Les bases théoriques du groupe toulousain « Sémantique et Corpus » : ancrages et perspectives

Anne Condamines*

Cet article présente les principaux courants de recherche dans lesquels s'inscrivent les travaux de l'opération « Sémantique et Corpus ». Le rôle des corpus est examiné selon quatre types d'approches : la sémantique « essentiellement » linguistique, la sémantique anglo-saxonne, la terminologie et l'informatique (TAL et IA). La présentation, bien que clairement personnalisée par la vision de l'auteur, vise à dégager les principales questions qui fédèrent les travaux de l'opération.

This paper presents the most important research trends in which the studies of the group « Semantics and corpora » are situated. The role of corpora is examined from the point of view of four approaches : « essentially » linguistic semantics, anglo-saxon semantics, terminology and data processing (NLP and AI). The presentation, although personalized by the view of the author, aims to bring out the main issues federating the work of the group.

* Equipe de Recherche en Syntaxe et Sémantique (UMR 5610 CNRS).

Cahiers de Grammaire 25 (2000), « Sémantique et Corpus », pp. 5-28

1. Introduction

Avant d'opter pour son nom actuel, « Sémantique et Corpus », en 1999, le groupe toulousain constitué en 1993 autour d'Andrée Borillo et de moi-même a pris successivement les noms suivants : « Langages spécialisés et terminologie », « Traitement automatique en langues de spécialité : Terminologie et organisation conceptuelle », « Traitement automatique des langues : terminologie et organisation conceptuelle ». Cette évolution des appellations est due à une évolution de la thématique, en grande partie en lien avec l'arrivée de l'un ou l'autre membre, recruté ou muté comme chercheur ou enseignant-chercheur. Aussi, si l'axe majeur qui nous réunit dans cette opération semble maintenant à peu près stabilisé, pour en faire comprendre l'originalité, il m'a semblé nécessaire de montrer comment il a pu se constituer à partir des apports historiques de chacun des membres. C'est par la présentation des principaux courants de recherche dont nous nous inspirons et des questions qui se posent en relation avec ces travaux que s'élabore cette présentation. Cette présentation n'est pas neutre ; pour chacun des travaux évoqués, je montrerai ce qui me semble intéressant à retenir et ce qui pose question pour exposer en conclusion les axes autour desquels s'organise notre réflexion.

Quatre directions fondamentales, un peu hétérogènes, seront ainsi examinées : les corpus dans la perspective d'une sémantique « essentiellement » linguistique (deux exemples français, les travaux de G. Kleiber et ceux de F. Rastier), les corpus dans les travaux anglo-saxons, qui s'inscrivent dans une sémantique reliée à la sociologie ou la psychologie, les corpus dans les études en terminologie, enfin, les corpus et le traitement automatique. J'essayerai de montrer en quoi ces divers travaux, malgré leur diversité et parfois leurs contradictions nous conduisent à une réflexion fructueuse et renouvelée sur le rôle des corpus dans l'étude du sens en linguistique.

2. Les corpus dans la perspective d'une sémantique « essentiellement » linguistique : deux exemples de travaux français

De manière générale, les travaux francophones en sémantique se situent surtout dans la perspective d'une sémantique essentiellement linguistique. Conséquence ou non de cette vision, les études sur l'utilisation des corpus en sémantique sont rares dans cette communauté. Cependant, même lorsqu'ils ne les utilisent pas, les sémanticiens prennent souvent position par rapport au rôle qu'ils attribuent au contexte dans le sens des mots. Les travaux de deux sémanticiens français inspirent particulièrement les réflexions de notre opération : G. Kleiber et F. Rastier.

2.1. La sémantique référentielle de G. Kleiber

Les travaux de G. Kleiber sont très souvent mentionnés par les chercheurs de l'ERSS (il était d'ailleurs l'un des membres du jury de ma thèse en 1990). Il est vrai que la production de cet auteur est importante et qu'elle s'intéresse à de nombreux phénomènes sémantiques : la polysémie, la généricité, les relations sémantiques… Sa rédaction est toujours claire et les nombreux exemples souvent convaincants : il joue ainsi un rôle très stimulant dans la réflexion en sémantique. Pourtant, cet auteur se situe nettement dans une approche référentielle, c'est-à-dire une approche qui revendique comme objet d'étude le lien entre la langue et l'extra-linguistique, (« la sémantique n'a aucun sens si elle n'est pas tournée vers (ce que nous croyons être) la réalité » (Kleiber 1999, 11)). Il s'agit, dans le cas de cet auteur d'une réalité « expériencée » (ibidem, 21), ou « modélisée » (ibidem, 22). Une sémantique ainsi conçue est stable parce que la modélisation est « intersubjectivement partagée », ce qui est un « facteur de stabilisation et d'objectivation » (p. 22). Un article rédigé en 1997 (*Langages*, n° 127) et repris tout récemment (Kleiber 1999) montre à quel point l'auteur éprouve le besoin de justifier son approche tant elle lui paraît menacée par celles des constructivistes. Cet article est tout à fait intéressant, d'abord parce qu'il fait le point sur les différentes approches existant en sémantique, ensuite parce qu'il rappelle, de façon opportune, que l'un des objectifs de la linguistique est de repérer des régularités dans la langue. Il y a pourtant deux points au moins sur lesquels l'expérience des corpus ne semble pas (ou pas complètement) compatible avec l'approche défendue par G. Kleiber.

A) D'une part, l'analyse de corpus, particulièrement de corpus spécialisés, met souvent au jour des fonctionnements qui ne correspondent pas à l'intuition que l'on peut avoir, ou d'autres qui, lorsqu'ils sont dénombrés, apparaissent comme secondaires par rapport à l'intuition première ce qui peut donner une analyse qui, sous couvert d'objectivité, ne rend pas compte du dynamisme des fonctionnements. Deux exemples :

- Dans un corpus fourni par le CNES, sur la base d'une analyse de contextes, nous avons pu mettre au jour qu'un mot comme *satellite* ne renvoyait pas toujours à un élément fixe et définissable *a priori,* qui aurait pu correspondre à une définition encyclopédique. Au contraire, nous avons montré que ce mot avait six types de fonctionnements syntaxico-sémantiques, non prédictibles, ni par nous-mêmes ni par les experts du CNES. En revanche ces six types de fonctionnements ont pu être corrélés avec des points de vue propres au CNES, en lien avec les différentes fonctions assurées par le satellite (le satellite comme corps

artificiel, comme mobile, comme plate-forme, comme véhicule, comme relais, comme hôte)[1].

- Autre exemple, lorsqu'une de nos collègues de l'opération « Sémantique du temps et de l'espace » a voulu s'aider du corpus du Monde Diplomatique pour rechercher des verbes de déplacement, elle s'est aperçue que dans la majorité des cas, ces verbes étaient utilisés de manière métaphorique (il s'agissait de mouvements d'argent, ce qui n'est pas étonnant, *a posteriori*, étant donné la nature du corpus). Cela n'invalide pas la pertinence de la description mais nécessite qu'elle soit replacée dans la perspective d'un usage réel de la langue.

B) D'autre part, il arrive souvent que les conclusions proposées par les tenants de l'approche référentielle et constitués sur la base d'exemples forgés, soient à revoir dès que l'on examine des corpus dans lesquels apparaissent des exemples correspondant à des cas *a priori* considérés comme non-valides. Prenons l'exemple du déterminant *chaque* que l'on considère comme inapte à exprimer la généricité, (le test du détachement et de la reprise par *c'est* ne semble pas fonctionner (**chaque N, c'est*). Il se trouve que dans sa thèse sur la définition (*Forme et fonction de la définition en discours*, Thèse de l'Université Toulouse Le Mirail, 2000), J. Rebeyrolle fait part d'un exemple avec *chaque* qui exprime la généricité : *On appelle terrassettes de petits gradins de quelques centimètres séparés par des replats dont la largeur est du même ordre. Chaque gradin est un plan de cisaillement entre deux paliers qui s'affaissent*.

