# Que cachent les « données textuelles » ?

François Rastier[1]

[1]CNRS-ERTIM, INALCO, 2 rue de Lille, 75 007 Paris

## Abstract

This study seeks to realign the triangular division which exists between discourse, text and documents. To further this end, it relies on corpus semantics and digital philology. In contrast to the Semantic Web programme, the goal here is to move beyond "data" back to documents, and to make use of their irreplaceable complexity in order to find specific information.

## Résumé

Cette étude invite à un remembrement de la tripartition de fait entre discours, texte, et document. Elle s'appuie pour cela sur la sémantique de corpus et la philologie numérique. Il s'agit en effet, à l'inverse du programme du Web sémantique, de revenir des « données » aux documents et d'exploiter pour la recherche d'informations leur irremplaçable complexité.

**Mots-clés :** texte, document, passage, complexité, textométrie, sémantique de corpus.

## 1. Introduction

Rien ne nous est donné. La notion de *donnée* prend toutefois un relief particulier si l'on s'avise qu'en promouvant le *Web sémantique*, le W3C, instance qui préside aux destinées du Web mondial, entend remplacer le « Web des documents » par le « Web des données » (cf. Tim Berners-Lee, 2007). En utilisant des ontologies, il s'agit de s'affranchir de la complexité des documents et de leur diversité linguistique et sémiotique. En accord avec l'objectivisme de la philosophie du langage issue du positivisme logique, la donnée est alors conçue comme une simple chaîne de caractères (ex. la donnée *pêche*, qui peut être reliée soit à *poisson*, soit à *fruit* ; Berners-Lee, *loc. cit*.). Nous proposerons ici un modèle moins sommaire de la donnée, qui tienne compte de la dualité sémiotique irréductible entre expression et contenu, ou plus généralement entre *phore* et *valeur*. Cela s'étend à toute chaîne de caractères, du signe de ponctuation au chapitre, sans égard pour le modèle apocryphe du signe prêté à Saussure par les rédacteurs du *Cours de linguistique générale* et contredit par les écrits autographes.

La dualité phore/valeur, qui constitue le corps sémiotique de la donnée, se trouve sous la rection d'une dualité de rang supérieur entre le *point de vue* et le *garant*. Le point de vue n'est pas un simple point d'observation : il est déterminé par une pratique et un agent individuel ou collectif ; dans un traitement de données, il dépend donc de l'application. Le garant est l'instance de validation qui fonde l'évaluation de la donnée : cette instance est une norme sociale qui peut être juridique, scientifique, religieuse ou simplement endoxale. En linguistique de corpus, le garant est l'autorité qui a présidé à la constitution du corpus ; certaines métadonnées documentaires, comme l'auteur ou l'éditeur, relèvent de cette instance.

Le point de vue est « subjectif » dans la mesure où il est occasionnel ; le garant, « objectif » dans la mesure où il est constitutionnel ou du moins constituant. La dualité du point de vue et

du garant définit deux régimes de pertinence, saillance pour le point de vue et prégnance pour le garant. Puisque les données sont bien ce qu'on se donne, elles sont ainsi les résultats initiaux d'un processus d'élaboration – et leur traitement produit des résultats ultérieurs, dans un cycle susceptible de récursivité.

Dans les termes de la sémiotique des zones anthropiques (cf. l'auteur, 1996, 2001a, 2002), le corps (phore+valeur) de la donnée, en tant qu'elle est objectivée, relève de la zone proximale de l'environnement ; le point de vue, de la zone identitaire ; enfin, le garant, de la zone distale où se situent les instances de normativité. L'axe sur lequel se répartissent ces zones est celui de la *médiation symbolique*, alors que l'axe subordonné qui relie le phore et la valeur relève de la *médiation sémiotique* (cf. l'auteur, 2001a). Soit en bref :

