été commercialisé, et CAMELEON, qui s'est inscrit dans une réelle perspective d'ingénierie linguistique). En tout cas, si un utilisateur était identifié, il s'agissait soit d'un expert du domaine, soit d'un terminologue, soit d'un ingénieur de la connaissance ; jamais, à ma connaissance, de linguistes de corpus. En réalité, ces outils sont difficilement utilisables dans le cadre d'une analyse de corpus pour constituer une BCT et ce pour différentes raisons :

- il est très difficile de se procurer, d'installer, de maintenir ces outils, existant seulement la plupart du temps sous la forme de prototypes ;
- ces logiciels sont souvent dédiés trop spécifiquement à une tâche ; quand on voit le nombre important de phénomènes à étudier pour constituer une BCT, on se rend compte que ce sont plusieurs de ces outils qu'il faudrait pouvoir utiliser, ce qui supposerait d'avoir une idée précise de la façon de combiner leur utilisation (d'où sans doute l'intérêt de concevoir une plate-forme comportant plusieurs fonctionnalités) ;
- conçus pour évaluer des hypothèses linguistiques, ces outils ne fournissent souvent pas des résultats assez fiables : trop de bruit mais aussi, ce qui est plus grave, trop de silence ;
- les hypothèses qui sont mises en œuvre dans ces outils n'ont souvent pas été travaillées suffisamment en linguistique. En revanche, tester ces hypothèses de manière automatique, même si elles sont insuffisamment mûres, peut donner des pistes de réflexion à la linguistique de corpus. C'est finalement cette dimension de mise à l'épreuve d'hypothèses qui est la plus intéressante pour les linguistes, beaucoup plus que l'utilisation réelle de ces outils. D'ailleurs, la plupart de ces outils ont été conçus dans un contexte pluridisciplinaire.

Au fond, l'utilisation réelle des outils, que l'on pourrait croire primordiale pour le TAL, ne constitue pas, au stade où en sont les recherches, la principale motivation des concepteurs

d'outils. La difficulté principale vient peut-être de ce que le contexte de conception de ces outils n'est pas stable :

- les besoins ne sont pas clairement identifiés en entreprise (si la plupart de ces outils ne sont pas commercialisés, c'est aussi parce que le marché n'est pas vraiment identifié ni, pour l'instant, suffisamment important);
- les linguistes sont rarement envisagés comme des utilisateurs potentiels : il y en a peu en entreprise ; quant aux chercheurs, leurs besoins, spécifiques, sont très rarement pris en considération et les outils ne sont donc pas conçus pour s'intégrer à une recherche en linguistique [42]. Il est vrai aussi que les besoins peuvent être différents selon les intérêts du linguiste ;
- la connaissance sur le fonctionnement terminologique en corpus n'est pas encore assez développée ; de ce point de vue-là, les recherches en lexico-sémantique et en TAL avancent au même rythme, ce qui explique parfois une confusion des problématiques ;
- la répartition entre résultats qui peuvent être obtenus automatiquement et résultats qui nécessitent une interprétation n'est pas claire, toujours pour la même raison d'une réflexion qui est encore très insuffisamment développée.

Il se peut aussi que la crainte des informaticiens de se transformer en simples ingénieurs les amène à occulter une utilisation réelle de leurs outils.

Enfin, il me semble que pour le linguiste de corpus, la donnée textuelle constitue son matériau premier, qu'il doit travailler jusqu'à élaborer un modèle ; il a donc besoin d'un contact très direct avec ce matériau et, parfois, l'utilisation d'outils qui produisent des résultats plus ou moins pertinents, plus ou moins complets, sans que la façon de les produire soit très claire, joue plutôt un rôle de parasite. Sont utiles en revanche des outils qui permettent de mettre en lumière des phénomènes difficilement visibles par une lecture linéaire, parce qu'ils mettent au jour des proximités de fonctionnement entre des formes linguistiques très éloignées dans le corpus, ce qui peut orienter la réflexion sur (et l'analyse plus approfondie de) ces formes-là.

3.4. Outils d'analyse de corpus

Les outils d'analyse de corpus sont eux aussi rarement commercialisés ; il est, en revanche, très courant que des informaticiens élaborent leur propre outil soit pour leur usage personnel, soit pour des collègues linguistes. Dans la mesure où ce type d'outils peut être employé de manières très diverses – ils sont d'ailleurs utilisés pour analyser des corpus dans plusieurs disciplines (analyse littéraire, sociologie, psychologie…) –, ils présentent souvent une grande souplesse d'utilisation. L'objectif est de repérer très rapidement toutes les occurrences de mots ou de structures plus ou moins complexes dans un corpus. Dans la perspective d'une linguistique de corpus pour construire des BCT, ces outils sont utilisés à divers moments :

- pour repérer les données à consigner dans la BCT, c'est-à-dire, effectuer les tâches énoncées en 2.3 ;
- pour travailler en amont, afin de caractériser certains phénomènes en corpus comme la polysémie, les marqueurs de relations en lien avec la nature du corpus, la recherche de variantes de termes.

Ils constituent des moyens d'accès aux données du corpus et sont de ce fait incontournables en linguistique de corpus en général, seuls ou pour valider des résultats proposés par des outils plus dédiés à la terminologie.

L'outil YAKWA, que nous utilisons dans l'Equipe de Recherche en Syntaxe et Sémantique, a été conçu spécialement pour nos besoins de linguistes de corpus par un collègue

[42] Cet état de faits est tout à fait regrettable tant il semble que seules, des études sur corpus approfondies pourront permettre de construire des outils eux directement utilisables par des terminologues et avec des résultats bien meilleurs que ceux obtenus actuellement.

informaticien, Ludovic Tanguy. Il a pour caractéristique principale de fonctionner sur un corpus étiqueté, ce qui permet de définir des requêtes qui contiennent des éléments grammaticaux (par exemple, Dét N V).
Non dédié à une tâche spécifique, ce type d'outils est très interactif : il permet d'interroger un corpus très rapidement et d'intégrer tout aussi rapidement les résultats obtenus dans la réflexion. Il permet ainsi d'affiner les hypothèses et d'éviter les impasses tout en permettant une prise en compte au plus juste des fonctionnements propres au corpus.
Finalement, l'outil idéal d'analyse de corpus pour constituer des BCT devrait prendre en compte à la fois un besoin de recherche et de caractérisation de phénomènes et un besoin d'élaboration de donnéees ; il devrait aussi combiner au minimum les fonctionnalités d'un extracteur de candidats termes « mixte », d'un outil d'extraction de relations candidates « mixte », d'un outil d'exploration de textes et d'un outil d'enregistrement et d'interrogation des données (comme GEDITERM ou TERMINAE, conçu au LIPN avec une dimension formelle plus importante (Biébow et Szulman, 1999)). Nous ne désespérons pas d'arriver à construire un tel outil à Toulouse.

4. Conclusion

Ce chapitre a permis de présenter les principales étapes et les principes méthodologiques qui permettent de construire une BCT à partir d'un corpus. Cette construction est présentée comme relevant à part entière d'une sémantique de corpus et se veut une exemplification de ce que peut être une sémantique doublement située. Ainsi, comme je le souhaitais dans le chapitre I, ce qui pourrait sembler relever seulement d'une linguistique appliquée (parce qu'on définit des objectifs d'étude, qui se manifestent en particulier par un modèle de données prédéfini) peut être considéré comme un véritable questionnement théorique, pas fondamentalement différent d'une approche plus traditionnelle en linguistique.
Pour continuer dans cette perspective, le prochain chapitre va aborder le traitement de certains phénomènes sémantiques qui concernent directement la construction de BCT mais qui intéressent aussi des questions plus théoriques sur l'apport de l'analyse des corpus à la sémantique en général.

Chapitre IV
D'un corpus à une BCT : nouvel éclairage sur des phénomènes linguistiques connus

Ce chapitre va me permettre de préciser ma réflexion sur l'analyse sémantique de corpus, à travers la présentation de travaux menés sur certains phénomènes sémantiques auxquels on est nécessairement confrontés lorsque l'on veut construire une BCT à partir d'un corpus. En effet, l'utilisation d'un corpus comme référence au travail d'élaboration de la BCT amène un éclairage nouveau sur des fonctionnements sémantiques souvent décrits sur des bases introspectives, et alimente la réflexion de manière originale. Ainsi que je l'ai dit dans le chapitre I, ce processus d'élaboration d'une BCT ne relève pas seulement d'une linguistique appliquée ; cadré d'une part par la situation de production des textes du corpus, d'autre part par l'objectif d'interprétation, toutes deux pouvant relever d'un genre, ce processus s'inscrit tout autant dans une perspective théorique.

Les phénomènes qui vont être présentés ici sont de trois ordres. Tout d'abord, je m'interrogerai sur ce que le fait de travailler à partir d'un corpus a comme conséquence sur la définition que l'on peut donner d'un terme. En effet, alors que l'approche que je propose consiste à prendre le texte, voire le corpus comme unité d'étude, le modèle de données privilégie, quant à lui, les unités lexicales et les relations qui les unissent. Je montrerai que cette situation n'est pas contradictoire, à condition que la modélisation ne se focalise pas d'emblée sur le repérage de termes mais que ce repérage apparaisse comme la dernière étape d'une mise en ordre interprétative qui concerne le texte ou le corpus dans son entier.

Dans un second temps, j'examinerai un phénomène longtemps considéré comme très secondaire dans les études en terminologie : la polysémie. Comme pour l'identification de termes, il s'agit de voir comment on peut élaborer une représentation centrée sur le lexique à partir de contextes d'utilisation ; or, cette perspective est très rarement prise en considération dans la sémantique « classique » qui décide *a priori* de l'existence d'une polysémie pour éventuellement étudier comment le contexte permet de décider en faveur d'un sens ou d'un autre. Je montrerai comment l'objectif d'interprétation joue un rôle majeur pour mettre au jour la polysémie.

Enfin, je ferai part des travaux que j'ai menés sur les nominalisations déverbales. Il s'agit d'évaluer la pertinence d'une intuition courante parmi les terminologues, qui est celle d'une

utilisation importante de la forme nominale dans les corpus spécialisés. Les résultats que je présenterai semblent confirmer une utilisation élevée de nominalisations, élément qui pourrait constituer un élément de validation linguistique de ce qui est couramment appelé « langue spécialisée ». Si cette hypothèse se confirme, on pourrait envisager de considérer que la langue spécialisée correspond à un genre particulier, qui s'opposerait à un genre non-spécialisé (qui se manifesterait par des discours moins riches en nominalisations).

1. Repérage des termes

Comme je l'ai déjà souligné, les définitions classiques du terme posent le concept comme préexistant, le terme servant de désignation (parfois au sens figé d'étiquette). Outre les exemples de ce type de définitions déjà donnés dans le chapitre III, citons encore la définition de l'OLF (Office de la Langue Française) de Québec :

> « Le terme se définit comme unité signifiante constituée d'un mot (terme isolé) ou de plusieurs mots (termes complexes) qui désigne un concept, de façon univoque à l'intérieur d'un domaine ».

J'ai montré en II que cette notion de concept préexistant, en plus des problèmes théoriques qu'elle soulève, n'est pas opérationnelle lorsque l'on est confronté à l'analyse d'un corpus. En effet, on n'a pas une liste de concepts prédéfinis dont il faudrait chercher les réalisations linguistiques en corpus. Si les concepts ne préexistent pas à l'analyse, il faut donc trouver un moyen de justifier leur élaboration au cours de l'étude et proposer une méthode qui permette cette élaboration. L'objectif de l'analyse joue un rôle majeur dans la définition des termes et des concepts puisqu'il permet de justifier la normalisation. On perçoit, même intuitivement, que ce qu'on appelle « terme » n'est pas la même chose selon que l'on se place du point de vue de l'indexation, de la traduction ou de la modélisation des connaissances, que l'on peut considérer comme relevant d'autant de genres interprétatifs.
Ma réflexion sur l'identification des termes en corpus est passée par l'exploration de deux voies : l'une vise à repérer des fonctionnements linguistiques « déviants », l'autre s'intéresse aux contextes jouant le rôle de marqueurs de relation. Cette seconde piste étant celle qui a été la plus productive pour moi, je ne ferai que l'évoquer dans ce chapitre et je lui consacrerai entièrement le dernier chapitre. Je ne parlerai ici que de la recherche des termes basée sur le repérage d'un fonctionnement « déviant ».

1.1. Repérage de termes basé sur la notion de « déviance »

Dans mon ambition de rapprocher la terminologie de la linguistique, il m'a semblé opportun, dans une première analyse (je veux dire, première dans la réflexion que j'ai menée), de définir les termes par rapport aux mots ou groupes de mots « généraux ». Au-delà de la seule caractérisation d'un phénomène, il s'agissait, dans une vision opérationnelle, d'essayer de comprendre comment travaillent les terminologues lorsqu'ils identifient des termes. Cette tâche m'a semblé relever souvent de l'identification d'un dysfonctionnement par rapport à un fonctionnement langagier considéré comme régulier, c'est-à-dire normé par des contextes d'utilisation connus et fréquemment rencontrés, c'est-à-dire aussi un fonctionnement qui correspondrait à la compétence des terminologues considérés comme locuteurs moyens.
A la suite de certains auteurs (voir ci-dessous), j'ai appelé ce dysfonctionnement « déviance », terme sans doute malencontreux car il suppose une norme valorisée, par rapport à laquelle on situe des fonctionnements non normés et donc dévalorisés. De mon point de vue, il s'agirait plutôt d'une norme produit d'une normaison, c'est-à-dire d'une régulation spontanée des locuteurs d'une langue.

En fait, cette notion de déviance apparaît dans différents travaux sur la langue. Je vais présenter quelques uns de ces travaux puis, je montrerai comment j'ai essayé de mettre en œuvre cette notion dans le repérage des termes.

1.1.1 Déviance et travaux sur la langue

La notion de déviance, d'écart par rapport à une norme, est assez couramment employée dans des travaux sur la langue apparemment très différents.

Déviance et rhétorique

Une des disciplines qui fait le plus appel à cette idée de déviance est certainement la rhétorique qui foisonne en termes exprimant le fait d'échapper à une norme *: transgression, écart, altération, délit, incorrection* :

> « Dans certains cas, (la métaphore notamment), la rhétorique transgresse visiblement le code lexical en même temps que la règle d'isotopie. » (groupe μ, 1982,38).
>
> « Au sens rhétorique, nous entendons l'écart comme altération ressentie du degré zéro.» (*ibid*, 41).
>
> « L'écart dont nous venons de parler est une altération locale du degré zéro. Il ne présente aucun caractère systématique, et est donc toujours imprévu. Il s'oppose à un autre type d'altération, systématique celui-ci, qui est la convention.» (groupe μ, 1982, 42-43).
>
> «Il est généralement admis que la métaphore repose crucialement sur une "incorrection" ou un "délit", mais ce trait, une fois dénommé et intégré, donne rarement lieu à une caractérisation satisfaisante. » (Kleiber, 1994, 177).

Dans ce type d'approche, l'effet de sens est donc clairement dû à un dysfonctionnement par rapport au code, dysfonctionnement voulu par le locuteur et reconnu comme tel par l'interlocuteur. C'est d'ailleurs le même type de décalage qui produit l'effet comique des mots d'esprits.

Déviance et néologie

On retrouve le recours à la même notion d'écart dans des travaux sur la néologie :

> « Si un changement intervient dans une règle, il se produit au niveau de la performance, sous la forme d'une déviation, d'une "faute", et sa transformation en règle nouvelle implique un usage répété, une longue évolution... le changement des règles grammaticales échappe à la création consciente. Aucun locuteur, en effet, n'a un comportement linguistique naturel qui le conduit à faire volontairement des fautes... Les déviations qui, accumulées, constituent l'usage nouveau, échappent à sa volonté, mais créent la règle nouvelle. » (Guilbert, 1972, 29).

Dans ce cas cependant, il est question d'un écart qui, en se répétant, en n'étant pas le fait d'un locuteur dans un discours particulier, s'intègre peu à peu dans le code lui-même. Ce type de création, contrairement au précédent, n'est alors pas voulu par un individu, mais peu à peu accepté par un ensemble de locuteurs, de manière non-consciente. La déviation n'est alors perçue que par un analyste attentif, le lexicologue par exemple, qui peut repérer le moment où elle est apparue dans le code.

Déviance et terminologie

Généralement, terminologie et néologie ne sont pas considérées comme relevant du même type de processus créatif. En effet, la néologie est considérée comme relevant plutôt d'une évolution inconsciente du système d'une langue, évolution qui relève soit de l'apparition de nouvelles formes (évolution morphologique) soit de l'apparition de nouveaux contextes. En terminologie, dans une vision traditionnelle, la création se fait par un processus volontaire voire normé de nomination ; il ne concerne pratiquement que les lexèmes. Mais ce distinguo ne tient plus si l'on base l'étude sur des productions réelles, qui ne sont pas seulement des

listes de termes mais des textes spécialisés. On peut alors opposer la terminologie à la « langue générale » en décidant que sera considéré comme terme dans un corpus spécialisé tout ce qui est nouveau par rapport au fonctionnement généralement attendu dans cette langue. D'ailleurs il est connu qu'un certain nombre de termes se créent par métaphorisation, c'est-à-dire (si on globalise les propositions présentées ci-dessus de la rhétorique et de la néologie) par création d'un écart (probablement volontaire à un moment donné) puis intégration de cet écart par un ensemble de locuteurs.
On trouve cette idée de déviance dans la modélisation du lexique choisie dans le projet européen Eurotra (Mac Naught et *al.*, 1991, 3) où cette notion, déclinée en trois modes, permet de rendre compte dans une même structure du lexique général et des lexiques spécialisés :

> « In describing SL [Sublanguage] and GL [General Language] one should make use of three complementary points-of-view or modes.
>
> - the restrictive mode: by excluding certain features of GL, SL can be described as a restricted form of language;
>
> - the deviant mode: SL can show specific features which are not found in GL and therefore can be considered a deviant form of GL;
>
> - the preferential mode: this approach of SL phenomena is complementary to the restrictive and deviant modes, and is expressed in terms of preferences. »

J'ai essayé d'exploiter cette idée d'un fonctionnement déviant non pour représenter le fonctionnement du lexique mais pour le repérer en corpus avec comme hypothèse que seraient considérés comme termes les éléments qui ne fonctionneraient pas conformément à la compétence d'un locuteur moyen.

1.1.2 Mise en œuvre de la notion de déviance

Cette étape de ma réflexion a consisté à donner une réalité linguistique et descriptive à cette idée de déviance. J'ai ainsi décrit 4 types de fonctionnements déviants.

Mots ou groupes de mots inconnus

Il s'agit de l' « écart » le plus manifeste, qui attire immédiatement le lexicologue non spécialiste du domaine : par exemple, dans le corpus MMS : *actionneur* ou *étagiste.*
Hormis dans des corpus très spécialisés, ces cas sont rares et la difficulté est bien plus grande de repérer un fonctionnement « étrange » dans un corpus où la quasi totalité des mots sont connus (comme le corpus SGGD, par exemple).
Ces mots nouveaux peuvent aussi provenir de l'application de règles morphologiques connues, sur des bases connues, par exemple *redocumenter* dans MOUGLIS.

Fréquence anormale

Qu'ils s'agissent de mots simples ou de combinaisons, leur fréquence est parfois si élevée qu'ils attirent l'attention ; par exemple *dossier de test* dans MOUGLIS ou *préparation de test* dans MMS. C'est sur cette idée d'une fréquence élevée que se sont constitués les outils d'extraction de termes candidats fonctionnant à partir de segments répétés (*cf.* chapitre III).

Structures elliptiques

L'ellipse peut concerner

- soit un « mot vide » : déterminant ou préposition : *alarme système, service support, décommutation télémesure* ;

– soit un mot plein par exemple l'objet d'un verbe comme dans l'énoncé issu du domaine bancaire : *vous pouvez déposer librement sur votre compte*.

Ce second type d'ellipse a été repéré par Noailly, qui ne travaille pas pourtant sur la terminologie :

> « Par "emploi absolu" d'un verbe, on entend des emplois où le complément du verbe transitif, direct ou indirect, est absent, sans que cela implique que le verbe en question ait globalement changé de sens...cet objet, s'il est nécessairement existant dans l'univers de référence, est linguistiquement considéré comme sans pertinence. » (Noailly, 1996a, 74).

Pour Harris aussi :

> « La forte probabilité d'un élément est aussi un point-clé de la contrainte de *réduction*. » (Habert et Zweigenbaum, 2002b).

On peut penser qu'un univers de référence récurrent, comme l'est par exemple une situation de travail, conduise à l'instauration régulière de l'ellipse, au point qu'elle fasse partie du code linguistique à l'œuvre dans cet univers de référence.

Combinaisons « anormales » de mots ou groupes de mots

– Relation entre verbe et arguments :

- apparition d'un nouvel argument (anomalie syntaxique)
par exemple, dans MMS : *geler un test* (pas d'objet attendu).
- anomalie sémantique d'un argument :
par exemple, *documenter un logiciel*, dans MOUGLIS (argument humain attendu).
Ces combinaisons seront considérées comme des termes.

– Coordination inattendue :

Par exemple dans le corpus MMS :
Il n'y a pas pour ces équipements de chien de garde ou de logique de reconfiguration.
Dans ce cas, *chien de garde* est considéré comme un terme parce qu'il est coordonné à une combinaison considérée comme un terme.

Il y aurait sans doute d'autres caractérisations possibles mais je n'ai pas poursuivi dans cette direction. En effet, je me suis rapidement rendu compte que la mise en œuvre systématique de cette approche supposait le recours à une référence (à défaut d'une norme) et que cela paraissait très difficile à réaliser. On peut espèrer, comme le font Habert et Zweigenbaum que :

> « Le recours désormais possible à des très gros volumes de données textuelles amène à questionner sur des bases partiellement renouvelées la validité de l'intuition d'acceptabilité […] » (Habert et Zweigenbaum, 2002b).

Mais il reste à mettre sur pied des études visant à comparer des corpus spécialisés avec des corpus équilibrés pour évaluer la pertinence de ce type d'approche.

1.2. Le recours systématique à une norme est-il possible ?

Si l'on accepte l'idée de fonctionnements perçus comme déviants, il faut donc accepter l'idée de norme. Mais pour que cette idée puisse être mise en œuvre de manière systématique, il faut aussi qu'on puisse, à un moment, considérer que l'on peut fixer des régularités et leur donner un contenu descriptif. C'est précisément la position que défend Saussure lorsqu'il parle de la langue comme objet de la linguistique puisqu'il parle du caractère de fixité de la langue :

> « [La langue] est à la fois un produit social de la faculté de langage et un ensemble de conventions nécessaires, adoptées par le corpus social pour permettre l'exercice de cette faculté chez les individus…La langue …est un tout en soi et un principe de classification ». (Saussure, 1982, 25)

« Si la langue a un caractère de fixité, ce n'est pas seulement parce qu'elle est attachée au poids de la collectivité, c'est aussi qu'elle est située dans le temps… » (*ibid*, 108)

Cette description du contenu de la norme linguistique se fait généralement par introspection, et on ne voit pas comment il pourrait en être autrement. La notion de norme va avec celle de compétence linguistique, qui serait quasi constante d'un locuteur à l'autre, dans la vision structuraliste. Cette analyse, propre à la sémantique lexicale, mobilise d'ailleurs de nombreux chercheurs. De mon point de vue, le problème n'est pas tant dans cette similitude de compétence supposée d'un locuteur à l'autre mais plutôt dans la croyance que l'introspection permettrait de décrire cette compétence. Bien souvent l'introspection en sémantique consiste à imaginer des contextes pour des éléments lexicaux afin d'en mettre au jour le sens ; la sémantique lexicale mais aussi la lexicologie par exemple mettent en œuvre ce principe. Il est alors étonnant de constater que ce qu'on appelle la compétence en sémantique semble s'articuler autour de mots partagés par l'ensemble des locuteurs, beaucoup plus que de fragments discursifs. L'identification du sens des mots consiste alors à imaginer « toutes les variantes possibles du sens selon la situation et selon les éléments de la situation pris en compte » (Ducrot, 1980, 18). Par cette introspection, ce que l'on recherche n'est pas tant de retrouver les traces mémorielles de notre contact avec ces mots mais à construire une série de contextes que nous soumettons au filtre de ce qui est généralement considéré comme langagièrement correct et qui nous permet d'appartenir à un groupe dont la cohésion est ainsi garantie.

Ainsi, appartenir à une communauté langagière (avoir une compétence linguistique au sens «classique » du terme), c'est accepter d'oublier son expérience individuelle pour participer à une expérience collective. C'est en référence à cette neutralisation de l'expérience individuelle, qui passe par un oubli des contextes d'apprentissage individuels et une adaptation à une vision collective, que l'on peut parler de compétence linguistique similaire d'un locuteur à l'autre. Or, on voit mal comment l'introspection, qui relève d'une activité spécifiquement individuelle, permettrait de n'identifier que les contextes auxquels tous les locuteurs d'une langue ont été confrontés et ainsi de délimiter les sens des mots. Nécessairement, les contours de ce qui relève de l'individuel et du collectif (ou du plus ou moins collectif) sont flous. La recontextualisation, le retour au discursif, fait se rencontrer expérience individuelle et expérience collective et le consensus est bien difficile à trouver, d'où les interminables discussions de linguistes sur le mode « ça se dit/ça ne se dit pas ». Ainsi, il est bien difficile de décider ce qu'est la langue générale, qui pourrait servir de norme. Il s'agirait sans doute d'une langue qui a été apprise dans des contextes (des expériences) culturellement importants dans la communauté linguistique, proches d'un individu à l'autre, fréquemment renouvelés et qui s'accompagne d'une minimisation des contextes non reconnus « collectivement » ; c'est-à-dire non seulement des contextes socialement dévalorisés, comme le pensent les sociolinguistes, mais aussi des contextes qui ne concernent qu'un petit groupe de locuteurs et qui mettent en œuvre des liens psychologiques autant que sociaux mais qui ne sont ni prédictibles ni contrôlables (contexte familial, scolaire, amical, sportif, associatif…). Il est pourtant probable que, dès qu'il y a interaction langagière, à condition que ce soit dans un contexte qui vise un partage d'objectif ou d'intention, des régularités s'instaurent qui échappent à tout contrôle.

Une certaine compétence linguistique commune existe parce que la langue constitue un facteur de cohésion et de structuration important d'une communauté, mais il n'est pas possible d'en fixer les contours au point de pouvoir l'utiliser comme référence dans le cadre d'une analyse de corpus. En effet, cette compétence ne concerne pas un « locuteur idéal », qui n'a pas d'existence mais un ensemble de locuteurs individuels qui ont été (et qui continuent à

être) à la fois confrontés à des expériences individuelles d'apprentissage et à la volonté d'appartenir à une communauté[43].
Cette position est assez proche de celle défendue par la praxématique :

> « Le sémantisme de tout terme se constitue par l'enregistrement, au fil de l'histoire, de diverses praxis associées à l'être ou à l'objet désigné ». (Siblot, 1996, 56).

Toutefois, la praxématique n'explique pas le passage de l'expérience individuelle à l'établissement d'une compétence collective. La notion même d'expérience n'est perçue que dans sa dimension « pratique » :

> « Notre rapport aux objets, ainsi que Bréal le souligne, est avant tout pratique ; nous ne nous intéressons pas à eux " pour eux-mêmes " mais en raison de l'intérêt qu'ils présentent « pour nous ». (*ibid*, 59).

Le « nous » de Siblot et de Bréal est-il individuel ou collectif ?
De mon point de vue, le « nous » individuel serait concerné par l'expérience personnelle du lien entre un phénomène langagier et un contexte d'apprentissage (locuteur, interlocuteur : qui, où, quand, pourquoi, conséquence…) ; le passage à un « nous » collectif se ferait par la décontextualisation individuelle ou plutôt par l'oubli de contextes individuels , qui seraient remplacés par des contextes reconnus (voire valorisés) collectivement. Ce n'est pas tant le réel qui est important mais l'expérience individuelle qu'on en a et qu'on essaie de rapprocher d'une expérience collective, dès lors que nous sommes des êtres parlants.
La confrontation avec un corpus constitue une nouvelle expérience langagière, qui correspond ou non à une expérience personnelle de locuteur-auditeur que nous avons déjà eue, et à une compétence langagière que nous avons élaborée pour participer à la cohésion langagière du groupe des locuteurs de la langue utilisée. La compétence linguistique (la compétence du linguiste, qui relève d'une compétence professionnelle) consiste alors à voir comment nos expériences précédentes (de locuteur et de linguiste) peuvent être convoquées pour mettre de l'ordre dans cette nouvelle expérience langagière et lui donner un sens, c'est-à-dire l'interpréter. La notion d'intuition trouve alors toute sa place dans ce processus et elle n'est pas contradictoire avec celle de rigueur et de scientificité.

Si la langue générale n'a pas de contenu précis, le partage entre langue générale et langue spécialisée est particulièrement difficile à faire. On voit bien par exemple l'embarras des lexicologues qui élaborent les dictionnaires : d'une part, ils ne savent pas s'ils doivent ou non retenir certains mots (sont-ils généraux ou non ?), d'autre part, ils ne savent pas dans quels « domaines » les ranger : pour le corpus MMS, pourtant assez resserré du point de vue de la thématique (simulation de satellites), certains termes étaient définis dans le Petit Robert mais ils étaient répartis dans douze disciplines différentes !

On peut essayer d'approcher les éléments qui constituent la compétence linguistique, c'est-à-dire ceux qui sont les mieux partagés par l'ensemble des locuteurs. Et la linguistique dite introspective a déjà permis la description de quelques phénomènes qui leur a donné un certain

[43] La notion d'expérience est souvent convoquée pour justifier l'existence d'universaux sémantiques (par exemple, dans (Martin, 1983, 88) : « Dès lors pointe l'idée des universaux d'expérience. Primitifs d'une autre nature, ils ne visent pas, en tant que tels, l'axiomatisation d'un système sémantique. Ils viennent plutôt de l'idée que certaines données du monde, physiques, physiologiques, anthropo-culturelles, exercent sur la vie des hommes une si forte contrainte qu'il est impensable qu'elles ne laissent aucune trace dans la langue. Et ces traces, du fait même, ont toutes les chances d'être des universaux »). C'est oublier que l'expérience, avant d'être collective est individuelle. C'est par immersion dans un contexte linguistique que cette expérience est filtrée et perçue comme partagée.

éclairage. Il n'est pas étonnant que les recherches en syntaxe soient ainsi bien plus développées qu'en sémantique. Les règles y paraissent plus figées et s'élaborent particulièrement sur la notion d'impossibilité (d'interdit ?) : certaines structures syntaxiques sont ainsi unanimement refusées par l'ensemble des locuteurs. En revanche, dès que l'on se rapproche d'une dimension plus sémantique, plus proche de l'expérience individuelle, l'unanimité est beaucoup moins sûre. Ainsi, dès que l'on examine introspectivement les mots dans leurs possibilités de rapport de cohésion avec d'autres, le consensus est nettement moins ferme : par exemple dans le cas de l'identification des structures argumentales de verbes : il n'est pas si aisé de définir quels arguments sont « obligatoires ». Il en va de même en morphologie, où les créations inattendues ne sont pas rares tout en étant parfaitement « acceptables ». Il y a des tendances générales, que l'on repère de mieux en mieux, mais rien qui puisse servir définitivement de référence.
Enfin, et cela ne simplifie pas la tâche, la compétence langagière est en perpétuelle évolution ; sans doute le noyau dur, le plus consensuel, évolue-t-il lentement ; mais, en matière de sémantique par exemple, il semble que la souplesse, le dynamisme et l'adaptation prévalent.
Pour revenir à mon propos initial, c'est peut-être cet élément de souplesse de la compétence sémantique qui limite le plus la mise en œuvre d'une identification des termes par reconnaissance de dysfonctionnements. En effet, un phénomène d'adaptation se produit très rapidement sur un nouveau corpus, qui fait qu'on ne repère plus les éléments qui, à la première lecture, sont apparus comme « dysfonctionnants ».

1.3. La normalisation comme recherche d'une cohésion interne

Lors de l'analyse d'un corpus spécialisé, la compétence linguistique (au sens de compétence de locuteur et compétence d'analyste de textes) ne se met pas en place sous la forme du recours à une norme, préexistante, mais dans la recherche d'une stabilisation des phénomènes propres à ce corpus en lien avec l'objectif de l'étude. Il s'agit moins d'apparier les phénomènes que l'on rencontre avec des phénomènes que l'on convoque de manière introspective mais plutôt, à l'aide de connaissances linguistiques plus ou moins latentes (relevant de la compétence de locuteur) d'essayer d'élaborer une cohérence interne, propre au corpus et souvent guidée par l'objectif de l'analyse (cette élaboration relevant de la compétence de linguiste). Ainsi, toute analyse linguistique de corpus (en particulier toute analyse sémantique) relève d'une élaboration nourrie à la fois par la compétence linguistique, par les manifestations du corpus et par l'objectif de l'analyse.

D'une certaine façon, la dimension spécialisée d'un corpus permet de mettre l'accent sur la façon dont l'analyse de corpus se met en place ; en effet, l'analyste n'a nécessairement pas la compétence de locuteur pour ce corpus. Ce qui est particulier aux corpus spécialisés vient de ce que le linguiste sait qu'il n'est pas locuteur expert, qu'il doit donc d'emblée ne pas considérer son analyse comme une simple projection de sa compétence de locuteur mais qu'il peut, dans un certain nombre de cas, s'appuyer à la fois sur cette compétence et sur sa compétence de professionnel de la langue pour mettre au jour une cohérence propre y compris dans les cas où les manifestations du corpus sont inconnues, voire incompatibles avec ce qu'est sa compétence de locuteur. Or, cette situation extrême permet de mettre en lumière un élément que tout linguiste devrait avoir en tête quel que soit le corpus qu'il étudie : il ne peut pas avoir d'*a priori* sur ce qu'il va trouver et il doit s'attendre à ce que sa compétence de locuteur ne corresponde pas aux fonctionnements linguistiques du corpus. Ainsi, la compétence du linguiste de corpus consiste, pour une bonne part, à se méfier de ce qu'il croit être sa compétence de locuteur et, en tout cas, à la maintenir à un niveau de latence qui laisse toute sa place à l'interprétation des phénomènes propres au corpus.

Dans une telle perspective, le problème du repérage des termes n'est pas premier : la recherche d'une cohérence convoque tous les éléments du corpus, syntaxiques, morphologiques et, bien sûr, sémantiques. Ce n'est finalement qu'au terme d'une analyse approfondie du corpus, quand tous les éléments s'ordonnent, que des éléments lexicaux, auxquels on peut donner un rôle prépondérant en fonction d'un objectif, apparaissent. Ces éléments pourront être considérés comme des termes et pourront certainement jouer un rôle important dans un système de représentation comme celui des Bases de Connaissances Terminologiques. Construire une BCT en commençant directement par identifier les termes revient ainsi à faire l'économie d'une réflexion approfondie sur la façon dont on les sélectionne. Et cette économie est parfois légitime : les terminologues et les ingénieurs de la connaissance n'ont ni le temps ni les moyens d'expliquer leurs choix et privilégient une sélection intuitive, quitte à adapter leurs résultats en fonction de la demande qui leur est faite. Il est naturel que la réflexion plus approfondie revienne aux linguistes et on peut espérer que les résultats obtenus facilitent, à terme, le travail des ingénieurs de la connaissance et des terminologues.
Dans une telle perspective, il m'a semblé que le plus important dans un premier temps était de focaliser l'attention sur la mise en relation des éléments dans le corpus plutôt que sur les éléments lexicaux eux-mêmes et donc de m'intéresser tout particulièrement aux marques de cette mise en relation. Au point que la définition de *terme* que je donnerai est la suivante :
« Terme : élément ou groupe d'éléments linguistiques mis en relation dans un corpus par d'autres éléments linguistiques que l'on peut utiliser comme marqueurs de cette relation. »
Cette problématique fera l'objet du cinquième chapitre.