C'est bien finalement l'idée de la totale fiabilité de l'approche introspective qui est à mettre en question. Même si l'on partage l'idée qu'il existe, en sémantique, des éléments stables qui peuvent fonctionner de manière régulière en corpus (c'est mon point de vue), le fait que cette régularité soit due à notre « commune nature d'homme » (Kleiber 1981, 27), est contestable parce qu'elle place la perception (même modélisée par la langue) comme prioritaire par rapport à la compétence linguistique acquise lors d'échanges langagiers contextualisés et sans cesse renouvelés. Vue dans son fonctionnement en contexte, la stabilité de la langue est toujours relative : le travail en corpus, surtout en corpus spécialisé, oblige à accepter ce phénomène. Ce qui alors peut guider la réflexion, c'est que s'il y a régularité, elle est observable au niveau d'un corpus, et peut être généralisable à l'ensemble des corpus qui ont les mêmes caractéristiques (domaine, genre, …).

Au bout du compte, les travaux de sémantique « introspective » sont souvent très utiles : ils mettent sur la voie de phénomènes tout à fait intéressants et les descriptions proposées sont souvent pertinentes (il ne s'agit

[1] Condamines A., Rebeyrolle J. : « Point de vue en langue spécialisée », META n°42-1, 1996.

pas par exemple de refuser les descriptions qui ont été faites des déterminants génériques)… à condition que l'on accepte de les réexaminer, à la lumière de l'analyse de corpus ; à condition au fond d'accepter qu'apparaissent des phénomènes qu'*a priori* nous n'aurions pas cru réalisables et qui semblent néanmoins intéresser fondamentalement la question du sens, et pas seulement des effets « locaux ».

2.2. La sémantique interprétative de F. Rastier

C'est de manière beaucoup plus récente que les travaux de F. Rastier sont utilisés comme une des références dans l'équipe. Cela correspond à l'arrivée de D. Bourigault puis de L. Tanguy dans l'équipe, tous les deux fortement nourris de la lecture des écrits de Rastier. Ces travaux s'inscrivent dans le courant de la sémiotique textuelle initié par Greimas (Greimas 1966) ; ils revendiquent à la fois une filiation avec Saussure, dont ils reprennent l'idée de système et de différences, et une vision nouvelle par le biais de la sémantique textuelle. Certaines des propositions de Rastier sont séduisantes et beaucoup sont partagées par les membres de l'opération « Sémantique et Corpus » :

- L'unité d'analyse est le texte : « L'essentiel demeure de pouvoir traiter par une théorie unifiée des paliers sémantiques du morphème, de l'énoncé, et du texte… Dans la perspective choisie, le texte demeure toutefois le palier primordial » (Rastier 1987, 10).
  La pratique de l'analyse de corpus fait en effet prendre conscience de l'importance de ce « palier du texte » et de la nécessité de prendre en compte, à un moment ou l'autre de l'analyse sémantique, ce qui fait l'homogénéité et la cohérence d'un texte.
- Le contexte fait partie de la construction du sens, « le contexte linguistique et non linguistique est, en tant qu'interprétant, constitutif du message. On doit convenir du moins que, globalement comme localement, le texte en reçoit de multiples déterminations » (Rastier 1991, 13). Dans cette perspective, il est pertinent d'identifier des corrélations entre des caractéristiques extra linguistiques du texte et des fonctionnements linguistiques (cf. travaux de Biber ci-dessous).
- Dans un domaine plus spécialisé, la construction d'une terminologie consiste en différentes opérations qui permettent de passer, à partir d'un corpus, du mot au terme et du terme au concept (Rastier 1995).

Sans doute, une des principales qualités des travaux en sémantique interprétative est de proposer une vision unifiée des recherches qui proposent de prendre les corpus comme objet d'étude et de situer ces travaux dans le domaine de la sémantique. En effet, l'étude des corpus sert de base à la sociolinguistique, à la dialectologie, à l'analyse littéraire, mais aussi à une grande partie des travaux en terminologie et il est tout à fait pertinent de

vouloir identifier les points communs entre ces différentes disciplines.

Sur certains points toutefois, les travaux de Rastier sont moins convaincants (ou moins explicites) :

- Le rôle du contexte textuel ne semble pas très constant. On ne mesure pas toujours très bien ce qui est considéré comme stable, quasiment inhérent au sens d'un mot et ce qui, au contraire, va être complètement dépendant du contexte, comme le montre les deux extraits suivants

  « De la même façon qu'on ne peut s'empêcher d'entendre, on ne peut s'empêcher de comprendre. Plus précisément, rappelons Posner et alii, un mot "peut activer ses **codages internes, visuel, phonologique, et même sémantique**[2], sans que la personne ait à y prêter attention"» (Rastier 1991, 213).

  « Après Schleiermacher on peut soutenir la thèse que toute occurrence sémantique est un hapax, et la compléter en affirmant que tout type n'est qu'une reconstruction… » (Rastier 1991, 114).

  Un certain nombre de chercheurs se sont particulièrement engagés dans l'exploration de cette voie (toute occurrence est un hapax) et sur ce point, il y a des divergences assez nettes dans l'opération « Sémantique et Corpus ».

- Le problème de la pluralité possible d'interprétations ne va pas sans poser de questions. Même si les diverses interprétations sont censées être guidées par le concept d'isotopie (« l'infini prétendu des lectures possibles ne permet pas d'admettre qu'un texte comporte un nombre infini d'isotopies » (Rastier 1987, 106)), on n'a guère de précisions sur les éléments qui vont pouvoir valider la pertinence des isotopies construites. Ceci amène une autre réserve qui est celle de la méthode d'analyse.

- Le point le plus délicat de l'approche proposée par la sémantique interprétative est certainement le problème de la mise en place d'une méthode d'analyse. Si l'empirisme revendiqué semble indissociable de l'analyse de corpus, l'utilisation de l'intuition, également revendiquée (« …l'intuition peut refléter une réalité objective » (Rastier 1987, 107)) pose, elle, plus de questions. Non parce que l'analyse linguistique devrait se passer d'intuition, mais parce que, à mon avis, une bonne partie de la recherche en linguistique devrait avoir pour but d'expliquer quelle intuition est mise en œuvre pour parvenir à telle ou telle interprétation et également, si elle pourrait être mise en œuvre de manière plus systématique. De fait, les livres et articles de Rastier contiennent peu d'exemples réels d'analyse (contrairement à ceux de Kleiber, ce qui est pour le moins paradoxal) et, lorsqu'il y a des

[2] C'est moi qui souligne.

analyses, elles portent presque toujours sur des textes littéraires qui, sans doute, se prêtent plus facilement à des interprétations multiples.

En fait, peu de linguistes français se situent clairement dans le champ de la théorie en sémantique textuelle et, il est un peu gênant que la théorie rastierienne soit la seule à occuper ce terrain. Cette théorie a séduit de nombreux chercheurs en linguistique, et même en IA, bien qu'elle ne soit pas d'un abord facile, à la fois parce que l'auteur, dans sa grande culture, fait appel à un grand nombre de références, et parce que les notions en jeu ne se laissent pas maîtriser aisément, y compris étant donné la manière dont elles sont formulées. Si bien que l'on peut facilement se réclamer de la sémantique différentielle ou de la sémantique interprétative sans être très sûr de se situer dans la théorie décrite par Rastier. Aucun de nous, dans l'opération, n'utilise de manière exhaustive l'appareillage de la sémantique interprétative. En revanche, beaucoup des grands principes élaborés par elle sont considérés comme pertinents et inspirent nos travaux.

