Sémantique et corpus, quelles rencontres possibles ?

Anne Condamines

Introduction

Sémantique et corpus, quelles rencontres possibles ?[1]

1.1. Présentation de la problématique

Quels sont les éléments nouveaux qui peuvent justifier un ouvrage intitulé *Sémantique et Corpus*. Ni la sémantique, ni l'analyse de corpus ne sont des problématiques réellement nouvelles (voir par exemple Rastier, 2001). Même si le terme de sémantique est assez récent (on le doit à Bréal, 1893), l'interrogation sur le sens traverse le questionnement des philosophes, des philologues et des linguistes depuis 2 500 ans. Quant aux corpus, en tout cas aux textes, ils constituent depuis déjà longtemps l'objet d'étude de disciplines des sciences du langage aussi diverses que la sociolinguistique, l'ethnométhodologie, l'analyse conversationnelle, l'analyse de discours, l'analyse linguistique de textes littéraires, la linguistique historique et, depuis plus longtemps encore, de l'analyse comparative particulièrement développée au XIX^{e} siècle par Bopp, Humboldt (Trabant,1995) ou Schleicher (Auroux *et al*., 2000).

C'est l'association des termes de « sémantique » et de « corpus » qui constitue la nouveauté du projet. Cette combinaison pourrait presque donner un effet d'oxymore où se verrait opposer la notion de possibilité de stabiliser le système (ce qui était un

---

Introduction rédigée par Anne CONDAMINES.

1. Je remercie Andrée Borillo, Benoît Habert, Marie-Paule Péry-Woodley et Josette Rebeyrolle pour leur relecture de cette introduction.

des objectifs de la sémantique à ses débuts) et celle de variation des usages. En fait, se trouve résumée dans ce titre une des problématiques majeures de la linguistique lorsqu'elle s'intéresse au sens : celle de la confrontation d'une élaboration introspective visant la mise au jour du « système » avec la réalité de la variation langagière ; cette confrontation créant souvent une vive tension. Mais ce qui stimule particulièrement l'analyse du sens en corpus provient d'une évolution technique sans précédent qui, tout à la fois, crée une dynamique importante et entraîne une déstabilisation des acquis. Cette évolution se manifeste par une mise à disposition de textes en très grand nombre *via* Internet et le développement d'outils pour les interroger. En lien avec cette abondance, la pression sociale est forte (en particulier en provenance des milieux professionnels) pour prendre en compte, interpréter, expliquer le sens de ces textes et trouver des modes d'enregistrement et d'accès à leur contenu. Il faut reconnaître aussi que cette possibilité d'accéder facilement à des textes intervient au moment où les linguistiques introspectives (structurale ou générativiste) sont fragilisées, justement parce que leurs postulats ne résistent pas toujours à la confrontation avec la réalité des usages. Cette situation crée une grande effervescence, renouvelle les interrogations dans la linguistique tout entière, modifie les relations entre les disciplines en rapprochant la sémantique de problématiques du traitement automatique de la langue ou de l'ingénierie des connaissances. Des questions comme la prise en compte de la variation, les possibilités d'établir des régularités à partir de corpus, le rôle des approches quantitatives, les modes d'évaluation des résultats sont pressantes. Pour la linguistique, le questionnement sur son statut épistémologique et même social devient incontournable. Mais cela suppose de renoncer à une forme de pouvoir du linguiste, qui est celui d'une connaissance *a priori*, coupée de toute réalisation réelle (Gadet et Pécheux, 1981 ; Auroux, 1998). Ainsi, la confrontation du linguiste avec la réalité des usages langagiers est bien souvent accompagnée d'un sentiment de limite et d'interrogation sur la nature de sa compétence. En comparaison avec une linguistique uniquement introspective, la prise en compte des corpus en sémantique suppose en effet une rencontre avec un principe de réalité qui, tenant compte de la difficulté de constituer un corpus, de la résistance à la systématisation que présentent parfois les faits langagiers et du travail long que suppose l'élaboration d'une interprétation, s'oppose souvent à la vision d'un locuteur « idéal » qui permettrait de décrire un modèle stable, contrôlé et prédictif.

Une autre difficulté d'une approche sur corpus est qu'elle oblige à s'intéresser aux liens entre textes et situation de production et aux liens entre constitution du corpus et objectifs de l'analyse. Si l'on ajoute à ces deux éléments la diversité des

méthodes d'analyse (à la main, avec des outils de TAL, avec des méthodes quantitatives, méthodes qui sont le plus souvent combinées), on se retrouve devant une situation comportant des éléments extrêmement imbriqués (nature des textes, objectif de l'interprétation, méthode d'analyse) qui interagissent les uns avec les autres, ce qui rend difficile l'établissement de descriptions stabilisées. Il est clair d'ailleurs qu'un des objectifs d'une analyse sémantique de corpus consiste à expliquer comment ces éléments s'organisent pour construire un sens et comment on peut essayer de stabiliser ces interactions, les expliquer et éventuellement les reproduire. Enfin, il est plus que probable que le degré de systématisabilité varie en fonction des phénomènes étudiés. Le départ entre phénomènes dont la description est pertinente pour toutes les occurrences (donc qui font partie du système de la langue et fonctionnent indépendamment de la nature des corpus), ceux dont la description est pertinente pour certains corpus (qu'il faut caractériser) voire pour certains objectifs, et enfin, ceux dont la description n'a de sens que pour un corpus et/ou un objectif donné, constitue un des enjeux de la sémantique de corpus. Seuls les deux premiers types de phénomènes seraient susceptibles de faire l'objet d'une généralisation ; pour le troisième type de phénomènes, des méthodes d'analyse adaptables devraient sans doute prendre le pas sur la recherche de la modélisation systématique. Si le thème des corpus en linguistique a donné naissance à la parution d'ouvrages assez nombreux dans le courant anglo-saxon, peu de livres existent en français sur le sujet (voir cependant Habert *et al.*, 1997 et Bilger, 2000).

Cette introduction permet de présenter le contexte global dans lequel s'élaborent les différents chapitres qui suivent. Elle s'organise en deux grandes parties, l'une concerne ce qui constitue la matière même des études : le corpus ; l'autre s'intéresse aux éléments qui sont étroitement liés à l'objet d'étude : les objectifs et les méthodes d'analyse.

1.2. L'objet d'étude : le corpus

Le corpus constitue un objet fondamentalement nouveau par rapport à celui que l'on dévolut souvent à la linguistique (la langue), principalement car il est limité. Ainsi, par rapport à un objet virtuel, que l'on pense pouvoir atteindre par introspection et *a priori* illimité, le corpus donne un effet de réel immédiat qui peut conduire à une remise en question fondamentale des connaissances linguistiques. La première partie de ce paragraphe s'interroge sur la constitution du corpus, c'est-à-dire sur les possibilités et les conséquences de la clôture du champ d'investigation. La seconde partie introduit les liens qui unissent le corpus à la situation

extralinguistique à travers la notion de textuel qui est très largement utilisée par les tenants d'une approche mettant en œuvre des productions attestées (Bouquet, 2004).

1.2.1. *Constitution du corpus*

Il est désormais acquis qu'un corpus n'est pas un ensemble de données langagières en vrac mais des données (en l'occurrence textuelles) qu'on décide de regrouper pour une étude particulière (Habert *et al*., 1997). Le corpus est ainsi à distinguer de la base textuelle, thématiquement assez homogène mais construite sans objectifs clairement définis comme le *Trésor de la Langue Française* (Viprey, cet ouvrage) ou la *Base des Textes de Français Ancien* ou la *Base du Dictionnaire du Moyen Français* (Prévost, cet ouvrage). La construction du corpus, parce qu'elle relève d'un choix, joue un rôle souvent crucial dans une analyse à partir de/en corpus.