1.4. Recherche de cohésion interne et acquisition de connaissances à partir de textes

Je n'ai fait ce constat de l'impossibilité de définir une norme *a priori* qu'après avoir moi-même tenté de constituer une sorte de grammaire sémantique qui avait pour vocation d'être utilisée pour apprendre du vocabulaire nouveau, dans le domaine bancaire[44].
Ma contribution était double. D'une part, il s'agissait d'élaborer un mode de représentation des données qui puisse être implémenté facilement ; c'est d'ailleurs ce qui a été fait par Patrick Saint Dizier (Saint Dizier et Condamines, 1990). Ce mode de représentation consistait en une organisation arborescente, aussi bien pour classifier les modes de fonctionnement syntaxiques que sémantiques. D'autre part, il s'agissait de caractériser le sens des verbes, des prépositions et des noms afin d'élaborer une méthode d'identification de la classe d'appartenance sémantique d'un argument inconnu, à partir de la connaissance de la classe de fonctionnement du verbe (classe sémantique et type d'arguments attendu, et mode de construction du verbe, directe ou pas) et du sémantisme de la préposition. Les propositions que j'ai pu faire alors se sont beaucoup inspirées d'une grammaire sémantique telle que celle décrite par Fillmore (Fillmore, 1968), qui place le prédicat au centre du processus de compréhension. La mise en oeuvre de ces connaissances s'opérait d'après la synthèse présentée à la page suivante (figure 5).
La mise en œuvre de ce « modèle » supposait que soient préalablement déterminés le fonctionnement syntaxique et argumental des verbes, la ou les classes d'appartenance possible(s) des noms et les types d'interprétations sémantiques possibles des prépositions (Condamines, 1993).

[44] Il s'agissait d'un étude réalisée dans le cadre d'un contrat Irit/Syseca dont le responsable était P. Saint Dizier. Cette expérience (1990) a servi de support à ma première réflexion en lexicologie spécialisée.

Tout au long de l'élaboration de ces classes sémantiques et sémantico-syntaxiques, j'ai eu conscience d'une difficulté à fixer les fonctionnements. En revanche, il m'a semblé qu'il y avait une certaine pertinence dans le fait qu'il était possible de mettre en œuvre une connaissance (une compétence) linguistique et s'appuyer sur elle pour émettre des hypothèses sur le sens de mots « inconnus ». Ainsi, le problème se situe plus dans la détermination de cette connaissance *a priori* que dans le fait qu'il faut chercher à dégager une cohésion interne en identifiant une cohérence relationnelle entre les éléments (en particulier mots et groupes de mots). Cette recherche de cohésion est sans doute assez proche de la recherche d'un état stable, au sens de Victorri et Fuchs :

> « L'image qui se dégage est donc celle d'un processus d'optimisation des interactions entre composantes de l'énoncé, qui conduit, chaque fois que cela est possible, à un état stable qui constitue le sens de l'énoncé et de ses composantes. » (Victorri et Fuchs, 1996, 43).

A mon avis cependant l'état stable n'est pas inhérent à l'énoncé mais le fruit d'une interprétation éclairée par différents paramètres.

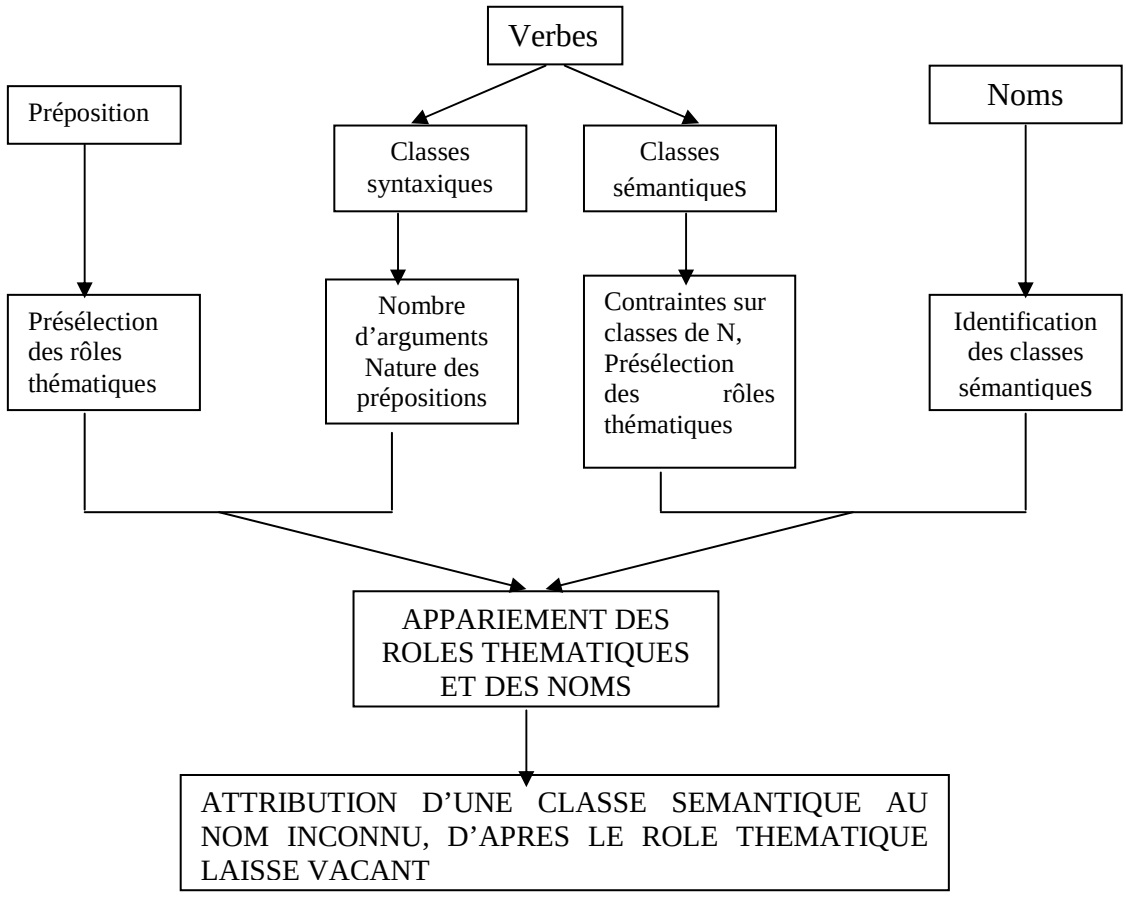


Figure 5 **:** Mode d'attribution d'une classe sémantique, dans le cas d'une grammaire sémantique

Finalement, si mon exploration de la « déviance terminologique» ne s'est pas poursuivie, c'est plus parce qu'il me semblait impossible d'établir une référence fixe qui pourrait être utilisée systématiquement plutôt que parce que l'idée elle-même n'était pas pertinente. Le fait d'être surpris par un fonctionnement langagier est un des éléments qui met en éveil le linguiste de corpus (tout comme le terminologue) et qui le pousse à s'intéresser à certains phénomènes.

Mais il est à craindre (à moins que ce soit un atout) que l'effet de surprise ne soit pas nécessairement le même d'un analyste à l'autre.

2. Polysémie en corpus spécialisé

Le deuxième phénomène sur lequel je me suis interrogée est celui de la polysémie. Dans une vision classique de la terminologie, la question de la polysémie est vite réglée puisque, par principe, il n'y a pas de termes polysémiques. On parle ainsi « d'univocité » entre le terme et le concept.
Ce point de vue revient à considérer que la terminologie ne relève pas d'un fonctionnement sémantique. En effet, en sémantique, la question de la polysémie fédère la quasi totalité des problèmes que posent le sens et, comme le dit Kleiber,

> « […] définir […] ce qu'on entend par sens différents et sens apparentés [deux éléments qui interviennent dans la définition de la polysémie] exige une prise de position sur la plupart des problèmes sémantiques généraux. » (Kleiber, 1999, 56).

Certes, comme Rastier, je crois que la polysémie est un artefact de linguiste :

> « Posé hors contexte, le problème de la polysémie, aussi lancinant qu'insoluble, est pour une large part un artefact des linguistes. Nous n'estimons pas que les lexies soient par elles-mêmes polysémiques en contexte : ce sont les parcours interprétatifs qui sont multiples.». (Rastier, 1996, 17).

En effet, l'« identification » d'une polysémie relève toujours d'une élaboration de linguiste ou de lexicographe qui, le plus souvent sur la base de sa compétence linguistique, décide d'identifier plusieurs sens « reliés », pour un même mot. Comme le dit Benveniste :

> « Ce qu'on appelle la polysémie n'est que la somme institutionnalisée, si l'on peut dire, de ces valeurs contextuelles, toujours instantanées, aptes continuellement à s'enrichir, à disparaître, bref, sans permanence, sans valeur constante.» (Benveniste, 1974, 227).

Dans le cadre de l'élaboration d'une polysémie à partir d'une analyse de corpus, il s'agit d'artefacts qui se justifient de deux manières, d'une part par le fait du repérage de fonctionnements syntaxiquement réguliers en corpus (c'est ce que Noailly note aussi dans son analyse du mot *fleuve*, *cf.* ci-dessous) et qui autorisent une stabilisation, d'autre part par le fait d'une interprétation sémantique guidée par l'objectif de l'analyse. Je montrerai ces deux éléments dans la présentation de l'analyse du mot *satellite* dans un corpus du CNES.

Dans la plupart des cas toutefois, la polysémie est considérée dans une vision descendante de son existence supposée en langue à sa réalisation en discours, qui permettrait de retrouver les différentes acceptions que l'on a identifiées par introspection :

> « Décrire la polysémie d'une expression, c'est d'une part dresser la carte des sens possibles de cette expression, et d'autre part prédire le sens qu'elle prend selon les contextes. » (Fuchs, 1997, 127).

Ainsi, si le rôle du contexte est examiné, c'est pour voir comment il contribue à résoudre l'ambiguïté que crée la polysémie mais pas pour étudier comment il peut contribuer à l'établir. Par exemple, lorsque Noailly étudie la polysémie du mot *fleuve*, elle définit *a priori* un ensemble de sèmes inhérents et étudie comment le contexte syntaxique permet d'identifier les sèmes afférents ; et elle constate que « la « polysémie » de fleuve repose sur des bases homogènes et cohérentes » » (Noailly, 1996b, 37) et que « contrairement aux sémanticiens en général, et à Rastier en particulier, [elle] ne lie pas (ou plutôt, [elle] lie moins) ces variations au contexte, c'est-à-dire aux isotopies en présence, et [elle] les considère comme déterminées principalement, et simplement, par la syntaxe.» (*ibid*, 37).

C'est seulement chez les analystes de discours que le phénomène de la polysémie est interrogé dans son rapport avec la construction du sens à partir des données et dans son rapport avec l'extra-linguistique :

> « J'ai montré comment la polysémie constitutive d'une notion comme " qualification " n'a pas été traitée de la même façon par le groupe social des enquêteurs et par celui des enquêtés […]. De façon indépendante de la sociologie j'ai établi des faits de divergence comme de convergence dans l'activité de construction du sens des sujets : convergence au sein du groupe ouvrier qui catégorise les situations de travail selon une même dimension, celle des activités ; divergence entre OS et OP dans l'expression des agents sociaux du travail ; divergence majeure enfin entre les enquêteurs et les enquêtés dans la construction du sens de " qualification ". » (Boutet, 1995a, 155-156).

La nature spécialisée du corpus crée, une fois de plus, un lieu d'expérimentation original, qui est peut-être la version agrandie de ce qu'est l'analyse de corpus en général : sans compétence de locuteur, il est impossible au linguiste de définir *a priori* quels éléments vont être polysémiques et combien de sens peuvent être repérés.
Si l'on accepte l'idée qu'il y a de la polysémie dans les corpus spécialisés mais que pour autant, tous les termes ne sont pas polysémiques au même titre, il est nécessaire de mettre en place des méthodes qui permettent de repérer la polysémie des termes en corpus. C'est dans cette entreprise que nous nous sommes lancées avec Josette Rebeyrolle, lors d'un stage de DEA qu'elle a effectué au CNES (Centre National d'Etudes Spatiales) sur la question des points de vue.
Le premier de nos constats a concerné le fait que très peu de travaux existent dans ce domaine, à la fois sur la polysémie dans les corpus spécialisés et sur l'identification ascendante d'une polysémie. Dans une telle vision, il faut repérer différentes acceptions possibles, que l'on élabore en une interprétation de polysémie, la polysémie n'étant que la caractérisation *a posteriori* d'un phénomène qui se manifeste par l'existence d'une polyacception.
Comme je l'ai déjà mentionné dans le chapitre III, la polysémie en lien avec l'analyse de corpus se manifeste par l'identification de classes de contextes que l'on peut considérer comme différentes et, en principe, par au moins une classe de contextes qui neutralise les différences.
Nous avons essayé de mettre en œuvre ce type d'approche dans l'analyse contextuelle du mot *satellite*.

2.1. Identification des acceptions possibles de satellite dans un corpus du CNES

Précisons d'emblée que ce travail sur la polyacception du mot *satellite* a été réalisé dans le cadre d'un projet avec le CNES qui concernait la notion de point de vue. Pour le CNES, il s'agissait de voir comment ces points de vue influençaient le processus de conception. Pour nous, il s'agissait de voir quel contenu linguistique pouvait avoir cette notion de point de vue.
Notre réflexion s'est rapidement orientée sur la question de la polysémie en lien avec l'analyse de corpus. Parmi les différentes analyses qu'a menées Rebeyrolle (*cf.* Rebeyrolle, 1995), l'une a porté plus précisément sur l'analyse contextuelle du mot *satellit*e. Il s'agissait dans un premier temps de relever les occurrences du mot *satellite* et de les classer sur des bases sémantico-syntaxiques afin de constituer des classes de contextes. La possibilité d'identifier plusieurs classes de contextes était considéré comme l'indice d'un fonctionnement polysémique.
Le corpus était constitué d'une part de documents techniques sur le projet DIODE (qui concerne l'autonomisation des satellites vis-à-vis des centres de contrôle) et d'autre part d'interviews de spécialistes. Une autre répartition nous a plus particulièrement intéressées : les documents, quel que soit leur genre, émanaient de trois Divisions du CNES : Observation

de la Terre, Systèmes Electriques et Automatiques, Mathématiques Spatiales. Nous verrons que cette partition a eu un rôle majeur dans l'étude des points de vue. Mais dans un premier temps, les occurrences ont été étudiées sans tenir compte de leur provenance.
7 classes de fonctionnements ont été repérées ((Rebeyrolle, 95) et (Condamines et Rebeyrolle, 1996)), en tenant compte essentiellement du rôle du mot *satellite* par rapport au prédicat ; ainsi, sur les 178 contextes d'apparition de *satellite*, 42 n'ont pas été pris en considération.

Remarques de notation : DVB : nominalisation déverbale, N1 : *satellite* ou un hyponyme de *satellite*.

– Structures dans lesquelles *satellite* est en position d'objet

A- V de fabrication + dét + N1

construire le satellite
réaliser le satellite
fabriquer le satellite

```
DVB de fabrication + de + dét + N1
```

construction du satellite
réalisation du satellite
fabrication du satellite

B- V de type *projeter dans l'espace* + dét + N1

lancer des satellites en orbite

```
Vsupport + dét + DVB de type projeter dans l'espace + de + dét + N1
```

effectuer le lancement du satellite
effectuer le tir de SPOT4

C- V de situation dans l'espace + dét + N1

positionner le satellite
localiser le satellite

```
Vsupport + dét + DVB + de + dét + N1
```

déterminer la position du satellite
calculer la position du satellite
prédire la position du satellite
calculer, délivrer une estimation de [la] position du satellite

– Structures dans lesquelles *satellite* est précédé d'une préposition

D- V de type *placer* + dét + N2 + sur + dét + N1

placer un système sur le satellite
placer un senseur stellaire sur le satellite
poser des fixations sur le satellite

```
Vsupport + dét + DVB placer + de + N2 + sur + dét + N1
```

réaliser l'embarquement du prototype sur le satellite SPOT4
faire l'intégration des instruments sur les satellites porteurs
faire l'intégration des essais électriques, essais mécaniques sur le satellite
réaliser l'embarquement du navigateur autonome sur le satellite

E- Vsupport + dét + DVB + de + N2 +à bord de + dét + N

faire le contrôle d'orbite à bord du satellite
reléguer les prétraitements à bord du satellite

réaliser les traitements à bord du satellite
fournir des estimations de navigation à bord d'un satellite
connaître les paramètres de navigation à bord de satellite
réaliser l'application navigation à bord du satellite
réaliser le contrôle d'orbite autonome à bord du satellite
réaliser la probation en vol du calcul d'orbite à bord du satellite

F- V de type *transmettre* + dét + N2 + à + dét + N1

envoyer les paramètres d'orbite au satellite
`Vsupport + dét + DVB de type transmettre + de + N2 + à + dét + N1`
réaliser l'envoi de télécommandes au satellite
faire la transmission de téléchargements / télécommandes aux satellites

G- V de mouvement + par/via + dét + N1

[*les données de DIODE*] *descendent via le satellite*
[*la voie de données*] *passe par le satellite.*

Ces classes ont été repérées par la mise en œuvre des connaissances linguistiques suivantes :

- équivalence verbe/verbe support + déverbal (noté ici DVB) : *positionner le satellite/déterminer la position du satellite*,
- récurrence de la nature des constructions argumentales : directe ou indirecte, à un ou deux arguments,
- équivalence de prépositions : *par* et *via* sont considérées comme équivalentes,
- équivalence « sémantique » de structures verbales : *construire/réaliser/fabriquer, positionner/localiser, réaliser l'envoi/faire la transmission de*.

Par ailleurs, une connaissance est présente dans une autre partie du corpus (ou un autre corpus, peu importe ici) qui concerne l'existence d'une hyperonymie entre *satellite* et *Diode* et *satellite* et *Spot4*.

On peut considérer qu'avec ces 7 classes, on se trouve dans le cas d'un fonctionnement polysémique tel que je l'ai décrit. En effet, il me semble qu'on peut dire que les classes C à F correspondent à des sens particuliers, les structures syntaxico-sémantiques dans lesquelles apparaissent le mot *satellite* permettent en effet de donner un sens assez précis à *satellite*. En revanche, les classes A et B font apparaître *satellite* comme objet d'un verbe très général, appartenant soit à la classe *concevoir*, soit à la classe *lancer* ; de très nombreux noms pourraient se trouver à cette place (de très nombreux objets peuvent être lancés ou conçus). On peut donc considérer que ces deux contextes neutralisent les spécificités des contextes B à F qui, eux, permettent de circonscrire des sens plus caractéristiques de *satellite :* objet sur lequel on peut placer des éléments bien précis, objet à l'intérieur duquel on effectue certaines actions définies, objet qui reçoit certaines informations, objet à travers lequel circulent certaines informations.

Même s'il faut recourir, indubitablement, à une interprétation des contextes pour construire ces classes, il est possible de leur donner une assise syntaxique et de trouver une homogénéité sémantique qui justifie leur pertinence. Il est évident aussi que le contexte de l'étude a certainement eu une influence sur l'élaboration de ces classes ; en effet, s'il est un objet sur lequel la notion de point de vue peut s'exercer au CNES, c'est bien sur le satellite ! Nous nous attendions donc à des acceptions différentes et nombreuses dans le corpus et il se peut que cette attente ait influencé en partie les résultats. Si l'on accepte la notion de genre interprétatif,

on peut considérer qu'il a eu une importance plus grande que le genre textuel, les corpus techniques étant réputés pauvres en manifestations polysémiques.
Les résultats de cette étude, au-delà de la seule analyse linguistique, ont pris une pertinence toute particulière sur la question des points de vue lorsqu'on les a proposés à nos partenaires du CNES.

2.2. Polysémie et point de vue

Nous avons proposé les 7 classes identifiées grâce à l'analyse des contextes d'apparition de *satellite* à un expert du CNES. L'examen de ces résultats à amené cet expert à donner un nom à chaque classe en lien avec le point de vue sur *satellite* qu'elle lui semblait manifester :

- Objet à concevoir,
- Corps artificiel
- Mobile
- Plate-forme (préposition *sur*) ; la plate-forme étant (considérée comme) une partie du satellite.
- Véhicule (préposition *à bord de)*
- Hôte (préposition *via/par*).

L'étude s'est poursuivie par la prise en compte de la répartition de ces 7 points de vue dans chacun des trois sous-corpus en provenance chacun d'une des trois Divisions : Observation de la Terre (D1), Systèmes Electriques et Automatiques (D2) et Mathématiques Spatiales (D3).
Les résultats obtenus sont les suivants (extrait de (Rebeyrolle, 1995)).

	Objet à concevoir	Corps artificiel	Mobile	Plate-forme	Véhicule	Hôte	Relais	
Interview	6 4,4%	0 0%	4 2,9%	4 2,9%	0 0%	6 4,4%	10 7,3%	30 22%
D1	0 0%	6 4,4%	3 2,2%	20 **14,7%**	0 0%	0 0%	0 0%	29 21,3%
D2	0 0%	0 0%	6 4,4%	8 5,8%	0 0%	0 0%	0 0%	14 10,2%
D3	0 0%	1 0,7%	28 **20,5%**	10 7,3%	13 **9,5%**	9 6,6%	2 1,6%	63 **46,3%**
	6 4,4%	7 5,1%	41 30,1%	42 30,8%	13 9,5%	15 11%	12 8,8%	136 100%

Tableau 2 : Répartition des points de vue dans le corpus du CNES

Les résultats présentés dans ce tableau permettent de nombreuses conclusions :

- Deux points de vue sont nettement majoritaires dans ce corpus (au CNES ?) : le satellite en tant que mobile et le satellite en tant que plate-forme qui correspondent chacun à un peu plus de 30% des utilisations.
- Ces deux points de vue sont les seuls présents dans la Division D2.
- Dans les deux autres Divisions, un point de vue est dominant : « plate-forme » en D1 (3 fois plus utilisé que le suivant : « corps artificiel »), « mobile » dans la Division D3 (2 fois plus utilisé que le suivant : « véhicule »).
- Le point de vue « objet à concevoir » n'apparaît que dans les entretiens ; le point de vue « véhicule » n'apparaît que dans la Division D3 ; le point de vue « corps artificiel » n'apparaît pratiquement que dans la Division 3.

Ces résultats ont été particulièrement intéressants pour nos interlocuteurs du CNES qui ont vu apparaître une répartition de points de vue élaborée sur des bases linguistiques. Inversement, cette répartition, à laquelle les experts du CNES ont attribué une réelle pertinence et une réelle utilité, est venue comme une sorte de validation de notre classement et, par conséquent, de la possibilité d'accorder un certain crédit à notre intuition linguistique (de locuteur et de linguiste) à condition qu'elle soit balisée d'une part par un corpus minutieusement constitué et d'autre part par un objectif clairement énoncé.

2.3. Repérage des mots polysémiques : recours à la statistique

Jusqu'à présent, il n'a été question que de la possibilité de constituer des classes de contextes qui, en mettant au jour une polyacception, permettent de justifier l'élaboration d'une polysémie. Le problème reste entier sur la façon de repérer les mots polysémiques. Soit on considère que tous les termes peuvent être également polysémiques et il n'est plus possible alors d'élaborer de classes de fonctionnement puisque cela suppose que chaque mot apparaît dans des contextes également polysémiques, ce qui rend impossible leur stabilisation. Soit on considère que certains mots sont particulièrement polysémiques, et il faut trouver un moyen de les repérer sans qu'il soit nécessaire, dans un premier temps, de recourir à une analyse linguistique détaillée.
Nous avons donc commencé à mettre en place une approche statistique qui nous permette de pointer sur les mots ou groupes de mots potentiellement polysémiques. Avec Max Reinert, statisticien et créateur du logiciel Alceste, nous avons commencé à réfléchir à la possibilité de mettre en place une classification ascendante hiérarchique. Sur ce corpus du CNES, les résultats étaient très préliminaires. Mais nous espérons reprendre notre réflexion sur un nouveau corpus du CNES, dans le cadre d'un projet sur l'évolution des connaissances dans le temps et les traces linguistiques de cette évolution (projet « CNES2 »).

Recadrée dans la perspective d'une analyse qui se fonde sur un corpus et qui a un objectif défini, la question de la polysémie s'éclaire d'un nouveau jour. Elle permet en particulier de mettre en évidence le lien qui s'instaure entre l'objectif de l'analyse (relevant d'un genre interprétatif) et les régularités que l'on observe. En effet, même les régularités dites syntaxiques s'appuient toujours sur des éléments sémantiques qui ne peuvent relever du seul fonctionnement immanent.

3. Les nominalisations

Les études sur les nominalisations qui sont présentées ici ne relèvent pas, à proprement parler, de la constitution de bases de connaissances terminologiques. Elles se situent plutôt dans la perspective de décrire un fonctionnement propre aux corpus spécialisés et donc, éventuellement, de repérer ces corpus spécialisés comme relevant d'un genre particulier. Comme dans toute définition d'un genre, il est nécessaire de passer par une première étape, intuitive, qui permet de considérer que les textes que l'on étudie relèvent du même genre afin de pouvoir, dans une deuxième étape, valider ou invalider l'existence de ce genre, à partir du repérage, ou non, de régularités lexico-syntaxiques[45].
Deux postulats sont omniprésents dans les travaux des terminologues classiques : d'une part il existerait des langues spécialisées[46], et d'autre part, les corpus spécialisés utiliseraient

[45] Le choix de travailler sur tel ou tel phénomène linguistique relevant lui-même d'une intuition.
[46] C'est une hypothèse qui existe non seulement chez les terminologues et les traducteurs mais aussi par exemple chez les enseignants de langues étrangères. On propose ainsi des cours « d'anglais technique », comme s'il y

abondamment les formes nominales[47]. Mais ces mêmes travaux sont très embarrassés pour définir une langue spécialisée :

> « Une langue spécialisée est une langue naturelle considérée en tant que vecteur de connaissances spécialisées.» (Lerat, 1995, 20).

Intriguée par l'intuition des terminologues, j'ai voulu évaluer ce postulat de la présence importante des noms dans un corpus « spécialisé ». Je ne me suis pas donné *a priori* de définition claire de ce qu'on appelle corpus spécialisé. Toutefois, il m'a paru indispensable d'une part de distinguer, dans les corpus spécialisés, les corpus scientifiques et les corpus techniques et, d'autre part, de n'utiliser que des corpus qui étaient, indubitablement « spécialisés ». Les corpus produits dans un contexte professionnel m'ont paru relever de ce « genre » ; je n'ai ainsi étudié que des corpus techniques sans tenir compte de leurs autres caractéristiques. Mon objectif, dans le cas où l'utilisation des nominalisations se révèlerait massive (voire originale) consistait à établir un lien entre nominalisations (examinées du point de vue de leur fréquence et de leur fonctionnement) et langue spécialisée, et donc, d'une certaine façon, de confirmer une intuition des terminologues. Je souhaitais aussi donner une matérialité linguistique à un genre « spécialisé » (en tout cas technique).
Je me suis intéressée aux nominalisations de trois manières : d'abord d'un point de vue strictement comptable et statistique, pour évaluer l'hypothèse d'un nombre de nominalisations particulièrement élevé en corpus technique ; j'ai ensuite étudié les nominalisations dans le cadre de leur fonctionnement syntagmatique ; enfin, dans un travail en collaboration avec Didier Bourigault, nous avons étudié le fonctionnement du sens des nominalisations par rapport aux verbes, toujours en corpus spécialisé.

3.1. Etude statistique des nominalisations dans un corpus spécialisé

Ce travail a nécessité la mise en œuvre de méthodes statistiques, l'une à travers l'utilisation du logiciel Hyperbase (Brunet, 1995), l'autre à travers un calcul de chi2. Il s'agissait de comparer un corpus technique avec, dans le premier cas, un corpus littéraire et dans le second, un corpus journalistique. D'une certaine façon, ces deux corpus sont censés garantir la représentativité de la non-technicité (à défaut de la langue générale) ; en tout cas, l'utilisation de deux corpus de comparaison très différents est une sorte de garantie que la caractérisation concerne bien le corpus technique. Les résultats détaillés de cette étude sont présentés dans (Condamines, 1998).

Je mentionnerai aussi une étude réalisée avec Didier Bourigault, qui permet de comparer trois corpus techniques à trois corpus littéraires et qui confirme les résultats des autres analyses.

3.1.1 Mise en place de l'étude

Le corpus technique étudié est un guide de rédaction de spécifications en génie logiciel (MOUGLIS) d'environ 350 pages et 50000 mots, fourni par EDF.
Les corpus de comparaison sont d'une part Frantext (XIXè et XXè siècles, essentiellement des textes littéraires), via le logiciel Hyperbase, et d'autre part 1000 pages du Monde, prises au hasard dans les numéros de février 1995 (CD-Rom), soit un peu plus d'un million de mots. Dans le premier cas, les résultats ont été fournis par le logiciel Hyperbase, dans le second cas, j'ai mis en œuvre un calcul de chi2 (*cf.* ci-dessous).

---

avait des régularités suffisamment importantes dans ce « genre » de discours pour qu'elles puissent être enseignées.

[47] Comme je l'ai montré dans le chapitre III, c'est, entre autres, pour cette raison que la plupart des outils d'extraction terminologique ne s'intéressent qu'aux formes nominales.

Le corpus technique fait apparaître 1151 noms dont 292 nominalisations (identifiées « à la main » sur la base de leur capacité supposée à exprimer un procés) ; seules 185 de ces nominalisations sont utilisées plus de 3 fois ; ce sont ces nominalisations qui ont été étudiées.

Trois caractéristiques ont été examinées.

Dénombrement d'occurrences, comparaison avec un corpus littéraire

Il s'agissait d'étudier le nombre d'occurrences de chaque nominalisation du corpus technique par rapport au corpus de référence, ici un corpus littéraire. Cette étude a été réalisée avec Hyperbase (outil d'analyse de textes construit par Etienne Brunet, de l'Université de Nice, (Brunet, 1995)) qui fournit une caractérisation statistique sur la base d'une comparaison du corpus à l'étude avec les textes de Frantext des XIXe et XXe siècles[48].
Trois classes de nominalisations ont été ainsi définies à partir des trois caractérisations proposées par Hyperbase :

- lexèmes en excédent,
- lexèmes en déficit,
- lexèmes absents du modèle.

Les deux premières classes sont obtenues à partir de « la valeur absolue de l'écart réduit afin de mettre en relief ce qui est le plus significatif » (Manuel d'hyperbase) ; la troisième concerne les lexèmes absents du fichier dit "REFER" qui contient les 10 000 formes les plus fréquentes de Frantext (XIXe et XXe siècles) c'est-à-dire celles dont la fréquence dépasse 500.
J'ai considéré l'appartenance à l'une de ces trois classes comme significative d'un fonctionnement des nominalisations propre au corpus spécialisé étudié ; mais j'ai également tenu compte des nominalisations qui n'avaient aucune spécificité, c'est-à-dire qui fonctionnaient comme dans le corpus de référence.

Proportion de formes nominales par rapport aux formes verbales

Le corpus du Monde a permis de mettre en œuvre un autre type d'analyse comparée sur deux points. Le premier point concerne la proportion de formes verbales par rapport aux formes nominales (par exemple, formes verbales de *gérer* par rapport à *gestion*). Cette première étude avait pour but de vérifier l'hypothèse d'une utilisation de la nominalisation plus fréquente que la forme verbale dans le corpus spécialisé MOUGLIS par rapport au corpus Le Monde. Dans cet objectif, pour chaque nominalisation, nous avons défini, avec Max Reinert, (statisticien dans l'ERSS), une mesure de chi2 qui met en oeuvre quatre chiffres : nombre de nominalisations dans MOUGLIS, nombre de nominalisations dans Le Monde, nombre de formes verbales dans MOUGLIS, nombre de formes verbales dans Le Monde.

Proportion de formes nominales au singulier par rapport aux formes nominales au pluriel

Comme dans l'étude précédente, le corpus MOUGLIS a été comparé avec le corpus du Monde. Cette étude avait pour but de mettre au jour un nombre éventuellement « anormal » de formes nominales au pluriel ou au singulier. Je voulais vérifier si l'utilisation éventuellement accrue des nominalisations dans le corpus technique pouvait être due au fait que l'une des valeurs sémantiques de la nominalisation était plus fréquemment utilisée. On considère généralement qu'une nominalisation peut avoir (au moins) deux valeurs sémantiques : elle peut renvoyer

[48] D'une certaine façon, cette étude visait à mettre un fonctionnement « déviant », au sens où je l'ai présenté en 1.1. Mais, le corpus de comparaison étant uniquement littéraire, on ne peut pas le considérer comme un corpus de référence, c'est-à-dire un corpus équilibré censé rendre compte du fonctionnement stable du système de la langue.

soit au procès, soit au résultat (état ou objet tangible du procès). Une nominalisation au pluriel renvoie le plus souvent à la deuxième valeur (résultat, objet tangible) ; ainsi, *les achats, les constructions* s'interprètent généralement avec cette deuxième valeur et s'intègrent mal dans des constructions qui contraignent l'interprétation « processive » comme : *pendant les achats, pendant les constructions.*
Ainsi un nombre significativement élevé de nominalisations au pluriel pouvait correspondre à un nombre important de concepts « objets créés » et montrer que l'utilisation des nominalisations ne vient pas remplacer l'utilisation des verbes.

3.1.2 Résultats de cette première étude

Fréquence de nominalisations dans MOUGLIS par rapport au corpus littéraire

Les résultats obtenus grâce à Hyperbase se répartissent dans 4 groupes :

- Premier groupe : 80 nominalisations hors modèle (non comprises dans la liste des formes les plus fréquentes),
- Deuxième groupe : 74 nominalisations en excédent (par rapport au corpus TLF),
- Troisième groupe : 31 nominalisations conformes au modèle,
- Quatrième groupe : 0 nominalisation en déficit (par rapport au corpus du TLF).

Ces premiers résultats mettent en évidence, d'emblée, une utilisation importante de la nominalisation dans le corpus spécialisé par rapport au corpus littéraire. En effet, 154 nominalisations (80 + 74) sur 185 apparaissent comme plus fréquentes que dans le corpus de référence, soit plus de 83 %. En revanche, aucune nominalisation n'apparaît en déficit. Cette première évaluation statistique confirme donc, de façon très nette, l'hypothèse d'une utilisation plus fréquente des nominalisations dans le corpus technique.
Une remarque importante doit cependant être faite. Elle consiste en une mise en garde d'Etienne Brunet lui-même : ces résultats sont à prendre avec des réserves. On peut s'étonner par exemple que ni *listage*, ni *interfaçage* ne soient caractérisés comme plus fréquents dans le corpus spécialisé.

Nominalisations par rapport aux formes verbales, comparaison de MOUGLIS avec le corpus du Monde

L'examen des résultats obtenus (sur la base de l'utilisation du chi2[49], dans le cadre d'une comparaison avec le corpus du Monde) met en évidence une convergence très nette entre les résultats obtenus ici et ceux obtenus avec le critère précédent.
En reprenant les 3 groupes obtenus avec Hyperbase et qui comportent au moins un élément, si l'on considère les résultats correspondant à un chi2 supérieur à 6.63, que nous avons considéré comme significatif, on constate les éléments suivants :

- Nominalisations du premier groupe (« hors modèle ») :
- 27 résultats sur 47 sont supérieurs à 6,63, c'est-à-dire que, pour le corpus MOUGLIS, dans 27 cas sur 47 examinés, le nombre de formes nominales est, de façon significative, supérieur au nombre de formes verbales, par rapport au corpus du Monde. Le chi2 moyen est de 80,93.
- Nominalisations du deuxième groupe (« en excédent ») :
  20 résultats sur 35 sont supérieurs à 6,63. Le chi2 moyen est de 35,3
- Nominalisations du troisième groupe (« conformes au modèle »):
  seuls, 5 résultats sur 18 sont supérieurs à 6,63.

[49] Max Reinert, statisticien dans l'ERSS, m'a aidée à mettre en œuvre l'analyse statistique sous la forme du test du chi2 qui permet de comparer des effectifs d'occurrences. Les résultats obtenus se sont avérés particulièrement significatifs.

