ENJEUX ÉPISTÉMOLOGIQUES DE LA LINGUISTIQUE DE CORPUS

François RASTIER
C.N.R.S.

http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html

(Ce texte est issu d'une conférence aux deuxièmes Journées de Linguistique de Corpus, Lorient, septembre 2002. Il sera recueilli dans les Actes à paraître sous la direction de Geoffrey Williams
aux Presses Universitaires de Rennes)

SOMMAIRE :

Résumé : Étendant le champ d'investigation de la linguistique générale, la linguistique de corpus lui permet tout à la fois une reconception de son objet et de ses théories. D'une part, elle permet de construire une observation des normes, chaînon manquant entre linguistique de la langue et linguistique de la parole. D'autre part, elle permet de mieux concevoir la sémiosis textuelle, en mettant en évidence des corrélations multiples entre plan du contenu et plan de l'expression. Elle périme enfin des conceptions obsolètes, en s'opposant au modèle newtonien de la science, et en permettant à la linguistique de s'intégrer pleinement aux sciences de la culture.

I. Pour approfondir le concept de corpus

En plein essor, la linguistique de corpus ne constitue aucunement un domaine de recherche unifié, mais j'ai choisi dans cette étude de privilégier les développements qui me paraissent revêtir une portée épistémologique particulière.

Deux conceptions du corpus : sac de mots ou archive de textes ? — Un auteur célèbre (Sinclair) définissait le corpus comme un « vaste ensemble de mots ». Cependant, l'objet empirique de la linguistique est fait de *textes* oraux ou écrits, non de mots ou de phrases - qui ne s'observent pas à l'état isolé, et, même quand on les isole, restent toujours relatifs à un genre et un discours. Certes la tradition logico-grammaticale, préoccupée des problèmes métaphysiques de la référence et de la vérité, a fait du mot (lieu de la référence) et de la proposition (lieu de la vérité) des horizons indépassables. Convenons toutefois que si le mot,

ou mieux le morphème, est l'unité élémentaire, le texte est pour une linguistique évoluée l'unité *minimale*, et le corpus l'ensemble dans lequel cette unité prend son sens.

Dans la tradition, la notion de corpus a d'abord été définie de manière canonique, dans les domaines religieux, juridique, voire littéraire. Elle a été élaborée par des disciplines injustement oubliées, du moins dans le domaine des Traitements automatiques du langage, que sont la philologie et l'herméneutique. À cette conception canonique, on semble préférer aujourd'hui une conception éclectique. Cependant, un corpus n'est pas plus un sac de mots qu'un nébuleux intertexte. Il est *structuré* d'une part en fonction d'une typologie des textes, qui se réflète dans leur codage, et d'autre part, dans chaque utilisation, par des sélections raisonnées de sous-corpus.

La structure du corpus dépend ordinairement de deux conceptions. L'une, *documentaire*, ne retient que des variables globales caractérisant les documents, sans tenir compte de leur caractère textuel, ni de leur structure. Dans cette conception *logico-grammaticale*, le corpus se résume à un échantillon de la langue, un réservoir d'exemples ou d'attestations.

En revanche, la conception *philologique-herméneutique* tient compte des rapports de texte à texte, ce qui n'est possible qu'au sein d'un discours. Damon Mayaffre (2002) parle à ce propos de *corpus réflexifs* : leur composition est réfléchie selon le principe critique qui est traditionnellement celui de la philologie, numérique ou non (cf. l'auteur, 2001, ch. 2).

Définir le corpus. — Convenons d'une définition positive.

Un corpus est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications.

Tout corpus suppose en effet une préconception des applications, fussent-elles simplement documentaires, en vue desquelles il est rassemblé : elle détermine le choix des textes, mais aussi leur mode de « nettoyage », leur codage, leur étiquetage ; enfin, la structuration même du corpus. Allons plus loin, un corpus doit « être aimé » : s'il ne correspond pas à un besoin voire un désir intellectuel ou scientifique, il se périme et devient obsolète.

Cette dépendance à l'égard d'une application ou d'une gamme d'applications permet de dédramatiser les problèmes récurrents de la représentativité et de l'homogénéité. Aucun corpus ne représente la langue : ni la langue fonctionnelle qui fait l'objet de la description linguistique, ni la langue historique, qui comprend l'ensemble des documents disponibles dans une langue. En revanche, un corpus est adéquat ou non à une tâche en fonction de laquelle on peut déterminer les critères de sa représentativité et de son homogénéité. La linguistique de corpus peut ainsi être objective, mais non objectiviste, puisque tout corpus dépend étroitement du point de vue qui a présidé à sa constitution.

Notre définition suppose qu'un corpus n'est pas un corpus de mots (cf. à l'inverse le projet européen Paroles) ; ni un corpus d'attestations ou d'exemples (comme Frantext, dès lors qu'on n'a pas accès aux textes-sources) ; ni un corpus de fragments (comme le *British National Corpus*, qui ne contient aucun texte complet, mais un échantillonnage).

De fait, tout regroupement de textes ne mérite pas le nom de corpus. Ainsi une *banque textuelle* peut regrouper des textes numériques de statuts divers : aucun critère linguistique ne permet cependant leur totalisation, sauf l'hypothèse que la langue leur confèrerait une unité *a priori* ; mais même organisée en base de données, une banque textuelle ne devient pas pour autant un corpus.

Un *hypertexte* n'est pas non plus par nature un corpus : soit c'est l'équivalent numérique d'un codex dont les renvois internes sont des liens hypertexte ; soit c'est l'hypertexte indéfini et il se confond avec le web — qui n'est pas un corpus mais, un euphémisme s'impose, une aire de stockage, voire une décharge publique. Or le texte ne se confond pas avec l'intertexte, et l'intertexte lui-même ne devient opératoire que s'il est structuré.

Enfin, des *œuvres complètes* ne constituent pas non plus un corpus, dans la mesure où elles contiennent des textes de statut hétérogène (ex. la traduction de Kleist dans les *Œuvres complètes* de Julien Gracq) [1]. Les travaux de Brunet et Muller ont d'ailleurs montré que la variable d'auteur est hiérarchiquement inférieure à la variable de genre [2].