Il ne faudrait pas pour autant en déduire qu'une fois l'objectif de l'étude clarifié, il ne reste plus qu'à trouver les textes pertinents pour la mener à bien. D'une part, cette notion de pertinence continue souvent à évoluer tout au long de l'analyse. Dans certains cas de description syntaxique par exemple, la définition de la nature exacte des corpus pour lesquels la description est pertinente se construit en même temps que la description des phénomènes eux-mêmes. D'autre part, il n'est pas toujours facile de constituer le corpus rêvé (« idéal » pour reprendre un terme lourdement chargé en linguistique) : les textes peuvent ne pas être disponibles en grande quantité (cas des textes de langues anciennes ou de textes d'entreprises), ils peuvent ne pas être disponibles sous le format électronique adapté (cas des textes tapés à la machine ou seulement sous la forme d'une image numérisée), ils peuvent être frappés de droit d'auteurs ou de confidentialité… Par ailleurs, beaucoup des « textes » disponibles ne comportent plus l'information typo-dispositionnelle qui contribue pourtant à l'instauration du sens (Bachimont, cet ouvrage). Mais cet état de fait semble évoluer. Dans des domaines où le linguiste n'est pas compétent, la disponibilité d'informateurs ou d'experts pour s'assurer de la bonne compréhension est quasiment indispensable ; elle n'est pourtant pas toujours possible, soit qu'il n'y ait plus d'informateurs comme dans le cas des textes anciens au sens large, (Prévost, cet ouvrage) soit que les experts aient disparu (mutation, déménagement, etc.).

Une distinction importante doit être faite entre les corpus qui sont constitués de textes écrits et ceux qui sont des transcriptions d'enregistrements. En règle générale, les premiers n'ont pas été rédigés à destination des linguistes ; ils préexistent à

l'analyse et ils sont donc détournés de leur finalité première. En revanche les seconds s'inscrivent assez souvent dans une situation d'analyse dans laquelle le linguiste est impliqué et ce, dès la constitution des productions (entretiens dirigés ou entretiens auxquels un analyste extérieur assiste). Dans ce cas de figure, l'analyste est confronté d'emblée à la nécessité de tenir compte de la situation de communication, au minimum parce qu'il doit s'intéresser au rôle de l'intonation et de la prosodie (et à sa prise en compte dans la transcription) mais aussi à celui du statut des locuteurs. Très souvent, ces échanges sont non seulement enregistrés mais aussi filmés. Il s'agit en effet de comprendre comment les gestes, les postures, les regards font sens au même titre que les communications verbales. Le cas le plus évident de l'implication du linguiste dans la constitution même des données est celui de l'ethnométhodologie qui fait du linguiste à la fois un participant et un interprète des échanges (Mondada, cet ouvrage). La nature orale des données a une autre conséquence, majeure. Contrairement à un texte qui n'est que le produit final d'un processus de rédaction, un discours oral est à la fois produit et élaboration : produit parce qu'à un moment il a été figé par la retranscription qui en a fait un texte mais aussi processus car la chronologie des communications, la « temporalité » (Mondada, cet ouvrage), contribue à la progression sémantique. On y voit à l'œuvre l'élaboration d'une pensée, d'une énonciation tout autant que d'un énoncé, on est en prise directe avec le « travail sémantique » (Blanche-Benveniste, cet ouvrage). Le fait que le sens continue à s'élaborer tout au long d'un échange oral a bien sûr des conséquences sur les modes d'analyse à mettre en place ; des questions très particulières se posent ainsi en ce qui concerne les possibilités de mettre en place des analyses automatiques.

1.2.1.1. *Le problème de la représentativité*

Le problème de la représentativité du corpus est totalement lié à celui de la généralisation des résultats. Lorsque celle-ci est posée comme un requis en amont de l'étude, la représentativité du corpus est l'élément garant de la possibilité de généraliser les résultats obtenus pour un corpus particulier à l'ensemble des textes qui auront les mêmes caractéristiques que ceux de ce corpus. Selon l'objectif de l'analyse envisagée, la représentativité du corpus se pose différemment. Trois cas de figure peuvent se présenter : le corpus existe préalablement à l'analyse qu'en fait le linguiste, le corpus est constitué pour représenter une langue ou un état de langue, le corpus est constitué pour la description d'un phénomène linguistique ou celle d'un phénomène de connaissance au sens large (cas de l'ingénierie des connaissances).

1.2.1.1.1. La représentativité n'est pas retenue comme notion pertinente pour l'analyse du corpus

Dans certains cas, l'étude n'a pas d'ambition de généralisation. C'est le cas lorsque le corpus est donné *a priori* et obéit à une cohérence décidée par un tiers ou par une situation objective (ou supposée telle). Relèvent de cette situation des corpus proposés à l'étude par une entreprise (souvent un seul long document) pour vérifier une cohérence, repérer des incomplétudes ou des ambiguïtés mais aussi l'ensemble des textes d'un auteur (textes littéraires : de nombreuses études sont ainsi réalisées sur les œuvres de tel ou tel auteur (Viprey, cet ouvrage)) dont on veut étudier le style ou encore les discours de tel homme politique. De manière générale, ce qu'on appelle l'analyse de discours à la française (dans la suite des travaux de Pécheux ou Foucault, 1966) relève de ce point de vue ainsi qu'une grande partie des travaux réalisés par l'ethnométhodologie ou l'analyse conversationnelle.

Ces études n'ont pas pour objectif de s'interroger sur les possibilités de décrire le système de la langue à partir d'usages mais plutôt sur la manière de dégager ce qui est propre au corpus étudié, ce qui en fait le style ou ce qui se manifeste comme des motifs récurrents dans ce corpus. Toutefois, ce type de travaux ne fait pas toujours l'impasse sur les possibilités d'extrapolation des résultats. Cette possibilité est envisagée à travers la notion de genre textuel qui est présentée dans la partie sur le genre textuel ci-dessous. La notion de genre textuel peut, de ce point de vue, être considérée comme une façon de rassembler des textes ayant les mêmes caractéristiques linguistiques et extralinguistiques.

1.2.1.1.2. Le corpus est représentatif d'une langue ou d'un état de langue

Particulièrement développées dans la linguistique de corpus anglo-saxonne (voir ci-dessous), les études qui visent à construire la grammaire ou le dictionnaire d'une langue prennent pour acquis que le corpus mis en œuvre est représentatif du noyau des usages de la langue. Dans le cas des *very large corpora*, cette représentativité est à comprendre en des termes quantitatifs : la quantité des données est censée pallier le risque d'insuffisance de la couverture de tous les registres. Ce type d'approche essaie de construire un système à partir des usages, de contrôler la variation en la repérant par des méthodes quantitatives qui ont le mérite de mettre en évidence les modifications de fréquence des phénomènes. Inévitablement se pose la question du sens que l'on va donner à ces variations de fréquence, la plupart du temps attribuables à la nature des textes (ou extraits de textes) dans lesquels elles apparaissent. Dans d'autres projets (Brown, LOB) les textes (voire les extraits de textes) ne sont généralement pas choisis au hasard mais en fonction de leur supposée

représentativité du genre dont ils relèvent, ce qui, évidemment, ne va pas sans poser de problèmes. En France, deux projets peuvent être situés dans ce type d'approche. D'une part, dans les années 1950, l'élaboration du *Français fondamental* à partir d'un corpus de 312 000 mots à destination des apprenants du français. D'autre part, la réalisation du *Trésor de la Langue Française*, à partir de la base Frantext, essentiellement composée d'ouvrages littéraires du XIX^{e} et XX^{e} siècles. Mais aucun de ces deux projets ne s'est inscrit dans la perspective d'une systématisation des méthodes comme cela a pu être le cas pour l'anglais.