3. Les corpus dans la perspective d'une sémantique reliée à la sociologie ou le psychologie : le courant anglo-saxon

Une caractéristique des travaux anglo-saxons qui ont les corpus pour base d'étude est qu'ils font appel, pour établir les bases de leurs analyses linguistiques, à des concepts qui proviennent de disciplines connexes des sciences humaines. Ainsi, si pour les linguistes post-saussuriens, même pour ceux qui s'intéressent à l'analyse de textes comme Greimas, la langue est un objet d'étude à part entière, les anglo-saxons font très tôt appel à la sociologie ((Bloomfield 1935), (Firth 1969)) ou à la psychologie ((Harris 1951), (Chomsky 1971)) pour ancrer leurs hypothèses. Les deux courants qui sont les plus nettement présents dans l'opération sont d'une part la théorie des sous-langages et d'autre part les analyses de corpus pratiquées par Biber.

3.1. Le distributionnalisme de Z. Harris

Les premiers travaux faisant appel à la théorie harrisienne dans l'ERSS sont ceux d'A. Borillo dans les années 70.

Dès le milieu du XX^{ème} siècle, le principal apport de Harris est d'avoir systématisé l'analyse des distributions des éléments afin de dégager la grammaire à l'œuvre dans un corpus. Un grand nombre d'opérations, de régularités, probablement mises en œuvre de manière intuitive par les comparatistes du XIX^{ème} siècle ont ainsi été mises au jour et explicitées. L'objectif était, par l'application systématique d'un ensemble d'opérations, sur la base de similitudes de distributions, aux niveaux phonologique et morphologique, de constituer des classes de fonctionnement qui permettaient de construire le noyau dur de la grammaire.
L'apport de Harris à la linguistique est double :

- d'une part, il fut un des premiers à prendre en compte de manière systématique l'analyse de corpus,
- d'autre part, il a théorisé la réflexion sur le rôle des distributions d'un mot.

De nombreux sémanticiens se sont élevés contre la théorie harrissienne, en tout cas, contre le postulat sur lequel elle est construite (pour n'en citer que quelques uns : (Benveniste 1966), (Lyons 1980), (Rastier 1991) …). En effet, la vision de Harris est clairement behavioriste….

> « It is emprirically 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).

Ainsi, comme le remarque Dachelet, Harris ne traite pas le sens mais l'information :

> « …comme on le voit, le concept de sémantique, de sens s'évanouit dans la conception harrissienne au profit du concept minimaliste - mais crucial - d'information. » (Dachelet 1994, 251).

Dans une telle perspective, sont considérées comme équivalentes une phrase à l'actif et au passif ou encore une phrase avec topicalisation ou non…, ce qui bien sûr n'est pas recevable pour un sémanticien. On peut peut-être penser aussi que, comme le note Dachelet, l'hypothèse harrissienne conduit à une vision positiviste de la langue, qui établit un lien direct entre les mots et les choses… tout particulièrement dans des sous-langages de domaine où cette équivalence est tentante (elle est d'ailleurs fréquemment établie dans les travaux en terminologie). L'interprétation des classes élaborées par l'analyse distributionnelle est ainsi faite sur la similitude de rôle informationnel auquel on attribue un statut sémantique à partir d'une interprétation.

Cette vision behavioriste du sens est sans doute l'élément qui donne le plus de prise à la critique de la théorie harrissienne. Dans une moindre mesure, de nombreuses questions sont posées par le distributionnalisme à la Harris :

- La notion des sous-langages suppose un découpage déterminé a priori, à l'intérieur duquel un corpus est sélectionné, qui permet d'identifier l'ensemble des phrases possibles de la grammaire de ce sous-langage. Le corpus a ainsi un rôle représentatif extrêmement important et, par principe, il n'y a pas de possibilité de création en dehors de la grammaire élaborée. Notons que cette question de la clôture du corpus et de son éventuel rôle représentatif se pose à tout linguiste qui prend

pour matériau un corpus. Il s'agit de savoir si les résultats seront valables seulement pour ce corpus ou bien s'ils seront généralisables à un ensemble de corpus. C'est la deuxième option qu'a choisie Harris, sans donner beaucoup de pistes sur la façon de construire un corpus représentatif : seule la notion de domaine est mentionnée sans être discutée.

- L'analyse se limite, au moins dans les premiers travaux de Harris, à la prise en compte du contexte immédiat, la phrase étant considérée comme l'unité d'analyse. Or, ce choix a au moins deux limites. D'une part, l'interprétation d'un mot peut nécessiter le recours à des éléments situés bien avant (voire bien après) dans le texte. D'autre part, le fait de considérer sur le même plan toutes les distributions d'un mot entraîne qu'on ne tient pas compte de la position de chacun des contextes dans le texte. Or, il n'est pas indifférent que tel contexte apparaisse par exemple dans une introduction et tel autre dans en conclusion. Les spécialistes de l'analyse littéraire savent bien aussi combien les phrases qui débutent et qui closent un roman sont importantes…

Le même élément qui a entraîné les critiques de la part des sémanticiens, le recours à la seule forme, a nettement inspiré les informaticiens pour traiter automatiquement la langue. En effet, l'idée proposée par Harris de partir de distributions de formes pour construire une grammaire avait de quoi séduire les informaticiens-linguistes, particulièrement dans le cadre d'une grammaire à états finis : elle a été par exemple reprise par les travaux du LADL sur le lexique-grammaire. Même dans des projets qui visent moins une construction totale du sens sur la seule base de l'étude des combinaisons de formes, on retrouve le rôle important attribué aux distributions. C'est le cas des travaux qui visent à construire des « cliques » lexicales : éléments rassemblés dans des classes censées être sémantiquement homogènes, sur la base de parentés distributionnelles (citons par exemple (Habert *et al* 1996)). En linguistique, même si peu de chercheurs se réclament encore de la théorie harrisienne, l'analyse distributionnelle, plus ou moins directement inspirée par les travaux de Harris, reste la base de l'analyse de corpus. On voit mal en effet quelle autre approche pourrait être mise en œuvre dans cet objectif. Cependant, il reste à s'interroger sur les rapports entre le sens tel que le conçoivent les sémanticiens et l'information qui correspond plus à l'apport de la théorie des sous-langages. Ce questionnement doit se faire tout particulièrement dans le cas des corpus fournis par des entreprises. La demande sociale en effet semble plutôt relever du traitement de l'information (par exemple : indexation, recherche d'information…) que du traitement du sens… Et on peut se demander si, au fond, la recherche du contenu informationnel, en lien avec une tâche particulière, ne correspond pas à une des interprétations sémantiques possibles pour un texte (interprétation certes importante mais pas unique).

C'est peut-être avec cette hypothèse que peut s'envisager de la manière la plus fructueuse une collaboration avec nos collègues informaticiens (voir ci-dessous).

3.2. L'analyse de corpus selon Biber

L'analyse de corpus selon Biber s'inscrit dans les travaux d'anglo-saxons qui, depuis bien plus longtemps qu'en France, ont considéré les corpus comme des objets d'étude à part entière, en particulier pour l'analyse lexicale. A son arrivée dans l'équipe, M.-P. Péry Woodley a introduit ces travaux, et tout particulièrement ceux de Biber.

L'analyse que propose Biber prend en compte des paramètres sociolinguistiques. Toutefois, et c'est ce qui le démarque d'une vision complètement sociologique de la langue comme celle de Firth (pour qui la langue est un fait sociologique comme un autre), l'ambition de Biber consiste à constituer une linguistique de l'usage et pour cela, de considérer les corrélations qui peuvent exister entre le genre du corpus (établi à partir d'éléments extra-linguistiques) et le type du texte (établi à partir d'éléments linguistiques).