Codage de variables globales. — Pour documenter un en-tête, il faut au moins référencer le texte à trois niveaux. En bref : les *discours* (ex. juridique *vs* littéraire *vs* scientifique), le *champ générique* (ex. théâtre, poésie, genres narratifs), le *genre* proprement dit (ex. comédie, roman « sérieux », roman policier, nouvelle, conte, récit de voyage). Le *sous-genre* (ex. roman par lettres) constitue un niveau encore subordonné. Les différences de statut épistémologique entre ces niveaux font qu'on ne peut, sauf simplification didactique, les représenter par une simple arborescence [3].

Un champ générique est un groupe de genres qui contrastent voire rivalisent dans une pratique : par exemple, au sein du discours littéraire, à l'époque classique, le champ générique du théâtre se divisait en farce, comédie, comédie héroïque et tragédie.

Pour étudier un texte, le « bon corpus » est d'abord constitué des textes qui partagent le même genre. Un corpus de champ générique est déjà hétérogène (par exemple, la tragédie et la comédie communiquent peu). Enfin, un corpus de discours comprend en général des textes écrits et oraux, ce qui instaure une hétérogénéité de principe.

De l'archive au corpus de travail. — Il semble utile de distinguer quatre niveaux.

1/ L'*archive* contient l'ensemble des documents accessibles. Elle n'est pas un corpus, parce qu'elle n'est pas constituée pour une recherche déterminée.

2/ Le *corpus de référence* est constitué par ensemble de textes sur lequel on va contraster les corpus d'étude.

3/ Le *corpus d'étude* est délimité par les besoins de l'application.

4/ Enfin le *sous-corpus de travail en cours* varie selon les phases de l'étude et peut ne contenir que des passages pertinents du texte ou des textes étudiés [4]. Par exemple, dans *L'analyse thématique des données textuelles — l'exemple des sentiments*, l'archive est la banque Frantext, le corpus de référence est constitué de 350 romans publiés entre 1830 et 1970, le corpus d'étude est constitué des passages contenant des noms de sentiments, et les corpus d'élection sont les corpus propres à tel ou tel sentiment.

Pour certaines applications informatiques, il faut encore distinguer, au cours de leur développement, le corpus d'apprentissage, le corpus de test, le corpus de validation.

Il faudrait enfin élaborer une description des parcours interprétatifs pour une tâche donnée — nous retrouvons là un des problèmes majeurs de l'herméneutique. Les propositions de Thomas Beauvisage (2004, ch. I) pour constituer une sémantique *percursive*, en suivant les parcours des internautes, pourraient avantageusement être étendues au travail sur corpus.

II. Pour une conception non-antinomique de la dualité langue / parole

Linguistique(s) de la langue, de la parole et des normes. — Traditionnellement, le rapport entre une grammaire et les productions linguistiques qu'elle règle est conçu comme un rapport entre la *puissance* et l'*acte* (dans la tradition aristotélicienne), ou encore entre *energeia* et *ergon* (selon Humboldt qui la reprend), ou enfin entre *compétence* et *performance* (selon Chomsky, qui se recommandait de Humboldt sur ce point). Or, faute peut-être de l'avoir posé de façon satisfaisante, aucune théorie linguistique n'est parvenue à résoudre le problème de leur articulation ; par exemple, les théories génératives n'ont pu restreindre la capacité générative des règles et se sont avérées trop puissantes pour être efficaces, c'est-à-dire pour générer des productions vraisemblables (grammaticales et acceptables).

Si l'on convient la puissance ne préexiste pas à l'acte (cf. l'auteur, 2003b), la langue ne préexiste pas à la parole : elle est apprise en son sein, et la compétence des sujets évolue au cours de leurs pratiques effectives.

Le "chaînon manquant" entre la langue et la parole est constitué par l'espace des normes (cf. Coseriu, 1969). Or, seule la linguistique de corpus peut offrir les moyens théorique et technique d'étudier l'espace des normes et de transformer en dualité l'antinomie entre compétence et performance. Pour cela il faut mener une étude comparative, tant des discours que des champs génériques et des genres, voire des styles – c'est là un aboutissement de la problématique de la linguistique comparée.

La lecture des leçons et manuscrits de Saussure confirme que Bally, dans son édition du *Cours de linguistique générale*, a gommé l'apport de Saussure à la linguistique de la parole. Cependant, même dans la tradition saussurienne, les deux linguistiques, celle de la langue et celle de la parole, sont restées séparées parce qu'une linguistique des normes n'a pas encore été édifiée.

On pense à tort qu'il ne peut y avoir de science des normes : elle serait une déontologie qui échapperait par son caractère relatif et conditionné à l'imaginaire logico-grammatical des règles, voire à l'imaginaire scientiste des lois. Aussi, le rapport entre langue et parole reste-t-il généralement conçu tantôt comme un passage du virtuel à l'actuel, tantôt comme le passage de contraintes à une liberté : dans les deux cas, on peine à concilier les virtualités impératives de la langue avec les libertés actuelles de la parole. En effet, passer de la langue, conçue abstraitement, à la parole, n'est pas seulement décliner des degrés de systématicité décroissants, mais aussi des statuts épistémologiques divers.

Si l'on prend la mesure des diversités effectives des discours, champs génériques et genres, le noyau invariant qu'on peut appeler *langue* se réduit drastiquement à l'inventaire des morphèmes, à des contraintes comme la structure de la syllabe, la structure du syntagme, etc. ;

par exemple, les lexèmes n'en font pas partie, car ils sont déjà des phénomènes de « discours » [5].

Aucun texte n'est écrit seulement « dans une langue » : il est écrit dans un genre et au sein d'un discours, en tenant compte évidemment des contraintes d'une langue. Au demeurant, l'analogie relative des pratiques et celle des discours et des genres qui en découle permet la traduction voire tout simplement l'intercompréhension. D'où la nécessité de tenir compte des genres et des discours dans toute étude de textes en linguistique contrastive.