On peut donc dire que, dans l'ensemble, les nominalisations qui avaient été identifiées comme n'ayant pas une fréquence spécifique dans le corpus journalistique par rapport au corpus littéraire ne sont pas utilisées de façon plus élevée que les formes verbales correspondantes dans le corpus du Monde. Inversement, pour les deux groupes de nominalisations qui avaient une fréquence élevée par rapport au corpus Frantext, on constate une utilisation de ces nominalisations bien plus grande que les formes verbales correspondantes dans le corpus du Monde. Il y a bien convergence de résultats pour ces deux premières caractérisations, obtenus avec deux méthodes différentes et par comparaison avec deux corpus différents : une forme nominale qui est utilisée plus fréquemment dans le corpus spécialisé que dans un corpus non-spécialisé est aussi utilisée plus fréquemment que la forme verbale, toujours par rapport à un corpus non-spécialisé.
Ces résultats semblent confirmer l'hypothèse d'une utilisation préférée des nominalisations par rapport aux formes verbales.

Cette hypothèse a été à nouveau confirmée par la mise en place d'une autre étude réalisée avec Didider Bourigault. Nous avons effectué un dénombrement des nominalisations par rapport à l'ensemble des noms d'une part et par rapport à l'ensemble des formes verbales d'autre part. Cette fois-ci, ce sont 6 corpus, 3 techniques (SGGD, MOUGLIS et GDP)[50] et 3 littéraires (Sartre, Balzac, Chateaubriand)[51] qui ont été comparés en utilisant l'outil Lexter (qui met en œuvre un étiqueteur grammatical) et la ressource Verbaction[52] (Condamines et Bourigault, 1999).
Les résultats parlent d'eux-mêmes :

	Noms Act		Autres noms		Verbes		Nom Act/Verbe
		%		%		%	
SGGD	12 243	25	25 455	54	10 271	22	1,14
MOUGLIS	5 212	20	16 082	63	4 294	17	1,18
GDP	14 293	24	32 685	55	12 662	21	1,14
SARTRE	1 648	8	9 548	48	8 680	44	0,18
BALZAC	4 751	7	35 070	53	26 359	40	0,18
CHATEAUB.	941	6	7 808	50	6 713	44	0,14

Tableau 3 : Répartition des noms d'action, des noms d'une autre nature et des verbes dans 3 corpus techniques et 3 corpus littéraires

[50] Les corpus Mouglis et GDP (Guide de Planification, fourni par EDF, environ 220 000 occurrences) peuvent être considérés comme des manuels). SGGD (environ 147 000 occurrences) est constitué de présentations de systèmes informatiques.
[51] Corpus de 200 000 occurrences pour Balzac, 50 000 pour Chateaubriand et 62 000 pour Sartre.
[52] Pour repérer les noms morphologiquement reliés à des verbes, nous avons utilisé le lexique *Verbaction*, réalisé par Nabil Hathout de l'INALF à partir des verbes de la nomenclature du Trésor de la Langue Française. Nabil Hathout a défini un ensemble de règles de dérivation qui, appliquées aux verbes de la nomenclature du TLF, ont généré une liste de formes. Celles-ci ont été comparées aux formes nominales de cette même nomenclature. Les couples ont été vérifiés manuellement ; seuls les couples sémantiquement apparentés ont été conservés. *Verbaction* constitue ainsi une ressource extérieure que l'on peut considérer comme représentative du fonctionnement général de la langue. Il est constitué d'un ensemble d'environ 7 000 couples, dont nous donnons un extrait : *(démarrage; démarrer), (démâtage; démâter), (démêlage; démêler), (démêlement; démêler), (démembrement; démembrer), (déménagement; déménager), (démenti; démentir).*

Le pourcentage des noms d'action, par rapport à l'ensemble des noms et des verbes, est de moins de 10 % dans les textes littéraires et de plus de 20 % dans les corpus techniques. Inversement, celui des verbes est de plus de 40 % pour les textes littéraires et seulement de moins de 20 % pour les corpus techniques. Ces deux types de résultats font que le rapport des noms d'action aux verbes passe de moins de 0,2 (textes littéraire) à plus de 1 (textes techniques).

Nominalisations au pluriel par rapport aux nominalisations au singulier, comparaison de MOUGLIS avec Le Monde

Dans les trois groupes de nominalisations repérés (« hors modèle », « en excédent », « conforme au modèle »), peu de résultats sont supérieurs à 6,63 ou inférieurs à (-)6,63, c'est-à-dire que peu de nominalisations ont une utilisation du pluriel beaucoup plus ou beaucoup moins élevée que dans le corpus de référence:

- 16/60 dans le premier groupe (« hors modèle »
- 13/57 dans le deuxième groupe (« en excédent »)
- 8/25 dans le troisième groupe (« conforme au modèle »).

Soit, au total 37 résultats "anormaux" sur 142 (un peu plus du quart).

Cela signifie que peu de nominalisations ont un comportement différent, dans le corpus MOUGLIS, du point de vue du pluriel, par rapport au corpus du Monde. L'examen des moyennes de chi2 confirme ces résultats. En effet, dans aucun des trois groupes la moyenne des chi2 est inférieure à (-) 6,63 ou supérieure à 6,63, c'est-à-dire une moyenne qui mettrait en évidence un fonctionnement spécifique :

- groupe 1 : moyenne des chi2 : (-)3,43
- groupe 2 : 0,33
- groupe 3 : 0,67

Notons cependant que, lorsqu'il y a fonctionnement différent, il va plutôt dans le sens d'une augmentation des formes plurielles (résultat inférieur à (-)6,63) (21 cas sur 37).

3.1.3 Conclusion

Au terme des trois types d'études réalisées sur un ou plusieurs corpus spécialisés, en comparaison avec des corpus littéraires ou journalistiques, utilisant des méthodes statistiques différentes et qui ont concerné :

- le nombre de nominalisations en corpus spécialisé ,
- le nombre de formes verbales par rapport aux nominalisations,
- le nombre de nominalisations au pluriel par rapport aux nominalisations au singulier,

les conclusions suivantes peuvent être tirées. Le corpus spécialisé qui a été étudié (et sans doute peut-on élargir cette constatation à l'ensemble des corpus techniques mais cela reste à vérifier) privilégie les nominalisations par rapport aux formes verbales mais ces nominalisations ne sont pas plus souvent au pluriel que dans les corpus non-spécialisés, ce qui tend à montrer que les nominalisations ne sont pas utilisées pour évoquer des objets mais bien en lieu et place des formes verbales.

Ces études confirment donc l'intuition d'une utilisation importante de la forme nominale dans les corpus techniques . Si ce même élément se vérifie dans des corpus scientifiques et que l'on accepte que caractère technique et caractère scientifique constituent les deux éléments possibles pour qu'un texte soit considéré comme relevant d'une « langue spécialisée » (c'est-à-dire en fait, un genre), alors on tient peut-être là une caractérisation linguistique du genre spécialisé. Il me semble en effet que c'est une caractéristique linguistique suffisamment importante pour considérer qu'elle permet d'identifier un genre textuel.

Mais il resterait à valider cette hypothèse avec d'autres corpus, techniques, scientifiques et non-spécialisés.

3.2. Etude du fonctionnement des nominalisations en syntagme, dans un corpus technique

L'objectif de cette étude a concerné le fonctionnement des nominalisations en syntagme, dans un corpus spécialisé. L'étude a été menée sur le même corpus technique que dans l'étude précédente (manuel de génie logiciel, MOUGLIS, fourni par EDF). Le fonctionnement syntagmatique des nominalisations dans ce corpus a été comparé avec ce même fonctionnement dans le corpus du Monde (également le même que dans l'étude précédente, soit 1000 pages de février 1995 prises au hasard). L'interprétation des contextes ne pouvant être réalisée « qu'à la main », je n'ai pas pris en compte un autre corpus de comparaison ce qui affaiblit un peu la démonstration.

Si les syntagmes contenant une nominalisation déverbale ont fait l'objet d'un grand nombre de travaux (voir par exemple (Samvellian, 1995), (Fabre, 1996), (Bartning, 1996)), ces travaux se sont rarement appuyés sur l'analyse de corpus réels. Or, l'analyse de corpus est riche d'enseignements, particulièrement sur cette question des nominalisations et de leur fonctionnement syntagmatique.

3.2.1 Les structures syntagmatiques étudiées

Trois structures ont été étudiées :

- Nominalisation + de + SN,
- Nominalisation + N,
- Nominalisation + Adjectif relationnel.

Il est bien connu que dans certains cas, ces structures peuvent être alternatives.

> « Il est certain que nos trois constructions sont directement concurrentes dans de nombreux cas [...] la construction prépositionnelle, de toute évidence, est à la fois la plus ancienne et la plus correcte, les groupes à adjectifs de relation, qui l'ont d'abord concurrencée, sont plus "modernes", plus "technocratiques", mais un peu lourd et d'un registre moins élégant [...]. La construction directe, enfin, appartient, au moins dans ses emplois les plus hardis, à un langage plus jeune et plus réservé... » (Noailly, 1990, 177).

Ces structures ont été étudiées les unes par rapport aux autres dans le corpus MOUGLIS et dans leur fonctionnement en contexte (syntaxique et sémantique) dans une perspective comparative MOUGLIS/Le Monde. Sur les 185 nominalisations de la première étude, seules 97 ont été examinées : celles qui apparaissent au moins 3 fois, dans au moins une des trois structures considérées.

Le repérage des structures à étudier a été effectué à l'aide d'un outil (SATO en l'occurrence) ce qui a nécessité un tri « manuel » qui visait à ne conserver que les structures dans lesquelles le N avait un rôle argumental ; par exemple, des structures comme *traitement de faveur* ou *solution de compromis* n'ont pas été conservées. Par ailleurs, la sélection des adjectifs relationnels n'a pas été facile ; on sait en effet que la distinction de ces adjectifs par rapport aux qualificatifs n'est pas toujours aisée :

> « Nombre de ces adjectifs, parallèlement à une interprétation relationnelle, donnent lieu à une analyse qualificative, avec, dans certains cas une répartition binaire franche des différents emplois et, dans d'autres, un continuum d'effets de sens, qui rendent la description très délicate.». (Bartning et Noailly, 1993, 27).

Enfin, dans le cas d'occurrences de mots tronqués, il est presque impossible de savoir si on doit les considérer comme des adjectifs ou des noms *: informations météo = de la météorologie ou météorologiques* ?

3.2.2 Résultats

Les résultats chiffrés sont résumés dans le tableau suivant :

	MOUGLIS	LE MONDE
1- Occurrences	50004	1720086
2- Nominalisations	4399 (8,09 % de 1-)	9670 (0,56 % de 1-)
3- + de N	1572 (35,73 % de 2-)	2869 (29,66% de 2-)
4- + Adj relationnel	215 (4,88 % de 2-)	1035 (10,70 % de 2-)
5- + N	330 (7,5 % de 2-)	19 (0,19 % de 2-)

Tableau 4 : Résultats chiffrés des structures contenant une nominalisation, dans un corpus technique et dans un corpus journalistique

Quantitativement et qualitativement, ces résultats portent essentiellement sur trois points – pour une analyse détaillée, *cf.* (Condamines, 1999) – :

- Les adjectifs relationnels sont peu utilisés dans le corpus technique comparativement au corpus du Monde.
  Il faut envisager que cette apparente sous-utilisation soit due à une utilisation très élevée de cette forme dans le corpus journalistique, c'est-à-dire, dans le corpus de comparaison. Enfin, il serait nécessaire de vérifier ce fonctionnement sur plusieurs autres corpus spécialisés avant de le considérer comme caractéristique des corpus techniques. Il n'en reste pas moins que cette sous-utilisation de la structure avec adjectif relationnel semble aller de pair avec une sur-utilisation de la structure NN (*cf.* ci-dessous).
- La structure « Nominalisation + N » est largement utilisée dans le corpus spécialisé (330 fois dans MOUGLIS, 18 fois dans Le Monde, qui comprend plus de 30 fois plus de mots, soit, 6,6/1000 occurrences dans MOUGLIS et 0 ,018/1000 occurrences dans Le Monde). Il faudra vérifier si elle est caractéristique de ce corpus ou bien si c'est une structure que l'on retrouve fréquemment dans des textes techniques (ce qui semble être une attente raisonnable).
- Dans la quasi totalité des cas où cette structure « Nominalisation + N » est utilisée dans le corpus spécialisé, le second N a un rôle d'objet, la structure correspondant à un effacement de la préposition "de" (il existe de très nombreux exemples où les deux structures coexistent dans le corpus spécialisé : *conception (de la) méthode, configuration (du)projet, intégration (du) produit*…). Ce fonctionnement, avec un N2 objet, ne semble pas correspondre à ce qui est généralement attendu. Ainsi dans la description des structures N1N2 avec N1 nominalisation, Noailly a pu écrire : « N2 = objet est plus rare mais pas interdit. » (Noailly, 1990, 120).
- L'absence quasi totale d'argument agent en position de N2 (je n'en ai trouvé qu'une occurrence : *acceptation client*) constitue une caractéristique du corpus MOUGLIS par rapport au Monde (où l'argument agent est moins fréquent que l'argument objet mais pas rare). Cependant, cet élément est peut-être plutôt à

mettre en relation avec la visée du corpus qu'avec sa nature technique. En effet, il s'agit de « recommandations » pour la spécification en génie logiciel et ce genre textuel nécessite sans doute une non-personnalisation des propos : on ne dit pas qui doit réaliser telle ou telle tâche mais qu'elle doit être réalisée. De là vient peut-être aussi cette sur-utilisation des nominalisations dans les corpus techniques qui permet de ne pas donner de sujets aux actions.

– Il semblerait que cette structure « Nominalisation + N » soit presque toujours utilisée dans le corpus à la place de la structure « Nominalisation + de N » et, dans quelques rares cas, à la place de la structure « Nominalisation + adjectif relationnel ». En effet, bien qu'en principe, les trois structures soient considérées comme substituables, cette substitution n'est pas toujours possible, en particulier semble-t-il parce que dans le corpus technique, le second N peut difficilement être dérivé en adjectif. Ainsi, si dans Le Monde, on trouve des paires comme *modification de la loi/modification législative* ou *construction de l'Europe/construction européenne*, cette alternance est presque absente dans le corpus MOUGLIS. Elle s'instaure bien plus nettement dans la structure avec préposition, comme je l'ai signalé ci-dessus. Ainsi, si l'on trouve beaucoup de constructions Nominalisation Nom avec nom = qualité ((*contrôle, évaluation, vérification*) *qualité*)…, parfois en alternance avec la construction Nominalisation de Nom (*contrôle de la qualité, vérification de la qualité*), on ne trouve aucune occurrence de la structure avec adjectif relationnel : *contrôle qualitatif, vérification qualitative*.

Ainsi, non seulement les nominalisations semblent être fréquentes dans les corpus techniques mais elles semblent aussi fonctionner de manière différente dans leur contexte syntagmatique. Il faut toutefois se garder de généraliser ce second type d'observations ; en effet, seul un corpus a été examiné et comparé avec un corpus relevant d'un seul genre.

3.3. Etude du fonctionnement sémantique des nominalisations en corpus spécialisé

Toujours dans la perspective de travailler sur la nominalisation dans des corpus techniques, nous avons, pour cette troisième étude, examiné d'un point de vue sémantique les cas d'alternance forme verbale/forme nominale dans les trois corpus techniques précédemment évoqués (SGGD, GDP et MOUGLIS)[53].

Ce type d'analyse suppose que soient interprétés tous les contextes d'apparition des verbes et de leur nominalisation, ce qui n'a pas été fait sur l'ensemble des cas d'alternance car cela aurait été beaucoup trop coûteux en temps. Comme dans tous les cas d'interprétation en contexte, il s'agit de classer les occurrences pour identifier des similitudes. Pour l'essentiel, ce classement s'est fait sur la base d'identification de similitudes sémantiques des arguments. Cette approche est généralement utilisée pour justifier la polysémie construite sur des bases introspectives :

> « Il est en effet admis depuis longtemps que la structure d'arguments est liée à la polysémie, c'est-à-dire que le sens des verbes varie avec la structure ou la structure avec le sens des verbes (Martin, 1979). » (Guy et Léard, 1996, 181).

Mais cette même approche est aussi souvent remise en question :

> « En somme, il existe des cas où les changements d'arguments (catégorie ou classe sémantique) sont associés à des changements sémantiques et d'autres où ils ne le sont pas.» (*ibid*., 184).

[53] Ce travail a été mené en collaboration avec Didier Bourigault, de l'ERSS.

Notre objectif n'était pas d'identifier le sens de tel ou tel verbe ou de telle ou telle nominalisation mais plutôt de mettre au jour, à partir de corpus réels, une sorte de panorama des cas d'alternance possibles qui montrent des fonctionnements souvent beaucoup plus difficiles à décrire que les fonctionnements qui sont caractérisés sur des bases introspectives. Nous avons utilisé pour notre démonstration les exemples les plus manifestes ; il est clair que très souvent l'interprétation des contextes pour identifier des sens clairement définis et représentables est très difficile, voire impossible. Et pourtant cette situation ne nuit pas à la compréhension. C'est au même type de constat que parvient Kayser :

> « la compréhension peut parfaitement se passer d'une dénotation » (Kayser, 1995, 33).

La présentation des résultats est faite en deux parties, l'une qui décrit quelques cas typiques, l'autre qui montre que les classements sémantiques habituellement donnés des nominalisations sont insuffisants.

3.3.1 Etude de quelques cas

3.3.1.1 Le verbe a deux significations/le nom n'est pas attesté

On trouve d'assez nombreux exemples de ce type :

– Le couple *assurance-assurer* dans le corpus GDP. Le nom n'est pas attesté ; le verbe est utilisé avec au moins deux significations. La première peut être généralisée dans la structure : `équipement assurer fonction`, la seconde dans la structure : `N assurer condition optimale,` le verbe *assurer* est équivalent à garantir.

Le réseau HTA assure la liaison entre les jeux de barres HTA.

…assurer la stabilité des groupes de production,…assurer la sécurité des matériels et la meilleure exploitation possible du système production-transport.

– Le couple *résolution-résoudre* dans le corpus GDP. Seul, le verbe est utilisé avec au moins deux significations qui peuvent être généralisées dans les deux structures suivantes : `Action résoudre contrainte, N résoudre problème.`

Cette seconde structure est moins spécifique au corpus ; les objets qui apparaissent sont du type *problème, difficulté…*

Le remplacement des réducteurs de mesure résout la contrainte de limitation due aux réducteurs de mesures des cellules d'extrémité.

Il existe en particulier des appareils monophasés en hydraulique pour résoudre les problèmes pratiques d'accès et pour l'évacuation des groupes nucléaires afin d'assurer la transportabilité des transformateurs.

3.3.1.2 Le verbe n'est pas attesté, le nom a deux significations

Le couple *encombrement-encombrer*. La particularité ici est que la différence sémantique se manifeste par une forme particulière ; en effet, lorsqu'il est au singulier, le nom concerne l'encombrement au sol d'un objet, et au pluriel la gêne produite par une circulation trop importante :

L'encombrement d'un poste de travail ne devrait pas dépasser 1,5 m de largeur et 1,5 m de profondeur

Les relevés des encombrements d'accès et de sortie de la rocade ont été réalisés de manière exhaustive.

3.3.1.3 Le nom a deux significations, le verbe n'en a qu'une des deux

C'est le cas par exemple du couple *échange-échanger* (corpus SGGD). En effet, l'analyse du nom révèle deux significations distinctes : l'une avec laquelle il est complément d'un autre déverbal et ce, directement (sans préposition *de*). Ce type d'occurrence correspond à la

dénomination d'un artefact. Dans le second cas, *échange* est utilisé avec des compléments prépositionnels (de N) qui constituent une classe sémantiquement homogène. La forme verbale est utilisée avec ce même type de compléments.
Elargissement à deux fois deux voies des échanges entre rocade est et rocade ouest
Echanges de données, d'informations ; échanger des données, des informations.

3.3.2 Interprétation aspectuelle des noms déverbaux : insuffisance de la distinction dynamique vs *non-dynamique*

Il est classique de présenter la sémantique des noms déverbaux comme correspondant à une répartition entre interprétation dynamique et interprétation non-dynamique (*cf. échange* ci-dessus). Souvent, la valeur non-dynamique du nom déverbal correspond à une utilisation pour la dénomination d'un artefact, par exemple, *enroulement* :
Certains de ces appareils ont des enroulements tertiaires raccordés en triangle.
Pour un grand nombre de nominalisations cependant, la distinction dynamique/non-dynamique n'est pas suffisante. C'est le cas du nom *jalonnement* dans le corpus SGGD, qui est utilisé avec trois significations différentes.
Elle s'adresse en priorité aux usagers connaissant le réseau urbain proche, puisqu'aucun jalonnement n'est mis en place sur celui-ci.
L'intégration de la signalisation dynamique dans le schéma général de jalonnement des VRU (Voies Rapides Urbaines) est particulièrement recherchée.
Les soumissionnaires devront proposer un emploi du temps détaillé…, ainsi que l'enchaînement et le jalonnement des tâches.
Dans le premier exemple, *jalonnement* renvoie à un sens non-temporalisé, un artefact. Dans le deuxième exemple, *jalonnement* renvoie à un sens temporalisé, qui concerne des éléments concrets : les voies en général et qui est équivalent à balisage. Le troisième exemple fait apparaître un sens lui aussi temporalisé, qui concerne des éléments abstraits : les *tâches*. Il n'est d'ailleurs pas impossible que l'usage propre au domaine dans le deuxième type d'exemple soit venu contaminer le troisième type, qui correspond à un usage plus général de *jalonnement* et dans lequel le mot attendu aurait plutôt été *échelonnement.*
Je n'ai pas, dans le commentaire de ces exemples, utilisé la distinction dynamique/non dynamique. C'est que cette distinction n'est pas pertinente pour ces exemples. On se rend compte en effet, que dans tous les exemples, *jalonnement* a un sens non-dynamique, soit parce qu'il concerne un artefact, soit parce qu'il concerne un état résultatif plutôt qu'un processus. S'il existe *un schéma général de jalonnement des VRU*, c'est que les VRU ont été jalonnées, au moins sur le papier ; il en va de même pour *jalonnement des tâches*. Ainsi, la différence entre le premier exemple et les deux autres vient de ce qu'il y a encore du temps dans ces deux derniers, plus exactement de l'aspect.
Ainsi, sur la base de ces seuls exemples, on constate qu'il faudrait déjà affiner la classification, qui pourrait être :
dynamique,
non dynamique non temporel,
non dynamique temporel.

L'analyse de l'alternance nominalisations/formes verbales en contexte fait ainsi apparaître d'une part qu'il est souvent impossible de définir (c'est-à-dire discrétiser) le sens des verbes et des déverbaux, et cela, même lorsqu'on se place dans la perspective de comparer leur fonctionnement et, d'autre part, que des nuances aspectuelles apparaissent, comparables à celles qui existent dans les verbes conjugués. Ces nuances seraient bien plus difficiles à mettre au jour sur la base d'exemples forgés.

3.4. Synthèse

Les différentes études sur les nominalisations qui ont été proposées ici relèvent de ce que j'appellerai une linguistique de corpus, c'est-à-dire une linguistique qui s'appuie sur l'analyse de données réelles en lien avec une hypothèse théorique et/ou applicative. Comme on l'a vu, ces résultats semblent confirmer des hypothèses :

- Les corpus techniques utilisent beaucoup de nominalisations. Cette caractéristique semble suffisamment stable pour définir un genre technique voire spécialisé. Cela signifie qu'un corpus qui utiliserait beaucoup de nominalisations devrait pouvoir être qualifié de « spécialisé ». Il faudrait toutefois vérifier cette corrélation sur un grand nombre de textes.
- Ces nominalisations semblent fonctionner différemment en corpus technique qu'en corpus journalistique. Ainsi, elles sont particulièrement utilisées pour modaliser la forme injonctive : comme elles peuvent fonctionner sans sujet (contrairement aux verbes), elles sont facilement mises en œuvre pour donner un ordre indirect, sans qu'il soit besoin de préciser l'agent : *la rédaction des informations relatives à chaque procédure peut suivre la découpe ci-après.*
- Le mode de fonctionnement des nominalisations par rapport aux formes verbales est bien plus complexe que ce qui est généralement décrit sur des bases introspectives.

Les premières bases d'une étude qui reste encore largement à faire ont été jetées.

4. Conclusion

Les phénomènes que j'ai évoqués dans ce chapitre : repérage de termes, polysémie, nominalisations, dépassent largement le problème de la constitution de bases de connaissances terminologiques. Je les ai abordés plutôt en m'interrogeant sur ce que l'utilisation de corpus d'une part et l'objectif d'étude d'autre part permettaient d'éclairer sur ces phénomènes. Ce faisant, je me suis située dans ce qui est, très clairement, une sémantique de corpus.

Ainsi donc, la constitution de BCT à partir de textes constitue une tâche particulière qui demande la mise en œuvre d'une sémantique de corpus. Ainsi que je l'ai dit plusieurs fois, cette tâche induit un type d'interprétation particulier qui vient de ce qu'on prend en compte d'emblée le besoin des utilisateurs que le linguiste essaie de comprendre et d'interpréter en des termes sémantiques. Lorsqu'on travaille sur des phénomènes linguistiques plus généraux, sur un corpus en tant qu'il fait partie d'un genre particulier, la situation pourrait paraître très différente. En réalité, dans ce cadre d'étude, l'analyste de corpus prend aussi en compte un objectif, une intention, collective celle-là, qui est celle de la communauté scientifique dans laquelle il s'inscrit. Il ne me semble pas y avoir de différence fondamentale entre une analyse de corpus pour constituer une BCT et une autre pour étudier tel ou phénomène : il s'agit très clairement dans les deux cas de se confronter avec des phénomènes sémantiques tels qu'ils apparaissent en corpus tout en s'inscrivant dans une visée qui oriente le type de questions que l'on se pose et le type de résultat que l'on obtient. Comme je le soulignais dans le premier chapitre, il n'y a pas, de mon point de vue, d'un côté une linguistique théorique et de l'autre une linguistique appliquée, qui se contenterait d'utiliser les résultats de la première (actuellement souvent bien mal adaptés à la réalité des fonctionnements en corpus).
La difficulté consiste surtout à voir comment les deux types de visées, les deux types d'expériences se nourrissent les unes les autres ; cette question se pose généralement à travers celle de la généralisation des résultats obtenus à partir d'un corpus avec une analyse

répondant à un besoin particulier vers des corpus avec une analyse sans application particulière. Sans application peut-être mais pas sans intention, comme je viens de le dire. Ainsi l'analyse de corpus, comme probablement toute analyse en sciences humaines (voire toute analyse), n'est jamais neutre ou objective ; elle s'inscrit toujours dans une intention particulière et il est nécessaire de prendre en compte cet élément.
Finalement, l'analyse de corpus spécialisés pour construire une BCT oblige à considérer deux éléments qui interviennent dans l'analyse de corpus, généralement occultés et qui s'imposent d'emblée dans ce type de contexte :

- la façon dont la compétence linguistique se met en œuvre et la nature de cette compétence,
- l'objectif, c'est-à-dire la demande qui sous-tend l'analyse, toujours plus ou moins extérieure au linguiste en tant qu'individu mais dont on peut faire l'hypothèse qu'elle participe d'un point de vue collectif que j'ai appelé « genre interprétatif ». Ces deux éléments sont incontournables dans le cas d'un corpus spécialisé et d'une demande extra-universitaire ; mais cette situation, qui semble particulière, n'est que la version grossie d'une situation qui existe toujours dans le cas de l'analyse de corpus.

Le dernier chapitre me permettra de montrer plus précisément pourquoi, en linguistique de corpus, analyse théorique et analyse appliquée ne constituent que différentes versions d'une même problématique, éminemment sémantique.

Chapitre V
D'un corpus à l'élaboration d'un réseau relationnel : la question des marqueurs

Dans la constitution de Bases de Connaissances Terminologiques tout comme dans celle d'ontologies, l'accent est mis sur la représentation sous forme de réseaux, dans lesquels les relations jouent un rôle au moins aussi important que les éléments qui sont mis en relation. Cette notion de relation, et donc de système, n'est pourtant pas apparue avec les besoins des informaticiens ou des terminologues ; c'est bien en effet sur cette idée que s'est élaborée la linguistique structurale. Il convient donc de situer cette problématique du point de vue de la linguistique pour mieux montrer que la constitution d'un réseau relationnel, bien que parfois possible, ne va pas de soi. Elle résulte d'une interprétation qui ne peut avoir de pertinence que si le cadre de son élaboration est clairement défini. D'emblée, on peut dire que la validité d'une représentation relationnelle élaborée à partir d'un corpus doit s'appuyer sur trois éléments :

- le corpus est représentatif d'un locuteur collectif, c'est-à-dire que les textes qui le composent relèvent tous d'une situation de communication que l'on peut considérer comme similaire et dans laquelle les caractéristiques locutives individuelles peuvent être neutralisées,
- l'objectif de la représentation (que l'on peut considérer comme relevant d'un genre interprétatif) est pris en compte très tôt dans le processus d'interprétation ; il influence la constitution du corpus, au moins au même titre que le genre textuel (homogénéité des textes pris en considération),
- la représentation est validée de manière consensuelle par les utilisateurs.

Ce mode de représentation n'est ainsi pas universel et pas pertinent pour n'importe quelle analyse sémantique. Il a peu d'intérêt par exemple pour une analyse littéraire et tout aussi peu pour l'analyse de textes en vrac (même si un certain nombre d'informaticiens pense pouvoir « extraire » de ces textes des connaissances à la volée, connaissances qui, en réalité, sont si hétérogènes qu'elles ne présentent pas grand avantage).

La possibilité de repérer des relations en corpus préoccupe à la fois les terminologues et les informaticiens et je montrerai que des hypothèses fortes imprègnent souvent ce thème de recherche, hypothèses qui ne tiennent pas compte de la réalité des fonctionnements en corpus.
La notion de marqueur de relation demande ainsi à être interrogée à la lumière de l'analyse de corpus. Les rapports de dépendance du marqueur, non seulement avec la relation qu'il est censé exprimer, mais aussi avec le corpus dans lequel il est utilisé sont très rarement interrogés. Pourtant, des analyses qui étudient les fonctionnements réellement attestés en corpus pour évaluer les possibilités d'élaboration sous forme relationnelle remettent en question les descriptions généralement admises des marqueurs, et permettent une ouverture vers la mise en œuvre d'une linguistique de corpus d'un nouveau type.
Ce chapitre comportera deux parties, l'une qui présente les questions qui se posent autour de la problématique des relations et des marqueurs, l'autre qui montre la réalité du fonctionnement des marqueurs étudiés dans des corpus et pour un objectif réels.

1. Relations conceptuelles, marqueurs : position du problème

Avant de faire intervenir le corpus dans l'étude du fonctionnement des marqueurs de relations, il est nécessaire d'examiner l'origine des notions de marqueurs et de relations et les postulats qu'elles véhiculent.

1.1. Notions de relation et de réseau relationnel

Contrairement à ce qu'on pourrait penser, en particulier dans certains courants de la terminologie ou de l'informatique, la représentation sous forme de relations n'est pas toujours une alternative idéale à la complexité de la linéarité des discours. En effet, elle peut introduire de l'implicite qui peut nuire à la compréhension. Avant de voir ce qu'il est possible d'envisager, il convient de cerner ce que suppose la notion de réseau relationnel construit à partir d'un corpus.

1.1.1 Relations et structuralisme

La notion de relation est omni-présente dans l'approche structuraliste, à tous les niveaux : phonologique, morphologique, syntaxique, lexical… Ainsi, toutes les représentations relationnelles qui visent à représenter la connaissance entretiennent une parenté évidente avec l'hypothèse saussurienne qui considérait la langue comme une structure, un système. Comme le montre Benveniste, cette hypothèse a constitué, au début du XXè siècle une évolution majeure par rapport aux approches antérieures :

> « On abandonne donc l'idée que les données de la langue valent par elles-mêmes et sont des « faits » objectifs, des grandeurs absolues, susceptibles d'être considérées isolément. » (Benveniste, 1966, 21).

La notion de structure est ainsi inévitablement corrélée à celle de relation. Toutefois, dans l'approche saussurienne, la notion de relation est associée à deux types de fonctionnements. L'un concerne plutôt la dimension paradigmatique, et constitue la base du système, l'autre concerne plutôt une dimension syntagmatique qui prend en compte la dimension discursive (même si ce mot n'est pas utilisé par Saussure). Ce qu'il est intéressant de noter, c'est que dans le premier cas, l'accent est mis sur l'éloignement, la distance que la relation instaure avec les autres éléments : les relations servent à maintenir les oppositions et les différences :

> « Chacune des unités d'un système se définit par l'ensemble des relations qu'elle soutient avec les autres unités, et par les oppositions où elle entre. » (Benveniste, 1966, 21).

« […] dans la langue, il n'y a que des différences. Bien plus, une différence suppose en général des termes positifs entre lesquels elle s'établit ; mais dans la langue il n'y a que des différences sans termes positifs. » (Saussure, 1982, 166).

Dans le paradigme relationnel, les éléments sont en opposition, éloignés par la relation. La mise en syntagme suppose un choix dans différents paradigmes ; le rapprochement syntagmatique, enfin considéré comme positif, est ainsi le résultat d'une série de choix qui rompent (déstructurent) les paradigmes. Ce n'est en effet que dans les contextes syntagmatiques, lorsque les unités apparaissent *in praesentia* que Saussure parle de rapports associatifs (rapports étant d'ailleurs le terme qu'il utilise en lieu et place de celui de relation) :

« Jusqu'ici, les unités nous sont apparues comme des valeurs, c'est-à-dire comme les éléments d'un système, et nous les avons surtout considérées dans leurs oppositions ; maintenant, nous reconnaissons les solidarités qui les relient ; elles sont d'ordre associatif et d'ordre syntagmatique, et ce sont elles qui limitent l'arbitraire du signe. » (Saussure, 1982, 182).

Il y a ainsi une double rupture : celle qui permet le passage du paradigmatique au syntagmatique, qui est nettement soulignée par Saussure mais aussi, me semble-t-il, celle qui permet le passage du syntagmatique au paradigmatique, du discours au système, au cours duquel les unités sont éloignées les unes des autres, les relations servant à maintenir les oppositions. Il semblerait donc que, dans le système paradigmatique, les relations servent à séparer les unités (en en faisant des éléments discrets) mais d'une manière qui rende l'ensemble cohérent et qui permette de constituer un système linguistique unique. Ces mêmes unités sont au contraire rapprochées dans le discours mais sous des formes qui peuvent être multiples voire incontrôlables.
Ces deux dimensions de la notion de relation sont malheureusement souvent oubliées par les utilisateurs de la représentation sous forme de réseau. Le réseau-système qui, en réalité, ne peut s'élaborer que par l'introduction d'une rupture dans le linéaire discursif est bien trop souvent considéré comme un préexistant que le discours se contente de mettre en forme à l'aide de relations syntagmatiques. C'est très clairement dans cette perspective que se situent les héritiers de Wuster en terminologie :

« Un domaine (ou une sous-section de domaine) n'est accessible mentalement que si le champ notionnel est structuré, c'est-à-dire s'il constitue ce qu'on appelle un système de notions. Dans cet ensemble, chaque notion révèle ses rapports avec les autres notions. » (Felber, 1987, cité par (Van Campenhoudt, 1994, 54)).

D'où l'idée assez répandue que la construction d'un réseau relationnel consiste à remplacer des éléments de forme linguistique, considérés comme parasites dans la diversité de leur expression syntagmatique, par des relations-types permettant de « retrouver » le système originel.
Il faut reconnaître aussi que la possibilité d'un système unique se décline sans doute différemment selon que l'on se situe au niveau phonologique, où elle semble opératoire et où elle a donné des résultats, aux niveaux morphologique ou syntaxique, où elle est déjà certainement beaucoup moins sûre, ou au niveau sémantique où elle constitue le cœur du questionnement sur la variation.