> « 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 » (Biber 1988, 70).

Des rapprochements de textes se font ainsi sur la base de corrélations de traits linguistiques (par exemple *passé + 3e personne + aspect accompli…*) qui conduisent à la mise en évidence de six dimensions qui font éclater les genres initialement identifiés ; en effet, les textes sont à nouveau classés (rapprochés ou éloignés) en fonction de leur coordonnées sur ces dimensions. Des rapprochements inattendus se font jour, ainsi que des éloignements ; ainsi, si l'on tient compte de certains fonctionnements linguistiques, les textes oraux et écrits n'apparaissent pas aussi aisément distinguables qu'on aurait pu le supposer *a priori*. L'apport de ce type de travaux est double. D'une part, il s'agit sans doute d'une des premières entreprises visant à considérer systématiquement les corpus comme des usages de la langue en lien avec des éléments « sociologiques ». Bien plus que chez Harris, qui s'en tenait à la notion de sous-langage de domaine, on trouve, chez Biber, la volonté de prendre en compte la réalité extra-linguistique des corpus en caractérisant de façon précise les corpus, avant toute analyse linguistique. Par rapport aux travaux des sociolinguistes, pas très éloignés dans leur objectif, on trouve chez Biber un travail sur des corrélations de phénomènes linguistiques, élaborés en dimensions essentiellement lexico-syntaxiques alors que les

sociolinguistes s'intéressent majoritairement à des éléments phonologiques, et constatent parfois l'échec de l'approche sociolinguistique pour des phénomènes syntaxiques (cf : « On dira sans doute que cette reconnaissance [envers Labov] devrait s'arrêter aux études de syntaxe, pour lesquelles il semble bien aujourd'hui que les pistes ouvertes par Labov conduisent à des impasses » (Deulofeu 1992, 66))[3], examinés sur une seule dimension (par exemple, la variation entre [θ] et [t] à New York).

D'autre part, les travaux de Biber mettent l'accent sur une approche statistique multidimensionnelle des phénomènes. En effet, il est plutôt rare de pouvoir dire que dans tel corpus on ne trouve jamais tel phénomène linguistique. En revanche, on peut dire que ce même phénomène apparaît plus souvent dans tel corpus que dans tel autre. La notion de dimension est claire sur ce sujet : il s'agit de situer des textes sur un axe. Dans une perspective strictement linguistique, mais aussi pour les traitements en TAL et, de manière générale pour les traitements en ingénierie linguistique où l'efficacité prime, ce genre de constatations est très important. Comme je l'ai déjà souligné, c'est sans doute le principal défaut que l'on peut reprocher à une approche purement introspective : elle ne permet pas (ou très peu) de rendre compte de la répartition des fonctionnements : toutes les intuitions que l'on peut avoir sont considérées comme équivalentes.

La question la plus fondamentale que pose l'approche biberienne est celle de la portée des résultats et de leur généralisation. Même s'il s'en défend à certains moments, la tentation est forte chez Biber de généraliser la portée de ces résultats à la langue et de considérer que, par son approche, il a décrit l'ensemble du système :

> « Although this study began as an investigation of speech and writing, the final analysis presents **an overall description of the relations among texts in English[4],** and, it can therefore be used as a basis for the investigation of several related issues. » (Biber 1988, 200).

Une telle généralisation ne serait possible qu'avec la certitude que tous les critères de constitution de genre ont été pris en considération et c'est une idée sous-jacente dans le travail de Biber. Or, la pratique de l'analyse de corpus nous conduit à penser que la définition exhaustive de ces critères peut être problématique tant ils peuvent varier en fonction du type d'étude que

[3] Remarquons que l'élément majeur sur lequel butent les sociolinguistes pour utiliser la thèse variationiste de Labov en syntaxe est le fait qu'elle suppose que les variations constituent « different ways of saying the same thing », c'est-à-dire qu'elles font intervenir des équivalences sémantiques. Or, si ces équivalences sont facilement acceptables en phonologie, elles le sont beaucoup moins si on se place du point de vue de la syntaxe. On retrouve une question que j'évoque plusieurs fois au cours de cet article : peut-on accepter qu'il y ait changement de forme sans changement de sens ?

[4] C'est moi qui souligne.

l'on veut faire et il paraît très difficile de dresser un inventaire complet de ces critères, en tout cas pour le moment. Il est vrai aussi que, dans la plupart des projets dans lesquels nous sommes impliqués, nous travaillons sur des corpus beaucoup moins volumineux que ceux de Biber (qui a mené son étude sur un corpus constitué de 481 textes représentant 23 genres), il faut donc caractériser beaucoup plus finement les sous-corpus éventuels pour faire émerger des distinctions. En revanche, et c'est ce qui peut être conservé de l'analyse « à la Biber », il est évident que le premier travail à faire est celui d'une caractérisation du corpus à étudier, *a priori*, visant à prendre en considération le maximum d'information sur l'objectif de sa rédaction, le statut des rédacteurs, des destinataires…

Finalement, là où on croyait tenir un élément stable (les critères extra-linguistiques de constitution des corpus), il s'avère qu'il faut aussi envisager la variabilité qui intervient surtout en fonction du rôle que l'on veut faire tenir au corpus.

4. Les corpus dans une discipline longtemps considérée comme marginale pour la sémantique : la terminologie

C'est par le biais de travaux que nous avons menés sur la terminologie (lors de notre participation au projet européen Eurolang), dans le cadre du laboratoire ARAMIIHS (1991-1993), que la problématique de l'analyse de corpus a émergé dans l'ERSS, ce qui a conduit à la création d'une opération sur cette thématique.

Le constat que j'ai pu faire à ce moment-là, sur les travaux en terminologie était assez inquiétant : la plupart d'entre eux avaient une visée strictement applicative, le plus souvent dans la perspective de la traduction. Aujourd'hui encore, la plupart des enseignements en terminologie se font dans le cadre de départements de LEA (Langues Etrangères Appliquées). Quelques travaux plus linguistiques existaient, soit en lexicologie scientifique et technique, soit sur les sous-langages. Mais, eux, en revanche, ne prenaient pas toujours en compte la dimension applicative, pourtant quasiment incontournable dans des travaux en terminologie. Dans leur majorité, les travaux existants (et c'est encore en partie vrai aujourd'hui) faisaient preuve d'une grande ignorance (délibérée ou non) du fonctionnement linguistique[5]. En effet, les langues spécialisées étaient considérées comme ayant un fonctionnement particulier, éloigné des pollutions prétendument propres à la langue générale (ambiguïtés, polysémies…). Ces travaux s'inscrivaient dans la suite des travaux de celui qui est considéré comme le fondateur de la terminologie dans les années 30 : Eugen Wuster, qui avait une vision très idéalisée des langues spécialisées.

[5] Quelques auteurs cependant se plaçaient clairement dans une perspective linguistique, comme P. Lerat ou J. Sager, mais le rôle des corpus dans la constitution de terminologie n'était pas problématisé.

Dans ce type de travaux en effet, le sens est dangereux car on ne peut le contrôler : « 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). Du coup, et c'est encore vrai aujourd'hui, une grande partie de l'énergie (et des capitaux) est employée à constituer des normes, tout particulièrement dans les pays où une langue semble menacée par l'hégémonie d'une langue dominante (par exemple, au Québec (Montréal) ou en Catalogne (Barcelone)). Beaucoup de travaux en terminologie sont ainsi issus de la traduction. Notons cependant qu'à côté de travaux normatifs, se développent aussi dans ces pays des travaux plus théoriques, qui bien souvent désormais s'ancrent dans une réflexion linguistique (voir dans ce numéro les articles de J. Pearson et M.-C. L'Homme).