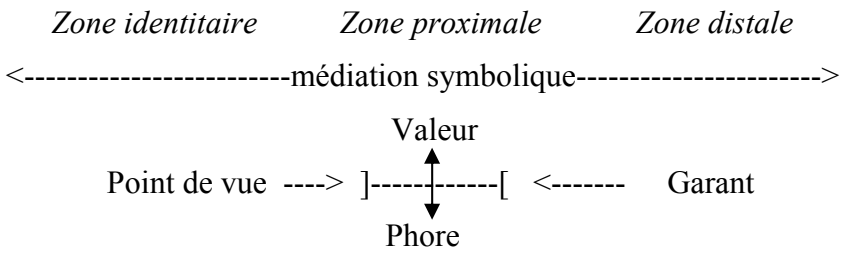


*Figure 1 : Les quatre instances et les trois zones de la donnée*

En ne percevant pas le caractère instituant de la valeur, du point de vue et du garant, en réduisant la donnée à la seule instance du phore, le positivisme ordinaire élude toute dimension critique et épistémologique. Recueil de données ainsi appauvries, un « corpus » sans point de vue ni garant n'est pas véritablement un objet scientifique mais un amas numérique inexploitable en tant que tel ; ainsi des pseudo-corpus recueillis par aspiration aléatoire de sites.

## 2. Modèles du texte

Le modèle général de la donnée relève de la sémiotique. S'agissant des données textuelles, la conception du texte mobilisée retentit évidemment sur leur statut, leur empan, leurs régimes de pertinence et de légitimité.

### *2.1. Dépasser l'opposition entre texte et discours*

Fondatrice pour l'école française d'Analyse du discours, l'opposition entre *discours* et *texte* transpose explicitement l'opposition entre l'*énoncé* (relevant de la pragmatique) et la *phrase* (relevant de la syntaxe). Elle privilégie cependant le discours sur le texte : le texte serait un discours privé de son contexte, selon l'équation texte+contexte=discours[1]. Aucune méthodologie déterminable ne permet cependant de soustraire un texte à son contexte. Tributaire de la tripartition syntaxe/sémantique/pragmatique, l'Analyse du discours peine ainsi à articuler la linguistique externe dont relève la pragmatique[2] à la linguistique interne dont relève la syntaxe. Elle estime que l'opposition *énonciation/énoncé*, qui prolonge l'opposition aristotélicienne puis scolastique *enargeia/ergon*, pourrait être résolue par le

[1] Dernière reformulation : Discours = texte + conditions de production et de réception (Adam, 2005).

[2] Il s'agit certes d'une pragmatique élargie, puisqu'elle entend déterminer les « conditions de production » socio-politiques des textes. Avec l'effacement des références marxisantes, les domaines de la pragmatique et de l'Analyse du discours sont devenus indiscernables, comme en témoigne le récent *Dictionnaire d'Analyse du discours*.

repérage dans l'énoncé de « marques de l'énonciation », ce qui fut jadis une des premières tâches de la lexicométrie appliquée notamment à l'analyse du discours politique. Si de telles marques existaient, on les trouverait partout (sauf à se limiter à une pragmatique des indexicaux). Au demeurant, le privilège singulier qui leur est accordé fait peu de cas de la complexité et de la systématicité des textes.

Si l'on s'attache à décrire les instances de normativité textuelles, en particulier les genres, les « conditions de production » trouvent un autre statut d'intelligibilité, car tout texte oral ou écrit appartient à la strate sémiotique d'une pratique sociale : prescrivant les régimes génétique, mimétique et herméneutique du texte, le genre relie le texte à un discours (politique, juridique, religieux, etc.). La typologie des genres et des discours, telles que l'appelle et la permet aujourd'hui la linguistique de corpus, permet ainsi de dépasser l'opposition figée entre texte et « discours » (au singulier). Si la langue est une formation sociale, le texte appartient d'abord à une « société de textes », selon la formule de Ioannis Kanellos : le corpus, notamment le corpus des textes de même genre, peut être alors reconnu comme la médiation nécessaire entre le texte et la caractérisation de la pratique sociale dans laquelle il prend place.

### *2.2. Modèles logico-grammaticaux : arbres et chaînes*

Si le texte n'est pas comparable à une phrase, il ne l'est pas non plus une suite de phrases. Les grammaires de textes ont transposé l'imaginaire grammatical de diverses manières non exclusives entre elles. Elles ont étendu la syntaxe de la phrase complexe à la macrosyntaxe, dans l'intention d'édifier une syntaxe textuelle : toutefois, le palier de complexité que concrétise le texte n'a pas véritablement permis cette extension, car aucune règle consistante n'a pu être formulée (comme l'a établi à la fin des années 1970 le débat dans *Cognitive Science*).