Les deux linguistiques, celle de la langue et celle de la parole, que Saussure cherchait explicitement à articuler, restent unies par l'espace des normes. Les niveaux des genres, champs génériques et discours sont bien les niveaux stratégiques qui permettent de passer de la généralité de la langue aux particularités des textes, car les relations sémantiques entre textes s'établissent préférentiellement entre textes du même genre, du même champ générique et du même discours [6].

Entre l'espace normatif des règles et le désordre apparent des usages, entre l'universel de la langue et la singularité des emplois, l'espace des normes s'étend de la généralité de la doxa jusqu'à la particularité du paradoxe. La dualité langue/parole n'est évidemment pas une contradiction. De fait, les règles de la langue sont sans doute des normes invétérées et les performances de la parole ne restent évidemment pas exemptes de normativité : elles instancient et manifestent les règles de la langue et diverses normes sociolectales.

Bref, la linguistique peut prendre *de droit* pour objet de description l'espace des normes : au lieu de les édicter, comme elle le faisait naguère en frappant d'inacceptabilité des énoncés, alors même qu'ils sont attestés, elle doit les décrire et pour cela exploiter des corpus [7].

Or, l'étude des corpus montre que le lexique, la morphosyntaxe, la manière dont se posent les problèmes sémantiques de l'ambiguïté et de l'implicite, tout cela varie avec les genres, les champs génériques et les discours. Les applications doivent tenir compte de ces spécificités.

Elles sont telles que les projets de systèmes universels semblent alors irréalistes, linguistiquement parlant. Pour parvenir à des traitements automatiques spécifiques et efficaces de corpus et adapter les stratégies applicatives, il convient de spécifier les fonctionnements propres aux différents discours, champs génériques et genres textuels. Dans un corpus homogène, connaître les régularités structurelles du genre peut permettre de simplifier les traitements ; par exemple, certaines parties des textes peuvent parfois être systématiquement éliminées, au profit des sous-corpus pertinents pour une tâche donnée.

On note certes des régularités transgénériques et transdiscursives. Par exemple, des domaines comme la littérature et les essais sont voisins : il arrive que les mêmes auteurs y transposent des thèmes comparables. Au palier morphosyntaxique et au plan de l'expression, ces régularités relèvent de la langue ; au plan sémantique, elles relèvent de l'idéologie ou de la doxa [8].

Cependant, alors que la morphosyntaxe reste pour l'essentiel affaire de règles (bien qu'elle ne soit aucunement indifférente aux normes), la sémantique demeure pour l'essentiel affaire de normes. Avec les méthodes de la linguistique de corpus, on dispose à présent de moyens nouveaux pour tester les hypothèses sur le rapport entre normes et règles, comme sur le

rapport entre les deux plans du langage (signifiant et signifié ; expression et contenu). Nous allons détailler ces points.

Paliers, niveaux, et remembrement. — À cause de la séparation infondée entre syntaxe, sémantique et pragmatique [9], on constate une double séparation, de fait et non de droit, entre niveaux du langage et entre paliers du langage.

(i) Entre *niveaux* du langage : la syntaxe et la sémantique sont longtemps demeurées séparées et l'articulation entre sémantique et syntaxe restait d'autant plus problématique. De fait, depuis vingt ans, l'articulation entre sémantique et syntaxe est de plus en plus étudiée (tant du côté des grammaires d'unification, des grammaires cognitives, de la sémantique verbale). Par ailleurs, dès lors qu'elles sont bien faites, sémantique et pragmatique deviennent indissolubles voire indiscernables.

(ii) Entre *paliers* (ou degrés de complexité) : on a conçu des lexicologies et des syntaxes indépendantes du palier du texte, en déléguant ce palier d'analyse à d'autres disciplines, notamment littéraires. Les acquis de la linguistique textuelle engagent à reconsidérer cette séparation. Même au palier du mot, les variations selon les discours et les genres sont fort importantes, tant au plan sémantique qu'au plan de l'expression. Par ailleurs, on avait sans doute opposé trop vite une morphologie et une syntaxe relevant d'une " linguistique de la langue " à une linguistique textuelle relevant de la " linguistique du 'discours' (ou de la parole) ".

En somme, centrée sur le problème du signe, la linguistique peine à concevoir le rapport entre signe et texte, et plus généralement le rapport du local au global, médiatisé par plusieurs paliers d'organisation de complexité croissante (le syntagme, la période, etc.).

Par ailleurs, la relation signifié/signifiant constitutive du signe ne va aucunement de soi, n'a rien d'inconditionné et reste médiatisée par les structures textuelles, considérées tant au plan du contenu qu'au plan de l'expression. Or, précisément la linguistique de corpus permet l'étude du rapport global / local tant au plan du contenu que de l'expression, que ces plans soient considérés séparément ou dans leur interrelation constitutive de la sémiosis textuelle [10].

Un peu comme l'imagerie cérébrale pour les neurosciences, les instruments de la linguistique de corpus permettent d'étendre le champ des observables et de tester des hypothèses. Sans prétendre ici à une synthèse, voici quelques exemples de corrélations entre paliers de complexité et entre plans du langage.