1.1.2 Réseaux relationnels

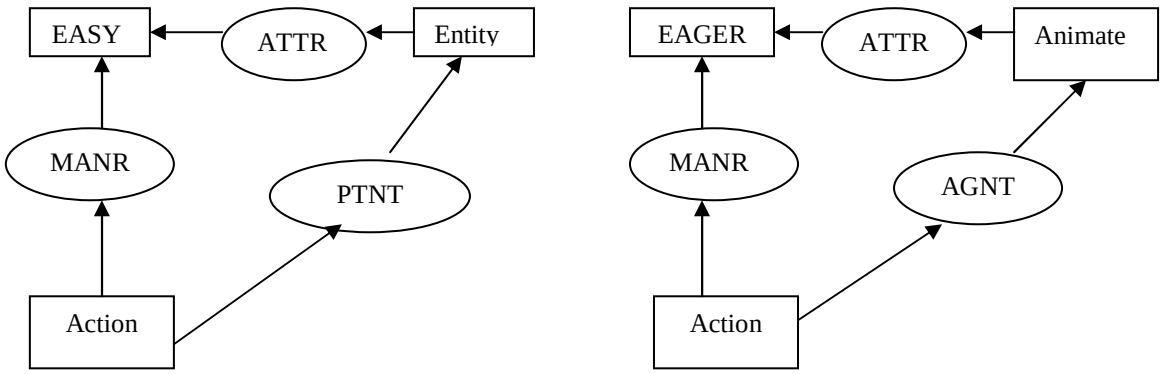
1.1.2.1 Représentations sous forme relationnelle

La représentation de « connaissances » sous forme de structures relationnelles, au sens large, est très fréquente. On la retrouve en terminologie sous le nom de « réseaux relationnels », en Intelligence Artificielle, sous celui « d'ontologies », en pédagogie sous celui de « cartes conceptuelles » (voir le numéro 5 de la revue Didaskalia). En sciences de l'information

même, les « thésaurus » relèvent d'une vision du même type, même si la sémantique des relations n'est pas explicitée. Enfin, dans la vision de la sémantique cognitive, la spatialisation des concepts est censée métaphoriser le *paysage mental :*

> « Within Linguistics and Artificial Intelligence, there has developped in the last ten or fifteen years a related notion : conceptual or cognitive space. An alternative, though still related, metaphor which is much used is that of the mental landscape ». (Werth, 1999, 6).

Ces représentations peuvent être utilisées pour différents types de connaissances : des concepts (figure 6), des phrases (figure 7), des connaissances « ontologiques » (figure 8).



AGNT : Agent, ATTR : Attribute, MANR : Manner, PTNT : Patient

Figure 6 : Graphe canonique pour EASY et EAGER (d'après Sowa, 1991, 47)

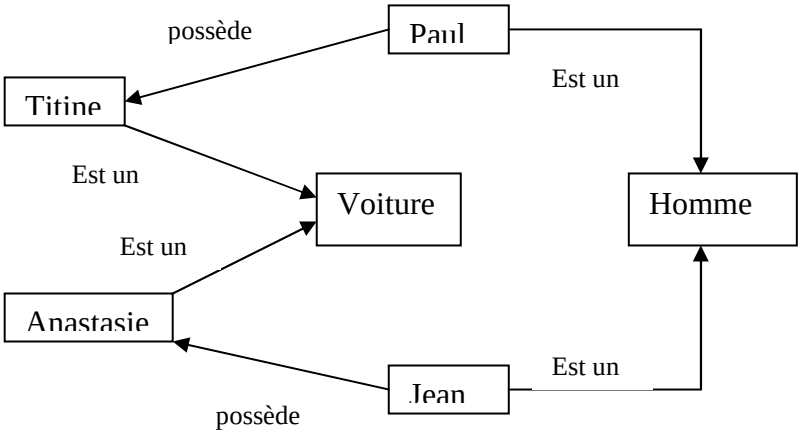


Figure 7 : Représentation sous forme de réseau de la phrase « Paul possède une voiture appelée Titine et Jean une autre appelée Anastasie » (d'après Sabah, 1988, 207)

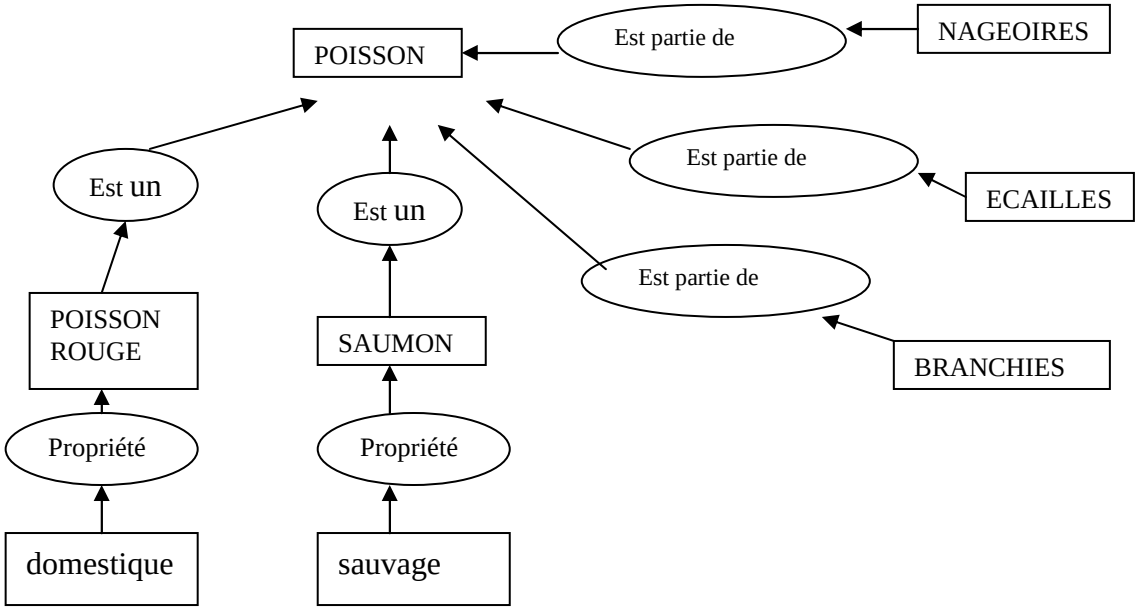


Figure 8 : Extrait du réseau ontologique de « poisson » (d'après Otman, 1996, 56)

Ces représentations partagent au moins trois constantes :

- elles comportent toujours des éléments en langue naturelle : les nœuds et les relations sont étiquetés par des éléments linguistiques,
- elles privilégient le plus souvent la forme nominale comme étiquette des nœuds, et la forme prédicative comme étiquette des relations,
- elles sont considérées à la fois comme plus simples et plus générales, plus faciles à comprendre et à manipuler (manuellement ou automatiquement) que la forme discursive.

Il n'est pas impossible que ce mode de représentation pour l'informatisation ait encore accentué l'apparence de clarté et de transparence souvent véhiculée par les travaux utilisant ce mode de représentation. Les voix qui s'élèvent pour mettre en garde contre l'illusion de cette pseudo transparence proviennent d'ailleurs rarement de la communauté informatique mais bien plus des linguistes ou des didacticiens :

> «[…] si ces représentations sont utilisées pour communiquer entre des individus ou des groupes qui ne partagent pas les mêmes connaissances, alors l'implicite sous-jacent aux choix devient important ; à notre avis il ne peut pas être éliminé complètement, il faut alors faire des hypothèses sur les connaissances partagées. » (Tiberghien, 1994, 61).

> « Mais le problème prend une tout autre ampleur lorsque les outils syntaxiques usuels sont remplacés par des symboles iconiques, notamment des flèches et des encadrés circulaires ou rectangulaires. Le propos ainsi schématisé, se présente comme une liste de substantifs fractionnés suivant un critère non explicite pour le lecteur-décrypteur […]. En outre, le code iconique, trop schématique, opacifie le propos. » (Cusin-Berche, 1997, 38).

Une fois encore, il est important de repérer que l'élaboration d'un réseau à partir d'un corpus relève d'une interprétation qui ne peut se justifier que par rapport à un objectif particulier. Cette nécessaire prise en compte est notée en didactique par exemple :

« Le type de connaissance cartographié dépend d'un énoncé scientifique initial délimité. Autrement dit, une carte correspond à un *corpus de connaissance.* Il est donc toujours nécessaire de s'accorder sur la délimitation de corpus… A titre d'exemple, si l'on réalise une carte conceptuelle sur le sang, la cartographie est différente si on a pour projet d'étudier la coagulation du sang, plutôt que la transfusion sanguine. » (Prévost et Jacobi, 1994, 122).

Ainsi, construire un réseau de connaissances à partir d'un corpus consiste à passer d'un système sémiotique linéaire, discursif et syntaxique à un autre système spatial et a-syntaxique. Ce passage se fait sur la base d'une association puis d'une catégorisation de contextes dont on considère qu'ils ont un fonctionnement sémantique similaire et interprétable sous la forme d'une relation. Un élément qui serait donc partagé par l'ensemble des locuteurs serait la capacité, plus ou moins consciente, à classifier (à « constituer des groupes »), c'est-à-dire non seulement la capacité à repérer des relations (des rapports) mais aussi à repérer la nature de ces relations :

« Les groupes formés par association mentale ne se bornent pas à rapprocher les termes qui présentent quelque chose de commun ; l'esprit saisit aussi la nature des rapports qui les relient dans chaque cas et crée par là autant de séries associatives qu'il y a de rapports divers. » (Saussure, 1982, 173).

Cette capacité à catégoriser, souvent de manière assez proche d'un interprétant à l'autre, est certainement la question la plus difficile pour la linguistique. J'essayerai de montrer que, dans le cas de l'utilisation des marqueurs, elle est due en grande partie à la capacité à associer structure linguistique et contenu relationnel, capacité probablement acquise lors de l'apprentissage linguistique en lien avec le constat inconscient de la fréquente apparition de ces éléments linguistiques avec ce contenu relationnel. Mais association ne signifie pas explication ; en d'autres termes, je montrerai que ce n'est pas l'élément linguistique qui explique (donne l'interprétation de) la relation mais seulement que sa présence, fréquemment associée à cette relation, a pu permettre de dire que l'élément marquait la relation.

1.1.2.2 Relations et définitions

Une des raisons pour lesquelles la représentation sous forme relationnelle est privilégiée vient de ce que cette modélisation est souvent considérée comme le moyen le mieux adapté pour rendre compte de la définition d'un mot. On a pu dire ainsi, dans une perspective distributionnelle, que le sens d'un mot était sa place dans un réseau lexical.
Dans la définition classique (aristotélicienne), on retrouve aussi une relation, la relation hyperonymique, qui s'instaure entre le *definiens* et le *genus* et permet au *definiens* d'être reconnu comme membre du lexique d'une langue, la relation d'hyperonymie étant considérée comme éminemment propre à la structuration linguistique. Les *differentiae* qui rendent compte des éléments qui distinguent le *definiens* du *genus* sont beaucoup plus rarement considérés comme manifestables sous une forme relationnelle. Pourtant, dans certains cas, cette représentation semble tout à fait valide :
Mégaphone : appareil servant à amplifier les sons (Le Petit Robert, 1987)
`Est-hyponyme-de` : appareil
`A-pour-fonction` : amplifier les sons
Instituteur : personne qui enseigne dans une école primaire (*ibid*.)
`Est-hyponyme-de` : personne
`A-pour-rôle` : enseigner dans une école primaire.

Ces définitions, éventuellement représentables sous une forme relationnelle, sont repérées par différents auteurs, par exemple par A.M. Loffler-Laurian qui les appelle « définitions par caractérisation » (Loffler-Laurian, 1994), mais aussi, en terminologie, par (Sager et Ndi-Kimbi, 1995).

Il faut toutefois noter que ce type de définitions semble surtout abondant dans les domaines techniques (même si le cas d'*instituteur* montre que ce n'est pas une règle absolue) dans lesquels il faut décrire des outils ou des processus et expliquer des fonctionnements. Ces définitions sont proches des définitions dites encyclopédiques, censées ne pas rendre compte du sens.

Définition en discours

Etant donné que la relation hyperonymique est une relation paradigmatique qui relève de l'axe des substitutions, on peut penser, dans une vision distributionnelle, que chaque mot qui apparaît dans un discours est le résultat d'un choix hyperonymique et, donc, qu'à chaque occurrence d'un mot correspondrait une relation hyperonymique non-exprimée, le contexte contribuant à exprimer les différences de ce mot avec son hyperonyme et avec ces co-hyponymes et donc à le définir. Ainsi la phrase *l'instituteur enseigne dans une école primaire* a l'allure d'une définition classique d'où l'on a simplement éliminé l'hyperonyme (*personne qui*). Mais la phrase *l'instituteur a parlé de ses élèves au directeur*, bien qu'elle n'ait pas l'allure d'une définition, contribue aussi à donner un sens à *instituteur*.

Cependant, même si l'expérience montre qu'en effet, chaque occurrence d'un mot constitue un élément qui aide à donner un sens à ce mot, il faut noter que la possibilité d'une représentation relationnelle directe des contextes dans lesquels apparaît un mot est plutôt rare.
Cela signifie à la fois qu'une définition, entendue comme l'ensemble des occurrences possibles d'un mot, ne peut pas toujours être représentée sous la forme de relations (et donc que ce mode de représentation appauvrit le sens) mais aussi que chacune des occurrences d'un mot, qui contribue pourtant à sa définition, n'est pas nécessairement représentable par une forme relationnelle.
En résumé, la représentation des connaissances sous forme relationnelle a certainement une pertinence réelle dans des domaines techniques et permet des manipulations intéressantes : lecture rapide, vérification de cohérence facilitée, vision d'ensemble d'un réseau de relations. Mais, dans le même temps, elle oblige à délaisser des occurrences importantes et peut-être pertinentes dans les textes faute de pouvoir les représenter. Le choix interprétatif est ainsi souvent guidé par le mode de représentation choisi.

Ce constat a deux conséquences majeures. D'une part, afin de rendre la plus efficace possible la construction d'un réseau, il est nécessaire de constituer un corpus qui soit riche en possibilité de représentation relationnelle, c'est-à-dire, un corpus qui contiennent beaucoup de « marqueurs » de relations. D'autre part, à l'intérieur du corpus, seules les occurrences dans un contexte représentable sous la forme d'une relation seront considérées, c'est-à-dire, des contextes qui jouent le rôle de marqueurs de relation. On voit combien cette notion de marqueur joue un rôle crucial. On voit aussi combien cette notion peut paraître paradoxale. En effet, si l'on se contente d'une définition très préliminaire de la notion de marqueurs (par exemple, éléments lexico-syntaxiques qui indiquent une relation), alors on fait le constat qu'un texte très spécialisé, rédigé par des experts pour d'autres experts, ne contient aucun marqueur et donc aucun terme si l'on considère que la notion de marqueur va permettre de repérer les termes. En réalité, l'expérience montre qu'il est possible, même avec des corpus très spécialisés, de mettre en œuvre une représentation relationnelle mais cela nécessite une analyse plus approfondie des occurrences et la définition beaucoup plus fine de la notion de marqueurs (*cf.* ci-dessous)[54]. Cette vision élargie de la notion de marqueur peut mettre en

[54] Finalement, la constitution d'un réseau relationnel nécessite toujours une analyse approfondie des corpus. Cela explique d'une part que cette analyse soit longue et donc coûteuse (au grand dam des entreprises qui financent ces travaux) et d'autre part, que les outils, tout en étant d'une aide appréciable, ne puissent se substituer à une

question la vision d'une représentation relationnelle seulement possible en cas de marquage explicite.

1.1.2.3 Y a-t-il des relations conceptuelles préexistantes ?

Que ce soit pour la constitution de terminologies, d'ontologies ou de réseaux sémantiques, un grand nombre de chercheurs travaillent à essayer d'établir des listes de relations que l'on convoquerait pour interpréter un corpus.

Ces constitutions de bibliothèques de relations me semblent relever de deux points de vue différents qui parfois se superposent. Dans un cas, il s'agit d'une vision philosophique ou psychologique qui considère que les concepts préexistent à leur mise en mots et qu'on peut non seulement retrouver mais aussi expliquer l'existence de telle ou telle relation. Les travaux d'inspiration wustérienne en terminologie mais aussi les approches ontologiques d'inspiration formelle en IA relèvent de cette approche. Dans un second cas, mais beaucoup plus rarement et essentiellement dans des approches linguistiques, la préexistence des relations n'est pas due à leur nature psychologique ou philosophique mais à leur très fréquente présence dans les textes et donc dans la langue. Cette façon de voir est surtout le fait de terminologues à orientation peu ou prou linguistique ((Cabré, 1998), (Kocourek, 1982), (Lerat, 1995), (Rey, 1979), (Sager, 1990)) et elle provient surtout du constat que, dans certains domaines ou applications, apparaissent des relations beaucoup plus spécialisées :

> « It is now recognised that for practical applications, virtually any number and type of conceptual relationship can be established and declared as required by a particular need. » (Sager, 1990, 29).

La question est alors moins de savoir si des relations préexistent mais, en faisant le constat de leur fréquente utilisation dans des textes de même genre, d'essayer de prédire presque à coup sûr leur présence dans tout nouveau corpus à l'étude.

Les relations les plus fréquemment rencontrées semblent ainsi être les relations d'hyperonymie, de méronymie et de causalité. C'est justement pour ces trois relations que les travaux de recherche de marqueurs sont les plus fréquents, aussi bien en termino/linguistique qu'en TAL, *cf.* par exemple : (Borillo, 1996), (Cabré et *al.*, 1997), (Garcia, 1998), (Jackiewicz, 1996), (Jouis, 1993), (Morin, 1999). En effet, de nombreux contextes associables à ces relations surgissent spontanément par réflexion introspective. Nous verrons toutefois que de nombreux autres contextes ayant le même type de fonctionnement existent, ils peuvent être décrits par la mise en œuvre d'une véritable linguistique de corpus.

Une fois encore, la notion de marqueur apparaît à la fois comme l'élément qui concrétise tous les espoirs et celui qui pose le plus de problèmes. Il est temps d'aborder ces difficultés.

1.2. La notion de marqueur de relations

Ce paragraphe fait le tour des problèmes que pose la notion de marqueur, tant en linguistique qu'en informatique (TAL, IA (Intelligence Artificielle ou en recherche d'information).

1.2.1 La notion de marqueur

Quel que soit le point de vue (lexical, morphologique, sémantique) où l'on se place, la notion de marquage est très couramment utilisée en linguistique : on dit qu'un temps marque telle ou telle valeur aspectuelle (l'imparfait marque l'imperfectivité), que le choix de telle structure marque tel ou tel fonctionnement (la position thématique marque le topique), que tel suffixe marque tel contenu (*et/ette* permet de construire le diminutif)…. Il est nécessaire de clarifier ce que cache cette abondante utilisation.

---

interprétation qui se construit au fil du temps pour viser une cohérence en fonction des données du corpus et de l'objectif de l'étude.

Dès les débuts du structuralisme, la notion de marqueur est indissociable de celle de relation (ou de rapports, dans les termes saussuriens). Que ce soit au niveau phonologique, morphologique, puis syntaxique et lexical, les rapports sont marqués (ou en tout cas peuvent être marqués). Ces marques peuvent être de nature physique (trait phonétique), linguistique (on parle d'élément marqué en morphologie comme *maisonnette* par rapport à *maison* ou *lionne* par rapport à *lion*) ou discursive (en particulier dans les relations sémantiques de toutes sortes qui peuvent être marquées dans le discours)[55].
On peut donc dire que la nature de ces marques n'est pas homogène et on peut s'interroger sur leur statut linguistique. Il semble que suivant les utilisations, le statut des « marques » oscille entre celui de signes linguistiques et celui d'éléments formels qui permettent d'instaurer un lien stable avec un contenu (une information). Plus le lien semble s'instaurer de manière systématique, plus le marqueur semble considéré comme clé d'accès à un contenu et perd son statut de signe linguistique : il devient un indice de telle ou telle information. C'est évidemment en sémantique que cette stabilité est la plus fragile, essentiellement parce que la notion de marque semble relever d'une interprétation, toujours renouvelable.

1.2.2 Les marqueurs de relations conceptuelles

Avec les marqueurs de relations conceptuelles, on retrouve cette dualité entre forme qui donne l'accès à un contenu d'un côté et signe linguistique auquel on confère un statut particulier de l'autre. Le premier de ces types de fonctionnement est nettement celui qui est privilégié par l'ingénierie des connaissances et la terminologie, le second par les analyses discursives, qui font intervenir la notion de métalangue.

1.2.2.1 Marqueurs comme indicateurs d'un contenu

Le postulat de l'existence d'éléments de forme permettant l'accès à un contenu est inséparable de l'objectif du traitement automatique de la langue. Dans cette perspective, le contenu relationnel préexiste nécessairement et il faut l'apparier avec des formes (en l'occurrence des chaînes de caractères, qui pourront être facilement repérées automatiquement) et qui sont considérées comme les marqueurs de ces relations. La plupart des outils qui aident à repérer les relations sont ainsi construits à partir de relations connues (parce que fréquentes comme je l'ai dit) et par identification le plus souvent introspective des marqueurs.
Cette approche, qui va des relations vers leur expression en discours est aussi celle qui est la plus souvent mentionnée par les terminologues et même par les lexicologues. Il s'agit de voir comment certaines relations s'énoncent en discours. Lyons, par exemple, parle de « formules » (Lyons, 1978). Souvent d'ailleurs, le choix des dénominations de ces marqueurs, aussi bien par certains linguistes que par certains informaticiens met l'accent sur leur fonction d'indices ; ainsi Cruse parle de « diagnostic frames » (Cruse, 1986) et Ahmad de « knowledge probes » (Ahmad et *al.*, 1992). Dans les deux cas, les marqueurs sont considérés comme des traces, des « sondes », qui permettent d'explorer un contenu plus profond, pour Ahmad ou de faire un diagnostic pour Cruse. Dans cette façon de voir, on attribue à un signe linguistique un rôle d'indice qui le désémantise, c'est alors essentiellement sa forme qui est considérée et ce, alors même qu'elle est souvent instable parce qu'elle s'adapte aux possibilités du discours : des marqueurs peuvent ainsi être modifiés par l'ajout d'adjectifs ou d'adverbes, ils peuvent être explosés en plusieurs phrases par des procédés d'anaphorisation… On a bien du mal

[55] Les marques dont il est question ne sont bien sûr pas toutes de la même nature. Mais le projet structuraliste a consisté à rendre compte d'un modèle unifié du fonctionnement de la langue, basé sur l'idée de système. Même si le système se met en œuvre de manière différente selon le point de vue que l'on adopte, l'utilisation de termes proches comme « marque », « marqueur », « élément marqué » me semble relever de la volonté de concevoir un système unifié, qui se manifeste par des éléments qui en sont les traces (les marques formelles).

dans ces cas-là à décrire les marqueurs sous la forme d'indices, ce qui montre qu'ils conservent leur fonctionnement de signes linguistiques.

1.2.2.2 Marqueurs comme signes linguistiques

Les travaux de linguistes qui portent sur la définition, considèrent les marqueurs comme des signes linguistiques qui ont un fonctionnement métalinguistique. Le fait de définir dans un dictionnaire relèverait ainsi d'une activité métalinguistique quasiment codée, la définition fonctionnant exactement comme un genre textuel :

> « La définition du dictionnaire doit être envisagée en fonction de la sémantique des langues naturelles, comme une manipulation de la quasi-synonymie, mais aussi en fonction de la production d'un discours didactique réglé, analogue à celui de la rhétorique, mais bien différent, et appartenant comme lui à la pratique sociale des discours. » (A.Rey, 1990, 21).

A l'intérieur de ces définitions, certains éléments, les marqueurs, joueraient plus particulièrement un rôle métalinguistique qui permettrait de structurer les éléments informationnels.

Mais l'activité définitoire ne se limite pas aux seuls dictionnaires. On peut la retrouver en discours (Péry-Woodley et Rebeyrolle, 1998), (Rebeyrolle, 2001), où l'intention définitoire est codée par des éléments de différentes natures, certains comparables aux marqueurs de définitions (marqueurs d'hyperonymie par exemple), d'autres exprimant la volonté définitoire: expressions de reformulation par exemple (Beacco et Moirand, 1995). On parle dans ces cas de véritables actes de langage définitoires et c'est sur la base d'éléments ayant un fonctionnement métalinguistique connu (codé) que s'établit la possibilité de reconnaître certains passages comme des actes de langage.

Mais dans certains cas, l'intention définitoire est nettement moins perceptible :

> « Traiter la définition comme un acte de langage ordinaire, c'est opérer d'emblée deux sortes de choix. Le premier consiste à s'intéresser en priorité aux énoncés définitoires que nous utilisons spontanément dans notre discours quotidien et dont la forme n'affiche pas ouvertement son caractère métalinguistique […]. » (Riegel, 1990, 97-98).

On peut se demander comment, dans les cas où l'intention définitoire n'est pas codée, s'élabore la notion de définition. Deux pistes peuvent intervenir dans la réflexion.

- On peut considérer que dans certains énoncés, les marques définitoires sont présentes mais relève plus de l'épilinguistique (au sens culiolien d'activité métalinguistique inconsciente) que du métalinguistique :

  > « […] la définition est non seulement une définition d'objets naturels, mais encore une définition formulée par les locuteurs eux-mêmes et non par le technicien qu'est le lexicographe. En ce sens, la définition naturelle est un des aspects de l'activité " épilinguistique " » (Martin, 1990, 87).

  Dans ce cas-là, le rôle du linguiste est de mettre au jour ces éléments, d'évaluer leur systématicité en particulier en fonction de la nature du corpus et finalement, de leur donner un statut métalinguistique de marqueur ; c'est ce que dit Auroux du savoir linguistique :

  > « Le savoir linguistique est multiple et il débute naturellement dans la conscience de l'homme parlant. Il est *épilinguistique*, non posé pour soi dans la représentation, avant d'être métalinguistique, c'est-à-dire représenté, construit et manipulé en tant que tel à l'aide d'un métalangage […] » (Auroux, 1989, 18).

- On peut aussi considérer, et ce n'est pas incompatible avec le premier point, qu'un troisième intervenant n'est pas pris en compte dans cette situation de définition, il s'agit de l'interprétant, qui n'est pas le destinataire du texte (pas l'interlocuteur)

> mais par exemple le linguiste ou l'ingénieur de la connaissance qui veut élaborer un réseau relationnel[56]. Il ne s'agit pas alors (ou pas seulement) de comprendre ce que le locuteur a voulu dire à l'interlocuteur mais d'utiliser certains contextes pour élaborer un réseau censé représenter une connaissance utile pour une application particulière et dont la présentation n'était pas nécessairement l'objet du locuteur. Cela signifie que la notion même de définition doit être réexaminée et que certains contextes peuvent sans doute apparaître comme définitoires pour certaines perspectives alors qu'ils n'ont aucune des caractéristiques formelles que l'on attribue en général à ce genre de contexte.

Cette utilisation du corpus a deux conséquences : tous les contextes qui paraissent définitoires ne sont pas systématiquement retenus pour élaborer le réseau : certains peuvent ne pas paraître pertinents, en fonction de l'objectif de la modélisation ; certains contextes, qui ne paraissent pas définitoires peuvent être retenus pour cette même élaboration.

L'intervention de ce type d'interprétant particulier (qui n'est pas l'interlocuteur) me semble capital dans le mode d'approche que nécessite la constitution de Bases de Connaissances Terminologiques. Il ne s'agit pas en effet (ou pas seulement) de rendre compte de ce qu'on pense que le locuteur a voulu dire, ni même de rendre compte de tout ce qu'il a voulu dire mais seulement des éléments qui semblent relever d'une connaissance consensuelle, modélisable sous forme relationnelle et pertinente pour un objectif. Cette façon de voir remet en question l'existence d'un métalangage fixé une fois pour toutes, qui me semble être la vision de Harris, par exemple. Dans ce cas, la métalangue est une sorte de langue qui met sous forme canonique l'ensemble des formulations possibles pour rendre compte de l'information, d'où sont extraites par transformation les phrases possibles :

> « The sequence of formulas in the articles in a given science can be looked upon as constituting discourses in a sublanguage of natural language, or alternatively as a new linguistic system structurally intermediate between natural language and mathematics. » (Harris et *al.*, 1989, 2).

> « Since all the segmentations and transformations mentioned above are paraphrastic, that is, do not change the meaning of their operand, the formulas are paraphrases of the original material, and can be considered as simply a canonical, inspectable, and processible form of the original material. » (*ibid*,3).

> « Harris montre qu'à toute phrase correspond une phrase du métalangage, et que les propriétés syntaxiques, morphologiques, ponctuationnelles, etc. de la première phrase sont des traces sur la langue de l'effacement du métalangage ». (Péry-Woodley, 2000, 65).

La position de Harris me semble trop radicale : même si la notion de sous-langage introduit une variation (il ne s'agit pas d'un système instauré pour toute la langue), cette vision du métalangage me semble trop figée : si le « métalangage est dans la langue » (comme on l'admet généralement), on voit mal comment il échapperait aux variations inhérentes au fonctionnement discursif, en tout cas à une partie d'entre elles. Comme le souligne Borillo, on peut se demander si « le métadiscours ne se dilue pas tout simplement dans le discours » (Borillo, 1985, 48).

Une fois encore, je pense que cette notion de métalangue n'a de pertinence que par rapport à l'objectif de la modélisation. L'élaboration d'un métadiscours ne me semble ainsi avoir de

[56] Remarquons aussi que la notion d'acte de langage est toujours attribué au locuteur mais l'interlocuteur, dans son processus d'interprétation du message, me semble tout aussi impliqué dans un acte de langage, dont on pourra trouver les traces dans son intervention en réponse à l'acte de langage du locuteur ; par exemple (*tu veux dire que, je ne comprends pas ce que tu veux dire, tu peux préciser*…). Bakhtine par exemple insiste beaucoup sur le rôle tout aussi actif de l'interlocuteur dans le dialogue (et pour lui, tout texte s'inscrit dans une perspective dialogique). Ainsi, l'activité de l'interlocuteur tout comme celle de l'interprétant qui n'est pas l'interlocuteur me semble relever d'une activité linguistique.

sens que, certes, par rapport à un genre discursif mais aussi par rapport à un objectif (qui relève d'un genre interprétatif). En d'autres termes, le métalangage, relève d'un construit, tout comme n'importe quelle modélisation, et l'analyse distributionnelle comporte une part d'interprétation dont le cadre doit être précisé. En revanche, le modèle est sans doute plus stable que le discours dont il est issu (c'est d'ailleurs ce postulat qui permet la modélisation quelle que soit sa nature). La possibilité de décrire des marqueurs (en tant qu'éléments métadiscursifs) s'appuie fortement sur cette hypothèse : la stabilité est plus grande puisque certains marqueurs peuvent fonctionner pour différents corpus que l'on peut caractériser d'après leur genre (voir partie suivante). La stabilité est aussi plus grande dans le temps : bien qu'aucun travail n'en ait fait la preuve, il semble pertinent d'utiliser les mêmes marqueurs pour comparer des textes rédigés à des périodes différentes. Je reviendrai sur cette question dans la partie suivante.

Ainsi, élaborer une BCT, c'est tout à la fois élaborer un modèle de connaissances dépendant d'un objectif et élaborer un modèle des marqueurs qui ont permis de construire cette connaissance. C'est bien pourquoi, un certain nombre de travaux, en particulier ceux qui sont issus du groupe TIA, préconisent de stocker dans la même base de données les résultats et les textes qui ont permis d'élaborer ces résultats.

Apparaissant en discours, les marqueurs sont nécessairement des signes linguistiques. Le rôle d'indices qu'on leur a attribué ne peut concerner que certains d'entre eux qui, fréquemment en lien avec des relations, elles mêmes fréquentes, semblent instaurer un lien très fort avec cette relation. L'analyse de corpus réels valide en partie cette hypothèse mais surtout, elle ouvre des perspectives tout à fait inédites sur le rôle des corpus dans l'interprétation de certains contextes comme des marqueurs.

2. Rôle du corpus dans la description des marqueurs de relations conceptuelles[57]

La plupart du temps, l'étude des marqueurs de relation se fait par introspection (ou en tout cas d'abord par introspection) à partir de relations connues, c'est-à-dire supposées fréquentes dans les corpus. La situation qui consiste à étudier des corpus pour élaborer des relations conceptuelles pose la question tout autrement. En effet, il faut s'attendre d'une part à ce que toutes les relations ne soient pas nécessairement connues et d'autre part, ce qui est beaucoup moins souvent évoqué, à ce que des marqueurs non-identifiés par introspection apparaissent. Il semble bien que la première à avoir essayé de voir ce qui se passe réellement dans les textes soit une informaticienne (Hearst, 1992), encore ce travail était-il réalisé à partir d'une relation connue, la relation d'hyperonymie. Par ailleurs, dans ce cas, l'objectif consistait à améliorer les performances d'un outil et pas à s'interroger d'un point de vue linguistique à la fois sur le fonctionnement et sur la nature des marqueurs. Peu de linguistes se sont intéressés à la question des marqueurs en corpus, à l'exception notable de Meyer (Meyer, 2001) au Canada et Pearson (Pearson, 1998) à Dublin qui, toutes les deux, s'inscrivent surtout dans la perspective de la traduction. C'est pourtant un thème qui, outre qu'il est très important à une époque où l'accès à l'information voire à la connaissance dans les textes est un défi majeur, permet une réflexion approfondie et un point de vue original sur la possibilité d'utiliser les corpus en sémantique. En effet, il ne s'agit pas seulement de vérifier des intuitions linguistiques mais bien d'essayer de comprendre ce que signifie, d'un point de vue linguistique (du point de vue d'une théorie linguistique) élaborer des relations à partir de corpus en utilisant certaines des parties de ce corpus (les fameux « Knowledge rich contexts »

[57] Les premières bases de la réflexion qui suit sont données dans (Condamines, 2002).

pour utiliser la terminologie de Meyer). Comme je l'ai déjà dit ci-dessus, deux types de questions peuvent se poser, selon le lien que le marqueur entretient avec le corpus :

- dans le cas où ce lien s'instaure avec un genre de corpus, comment passer d'une connaissance épilinguistique à une connaissance métalinguistique ; c'est-à-dire pour le dire plus simplement comment rendre conscient (et donc éventuellement systématisable) l'utilisation de certaines portions de corpus pour élaborer des relations ;
- dans le cas où ce lien ne s'instaure qu'avec un corpus particulier, comment passer directement d'une connaissance épilinguistique à une représentation relationnelle, sans que ce passage soit systématisable car il n'est pertinent que pour un corpus particulier et un objectif particulier.

Finalement, à travers ces questionnements, se profile la question qui interpelle tout linguiste : qu'est-ce que le « savoir linguistique », pour reprendre le terme d'Auroux ?
Cette question prend une acuité particulière puisque, dans le cas de la constitution de réseaux relationnels, la dimension interprétative est incontournable. Il s'agit donc de comprendre comment cette élaboration se met en place, en faisant appel à quelles connaissances et comment la méthode d'interprétation peut être généralisée.

Alimentés à la fois par certains terminologues et certains informaticiens, un certain nombre de postulats sont généralement véhiculés autour de la question des relations et des marqueurs de relations qui influencent beaucoup les analyses mises en place. Je présenterai ces postulats, puis, je montrerai comment on peut catégoriser les fonctionnements des marqueurs en corpus. Enfin, je reviendrai sur la question de l'application, de son rôle sur la représentation relationnelle, et même sur le fonctionnement des marqueurs.

2.1. Postulats sous-tendant la constitution de réseaux, en terminologie et en intelligence artificielle

Que ce soit en terminologie, avec les réseaux notionnels, ou en intelligence artificielle avec les ontologies, un certain nombre d'idées sont sous-jacentes aux travaux et réflexions, idées qui ne sont pas toujours très clairement présentes à l'esprit de ceux qui les utilisent tant ils en sont imprégnés. Pourtant, elles constituent un cadre de pensée préétabli qui a une influence importante sur la façon de considérer la question des marqueurs de relation. Ces postulats peuvent être résumés en points majeurs.

- Un seul système linguistique, des marqueurs accessibles par introspection

Pour les terminologues post wustériens, mais aussi pour beaucoup de ceux qui s'inscrivent dans une vision plus linguistique, le problème de la variation n'est pas pris en considération. Dans une vision descendante, qui va des relations vers leurs réalisations discursives, on considère qu'un seul système, fort et stable, est à l'œuvre. Ce sentiment est certainement renforcé par l'idée que les marqueurs ne sont pas des éléments lexicaux comme les autres puisqu'ils ont un rôle particulier, métalinguistique voire indiciel, qui renforce leur stabilité. Corrélée à cette idée d'un système stable, on trouve la possibilité de décrire ces marqueurs par introspection, ce qui suppose que les relations, elles aussi, sont stables et accessibles par introspection.