Les travaux se basant sur une approche « à la Wuster » mettent la priorité sur les domaines, définis *a priori*, et sur les concepts, également définis *a priori* et considérés comme stables. Dans les cas extrêmes, les termes ne sont plus que des étiquettes de concepts quasiment vidés de leur contenu sémantique (exemple, cette définition de l'OLF (Office de la Langue Française) : « 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** ». Dès lors, l'étude de la terminologie s'éloigne très nettement d'une perspective linguistique qui tient compte de la réalité des faits et non de postulats, et l'étude des corpus joue les trouble-fête. On préfère demander à des experts de dévoiler les concepts du domaine, qu'ils sont censés maîtriser parfaitement.

Seules, quelques catégories de travaux ont pris en compte les corpus dans l'étude terminologique (par exemple ceux de Kocourek 1991). Je m'attarderai plus particulièrement sur les travaux de R. Kittredge, sur les sous-langages, ou ceux de L. Guespin puis de F. Gaudin (Rouen) en socioterminologie. Ces travaux sont beaucoup plus proches de la réalité linguistique du fonctionnement terminologique que ceux précédemment mentionnés.

Les travaux en socioterminologie ont d'emblée pris en compte la dimension sociale de l'analyse terminologique et le nécessaire recours à l'étude de corpus :

> « …Ceci débouche sur la nécessité d'une analyse de discours spécialisés vers laquelle doit s'orienter la socioterminologie. Pour ce faire, il importe d'exploiter les acquis de la lexicologie structurale pour mener des analyses fines sur des corpus spécialisés » (Gaudin 1993, 180).

Les travaux sur les sous-langages présentent un intérêt certain ; bien qu'ils ne soient pas particulièrement orientés vers l'analyse lexicale mais vers la mise au jour de grammaires spécifiques à ces sous-langages, ils peuvent contribuer au repérage d'éléments lexicaux et de relations entre ces éléments :

> « The sublanguage grammar is more than just a linguistic characterization of the texts. The lexical classes and the hierarchical relations between the classes usually reflect the accepted taxonomy which the specialized field of knowledge imposes on the objects of its limited domain of discourse. » (Kittredge 1982, 112).

Malheureusement, malgré leur qualité, et malgré les revendications de leurs auteurs, ces travaux sont souvent considérés comme marginaux par les sémanticiens, au mieux comme relevant d'une linguistique appliquée, nécessairement moins noble qu'une linguistique théorique. Or, on peut considérer que l'analyse de corpus spécialisés pour en identifier la terminologie n'est pas autre chose que de l'analyse de corpus, tout particulièrement examinés dans leur dimension sémantique. Corpus spécialisé ou non, le même type de questions se pose :

- critères de constitution du corpus,
- objectif de l'étude,
- généralisation des résultats,
- lien entre stabilité et variation.

Dans les travaux que nous menons à Toulouse, la nature spécialisée des corpus n'est qu'une des caractéristiques, certes importante, des corpus. Mais c'est une caractéristique qui interpelle la linguistique sur des questions qu'elle a soigneusement évitées comme la prise en compte de l'objectif de l'étude dans l'interprétation sémantique ou même le rôle social de la linguistique. On retrouve peut-être là le lien avec le réel que G. Kleiber appelle de ses vœux et pourtant, cette façon de voir le réel éloigne encore un peu plus d'une vision stable du sens puisqu'elle suppose de prendre en compte, pour l'interprétation sémantique, non seulement le co-texte et le contexte du (des) rédacteur(s) et du (des) lecteur(s) mais aussi le contexte de l'interprétation (on retrouve cette idée dans Rastier (1994) : « L'interprétation elle-même est située. Elle prend également place dans une pratique sociale, et obéit aux objectifs définis par cette pratique. Ils définissent à leur tour les éléments retenus comme pertinents », p. 13).

L'élément récurrent dans les travaux en terminologie est la notion de concept qui me semble poser des questions nouvelles à la linguistique, justement en lien avec la prise en compte du réel. On peut penser en effet qu'on passe du sens au concept grâce à une interprétation qui prend en compte l'objectif de l'étude. Cet objectif est lui-même pris en considération au moment de la constitution du corpus. Ainsi l'analyse de corpus pour la constitution de terminologies peut être considérée comme une analyse

sémantique qui vise à mettre au jour les concepts, c'est-à-dire de contenus informationnels, en lien avec des applications, c'est-à-dire de signifiés retenus pour leur pertinence par rapport à l'objectif visé. La question est alors de savoir si la diversité des applications est telle que chacune d'elles nécessite que soit reconsidéré le corpus et l'analyse refaite ou bien si l'on peut imaginer une « stabilité » dans les applications qui permette de faire un travail de repérage maximal de termes (c'est-à-dire d'éléments linguistiques ayant un contenu informationnel) une fois pour toutes, quitte à refaire un tri éventuel pour des applications moins exigeantes. Nous retrouvons ici une problématique très semblable à celle de l'Ingénierie des Connaissances (cf. 5.2).

En dix ans, beaucoup de progrès ont été réalisés sur l'analyse linguistique de corpus spécialisés dans une perspective terminologique. De nombreux travaux désormais revendiquent une analyse de corpus systématique pour repérer la terminologie. Cependant, il reste beaucoup à faire car d'un côté, la tradition positiviste en terminologie est encore très vivace, de l'autre, l'analyse de corpus, tout particulièrement l'analyse sémantique, commence seulement à être acceptée comme posant des questions à la théorie linguistique alors qu'elle a été reléguée pendant longtemps du côté de la linguistique appliquée.

5. Sémantique et informatique : les corpus et le traitement informatique

Les liens de l'opération « Sémantique et Corpus » avec l'informatique sont très étroits. Nous travaillons tous de manière suivie avec des informaticiens, tout particulièrement des informaticiens de l'Irit (Institut de Recherche en Informatique de Toulouse). Par ailleurs, plusieurs d'entre nous ont soutenu des thèses en informatique (C. Fabre, N. Hathout, L. Tanguy). Nous sommes également très actifs dans la communauté TAL et IA puisque nous participons régulièrement à des colloques de TAL ou d'Ingénierie des Connaissances, à des groupes de recherche et à des revues dans ces mêmes domaines, soit comme intervenants soit comme responsables. Enfin, une filière TAL, rattachée au Département de Sciences du Langage, s'est mise en place depuis 1999 à l'Université Toulouse-Le Mirail.

5.1. Corpus et TAL

Le matériau textuel constitue la matière première du traitement automatique de la langue. En effet, même s'ils sont élaborés de manière introspective, les modèles et les formalismes proposés par le TAL ont pour vocation d'être testés sur des données réelles. Ce matériau textuel prend des formes différentes selon les perspectives :

A) Pour l'analyse morphologique et syntaxique, la « dimension textuelle » d'une langue est souvent considérée comme homogène ; aucune différence

n'est alors faite en lien avec la nature du corpus à traiter. Les études pour la construction du formalisme ou du modèle sont le plus souvent réalisées d'abord par introspection et parfois ajustées en fonction de fonctionnements attestés, l'objectif étant la plus grande efficacité. Dans cette perspective d'efficacité d'ailleurs, les approches statistiques les plus récentes tiennent compte du fonctionnement en corpus en recourant à un entraînement sur une partie du corpus à analyser.