Les corrélations entre paliers de complexité. — On distingue les paliers microtextuel (morphème, lexie), mésotextuel (de la période au chapitre), macrotextuel (texte complet dont péritexte et paratexte), intertextuel (le corpus). Voici divers exemples de corrélations entre ces paliers.

a- *Entre lexique et genres*. — Dans le roman, *amour* a pour antonyme *mariage* ou *argent*. En poésie, point de mariage, ni d'argent (pour une analyse en corpus dans la banque Frantext, cf. Bourion, 2001, pp. 42-45). Les acceptions dépendent ainsi des genres et champs génériques.

b- *Entre lexies et formes textuelles*. — Bourion (2001) a montré que dans la banque Frantext *au pied de* et *aux pieds de* n'ont aucun contexte commun. Le singulier correspond toujours à

une localisation, et souvent à une description, alors que le pluriel se place toujours dans une scène d'imploration, et donc dans une configuration narrative. En bonne règle, il faudrait donc prévoir deux entrées de dictionnaire distinctes pour cette expression, selon qu'elle se trouve au singulier ou au pluriel.

c- *Entre thématique et positions dans le texte.* — Dans *Le Père Goriot*, le mot *amour* revient, sauf au chapitre IV, ou pourtant ce trop bon père meurt pour l'amour de ses filles : mais il est « remplacé » par le pronom *elles*, qui revient alors avec une fréquence anormale [11].

d- *Entre variables globales de discours et de genre et variables morphosyntaxiques*.— Des résultats récents confirment l'incidence du genre sur les variations morphosyntaxiques (cf. Malrieu et Rastier, 2001). À partir d'un corpus de 2.600 textes complets classés par genres et discours et étiquetés par 251 types d'étiquettes, morphosyntaxiques pour la plupart, on a retrouvé et validé les différents niveaux de classification présentés ci-dessus, en utilisant des pourcentages calculés sur les étiquettes. On a conduit des analyses univariées pour qualifier les variations selon les catégories d'étiquettes, puis une analyse multivariée utilisant des méthodes de classification automatique. Les résultats, probants, confirment la corrélation entre les variables globales de genre, champ générique et discours d'une part, et d'autre part les variables morphosyntaxiques, locales par définition.

Malgré les théories formalistes sur l'autonomie de la syntaxe, la *corrélation syntaxe / sémantique* est ainsi assez forte pour qu'une catégorisation morphosyntaxique permette de recouper massivement une classification préalable des genres, qui relève essentiellement de critères sémantiques globaux. La relation confirmée entre local et global témoigne évidemment de l'unité de fait et de droit entre la linguistique du signe et de la phrase et la linguistique des textes. Plus généralement, la corrélation entre descriptions locales et description globale permet de préciser l'articulation entre la problématique du signe et la problématique du texte, en subordonnant la première à la seconde.

e- *Entre sections du texte et catégories morphosyntaxiques.*— Le balisage des sections du texte reste une condition indispensable à toute analyse fine : par exemple, dans un roman classique, les parties attribuées au narrateur et les paroles des personnages obéissent évidemment à des régimes morphosyntaxiques très différents.

De même, dans un article de linguistique, le système des temps et des personnes diffère considérablement selon qu'il s'agit des exemples ou du corps du texte : dans les exemples, on relève le double de *je*, moitié moins de *nous*, mais le triple de passés composés. Cela rapproche d'autant plus les exemples de linguistique des textes romanesques ou autobiographiques que les mots concrets y sont en outre plus nombreux ; or, comme l'a montré Biber, ils restent paradoxalement plus fréquents dans les textes de fiction [12].

f- *Entre texte et intertexte.*— L'accès à de grands corpus permet d'étudier avec des moyens nouveaux la stéréotypie textuelle et les normes de la doxa. L'exemple le plus simple est celui de la canonicité : dans le corpus roman 1830-1970 de la banque Frantext, qui compte environ 350 œuvres, on trouve seulement 5 sortes de fractions de seconde, et 12 nombres de secondes (sur une infinité théoriquement possible). Sur 4488 mentions d'âge — 2650 hommes (59 %) et 1838 (41%) femmes — , certains âges n'apparaissent pas : 41 ans pour les femmes (en revanche 40 ans est un âge canonique), 49 ans pour les hommes (en revanche 50 ans est un âge canonique), 71 ans ou encore 92 ans ; d'autres sont sur-représentés, par exemple 15, 18 et

20 ans pour les deux sexes ; 16 ans pour les personnages féminins (résultats dus à Nathalie Deza, 1999). Dans le roman français, on a presque toujours vingt ans..

Les complémentarités entre paliers de complexité sont ainsi illustrées par des phénomènes de *solidarité d'échelle* jusqu'ici peu étudiés en linguistique.

Les corrélations entre plans du contenu et de l'expression.— Elles ne sont pas moins remarquables que les corrélations entre paliers, et bien plus complexes que ne le laisse supposer le modèle binaire du signe (signifiant/signifié, expression/concept, etc.).

a- *Au plan graphique*. — Alors que la ponctuation n'est pas considérée comme sémantique et qu'elle est tout simplement absente des grammaires formelles, l'étude en corpus permet de souligner les corrélations entre contenus lexicaux et ponctèmes. Par exemple, dans un corpus romanesque, Evelyne Bourion (2001) a ainsi pu confirmer la corrélation entre des noms de sentiments et les ponctuations dans les contextes où ces noms apparaissent. Ainsi les sentiments ponctuels, brusques, comme la colère ou la joie, sont significativement associés aux points de suspension.

Pour sa part, dans une étude comparative sur Baudelaire, Maupassant, Proust et Duras (à paraître), Denise Malrieu a ainsi spécifié les contextes de *mer* : « On constate que les ponctèmes après *mer* sont plus fréquents qu'avant *mer*, ces derniers variant de 6% chez Proust à 32% chez Rimbaud, les ponctèmes après variant de 36,2% chez Rimbaud à 53% chez Proust ». Il s'agit sans doute, dans la langue littéraire, d'une convention stylistique qui engage à finir la phrase ou le membre de phrase par un sème /non-borné/ (des sémèmes comportant le sème /borné/ comme 'mur' n'entraînent aucun résultat comparable). Ce qui vaut pour les lexèmes vaut aussi pour les grammèmes. Par exemple, dans le corpus littéraire de la société Synapse (2600 textes), le point-virgule et l'imparfait du subjonctif sont associés par une corrélation de 0.44. Cela tient sans doute à leur emploi commun dans les passages d'analyse psychologique : l'imparfait du subjonctif comme le point-virgule sont pour ainsi dire imperfectifs et peuvent supposer un suspens critique.

b- *Au plan phonétique*. — On constate également des effets de solidarité entre paliers : ainsi, dans les tragédies de Racine, les phonèmes du nom du personnage principal, qui sert aussi son titre, sont significativement diffusés sur l'ensemble du texte (Valérie Beaudouin, 2002, 8.3.2). Ainsi, les éléments d'une forme phonique locale se trouvent diffusés pour constituer un fond perceptif global.