- Binarité des relations

A ma connaissance, les ontologies et les terminologies sont toujours organisées à partir de relations binaires. On peut prendre en compte que A est en relation avec B et que A est en relation avec C mais pas que A est en relation avec B seulement s'il est aussi en relation avec C. Par exemple, on ne peut pas représenter une méronymie si elle est conditionnée par le lieu ou le temps, ce qui est pourtant très fréquent. Par exemple (phrase forgée) *: En 1995, l'entreprise X comprenait 2 succursales : Y et Z.*

La prise en compte de relations n-aire est bien sûr possible en IA (*cf.* ci-dessous) mais elle apparaît rarement dans les ontologies. Quant aux terminologies, je ne crois pas que cette possibilité ait été prise en compte. C'est un fait étonnant et, en tout cas pour ce qui concerne la terminologie, il est possible que cette vision rigide ait été en lien avec la volonté de maintenir éloignées les constructions terminologiques du fonctionnement discursif qui s'élabore de manière bien plus complexe que la seule binarité (rappelons-nous que, pour Wuster, les possibilités créatives de la langue étaient considérées comme une véritable menace pour l'activité terminologique).

– Le marqueur donne le sens de la relation

C'est peut-être le postulat le plus répandu, même chez les linguistes : non seulement le marqueur indique la présence de la relation mais il donne en plus le mode d'interprétation de cette relation. D'une certaine façon, cette interprétation est le sens du marqueur, qu'il transporte avec lui dans tous les contextes.

L'analyse détaillée des contextes à partir desquels peuvent s'élaborer des relations montre que tous ces postulats doivent être remis en question : oui il y a de la variation, qui, dans certains cas, peut être corrélée avec le genre textuel ; non, les relations binaires ne suffisent pas, non les marqueurs de relation ne donnent pas (toujours) le mode d'interprétation de la relation.

2.2. Etude du fonctionnement des marqueurs en corpus[58]

Les études que j'ai menées sur les rapports entre marqueurs de relation et corpus m'amènent à penser que l'on peut considérer trois types de fonctionnement : dans les cas extrêmes, il n'y a quasiment pas de dépendance ou inversement, cette dépendance est très forte ; dans les cas intermédiaires, on peut établir cette dépendance non pas avec un corpus mais avec un corpus en tant qu'il appartient à tel ou tel genre.

2.2.1 Cas extrêmes : lien marqueur/corpus quasi inexistant ou quasi total

Dans le continuum de dépendance qui s'établit entre marqueurs et nature du corpus, on peut considérer que les deux cas extrêmes concernent d'une part les cas où il semble que la dépendance soit faible, et d'autre part ceux où la dépendance avec un corpus est quasi totale.

2.2.1.1 Dépendance marqueur/corpus quasi inexistante

Il s'agit d'abord de comprendre ce que recouvre une telle indépendance. Elle signifie que, dans le triplet relation/marqueur/corpus, la seule dépendance pertinente s'instaure entre relation et marqueur, sans que le corpus (ou son genre) influence cette relation de quelque façon que ce soit.

Mon expérience d'analyse de nombreux corpus me permet de dire qu'en effet, cette relation directe entre relation et marqueur semble s'établir dans certains cas pour des relations très fréquentes et certains de leurs marqueurs. Quelques exemples :

Hyperonymie : [dét N1 être dét N2],

Les gestes sont des actions des extrémités, sans implication de transfert ou de support du poids du corps (Corpus de danse)[59].

[58] Dans toute la suite de la présentation, j'utiliserai la notation suivante :
les relations sont indiquées en police `courrier`, les marqueurs sont indiqués [entre crochets] (ils correspondent à une abstraction d'un ensemble de formes discursives, qu'il faut donc identifier lorsque l'on souhaite utiliser ces marqueurs pour faire de la recherche en corpus), les exemples issus ou non de corpus réels sont donnés *en italiques*.

[59] Les exemples cités dans ce paragraphe sont extraits des corpus dont nous disposons à l'ERSS, interrogés grâce au logiciel Yakwa, développé par L.Tanguy ; il ne me semble pas utile de préciser plus avant leur contenu.

L'électron est une particule en mouvement (Corpus de chimie).

Méronymie : [dét N1 comprendre dét N2 (et dét N3)]
L'acceptation client comprend l'ensemble des opérations de contrôle (corpus MOUGLIS)
Un système vertical comprend deux piliers antérieurs (naso-ethmoïdo-frontal), deux piliers latéraux (malaire et zygomatique), deux piliers postérieurs (ptérygoïdiens) (Corpus de médecine).

Dans ces cas-là, on rejoint l'approche introspective qui neutralise le contexte d'utilisation des discours pour se concentrer sur l'expression possible de telle ou telle relation. Comme ce lien relation/marqueur se trouve fréquemment en discours, il présente une grande stabilité et on a tendance à penser qu'il s'instaure systématiquement entre un contenu et une forme, indépendamment de tout contexte, ce qui est exact d'une certaine façon si l'on se place du strict point de vue de la recherche d'information.
Toutefois, il faut prendre garde à ne pas penser que ce mode de fonctionnement qui consiste à établir une équivalence entre relation/marqueur d'une part et forme/contenu d'autre part représente le type même du fonctionnement des marqueurs.
Si l'on y regarde de plus près en effet, on s'aperçoit que même avec des relations très générales, le fonctionnement discursif fait que l'instauration du lien forme/contenu qui permettrait d'utiliser le premier pour accéder au second n'est pas si simple. Trois types d'éléments sont à considérer.

– Même si la dépendance marqueur/relation est forte, cela ne signifie pas que ce marqueur va se retrouver dans n'importe quel corpus, en effet cette relation peut ne pas apparaître dans tous les corpus. Ainsi, dans le corpus de Matra Marconi Space, qui était constitué de spécifications de satellites (donc rédigées par des experts pour d'autres experts), nous n'avons trouvé aucun marqueur connu de la relation d'hyperonymie. Tout simplement parce que la relation d'hyperonymie n'apparaissait pas dans ces contextes.

– Même si le marqueur se trouve dans le corpus, la relation n'est parfois pas retenue. Cela peut venir de différentes raisons. Dans certains cas, la relation semble relever d'une appréciation personnelle plus que d'une connaissance partagée par un collectif de locuteurs[60]. Par exemple la phrase suivante, extraite de Germinal :

Le brigand est le vrai héros, le vengeur populaire, le révolutionnaire en acte.

Cette phrase correspond exactement à une phrase contenant le marqueur d'hyperonymie classique [dét N1 être dét N2 + X] (X pouvant être un pronom relatif, un adjectif ou un participe passé…). Si on décide de ne pas la modéliser sous la forme relationnelle, c'est parce qu'on considère qu'elle ne relève que d'un jugement individuel. Cette appréciation est renforcée par la présence de *vrai* qui, justement, souligne qu'on pourrait penser que le brigand n'est pas un héros.

Dans d'autres cas, la relation n'est pas retenue parce qu'il est quasiment impossible de décider quels sont les éléments qui sont mis en relation : il est impossible de les maintenir dans une forme enregistrable dans une structure en réseau. C'est le cas dans : *L'athérosclérose est une lésion focale de la paroi des artères de gros et moyen calibre qui consiste en un épaississement localisé de l'intima, formant une plaque où s'associent l'athérome (dépôts graisseux) au centre et la sclérose au pourtour, enchassant le " coeur lipidique " de la plaque.* (Corpus de médecine).
Cet énoncé contient une structure définitoire classique de type [dét N1 être dét N2 + X] : *l'athérosclérose est une lésion.*…Si l'on distingue aisément quel terme va être défini :

[60] La nature du corpus réapparaît dans ce cas ; on a par exemple, plus de chances de trouver des connaissances partagées dans un manuel technique que dans un article de critique cinématographique ou une œuvre littéraire.

athérosclérose, on a beaucoup plus de difficulté à identifier, dans le contexte droit, les éléments qui constituent un éventuel hyperonyme : *lésion ?, lésion focale ?, lésion focale de la paroi des artères ?* Tout dépend si ces éléments vont être retrouvés dans la même position d'hyperonymes ailleurs dans le corpus ; si c'est le cas, c'est un argument pour retenir un composé plutôt qu'un autre. Mais ce « test » n'est pas toujours suffisant et il arrive assez souvent que des énoncés de ce type ne soit pas retenus parce qu'on n'arrive pas à les intégrer dans une structure de BCT.

– Enfin, cette dépendance marqueur/relation n'existe réellement qu'avec certains marqueurs, très généraux. Dans sa thèse, Séguéla (Séguéla, 2001) montre que certains patrons connus d'hyperonymie et de méronymie ne se répartissent pas également dans tous les corpus ; Rebeyrolle et Tanguy (Rebeyrolle et Tanguy, 2000) arrivent au même résultat avec les marqueurs définitoires.

Finalement, les cas d'indépendance totale entre corpus et marqueur sont rares, c'est-à-dire des cas où un lien direct, très stable, s'instaure entre marqueur et relation. Même dans ces cas-là, nous le verrons, il est difficile de dire que le marqueur donne le sens de la relation ; c'est seulement par l'instauration d'une régularité entre marqueur et relation qu'on a eu tendance à interpréter cette régularité comme un système dans lequel le marqueur donne accès à la relation.

Marqueur et compétence de locuteur

Dans le cas où un lien marqueur/relation indépendant d'un corpus est valide, il semble qu'il soit en rapport d'une part avec la capacité d'introspection de l'interprétant et d'autre part avec la fréquence d'apparition de ce doublet. Voyons ce que cela signifie plus précisément. On peut parler d'introspection puisque les marqueurs concernés sont ceux qui sont les plus souvent cités par les linguistes mais aussi parce que ce sont ceux qui sont le plus souvent repérés en discours. J'ai ainsi pu mener une expérience avec des élèves ingénieurs en informatique auxquels j'ai proposé des exemples où l'on pouvait repérer une relation d'hyperonymie comme *le processus de conception est un processus continu d'affinage* ou *les éléments d'infrastructure logiciel tels que les outils de test, les scénarii et procédures de test* (corpus MOUGLIS). Ces élèves ont su retrouver les marqueurs de cette relation ainsi que les éléments reliés avec, me semble-il, autant de facilité que des étudiants linguistes. Cette compétence métalinguistique semble donc commune aux locuteurs d'une langue. On peut faire l'hypothèse que, lors de l'apprentissage de la langue, l'association fréquente entre ces marqueurs et cette relation a été assimilée quasi inconsciemment et sans que la situation d'apprentissage (et donc la nature du discours) soit enregistrée simultanément. Dans ce type de cas, rare en réalité car il concerne peu de relations et peu de marqueurs de ces relations, le rôle du linguiste (qui possède ce que j'appelle la **compétence de linguiste** par opposition à **compétence de locuteur**) n'est pas différent de celui d'un locuteur qui s'interroge sur sa compétence langagière et sur ce terrain se retrouvent aussi bien les terminologues que les ingénieurs de la connaissance qui travaillent à partir de textes, voire ceux qui travaillent par introspection.

En revanche, ce qui est beaucoup moins facile à repérer, c'est le moment où la dépendance relation/marqueur cesse d'être pertinente si l'on ne fait pas intervenir d'autres éléments qui prennent en compte à la fois la nature du corpus et l'objectif de la modélisation. Là me semble être fondamentalement le rôle du linguiste construisant un réseau relationnel : examiner comment, en tant qu'élément discursif (et pas seulement forme), un « marqueur » participe à l'élaboration d'un sens qui n'est pas un donné mais bien un construit qui fait appel à différentes connaissances qu'il faut essayer d'identifier.

Le métadiscours est moins spécialisé que le discours

Une autre difficulté est soulevée par ces marqueurs non-dépendants d'un corpus, difficulté que l'on rencontre d'ailleurs avec tous les marqueurs : les marqueurs, tout comme les relations qui sont élaborés grâce à eux, semblent moins spécialisés que les termes qu'ils sont censés relier. En effet, il semble bien que, dans les systèmes relationnels (thésaurus, BCT, ontologies…), les relations soient moins expertes que les termes reliés. La relation la plus nettement privilégiée est la relation `est-un`, mais même lorsque des relations associatives apparaissent, elles restent compréhensibles par des locuteurs non-experts et cela, même lorsque la représentation est faite dans un domaine spécialisé, par et pour des spécialistes. Ainsi, dans UMLS (Unified Medical Language System), un thesaurus médical réalisé par la NLM (National Library of Medicine), aucune des 54 relations retenues n'est incompréhensible pour un non-expert[61]. Certaines sont assez nettement typiques du domaine médical comme `diagnoses` ou `developmental-form-of` mais elles restent accessibles à des non-experts ; la plupart sont très générales, indépendantes du domaine : `ingredient-of, location-of, produces`... La spécificité du domaine se trouve ainsi pratiquement tout entière dans les termes reliés et, éventuellement, les qualificatifs. On a l'impression ainsi que l'élaboration d'une modélisation relationnelle suppose une simplification, une généralisation des relations au bénéfice des éléments reliés qui, de ce fait, contiennent toute la spécificité d'une connaissance. La question qui reste entière est celle de la raison de cette simplification : on peut envisager deux possibilités : soit les possibilités relationnelles sont linguistiquement beaucoup plus restreintes que les constructions dénominatives, soit, ces constructions sont (et ont été), dans la plupart des cas, faites par des interprétants moins experts que les rédacteurs du corpus étudié. On peut supposer ainsi que, dans la grande majorité des cas, les thésaurus et autres systèmes relationnels sont construits par des non-experts ou des semi-experts. Lorsqu'ils élaborent les représentations à partir de corpus, ils se basent sur les éléments qu'ils connaissent c'est-à-dire sur des marqueurs qui leur sont familiers et des relations tout aussi familières.

Ainsi, soit parce que les relations et leurs marqueurs sont linguistiquement moins nombreux que les éléments à relier, soit parce que les interprétants qui les construisent sont moins experts que ceux qui ont écrits les corpus, la représentation sous forme relationnelle est nécessairement une simplification du contenu d'un corpus. C'est à ce prix seulement que peut s'élaborer la modélisation mais aussi l'étude linguistique des marqueurs et probablement la définition d'un métalangage. Il s'agit en effet de rassembler sous une même relation des structures linguistiques qui peuvent être très différentes.

Mais cette simplification n'est pas forcément problématique si elle permet de rendre compte d'une connaissance en lien avec le contenu d'un corpus et un objectif particulier, clairement identifié. C'est d'ailleurs cette simplification qui permet de stabiliser certains marqueurs et qui autorise à penser que certains marqueurs traversent les domaines et les époques sans dommage (si l'on pense par exemple aux contextes définitoires qui sont déjà décrits dans Aristote). C'est, en tout cas, la seule hypothèse qui permette aux linguistes d'essayer de systématiser l'analyse des marques linguistiques pour accéder à tel ou tel contenu. Il y a simplification (abstraction diraient certains) et donc à la fois perte de sens et gain en efficacité.

2.2.1.2 Dépendance relation/marqueur quasi totale

A l'inverse du cas précédent, il semble que dans certains énoncés, la possibilité du lien marqueur/relation soit très dépendante du corpus. C'est cette situation que nous avons rencontrée dans l'exemple suivant. Le corpus d'étude était MOUGLIS. Rappelons-le, il

[61] http://www.nlm.nih.gov/research/umls/META3.HTML

s'agissait d'un manuel rédigé par le service qualité d'EDF et qui était censé être utilisé par les ingénieurs pour spécifier des logiciels et pour rédiger la documentation qui accompagne ces spécifications. La consultation de ce manuel n'était, en principe, pas laissée au libre choix des ingénieurs ; elle était au contraire oligatoire et ce mode d'utilisation a une influence à la fois sur le mode de rédaction et sur le mode d'interprétation.
Dans ce corpus, la condition est marquée de la manière suivante (Condamines et Rebeyrolle, 1997) :

[((phase, étape), déverbal) + (lorsque, dès que) + V au passif]

comme dans :
La phase d'intégration du composant peut commencer lorsque l'ensemble des éléments logiciels ont été codés.
Cette transition est déclenchée par le responsable des développements dès lors que toutes les manifestations spécifiées sont implémentées.
Ces formules qui semblent relever de la simple expression d'une succession relèvent en réalité de l'injonction ; ainsi, les exemples précédents doivent être interprétés de la manière suivante :
La phase d'intégration ne peut commencer que lorsque l'ensemble…
Cette transition ne peut être déclenchée par le responsable … que lorsque …
Il faut noter que dans ce même corpus, les subordonnées temporelles peuvent aussi marquer seulement la succession :
V est incrémenté lorsque le logiciel subit une évolution ou une modification majeure.
Cela montre bien que dans les phrases à interprétation injonctive, le marqueur n'est pas constitué par les seules conjonctions de subordination mais par une structure bien plus complexe, qu'il a fallu préciser. Il faut reconnaître aussi que notre attention a été attirée sur cette relation par des énoncés où le marquage de la condition était lui, explicite :
Conditions de passage à la phase suivante :
La phase d'intégration du produit est achevée lorsque tous les tests prévus au plan d'intégration ont été exécutés avec succès.

En revanche, le marqueur décrit ci-dessus était imprédictible. Sans être très fréquent dans le corpus (il apparaît 11 fois) il est très efficace puisque tous les énoncés le contenant ont été retenus et représentés soit par la relation `conditionne-le-début-de` soit par la relation `conditionne-la-fin-de.`
Ce marqueur fonctionne sur le mode épilinguistique. En effet, il peut être repéré et décrit (en tout cas par des linguistes) mais il n'est pas présent à la conscience et ne peut être spontanément proposé.
Par ailleurs il semble difficile de circonscrire un genre textuel dans lequel il serait susceptible d'apparaître : ce n'est ni le fait que ce corpus soit un manuel ni le fait que le style soit injonctif qui suffisent à expliquer ce mode de fonctionnement. Il semble que ce type de marqueur, instable, puisse apparaître dans des groupes de locuteurs plus ou moins éphémères. Il faut noter en effet que le manuel dont il est question ci-dessus n'a pas été écrit par une seule personne mais bien par un groupe de rédacteurs qui ont ajusté leur façon de rédiger, certainement de manière inconsciente.
Dans ce cas, la notion d'appartenance à un genre n'est pas pertinente et les études linguistiques ne peuvent consister qu'à expliquer comment on peut élaborer des méthodes pour repérer ces marqueurs en corpus. Cela ne signifie pas que d'autres fonctionnements linguistiques ne sont pas caractéristiques des manuels par exemple (*cf.* le chapitre IV sur l'utilisation massive de nominalisations et la non-apparition d'argument agent dans le corpus

MOUGLIS par exemple), mais que certains marqueurs de relation échappent à la caractérisation par rapport à un genre de corpus et ne sont pertinents que pour un corpus particulier, plus précisément pour un texte. Tout en étant interprétables *a posteriori*, et donc sans doute pas en nombre indéfini, ces marqueurs ne sont pas prédictibles.
Leur fonctionnement épilinguistique ne peut pas, dans ce cas, être systématisé, c'est-à-dire acquérir un statut métalinguistique.

2.2.2 Dépendance en fonction du genre du corpus

Dans certains cas, probablement plus nombreux que ce que l'on croit, le fonctionnement de certaines formes ou structures est lié au genre du corpus.
Deux prépositions correspondent à ce mode de fonctionnement : *avec* et *chez* qui peuvent être associées à la relation partie-tout (composant).

2.2.2.1 Le cas de avec

La possibilité pour *avec* de marquer la méronymie est généralement repérée dans les descriptions introspectives, par exemple dans (Cadiot, 1997) ou (Choi-Jonin, 1995). Dans le TLF, cette possibilité est associée à l'idée d'accompagnement (*une robe avec des dentelles*), sans doute par fidélité à l'étymologie. Ce qui n'est en revanche jamais évoqué, c'est le fait que certains genres de corpus utilisent de manière privilégiée cette préposition pour marquer la méronymie. Pourtant, sans mener une investigation très approfondie, on peut trouver des corpus dans lesquels cette utilisation privilégiée est nette. Trois de ces genres corporaux sont présentés ici.

1- Les catalogues de jouets

C'est un peu par hasard que nous avons rencontré la possibilité de marquer la méronymie par *avec* de manière régulière, lors d'une étude des catalogues de jouets (thèse de Christine Pernet[62], en cours). Une des informations régulières données dans ce type de support concerne en effet les composants du jouet. Cette information apparaît de deux manières : soit directement, sans aucune précision :
Voiture à pédales : châssis en acier et carrosserie en plastique (La Grande Récré).
Soit introduite par un marqueur. A coté de marqueurs classiques comme *comporte* ou *contient*, peu utilisés[63]:
Une malette de beauté, contient 7 rouges à lèvres (La Grande Récré),
Circuit routier électrique, comprend 2 beetles avec phares (La Grande Récré),
le marqueur le plus fréquent est *avec :*
Cuisine avec coin-cuisine, égouttoir, plaque chauffante, four… (Leclerc).

Dans quelques cas minoritaires, *avec* apparaît dans des structures en complément de prédicat où il ne marque plus la méronymie ; ces cas semblent assez faciles à discriminer sur des bases syntaxiques :
En le guidant avec la télécommande, ce bébé marche à 4 pattes.
Appuie avec le stylo sur la fiche.

[62] C. Pernet a constitué le corpus sur les catalogues de jouets dont sont extraits les exemples proposés.
[63] On peut considérer que la structure [N : N] correspond à une structure définitoire (et on peut d'ailleurs s'interroger sur la nature « définitoire » de ces descriptions de jouets). Mais cette structure ne dit rien sur la nature du second N : il se trouve qu'ici, ce N renvoie à un méronyme mais dans les définitions classiques, il est plutôt un hyperonyme.

On peut retenir que le marqueur privilégié de la composition dans ce genre de corpus est *avec*. Il n'est pas impossible que le choix de ce marqueur soit lié à sa brièveté : les descriptions de catalogues doivent être courtes !

2- Les petites annonces immobilières

C'est probablement pour la même raison de brièveté que cette même préposition *avec* apparaît de manière très régulière dans les petites annonces immobilières, pour introduire des parties. On trouve d'ailleurs, dans ces petites annonces des version raccourcies (!) d'*avec* : *av* ou *avc*.

Comme dans les catalogues, les parties sont soit données directement (*T3, cuisine équipée*), ces cas étant les plus nombreux, soit introduites par un marqueur. Ce marqueur est rarement différent d'*avec* mais il peut être *comprendre* par exemple : *villa T4 comprenant séjour avec cheminée.*
En revanche, contrairement aux catalogues de jouets, dans les petites annonces, il ne s'agit pas de donner une information que les lecteurs sont censés avoir déjà, c'est-à-dire qu'on ne trouve pas indiquées des parties que l'on rencontre toujours dans une habitation. Ainsi on ne trouve pas : *un appartement avec toilettes*. Si une partie est indiquée, c'est, soit parce qu'elle ne fait pas nécessairement partie des composants habituels d'une maison :
T5 avec garage,
Belle villa avec piscine,
Séjour avec cheminée
soit parce qu'elle a une caractéristique particulière, manifestée syntaxiquement par un modifieur:
T2 avec chauffage au gaz,
T4 avec cuisine aménagée.

Dans les deux modes de fonctionnement, la partie est considérée comme un atout et cet intérêt est fréquemment manifesté par +, qui apparaît soit à la place de *avec* soit en coordination avec lui :
T3 de 63 m2 hab. + terr. de 21 m2 + pkg s/sol.
T2/3 avec balcon + cellier.

Il semble d'ailleurs que ce + soit beaucoup plus rare dans les annonces immobilières d'achat que dans les annonces de ventes : ce sont surtout les vendeurs qui veulent mettre en évidence les avantages de la maison qu'ils vendent. En revanche, *avec* est utilisé dans les annonces d'achat.

Avec non marqueur de méronymie

Tout comme dans les catalogues, *avec* ne marque pas toujours la méronymie mais ces cas sont beaucoup plus difficiles à discriminer syntaxiquement que dans les catalogues. En effet, les cas où *avec* apparaît après un verbe sont très rares bien que pas complètement impossibles :
Belle ferme rénovée avec grand confort et goût
(il est probable que cette annonce provient d'une agence ; une telle affirmation serait peu recevable de la part d'un propiétaire).
En revanche, dans les petites annonces, *avec* peut apparaître dans des structures en tous points identiques à celles où il marque la méronymie [(dét)N1 avec (dét)N2] alors qu'il n'a pas le sens de composant (je reviendrai sur ces cas) :
Très belle villa avec prestation haut de gamme
T3 dans immeuble avec gardien
Villa type 5 avec vue panoramique sur les Pyrénées.

Les exemples de ce type restent peu nombreux.
On peut donc constater que, dans les petites annonces immobilières, lorsqu'il y a un marqueur de méronymie (c'est-à-dire dans environ 10% des cas (63 fois sur 665 annonces consultées)), ce marqueur correspond à la structure [(dét)N1 avec (dét)N2].

Qu'en est-il d'autres types d'annonces ?

Il est intéressant de noter que ce marquage de la méronymie par *avec* ne se retouve pas dans tous les types d'annonces. On le trouve, très rarement, dans des annonces de vente d'objets divers :
Siège coquille réglable av harnais sécu.
Offre 4 volets projection en Sipo avec armature.
En revanche, *avec* n'est jamais utilisé pour marquer la méronymie dans les annonces de vente de voiture où la nature de cette information, comme de toutes les autres d'ailleurs, n'est tout simplement pas marquée. Les informations sont données dans une énumération qui semble parfois totalement ésotérique au lecteur non averti, sans doute parce que le lecteur censé être intéressé est considéré comme initié (ce qui permet aussi des ellipses et réductions morphologiques drastiques !) :
BMW 525 TDS pack, clim, ABS, cuir, an 94…
530 D pack 193ch 01 7000 kms gris clim. JA rad ABS DSC BM5 opts ét nf int gar usin pos cred rep. (sic !).

3- Les descriptions d'itinéraires

Parmi les corpus disponibles dans le laboratoire ERSS, il y a un corpus d'enregistrements retranscrits de description d'itinéraires[64].
Le marquage de la relation méronymique par *avec* est manifeste. Toutefois, comme pour les petites annonces, les parties décrites ne constituent pas des parties essentielles des objets mais, l'objectif de ces descriptions étant de donner des points de repères à un interlocuteur ne connaissant pas les lieux, les parties citées sont des éléments visuellement saillants :
Il y a une rue piétonne avec des pavés par terre,
Vous la repérez grâce à une église avec une coupole,
Il y a des entrées de métro, avec de grands M.

Cette dimension de saillance fait que, dans certains cas, le N1 dans la structure [détN1 avec détN2][65] sert de point de référence plutôt que de tout englobant :
Traverser un mini-parking, avec un plus grand parking sur la gauche.
Dans quelques cas, la « méronymie » est marquée par le verbe *avoir* ou la structure *il y a* :
C'est une rue qui a deux angles
Un grand rond-point où il y a une allée au milieu.

Quoi qu'il en soit, la structure [dét N1 avec dét N2] est nettement majoritaire puisqu'elle apparaît 106 fois sur les 131 occurrences de *avec*. Dans les autres cas, *avec* apparaît après un verbe, dans des structures variables :
On se retrouve avec le Pont St Michel à droite,
La rue des Changes qui est souvent confondue avec la rue Pargaminière,
Le Bar Basque qui fait angle avec la rue Pargaminière.

[64] Ce corpus a été constitué par K. Ricalens pour une thèse en psycholinguistique.
[65] La structure dans laquelle s'insère *avec* comprend nécessairement des déterminants ; en effet, il s'agit d'une structure discursive classique. Dans les deux genres précédents, sans doute essentiellement pour des raisons de place, il est courant que les noms fonctionnent sans déterminant.

Remarques à propos de la méronymie en discours

Sans entrer dans le détail du fonctionnement de la méronymie en discours, il me semble important de noter combien le fait de s'intéresser aux éléments qui la marquent amène à s'interroger sur la relation méronymique elle-même.

Composant essentiel vs composant accessoire

Le genre « catalogue de jouets » semble correspondre à un fonctionnement particulier de la méronymie, particulièrement notable avec la préposition *avec* (mais également présent avec les autres marqueurs de méronymie). Il semble bien en effet que, dans ce genre de corpus, et contrairement me semble-t-il, à ce que disent les descriptions habituellement données, cette préposition introduit souvent les composants principaux de l'objet décrit, c'est-à-dire lorsqu'il n'est pas un jouet mais un élément du monde des adultes. Il s'agit d'insister sur la similitude du jouet avec son modèle adulte en rappelant ses composants. D'ailleurs, d'autres éléments lexicaux montrent que l'objectif de la description est de montrer à quel point le jouet ressemble à l'objet grandeur nature et l'on trouve fréquemment utilisés des termes comme *vrai* ou *véritable* :
Véritable boutique de mode avec son miroir (La Grande Récré).
(le possessif renforce encore ici l'idée d'une partie inaliénable).
Cet exemple serait à opposer par exemple à *un chapeau avec des rubans* à partir duquel on pourrait penser que les rubans ne sont pas une partie indispensable du chapeau.
Cette situation semble propre aux catalogues de jouets et il n'est pas certain que l'on retrouve le même type de fonctionnement dans un autre type de catalogue. En tout cas pas des structures [dét N1 avec dét N2] sans aucun modifieur de N2. En revanche, on trouve des structures avec N2 modifié dans d'autres catalogues (par exemple dans La Redoute, été 2002 *: coussin avec boutons recouverts*).

Quand le marqueur crée la méronymie

Si, dans certains exemples, on a le sentiment que ce qui se marque, ce sont des parties que l'on aurait pu définir à l'avance comme essentielles ou secondaires, le discours utilise aussi les mêmes structures contenant *avec* pour renvoyer à des éléments que l'on a du mal à identifier comme étant des composants. On peut avoir deux attitudes face à ces éléments : soit, dans une vision très discontinue du sens, considérer que l'on a affaire au marquage d'une autre relation (qu'on aurait d'ailleurs bien souvent du mal à identifier), soit considérer que l'on a affaire à une méronymie au sens large, qui prend sens dans le cadre du point de vue du genre du texte.
Prenons deux exemples, l'un issu d'un catalogue de jouets, l'autre déjà mentionné, issu d'une petite annonce immobilière :
VTT à roues libres avec un panier, un chien et ses lunettes, un casque pour Betty.
Immeuble avec gardien.
Dans aucun de ces énoncés, on peut considérer que l'on a une partie dite essentielle de l'objet principal décrit. Pourtant, ce n'est pas par hasard que le même marqueur de méronymie est utilisé. Ce qui peut paraître comme une énumération à la Prévert dans le premier cas correspond à des éléments qui font sens pour l'enfant censé être intéressé par le VTT et qui, de ce fait, prennent une importance particulière au point de devenir des éléments majeurs du VTT décrit : la description ne s'appuie pas ici sur une méronymie « préexistante », cette méronymie est construite par le discours. De même, nul ne pense qu'un *gardien* constitue une des parties d'un immeuble ; mais, dans la perspective de la description valorisée d'un logement, la présence d'un gardien est présentée au même titre qu'une des pièces du logement ; avec le point de vue qui consiste à montrer tous les éléments positifs d'un appartement pour le vendre, le gardien est mis sur le même plan que la cuisine équipée. Là

encore, le discours construit sa propre méronymie. Il reste que c'est le rôle de l'interprétant de décider de conserver ou non la méronymie construite par le discours. Si l'on voulait construire une ontologie à partir des petites annonces immobilières, il est probable que l'on ne mettrait pas sur le même plan la cuisine et le gardien mais ce serait parce que l'on fait intervenir une connaissance qui vient contredire les données du texte : compétence de locuteur, objectif de l'étude qui viendraient rétablir un ordre dans ce qui apparaît comme une présentation hétéroclite.

Conclusion

Dans ces trois genres de corpus : catalogues de jouets, petites annonces immobilières, description d'itinéraires, *avec* fonctionne majoritairement comme un marqueur de méronymie et, qui plus est, quasiment comme le seul marqueur de méronymie. Il est probable que ce même type de fonctionnement se retrouverait dans d'autres genres de corpus. Inversement, je l'ai trouvé moins de quinze fois sur 616 occurrences de *avec* dans Bel Ami de Maupassant :
La figure du Comte de Vautrec, un peu vieux déjà, avec des cheveux gris,
De jeunes gens en costume d'assaut, minces, avec des membres longs.

Toute la difficulté est alors d'arriver à déterminer les genres de corpus pertinents. C'est par hasard que le genre catalogue de jouets a été trouvé comme pertinent, pour les petites annonces, c'est l'intuition qui a joué et, pour les itinéraires, le tâtonnement (je n'avais pas d'attente particulière pour ce genre de corpus).
Ce problème de la détermination des corpus potentiellement pertinents pour décrire tel ou tel phénomène linguistique, en lien avec un genre, reste la pierre d'achoppement de cette forme de linguistique de corpus. Il n'est pas satisfaisant que la détermination d'un lien entre tel phénomène linguistique et tel genre, dans une perspective de systématisation, relève de la seule intuition, pas assez fiable, mais elle ne peut se baser non plus sur le seul tâtonnement, complètement dépendant des corpus disponibles. Il faut espérer que le nombre et la variété des corpus disponibles ira en augmentant ce qui permettra au moins de disposer d'un fonds documentaire important, qui limitera les biais.

Hormis le lien entre genre du corpus et interprétation, une autre conclusion de ce travail sur *avec* (qui mériterait d'être approfondi) est que la consultation de données textuelles et la prise en compte du genre des corpus, donne un éclairage qui semble majeur sur le fonctionnement de telle ou telle préposition. Dans ce type de perspective, il ne s'agit pas d'étudier tous azimuts le fonctionnement sémantique de telle préposition mais plutôt, après un premier balayage intuitif, d'essayer d'associer tel marquage à tel genre de corpus. Evidemment, cette association est sans doute parfois impossible, peut-être aussi n'a-t-on pas toujours d'intuition fiable concernant ce type d'association. Mais lorsque le lien marqueur/genre du corpus apparaît comme manifeste, il permet de bien mieux comprendre et de décrire de manière dynamique le fonctionnement de tel ou tel élément sémantique. Cela permet en effet de cadrer l'étude et de mettre en lumière des éléments totalement ignorés par des descriptions générales et basées sur la seule introspection.

Si l'on essaye de généraliser les résultats obtenus pour *avec*, on peut dire que la possibilité pour un élément de marquer une relation fait intervenir 5 types d'éléments :

- Un élément linguistique, déclencheur (ici, *avec*),
- Des structures syntaxiques dans lesquelles apparaît cet élément (par exemple, [(dét)N1 avec (dét)N2]),

– Un contenu (une relation) (ici, la relation méronymique),
– Un genre de corpus (par exemple, catalogues de jouets),
– La fréquence d'apparition entre structure syntaxique et interprétation relationnelle.

Ces éléments se combinent de la façon suivante pour garantir le marquage :

> Si dans un corpus d'un genre identifié, une forme (apparaissant ou non dans une certaine structure syntaxique) est souvent associée à une interprétation relationnelle, alors, on dira que cette forme (ou cette structure) sert de marqueur pour le repérage de cette relation dans les corpus de ce genre.

Dans certains cas, la structure dans laquelle apparaît l'élément déclencheur permet de discriminer les phrases non pertinentes du point de vue du marquage (par exemple, [V avec détN] permet d'éliminer les cas où *avec* ne marque pas la méronymie) mais cette discrimination n'est parfois pas possible (ainsi [détN1 avec détN2] n'est pas suffisant pour discriminer les cas de méronymie dans les petites annonces). Seule alors une étude préalable sur un corpus considéré comme un échantillon permet de mettre au jour des tendances fortes. Ce type d'étude relève à part entière d'une linguistique de corpus. La notion de fréquence du lien structure/relation est alors primordiale.
Enfin, contrairement à ce que l'on pourrait penser, la définition proposée ci-dessus ne signifie pas que c'est la forme qui donne accès au contenu de la relation mais, plus exactement que cette forme permet le repérage de cette relation. Ce point est fondamental pour le fonctionnement de la préposition *chez*.

2.2.2.2 *Le cas de* chez

Chez apparaît dans certains énoncés où « s'exprime » une relation de méronymie. Par exemple :
Chez les colobinés, le nez fait saillie sur la lèvre supérieure.