B) Pour l'analyse sémantique, deux approches s'opposent. L'une, fonctionnant sur la même idée d'une homogénéité de la langue, essaie de construire des systèmes généraux d'interprétation sémantique. Les projets visant à construire des grands réseaux sémantiques utilisables comme entrée pour des outils de TAL, comme Wordnet et maintenant Eurowordnet, s'inscrivent clairement dans cette perspective. Les outils construits ont alors pour ambition de permettre l'interprétation de n'importe quel corpus, considéré comme une instance du modèle de la langue pré-construit (vision descendante). Les résultats sont à la mesure de la vision uniforme des problèmes : souvent inadaptés ou ne correspondant qu'en partie à la demande.

Ce point de vue trouve un soutien considérable dans la demande sociale concernant la recherche d'informations sur internet. En effet, étant donné que quasiment n'importe quel type de demande peut être faite (dans n'importe quel type de domaine, avec n'importe quel niveau de connaissances), il est impossible de tenir compte de toutes les variations possibles. Dans ce type d'approche, la solution qui semble la mieux adaptée, en tout cas pour l'instant, est de proposer un système homogène, supposé robuste car adaptable quelle que soit la question.

L'autre approche, en lien avec la prise en compte de caractéristiques des corpus à traiter, s'oppose à cette vision globalisante du fonctionnement sémantique qui peut occulter le fonctionnement propre d'un corpus. Comme la caractérisation *a priori* de tous les genres de corpus n'est pas possible, les outils constitués dans ce courant visent à faire émerger des fonctionnements propres au corpus à l'étude en laissant à l'utilisateur le soin d'en donner une interprétation sémantique. Les deux outils réalisés dans l'Equipe : Syntex, qui repère les syntagmes nominaux et les syntagmes verbaux (voir l'article de Didier Bourigault et Cécile Fabre dans ce numéro), et Yakwa (Ludovic Tanguy), qui permet de faire des interrogations sur corpus étiqueté, s'inscrivent dans cette vision. Une cohérence se crée ainsi entre les travaux des linguistes de l'opération, qui partent du postulat d'un fonctionnement propre à chaque corpus, et ces outils qui peuvent être utilisés pour analyser ce fonctionnement. Ce point de départ de l'analyse sémantique de corpus ne se basant sur aucun présupposé sémantique nous permet d'ancrer nos approches sur des observables, ce qui donne une assise solide à nos conclusions. Nous sommes bien conscients dans le même temps des questions que soulève cette

approche : séparation entre syntaxe et sémantique, problème de la généralisation des résultats.

- La séparation entre syntaxe et sémantique. Les deux outils construits dans l'équipe utilisent en entrée un corpus étiqueté avec un étiqueteur « général » qui ne tient pas compte de la nature du corpus (même si les informaticiens de l'équipe interviennent parfois pour forcer les analyses de l'étiqueteur afin de tenir compte des caractéristiques du corpus). Une contradiction apparaît ainsi entre un fonctionnement sémantique considéré comme dépendant du corpus et un fonctionnement syntaxique considéré comme indépendant, comme si syntaxe et sémantique pouvaient être à ce point déconnectées, comme si la stabilité était toute entière du côté de la syntaxe et la variation du côté de la sémantique. Or, nous savons bien que la variation sémantique a une influence sur le comportement syntaxique. Méthodologiquement, nous sommes obligés de partir d'une stabilité syntaxique, considérée comme plus grande qu'en sémantique, pour nous appuyer sur elle et mettre au jour des variations sémantiques, variations qui, à leur tour, sont étudiées dans leur dimension syntaxique. La mise au jour de variations sémantiques est un de nos principaux objectifs. Il ne s'agit pas pour autant d'écarter toute régularité sémantique, d'une part parce que nous visons souvent à décrire tout le fonctionnement sémantique d'un corpus, d'autre part parce que nous nous appuyons sur cette régularité d'un corpus à l'autre, pour mieux mettre en évidence les variations. La difficulté vient de ce que nous ne savons pas toujours *a priori* quels éléments vont fonctionner de manière régulière et lesquels vont avoir un fonctionnement inattendu. Ce constat est évidemment frustrant pour le linguiste qui cherche à identifier des régularités. Le problème de la généralisation des résultats est ainsi crucial pour nous.
- Le problème de la généralisation. Pouvoir généraliser un résultat suppose que les résultats obtenus pour un corpus peuvent être utilisés en entrée d'une analyse d'un corpus du même type, voire, en entrée d'un outil ayant pour objectif une analyse sémantique. La difficulté est alors de s'assurer de la similitude des corpus. En fait, la généralisation ne semble possible que lorsque l'étude a été réalisée sur un corpus constitué spécifiquement pour cela : les tests sont faits sur des sous-corpus constitués en faisant varier différents éléments de domaine ou de genre ; l'objectif est d'identifier une corrélation entre tel fonctionnement et tel(s) trait(s) du corpus.

Le traitement des corpus en TAL est un thème d'actualité. En peu de temps, (et on note une évolution assez nette par rapport à la description faite dans leur livre par Habert/Nazarenko/Salem sur les linguistiques de corpus (1997)), on ne compte plus, en France et à l'étranger, les publications ou les colloques qui ont ce thème pour axe principal. Il est d'ailleurs assez étonnant

que cette préoccupation ne prenne son importance que depuis les années 90 alors que la dimension textuelle pour le TAL est à ce point incontournable. Il est étonnant aussi que cet engouement s'accompagne d'un sentiment d'innovation alors que les corpus sont utilisés en linguistique depuis bien longtemps : au moins depuis le XIX^{ème} siècle pour ce qui est de la description comparée des langues et à tout le moins depuis le début ou la moitié du XX^{ème} siècle en dialectologie ou en sociolinguistique, disciplines qui, il est vrai, n'étaient pas considérées comme les plus centrales de la linguistique[6]. Ce qui est nouveau sans doute, c'est que le travail sur corpus vient maintenant interroger la linguistique (mais aussi le TAL) au cœur même de sa problématique : l'étude sémantique. Il y a ainsi une évolution parallèle entre le TAL et la linguistique qui s'intéressent tous deux à l'analyse de corpus. Pour autant, il serait dangereux de penser qu'il n'y a aucune différence entre « linguistique de corpus » et « TAL et corpus ». La problématique de la linguistique de corpus est une problématique avant tout descriptive, qui s'interroge sur le rôle des corpus. Dans cet objectif, et surtout lorsqu'il s'agit de traiter des grands corpus, l'utilisation d'outils est indispensable, elle permet d'ailleurs souvent de mettre en évidence rapidement des phénomènes (rapprochements de distributions, par exemple) qui seraient difficilement visibles « à l'œil nu ». Pour le TAL, l'utilisation des corpus est d'emblée associée, d'une part à une recherche d'efficacité (parce que le postulat d'un système sémantique général n'est pas toujours satisfaisant) et d'autre part à une application informatique. Or, même dans un contexte industriel, la demande ne concerne pas toujours la réalisation d'un outil ; elle concerne souvent l'analyse de corpus en tant que telle, pour identifier des cas d'ambiguïtés, d'incohérences ou pour faire des comparaisons de corpus. L'analyse à effectuer se rapproche alors de ce qu'on a pu appeler l'ergonomie linguistique (Rastier 1994) (et la linguistique de corpus aurait sans doute intérêt à se situer aussi par rapport à cette discipline - l'ergonomie - qui a, depuis longtemps, l'expérience du terrain). Les outils n'interviennent alors que comme aides dans une analyse qui relève à part entière de l'analyse sémantique.