1.2.1.1.3. Le corpus est constitué pour étudier un fonctionnement linguistique particulier ou pour acquérir des connaissances.

Dans ces cas-là, ce n'est pas l'ensemble du corpus qui est étudié en tant qu'échantillon de langue mais certains phénomènes, prédéfinis en fonction du type d'objectif. Lorsqu'il s'agit de travailler sur un phénomène linguistique particulier (syntaxique, lexical ou discursif), l'étude est focalisée sur ce phénomène et sur les éléments du corpus qui contribuent à le décrire. Dans un premier temps, peu d'hypothèses existent sur le rôle du corpus, le seul critère pris en compte étant que le corpus doit être homogène. Or, cette notion pose des problèmes car elle dépend du point de vue adopté. De fait, il y a une dizaine d'années, pour le français, c'est souvent la base Frantext qui était utilisée et on a reproché aux chercheurs de considérer cette base comme représentative du français. A présent, c'est souvent la base d'articles du quotidien *Le Monde* qui est étudiée ; elle a en effet le mérite d'être en partie disponible dans une version étiquetée grammaticalement, ce qui facilite les interrogations mais qui ne règle pas la question de sa représentativité. Cette étape permet de préciser les descriptions et de mettre au point des modes d'interrogations automatiques, ce qui est loin de se faire de manière aisée. Mais l'étape suivante, qui consiste à étudier de plus près les liens entre le mode d'instauration de tel ou tel phénomène et la nature du corpus reste très difficile à mettre en œuvre pour plusieurs raisons qui sont présentées ci-dessous.

Un autre type de travaux utilise un corpus pour acquérir et représenter des connaissances propres à un domaine. Il s'agit le plus souvent d'élaborer des connaissances sous formes de réseaux de termes, réseaux qui, dans leur version formelle, sont appelés ontologies (voir ci-dessous). Dans ce cas, l'analyse se focalise sur les parties de textes qui peuvent être représentées sous une forme relationnelle, c'est-à-dire sur les parties qui contiennent des marqueurs de relations. Mais pour que ces constructions soient possibles, il faut que le corpus soit constitué de manière très homogène, c'est-à-dire en respectant des caractéristiques extralinguistiques très

stables mais dont la définition est souvent liée à l'objectif d'analyse. Par exemple, pour une étude sur l'évolution terminologique, les textes du corpus doivent provenir de la même entreprise, doivent relever du même genre ; dans le cas de textes en anglais, ils doivent avoir été rédigés par des anglophones. Autant d'éléments qui garantissent que, tant du point de vue des modes de rédaction que des connaissances qui sont mises en œuvre, le corpus est constitué de manière linguistiquement homogène. La question sur les possibilités de généraliser les résultats se pose lorsque la construction d'outils est envisagée. En effet, la dépendance des marqueurs avec le texte varie en fonction des marqueurs : certains fonctionnent indépendamment de la nature du texte, d'autre sont dépendants d'un genre textuel, d'autres ne semblent fonctionner que pour un texte en particulier (Biber, 1993 ; Condamines, 2002).

1.2.1.2. *Clôture*

Associé à la question de la représentativité mais ouvrant sur d'autres interrogations, se pose le problème de la clôture du corpus. Il s'agit à la fois de s'interroger sur la position du corpus par rapport à d'autres corpus et sur l'interaction des données du corpus avec la connaissance de l'analyste. En d'autres termes, ce qui fait problème est le rôle du corpus comme objet d'étude circonscrit et la distance que l'analyste peut raisonnablement établir avec lui. Le linguiste (comme n'importe quel locuteur) qui explore un corpus ne fait pas table rase de ses connaissances linguistiques antérieures, mais au contraire, il les met en œuvre, consciemment ou inconsciemment. Pour autant, l'analyse sémantique d'un corpus ne consiste pas en une simple projection de sa connaissance pour faire émerger un sens définitif, voire préexistant à la mise en mots. Pour la plupart des chercheurs travaillant à partir de corpus, il est désormais acquis que l'élaboration d'un sens relève d'une construction. Mais cette construction ne peut être le fruit de l'ordonnancement spontané d'éléments immanents du corpus. Le linguiste nourrit son interprétation à la fois de sa mémoire de phénomènes langagiers antérieurs (intertextualité) et de son objectif d'étude tout en ancrant sa réflexion dans des éléments textuels (d'où l'importance du corpus). Une des tâches du linguiste est d'ailleurs sans doute de comprendre comment s'élabore son interprétation en faisant appel à ces différents éléments.

La question est alors de savoir jusqu'où le linguiste s'autorise à faire intervenir sa connaissance pour construire l'interprétation. Elle est particulièrement patente dans les cas des corpus spécialisés. Prenons le cas d'un corpus médical dans lequel on trouve les termes *lésion, obstruction, sténose, occlusion, réocclusion* et les

composés *artère lésée, artère sténosée, artère occluse*. Que doit-on penser du fait que l'on ne trouve ni *artère obstruée* ni *artère réoccluse* ? Soit on s'interroge sur la représentativité du corpus, soit on considère que puisqu'ils n'apparaissent pas, ces termes n'existent pas dans ce sous-langage, soit on s'autorise à utiliser sa compétence de locuteur non spécialiste par exemple, pour ajouter ces deux termes à un système d'extraction d'information. Cet exemple, parce qu'il relève seulement du lexical peut paraître simple à résoudre mais ce même phénomène se produit quel que soit le phénomène linguistique étudié et quelle que soit la nature du corpus.

Autre difficulté, la linguistique de corpus peut-elle être une linguistique de l'astérisque, c'est-à-dire une linguistique qui prend en compte non seulement les données présentes ou les données qui peuvent être raisonnablement extrapolées (comme ci-dessus) mais aussi aux éléments dont l'analyste décide qu'ils ne peuvent pas apparaître en lieu et place d'un autre élément ? En d'autres termes, jusqu'à quel point la norme intégrée par le linguiste peut-elle intervenir dans l'analyse de corpus ? Une fois encore, il ne peut y avoir de position définitive ; la nature du corpus, l'objectif de l'étude mais aussi l'étape de l'analyse à laquelle on se trouve peuvent intervenir. Par exemple, si l'on est en train de construire une hypothèse, il est acceptable de tolérer l'utilisation des astérisques parce que c'est par discrimination par rapport à des phénomènes qui paraissent impossibles que s'élabore l'hypothèse (Lamiroy et Charolles, cet ouvrage). Il en va de même pour les outils d'apprentissage en TAL, qui demandent que soient proposés des contre-exemples (Nazarenko, cet ouvrage). En revanche, pour toutes les méthodes et les objectifs qui sont basés sur la mise au jour des spécificités du corpus, le recours à l'astérisque n'a pas de sens : certaines méthodes statistiques, analyse conversationnelle, analyse de discours, etc.