En outre, dans son analyse de Racine, Beaudouin (2002, § 8.3.3) a pu montrer que le champ sémantique de la mort était associé à des mètres anapestiques, et le champ sémantique de l'amour à des mètres iambiques (la mort est repos, donc les accents sont plus rares, alors que l'amour est passion, et se trouve associé à des accents plus fréquents). Mieux encore, le taux d'hémistiches irréguliers selon les actes semble corrélé à la structure narrative globale (Beaudouin, 2002, § 8.3.4).

Les corrélations entre plans du contenu et de l'expression rendent ainsi licite la notion de *contextualité hétéroplane* : le contexte d'une unité sur un plan, expression ou contenu, est constitué par d'autres unités sur le même plan, mais aussi sur l'autre. On ressent le besoin d'une théorie qui puisse penser ces corrélations, c'est-à-dire d'une linguistique informée par une sémiotique textuelle [13].

L'étude des textes littéraires est ici particulièrement révélatrice, car ils multiplient les rapports global / local (par des structures en abyme) et les rapports entre plans du contenu et de l'expression (cela est évident en poésie, notamment dans l'usage de la rime). Ainsi, par leur complexité, ils « signifient » plus, ce qui leur vaut sans doute d'être relus.

Mais les mêmes types de corrélations sont aussi à l'œuvre dans des corpus non littéraires. Ainsi le projet européen Princip.net de détection automatique de sites racistes met-il à profit des critères de « bas niveau » comme la ponctuation (un antiraciste ne redouble jamais un point d'exclamation), la casse (un antiraciste n'écrit jamais une phrase en majuscules), les polices de caractères, voire les codes html (les images sont caractéristiques des sites racistes, qui comptent également plus de bannières, etc.) [14].

Or les corrélations entre plans du contenu et de l'expression ont aussi un enjeu immédiat pour les applications comme la catégorisation de documents, la détection automatique de sites, etc. En pratique, elles permettent, dès lors que la catégorisation des documents du corpus d'apprentissage tient compte d'une classification évoluée, d'éviter des traitements sémantiques complexes et aléatoires. En effet, si l'on a identifié des éléments locaux de l'expression corrélés à des variables globales du contenu, on peut parvenir à des catégorisations automatiques fiables des textes.

Les méthodes de recueil et d'analyse de corpus ainsi mises au point s'adaptent aussi à d'autres sémiotiques. Ainsi, dans une tâche de catégorisation d'images fixes, on peut arriver à une identification du genre à partir d'un élément local de l'expression : par exemple, une photo *people* à gros grain et faible définition suppose une prise au téléobjectif et entre donc dans le genre de l'indiscrétion (cf. le projet Semindex, ENSTB).

Dépasser l'opposition entre « formel » et sémantique.— Un des problèmes fondamentaux que rencontre la linguistique de corpus reste l'interprétabilité des résultats, notamment ceux qu'obtiennent les méthodes quantitatives. Une sémantique de l'interprétation nous semble indispensable pour qualifier les résultats obtenus, car le détour interprétatif est une condition première de l'objectivation.

La première difficulté consiste à passer des chaînes de caractères à des constructions de formes sémantiques, ce qui suppose une déontologie et une méthodologie, sans quoi l'on en resterait à une lexicométrie somme toute limitée. Par exemple, les cooccurents lexicaux d'un mot-pôle doivent être qualifiés comme des corrélats sémantiques pour pouvoir être considérés comme des lexicalisations partielles d'un thème (cf. l'auteur, 2001, ch 8). Mais, si l'on réussit à construire des formes sémantiques, on peut parvenir en retour à sémantiser les unités de l'expression, même « de bas niveau » comme les ponctèmes ou les phonèmes. Ces rapports inaperçus entre les deux plans du langage sont l'objet privilégié de la sémiotique du texte.

La relation même entre les *plans du langage*, signifiant et signifié, contenu et expression, reçoit ainsi un éclairage nouveau. Certes, l'analyse morphosyntaxique par étiquetage automatique n'est pas purement « formelle » et s'appuie généralement sur un lexique qui contient des informations sémantiques. Mais cette étape franchie, elle ouvre la possibilité de mettre à jour des corrélations fortes entre régularités de l'expression et régularités du contenu.

Loin de se limiter aux textes littéraires, la corrélation confirmée entre variables globales comme le discours, le champ générique, le genre et les variables locales tant morphosyntaxiques que graphiques ou phonologiques, nous conduit à poser le problème de la

sémiosis textuelle. On définit ordinairement la sémiosis au palier du signe, et comme un rapport entre signifié et signifiant ; mais on ne s'interroge guère sur les paliers supérieurs, comme si leur sens se déduisait par composition de la signification des signes. Or, un genre définit précisément un rapport normé entre signifiant et signifié au palier textuel : par exemple, dans le genre de l'article scientifique, le premier paragraphe, sur le plan du signifiant, correspond ordinairement une introduction, sur le plan du signifié ; dans le genre de la nouvelle, il s'agit le plus souvent d'une description.

En somme, la sémiosis conditionnelle proposée par la langue aux paliers de complexité inférieurs, du morphème à la phrase, ne devient effective que si elle est compatible avec les normes de genre voire de style qui assurent la sémiosis textuelle.

Enfin, l'opposition humboldtienne entre la *forme intérieure* et la *forme extérieure* des textes, qui a fait couler tant d'encre chez les stylisticiens, pourrait recevoir une nouvelle formulation qui la relativise : la forme intérieure, loin d'être un mystère esthétique, est constituée par les régularités jusqu'à présent imperceptibles de la forme extérieure, celle de l'expression, que les moyens théoriques et techniques de la linguistique de corpus permettent à présent de mettre en évidence. En d'autres termes, le contenu d'un texte ne se réduit certes pas une mystérieuse représentation mentale : un texte est fait de deux plans, celui des formes sémantiques et celui des formes expressives [15], dont le genre notamment norme la mise en corrélation. Au sein de chaque plan s'établissent des relations forme / fond, de type gestaltiste, qui permettent la perception sémantique et phonologique.