Contrairement à *avec*, pour *chez*, une étude complète a été menée, qui est décrite dans (Condamines, 2000). Je ne donne ici que les principaux résultats.
Afin de suivre mon intuition selon laquelle l'interprétation méronymique des phrases avec *chez* est particulièrement liée au domaine des sciences naturelles, j'ai étudié une série d'exemples contenant *chez* en provenance de textes qui varient quant à leur genre. Le corpus d'étude contient ainsi trois séries de phrases :

– les unes extraites de l'Encyclopedia Universalis (387 occurrences de *chez*), ensemble noté EU1,
– d'autres extraites du journal Le Monde (100 occurrences de *chez*), ensemble noté LM,
– d'autres extraites du Cahier de notes de Claude Bernard de 1860 (127 occurrences de *chez*), ensemble noté CB.

Ces trois séries d'occurrences ont été sélectionnées dans des rubriques (dans un livre pour la troisième) relevant des sciences naturelles. Un quatrième exemplier a été utilisé comme élément de comparaison, il s'agit d'extraits de l'Encyclopaedia Universalis comportant *chez*, en lien avec le domaine des idées et de la création (78 occurrences), noté EU2. Au total, ce sont donc 692 exemples qui ont été étudiés.

Structure et interprétation des phrases en *chez*

Les phrases avec *chez* correspondent à trois structures (la position du syntagme avec *chez* n'est pas prise en compte car elle n'a pas de pertinence pour l'expression de la méronymie).
1- *Chez* (det1)N1, structure présentative det2N2

> *Chez les lémurs, il existe des zones glandulaires circumgénitales (*N1 *: Lémurs,* N2 *: zones glandulaires circumgénitales).*

2- *Chez* (det1)N1, det2N2 prédicat

Les callosités ischiatiques sont séparées chez les mâles comme chez les femelles (N1 *: mâles, femelles,* N2 *: callosités ischiatiques).*

3- *Chez* (det1)N1, (det3N3) prédicat det2N2

L'arrivée du printemps crée une sorte de fièvre chez les observateurs d'oiseaux
(N1 : *observateurs d'oiseaux*, N2 : *fièvre,* N3 : *arrivée du printemps*).

Il faut tenir compte du fait que *chez* n'est pas toujours associé à une interprétation méronymique mais parfois à une interprétation d'une autre nature : localisation concrète (spatiale), ou relation d'une autre nature, souvent difficile à identifier. Le décompte pour chacune des quatre séries d'exemples (EU1, LM, CB et EU2), pour chacune des trois structures et pour chacune des trois interprétations de la relation entre N1 et N2 permet d'aboutir aux conclusions suivantes.

Aucune structure syntaxique n'est particulièrement habilitée à renvoyer à une interprétation plutôt qu'une autre.

En revanche, la nature du corpus intervient dans l'interprétation de la relation entre N1 et N2 : les corpus didactiques de sciences naturelles (EU1 et CB) favorisent l'interprétation méronymique (plus de 50% des cas) alors que le corpus didactique sur un autre domaine (EU2) l'empêche complètement, ce qui confirme mon intuition première. Le corpus journalistique sur les sciences naturelles (LM) est plutôt favorable à l'interprétation d'une relation non méronymique entre N1 et N2 (environ 70 % des cas), ce qui prouve que la notion de didacticité joue un rôle majeur dans l'expression de la méronymie en lien avec chez ; les deux caractéristiques sont ainsi importantes : le corpus doit à la fois appartenir aux sciences naturelles et il doit être didactique.

On retrouve le même mode de fonctionnement qu'*avec* dans la mesure où il fait intervenir une forme, une interprétation et un genre de corpus. Toutefois, deux changements majeurs sont à noter.

D'une part, la structure dans laquelle apparaît *chez* ne semble pas pertinente pour repérer la relation puisqu'aucune des trois structures identifiées n'est discriminante (alors que pour *avec*, seules les structures de type [détN1 avec détN2] étaient pertinentes).

D'autre part, dans le cas de *avec*, on avait le sentiment que c'était cette préposition même qui permettait l'accès à l'interprétation relationnelle et donc qu'un rapport très stable s'établissait entre l'élément et l'interprétation. Il en va tout autrement avec *chez*. En effet, si l'on considère des exemples qui ont une structure parfaitement identique et qui sont issus du même genre de corpus (sciences naturelles et didactique), on constate que certains, majoritaires, ont une interprétation méronymique et d'autres non. Ainsi, il y a méronymie dans :

Chez la majorité des bivalves…, le caecum du stylet s'isole de la poche stomacale,

mais pas dans :

La ménarche chez les Cercopithecoidea et les Hominoidea est généralement suivie par une période de « stérilité d'adolescence »…

Dans ces deux exemples, seule varie la position du syntagme avec *chez*, dont j'ai montré qu'elle n'était pas pertinente (*cf.* l'article complet). Or, dans le premier exemple, il y a une relation de partie à tout entre *caecum du stylet* et *bivalves* alors qu'il n'y a pas cette relation entre *ménarche* et *cercopithecoidea*. L'interprétation est peut-être guidée par certains autres éléments des énoncés, que l'on peut caractériser sémantiquement : *poche stomacale* dans le premier cas, qui oriente vers une interprétation méronymique, *période de stérilité* dans le second, qui oriente vers une interprétation non-méronymique, mais dans certains exemples, rien ne permet de décider ou pas en faveur de l'interprétation méronymique :

Chez les cercopithécidés, le magot (Macaca sylvanus) occupe une place à part du fait de l'intensité des interactions entre les mâles adultes et les jeunes.
D'autres occurrences de ces noms, dans d'autres contextes, peuvent permettre d'aider à l'interprétation mais en aucune manière, on ne peut considérer que la présence de la préposition *chez* suffit pour discriminer les énoncés méronymiques des autres. En d'autres termes, alors que *chez* apparaît très souvent dans des contextes où l'interprétation de relation méronymique est possible, on ne peut pas dire que ce soit *chez* qui donne accès à cette interprétation mais un ensemble d'autres éléments, présents ou non dans le corpus entier[66].
On peut essayer de comprendre comment ce fonctionnement se met en place avec *chez.*
En fait, *chez* permet de mettre un élement en évidence, le reste de la phrase donnant une information sur cet élément. Ce qui reste vrai, c'est que, dans le contexte des sciences naturelles, cette information porte majoritairement sur l'anatomie, d'où une abondance de phrases qui mentionnent un élément « méronymique »[67]. On ne peut pas dire ainsi que *chez* « marque » la relation méronymique entre deux éléments ; en revanche, si l'on se place du point de vue de la recherche d'information, on peut dire que *chez*, dans un corpus de sciences naturelles, didactique, peut être utilisé comme marqueur de la possible présence d'une relation de méronymie : d'une certaine façon, il sert de marqueur (d'indice) de la possibilité de trouver une relation mais il n'en donne pas l'interprétation. Dans une perspective de recherche d'information, il est intéressant de se fonder sur cet élément puisque, si le corpus est didactique dans le domaine des sciences naturelles, dans un cas sur deux la relation sera méronymique.

On pourrait penser que ce fonctionnement est propre à *chez*. Or, ce n'est pas le cas. J'ai mentionné pour *avec* des cas où [détN1 avec détN2] « n'exprimait pas » une méronymie par exemple, *T2 avec vue sur les Pyrénées*.
On pourrait trouver bien d'autres exemples, comme ces phrases avec *comme* extraites du même Atlas scolaire :
Un département comme la Seine bénéficie à la fois d'arrivées d'enfants et de scolaires.
On comprend que les lycées professionnels et d'enseignements général, comme l'université, soient très peu tournés vers les formations scientifiques et technologiques.
Dans le premier cas, il existe une relation d'hyperonymie entre *La Seine* et *département* mais ce n'est pas cette relation qui existe dans le second exemple. Les structures étant identiques, elles ne peuvent être utilisées pour discriminer les deux sens. Pour un corpus sur un domaine inconnu, le recours à ce seul marqueur ne permet pas de décider pour l'un ou l'autre sens.

Dans certains cas toutefois, on peut arriver à trouver d'autres éléments de forme qui contraignent l'interprétation. Le repérage des déverbaux peut ainsi permettre de sélectionner les exemples sans relation partie à tout, aussi bien dans le cas de *chez* que dans celui d'*avec* :
*Chez les embryons , l'*excitation *des nerfs donne lieu à des convulsions dans les muscles.*
Séjour avec accès *direct à la piscine.*
Mais il reste des phrases qui échappent à toute possibilité de contrôle par la seule forme :
La frugivorie est le régime le plus répandu chez les primates.
Résidence avec gardien.

[66] Lorsque ces éléments ne sont pas présents dans le corpus complet, on peut dire qu'ils proviennent d'une connaissance antérieure, « présupposée », c'est-à-dire en fait, rencontrée dans d'autres discours, par exemple, lors de l'apprentissage de la langue.
[67] Il faut noter d'ailleurs que les cas d'interprétation méronymique apparaissent majoritairement au début des articles de l'Encyclopaedia : c'est en effet, par la description anatomique que commencent généralement ces articles, qui se poursuivent par la description du mode de vie : alimentation, habitat, reproduction…

Comme je l'ai déjà signalé, seule une analyse sémantique approfondie qui s'appuie à la fois sur la compétence de locuteur, sur les autres occurences des termes à analyser (souvent dans d'autres textes que celui à l'étude d'ailleurs), sur l'objectif de la modélisation, voire, sur le recours à un expert, peut permettre d'interpréter certains exemples.

D'un point de vue informatique, donc strictement formel, on ne peut pas arriver à discriminer les cas qui, bien qu'apparaissant dans une structure particulière peuvent avoir au moins deux interprétations possibles[68]. Dans un souci d'efficacité, on peut cependant constater, après analyse linguistique approfondie d'un corpus échantillon, qu'une des interprétations est largement favorisée et donc que le bruit (produit par un outil) à gérer par l'interprétant ne sera pas trop important. Par exemple, il n'est certainement pas pertinent de chercher à repérer la méronymie dans un corpus littéraire (au cas où cela pourrait avoir un intérêt mais qui sait ?) en recherchant les phrases contenant *avec* ; en revanche, cela a une pertinence si on utilise un corpus de genre « catalogue ».

2.3. Quand le fonctionnement des marqueurs se complexifie

2.3.1 Marqueurs lexico-syntaxiques

Jusqu'à présent, les seuls marqueurs qui ont été examinés étaient des éléments lexicaux qui, dans certains cas, jouent un rôle central, dans une structure plus complexe (comme *avec* dans la structure [(dét)N1 avec (dét)N2]).
Certains marqueurs font intervenir des positions argumentales ce qui a pour conséquence de compliquer leur repérage et de faire éclater la notion de binarité des relations.
Nous avons rencontré ce problème lors du projet SGGD (Système de Gestion Globale des Déplacements) qui visait à faire collaborer différents organismes sur une même tâche : gérer la circulation dans l'agglomération toulousaine. Conscients des problèmes langagiers, et en particulier terminologiques qu'ils pouvaient rencontrer, ces organismes nous ont sollicités pour que nous construisions un référentiel terminologique.
Il nous est apparu que beaucoup des informations véhiculées dans les corpus de chacun de ces partenaires concernaient le problème de la communication. D'ailleurs, nous avons trouvé de nombreux énoncés où on pouvait identifier des relations binaires sur ce thème :
La ville de Toulouse et les ASF sont les interlocuteurs principaux de la DDE.
Les appels sont transférés via des liaisons spécialisées.
Le retour en C1 fera l'objet d'une information en temps réel par télécopieur.
Dans le premier cas, on peut décider que l'on a une relation :

`X-communique-avec-Y`

DDE `communique-avec` ville de Toulouse
DDE `communique-avec` ASF
Dans les deuxième et troisième cas, on peut décider que l'on a une relation :

X-`médiatise`-Y

Liaisons spécialisées `médiatise` appels
Télécopieur `médiatise` retour en C1.
La difficulté vient de ce que dans les deux cas, le « marqueur » est un prédicat par rapport auquel s'organisent les arguments entre lesquels s'instaure une relation particulière. Dans le deuxième exemple en tout cas, il n'est plus possible de parler seulement de position par rapport à une structure lexicale comme cela a été fait jusqu'à présent, on est obligé de faire

[68] Evidemment, dans un domaine que l'on connaît, on peut avoir l'impression qu'il suffit d'ajouter une caractérisation sémantique pour filtrer les exemples pertinents mais dans le cas de corpus spécialisés, cette caractérisation sémantique *a priori* n'a pas de sens puisque l'on cherche justement à la faire émerger par l'utilisation de marqueurs non spécifiques.

intervenir la notion d'argument et aussi, ce qui est plus délicat, la notion de classe sémantique de verbe.
Pour la seconde phrase par exemple, le marqueur pourrait être
M est un marqueur de la relation `X-médiatise-Y` si
M un verbe de la classe « transfert d'information »,

- à la voie active, Y est sujet du verbe (placé à gauche) et X COD (placé à droite)
- à la voie passive, X est sujet du verbe et Y est introduit par *via* ou *par*.

Ce genre de « marqueur » n'est pas inconnu des informaticiens. Dans les années 80-90, on a ainsi cherché à construire des systèmes de dialogues Homme-Machine sur des domaines restreints, par exemple, à Nancy sur les informations administratives des pages roses de l'annuaire. Avec fortement présente l'idée des sous-langages, les chercheurs ont essayé de reconstituer la grammaire de l'expression de ces informations administratives (voir par exemple : (Deville,1989)). Cette grammaire accorde la prépondérance au prédicat (comme dans les grammaires sémantiques à la Fillmore) qui joue le rôle pivot autour duquel s'organisent les arguments (au sens large) qui portent les informations principales et sur lesquels aussi portent les questions des usagers (par exemple, *où puis-je faire renouveler ma carte d'identité ?*). L'objectif n'était pas alors la constitution d'un réseau relationnel mais la méthode pour repérer automatiquement les informations était en réalité très proche de celle que l'on utilise maintenant pour repérer des relations dites conceptuelles : les éléments mis en relation sont en fait des éléments occupant une place dans une structure prédéfinie appelée marqueur.
Autre exemple de travaux informatiques mettant en œuvre la structure argumentale : les travaux en extraction d'information qui consistent à retrouver des informations dans un certain type de textes, prédéfini (par exemple les dépêches d'agence). Dans ce type de travaux, on sait quels types d'informations sont recherchées, ils sont modélisés dans un « formulaire » (par exemple, qui a fait un attentat, où et pourquoi) et la réflexion porte sur les formes que peut prendre cette information (voir par exemple (Gaizauskas et Wilks, 1998). Une analyse complète du texte n'est la plupart du temps pas utile : il suffit de trouver des structures qui permettent de repérer ces informations. Ici encore, ces structures, hormis le fait qu'elles n'ont pas pour but de repérer des relations, sont très proches de ce que l'on peut appeler des marqueurs de relations (du point de vue du traitement automatique) : il s'agit de formes qui permettent d'accéder à un contenu : Riloff parle d'« extraction patterns ». Dans cette perspective d'ailleurs, le lexique est considéré comme une forme de surface permettant d'accéder à un contenu (Riloff, 1996) et (Basili et Pazienza, 1997) :

> « Another valid and slightly more comprehensive definition describes the lexicon as an association between surface forms and linguistic information » (Basili et Pazienza, 1997, 44).

La recherche de formes en lien avec un contenu est ainsi fréquente en informatique, pour retrouver différents types d'information « discursive ». Et des langages sont possibles pour représenter cette information, par exemple, les graphes conceptuels.
En revanche, la mise en œuvre de formes élaborées, qui prennent en compte la notion d'arguments, pour repérer des relations conceptuelles n'est, à ma connaissance, jamais faite, pas plus en ingénierie des connaissances qu'en terminologie.
Cela tient sans doute au fait que ce mode de fonctionnement ouvre des perspectives qui remettent en cause la notion de binarité des relations. En effet, les structures argumentales font souvent intervenir plus de deux arguments et, dans ces cas-là, la représentation relationnelle sous la forme binaire n'est pas pertinente. Soit l'exemple extrait du même corpus SGGD :
Chaque subdivision transmet une fiche aux chantiers qui se terminent.

Cette phrase fait intervenir trois arguments, également dépendants les uns des autres. On peut représenter les dépendances deux à deux, sous la forme de relations binaires :

X-émet-Y

Subdivision émet fiche relative aux chantiers qui se terminent

X-envoie-à-Y

Subdivision envoie au CIGT.

Mais on perd alors une partie importante de l'information.

Ces dépendances « concomitantes » peuvent s'exprimer sous la forme d'un schéma de la communication, qui semble saturer le nombre possible d'arguments pour le corpus[69]:

X-envoie-Y-à-Z-via-W,

Dans lequel le verbe joue ce rôle de pivot. On trouve des exemples qui correspondent à tout ou partie de cette structure :

L'opérateur est celui qui est chargé de diffuser auprès des usagers l'information routière.

La fonction de guidage des usagers se fait au travers des messages diffusés sur les PMV.

Comme on le voit, les modes discursifs que peuvent prendre la mise en œuvre de ce schéma peuvent être très complexes, certains étant certainement imprédictibles.

Ce mode de fonctionnement des marqueurs est beaucoup plus proche du fonctionnement discursif et leur mise au jour relève pratiquement de l'établissement d'une grammaire de discours, évidemment plus longue que la seule recherche de marqueurs binaires, prédéfinis. Cette constatation met en évidence que la modélisation à partir d'un corpus, quelle que soit la forme qu'elle prend (Bases de Connaissances Terminologiques ou grammaire), relève, pour le linguiste, d'une étude approfondie qui, certes, s'appuie sur des éléments de forme mais pour mieux élaborer un contenu, c'est-à-dire donner une forme nouvelle qui prend sens en fonction d'un objectif[70].

2.3.2 Marqueurs et interprétation

A plusieurs occasions, j'ai évoqué la prise en compte de l'objectif pour justifier des choix : choix du mode de représentation, choix des relations retenues, choix des termes mis en relation, choix des contextes pertinents… Je terminerai ce chapitre en insistant sur deux éléments concernant directement le rôle de l'objectif d'interprétation. Le premier me semble montrer le bien-fondé de ma position. Il s'agit d'un cas où l'objectif d'application, au sens large, s'inscrit dans le fonctionnement textuel même. Le second consiste en une réflexion sur les notions de genre textuel et de genres interprétatifs.

2.3.2.1 « L'application » s'inscrit dans les fonctionnements textuels

Il me semble avoir identifié un cas où « l'application » (que l'on peut considérer comme relevant d'un genre interprétatif) a une influence directe sur la forme du marqueur. Plus précisément, pour une même relation, plusieurs marqueurs existent qui ne donnent pas le même type de résultats, cette variation pouvant être liée au type de représentation que l'on souhaite faire.

---

[69] Les notions de temps et d'espace n'ont pas été considérées dans ce cas, probablement à tort car il aurait fallu rendre compte de phrase comme *Cette information [practicabilité du réseau] est communiquée du lundi au samedi.*

[70] Ainsi, la collaboration que nous menons avec le CENA concernant l'étude des modes d'expression du dysfonctionnement dans un dialogue entre deux populations d'experts conduit à une réflexion linguistique qui n'est pas fondamentalement différente de celle qui permet de construire une BCT (*cf.* thèse en cours de P. Vergely (Vergely, 2002)).

Il s'agit d'un phénomène décrit par différents auteurs (par exemple (Milner, 1976) ou (Lerat, 1981)) mais rarement comme possible marqueur de l'hyperonymie. Le phénomène que je vais décrire est le suivant : présence dans une phrase d'un SN qui reprend une partie d'un énoncé précédent. Dans certains cas, il y a une relation d'hyperonymie entre le SN du second énoncé et un SN du premier. Par exemple, dans le corpus d'EDF (MOUGLIS) :
Préparation des activités de liaison.
Ce processus débute par la fourniture d'une copie de l'Etat de configuration du logiciel.
Dans cet exemple, *processus* est un hyperonyme de *préparation des activités de liaison*.
Afin d'étudier plus en profondeur la possibilité d'un marquage de l'hyperonymie, j'ai essayé de la mettre en œuvre automatiquement, ce qui n'est pas facile car elle fait intervenir l'absence (le second SN ne doit pas apparaître dans un énoncé précédent, sinon, il est sûr qu'on n'a pas à faire à une hyperonymie).
Ce marqueur, complexe pourrait être :
[(défini ou démonstratif) + N] à la condition que :

- Ce N n'apparaît pas dans le paragraphe précédent,

Cette restriction sert à éliminer les exemples comme :
Espace de réception. Cet espace est un espace de transition destiné aux fournitures externes.
En effet, la reprise par *espace*, la tête de *espace de réception*, relève d'un fonctionnement discursif, toujours possible sauf en cas de syntagme très figé (par exemple, on ne peut reprendre *pomme de terre* par *pomme*). Cela revient à considérer que la tête d'un SN constitue l'hyperonyme de ce syntagme.

- Le déterminant n'a pas un rôle déictique.
- Le déterminant a un rôle anaphorique.

Parmi les déterminants potentiellement anaphoriques (définis et démonstratifs), seul le cas des démonstratifs qui ne peuvent être que déictiques ou anaphoriques a été considéré, la valeur des définis étant beaucoup trop difficile à identifier sur des bases formelles.
Enfin, afin d'éliminer les démonstratifs déictiques, j'ai supprimé de l'étude les cas où [démonstratif N] renvoie à coup sûr à un fonctionnement déictique : *cette partie, ce chapitre, cet ouvrage …*
J'ai également supprimé les expressions plus ou moins figées (bien qu'anaphoriques) comme *dans ce but, à ce moment…*[71]

Deux types de résultats ont été obtenus à partir de l'application du marqueur sur le corpus EDF (MOUGLIS).

Marquage de l'hyperonymie ?

Le premier type de résultats concerne les interprétations possibles pour la relation entre [démonstratif N] et le contexte précédent (relation anaphorique).
Sur les 78 cas identifiés, la répartition se fait de la manière suivante :

- 44 concernent une relation hyperonymique :

par exemple : *Archivage de l'Etat de Configuration logiciel. Cette activité est à la charge du responsable de la gestion de configuration.* (*Activité* est un hyperonyme de *Archivage de l'état de configuration*)

- 27 visent à reprendre un prédicat sous une forme nominale :

Ce chapitre s'attache à décrire *les différents processus de gestion de configuration. Cette* description *ne présume pas de la méthodologie de développement utilisée*.

- Dans 4 cas, le SN reprend tout un énoncé :

[71] La mise en œuvre informatique de cette étude a été assurée par L. Tanguy, enseignant-chercheur à l'ERSS que je remercie pour son aide.

Il se présente sous la forme d'un arbre dont les feuilles correspondent à des produits ou composants à réaliser. Cette décomposition, qui suit l'arborescence des produits définie au paragraphe 4 .1.

– 2 correspondent à une relation synonymique :

Le produit développé par l'équipe projet peut s'appuyer sur un noyau logiciel *développé par une autre équipe projet. Cette dernière livre alors toutes les Unités de Configuration permettant à l'équipe projet de modifier cette* souche logiciel *en fonction des nouvelles fonctionnalités à implémenter.*

– 1 consiste en la répétition d'un SN sous la forme d'un acronyme :

Le Plan de Développement *standard définit les mécanismes de vérification et de validation de ces produits… Liens de traçabilité entre ce* PDL *standard et les propositions du MAQL-ST DER.*

La plupart des cas rencontrés dans ce corpus sont décrits dans (Lerat, 1981). Mais ce qui est étonnant est la fréquence de la reprise par un hyperonyme : plus de 56 % des exemples. On peut donc penser que cette structure, particulièrement élaborée, peut être considérée comme un marqueur d'hyperonymie. Il faudrait toutefois vérifier cette possibilité sur un plus grand nombre d'exemples et parvenir à circonscrire un genre de corpus pertinent pour ce marquage. On ne sait pas par exemple si le genre « manuel technique » est particulièrement lié à la possibilité de ce marquage ou bien si tous les corpus sont concernés.

Remarquons que nous retrouvons le même type de fonctionnement qu'avec *chez* : ce n'est pas la structure qui indique la relation ; cette relation est, en quelque sorte, présupposée (ou explicitée ailleurs dans le texte ou dans un autre texte). Dans (Péry-Woodley, 2000), Péry-Woodley note le même type de constat chez Mann et Thompson, à propos des relations discursives:

> « W.Mann et S.Thompson posent la question du rôle interprétatif de l'analyste dans l'élaboration d'une représentation de la structure rhétorique d'un texte. […] Cet aspect interprétatif est d'autant plus important que les auteurs rejettent le recours à des marqueurs pour l'identification des relations. » (Péry-Woodley, 2000, 43).

Quelle hyperonymie ?

Hormis la possibilité de marquage de l'hyperonymie par cette structure, un autre résultat est particulièrement intéressant du point de vue de la prise en compte de l'objectif applicatif. Il concerne la nature des éléments mis en relation.

Première observation : alors que l'on considère en général que les termes sont le plus souvent des combinaisons, les hyperonymes trouvés avec ce marqueur sont le plus souvent des termes simples. Ainsi, sur les 44 hyperonymes repérés, 12 seulement sont des polytermes. Cette faible présence des polytermes trouvés grâce à ce marqueur est confirmée par l'application de ce marqueur sur un autre corpus : un livre sur l'ingénierie des connaissances. Cette fois-ci, ce sont seulement 74 des 376 hyperonymes qui sont constitués d'un syntagme composé.

Deuxième observation : les termes mis en relation d'hyperonymie par cette structure ne sont pas les mêmes que ceux que l'on trouve avec un marqueur plus classique de l'hyperonymie, celui que l'on a dans les structures définitoires comme [dét N1 être dét N2 + différences]. En effet, aucun des termes mis au jour par le marqueur anaphorique ne se trouve ni en position de *definiens* (N1) ni en position de genre (N2). C'est-à-dire que non seulement ils ne sont pas définis dans le corpus mais ils ne sont pas non plus utilisés pour définir d'autres noms. En revanche, 8 des monotermes trouvés grâce à la structure avec démonstratif correspondent à des têtes de polytermes trouvés soit en position de *definiens*, soit en position de genres dans les structures définitoires :

– *Acteur* (tête d'un terme),

- *Activité* (tête de deux termes),
- Composant (tête de deux termes),
- Décomposition (tête de deux termes),
- Espace (tête de trois termes),
- Phase (tête de deux termes),
- Processus (tête de deux termes),
- Revue (tête d'un terme).

Par exemple, *phase* est trouvé comme hyperonyme dans la structure avec démonstratif dans le contexte suivant :
Description générale de l'activité. Cette phase recouvre trois activités principales.
En revanche, avec la structure définitoire classique, on retrouve seulement des syntagmes composés avec *phase* mais jamais *phase* tout seul.
La phase de conception est une activité de conception du produit logiciel sur la base d'un ensemble de spécifications fonctionnelles et non-fonctionnelles.
Ainsi, plusieurs éléments montrent que les termes, et en particulier les hyperonymes, mis au jour grâce à la structure avec démonstratif sont différents des hyperonymes découverts avec des marqueurs plus habituels.

Du point de vue linguistique, ces hyperonymes ont peut-être à voir avec les termes de base, décrits par la psychologie cognitive (sémantique du prototype). Les termes de base sont ceux qui, dans une catégorisation en trois niveaux (niveau superordonné (par exemple *animal*), niveau de base (par exemple *chien*), niveau subordonné (par exemple *doberman*)) portent la plus grande saillance cognitive (Kleiber, 1994). Selon le courant auquel ils appartiennent, les sémanticiens sont partagés sur la pertinence de ces termes de base. Ainsi, là où Rastier pense que cette idée n'a pas d'intérêt pour l'étude du lexique (Rastier, 1991), Kleiber soutient la thèse inverse en proposant des tests pour mettre en évidence linguistiquement l'existence de ces termes. Par exemple, il pense que c'est par un terme de base que l'on répond prioritairement à la question : *qu'est-ce que c'est ça* ? C'est-à-dire que pour lui (et pour les psychologues cognitivistes), les termes de base font partie d'un fonctionnement cognitif mais qui se manifeste linguistiquement. Une fois de plus, il me semble que l'étude des corpus apporte un éclairage nouveau sur cette question. Il ne s'agit pas de projeter une connaissance supposée préexistante car psychologique mais de montrer qu'une connaissance peut être mise au jour à partir d'une analyse de corpus. En effet, les termes trouvés par la structure avec démonstratif ressemblent à ceux qui sont décrits comme termes de base au sens où, d'après leur fonctionnement discursif, on peut constater qu'ils se comportent différemment des termes spécifiques du corpus (puisqu'ils ne sont pas modifiés et pas non plus utilisés pour définir comme le sont les termes spécifiques) bien qu'ils soient utilisés comme hyperonymes à l'intérieur du corpus (puisqu'ils sont mis au jour grâce à un marqueur d'hyperonymie). Ainsi, tout en étant des termes du corpus, ils paraissent plus généraux, plus hauts dans la hiérarchie que les autres termes du corpus, qui seraient, eux, des termes subordonnés.

Du point de vue de la construction de représentations relationnelles, ces termes pourraient être particulièrement utiles pour rassembler des fragments de taxinomies sous un même « père ». En effet, une situation courante est celle où l'application de marqueurs classiques d'hyperonymie permet de constituer des fragments de taxinomies isolées qu'on a parfois du mal à rapprocher. Si d'autres marqueurs, comme cette structure avec démonstratif, permettaient de repérer quelques racines de taxinomies, cela représenterait une aide importante.

C'est ici aussi qu'intervient la notion d'application. En effet, la profondeur des taxinomies à construire peut dépendre de l'application. Si l'on veut comparer dans le détail le contenu de deux corpus, il est important d'aller aussi profond que possible dans la construction des hiérarchies. En revanche, si l'on veut savoir le thème principal d'un texte (par exemple pour la recherche d'information) ou bien si l'on veut indexer un texte, on a plutôt besoin de n'avoir accès qu'au haut de la hiérarchie et l'utilisation des seules structures avec démonstratif peut être plus efficace dans un premier temps.

La diversité des marqueurs pour une même relation est connue et a pu conduire à organiser les relations en sous-relations par exemple la cause (Cabré et *al.*, 1997), (Garcia, 1998) ou, plus classiquement, la méronymie ; par exemple, la méronymie temporelle peut être marquée par des patrons spécifiques qui structurent la durée (par exemple [X précède Y] (Condamines et Rebeyrolle, 1997). Mais dans ce cas précis, le marqueur concerne l'hyperonymie, généralement considérée comme permettant la principale structuration, autour de laquelle s'organisent les autres relations. Or, c'est justement la finesse du maillage hyperonymique qui permet de distinguer des modélisations de haut niveau (thésaurus ou index) de celles qui sont plus près du contenu discursif. On voit ainsi se dessiner une continuité entre problématique appliquée (construction d'une modélisation relationnelle (thésaurus, index, BCT…)), problématisation linguistique et étude des manifestations linguistiques. Dans l'étude réalisée, le parcours s'est fait dans le sens contraire : étude d'un phénomène en corpus, problématisation linguistique, établissement d'un lien avec une problématique appliquée ; ce qui prouve qu'il n'y a pas un hiatus entre linguistique théorique et linguistique appliquée mais bien plutôt prise en compte du réel dans l'interprétation des données corporales.

2.3.2.2 Genres textuels et genres interprétatifs

Dans une perspective de sémantique textuelle, la nécessaire vision variationiste prend corps dans la notion de genre textuel : ensemble d'éléments extra-linguistiques qui s'élaborent conjointement avec des régularités linguistiques pour constituer une sorte de norme qui vient se superposer à la norme linguistique (Bakhtine, 1984). Mais pour rendre compte au plus juste de la construction du sens, il est certainement nécessaire d'envisager des genres interprétatifs qui pourraient permettre de systématiser les interprétations. La difficulté vient de ce que genres textuels et genres interprétatifs ne coïncident pas toujours. Reprenons l'exemple de *avec* et les trois genres de corpus où cette préposition apparaissait comme « marqueur » de méronymie : petites annonces immobilières, catalogues de jouets, description d'itinéraires. Dans les deux premiers cas, le genre « catalogue » ou « petites annonces » n'est pas suffisant pour rendre compte de la possibilité de « marquer » la méronymie par *avec* ; ce genre doit être plus finement caractérisé. En effet, les petites annonces de voitures n'utilisent pas ce marquage. De la même façon, (mais je n'ai consulté que quelques dizaines de descriptifs), il semble que les catalogues de vêtements utilisent peu le marquage par *avec* (pas impossible cependant comme dans *: sac cabas avec anses contrastantes* (La Redoute printemps-été 2002) et préfèrent soit pas de marquage du tout (*le cardigan, manches longues, emmanchures raglan diminuées* (*ibid*.)) soit un marquage par *à* (*Débardeur à bretelles « spaguetti »*, (*ibid*.)) Ce marquage par *à*, sans être impossible, est rare dans les catalogues de jouets (*voitures à pédales*).
Ainsi, l'objectif de l'interprétation fait que ce qui apparaissait comme genre textuel stable peut éclater. Par exemple, pour ce qui est du marquage de la méronymie par *avec*, une description d'itinéraires a un fonctionnement plus proche des petites annonces immobilières que les petites annonces immobilières n'en ont avec les petites annonces de vente de voitures. C'est au même type de conclusion que parvient Biber lorsqu'il constate que, selon la dimension considérée, il y a plus de similarité de fonctionnements linguistiques entre X et Y

qu'entre Y et Z alors que Y et Z paraissaient intuitivement proches ; par exemple, du point de vue de la dimension narrative, les dialogues spontanés sont plus proches des biographies que des conversations téléphoniques (Biber, 1988, 136). Dans le cas du marquage méronymique, le mode de fonctionnement linguistique qui est considéré demande une analyse interprétative plus sophistiquée, qui ne peut être réalisée automatiquement mais qui pourrait peut-être être considérée comme une dimension, selon la terminologie de Biber.

La non-coïncidence entre genres textuels intuitifs et régularités linguistiques que l'on repère en fonction du point de vue que l'on adopte est une difficulté majeure pour la mise en place d'analyses sémantiques de corpus. Plusieurs types d'études devraient se développer pour pallier cette difficulté.

- Tout d'abord, il est peut-être encore possible de se lancer dans une caractérisation des genres textuels : il se peut que la mise à disposition de textes sur internet, en permettant l'accès à une grande diversité de genres texuels, autorise un classement sur un volume de textes très important. En effet, les tentatives de définition de genres ont été réalisées jusqu'à maintenant de manière introspective : on essayait d'imaginer toutes les situations possibles d'énonciation collective et de les systématiser. Tout comme pour les études concernant des phénomènes locaux, l'analyse de productions réelles pourrait pallier les manques de cette approche uniquement introspective. On pourrait commencer à travailler sur les genres qui sont nommés dans la langue. Mais il est probable aussi qu'on se heurtera à d'autres problèmes : d'une part, tous les genres textuels ne sont pas représentés sur internet et, d'autre part, des genres textuels apparaissent, qui sont propres à internet, par exemple les pages de présentation commerciales des entreprises, qui ne sont pas toujours la copie de la version papier ou encore les pages de présentation personnelle.
- Indépendamment d'une éventuelle définition de genres mais dans cette perspective, il paraît absolument nécessaire de dresser un panorama des études menées sur corpus. Cela supposerait de mettre en place un observatoire qui recenserait les caractéristiques des corpus utilisés et des analyses menées. Il faudrait alors que chaque corpus d'étude soit caractérisé le plus précisément possible ; pour chacun des textes collectés : caractérisation du support, compétence de (des) écrivain(s), compétence et attente des interlocuteurs, objectif de la rédaction ; pour le corpus : hypothèse poursuivie, études réalisées… Il est probablement impossible de penser à toutes les caractérisations qui pourraient être pertinentes en fonction du point de vue que l'on adopte (ou, surtout, que l'on pourrait adopter). Ainsi, pour l'exemple des petites annonces, on pourrait tenir compte de la nature de l'objet à vendre ou en tout cas à vanter, du support de l'annonce (quotidien, hebdomadaire, web…) de la compétence du locuteur et de celle de l'interlocuteur (simple acheteur, agence immobilière, garagiste…), de l'espace attribuée pour chaque annonce (qui peut, comme on l'a vu, avoir un rôle sur l'utilisation de mots courts), peut-être aussi du lieu de l'annonce (ville/campagne, régions …), de la période de l'annonce… Et sans doute de beaucoup d'autres éléments encore, probablement à adapter à chaque étude. Mais une fois encore la réflexion à partir d'études réellement menées, sur des corpus réels serait bien plus riche d'enseignements qu'une réflexion introspective sur des études éventuelles.
- Un autre type d'études, menées en parallèle avec les travaux sur le genre textuel, pourrait concerner les types d'interprétation. Il me semble que la réflexion sur ce

sujet est peu développée. Or, plusieurs éléments encouragent à penser qu'elle serait nécessaire :

- Si un corpus est construit pour étudier tel ou tel phénomène, c'est que l'objectif de l'étude prévaut à la constitution du corpus ; il est donc pertinent de s'interroger sur la nature des objectifs d'étude et sur la possibilité de les élaborer en classes (j'ai donné quelques pistes en ce sens dans le premier chapitre).
- L'interprétation, au même titre que la rédaction d'un texte, consiste en la construction d'un sens ; elles relèvent donc toutes les deux d'une activité sémantique. Il ne semble donc pas sans pertinence d'étendre la notion de genre à la situation d'interprétation. Il est probable que ces genres interprétatifs seraient aussi stables (ou peu stables) que les genres textuels. Comme pour les genres textuels, on aura des genres interprétatifs évidents (la construction de BCT, par exemple) et d'autres qui n'apparaîtront qu'au cours du développement d'une pensée individuelle ou collective.