1.2.2. *Le genre textuel*

La notion de genre ou de registre est présente dans la quasi-totalité des chapitres de cet ouvrage. C'est dire si elle est majeure pour la caractérisation du sens en corpus. S'il en est ainsi, c'est que beaucoup d'espoirs sont mis dans la capacité du genre textuel à stabiliser les descriptions, particulièrement les descriptions sémantiques. Le genre textuel pourrait ainsi permettre d'associer situation de production et réalisations langagières au point d'établir des corrélations qui entreraient dans les descriptions linguistiques (Branca-Rosoff, 1999).

Historiquement, la notion de genre a été travaillée dans des communautés bien distinctes. L'une (qui parle plutôt de registre), anglo-saxonne, a émergé dans la perspective de la prise en compte de la dimension sociale du langage avec des auteurs comme Firth (Firth, 1957) ou Bernstein. De très nombreux courants se sont ainsi intéressés à l'aspect fonctionnel du langage (analyse conversationnelle (Hymes), analyse fonctionnelle (Dik, Halliday), ethnométhodologie (Garfinkel, Schegloff)) et à la définition de groupes de locuteurs poursuivant des objectifs communs :

« *A genre comprises a class of communicative events, the members of which share some set of communicative purposes. These purposes are recognized by the expert members of the parent discourse community, and thereby constitute the rationale for the genre* » (Swales, 1990, 58).

Une autre communauté, originellement essentiellement russe, s'est aussi intéressée à la notion de genre, antérieurement à la communauté anglo-saxonne. Principalement inspirée par les travaux de Bakhtine, cette communauté s'inscrit, au moins initialement, dans une perspective à la fois plus historique et plus littéraire que la communauté anglo-saxonne. Telle qu'elle est définie, la notion de genre met en avant la dimension dialogique de la communication. Le sens d'un discours apparaît ainsi comme une coconstruction dans laquelle les deux protagonistes interviennent à part égale comme protagonistes socialement situés.

« *Aucun énoncé en général ne peut être attribué au seul locuteur : il est le produit de l'interaction des interlocuteurs et, plus largement, le produit de toute situation sociale complexe, dans laquelle il a surgi* ». (Bakhtine, cité par Todorov, 1981).

Ce point de vue du fonctionnement discursif éloigne de la perspective de la dimension strictement fonctionnelle du langage. En revanche, le fait qu'il existe à tout le moins une corrélation entre situation extralinguistique et réalisations effectives est très présente aussi chez Bakhtine :

« *Tout énoncé particulier est assurément individuel, mais chaque sphère d'usage du langage élabore ses types relativement stables d'énoncés, et c'est ce que nous appelons les genres discursifs* » (Bakhtine, 1984).

Les différences initiales entre les deux courants tendent à s'atténuer en particulier parce que la linguistique quantitative a fait émerger l'urgence de définir des paliers d'organisation des faits langagiers qui permettent de rendre compte

(voire d'expliquer) la variation des usages (qu'ils soient lexicaux, syntaxiques ou qu'ils concernent l'organisation textuelle). On peut ainsi considérer que la notion de genre fait intervenir trois éléments :

– tout d'abord, le genre préexiste à l'énonciation ; il constitue une façon de s'inscrire, socialement et linguistiquement, dans une communauté qui existe déjà ;

– du fait de la régularité qu'elle instaure, la notion de genre est associée à une idée de normativité ; comme le signale Todorov, « le genre forme un système modélisant qui propose un simulacre du monde » (Todorov, 1981, 128) ;

– enfin, la mise en œuvre des règles linguistiques propre à un genre se fait la plupart du temps à l'insu des locuteurs. C'est peut-être cet élément qui constitue la plus grande différence avec la caractérisation d'une langue qui peut être faite en utilisant les notions de préexistence et de normativité : on a conscience de parler dans une langue, beaucoup moins de s'inscrire dans un genre donné.

L'intérêt de l'existence de genres est qu'ils permettent de constituer des catégories de textes dont on suppose qu'ils ont les mêmes caractéristiques linguistiques et extralinguistiques. Avec une telle hypothèse, un texte devient représentatif d'un ensemble d'autres textes et il suffit de décrire un phénomène dans un des ces textes pour qu'on puisse envisager que la description soit valable pour tous les textes du même genre. Evidemment, les difficultés à propos de la définition des genres restent très nombreuses (Adam, 1999 ; Bronckart, 1996) : difficulté à définir des genres dans des situations discursives toujours mouvantes, combinaisons de plusieurs genres à l'intérieur d'un discours, points de vue de descriptions variables en fonction des objectifs. Un progrès est certainement venu de la distinction entre les régularités extralinguistiques dont on suppose qu'elles s'accompagnent de régularités de faits langagiers et régularités intralinguistiques qui peuvent conduire à réorganiser les textes initialement considérés comme étant du même genre :

« *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).

Dans le même temps, les méthodes d'analyse automatique se sont développées et ont permis d'évaluer rapidement la similarité entre textes supposés du même genre.

Cependant les méthodes automatiques ne permettent d'identifier que des régularités de formes ou de distributions. Lorsque l'analyse s'intéresse à l'interprétation sémantique, elle ne peut être faite de manière automatique.

C'est certainement autour de cette notion de genre et de sa pertinence dans la description des phénomènes en corpus que devraient se développer les travaux dans les prochaines années.

1.3. Objectifs et méthodes d'analyse

Il ne suffit pas de se donner un objet d'étude pour constituer un cadre d'analyse parfaitement clair. L'objectif et la méthode mis en place ont une influence majeure sur la nature des résultats qui sont obtenus.

1.3.1. *Objectifs d'analyse*

Les objectifs d'une analyse de corpus peuvent être très divers. Ils ne constituent pas un élément second dans la caractérisation de ce que peut être une sémantique de corpus. En effet, étant donné que la dimension interprétative est omniprésente dans les études à partir de corpus (les possibilités de contrôle (ou pas) de l'interprétation constituant pratiquement la principale interrogation d'un point de vue scientifique dans ce domaine), il est nécessaire de prendre en compte l'objectif d'étude pour mieux envisager une généralisation des résultats. Il n'est pas possible de faire un recensement exhaustif de tous les types d'objectifs qui président aux études qui, partant d'un corpus, en construise une interprétation fortement guidée par cet objectif même. Afin de situer les différents chapitres qui composent cet ouvrage, nous avons regroupé ces objectifs sous cinq rubriques.

1.3.1.1. *Analyse de discours*

L'analyse de discours relève bien d'une sémantique de corpus (il s'agit bien d'élaborer un sens (voire plusieurs sens) à partir d'un corpus même s'il n'est constitué que d'un seul texte). Pour autant, elle ne s'inscrit pas dans la perspective d'une généralisation des résultat ; en effet, es résultats sont uniquement relatifs à ce discours. Pourtant, le problème de l'interprétation s'y pose de manière cruciale avec le risque, dans une situation qui se cantonne à un discours isolé de tout autre discours, de produire des résultats eux-mêmes isolés, propres seulement à une situation d'interprétation et de nombreux auteurs sont conscients de ce risque :

« *L'enjeu crucial [de l'analyse de discours] 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, 17).