Concevoir le texte comme une grande phrase conduisit à proposer un modèle arborescent : par exemple, celui de Van Dijk et Kintsch, hégémonique en psychologie cognitive jusqu'au début des années 1990, étageait des propositions d'une généralité croissante jusqu'à le résumer à une « macroproposition ». Les réseaux sémantiques puis les ontologies ont radicalisé ce type de réduction, pour ne conserver que la structure arborescente, étiquetée par des concepts extraits du texte ou des prédicats élémentaires (en RDF par exemple). En outre, avec le développement de la *Text Encoding Initiative* et la généralisation d'XML, la structure de données XML est utilisée à présent comme un modèle arborescent du texte : un standard de codage de données se trouve ainsi promu au rang de théorie (cf. Loiseau, 2007)[3].

Complémentaire de l'arborescence, l'image de la chaîne d'unité discrètes reste prégnante. Codant successivement toutes les phrases du texte, le format propositionnel engageait jadis, au temps des grammaires générales, à considérer le texte sur le modèle d'un syllogisme étendu ; il conduit aujourd'hui à définir la syntagmatique textuelle à partir d'enchaînement logiques élémentaires (comme les « blocs sémantiques » selon Ducrot et Carel ; ou les Représentations discursives selon Kamp puis Asher) ou plus simplement de chaînes de coréférence. En élargissant cette perspective syntagmatique à des unités de plus grande taille

[3] Par exemple, dans leur quatrième version, les Recommandations de la TEI (Sperberg-McQueen & Burnard, http://www.tei-c.org/P4X/SG.html), représentent une anthologie poétique par un arbre où les titres sont au même rang que les strophes, lesquelles subsument les vers. Toutefois la relation entre un poème et son titre n'a rien de commun avec celle qui relie les strophes entre elles ; celle qui ordonne les poèmes d'une anthologie diffère profondément de celle qui relie les vers entre eux, etc.

que des propositions, on a pu considérer le texte une succession de « séquences » (Adam, 1992), concrétisant des fonctions du langage (comme *décrire* ou *argumenter*) et correspondant pour la plupart à des figures non tropes de l'ancienne rhétorique. Malgré la faveur de l'appareil pédagogique, ce modèle est en voie d'être abandonné, car le codage de séquences *a priori* reste problématique.

Les images de l'arbre et la chaîne se complètent : la chaîne représente la succession d'entités empiriques discrètes et isonomes, à l'image des mots ; l'arbre, la hiérarchie vers l'abstraction qui inclut ces unités dans des groupements. Multiplicité des objets déjà donnés et clos sur leurs substances, passage vers l'unité totalisante par l'abstraction progressive, ces deux fondements de la tradition ontologique occidentale continuent d'informer la tradition logico-grammaticale étendue ainsi aux modèles du texte.

### *2.3. Modèles non-hiérarchiques*

Plutôt que la grammaire ou de la logique, les modèles alternatifs du texte sont issus des mathématiques, des sciences de la vie, voire de l'herméneutique et de la philologie.

#### *2.3.1. Sacs de mots et merveilleux nuages*

Définir le corpus comme un « vaste ensemble de mots » (comme le faisait naguère Sinclair) fut sans doute une projection imaginaire d'une méthodologie qui recourt massivement aux *word-crunchers*. Elle reflétait sinon un état de l'art, du moins une image spontanée des textes pour une linguistique de corpus qui privilégie le lexique, conformément à une épistémologie empiriste de tradition nominaliste qui privilégie les objets atomiques. Les méthodes d'analyse des données issues des travaux de Salton ont rencontré naturellement cette préconception : les textes sont représentés par des nuages de vecteurs (les « nuées dynamiques »).

Par leur puissance et leur adaptabilité, les méthodes mathématiques contrastent heureusement avec la logique appauvrie des réseaux sémantiques classiques et des ontologies, souvent réduite à un calcul des prédicats du premier ordre sans opérateurs modaux, comme RDF par exemple.