III. Incidences sur la théorie linguistique

« Corpus linguistics does *not* exist » affirmait naguère Chomsky (entretien avec Baas Aarts, 1999). Amusant petit « meurtre symbolique », cette dénégation est l'aveu de l'échec épistémologique et pratique de la linguistique computationnelle, alors qu'elle s'était précisément constituée autour des théories chomskiennes successives, en assumant l'objectif explicite de les valider.

D'autres auteurs, plus soucieux de débat épistémologique, prennent la peine d'argumenter : « La recherche d'attestations dans des textes (quelles que soient sa sophistication et l'utilisation de moyens techniques coûteux, voire informatiques), la constitution d'un *corpus*... ne relèvent pas directement des protocoles expérimentaux. À cela deux raisons : i) elles ne sont pas en relation directe avec une hypothèse explicite à tester ; ii) elles ne correspondent pas à la production d'un phénomène » (Auroux, 1998 : 183). Ces affirmations, contredites par la pratique et la théorie de la linguistique de corpus, reflètent plusieurs mécompréhensions courantes qu'elles érigent en obstacles théoriques de principe. Ainsi, la linguistique de corpus reste sans rapport défini avec la recherche d'attestations. Et si faute de paramètres historico-culturels reproductibles, l'expérimentation reste impossible dans les sciences sociales [16], les observables produits par la linguistique de corpus sont bien des phénomènes nouveaux.

Au milieu des années 90, la linguistique de corpus, issue des humanités (et notamment de la collectivité réunie autour de la revue *Computers and the Humanities*), l'a emporté sur le courant orthodoxe-néochomskien de la linguistique computationnelle. Pour la linguistique de corpus, l'informatique n'est qu'un instrument, non un modèle théorique, car la linguistique appartient pleinement aux sciences de la culture.

Il serait fallacieux de postuler une épistémologie propre et donc une autonomie scientifique des « Traitements Automatiques du Langage ». Il n'est aucunement certain en effet que l'informatique, technologie sémiotique, soit en outre une science, car le traitement de l'information n'est pas un objet scientifique, mais un objectif ; et l'information reste un objet des mathématiques.

Un traitement constitue évidemment un objectif technique et non un objet scientifique : confondre les deux, c'est instituer une technoscience qui trouverait sa légitimité dans les outils. De fait, les TAL doivent leur scientificité à la linguistique dont ils constituent un secteur d'application et dont ils doivent pour ainsi dire « hériter les propriétés ».

La conception exclusive sinon fétichiste de la théorie s'est révélée néfaste, car elle a conduit à multiplier les chapelles, voire les grands prêtres. Elle est fausse, car elle suppose qu'une théorie linguistique puisse expliquer la totalité des faits de langage. Cet exclusivisme a connu un échec pratique et un échec théorique, souligné plutôt que masqué par la multiplication de modèles « formels » partiels et revendiqués comme tels. Paradoxalement, les TAL n'ont pris leur essor qu'en s'affranchissant des prétentions théoriques des grammaires qui entendaient les utiliser à des fins de validation.

Il faut donc passer du principe de plaisir théorique au principe de réalité philologique. Un nouveau rapport à l'empirique change non seulement l'étendue, mais la nature des faits, et rend nécessaire l'innovation théorique. Il permet de produire de nouveaux faits, qui naissent de la rencontre entre les modes d'observation et les modes d'explication.

Le nouveau rapport à l'empirique entraîne un nouveau rapport au théorique. Au plan épistémologique, il s'agit d'articuler plus clairement les rapports entre théorie et pratique. De fait, le face-à-face entre les règles et les exemples reste intrathéorique et ne débouche pas sur des applications. En revanche, dès lors que les applications deviennent déterminantes, les théories sont jugées et utilisées selon leur applicabilité, et deviennent modifiables en fonction des besoins. Elles doivent en outre prévoir les conditions de leur simplification, même drastique, en fonction des applications.

L'accès aux corpus conduit ainsi à modifier le rapport entre théorie et pratique, tant en amont du processus de recherche, dans la formulation des hypothèses, qu'en aval, dans la recherche de contre-exemples ou de variations. C'est aussi le moyen de sortir des apories théoriques suscitées par la philosophie du langage : par exemple, l'analyse de corpus reste le seul moyen éprouvé pour relativiser la polysémie et contrôler l'ambiguïté ; ou encore, pour déterminer les valeurs des formes grammaticales (par exemple, le futur n'a pas les mêmes valeurs dans le discours juridique que dans le roman).

Enfin, la linguistique de corpus, dès lors qu'elle adopte un point de vue réflexif à l'égard de ses propres démarches, peut permettre de rompre avec l'objectivisme candide : elle ne pratique pas d'analyse automatique des *données*, dans la mesure où elles doivent d'abord être qualifiées comme données, puis interprétées avant et après traitement : les données sont ce qu'on se donne.

Tout cela conduit à un remembrement de la linguistique et à un nouveau régime de l'interdisciplinarité. Ils sont favorisés par la " dé-idéologisation " qui conduit à distinguer des préconceptions implicites de l'ontologie et de la cognition les exigences théoriques et méthodologiques propres à la discipline. Les thèses " réalistes " qui supposaient un fondement

cognitif aux modèles linguistiques ont généralement été abandonnées. Dans la fusion progressive des TAL avec la linguistique de corpus, les points de débat deviennent pour l'essentiel méthodologiques.