1.3.1.2. *Construction et/ou vérification d'une hypothèse, analyse de la structure des textes*

Cette situation correspond à celle du linguiste qui, conscient des limites de l'introspection, décide d'élaborer ou de vérifier une hypothèse dans des textes (voir Lamiroy *et al*., cet ouvrage). Désormais, la nécessité et l'intérêt de constituer un corpus sont suffisamment connues pour que ces études soient faites sur des données qui ne sont pas constituées de toutes les occurrences rencontrées fortuitement au cours de lectures mais bien des données dont on considère qu'elles ont une réelle homogénéité. Ce passage des données « en vrac » à un corpus correspond à une évolution fondamentale qui s'accompagne de la conscience (parfois plus ou moins claire) que l'on n'est plus dans l'élaboration du système de la langue mais de résultats qui sont relatifs au corpus d'étude. Dans la grande majorité des cas, les résultats sont présentés comme devant être évalués sur d'autres corpus, relevant d'autres domaines ou d'autres genres discursifs. Cette première étape d'analyse sur un corpus réel conduit aussi à la confrontation avec deux éléments qui font difficulté. D'une part, il s'agit d'un travail extrêmement long, qui nécessite l'encodage à la main d'un grand nombre d'informations afin de caractériser (d'annoter) à la fois l'élément à étudier et les éléments de son contexte, pertinents pour une description systématique. D'autre part, cette caractérisation apparaît souvent difficile à faire, les linguistes impliqués n'étant pas toujours d'accord pour opter en faveur de l'une ou de l'autre catégorie. L'évaluation sur un autre corpus n'est pas non plus chose aisée. On souhaiterait alors utiliser les résultats de la première étude pour mettre en place des analyses automatiques mais il faut alors d'une part, élaborer des patrons de recherche qui comportent des catégories plutôt que des formes (ce qui suppose de constituer des listes de formes à associer à ces catégories *a priori*) et d'autre part (et en conséquence), admettre une part de silence (éléments qui auraient été pertinents mais qui ne sont pas retrouvés, faute d'avoir été prélistés). Il s'agit en fait de s'interroger sur les liens entre analyse de texte et linguistique de corpus (Péry-Woodley, cet ouvrage ; Lagerwerf *et al*., 2003). En revanche, la mise en place d'une approche automatique permet de tester les hypothèses sur des volumes de données importants et de mettre au jour des phénomènes insoupçonnés ou des variations significatives d'un corpus à l'autre (Biber, 1996 ; Van Dijk, 1997).

1.3.1.3. *Description systématique d'une langue*

Cet objectif reste majoritairement associé à ce que l'on a appelé linguistique de corpus, dans le courant anglo-américain. Qu'il vise la description des règles de grammaire (Quirk *et al*., 1985 ; Biber *et al*., 1999) ou la construction de dictionnaires (Sinclair, 1991), cet objectif fait toujours intervenir deux éléments : un corpus très volumineux constitué, pour être représentatif, de la langue à étudier et la mise en œuvre de méthodes quantitatives.

1.3.1.4. *TAL*

Les objectifs du TAL s'organisent en deux types bien distincts ; l'un vise des applications définies et met en œuvre des ressources préexistantes : lexiques, étiqueteurs ; l'autre concerne la production d'outils permettant de créer des ressources qui seront éventuellement ensuite utilisées par les outils du type précédent.

Les applications du TAL sont bien répertoriées (Nazarenko ; Péry-Woodley, cet ouvrage ; voir aussi Pierrel, 2000) : recherche d'information, extraction d'information, système de question-réponse, résumé automatique, etc. Dans la plupart de ces applications, on connaît, avant la constitution de l'outil, le type de données (qui correspondent à des informations) qui doit être recherché ; le travail consiste à identifier *a priori* des formes (reconnaissables automatiquement) qui vont permettre de repérer ces données ou des éléments particulièrement pertinents pour identifier ces données (éléments appelés « marqueurs »). C'est finalement la pertinence de l'interprétation que l'on peut faire des formes repérées par rapport à l'objectif qui valide la pertinence de la méthode, d'où un point de vue de pragmatisme qui est maintenant revendiqué par les chercheurs de ce domaine (Nazarenko, cet ouvrage). Certains objectifs ne concernent pas des applications précises mais visent à constituer des outils généralistes d'analyse de textes ou de production de ressources (par exemple, extracteurs de termes ou de relations conceptuelles). Bien qu'ils ne leur soient pas spécifiquement dédiés, ces outils peuvent être utilisés par les linguistes pour élaborer des règles de fonctionnement ou vérifier des hypothèses sur des données textuelles volumineuses.

1.3.1.5. *Ingénierie des connaissances, acquisition de connaissances*

Dans ce type de perspective, les textes sont considérés comme des traces de connaissances. On fait donc l'hypothèse que l'on va pouvoir les utiliser pour construire des représentations qui pourront être utilisées soit par des humains

(terminologies, thesaurus), soit par des outils d'aide au raisonnement (Aussenac-Gilles *et al*., 2003). Une grande partie des travaux de l'ingénierie des connaissances se focalise maintenant sur la constitution « d'ontologies » (Charlet *et al*., 2000), c'est-à-dire de représentations formelles de la connaissance (Bachimont, cet ouvrage). Cette problématique présente en réalité des points communs très nets avec le problème des liens entre langue et connaissance et, plus précisément, avec les possibilités de combiner connaissances *a priori* (générales) et connaissances locales (Nazarenko, cet ouvrage). En effet, deux théories s'affrontent, l'une qui voudrait construire des ontologies générales, c'est-à-dire des ontologies stables, réutilisables car indépendantes des applications et élaborées par introspection. L'autre qui, dans une vision ascendante, envisage de construire des ontologies à partir de textes. Il s'agit bien alors d'élaborer une construction sémantique (un réseau de termes) à partir de textes. D'une certaine façon, il s'agit de s'interroger sur les possibilités de combiner sémantique logicoréférentielle et sémantique textuelle. On se trouve ainsi au cœur même de la problématique qui sous-tend la rencontre entre corpus et sémantique, entre connaissances préexistantes et connaissances immanentes, entre stabilité et variation.

1.3.2. *Méthodes d'analyse*

Ainsi que nous l'avons vu, les méthodes sont étroitement liées avec l'objectif, le point de vue qui préside à l'étude. On peut considérer qu'il existe trois types de méthodes, qui sont souvent d'ailleurs utilisées en parallèle ou successivement : méthodes manuelles, méthodes inspirées du TAL parmi les méthodes du TAL, méthodes quantitatives.