3. Conclusion

La prise en compte de corpus dans l'étude des marqueurs de relation ouvre des perspectives sémantiques nouvelles, à la condition de poser des bases de réflexion plus en conformité avec la réalité des fonctionnements. Ces fonctionnements peuvent être décrits en plusieurs points.

– La représentation relationnelle ne permet pas de rendre compte de tout le sens d'un corpus.

D'une part parce que l'interprétation consiste en la construction d'un sens en fonction d'un point de vue, d'autre part parce que la représentation en réseau contraint l'interprétation.

– Discursivité des marqueurs.

Les marqueurs sont fondamentalement des signes linguistiques comme les autres qui s'intègrent donc dans le discours comme n'importe quel signe linguistique. Ils n'ont ainsi que le fonctionnement métalinguistique qu'on leur attribue en fonction d'un objectif particulier. Une structure linguistique acquiert ainsi le statut de marqueur lorsqu'on a constaté qu'elle était fréquemment associée à des éléments que l'on peut représenter sous une même forme relationnelle. Fréquemment, c'est-à-dire soit dans un corpus, soit dans un genre de corpus, soit indépendamment du corpus. Dans la majorité des cas, les marqueurs sont moins spécialisés que les termes qu'ils réunissent mais il se peut que cette caractéristique soit due à la compétence des interprétants, souvent moins experts que les rédacteurs des textes constituant le corpus.

Si on a pu concevoir les marqueurs comme des indices, particulièrement en TAL, c'est seulement dans des cas où la dépendance avec le corpus est faible mais ces cas sont rares et il n'en reste pas moins que ces structures ne peuvent être complètement désémantisées ; elles peuvent ainsi subir des variations contextuelles.

– Les marqueurs ne relèvent pas du seul fonctionnement lexical.

Il n'y a pas de justification à ne s'intéresser qu'aux seuls marqueurs lexicaux. Dans une telle perspective, n'importe quelle structure typologique, morphologique, syntaxique ou discursive peut jouer le rôle de marqueur[72]. Si l'on s'intéresse aux fonctionnements sémantico-syntaxiques, il est nécessaire de pouvoir représenter des structures n-aires, qui rendent compte de structures argumentales entretenant des relations de dépendance et pas seulement des

[72] Péry-Woodley arrive aux mêmes conclusions à propos des marqueurs de relations discursives (Péry-Woodley, 2000).

relations binaires. La constitution de Bases de Connaissances Terminologiques requiert ainsi toujours une analyse sémantique approfondie.

– Les marqueurs ne donnent pas l'interprétation de la relation.

Ce que marque les marqueurs, c'est la possibilité d'interpréter telle portion d'un corpus sous la forme d'une relation binaire ou n-aire. Ils ne donnent pas eux-mêmes le mode d'interprétation. C'est seulement dans les cas, rares, où l'interprétation s'est systématisée (à la suite d'une utilisation très fréquente) qu'on a pu penser possible d'associer directement une forme à un contenu. Dans la plupart des cas, cette association n'est possible que par une interprétation humaine. Cela signifie aussi qu'en dehors de ces relations et de ces marqueurs, il est difficile de concevoir des outils d'aide au repérage des relations dans un corpus.

– Variation des marqueurs en fonction du genre textuel.

La possibilité pour telle ou telle structure de marquer telle ou telle relation peut être due au genre du corpus. Il est donc possible d'envisager une linguistique de corpus qui s'attache non seulement à décrire les structures au plus juste de leur pertinence pour permettre de repérer une relation mais aussi d'associer cette possibilité à des genres de corpus. Le marquage d'une relation en particulier n'est pas alors exclusif d'une autre relation mais ce marquage est majoritaire dans le corpus étudié. Il s'agit donc de construire un corpus comme un échantillon représentatif d'un genre textuel et de dégager des régularités de fonctionnement. Lorsqu'une interprétation relationnelle est fréquemment associée à une structure linguistique, on peut considérer qu'il est pertinent, si l'on souhaite construire un réseau relationnel, d'utiliser cette structure dans les corpus de ce genre pour repérer cette relation, sans que cela génère trop de bruit.

Dans la perspective de la systématisation des résultats, c'est la notion de genre qui semble la plus prometteuse, à condition qu'on la décline d'une part en genres textuels et d'autre part en genres interprétatifs. La réflexion sur les genres textuels est déjà engagée depuis longtemps mais pourrait être re-dynamisée par les possibilités offertes par internet. La réflexion sur les genres interprétatifs me semble, quant à elle, nécessaire.

Conclusion

La constitution de bases de connaissances terminologiques à partir de corpus m'a servi de fil conducteur pour aborder un ensemble de questions qui apparaissent lorsqu'on met en œuvre un corpus pour mener une analyse sémantique. En plusieurs occasions, j'ai souligné que cette tâche de construction de BCT, qui pouvait paraître très spécifique, n'est en fait que la version grossie de ce qu'est toujours une analyse sémantique de corpus : une interprétation qui, sans être totalement libre, peut être très diverse en fonction de la nature des textes du corpus étudié et de l'objectif de l'analyse. Pour le linguiste, une des façons de comprendre, voire de généraliser les interprétations possibles pourrait consister à travailler la notion de genre, qui, de mon point de vue, se décline en genre textuel, pour caractériser la situation de production des textes et en genre interprétatif, pour caractériser la situation d'interprétation. J'ai donné des exemples de régularités linguistiques qui ne sont observables qu'en prenant en compte l'un ou l'autre de ces genres, voire les deux.

La constitution de BCT à partir de corpus a cependant un caractère très particulier. Il s'agit en effet d'élaborer une représentation relationnelle qui privilégie les formes lexicales : noms pour les nœuds et prédicats pour les relations. Concrètement, il s'agit de passer d'une forme linéaire et syntaxique à une structure spatiale et quasiment a-syntaxique. Ce mode de représentation, particulièrement adapté à l'objectif informatique, ne va pas de soi ; il ne permet de rendre compte que d'un point de vue sur le corpus et peut constituer une contrainte pour le linguiste qui doit faire des choix qui sont parfois dépendants de cette représentation ; quelques-uns des problèmes que ce mode de structuration peut soulever ont été présentés.

Enfin, la constitution de BCT rend incontournable la rencontre entre sémantique de corpus et informatique, parce que l'informatique met à disposition des corpus électroniques et des outils pour les traiter mais aussi des outils pour traiter les représentations construites. Mais cette possible interdisciplinarité doit être interrogée en profondeur ; en effet, cette rencontre ne peut être profitable que si chaque discipline identifie ses intérêts, ses objectifs et ses présupposés. Dans un contexte de demande sociétale accrue en matière d'analyse de corpus et d'explosion des possibilités de calcul dans un cadre automatique, il reste capital que les questions sémantiques soient examinées non seulement dans la perspective d'alimenter des outils d'ingénierie ou de TAL mais aussi dans celle de mieux comprendre ce qui constitue le sens (les sens) d'un corpus. Il est donc nécessaire que les sémanticiens mènent une réflexion

poussée sur les perspectives profondément renouvelées qu'autorisent les possibilités de l'analyse de corpus assistée par des outils.

La sémantique de corpus n'est certainement qu'au début de son histoire, en tout cas au début d'une nouvelle histoire du fait de la mise à disposition de corpus nombreux et d'outils qui permettent de tester des hypothèses. Pour ce qui est de la façon dont je m'intéresse à cette problématique, qui consiste pour l'essentiel dans le passage d'un corpus à une modélisation de ce corpus, différentes pistes de réflexion, à plus ou moins long terme, s'ouvrent. L'une concerne les réseaux relationnels et la possibilité de définir des marqueurs en lien avec un genre ; l'autre consiste en une exploration de la possibilité d'asseoir les notions de genre, genre textuel et genre interprétatif sur l'observation de régularités linguistiques. Le troisième concerne les complémentarités entre analyse sémantique introspective et analyse sémantique à partir de corpus.

1. Représentation relationnelle

La représentation sous forme de réseaux relationnels est privilégiée par bon nombre de disciplines, on l'a vu. Lorsque cette représentation est construite à partir d'un corpus, elle entraîne nécessairement un ensemble de questions sur les marqueurs de relation et, une fois encore, sur la variation. Variation dans le lien entre relation conceptuelle et marqueur mais aussi entre marqueur et genre de corpus.
Dans le même temps, la recherche d'une représentation sous forme de relation permet d'examiner un certain nombre de phénomènes, pas en tant que catégories de telle ou telle nature grammaticale mais d'après leur capacité à marquer telle ou telle relation. C'est le cas par exemple de *chez* ou de *avec*, qui sont généralement décrits par rapport à l'ensemble de leurs sens supposés (que l'on essaie, le plus souvent, de rassembler sous un ou deux méta-sens censés expliquer et prédire tous les effets possibles). Lorsqu'on examine ces prépositions, d'une part dans un corpus et d'autre part dans la perspective d'étudier leur capacité à « marquer » une relation particulière, les résultats sont éclairants à plusieurs titres. Non seulement on repère le rôle du genre textuel dans le marquage de cette relation mais on découvre aussi un fonctionnement de cette relation beaucoup plus complexe que ce que l'on peut imaginer par introspection. C'est le constat que j'ai fait avec la méronymie en lien avec *avec*.
A moyen terme, certaines études présentées dans ce mémoire mériteraient d'être réexaminées et développées et d'autres pourraient se mettre en place dans le même type de perspective. Voici celles qui semblent les plus évidentes.

Préposition/relation/genres

Plusieurs des résultats présentés concernent le marquage de la méronymie par une préposition, en lien avec certains genres textuels. Dans cette perspective, voici quelles seraient les pistes à explorer :

- Confirmation du fonctionnement de *chez* et de *avec* sur d'autres corpus, de même genre. Les expériences dont j'ai rendu compte ne concernent le plus souvent qu'un corpus (considéré comme représentatif d'un genre). Ce fonctionnement de marquage devrait être vérifié sur d'autres corpus (la vérification est en cours pour *chez ;* elle est réalisée par une étudiante de maîtrise que j'encadre).
- Identification d'autres genres textuels pertinents pour le marquage de la méronymie par *avec*. Il est probable que d'autres genres textuels, associés à

certains domaines favorisent l'interprétation méronymique de *avec*. Je pense à des corpus d'architecture ou de tourisme par exemple mais d'autres domaines seraient sans doute pertinents, identifiables seulement par tâtonnement (comme je l'ai souligné, c'est une des difficultés majeures de la linguistique de corpus).

- Travail sur le genre catalogue. Le genre textuel *catalogue* semble suffisamment riche pour mériter une analyse approfondie, non seulement sur la possibilité du marquage de la méronymie mais aussi, par exemple, sur le lien entre les rubriques de catalogues et les définitions de dictionnaire (ces rubriques ont en effet l'allure de définitions classiques tout en s'en démarquant par leur visée) et sur les possibilités de réorganiser ce genre en sous-genres. Une thèse est en cours sur une partie de ces problématiques (C. Pernet, sous la direction commune de M. Roché et de moi-même).
- Etude d'autres prépositions et de leur possibilité de marquer la méronymie en lien avec certains genres de textes. La possibilité de marquer la méronymie grâce à *à* a été repérée depuis longtemps dans des exemples comme *débardeur à bretelles*, que j'ai noté dans le chapitre V (voir Borillo, 1996). Mais il est probable que cette possibilité peut être corrélée avec certains genres textuels, comme le genre catalogue mais d'autres encore, qu'il faudra identifier.

Stabilité des marqueurs et des relations

L'hypothèse de la stabilité des marqueurs est fondamentale du point de vue des possibilités d'analyse linguistique. En effet, c'est elle qui, en faisant intervenir le genre textuel, l'objectif d'interprétation et en se basant sur des éléments linguistiques, permet de maîtriser les interprétations possibles et de les expliquer. En lien avec les marqueurs, la stabilité des relations cadrée par une situation doublement caractérisée est, elle-aussi, fondamentale. Il reste qu'il faut mettre cette hypothèse à l'épreuve de la réalité des faits langagiers pour confirmer qu'une certaine stabilisation est possible.

Un projet qui démarre avec le CNES devrait nous permettre de tester cette hypothèse (j'encadrerai deux post-doctorants sur ce projet : Maarten Janssen et Josette Rebeyrolle). Il s'agit d'identifier des moyens linguistiques qui permettraient de repérer des évolutions de connaissances (*cf.* description du projet en annexe). L'hypothèse de la stabilité des marqueurs et des relations va être examinée de deux manières :

- **Stabilité des relations d'une langue à l'autre**

Etant donné la nature internationale du projet, les corpus sur lesquels nous allons travailler sont en anglais, nous allons donc devoir identifier des marqueurs en anglais ; or, les marqueurs que nous avons déjà élaborés sont en français. Pour identifier leurs équivalents anglais, nous allons mettre en place une étude sur des corpus parallèles, à la manière de (Pearson,2000). Il s'agit d'essayer de faire apparaître des marqueurs qui n'ont pas été repérés antérieurement pour l'anglais, en examinant la manière dont les marqueurs français sont traduits. L'hypothèse qui sous-tend cette approche est que les relations sont stables d'une langue à l'autre, pour une situation d'analyse en tous points identiques.

- **Stabilité dans le temps**

Une des pistes de réflexion pour repérer l'évolution des connaissances consiste à prendre appui sur la supposée stabilité des marqueurs pour identifier des changements à la fois dans les termes (phénomènes d'ellipse en particulier) mais aussi dans l'organisation sous forme de relations. Il se peut ainsi que les éléments mis en relation par un marqueur défini pour un

corpus à T0 ne soient pas les mêmes que ceux de T+n. Par exemple, on peut voir apparaître des changements dans l'organisation taxinomique.

Stabilité de l'expression d'une information discursive
Dans un tout autre ordre d'idée, puisqu'il ne s'agit pas du marquage d'une relation conceptuelle mais du « marquage » d'une information, nous sommes en train de tester la stabilité dans l'expression d'une information précise : l'expression d'un dysfonctionnement technique (projet CENA, décrit en annexe). Nous souhaiterions vérifier si, quel que soit le domaine, mais toujours dans une situation dialogique, l'expression d'un dysfonctionnement (technique mais aussi peut-être d'une autre nature, par exemple l'expression de symptômes) se fait toujours en utilisant les mêmes structures discursives (thèse de P. Vergely, en cours, co-dirigée par A. Borillo et moi-même).

Stabilité du marquage de l'hyperonymie par rapport à l'application
Un des résultats qui me semblent les plus prometteurs en matière de marquage a été fourni par l'étude sur l'anaphore dite associative. Rappelons que ces résultats ont montré que l'hyperonymie repérée par ce mode de marquage peut correspondre à des constructions taxinomiques différentes de celles repérées avec des marqueurs plus classiques, ces constructions pouvant correspondre à des applications différentes (par exemple thésaurus ou modélisation d'une connaissance plus précise), *cf*. chapitre V. Il est évident que si ces fonctionnements sont confirmés sur d'autres corpus, il s'agira là d'un résultat majeur. Mais il reste à étudier si ce type de fonctionnement (anaphore associative correspondant majoritairement à une hyperonymie) est propre à un genre textuel particulier ou pas.

2. La question du genre

Il me semble que la réflexion sur les genres textuels n'a pas fourni tous les résultats que l'on peut en attendre. Tout d'abord, comme je le soulignais dans le chapitre V, on peut envisager que les genres de textes ne soient plus seulement étudiés sur des bases introspectives.

- D'une part, le nombre de corpus disponibles sous un format électronique et interrogeables de manière automatique va croissant, sur internet, mais aussi, et peut-être surtout, dans les laboratoires de linguistique, de TAL, d'ingénierie des connaissances, dans les entreprises etc. Cette abondance (relative tout de même) rend moins aléatoire la définition de genres car elle peut se baser sur des textes existants plutôt que sur une idée *a priori* de genres textuels.
- D'autre part, et corrélativement, ces textes peuvent être analysés grâce aux outils de TAL. On peut donc rapidement confirmer ou infirmer des intuitions grâce à l'étude automatique de phénomènes linguistiques et en tout cas développer le profilage de textes, c'est-à-dire le classement de textes sur la base du repérage d'un certain nombre d'éléments morpho-syntaxiques communs (Illouz et al, 1999). On peut ainsi envisager de donner une assise linguistique à des intuitions et diminuer le hiatus, depuis longtemps identifié entre genre (intuitif) et type de texte (associé à des régularités linguistiques), pour reprendre la terminologie de Biber.

Le problème reste cependant que ces textes ne sont pas facilement accessibles car ils sont disséminés dans différents lieux ; on peut espérer que des projets d'envergure permettront de les rassembler et de les formater pour les mettre à disposition des chercheurs.
Au-delà de ces possibilités, qui permettront d'avoir une bien meilleure idée de la réalité linguistique des genres textuels, je pense que cette question sera loin d'être épuisée si on ne s'interroge pas aussi sur la possibilité de prendre en compte le point de vue d'analyse sur les

textes et d'ériger ces points de vue en genres interprétatifs. En effet, aussi élaborée soit-elle, l'analyse des textes « à la Biber », tout en concernant des corrélations de phénomènes linguistiques, ne peut concerner que des phénomènes qui sont accessibles sous des formes identifiables automatiquement, c'est-à-dire des phénomènes morpho-syntaxiques tels que l'utilisation de tel ou tel temps, de telle ou telle structure, de tel ou tel déterminant…. Ce faisant, ce mode d'interrogation n'est pas neutre ; il constitue un point de vue, certes pertinent et certes très utile, en particulier s'il vise à confirmer des intuitions sur les genres textuels. Mais il reste qu'il s'agit d'un point de vue. Comme on l'a vu, on peut avoir d'autres types de points de vue, par exemple, vouloir construire un réseau relationnel (voire des réseaux relationnels en fonction du mode d'utilisation envisagé par la suite). Et dans ce cas-là, les rapprochements entre textes se font selon une autre dimension, par exemple selon leur richesse en marqueurs de relation. Or, la capacité des éléments linguistiques à « marquer » telle ou telle relation ne peut être repérée automatiquement mais seulement par une interprétation, ce qui suppose que soit mise en œuvre une linguistique de corpus qui ne pourra qu'être assistée par des outils de TAL et qui ne sera efficace que si elle prend d'emblée en compte l'objectif de la modélisation.

Tout comme les recherches sur la réalité linguistique des genres textuels doivent être développées, il paraît nécessaire de développer une réflexion sur les genres interprétatifs. D'une part, de manière indépendante de la réflexion sur les genres textuels, et certainement dans une perspective interdisciplinaire, en interrogeant les différentes attentes par rapport à l'interprétation des textes (ergonomie, ingénierie de la connaissance, documentation, recherche d'information mais aussi sémantique formelle, sémantique cognitive…). Cette façon de procéder permettrait de définir des genres interprétatifs intuitifs, un peu comme on le fait pour les genres textuels. D'autre part, de manière étroitement reliée avec le problème des genres textuels, l'objectif pourrait être d'arriver à une caractérisation des textes non seulement en fonction des régularités linguistiques de surface mais aussi en fonction de l'objectif de l'analyse. Par exemple, on pourrait peut-être arriver à caractériser la didacticité des textes d'une part grâce à l'identification de régularités linguistiques de surface mais aussi grâce au repérage d'un grand nombre de marqueurs de relations, ce qui manifeste une possibilité de représenter ces textes sous une forme relationnelle. En effet, comme je l'ai souligné, les marqueurs relèvent d'une compétence moins experte que les termes reliés, et on peut penser que cette caractéristique est propre à des textes destinés à des lecteurs moins experts que les rédacteurs, c'est-à-dire des textes « didactiques ». On pourrait avoir ainsi plusieurs caractérisations possibles pour un même texte, selon le point de vue interprétatif où l'on se place.

Même enrichie et développée, il est évident que la réflexion sur le genre n'expliquera pas tout le phénomène de la variation en sémantique textuelle. Par essence inhérent au processus discursif, on ne peut espérer maîtriser et prédire parfaitement ce phénomène. Toutefois, il existe une grande part d'inconnu sur les résultats que peut encore produire l'analyse réelle de textes. De nombreuses recherches se développent dans ce paradigme, ce qui permet d'espérer que l'on comprenne mieux le phénomène de la variation ; mais, malheureusement, aucun recensement des résultats n'existe. Or, pour arriver à mieux cerner ce phénomène, il serait indispensable que des bases de données soient constituées, qui comporteraient un grand nombre d'informations à la fois sur les conditions de productions des textes mais aussi sur les conditions et les objectifs d'interprétations et sur les résultats obtenus (c'est le même type de souhait que manifeste Habert (Habert, 2000 (b), 20). Alors seulement on pourrait espérer voir se dessiner un premier balisage du phénomène de la variation en corpus.

3. Analyse introspective, analyse de textes, quelles complémentarités possibles ?

Au terme de ce mémoire, je fais le constat qu'une question a traversé toute ma réflexion : dans l'interprétation d'un texte ou d'un corpus de textes, qu'est-ce qui préexiste et qu'est-ce qui est nouveau. J'ai proposé de systématiser les interprétations avec la notion de genre déclinée en genre textuel et genre interprétatif ; c'est sûrement une bonne façon pour le sémanticien d'essayer de comprendre voire d'expliquer et d'anticiper les phénomènes sémantiques. Mais cela laisse de côté deux éléments : d'une part, le fait que, avant les genres textuels il existe la langue et ses systématicités propres et, d'autre part, qu'il existe le discours individuel (c'est-à-dire aussi l'interprétation individuelle). En linguistique, on ne peut sortir de la tension qui existe entre individualité et appartenance à un collectif. Si l'analyse de textes réels permet de mettre l'accent sur des phénomènes tout à fait occultés par l'approche introspective, je ne crois pas qu'elle puisse être le moyen idéal pour décrire une langue. En effet, il est très difficile de situer un texte ou un corpus en termes de représentativité par rapport à une langue : la langue consiste autant en potentialités toujours mouvantes, qu'en réalisations effectives. Il me semble très difficile par exemple de décrire tel ou tel phénomène syntaxique à partir de l'analyse d'un corpus censé représenter le fonctionnement d'une langue ; si bien que j'ai des doutes sur la pertinence de la constitution de dictionnaires ou de grammaires à partir de corpus dits équilibrés. Avec ce genre d'objectif, je comprends la démarche introspective ou celle qui utilise les données textuelles comme de simples attestations, le problème vient plutôt peut-être de la nature de cet objectif dont j'ai du mal à penser qu'il n'a pas, un tant soit peu, une vocation normative[73]. Or, toute analyse sémantique met en œuvre une dimension introspective ou intuitive, qui fait appel à la mémoire, plus ou moins consciente[74], que nous avons conservée des précédentes occurrences de certains phénomènes ; mais il ne s'agit pas d'une compétence que l'on peut définir de manière définitive ni surtout comme similaire d'un interprète à l'autre. On ne peut donc compter sur elle pour systématiser les interprétations, encore moins pour les automatiser. La langue reste ainsi le système sur lequel s'appuie tout locuteur mais aussi tout interprète : c'est un système aux contours mouvants, ce qui a un effet de frustration pour le linguiste à la recherche de résultats fermes et définitifs. A l'autre extrémité de la tension individu/collectivité, au-delà des limites de l'appartenance à un genre (textuel et/ou interprétatif) les productions individuelles, que l'on rejette généralement comme autant d'idiolectes (souvent considérés comme sans intérêt pour le linguiste) montrent que la construction d'un sens peut échapper à toute prévision. Et peut-être peut-on s'en féliciter : il s'agit de lieux où l'on échappe un peu aux règles préétablies ; on est en effet dans le lieu de création de nouvelles productions qui, poussée à son extrémité, relève de la poésie ; on est aussi dans le lieu de la création de nouvelles interprétations, c'est-à-dire, pour un sémanticien, dans le lieu de la création « scientifique ». On en arrive donc au paradoxe suivant : le sémanticien cherche à repérer des régularités de sens, mais c'est parce qu'il s'autorise à ne pas être dans le type de régularités de sens généralement admises qu'il peut proposer des pistes nouvelles de réflexion. De là vient sans doute l'impression de vertige qui saisit parfois lorsqu'on fait de la recherche en sémantique textuelle.

[73] Il suffit de voir l'usage qui est fait du dictionnaire dans les collèges : il est le garant de ce qui est acceptable dans telle ou telle langue.

[74] Ainsi que je l'ai déjà noté, il me semble que nous avons surtout perdu la mémoire des contextes (co-textes et situations de production) dans lesquels nous avons rencontré ces phénomènes.

BIBLIOGRAPHIE

ADAM J.-M., 1999 : *Linguistique textuelle. Des genres de discours aux textes*. Paris : Nathan.

AHMAD K., 1993 : « Terminology and Knowledge Acquisition : A Text Based Approach ». K.D Schmitz (ed) : *Proceedings of TKE'93 : Terminology and Knowledge Engineering*, Frankfurt : Indeks verlag. pp. 56-70.

AHMAD K., FULFORD H., 1992 : « Knowledge Processing : Semantic Relations and their Use in Elaborating Terminology ». *Computing Sciences Report CS-92*-Guildford : University of Surrey.

ASSADI H., 1998 : *Construction d'ontologies à partir de textes techniques : Application aux systèmes documentaires*. Thèse d'informatique de l'Université Paris 6.

AUGER P., L'HOMME M.-C, 1995 : « La terminologie selon une approche textuelle : une représentation plus adéquate du lexique dans les langues de spécialité ». *Terminologie et langues de spécialité*, 7/8, Dalhousia, Halifax, Nova Scotia. Canada : *ALFA (Actes de Langue Française et de Linguistique).* pp. 17-21.

AUROUX S., 1989 : « Introduction ». S.Auroux (ed) : *Histoire des idées linguistiques, Tome 1 : la naissance des métalangages*. Liège-Bruxelles : Mardaga. pp. 13-35.

AUROUX S., 1998 : La raison, le langage et les normes. Paris : PUF.

AUSSENAC-GILLES N., BIÉBOW B., SZULMAN S., 2000 : « Corpus Analysis for conceptual modelling ». *Proceedings of the workshop on Ontologies and texts, EKAW'2000* (European Knoweldge Acquistion Worshop). Juan les Pins. pp. 13-20.

AUSSENAC-GILLES N., CONDAMINES A., 2001 : « Entre textes et ontologies formelles : les Bases de Connaissances Terminologiques ». M.Zacklad et M.Grundstein (eds) : *Ingénierie et capitalisation des connaissances*. Paris : Hermes, pp. 153-176.

AUSSENAC-GILLES N., SEGUELA P., 1999 : « GEDITERM, un logiciel de gestion de bases de connaissances terminologiques ». Enguehard C. et Condamines A. (Eds.) *: Actes des 3es Rencontres "Terminologie et intelligence artificielle"* (Nantes, 10 et 11 mai 1999), *Terminologies Nouvelles* n°19. pp. 111-123.

BACHIMONT B., 1995 : « Ontologie régionale et terminologie : Quelques remarques méthodologiques et critiques ». G.Otman (ed) : *La Banque des Mots* n°7. pp. 65-84.

BACHIMONT B., 2000 : « Engagement sémantique et engagement ontologique : conception et réalisation d'ontologie en ingénierie des connaissances ». J.Charlet, M.Zacklad, G.Kassel, D.Bourigault (eds) : *Ingénierie des Connaissances, Evolution récentes et nouveaux défis*. Paris : Eyrolles. pp. 305-324.

BAKHTINE M., 1984 : *Esthétique de la création verbale*. Paris : Gallimard, Tel.

BARTNING, I., 1996 : « Les nominalisations déverbales dans les SN complexes en *de* envisagées sous l'angle des traits processifs et résultatif ainsi que de l'opposition abstrait/concret ». N.Flaux, M.Glatigny, D.Samain, (eds) *Les noms abstraits*. Lille : Presses Universitaires du Septentrion. pp. 323-336.

BARTNING I., NOAILLY M., 1993 : « Du relationnel au qualificatif : flux et reflux ». *L'Information Grammaticale* n°58. pp. 27-33.

BASILI R., PAZIENZA M.-T., 1997: « Lexical Acquisition for Information Extraction ». *Lectures Notes in Computer Science.* Berlin : Springer-verlag.

BEACCO J.-C., MOIRAND S., 1995 : « Autour des discours de transmission de connaissances ». D.Maingueneau (ed) : *Langages* n° 105, *Les Analyse de discours en France*. pp. 32-53.

BEJOINT H., THOIRON P., 2000 : *Le sens en terminologie*. Lyon : PUL.

BENVENISTE E., 1966 : *Problèmes de linguistique générale, 1*. Paris : Gallimard.

BENVENISTE E., 1974 : *Problèmes de linguistique générale, 2*. Paris : Gallimard.

BIBER D., 1988 : *Variation Across Speech and Writing* . Cambridge University Press.

BIBER D., JOHANSSON S., LEECH G., CONRAD S., FINEGAN E., 2000 *: Grammar of Spoken and Written English*. London : Longman.

BIÉBOW B, SZULMAN S**.,** 1999 ***:*** « TERMINAE : A linguistic-based tool for the building of a domain ontology». *11th European Workshop, Knowledge Acquisition, Modeling and Management (EKAW' 99),* Dagstuhl Castle, Germany. pp. 49-66.

BLACHE P., 2000 : « A quoi sert l'annotation syntaxique de corpus ? ». M.Bilger (ed) : *Corpus : Méthodologie et applications linguistiques*. Paris : Honoré Champion. pp. 82-93.

BLANCHE-BENVENISTE C., 1996 : « De l'utilité du corpus linguistique ». *Revue française de linguistique appliquée*, 1-2. pp. 25-42.

BORILLO A., 1985 : « Discours ou métadiscours ? ». *DRLAV*, n°32. pp. 47-61.

BORILLO A., 1996a, : « Exploration automatisée de textes de spécialité : repérage et identification automatique de la relation lexicale d'hyperonymie ». *Linx,* n°34-35. pp. 113-121.

BORILLO A. 1996b : « La relation partie-tout dans la structure N à N en français ». *Faits de langue* n°7. pp. 11-120.

BOUQUET S., 1998 : « Linguistique textuelle, jeux de langage et sémantique du genre ». S.Bouquet (ed), *Langages* n°*129 : Diversité de la (des) science(s) du langage aujourd'hui. Figures, modèles et concepts épistémologiques*. pp. 112-124.

BOURIGAULT D., 1996 : « Lexter, a Natural Language Processing Tool for Terminology Extraction ». *Proceedings Euralex'96*, Göteborg University, Department of Swedish. pp. 771-779.

BOURIGAULT D., CONDAMINES A., 1995 : « Réflexions sur le concept de base de connaissances terminologiques ». *Journée du PRC IA*, 1-2 février 1995, Nancy : Teknea, pp. 425-445.

BOURIGAULT D., FABRE C., 2000 : « Approche linguistique pour l'analyse syntaxique de corpus ». A.Condamines (ed) : *Les Cahiers de Grammaire* n°25 : Sémantique et Corpus. pp. 131-152.

BOURIGAULT D., JACQUEMIN C., 2000 :« Construction de ressources terminologiques ». J.-M. Pierrel (ed) : *Ingénierie des langues*, Traité I2C, Paris : Hermes. pp. 215-233.

BOURIGAULT D., SLODZIAN M., 1999 : « Pour une terminologie textuelle ». *Terminologies Nouvelles* n°19. pp. 29-32.

BOUTET J., 1992 : « La linguistique variationiste face à l'expertise linguistique et au sens ». F.Gadet (ed) : *Langages n°108 : Hétérogénéité et variation : Labov, un bilan*. pp. 90-100.

BOUTET J., 1995a : *Construire le sens*. Bern, Berlin : Peter Lang.1994.

BOUTET J., 1995b : « Le travail et son dire ». J.Boutet (ed) : *Paroles au travail*. Paris : L'Harmattan. pp. 247-267.

BOUTET J., GARDIN B., LACOSTE M., 1995 : « Discours en situation de travail ». D.Maingueneau (ed) : *Langages n°117 : Les analyses de discours en France*. pp. 12-31.

BOUVERET M., DELAVIGNE V., 1999 : Sémantique des termes spécialisés. Rouen : Publications de l'université de Rouen, collection Dyalang,

BOWDEN P.R., HALSTEAD P., ROSE T.G., 1996 : « Extracting Conceptual Knowledge From Text Using Explicit Relation Markers ». *Proceedings of the European Knowledge Engineering Workshop (EKAW-96), Lectures notes in Artificial Intelligence*, n°1076, Springer Verlag. pp. 146-162.

BOWKER L ., 1997 : « Multidimensional Classification of Concepts and Terms » . S.E. Wright, G.Budin (eds) : *Handbook of Terminology management*. Amsterdam/Philadelphia : John Benjamins. pp. 131-143.

BOWKER L, PEARSON J., 2002 : *Working with Specialized Language. A practical guide to using corpora*. London, New York : Routledge.

BRANCA-ROSOFF S., 1999 : « Types, modes et genres : entre langue et discours ». S.Branca-Rosoff (ed) : *Langage et Société* n°87, *Types, modes et genres de discours*. pp. 5-24.

BRANCA-ROSOFF S., COLLINOT A., GUILHAUMOU J., MAZIERE F., 1995 : « Questions d'histoire et de sens ». D.Maingueneau (ed) : *Langages* n°117, *Les analyses de discours en France.* pp. 54-66.

BRONCKART J.-P., 1996 : *Activités langagières, textes et discours* . Lausanne : Delachaux et Niestlé.

BRUNET E., 1995 : « Un hypertexte statistiques pour grands corpus: HYPERBASE ». Lexicomatique et Dictionnairique. *Actes des IVes journées scientifiques du réseau lexicologie, Terminologie, Traduction*, Lyon.

CABRE M.-T., 1998 : *La terminologie, Théorie, méthode et applications*. Paris : Armand Colin, Les Presses de l'Université d'Ottawa.

CABRE M.-T. MOREL J., TEBE C., 1997: « Las relaciones conceptuales de tipo causal : un caso practico ». *Actes du V Simposio de Terminologia*. RITERM. Mexico.

CADIOT P., 1997 : « *Avec*, ou le déploiement de l'éventail ». C.Guimier (ed) : *Co-texte et calcul du sens* ; Caen : Presses Universitaires de Caen. pp .135-155.

CHARAUDEAU P., MAINGUENEAU D., 2002. : *Dictionnaire d'analyse du discours* . Paris : Editions du Seuil.

CHARLET J., BACHIMONT B., BOUAUD J., ZWEIGENBAUM P., 1996 : « Ontologie et réutilisabilité : expérience et discussion ». N. Aussenac-Gilles, P. Laublet, C. Reynaud (eds) : *Acquisition et Ingénierie des Connaissances*. Toulouse : Cépaduès-Editions. pp.69-88.

CHARLET J., ZACKLAD M., KASSEL G., BOURIGAULT D., 2000 : *Ingénierie des Connaissances, Evolutions récentes et nouveaux défis*. Paris : Eyrolle et France télécom.

CHOI-JONIN I., 1995 : « La préposition " avec " : opérateur de (dé)composition ». *Scolia* n°5. pp. 109-129.