Les faits nouveaux naguère inaperçus et à présent objectivés que nous avons mentionnés dans la seconde partie de cette étude revêtent une portée scientifique, car ils sont inconcevables pour les théories linguistiques les plus répandues. La plupart se fondent en effet sur la tripartition sémantique / syntaxe / pragmatique, pour les plus communes, et ne peuvent que renvoyer ces phénomènes hors de la linguistique, vers des études rhétoriques ou stylistiques. Cependant, l'exigence scientifique préfère encore les faits sans explication, défis et gages de renouvellement théorique, aux explications sans faits que prodiguent les grammaires universelles.

Même les théories qui se recommandent de la tradition saussurienne et admettent la bipartition entre plans du langage éprouvent des difficultés à décrire les relations qui unissent ces plans. Or, remarquait Saussure dans un manuscrit récemment retrouvé : « l'entreprise de classer les faits d'une langue se trouve donc devant ce problème : de classer des accouplements d'objets hétérogènes (signes-idées), nullement, comme on est porté à le supposer, de classer des objets simples et homogènes, ce qui serait le cas si on avait à classer des signes ou des idées. Il y a deux grammaires, dont l'une est partie de l'idée, et l'autre du signe ; elles sont fausses ou incomplètes toutes deux. » (2002 : 20).

C'est pourquoi il faut élaborer une théorie de la sémiosis textuelle, qui, loin d'être une lointaine extension de la linguistique, y occupe un rôle central, non seulement parce que le texte (oral ou écrit) est l'unité minimale d'étude, mais parce c'est elle qui détermine la sémiosis des paliers inférieurs et permet de concevoir l'unité du contenu et de l'expression (cf. l'auteur, 2004a).

Dans cet agenda, la sémantique des textes en corpus met ainsi l'accent sur deux complémentarités générales : (i) celle des niveaux de langage ou plans de description (ponctuation, morphologie, syntaxe, sémantique), et (ii) celle des paliers d'organisation et de complexité : mot, phrase, texte, intertexte.

On peut formuler l'hypothèse que paliers et niveaux correspondent à des variations objectives de complexité. Il reste cependant à problématiser ces variations sans explorer dans l'abstrait les complémentarités entre niveaux et paliers. En effet, les applications qui font l'objet d'une demande sociale croissante requièrent la mise en évidence de ces complémentarités : par exemple, reconnaître un type de texte par des caractéristiques lexicales ou morphologiques ; détecter un type de site web ; faire de l'analyse thématique assistée ; faire de la diffusion ciblée en définissant des proximités entre textes, etc. La plupart des applications supposent aujourd'hui des tâches de caractérisation : au sein d'un corpus, il s'agit de singulariser les éléments pertinents pour l'application. Dès lors, la linguistique renoue, par une voie nouvelle, avec la problématique de description des singularités, qui est propre aux sciences de la culture ; la description de lois, qui fut longtemps jugée la condition nécessaire de toute scientificité, se subordonne alors à l'étude systématique des usages effectifs.

NOTES

[1] Alors par exemple que le mot *femme* est resté rare dans les œuvres romanesques et critiques de Gracq, sa traduction de la *Penthésilée* de Kleist lui rend une fréquence rassurante (car l'héroïne était la reine des Amazones).

[2] Aussi, malgré l'intérêt des recherches récentes de Dominique Labbé, il nous semble impossible de prouver que Molière a écrit l'œuvre de Corneille ; non qu'il n'en ait pas été capable, mais la différence des genres entre les deux séries d'ouvrages est trop prégnante pour qu'un rapprochement puisse être concluant (cf. Brunet, 2004).

[3] Cf. l'auteur, 2001, ch. 8.

[4] Quand elles sont faites, ces distinctions philologiques sont souvent masquées : par exemple, Foucault partage avec la tradition nietzschéenne quelque mépris pour la philologie et masque sa dette théorique à son égard. Ce brouillage a eu d'importantes conséquences, puisqu'à sa suite, l'école française d'Analyse du Discours reprend de façon voilée les distinctions philologiques ; ainsi, Maingueneau (1991, pp. 159-159) nomme *univers discursif* l'archive, *champ discursif* le corpus de référence et *espace discursif* le corpus d'étude (cf. Pincemin, 1999). L'archéologie de ce brouillage mérite l'attention.

[5] C'est pourquoi le lexique, du moins celui des lexies, n'appartient pas à la langue. De fait, comme la syntagmatique relève de la parole, les mots sont aussi des formations textuelles (à la différence des morphèmes).

[6] Cela n'exclut pas, bien entendu, les relations entre discours.

[7] Elle n'a pas à se prononcer sur l'inacceptabilité : ce sont les théories normatives qui la créent et se rendent ainsi inacceptables. C'est par l'étude comparative systématique des textes que l'on peut restituer les normes linguistiques en vigueur. La langue est faite des invariants qui rendent comparables les éléments du corpus : il faut pour établir ses régularités vérifier des hypothèses d'isonomie (dans une synchronie) et d'homogénéité (malgré les variations de lieu et de registre).

[8] Cf. l'auteur, 2004b.

[9] Due à Morris et Carnap, elle est issue de la philosophie du positivisme logique et reste un obstacle épistémologique majeur pour la linguistique (cf. l'auteur, 1990).

[10] Cf. *infra*, et l'auteur, 2004a.

[11] Cf. Bourion, 2001. Il y a là une mise en garde à l'égard d'une thématique du mot-clé. Là où le thème revêt sa plus grande efficacité narrative, il perd sa lexicalisation privilégiée. La Duchesse de Guermantes disait d'ailleurs, à propos d'une forme d'amour moins dramatique : « on le fait, mais on n'en parle pas ».

[12] Au risque de persifler, on pourrait en conclure que l'exemple est ainsi un sous-genre « littéraire », transposé dans le discours linguistique, où la fiction « réaliste » tient lieu de réel empirique. On comprend mieux pourquoi, dans un corpus de 250 articles de linguistique française (1990-2000), seulement 0,5% des exemples sont attestés (cf. Poudat, 2003) : il ne

s'agit pas seulement de se protéger des « irrégularités » constatées dans les corpus, il faut encore concrétiser le monde de sens commun qui sert de norme aux théories de la référence.