1.3.2.1. *Méthode manuelle*

La méthode d'analyse manuelle consiste en un parcours « à la main » des textes pour y repérer les phénomènes jugés intéressants en fonction d'un objectif. Cela suppose un premier temps au cours duquel une hypothèse de fonctionnement a été élaborée sur le phénomène à étudier : telle structure informationnelle, telle construction syntaxique, telle relation sémantique (synonymie, hyperonymie, etc.). Le repérage dans les textes permet de confirmer ou d'affiner l'hypothèse sur les modes de fonctionnement, plus rarement de l'infirmer (ce qui nécessite de remettre en question radicalement son intuition linguistique !). La plupart du temps, la confrontation avec la réalité des fonctionnements produit le sentiment que la description est plus compliquée qu'on ne le pensait. En particulier les dépendances

entre les fonctionnements, les effets sémantiques produits, la diversité des répartitions d'un texte à l'autre apparaissent comme plus importants qu'on ne l'imaginait. Cette étape manuelle est quasiment indispensable pour certains types d'études et il est difficile de l'assister avec des outils, en tout cas des outils qui utilisent une catégorisation *a priori*. En effet, la caractérisation des éléments constituent l'objectif même de l'étude, il est donc difficile de partir de corpus déjà annotés pour élaborer la description. Cela supposerait d'ailleurs qu'un « corpus de référence » soit disponible (ce qui n'est pas le cas pour le français) et que des jeux d'étiquettes soient élaborées. On peut formuler l'espoir que, pour la description de certains phénomènes assez stables en tout cas, il soit possible de stabiliser les annotations pertinentes. Cette stabilisation pourrait être le fruit de la cumulation d'analyses sur des corpus, ce qui supposerait de fédérer les travaux qui existent sur l'analyse de textes et de travailler à établir des catégories. Pour des perspectives morphologiques et/ou syntaxiques, qui ne nécessitent pas l'utilisation d'étiquettes sémantiques (ce qui est sans doute assez rare), cette perspective est peut-être envisageable. Pour des analyses qui visent uniquement une caractérisation sémantique ou qui nécessite des caractérisations sémantiques, ce jeu d'étiquettes est évidemment beaucoup plus difficile à envisager et il est probable que, dans ces cas-là, il vaudrait mieux mettre en œuvre des méthodes qui visent à travailler sur les régularités qui émanent du corpus (méthodes dites endogènes). La question reste entière de savoir quelle connaissance est *a priori* pertinente et laquelle doit être élaborée à partir du corpus.

1.3.2.2. *TAL*

Lorsqu'on l'examine du point de vue sémantique, le TAL quel que soit son objectif, vise à apparier des formes avec des contenus. Ces formes peuvent avoir fait l'objet d'un étiquetage préalable très poussé, c'est-à-dire d'une interprétation ou encore d'une catégorisation, grammaticale et parfois sémantique, ou au contraire d'une catégorisation presque inexistante (cas de la recherche d'information par exemple). Les outils utilisent aussi la répartition de ces formes pour repérer des distributions auxquelles on peut donner un sens. La plupart s'inscrivent ainsi dans une parenté pas toujours revendiquée avec l'approche harissienne, même si l'analyse automatique n'est pas toujours mise en œuvre sur des corpus censés représenter des sous-langages. Selon l'objectif toutefois, l'étiquetage préalable ne joue pas le même rôle. Lorsque un objectif très défini préside à la construction de l'outil, le pragmatisme joue un rôle déterminant (Nazarenko, cet ouvrage) qui justifie des choix qui, dans certains cas, peuvent paraître arbitraires du point de vue de la linguistique (par exemple, la troncation (plutôt que la lemmatisation) est mise en

œuvre en recherche d'information parce qu'elle semble plus efficace que l'application de règles morphologiques). Mais de manière générale, on cherche à justifier les connaissances utilisées par les outils d'un point de vue linguistique. Tout étiquetage est un choix d'interprétation. Lorsque cette interprétation est supposée consensuelle, il est évident qu'il y a un gain à l'utiliser car l'outil ne travaille plus sur des formes mais sur des catégories. Mais certains choix d'étiquetage, en particulier sémantique, peuvent constituer des biais. En effet, plus l'étiquetage sémantique est poussé, plus on peut retrouver d'information jugée *a priori* pertinente mais plus on risque aussi de s'éloigner des spécificités d'un corpus.

Ainsi que nous l'avons déjà souligné, cette question de l'étiquetage *a priori* devient cruciale lorsque l'objectif n'est pas défini en termes d'applications mais concerne la vérification d'une hypothèse linguistique sur un grand volume de données. Une tension se crée alors entre nécessité de catégoriser *a priori* pour pouvoir considérer un ensemble de formes linguistiques dans un même patron de recherche et réalité des phénomènes langagiers qui échappent parfois à la possibilité d'être (facilement) catégorisés. On se rend compte aussi que la catégorisation peut varier en fonction des points de vue d'étude. Tous les chapitres qui, dans cet ouvrage, concernent la possibilité de passer d'une analyse linguistique qualitative à une analyse quantitative, qui permettrait l'évaluation rapide et massive des hypothèses, évoquent les difficultés majeures qui sont rencontrées (Mondada, Péry-Woodley, Prévost). Enfin, certains types d'analyse, qui s'intéressent plus au processus d'élaboration du sens au cours d'échanges verbaux, par exemple ceux qui prennent en compte la temporalité de l'oral (Mondada, cet ouvrage), ne peuvent que difficilement relever d'un étiquetage *a priori*.

1.3.2.3. *Méthodes quantitatives*

Le dénombrement de faits langagiers dans les textes (quelle que soit leur nature) n'est pas nouveau ; pensons par exemple aux comptages sur la Bible qui existent depuis déjà longtemps (Guiraud, 1960) ; mais c'est seulement dans les années 1950-1960 que cette perspective a été intégrée dans l'analyse linguistique. Deux courants majeurs se sont alors profilés, tout particulièrement dans la communauté française : l'analyse lexicale (Guiraud, 1960 ; Muller, 1967) et l'analyse multidimensionnelle (Benzécri *et al*., 1984). Pour la statistique lexicale dans sa version initiale, « tout discours, tout texte est un échantillon d'un état de langue dont il reflète la structure numérique aussi bien que les possibilités de réalisations sémantiques » (Guiraud, 1960, 18). En lui offrant une assise expérimentale, la statistique est ainsi une façon de donner à la linguistique le statut de science qui lui faisait défaut : « [la

statistique] offre ce qui a manqué [au linguiste] jusqu'ici, l'appareil de mesure sans lequel il n'y a pas de science. » (*ibid*, 22). La statistique multidimensionnelle, initiée en France par Benzécri, s'est nettement positionnée contre la vision d'une langue idéale de Chomsky et a proposé une méthode inductive de construction d'abstractions à partir de régularités langagières. Dans les deux cas, on reste dans une vision de la langue unifiée, avec des fonctionnements stables que l'on va retrouver dans les corpus. L'approche distributionnelle, en tout cas au sens où la valeur linguistique est dépendante du cotexte où se trouvent les signes, est présente dans les deux approches mais là s'arrête la parenté avec les travaux de Harris ; en effet, dans aucun de ces deux courants, en tout cas à leurs origines, on ne trouve la notion de sous-corpus et plus généralement la nécessité de constituer un corpus avec des critères d'homogénéité bien définis. Ce n'est qu'avec l'approche anglo-saxonne de la linguistique de corpus que cette nécessité est apparue comme majeure, remettant en cause les ambitions théoriques de la statistique linguistique, qu'elle soit lexicale ou multidimensionnelle. Malgré cette remise en cause, qui a permis une évolution vers la recherche de régularités immanentes plutôt que vers la mise au jour de règles préexistantes, la dimension quantitative est aujourd'hui incontournable pour qui s'intéresse à l'utilisation des corpus en linguistique et plus particulièrement en sémantique. Cette dimension se décline selon deux modes de prise en compte. Dans un premier type de travaux, l'approche quantitative est mise en œuvre pour essayer de faire émerger des régularités immanentes d'un corpus constitué, soit en mettant en place des méthodes sur des formes très peu interprétées mais sur des corpus très homogènes (approches apparentées à la statistique lexicale) (Viprey, cet ouvrage), soit en combinant des méthodes statistiques avec des méthodes TAL, elles-mêmes de plus en plus sophistiquées (Nazarenko, Habert *et al*., cet ouvrage). Dans un second type de travaux, plus linguistiques, la quantification est plutôt vue comme une façon de valider des résultats sur un grand volume de données (Péry-Woodley, Prévost, cet ouvrage). Se pose alors la question du passage du qualitatif au quantitatif, qui est un problème bien connu de la linguistique de corpus (McEnery *et al*., 1996) et celle du problème, déjà évoqué, de l'étiquetage permettant la généralisation.