Les deux types d'approches : « linguistique de corpus » et « TAL et corpus » ne sont ainsi pas toujours compatibles dans leurs objectifs : un outil peut bien fonctionner sans beaucoup de connaissances linguistiques ; inversement, les outils sont parfois peu utiles pour des analyses sémantiques très fines (voir l'article de M.-P. Jacques dans ce numéro). Il est nécessaire d'avoir conscience de ces limites ce qui permet de définir les points de

[6] Il est probable que les travaux des générativistes, à la suite de Chomsky qui s'opposait formellement à la prise en compte des variations préférant considérer un locuteur idéal, n'ont pas contribué à la prise en compte de l'analyse de corpus réels. L'évolution actuelle est due, en grande partie, à l'évolution des moyens techniques (corpus disponibles en format électronique et outils d'analyse), et à la demande sociale d'analyse de ces corpus.

collaboration possibles et de ne pas introduire de confusion entre les problématiques du TAL et de la linguistique. C'est l'objectif que nous visons dans l'opération : arriver à définir au plus juste ce que peut être une sémantique de corpus et, dans le même temps, circonscrire les champs de collaboration possible avec le TAL (cf. article Tanguy/Rebeyrolle dans ce numéro).

5.2. Corpus et IA

En IA, les corpus, particulièrement les corpus spécialisés, concernent les chercheurs qui travaillent sur l'acquisition de connaissances à partir de textes (voir l'article de N. Aussenac-Gilles et P. Séguéla dans ce numéro). Il s'agit pour eux de repérer à la fois les concepts qu'ils vont pouvoir retenir pour modéliser le domaine de connaissances sur lequel ils travaillent et les noms qu'ils pourraient utiliser comme étiquettes de ces concepts.

On retrouve pour l'IA une évolution assez comparable à celle du TAL dans la prise en compte des corpus. De nombreux chercheurs considèrent désormais qu'il ne s'agit plus de traiter le raisonnement dans son ensemble mais d'identifier des applications, que l'on prend en compte très tôt dans le processus de constitution de la base de connaissance (voir Charlet *et al* 2000). Si bien que l'on ne parle plus à présent d'Intelligence Artificielle mais d'Ingénierie des Connaissances (IC). Il ne s'agit plus de faire des systèmes généraux de raisonnement mais, étant donné un besoin, de trouver les moyens informatiques qui vont permettre d'aider à la résolution de ce besoin. Dans cette évolution, le rôle des textes s'est trouvé renforcé, tout particulièrement en France, sous l'impulsion du groupe TIA (Terminologie et Intelligence Artificielle)[7], que D. Bourigault et moi-même avons constitué en 1993, et dans lequel s'est élaborée une réflexion conjointe sur le rôle des corpus pour acquérir de la connaissance en IC et sur le rôle de l'application pour interpréter un texte en linguistique.

En effet, à côté d'ontologies générales, censées représenter le monde, commun à l'ensemble des humains supposés avoir le même type de perception et donc de représentation, la prise en compte des corpus et de leur nature dans la recherche des concepts a conduit à la notion d'ontologie régionale « valable seulement localement, régionalement, dans le cadre d'un domaine et d'une tâche » (Bachimont 2000, 315).

Le parallélisme entre le TAL et l'Ingénierie des Connaissances à partir de textes est ainsi très net et deux approches s'opposent entre réseau sémantique général/ontologie générale d'une part et réseau sémantique propre à un corpus/ontologie régionale d'autre part (même si les formalismes mis en œuvre dans l'un et l'autre cas ne sont pas les mêmes : formalismes syntaxiques pour le TAL (HPSG, LFG…) et formalismes de représentation

7 http://www.biomath.jussieu.fr/TIA/

des connaissances pour l'IC (logiques de description, Graphes conceptuels…)). La question qui se pose à la linguistique de corpus sous la forme de la généralisation des résultats se pose pour le TAL et l'IC en terme de réutilisabilité. En revanche, la vision dichotomique à propos du rôle des corpus, qui oppose assez nettement les chercheurs aussi bien en IA qu'en TAL, ne me semble pas aussi nette en linguistique, peut-être parce que la question s'y pose depuis bien plus longtemps qu'en informatique, en tout cas de manière sous-jacente, à travers le problème de la variation. Pour emprunter la terminologie de Récanati (Récanati 1997), entre les fixistes radicaux et les contextualistes tout aussi radicaux, finalement assez rares, un large éventail de points de vue existe et chaque chercheur, en fonction de son expérience et de sa réflexion se situe plutôt d'un côté ou de l'autre, les linguistes ayant une pratique des corpus se situant nettement du côté des contextualistes. Pour l'informatique, ce qui prime est l'efficacité des traitements, ce n'est pas la qualité de la description linguistique qui est recherchée : une description linguistique trop fine peut même être un handicap dans certains cas et des approches probabilistes, sans aucune connaissance linguistique peuvent être plus efficaces que des approches utilisant les résultats de la linguistique.

La difficulté vient alors du problème de la validation d'une description linguistique. Là où l'informatique peut avoir pour validation un « ça marche » encourageant, y compris pour les linguistes qui ont participé à la conception de l'outil, la validation en linguistique, tout particulièrement en sémantique est bien plus difficile à obtenir, ce qui rend sans doute cette discipline plus austère et moins ostentatoire que l'informatique. La psycholinguistique est peut-être une manière de validation, mais il peut y avoir dans cette discipline une recherche de reconnaissance « scientifique » qui peut biaiser les expérimentations[8]. D'autres moyens, sans doute moins radicaux, me semblent à retenir :

- Le fait que des résultats obtenus pour un corpus ayant telle caractéristique de domaine et de genre se retrouvent dans un corpus ayant les mêmes caractéristiques.
- Lorsque l'analyse a été faite en fonction d'une demande précise, le fait que les résultats obtenus sur un corpus satisfassent les demandeurs et leur apportent un plus dans la compréhension de leurs problèmes en lien avec du matériau textuel.
- Le fait qu'une communauté de linguistes, partageant une pratique et une connaissance du matériau textuel accordent une pertinence réelle aux résultats obtenus.

[8] Au fond, on peut se demander si la recherche de validation scientifique est compatible avec le travail en sémantique tant le sens paraît parfois éloigné de la notion de vérité unique.

Au fond, assez peu d'éléments qui garantissent la « vérité » des résultats mais beaucoup qui encouragent à la modestie et surtout, à la mise en place de projets nouveaux qui permettent de développer une approche qui n'en est qu'à ses balbutiements. En tout cas, beaucoup d'éléments qui rappellent que la linguistique, tout particulièrement examinée du point de vue de la sémantique est avant tout une discipline qui relève des sciences humaines et sociales.

6. Conclusion

Au terme de cette réflexion sur les travaux qui balisent le thème de recherche de l'opération « Sémantique et Corpus », il me semble que l'on peut résumer ainsi les points qui nous fédèrent :

Le corpus comme lieu d'étude

Que ce soit pour la constitution de produits terminologiques, pour la construction de logiciels, pour l'analyse de certains phénomènes sémantiques (nominalisation, polysémie, ellipses, repérage de relations conceptuelles…), notre matériau d'étude est un corpus. Il ne s'agit pas seulement de vérifier des intuitions mais essentiellement de faire émerger des régularités à partir du matériau textuel. Dans cette perspective, la constitution du corpus est une question à part entière selon que l'on vise une application particulière ou bien une généralisation des résultats obtenus à d'autres corpus ayant les mêmes caractéristiques extra-linguistiques.