[13] Dans les manuscrits de Saussure, les figures de contextualité hétéroplane sont remarquables ; soient par exemple deux signes voisins *A* et *B* : le signifié *a* pourra être corrélé au son *b*, et le signifié *b*, au son *a* (cf. 2002 : 290).

[14] Cf. l'auteur, à paraître.

[15] Cf. l'auteur, 2003a.

[16] Le refus implicite de tenir compte de la spécificité des sciences sociales reste lié au scientisme, qui a toujours postulé l'unité de la science.

BIBLIOGRAPHIE

Auroux, S., 1998, *La raison, le langage et les normes*, Paris, Presses Universitaires de France.

Beauvisage, T. 2004, Sémantique des parcours des utilisateurs sur le web, thèse de doctorat, Université Paris X.

Berkenkotter, C. & Huckin, T. N., éd., 1995, *Genre Knowledge in Disciplinary Communication*. Hillsdale (N. J.), Lawrence Erlbaum.

Bhatia, V. K., 1993, *Analysing Genre : Language Use in Professionnal Settings*, Londres, Longman.

Biber, D., 1988, *Variations across Speech and Writing*, Cambridge, CUP.

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

Biber, D., 1995, *Dimensions of register variation : a cross-linguistic comparison*. Cambridge: Cambridge University Press.

Bommier-Pincemin, B., 1999, *Diffusion ciblée automatique d'informations : conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*, Thèse de l'Université Paris IV, 805 p.

Bourion, E., 2001, *L'aide à l'interprétation des textes électroniques*, Thèse, Université de Nancy II. Ed. pdf. http://www.texto-revue.net

Brunet, E., 2004, Où l'on mesure la distance entre les distances, *Texto !* mars 2004 [en ligne]. Disponible sur : <http://www.revue-texto.net/Inedits/Brunet/Brunet_Distance.html>.

Coseriu, E., 1969, Sistema, norma, et 'parola', Studi linguistici in onore Vittorio Pisani, Brescia, Paideia Editrice, pp. 235-253.

Deza, N., 1999, *L'accès sémantique aux banques textuelles*, Thèse, Université de Nancy II.

Fløttum, K. et Rastier, F. éds., 2003, *Academic Discourse - Multidisciplinary Approaches*, Oslo, Novus.

Habert, B., Nazarenko, A. & Salem, A., 1997, *Les linguistiques de corpus*. Paris, Armand Colin — Masson.

Habert, B., Fabre, C. & Issac, F., 1998, *De l'écrit au numérique : constituer, normaliser, exploiter les corpus électroniques*. Paris, InterEditions — Masson.

Maingueneau, D., 1991, *L'analyse du discours*, Paris, Hachette.

Malrieu, D., à paraître, *Analyse sémantique du lexème* mer *dans un corpus de textes littéraires,*17 p.

Malrieu, D. et Rastier, F., 2001, Genres et variations morphosyntaxiques, *Traitements automatiques du langage*, 42, 2, pp. 547-577.

Mayaffre, D., 2002, Les corpus réflexifs : entre architextualité et intertextualité, *Corpus*, I, 1, pp. 51-70.

Poudat, C., 2003, Characterization of French linguistic research articles with morphosyntactic variables, in Fløttum, K. et Rastier, F. éds.

Rastier, F., 1987, *Sémantique interprétative,* Paris, Presses Universitaires de France.

Rastier, F., 1989, *Sens et Textualité,* Paris, Hachette.

Rastier, F., 1990, *La triade sémiotique, le trivium et la sémantique linguistique, Nouveaux actes sémiotiques,* 9, 54 p.

Rastier, F., 1991, *Sémantique et Recherches Cognitives,* Paris, Presses Universitaires de France.

Rastier, F., 1995, *L'analyse thématique des données textuelles — L'exemple des sentiments*, Paris, Didier.

Rastier, F., 2001, *Arts et sciences du texte*, Paris, PUF.

Rastier, F., 2003a, Formes sémantiques et textualité, in Unité(s) du texte, *Cahiers du Crisco*, Université de Caen, 12, pp. 99-114.

Rastier, F., 2003b, Parcours de production et d'interprétation : pour une conception unifiée dans une sémiotique de l'action, in Ouattara, A. (éd.), *Parcours énonciatifs et parcours interprétatifs. Théories et applications*, Paris, Ophrys (coll. HLD), pp. 221-242.

Rastier, F., 2004a, Poétique et textualité, *Langages*, 153, pp. 200-206.

Rastier, F., 2004b, Du lexique à la doxa — pour une sémantique des idéologies, in Actes des Journées Scientifiques en linguistique 2002-03, J. Pauchard et F. Canon-Roger (éds.), *CIRLLLEP*, Presses Universitaires de Reims, n° 22.

Rastier, F., à paraître, Sémiotique du discours raciste — application au filtrage automatique de sites, *Mots*.

Rastier, F. et coll., 2002, *Semantics for Descriptions,* Chicago, Chicago University Press.

Saussure, F. de, 2002, *Écrits de linguistique générale*, éd. Rudolf Engler et Simon Bouquet, Paris, Gallimard.

Swales, J. M., 1990, *Genre Analysis. — English in Academic and Research Settings*. Cambridge, Cambridge University Press.

Tanguy L., 1997, *Traitement automatique de la langue naturelle et interprétation : contribution à l'élaboration informatique d'un modèle de la sémantique interprétative*, thèse, Université de Rennes I.

Thlivitis T., 1998, *Sémantique Interprétative Intertextuelle : assistance informatique anthropocentrée à la compréhension de textes*, thèse de doctorat, Informatique, Université de Rennes I.

Vous pouvez adresser vos commentaires et suggestions à : Lpe2@ext.jussieu.fr

Référence bibliographique : RASTIER, François. Enjeux épistémologiques de la linguistique de corpus. *Texto !* [en ligne], juin 2004.Rubrique Dits et inédits. Disponible sur : <http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html>. (Consultée le ...).