Même dans sa version la moins évoluée (simple dénombrement d'occurrences), l'approche quantitative est omniprésente dans les travaux de linguistes qui s'appuient sur des données textuelles. Il n'y a pas dans cet ouvrage un seul chapitre qui n'évoque la notion de fréquence des phénomènes. Il est clair que la quantification et l'étude de la répartition des phénomènes contribuent à décrire les fonctionnements, et il convient de s'interroger sur les relations entre fréquence de

phénomènes en corpus et possibilité d'interpréter cette fréquence comme manifestation de règles d'un système linguistique. Pour le dire autrement, il s'agit d'étudier les possibilités de remplacer la notion de règles qui a toujours été utilisée pour ériger la linguistique comme science par celles de régularités. D'une certaine façon, l'ambition initiale des statisticiens continue à avoir du sens même si elle a été largement révisée par le constat de variations parfois flagrantes qui émanent de corpus et par la difficulté à trouver des modèles qui soient de bons modèle de la « réalité » à décrire (Habert *et al*., cet ouvrage). Beaucoup d'espoir est mis sur l'étude des genres pour expliquer ces variations, mais il est probable qu'il faudra encore de longues études pour comprendre comment des régularités s'instaurent parfois de manière très stable, pourquoi certaines régularités, qui ne se révèlent qu'à l'analyse, échappent à toute prédiction et enfin pourquoi parfois, il n'est pas possible de faire émerger des régularités.

1.3.3. *Prise en compte de la situation de production et d'interprétation*

Un corpus étant constitué de textes ou d'extraits de textes, il est difficile de faire totalement l'impasse sur le fait que ces textes ont été rédigés dans des situations particulières qui impliquaient des protagonistes ayant des intentions particulières. Mais il est tout aussi difficile de ne pas tenir compte du fait que l'interprétation des résultats est elle-même située, c'est-à-dire qu'elle obéit à une intention qui intervient dans la lecture des résultats. Le mode de prise en compte de ces situations permet de mettre en lumière des points de vue d'analyse différents.

La distance qu'il est possible d'établir entre les données textuelles et la situation dans laquelle elles sont produites constitue une des problématiques majeures de l'analyse linguistique à partir de corpus. Il s'agit de comprendre s'il est possible de décontextualiser les phénomènes dans la perspective de constituer des règles de fonctionnement (d'une certaine façon, la question est de savoir si l'on peut constituer une compétence de locuteur à partir de productions réelles, ce à quoi Chomsky ne croyait pas (McEnery *et al*., 1996)) ou bien si cet objectif est voué à l'échec parce que les phénomènes langagiers sont totalement imbriqués avec la situation dans laquelle ils ont été produits et analysés. Cette dernière position est assez proche de celle de l'ethnométhodologie ; pour Garfinkel par exemple « le langage naturel ne peut faire sens indépendamment de ses conditions d'usage et d'énonciation » (Coulon, 2002). Cette position ne va pas sans poser de problèmes sur les possibilités de systématiser et de généraliser les résultats. Deux voies semblent possibles : soit essayer de généraliser conjointement les situations et les

productions langagières, ce qui revient à travailler la notion de genre, soit distinguer, parmi les phénomènes langagiers, ceux qui peuvent être décrits de manière systématique parce qu'ils sont moins dépendants du contexte (par exemple, certains fonctionnement syntaxiques ou morphologiques et certains phénomènes sémantiques) de ceux qui sont totalement liés à une situation donnée (tous ceux qui relèvent de la négociation sur le sens des mots par exemple) (Mondada, cet ouvrage). A l'opposé de cette vision très située du fonctionnement langagier, qui, dans la perspective d'une généralisation, oblige à des études très coûteuses en temps (comparaison de corpus) pour étudier les possibilités de décontextualisation, on trouve des travaux qui visent plutôt la validation d'une hypothèse et qui, dans un premier temps tout au moins, tiennent assez peu compte de l'origine et de la nature du corpus qu'ils mettent en œuvre. A ce moment de l'analyse, il s'agit de confirmer ou d'affiner une hypothèse sur un corpus, certes considéré comme homogène mais sans que l'objectif de l'étude concerne le rapport entre la nature des résultats et la nature du corpus (Lamiroy et Charolles, cet ouvrage). C'est souvent dans un second temps que cette question se pose, au moment où l'on souhaite vérifier les résultats sur un autre corpus.

La prise en compte de la situation de production des textes oblige à s'interroger aussi sur leur situation d'interprétation, c'est-à-dire sur le mode de production des résultats et leur validation. Tout d'abord, les résultats à produire relèvent parfois d'un choix (Prévost, cet ouvrage) ; ensuite, dans le cas d'analyses assistées par des outils, le mode de visualisation des résultats est particulièrement important car il peut influencer directement le mode d'interprétation ; c'est tout particulièrement le cas avec les outils d'analyse statistique (Viprey, Habert *et al*., cet ouvrage). L'interprétation des résultats est elle-même à replacer dans le cadre général de la validation. Il faut reconnaître que traditionnellement, cette question est assez peu discutée dans le milieu académique des sciences du langage pour qui elle est assimilée à l'obtention d'une sorte de consensus parmi les pairs. La confrontation avec la réalité des usages oblige le linguiste à une modification de ce point de vue même si, au bout du compte, dans la perspective de l'inscription dans une discipline, c'est quand même l'appréciation des pairs qui décide de la validité d'une étude. Mais cette validation ultime aura été précédée d'un grand nombre d'études qui passent inévitablement par l'examen de nouveaux corpus, censés avoir les mêmes caractéristiques, et par l'obtention de résultats similaires. Encore faut-il préciser ce qu'on appelle « mêmes caractéristiques » pour un corpus et « résultats similaires ». La notion de similarité est sans doute à nuancer en fonction de la nature de l'étude : une variation sur un élément lexical (mot qui n'avait pas été prévu par le modèle) est

acceptable dans le cas d'une étude sur un fonctionnement d'ordre structurel (par exemple structures correspondant au marquage de relations conceptuelles) mais l'apparition d'une structure nouvelle correspondant à une certaine information, dans un nouveau corpus, est plus problématique et amène à s'interroger à la fois sur les résultats obtenus et sur la caractérisation du corpus. On le voit, par rapport à une vision plus classique (introspective) la notion de connaissance linguistique est à relativiser, ce qui peut produire un effet de déstabilisation chez le linguiste, qui a l'impression que ce qu'il pensait être son savoir lui échappe. Pour le TAL ou l'ingénierie des connaissances , l'objectif applicatif constitue d'emblée un mode de validation : lorsque la satisfaction d'un besoin identifié initialement a été atteinte, la validation de la pertinence des méthodes et des outils est acquise. Ce point de vue est très orienté vers l'ingénierie mais lorsqu'il s'agit d'outils à destination des linguistes, la question est déjà moins facile à traiter puisque le linguiste est dans une perspective de construction de connaissances dont il n'a pas une idée précise *a priori* (Habert, 2004). Un bon outil est alors sans doute celui qui lui permet d'accéder au maximum d'éléments pertinents pour son étude sans en laisser de côté. Dans une telle perspective, il est évident qu'une part importante de bruit est acceptable puisqu'il s'agit aussi de comprendre quelle est la nature de ce bruit mais le silence est inacceptable.