Le problème de la généralisation des résultats

La possibilité de généraliser les résultats à d'autres corpus se fonde sur l'idée que les régularités observées sur un corpus se retrouvent aussi sur un autre corpus du même type. Cette possibilité permet une prédictibilité des fonctionnements et rend possible le traitement automatique et l'enseignement de ces régularités. Ce que l'on ne sait pas bien pour l'instant, c'est si une telle méthode ascendante a un sens pour décrire le fonctionnement sémantique d'une langue, même si l'on prend en compte des notions comme le domaine, le genre… Il semble qu'il existe un continuum entre fonctionnements totalement imprédictibles et liés à un corpus et/ou une application et fonctionnements totalement prédictibles, indépendants de caractéristiques extra-linguistiques. On mesure mal actuellement la part de prédictible et d'imprédictible que l'on va trouver dans un corpus. On peut espérer que les travaux sur corpus se développant, la part de prédictible augmentera mais il est vraisemblable que la part d'imprédictible ne disparaîtra jamais complètement (ce qui ne veut pas dire que cette part-là sera inaccessible mais qu'il faudra mettre en œuvre des études chaque fois nouvelles pour y accéder ; en revanche la mise en œuvre de ces études pourra, elle, être beaucoup plus systématisée).

Le rôle du contexte dans l'interprétation sémantique

Le problème de la constitution du corpus oblige à s'interroger sur le statut qu'on lui donne et par conséquent sur l'objectif de l'étude que l'on va mener. En effet, selon que l'on répond à une demande d'entreprise ou que l'on construit un corpus pour caractériser tel fonctionnement linguistique, les possibilités et les contraintes (disponibilité des corpus par exemple) ne sont pas les mêmes. Dans tous les cas pourtant, on aborde des questions de sémantique, qui apparaissent comme indissociables du contexte (social, scientifique…) dans lequel et pour lequel les textes du corpus ont été élaborés. Dans cette perspective, si un fonctionnement sémantique peut apparaître comme indépendant d'un contexte, c'est plutôt parce que l'on a identifié une régularité de fonctionnement d'un corpus à l'autre, c'est-à-dire d'un contexte à l'autre. On peut d'ailleurs penser que lorsqu'on essaie d'identifier le sens d'un mot, dans l'absolu, on essaie de classifier l'ensemble des contextes où on l'a rencontré pour décider en faveur d'une ou plusieurs classes de fonctionnements, c'est-à-dire en fonction d'un ou plusieurs sens. Ainsi, dans le cas d'un dictionnaire de langue générale, on suppose que tous les locuteurs d'une langue ont peu ou prou le même type d'expériences sémantiques et donc qu'ils neutralisent de la même façon le rôle du contexte. D'où un certain nombre d'inadéquations avec les fonctionnements réellement constatés lorsqu'on travaille sur des corpus, en particulier des corpus spécialisés.

Les rapports entre information et sens

L'analyse de corpus passe nécessairement à un moment ou à un autre par une étude et même un classement des contextes. En ce sens, on peut dire que l'analyse de corpus est toujours de type distributionnel. Il est nécessaire alors de rapprocher ou de distinguer des contextes afin d'identifier des classes de fonctionnements, soit sur des bases strictement formelles, comme le fait l'analyse automatique, soit en faisant intervenir des connaissances sémantiques *a priori* pour repérer des similarités malgré des différences de forme. Dans cette seconde optique, il est probable que la similarité concerne plutôt le contenu informationnel que le contenu sémantique. C'est encore plus net lorsqu'on travaille sur des corpus spécialisés, en lien avec des applications. On gagne sans doute en efficacité mais on s'éloigne du sens qui, au bout du compte, peut sembler totalement hors d'accès (c'est d'ailleurs l'hypothèse que faisait Harris, en tout cas dans ses premiers travaux).

Le problème de la validation des résultats

Comme je l'ai dit, on peut trouver dans la confrontation avec nos pairs linguistes ou avec nos interlocuteurs sociaux une validation ou une invalidation de nos résultats. Reste à savoir si nous validons alors des résultats de sémantique. En tout cas, cette question de la validation, problématique sans doute pour toutes les disciplines « humaines », qui

aimeraient se voir accorder le statut de sciences, interpelle tout particulièrement la sémantique, spécialement lorsqu'elle accepte de prendre en compte la réalité des corpus[9].

Nous avons bien conscience de possibilités d'études nouvelles qui sont offertes par la mise à disposition de corpus et surtout d'outils pour les étudier. La sémantique, déjà la plus récente des disciplines linguistiques, se trouve ainsi interpellée sur des questions fondamentales. Nous avons la chance à Toulouse d'avoir pu constituer un groupe homogène, qui a maintenant une expérience suffisamment longue d'analyse de corpus pour avoir un peu de recul sur ces questions. L'aventure ne fait que commencer et elle est riche de promesses.

Références bibliographiques

Bachimont, B. (2000), « Engagement sémantique et engagement ontologique : conception et réalisation d'ontologie en ingénierie des connaissances », in J. Charlet, M. Zacklad, G. Kassel & D. Bourigault (éds), *Ingénierie des Connaissances, Evolutions récentes et nouveaux défis*, Paris, Eyrolles.

Benveniste, E (1966), *Problèmes de linguistique générale*, 1, Paris, Gallimard.

Biber, D. (1988), *Variation across speech and writing*, Cambridge University Press.

Bloomfield, L. (1965), *Language*, Printed in Great britain By the Compton Printig Works (first published in Great Britain, 1935).

Charlet, J., Zacklad, M., Kassel, G. & Bourigault, D. (2000), *Ingénierie des Connaissances, Evolutions récentes et nouveaux défis*, Paris, Eyrolle et France Télécom.

Dachelet, R. (1994), *Sur la notion de sous-langage*, Thèse en sciences du langage de l'Université Paris VIII.

Deulofeu, J. (1992), « Variation syntaxique : Recherche d'invariants et étude des attitudes des locuteurs devant la norme », in *Langages* 108, « Hétérogénéité et variation : Labov, un bilan », pp. 66-78.

Firth, J.R. (1969), *Papers in Linguistics 1934-1951*, Oxford University Press, (première édition, 1957).

Gaudin, F. (1993), *Pour une socioterminologie*, Publications de l'Université de Rouen n°182.

Greimas, A. (1966), *Sémantique structurale*, Paris, Larousse.

[9] Il se peut d'ailleurs que la crainte de s'éloigner d'une vision considérée comme scientifique joue un rôle de frein dans l'utilisation des corpus en linguistique. Chomsky avait clairement fait le choix d'un locuteur idéal pour garantir la scientificité de l'approche linguistique.

Habert, B., Naulleau, E., & Nazarenko, A. (1996), « Symbolic Word Clustering for Medium-Size Corpora », in *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.

Harris, Z. (1966), *Structural linguistics*, The University of Chicago Press, seventh edition (first edition, 1951).

Kittredge, R. (1982), « Variation and Homogeneity of Sublanguages », in R. Kittredge & J. Lehrberger (eds), *Sublanguage : Studies of language in Restricted Semantics Domains*, Berlin, New York, de Gruyter, pp. 107-137.

Kleiber, G. (1981), *Problèmes de référence : Descriptions définies et noms propres*, Paris, Klincksieck.

Kleiber, G. (1999), *Problèmes de sémantique, la polysémie en question*, Paris, Villeneuve d'Asq, Presses Universitaires du Septentrion.

Kocourek, R. (1991), *La langue française de la technique et de la science*, (2ème édition, 1982), Wiesbaden, Brandestetter.

Lyons, J. (1980), *Sémantique linguistique*, Paris, Larousse.

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 », in *La Banque des mots* 7, Numéro spécial « Terminologie et Intelligence Artificielle », pp. 35-64.

Rastier, F., Cavazza, M., & Abeillé, A. (1994), *Sémantique pour l'analyse, De la linguistique à l'informatique*, Paris , Masson.

Récanati, F. (1997), « La polysémie contre le fixisme », in *Langue Française*, P. Cadiot, & B. Habert (éds), « Aux sources de la polysémie nominale », n° 113, pp. 107-123.

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 », in G. Rondeau & H. Felber (éds), *Textes choisis de terminologie*, GIRSTERM, Université de Laval, Québec, pp. 55-108.