#### *2.3.2. Réseaux*

Depuis le milieu des années 1960, le modèle logique de l'arborescence se voit concurrencé par le modèle du réseau, d'inspiration biologisante : notamment, le modèle de mémoire sémantique de Quillian, associé aux neurones formels de McCulloch et Pitts, inspirera dans les années 1980 les premiers modèles connexionnistes de la phrase. Eco, dès 1974, opposera par ailleurs, pour la description des textes, le modèle arborescent dit modèle *K* (pour Katz et Fodor) et le modèle réticulaire dit modèle *Q* (pour Quillian). Le caractère non hiérarchique du réseau sera prisé tant par les théories critiques (cf. le *rhizome* chez Deleuze et Guattari) que par les théoriciens de l'hypertextualité comme Ted Nelson.

Il convient de préciser toutefois si les réseaux sont fixes (auquel cas ils se réduisent à des versions non hiérarchiques des réseaux sémantiques) ; s'ils sont dynamiques et se reconfigurent par apprentissage ; enfin, corrélativement, si leurs liens sont typés et/ou s'ils supportent des propagations d'activations.

### *2.4. Un modèle textuel issu de la sémantique interprétative*

On confond trop souvent les modèles théoriques et les formats de représentation, voire les modes d'implémentation. En raison de la complexité des textes, il reste indispensable

d'édifier des modèles spécifiques du texte et du corpus issus de la linguistique et de la philologie et qui soient fondés sur une théorie des pratiques sémiotiques.

*2.4.1. Hétérarchie des composantes textuelles*

La sémantique interprétative présente le contenu du texte comme une hétérarchie de composantes sémantiques (thématique, dialectique, dialogique, tactique). Le plan de l'expression est décrit par d'autres composantes (médiatique, etc.). Le genre se définit par un type d'interaction entre composantes au sein des deux plans du contenu et de l'expression, ainsi qu'entre ces deux plans : elles norment ainsi la *sémiosis textuelle*. Alors que la sémiosis au palier du mot reste trivialement problématique (en raison des faux problèmes induits par la polysémie, la synonymie, etc.), la sémiosis des paliers supérieurs comme le paragraphe dépend de la sémiosis textuelle, telle qu'elle est normée par le genre ; par exemple, *amour* n'a pas le même sens en poésie et dans le roman et ne partage aucun de ses cooccurrents (cf. l'auteur 2004).

*2.4.2. Formes sémiotiques*[4]

Dans l'hypothèse de la perception sémantique, l'opposition entre fond et forme qui détermine le traitement de l'expression vaut aussi pour le plan du contenu. Les fonds sémantiques sont des isotopies et faisceaux d'isotopies. Les formes sémantiques, comme les thèmes, les acteurs, les foyers énonciatifs, s'apparient avec des formes expressives (périodes) pour constituer des *formes sémiotiques*.

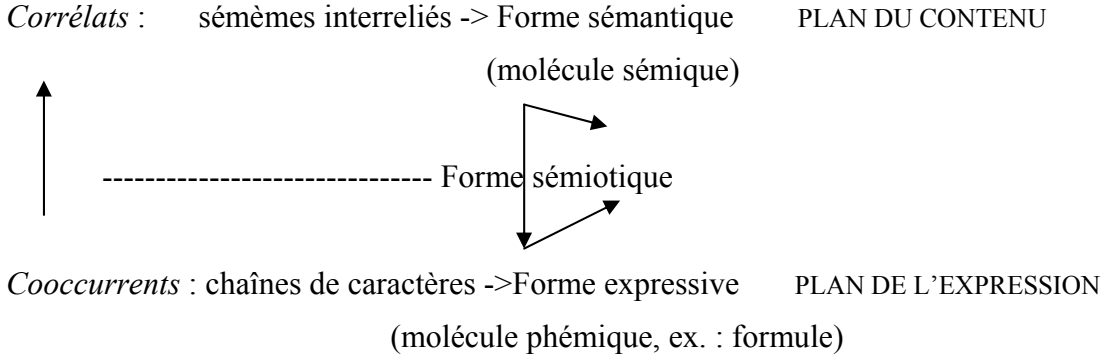


*Figure 2 : Constitution d'une forme sémiotique*

Cet appariement ne va pas de soi et suppose une interprétation : quand on utilise des méthodes lexicométriques, les groupes de *cooccurrents*, qui, en tant que chaînes de caractères, relèvent du plan de l'expression, doivent être qualifiés comme des *corrélats* sémantiques. L'interprétation locale est ainsi constitutive des signes, et dépend du régime herméneutique global propre au genre et au texte.