1.4. Conclusion

L'apparition et le développement rapide des techniques d'analyse automatique de textes crée une situation inédite pour la recherche, dans toutes les disciplines scientifiques (et en particulier les disciplines des sciences humaines) qui ont le discours pour objet ou pour support de réflexion. L'accès très rapide et parfois très sophistiqué à des données textuelles autorisé par les outils de TAL ou d'analyse statistique (même s'il peut être biaisé par les possibilités de ces outils ou les modes de mesure choisis) produit une accélération dans l'évaluation de certaines hypothèses et une déstabilisation des acquis en linguistique. Mais la mise à distance produite par l'utilisation des outils donne un effet de réel salutaire. Cependant, cet éloignement rend plus cruciales encore des questions sur le fonctionnement sémantique qui se posent dès que l'étude des phénomènes langagiers ne se fait plus seulement de manière introspective mais à partir de productions réelles. Ces questions peuvent être synthétisées en une seule : quelle est la validité des résultats obtenus à partir d'un corpus au regard d'une possibilité de systématisation et de prédiction du fonctionnement linguistique (la question corrélative étant : est-ce que

des résultats qui ne sont valides qu'au regard d'un seul corpus (voire d'un unique objectif) relèvent d'une approche scientifique ?). Concrètement, dans la perspective de l'utilisation des méthodes automatiques, cette question trouve son expression la plus nette dans le problème de la catégorisation. En effet, la généralisation des résultats, qui constitue une forme de garantie de systématicité, passe par la possibilité ou non de considérer une occurrence, au sens large, comme ressortissant d'un ensemble d'autres qui lui sont équivalentes. Ce problème se décline en deux points de vue d'ailleurs complémentaires ; l'un concerne la nature des éléments à catégoriser, l'autre concerne le moment de la catégorisation. Pour ce qui concerne la nature des éléments, deux types peuvent être considérés, les mots d'une part, les textes d'autre part. Pour les mots, la catégorisation se réalise par l'attribution d'une valeur grammaticale ou sémantique ; pour les textes, elle se fait par le rattachement à un genre. La véritable question pour la linguistique de corpus, question qui est particulièrement mise en lumière lors de la réflexion sur les possibilités d'automatisation des études sur corpus, concerne le moment où la catégorisation se fait : avant ou après l'occurrence. Il s'agit donc de savoir si la catégorisation est un point de départ et préexiste à l'analyse ou si elle est un résultat de l'analyse, ce qui revient à s'interroger sur la connaissance qui préexiste à l'étude et sur celle qui est construite par elle. On peut ainsi considérer que l'objectif d'une sémantique de corpus consiste à comprendre les liens entre catégorisation des mots et catégorisation des textes (genres) d'une part et, d'autre part entre catégorisation *a priori* et catégorisation *a posteriori*.

Les différents chapitres de cet ouvrage donnent un éclairage sur différentes facettes de la question des relations entre sémantique et corpus. Il s'agit d'un problème en pleine évolution qui voit foisonner les approches et les résultats d'où, peut-être, l'impression de grande diversité donné par l'ensemble de l'ouvrage. Nous espérons que cette introduction aura montré en quoi ce foisonnement est cohérent et prometteur.

1.5. Bibliographie

Adam, J.-M. (1999). Linguistique textuelle. Des genres de discours aux textes. Paris : Nathan.

Auroux, S. (1998). *La raison, le langage et les normes*. Paris : PUF.

Auroux, S., Bernard, G. & Boulle, J. (2000). Le développement du comparatisme indo-européen. In S. Auroux (Ed.), *Histoire des idées linguistiques* (pp. 155-172). Liège-Bruxelles : Pierre Mardaga.

Aussenac-Gilles, N. & Condamines, A. (2003). *Rapport final de l'action spécifique* « *Corpus et Terminologie* ». http : http : //www.irit.fr/ASSTICCOT/.

Bakhtine, M. (1984). *Esthétique de la création verbale*. Paris : Gallimard, Tel.

Benzécri, J.-P. & Benzécri, F. (1984). *Pratique de l'analyse des données*. Paris : Dunod, 2è édition.

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press.

Biber, D. (1993). *Using register-diversified corpora for general language studies*, *Computational Linguistics*, 19(2), 243-258.

Biber, D. (1996). Investigating language use through corpus-based analyses of association patterns. *International Journal of Corpus Linguistics*, 1(2), 171-197.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Grammar of Spoken and Written English*. London : Longman.

Bilger, M. (Ed.). (2000). *Corpus : Méthodologie et applications linguistiques*. Paris : Honoré Champion.

Bouquet, S. (Ed.). (2004). Les genres de la parole. *Langages*, 153, mars 2004.

Bronckart, J.-P. (1996). *Activités langagières, textes et discours*. Lausanne : Delachaux et Niestlé.

Branca-Rosoff, S. (Ed.). (1999). *Langage et Société n°87, Types, modes et genres de discours*.

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

Condamines, A. (2002). Corpus Analysis and Conceptual Relation Patterns. *Terminology*, 8(1), 141-162.

Coulon, A. (2002). *L'ethnométhodologie*. Paris : PUF, collection Que sais-je.

Foucault, M. (1966). *Les mots et les choses*. Paris : Tel, Gallimard.

Gadet, F. & Pécheux, M. (1981). *La langue introuvable*. Paris : Pierre Mardaga.

Guiraud, P. (1960). *Problèmes et méthodes de la statistique linguistique*. Paris : PUF.

Habert, B. (Ed.). (2004). Linguistique et Informatique : nouveaux défis. *Revue Française de Linguistique Appliquée*, 2004-1.

Habert, B., Nazarenko, A. & Salem, A. (1997). *Les linguistiques de corpus*. Paris : Armand Colin.

McEnery, T. & Wilson, A. (1996). *Corpus Linguistics*. Edinburg : Edinburgh University Press.

Lagerwerf, L., Wilbert, S. & Degand, L. (Eds). (2003). *Mutidisciplinary Approaches to Discourse. Determination of Information and Tenor in TextsStichting*. Neerlandistiek Amsterdam and Nodus Publikationen Münster.

Muller, C. (1967). *Etude de statistique lexicale, Le vocabulaire du Théâtre de Pierre Corneille*. Paris : Larousse.

Pécheux, M. (1984). Sur le contexte épistémologique de l'analyse du discours. *Mots,* 9, 7-17.

Pierrel, J.-M. (Ed.). 2000. *Ingénierie des langues*. Paris : Hermès, Traité IC.

Quirk, R., Greenbaum, S., Leech, G & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London : Longman.

Rastier, F. (2001). *Arts et Sciences du texte*. Paris : PUF, formes sémiotiques.

Sinclair, J. (1991). *Corpus, Concordance. Collocation*. Oxford University Press.

Todorov, T. (1981). *Mikhaïl Bakhtine, le principe dialogique*. Paris : Seuil.

Trabant, J. (1995). *Humboldt ou le sens du langage*. Paris : Pierre Mardaga.

Swales, J.-M. (1990). *Genre Analysis, English in Academic and research settings*. Cambridge University Press.

Van Dijk, T. A. (Ed.). (1997). *Discourses Structure and Process*. London : Thousand Oaks, New Delhi : Sage Publications.