CHOMSKY N., 1971 : *Aspects de la théorie syntaxique*. Paris : Seuil

CHOMSKY N., 1977 : *Langue, linguistique, politique ; Dialogues avec Mitsou Ronat*. Paris : Flammarion.

CONDAMINES A., 1988 : « Intelligence Artificielle, Linguistique Informatique, deux approches pour le traitement automatique du Langage Naturel ». *Actes du 3ème colloque de l'ARC (Association pour la Recherche Cognitive)*, Mars 1988, Toulouse.

CONDAMINES A., 1995 : « Analyse de textes spécialisés pour le recueil de données terminologiques ». *Terminologies Nouvelles*, n°14. décembre 1995. pp. 35-42.

CONDAMINES A., 1996 : « Aide à l'acquisition de connaissances par l'étude de la terminologie ». Aussenac-Gilles, N., Laublet, P., Reynaud, C. (eds) : *Acquisition et ingénierie des connaissances*, Toulouse : Cépaduès-Editions. pp. 247-266.

CONDAMINES A., 1997 : « Langue spécialisée ou discours spécialisé *? »*. L.Lapierre, I. Oore, H.R. Runte (eds) : *Mélanges de linguistique offerts à Rostislav Kocourek*, Université de Dalhousie : Les presses d'Alfa. pp. 171-184.

CONDAMINES A., 1998 : « Analyses des nominalisations dans un corpus spécialisé : comparaison avec le fonctionnement en corpus "général" ». A.Clas, S.Mejri, T.Baccouche (eds), *La Mémoire des mots, Actes Ves Journées Scientifiques du réseau Lexicologie, Terminologie, Traduction*, Tunis, 25-27 septembre. Montréal : Aupelf. pp. 351-368.

CONDAMINES A., 1999 : «Analyse des structures " Nominalisation + de N ", " Nominalisation + N ", " Nominalisation + Adjectif relationnel " dans un corpus spécialisé ». M.Aurnague, A.Condamines, J.-P. Maurel, C.Molinier, (eds) : *L'emprise du sens*, Amsterdam, Atlanta : Rodopi. pp. 61-82.

CONDAMINES A., 2000 : « *Chez* dans un corpus de sciences naturelles : un marqueur de méronymie ? ». *Cahiers de Lexicologie* n° 77. pp. 165-187.

CONDAMINES A., 2002 : « Corpus Analysis and Conceptual Relation Patterns ». *Terminology*, volume 8 number 1. pp. 141-162.

CONDAMINES A., AMSILI P., 1993 : « Terminology between Language and Knowledge: an example of Terminological Knowledge Base ». K.-D. Schmitz (ed) : *TKE 93* Terminology and Knowledge Engineering. Frankfurt : Indeks Verlag. pp. 316-323.

CONDAMINES A., REBEYROLLE J., 1996 : « Point de vue en langue spécialisée ». *META*. 42, 1, pp. 174-184.

CONDAMINES A., REBEYROLLE J., 1997 : « Utilisation d'outils dans la constitution de bases de connaissances terminologiques: expérimentation, limites, définition d'une méthode ». *Actes des 1ères Journées Francil*. pp. 529-535.

CONDAMINES A., REBEYROLLE J., 2000 : « Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode » J. Charlet, M. Zacklad, G. Kassel & D. Bourigault, (eds). *: Ingénierie des Connaissances, évolutions récentes et nouveaux défis*. Paris : Eyrolles. pp. 225-242.

CONDAMINES A., FABRE C., PERY-WOODLEY M.P., 1999 (EDS) : Actes de l'atelier « Corpus et TAL : pour une réflexion méthodologique », TALN'99, Cargèse, Corse.

CONEIN B. : « Hétérogénéité sociale et hétérogénéité linguistique ». F.Gadet (ed) : *Langages n°108 : Hétérogénéité et variation : Labov, un bilan*. pp. 101-113.

CORBIN P., 1980 : « De la production des données en linguistique introspective » . A.M. Desseaux Berthoneau (ed) : *Théories linguistiques et traditions grammaticales.* Lille : PUL. pp. 121-179.

CRUSE D.A., 1986 : *Lexical Semantics.* Cambridge : Cambridge University Press.

CUSIN-BERCHE F., 1997 : « A la recherche de quelques caractéristiques linguistiques des textes spécialisés et de la rédaction technique ». *Le langage et l'homme*, Vol. XXXII n°4, *Langues de spécialité et terminologie*. pp. 21-55.

DACHELET R., 1994 : *Sur la notion de sous-langage*. Thèse en sciences du Langage de l'Université Paris VIII.

DAILLE B., 1994 : *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*. Thèse d'informatique. Université Paris 7.

DAVID S., PLANTE P., 1990 : *Termino* version 1.0, Rapport du centre d'Analyse de textes par Ordinateur. Université du Québec à Montréal.

DAVIDSON L., 1998 : *Knowledge Extraction Technology for Terminology*. Master of Art Thesis, University of Ottawa.

DEFRANCQ B., WILLEMS D., 1996 : « De l'abstrait au concret, Une réflexion sur la polysémie des noms déverbaux ». *Les noms abstraits, histoire et théorie*. Presses Universitaires du Septentrion. pp. 221-223.

DELHAY C., 1996 : « Texte, contexte, contextualisation : A-t-on progressé ? » P.Schmoll (ed), *Scolia* n°6, Contextes . pp. 61-86.

DEPECKER L., 2000 : « Le signe entre signifié et concept ». H.Béjoint et P.Thoiron (eds) : *Le sens en terminologie*. Lyon : PUL. pp. 86-126.

DESCAMPS J.L., MOCHET M.A, LEWIN T., LAMIZET B., COSTES D., 1992 : *Sémantique et concordances.* Publication de l'INALF, Collection "St Cloud", Paris : Klincksieck.

DESCLES J-P., CARTIER E., JACKIEWICZ A, MINEL J.-L., 1997 : « Textual Processing and Contextual Exploration Method ». *Actes de CONTEXT'97*. Rio de Janeiro, Brésil, pp. 189-197.

DEULOFEU J., 1992 : « Variation syntaxique : recherche d'invariants et étude des attitudes des locuteurs devant la norme ». F.Gadet (ed) : *Langage* n°108, *Hétérogénéité et variation : Labov, un bilan*. pp. 66-78.

DEVILLE G., 1989 : *Modelization of Task-oriented Utterances in a Man-Machine Dialogue System* ; Thèse de Philologie, Université d'Antwerpen.

DUBOIS J., 1966 : « Les problèmes du vocabulaire technique ». *Cahiers de lexicologie*, Vol.IX-2. pp. 103-112.

DUBOIS J., GUESPIN L., GIACOMO M., MARCELLESI C., MARCELLESI J.B., MEVEL J.P., 1973 : *Dictionnaire de linguistique*. Paris : Larousse.

DUBOIS J. ET AL., 1994 : *Dictionnaire de linguistique et des sciences du langage*. Paris : Larousse.

DUBUC R., 1985 : *Manuel pratique de terminologie*. Montréal : Linguatech.

DUCROT O., 1980. : *Les mots du discours.* Paris : Editions de Minuit.

ENCREVE P., 1976. : « Introduction ». W. Labov : *Sociolinguistique*. Paris : Editions de Minuit.

ENGUEHARD C., PANTERA L., 1995 : « Automatic natural acquisition of terminology ». *Journal of Quantitative Linguistics*, vol.2, n°1. pp. 27-32

FABRE C., 1996 : *Interprétation automatique des séquences binominales en anglais et en français. Applications à la recherche d'informations*. Thèse d'informatique de l'Université de Rennes 1.

FALL K ., LEARD M., SIBLOT P., 1996 : *Polysémie et construction du sens*. Montpellier : Praxiling,

FELLBAUM C. ET AL, 1999 : *Wordnet. An Electronic Lexical Database.* Cambridge, London : The MIT Press.

FELBER H., 1987. : *Manuel de terminologie*. Paris : Unesco.

FELBER H., 1995 : « Terminology Research : Its Relation to the Theory of Science ». *Terminologie et langues de spécialité*, 7/8, Dalhousia, Halifax, Nova Scotia. Canada : *ALFA (Actes de Langue Française et de Linguistique).* pp. 163-171.

FILLMORE C.J., 1968 : «The case for case ». E. Bach and R.T. Harms (eds) : *Universals in Linguistic Theory*. New York : Holt, Rinehart, Winston. pp. 1-90.

FOUCAULT M., 1966. : *Les mots et les choses*. Paris : Tel, Gallimard.

FIRTH J.R, 1957 : *Papers in Linguistics 1934-1951*. Oxford University Press. 1969. (première édition : 1957).

FRATH P., OUESLATI R., ROUSSELOT F., 2000 : « Identification de relations sémantiques par repérage et analyse de cooccurrences de signes linguistiques ». J.Charlet, M.Zacklad, G.Kassel, D.Bourigault : *Ingénierie des connaissances : Evolutions récentes et nouveaux défis*. Paris : Eyrolles. pp. 291-304

FUCHS C., 1997 : « L'interprétation des polysèmes en contexte ». G.Kleiber et M.Riegel (eds) : *Les formes du sens* . Paris : Duculot. pp. 127-134.

GADET F., 1992 : « Variation et hétérogénéité ». F.Gadet (ed) : *Langages* n°108 : *Hétérogénéité et variation : Labov, un bilan*. pp. 5-15.

GAIZAUSKAS R., WILKS Y., 1998 : « Information Extraction : Beyond Document Retrieval ». *Journal of Documentation*, vol.54, n°1. pp. 70-105.

GARCIA D., 1998 : *Analyse automatique des textes pour l'organisation causale des actions, Réalisation du système Coatis*. Thèse d'informatique, Université Paris IV.

GAUDIN F., 1993 : *Pour une socioterminologie*. Publications de l'Université de Rouen n°182.

GAUDIN F., 1995 : « Champs, clôtures et domaines : Des langues de spécialités à la culture scientifique ». *Meta*, XL.2. pp.229-238.

GEERAERTS D., 1991 : « Grammaire cognitive et sémantique lexicale ». *Communications* n°53, *Sémantique Cognitive*. pp. 17-50.

GOMEZ-PERREZ A., 1999 : « Développements récents en matière de conception, de maintenance et d'utilisation des ontologies ». *Terminologies Nouvelles* n°19. pp. 9-20.

GOUGENHEIM G., MICHEA R., RIVENC P., SAUVAGEOT A., 1958 *: L'élaboration du français fondamental*. Paris : Didier.

GREIMAS A., 1966 : *Sémantique structurale*. Paris : Larousse.

GROUPE μ, 1982 : *Rhétorique générale*. Paris : Editions du Seuil.

GUARINO N., 1995 : « Formal Ontology, Conceptual Analysis and Knowledge Representation". International Journal of Human-Computer Studies, 43. pp. 625-640.

GUILBERT, F., 1973 :« La spécificité du terme scientifique et technique ». *Langue Française*, n°17. pp. 5-17.

GUILBERT L., 1972 : *La créativité lexicale*. Paris : Larousse, collection "langue et langage".

GUSDORF G., 1988 : *Les origines de l'herméneutique*. Paris : Payot.

GUY S., LEARD J.-M., 1996 : « Polysémie verbale et structure d'arguments : les rapports entre SN et que P ». Fall K ., Léard J.-M., Siblot P. (eds) : *Polysémie et construction du sens*. Montpellier : Praxiling. pp. 181-193.

HABERT B., 2000a : « Détournements d'annotation : armer la main et le regard ». M.Bilger (ed) : *Corpus : Méthodologie et applications linguistiques*. Paris : Honoré Champion. pp.106-120.

HABERT B. 2000b : « Des corpus représentatifs : de quoi, pour quoi, comment ? » *Cahiers de l'Université de Perpignan* (31). pp. 11-58.

HABERT B., FABRE C., ISAAC F., 1998 : *De l'écrit au numérique : constituer, normaliser, exploiter les corpus électroniques.* Paris : InterEditions/Masson.

HABERT B, NAULLEAU E., NAZARENKO A., 1996 : « Symbolic Word Clustering for Medium-Size Corpora ». *Proceedings of the 16 th International Conference on Computational Linguistics* (Coling'96), Copenhague, vol.1. pp.490-495.

HABERT B., NAZARENKO A., SALEM A., 1997 : *Les linguistiques de corpus*. Paris : Armand Colin.

HABERT B., ZWEIGENBAUM P., 2002a : « Contextual Acquisition of Information Catégories : what has been done and what can be done automatically ? ». Nevin B. (ed) : *The Legacy of Zellig Harris : Language and Information into 21st century*, volume 2. Amsterdam : John Benjamins.

HABERT B, ZWEIGENBAUM P., 2002b : « Régler les règles ». *TAL*, 43(3). pp. 83-105.
HAMON T., 2000 : *Variation sémantique en corpus spécialisé : Acquisition de relations de synonymie à partir de ressources lexicales*. Thèse d'informatique, Université Paris13.
HAMON T., NAZARENKO A., (EDS), 2002 : *Structuration de terminologie*. TAL volume 43 – n°1/2002.
HARRIS Z., 1966 : *Structural linguistics.* The University of Chicago Press, first edition : 1951, seventh edition.
HARRIS Z.S., GOTTFRIED M., RYCKMAN T., MATTICK J., DALADIER A., HARRIS T.N., 1989 : *The form of information in science. Analysis of an immunology sublanguage.* Dordrecht : Kluwer Academic Publishers.
HEARST M.A., 1992 : « Automatic Acquisition of Hyponyms From Large Text Corpora ». *Proccedings, 14th International Conference on Computational Linguistics,* Nantes, France. pp. 539-545.
ILLOUZ G, HABERT B., FLEURY S., FOLCH H., HEIDEN S., LAFON P., 1999 : « Maîtriser les déluges de données hétérogènes ». A.Condamines, C.Fabre, M.-P.Péry Woodley (eds) : Actes de l'Atelier « *Corpus et TAL, Pour une réflexion méthodologique* ». TALN'99. pp. 37-46.
JACKIEWICZ A., 1996. « L'expression lexicale de la relation d'ingrédience (partie-tout) ». *Faits de Langues*, 7. Paris : Ophrys. pp. 53-62.
JACQUEMIN C., 2001 : *Spotting and Discovering Terms through NLP*. Cambridge MA : MIT Press.
JACQUES M.-P., 2000 : « La réduction du syntagme terminologique au fil du discours ». A.Condamines (ed) : *Les Cahiers de Grammaire* n°25. pp. 93-114.
JACQUES M.-P., SOUBEILLE A.-M., 2000 : « Partage des termes, partage des connaissances ? Construire une modélisation unique de plusieurs corpus ». *Actes de IC'2000*, Toulouse, IRIT. pp. 313-324.
JOUIS C., 1993 : *Contribution à la conceptualisation et à la modélisation des connaissances à partir d'une analyse de textes. Réalisation d'un prototype : le système Seek.* Thèse d'informatique, EHESS, Paris.
KAVANAGH J., 1996 : *The Text Analyzer: a tool for extracting knowledge from text* . Master Thesis, Université d'Ottawa.
KAYSER D., 1995 : « Terme et dénotation ». *La Banque des Mots*. Numéro spécial. pp. 19-34.
KELLER E., 1985 : *Introduction aux systèmes psycholinguistiques*. Chicoutimi, Québec : Gaëtan Morin.
KENNEDY G., 1998 : *An introduction to Corpus Linguistics*. London and New York : Longman.
KERBRAT-ORRECHIONI C., 1996 : « Texte et contexte ». P.Schmoll (ed) : *Scolia* n°6, *Contextes*. pp. 39-60.
KLEIBER G., 1981 : *Problèmes de référence : Descriptions définies et noms propres*. Paris : Klincksieck.
KLEIBER G., 1990 : *La sémantique du prototype*. Paris : PUF.
KLEIBER G., 1994 : *Nominales, essai de sémantique référentielle*. Paris : Armand Colin.
KLEIBER G., 1994 : « Lexique et cognition : y a-t-il des *termes* de base ? ». *Rivista di Linguistica* 6, 2, pp. 237-266.
KLEIBER G., 1999 : *Problèmes de sémantique, la polysémie en question*. Paris, Villeneuve d'Asq : Presses Universitaires du Septentrion.
KOCOUREK R. : « Textes et termes », *META*, 36, 1, 1991. pp. 71-76.
KOCOUREK R., 1991 : *La langue française de la technique et de la science*. Wiesbaden : Brandestetter. 1ère édition. 1982.
LABOV W., 1976 : *Sociolinguistique.* Paris : Editions de Minuit.

LECOLLE M., 2000 : « Figures et référence plurielle, en corpus journalistique ». A.Condamines (ed), *Cahiers de Grammaire n°25, Sémantique et Corpus*. pp. 29-52.

LEGLISE I., 2000 : « Lorsque les linguistes interviennent : écueils et enjeux ». *Revue Française de Linguistique Appliqué.* Volume V-1, juin 2000. pp. 5-14.

LENAT D.B., GUHA R.V., PITTMAN K., PRATT D., SHEPERD M., 1990 : « Cyc : Towards Programs With Common Sense ». *Communications of the ACM*, vol.33, N° 8, August 1990. pp. 30-49.

LERAT P., 1981 : « Les noms de relation », *Cahiers de lexicologie*, 39-2. pp. 55-65.

LERAT P., 1989 : « Les fondements théoriques de la terminologie ». *La Banque des Mots*, numéro spécial. pp. 51-62.

LERAT P., 1983 : *Sémantique descriptive*. Paris : PUF.

LERAT P., 1995 : *Les langues spécialisées*. Paris : PUF.

L'HOMME M.-C., 1998 : « Le statut du verbe en langue de spécialité et sa description lexicographique ». *Cahiers de lexicologie*, n°73. 1998-2. pp. 61-84.

L'HOMME M.-C., GEMME R., 1997 : « Modèle d'accès informatisé aux combinaisons lexicales spécialisées : verbes + nom (terme) et extension aux noms (déverbal) + préposition : nom (terme) ». L. Lapierre, I. Oore et H.R. Runte : *Mélanges de linguistique offerts à Rostlislav Kocourek* . Université de Dalhousie : Presses ALFA. pp. 89-103.

LOFFLER-LAURIAN A.-M., 1994 : « Les définitions dans la vulgarisation scientifique (presses, musées) ». *Etudes de Sémantique lexicale*, n°2. pp. 93-112.

LYONS J., 1978 : *Eléments de sémantique*. Paris : Larousse Universités.

LYONS J., 1980 : *Sémantique linguistique*. Paris : Larousse.

MC NAUGHT J., B. NKWENTI-AZEH, W. MARTIN AND E. TEN PAS, 1991 : « Eurotra 7-1 : Feasibility of Standards for Terminological Description of Lexical Items ». Luxembourg : *Commission of the European Communities, Directorate-General, Telecommunications, Information Industries and Innovation.*

MAINGUENEAU D., 1996 : *Les termes de l'analyse de discours*. Paris : Seuil.

MAINGUENEAU D., 1995. : « Présentation du numéro : Les analyses du discours en France ». *Langages* n°117. pp. 5-11.

MANNING C.D., À paraître. : «Probabilistic Syntax ». R. Bod, J. Hay and S. Jennedy (eds) : *Probabilistic Linguistics*. Cambridge : MIT Press.

MARTIN R., 1983 : *Pour une logique du sens*. Paris : PUF.

MARTIN R., 1990 : « La définition « naturelle » ». J. Chaurand et F. Mazière (eds) : *La Définition*. Paris : Larousse, collection Langue et Langage. pp. 86-95.

MEYER I., BOWKER L., ECK K., 1992a: «Cogniterm: An Experiment in Building a Terminological Knowledge Base». *Proceedings 5th EURALEX International Congress on Lexicography*, Tampere, Finland.

MEYER I., DOUGLAS S., BOWKER L., ECK K., 1992b : «Towards a new generation of terminological resources: An experiment in building a terminological knowledge base». *Proceedings 16th International Conference on Computational Linguistics*, Nantes. pp. 956-957.

MEYER I., MACKINTOSH K ., 1996 : « The Corpus from a Terminographer's Viewpoint. ». *International Journal of Corpus Linguistics*. vol.1, n°2.

MEYER I., 2000 : « Extracting Knowledge-rich Contexts for Terminography : A Conceptual and methodological Framework ». D.Bourigault, M.C. L'homme, C.Jacquemin (eds) : *Recent Advances in Computational Terminology*, John Benjamins. pp. 279-302.

MILNER J.C., 1976 : « Réflexions sur la référence ». *Langue française* n°30. pp. 63-73.

MILNER J.C. : *L'amour de la langue*. Paris : Seuil. 1978.

MORIN E., « Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique ». *TAL* (Traitement Automatique des Langues), vol.40, n°1. Paris : Université Paris VII. 1999, pp. 143-166.

NOAILLY M., 1990: *Le substantif épithète*. Paris : PUF.

NOAILLY M., 1996a : « Le vide des choses ». *Cahiers de Praxématique* n°27. pp. 73-90.

NOAILLY M., 1996b : « Dans le sens du fleuve : syntaxe et polysémie ». K.Fall, J.-M. Léard, P.Siblot (eds) : *Polysémie et construction du sens* ; Université Paul Valéry : Praxiling. pp. 25-40.

NORMAND S., 1999 : « Construction du sens dans un échange professionnel lié à la dégustation ». V.Delavigne, M.Bouveret (eds) : *Sémantique des termes spécialisés*. Publications de l'Université de Rouen, Collection Dyalang. pp. 119-127.

O.L.F, 1985 : *Vocabulaire systématique de la terminologie*.

OTMAN G., 1995 : *Les représentations sémantiques en terminologie*. Thèse en Sciences du langage de l'Université Paris IV Sorbonne.

OTMAN G., 1996 : *Les représentations sémantiques en terminologie*. Paris : Masson.

PARENT R., 1989 : « Recherche d'une synergie entre développement linguistique informatisé et systèmes experts : importance de la terminologie ». *Meta*, vol.34-3. pp. 611-614.

PEARSON J., 1998 : *Terms in Context,* Amsterdam and Philadelphia: John Benjamins.

PEARSON J., 2000 : « Une tentative d'exploitation bidirectionnelle d'un corpus bilingue ». A.Condamines (ed) : *Cahiers de Grammaire* n°25. pp .53-70.

PERY-WOODLEY M.-P., 1995 : « Quels corpus pour quels traitements automatiques ». *TAL* (Traitement Automatique des Langues), n°36 (1-2). pp. 213-232.

PERY-WOODLEY M.-P., 2000 : *Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle*. Habilitation à diriger les recherches, Carnet de Grammaire n°8. juillet 2000.

PERY-WOODLEY M.-P., REBEYROLLE J., 1998 : « Domain and genre in sublanguage text: definitional microtexts in three corpora ». *First International Conference on Language Resources and Evaluation,* Grenade (28-30 Mai 1998). pp. 987-992.

PICOCHE J., 1992 : *Précis de lexicologie française : l'étude et l'enseignement du vocabulaire*. Paris : Nathan Université.

PREVOST P., JACOBI D., 1994 : « Les cartes conceptuelles : outil cognitif, instrument de communication ou moyen de recherche ? ». *Didaskalia* n°5. Décembre 1994. pp. 119-124.

QUIRK R., GREENBAUM S., LEECH G, SVARTVIK J., 1985 : *A Comprehensive Grammar of the English Language*. London : Longman.

RASTIER F., 1987 : *Sémantique interprétative*. Paris : PUF.

RASTIER F., 1991 : *Sémantique et recherches cognitives*. Paris : PUF.

RASTIER F., 1995 : « Le terme : Entre ontologie et Linguistique ». *La Banque des Mots* n°7, Numéro spécial. pp. 35-64.

RASTIER F., 1996 : « Le défigement des expressions figées et leur interprétation ». K. Fall, J-M Léard, P. Siblot (eds) : *Polysémie et construction du sens*. Montpellier : Praxiling, Université Paul Valéry. pp. 17-24.

RASTIER F., 1998 : « Le problème épistémologique du contexte et le statut de l'interprétation dans les sciences du langage ». S. Bouquet (ed) : *Langages* n°129, *Diversité de la (des) science(s) du langage aujourd'hui.* pp. 97-11.

RASTIER F., 2001 : *Arts et Sciences du texte*. Paris : PUF, formes sémiotiques.

RASTIER F., CAVAZZA M., ABEILLE A., 1994. : Sémantique pour l'analyse, De la linguistique à l'informatique. Paris : Masson.

RASTIER F., MALRIEU J., 2001 : « Genres et variations morphosyntaxiques ». B. Daille, L. Romary (eds) : TAL, vol.42/2. pp. 547-577.

REBEYROLLE J., 1995 : *Polysémie dans les langues spécialisées*. Mémoire de DEA en linguistique. Université Toulouse Le Mirail.

REBEYROLLE J., 2000 : *Forme et fonction de la définition en discours.* Thèse de Sciences du langage, Université Toulouse Le Mirail.

REBEYROLLE J., TANGUY L., 2000 : « Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires ». A.Condamines (ed) : *Les Cahiers de Grammaire* n°25 : *Sémantique et Corpus*. pp. 153-174.

REY A., 1979 : *La terminologie, noms et notions*. Paris : PUF, Que sais-je ?.

REY A., 1990 : « Polysémie du terme *Définition* ». J. Chaurand et F. Mazière (eds) : *La Définition*, Paris : Larousse. pp.13-22.

RIEGEL M., 1990 : « La définition, acte du langage ordinaire ». J. Chaurand et F. Mazière (eds) : *La Définition*. Paris : Larousse. pp. 97-111.

RILOFF, E., 1996 : « Automaticaly Generating Extraction Patterns from Untagged Text ». *Proceedings of the Thirteenth National Conference on Artificial Intelligence* (AAAI-IAAA; vol.2. pp. 1044-1049.

ROUSSELOT F., FRATH P., OUESLATI R., 1996 : « Extracting Concepts and relations from corpora ». *Proceedings ECAI'96, 12th European Conference on Artificial Intelligence,*.

SABAH G, 1988 : *L'intelligence artificielle et le langage . volume 1, Représentation des connaissances*. Paris : Hermes.

SAGER J.C., 1990 : *A practical Course in Terminology Processing*. Amsterdam/Philadelphie : John Benjamins Publishing Company.

SAGER J.C., NDI-KIMBI A., 1995 : « The Conceptual structure of terminological definitions and their linguistic realisations ». *Terminology*, Vol2(1). pp. 61-81.

SAGER N., 1987 : « Computer Processing of Narrative Information ». N. Sager, C. Friedman, M.S. Lymann (eds) : *Medical language Processing. Computer Management of Narrative data*. Reading, MA : Addison-Wesley Publishing Company. pp. 3-21.

SAINT-DIZIER P., CONDAMINES A., 1990 : «An environment for the incremental acquisition of lexical semantic data », *Actes du Colloque Knowledge-based computer system*, 13-15 décembre 1990, Pune, Inde

SAMVELLIAN P., 1995 : *Les nominalisations en français : arguments sémantiques et actants syntaxiques*. Thèse de l'Université de Paris 7.

SAUSSURE DE F., 1982 : *Cours de linguistique générale*. Paris : Payot.

SEGUELA P., 2001 : *Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques.* Thèse d'informatique, Université Paul Sabatier, Toulouse.

SIBLOT P., 1996 : « La polysémie en question ». K.Fall, J.-M. Léard, P.Siblot (eds) : *Polysémie et construction du sens* ; Université Paul Valéry : Praxiling. pp. 41-62.

SKUCE D., MEYER I., 1991: « Terminology and Knowledge Engineering: Exploring a Symbiotic Relationship ». *6th International Worshop on Knowledge Acquisition for Knowledge-Based Systems* (Banff).

SLODZIAN M., 1995 : « La doctrine terminologique, nouvelle théorie du signe au carrefour de l'universalisme et du logicisme ». *ALFA (Actes de Langue Française et de Linguistique : Terminologie et langues de spécialité*, 7/8, Dalhousiana, Halifax, Nova Scotia, Canada. 1994-1995. pp. 121-136.

SLODZIAN M., 2000 : « L'émergence d'une terminologie textuelle et le retour du sens ». H.Béjoint et P.Thoiron (eds) : *Le sens en terminologie*. Lyon : PUL. pp. 61-85.

SMADJA F., 1993 : « Retrieving collocations from text : Xtract ». *Computational Linguistics*, 19(1). *Special Issue on Using Large Corpora*. pp. 143-177.

TIBERGHIEN A., 1994 : « Choix sous-jacents à la construction de représentation spatiale de concepts ». *Didaskalia* n°5. Décembre 1994. pp. 53-62.

VAN CAMPENHOUDT M., 1994 : *Un apport du monde maritime à la terminologie notionnelle multilingue ; Etude du dictionnaire du Capitaine Henri Paasch*. Thèse de linguistique de l'Université de Paris XIII.

VANDELOISE C., 1991 : « Autonomie du langage et cognition ». *Communications* n°53, *Sémantique Cognitive*. pp.69-102.

VERGELY P., 2002 : « Régularités linguistiques d'un langage opératif : le cas de la Navigation Aérienne ». *Actes du GLAT (Groupe de Linguistique Appliquée des Télécommunications)*. Mai 2002. pp. 59-70.

VERONIS J., 1998 : « A study of polysemy judgements and inter-annotator agreement ». *Programme and advanced paper of the Senseval workshop*. Hersmonceux Castle (England). http:///www.up.univ-mrs.fr/tildeveronis/

VICTORRI B., FUCHS C., 1996. : *La polysémie, construction dynamique du sens*. Paris : Hermes.

WERTH P., 1999 : *Text Worlds : Representing conceptual space in discourse*. London : Longman.

WIJNANDS P., 1989 : « Systèmes-experts et terminologie ». *Méta* 34-3. pp. 502-508.

WIJNANDS P., 1993 : « Terminology versus artificial intelligence ». H.B. Sonneveld, K.L. Loening (eds) : *Terminology, applications in interdisciplinary communication*. Amsterdam/Philadelphia : John Benjamins. pp. 165-179.

WOODS W.A., 1991 : « Understanding Subsumption and Taxonomy : A Framework for Progress ». J.Sowa (ed) : *Principles of Semantic Networks*, San Mateo California : Morgan Kaufman Publishers. pp. 45-94.

WUSTER E., 1981 : « L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et les sciences des choses ». G.Rondeau et H.Felber (eds) : *Textes choisis de terminologie*, GIRSTERM, Université de Laval, Québec. pp. 55-108.

ZWEIGENBAUM P., GRABAR N., 2000 : « Liens morphologiques et structuration de terminologie ». Actes d'IC'2000 (Ingénierie des Connaissances). Toulouse, Irit. pp.325-334.

Index

Annexe

Présentation des projets menés en lien avec une demande sociétale et cités dans le mémoire.
Chaque rubrique présente le contexte applicatif, les études qui ont été menées et les problématiques théoriques qui en ont émergé.

1. MMS (Matra Marconi Space) (1991-1993)

Demande : Aide à la rédaction, aide à la formation des nouveaux arrivés dans l'entreprise.

Contexte : Laboratoire mixte ARAMIIHS (Action Recherche et Application en Interface Homme-Système) laboratoire mixte du CNRS. Projet européen EUROLANG d'aide à la traduction.

Corpus : Volume inconnu. Deux sous-corpus ont été constitués, l'un français, l'autre anglais, sur le même thème : simulation segment-sol. Deux équipes corpus comportant chacune des linguistes terminologues et des experts.

Résultats obtenus :

- Pour l'entreprise : mise à disposition d'une BCT sur le domaine segments-sols simulation.
- Du point de vue de la recherche :
  - Ebauche d'une méthode linguistique d'élaboration de BCT,
  - Prise de conscience de l'importance d'une terminologie textuelle,
  - Définition d'un modèle de BCT (qui a été repris dans les outils GEDITERM, de l'Irit et, en partie, TERMINAE, du LIPN).
  - Création du groupe TIA (Terminologie et Intelligence Artificielle), visant à mettre en œuvre une interdisciplinarité autour de la constitution de terminologies à partir de corpus.

2. CNES 1 (Centre National d'Etudes Spatiales) (1995)

Demande : Contribuer à la mise au jour de points de vue, par une analyse linguistique
Contexte : Division Recherche et Développement du CNES ; réflexion sans recherche de résultats immédiats.

Corpus : volume inconnu. Corpus sur le projet DIODE, constitué de 3 sous-corpus
Résultats pour la recherche :

- Etablissement d'un lien entre polysémie et point de vue,
- Mise en place d'une méthode pour étudier la polysémie
- Prise de conscience du rôle de l'objectif d'étude pour construire le sens

3. EDF (1997-1998)

Demande : Améliorer l'accès au contenu d'un manuel (demande faite par le Service Qualité de la Direction des Etudes et Recherches d'EDF).
Réponse : Constituer une BCT pour rendre compte de la connaissance du manuel et repérer les incohérences.
Corpus : MOUGLIS (Méthodes et Outils de Génie Logiciel pour l'Informatique Scientifique) (le corpus est un texte fourni par l'entreprise), environ 60000 mots.
Résultats :

- Pour l'entreprise : réseau terminologique permettant de mettre au jour les principaux thèmes du manuel. Constitution d'un memento assistant la consultation de ce manuel.
- Du point de vue de la recherche :

- Elaboration d'une méthode précise de construction de BCT.
  - Réflexion sur la notion de marqueurs de relations,
  - Réflexion sur la définition en corpus (thèse de J. Rebeyrolle).

4. Système de Gestion Globale des Déplacements (SGGD), DDE (Direction Départementale de l'Equipement) de la Haute Garonne (1998-1999)

Contexte : 5 organismes impliqués dans la gestion de la circulation dans l'agglomération toulousaine ont décidé de collaborer plus étroitement.
Demande : Constituer un référentiel terminologique.
Réponse : Constitution de 4 BCT et fusion pour repérer les incompatibilités
Corpus : Environ 478000 mots, constitué de 4 sous-corpus (la police ayant refusé de donner le sien).
Résultats :

- Pour l'entreprise :
  - Mise à disposition d'une BCT rendant compte des différences d'usages dans les 4 organismes,
  - Réunions d'information sur les différences d'usage.

- Du point de vue de la recherche :

- Poursuite du travail sur les marqueurs de relations : mise au jour de marqueurs propres à un corpus,
- Constitution de l'outil GEDITERM par l'Irit, directement en lien avec les besoins identifiés par les linguistes ; cet outil permet de rendre compte des résultats de l'analyse terminologique,
- Travail sur l'ellipse terminologique (thèse de M.-P. Jacques).

5. CENA (Centre d'Etude de la Navigation Aérienne) (1999-2003)

Demande : Etudier un corpus de dialogues concernant le dysfonctionnement technique afin d'une part de repérer d'éventuelles difficultés de communication et, d'autre part, d'étudier l'impact d'une situation simulée sur les échanges

Réponse : Constituer une grammaire de l'expression du dysfonctionnement (thèse de P. Vergely).

Corpus : Environ 110000 mots, organisé en trois sous-corpus :

- situation réelle 1995,
- situation simulée, 2000
- situation réelle, 2001.

Le corpus varie ainsi selon deux critères : la période de rédaction et la situation (réelle vs simulée).

Résultats pour la recherche :

- Réflexion sur la spécificité des corpus oraux : difficulté de traitement par les outils, difficultés d'interprétation,
- Vérification de la stabilité des moyens d'expressions du dysfonctionnement lorsque des éléments extra-linguistiques varient.

6. CNES 2 (2002-2004)

Demande : Proposer des méthodes linguistiques pour repérer l'évolution des connaissances dans le temps.

Contexte : Une difficulté identifiée pour certains projets : entre le moment où certaines sondes sont lancées et le moment où elles atteignent leur objectif, une longue période (plus de 10 ans) peut se passer. Pendant ce délai, les équipes peuvent changer et les connaissances évoluer. Il faut alors essayer de trouver ces ruptures dans la connaissance.

Corpus : en cours de constitution.

Réponse :

- Etude de l'évolution de la forme des termes (phénomènes d'ellipse, de siglaison…)
- Etude de l'évolution de la mise en réseau (l'hypothèse forte qui sera testée est celle d'une stabilité des marqueurs de relations et des relations elles-mêmes ;
- Etude de la complémentarité entre approche statistique et approche linguistique, en particulier en ce qui concerne le phénomène de la polysémie.