*2.4.3. Métamorphismes*

Dans cette conception morphosémantique du texte, les transformations se spécifient en *métamorphismes* (changements de forme), *métatopies* (changements de fond) et *transpositions* (changements des rapports entre forme et fond : une forme peut se diffuser dans un fond). Les transformations des formes sémantiques et des formes expressives se

[4] Dans les paragraphes qui suivent nous reprenons des éléments de l'auteur (2001b, 2007).

manifestent par des changements de contexte, comme par des modifications corrélatives du rapport entre contenu et expression :

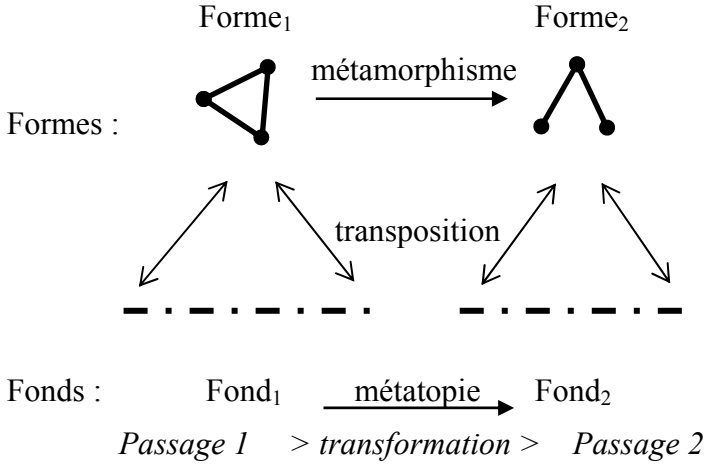


*Figure 3 : Métamorphismes, métatopies et transpositions*

Le principe des transformations est analogue pour le contenu et pour l'expression[5]. Ces deux plans étant solidaires en raison du principe même de la sémiosis, qui s'étend à des relations de contextualité entre plans, toute transformation sur un plan s'accompagne d'une transformation sur l'autre.

Rompant avec l'ontologie prégnante dans la tradition logico-grammaticale, la conception du texte comme série de transformations relève d'une *praxéologie* : non seulement tout texte est inscrit dans une pratique, mais cette pratique s'inscrit en lui, car il est produit par une activité de réécriture constante qui assure sa cohésion et définit la textualisation comme un cours d'action que les parcours interprétatifs peuvent avoir ultimement l'ambition de restituer[6]. Comme le texte est lui-même produit par un processus de reformulation intralinguistique, la traduction peut restituer ainsi la réécriture interne qui établit la textualité. En effet, malgré leur diversité, les métamorphismes internes qui définissent la textualité ne connaissent pas de différence de principe avec les métamorphismes externes qui définissent l'intertextualité, qu'ils soient unilingues ou plurilingues. Ainsi, les parcours au sein du texte et entre textes d'un même corpus peuvent-ils être problématisés et décrits selon les mêmes principes.

### *2.4.4. Passages et problématisation des unités*

Pour rendre opératoire le concept de forme sémiotique et permettre l'extraction assistée de telles formes, il convient de revenir sur la notion d'unité textuelle. Dans la perspective néo-saussurienne qui est la nôtre, les grandeurs ne sont pas des unités empiriquement constatables ou des données d'évidence. Elles ne sont pas des unités simplement isolables car discrètes ni déterminables à un seul plan d'analyse (du contenu ou de l'expression). Les signes et plus généralement les grandeurs sémiotiques sont construits dans l'interprétation qui leur assigne leur valeur. Si l'ontologie logico-grammaticale attribue aux grandeurs textuelles la discrétion et la présence, l'identité à soi et l'isonomie, sans doute à l'image naïve des objets physiques,

[5] Pour l'analyse détaillée d'un exemple de transformations entre passages de deux romans de Balzac, cf. l'auteur (2007b).

[6] Notamment dans les textes qui peuvent prétendre au rang d'œuvres, chaque page révise à sa manière la précédente ; cette reformulation créatrice affermit la textualité et donne à la langue sa valeur critique de culture.

la conception rhétorique / herméneutique admet en revanche que les grandeurs qu'elle construit soient continues, parfois implicites, varient dans le temps et selon leurs occurrences et leurs contextes, connaissent entre elles des inégalités qualitatives et ne relèvent pas uniformément des mêmes règles.

Le texte ne présente pas de signifiant identifiable par des procédures de segmentation systématique, sinon par les démarcations fortes, comme les pauses longues ou les changements de chapitre. C'est une raison fondamentale pour échapper au modèle du signe : les grandeurs sémantiques textuelles n'ont pas de signifiants uniformément isolables comme des parties du discours ; elles sont constituées par des connexions de signifiés et d'expressions des paliers inférieurs de la période, du syntagme, de la sémie. L'articulation de l'expression qui détermine l'identité de la grandeur, et du contenu qui détermine sa valeur s'opère au sein du *passage*. Dans la perspective interprétative, il convient donc de définir cette grandeur locale, qu'elle corresponde indifféremment à un signe, à une phrase, ou par exemple à un paragraphe. Au plan du signifiant, le passage est un *extrait*, entre deux blancs s'il s'agit d'une chaîne de caractères ; entre deux pauses ou ponctuations, s'il s'agit par exemple d'une période. Au plan du signifié, le passage est un *fragment* qui pointe vers ses contextes gauche et droit, proches et lointains. Cela vaut pour le sémème comme pour le contenu du syntagme ou de la période. On peut ainsi substituer à la monade sémiotique apocryphe du *Cours de linguistique générale* cette figure du *passage* :

$$\leftarrow \supset \textit{fragment du contenu} \subset \rightarrow$$

- - - - - - - - - - -

$$\leftarrow \supset \textit{extrait de l'expression} \subset \rightarrow$$

*Figure 4 : Le passage*

Le passage, comme le figurent les signes convexes symbolisant son ouverture et les flèches droite et gauche, renvoie aux étendues contiguës ou plus lointaines. L'extrait peut renvoyer aux étendues connexes, par exemple par des règles d'isophonie ou de concordance de morphèmes : ce sont des *cooccurrents* expressifs. Le fragment se relie à d'autres par des phénomènes d'isotopie : ils ont le statut de *corrélats* sémantiques. Pour ce qui concerne leur connectivité externe, on distinguera l'*incidence* de l'extrait et la *portée* du fragment. Un extrait peut être conventionnellement isolé, car les structures de l'expression relèvent pour l'essentiel de la mésolinguistique ; en revanche, un fragment ne peut l'être sans perte, car les structures du contenu sont macrolinguistiques.

La sélection d'un passage et *a fortiori* l'isolation d'un « signe » exigent deux opérations : faire l'hypothèse qu'à un extrait minimal correspond un fragment, de façon à pouvoir les isoler ; puis, en les décontextualisant, leur assigner un rapport terme à terme entre signification et expression qui littéralise la première et fixe la seconde. Les méthodes statistiques[7] permettent de proposer, pour chaque passage, des *cooccurrents* expressifs de l'extrait qui restent à qualifier comme des *corrélats* sémantiques du fragment.

[7] Quand il s'appuie sur des corpus de textes appartenant au même genre et au même discours que le texte analysé, le test de l'écart réduit permet de repérer des groupements de cooccurrents qui sont de bons candidats pour la constitution de passages (cf. la fonction Thème du logiciel Hyperbase, obligeamment ménagée par Étienne Brunet). Les travaux de Viprey sur les *isotropies* peuvent également être utilisés pour la détection de passages. Enfin, la thèse de Mauceri (2007) ouvre des perspectives fort intéressantes.

*Plan du contenu*

⊃ frag. corrélat $_1$⊂ ⊃ *fragment* ⊂ ⊃ frag. corrélat $_n$⊂

---

⊃ ext. cooccurrent $_1$⊂ ⊃ *extrait* ⊂ ⊃ ext. cooccurrent $_n$⊂

*Plan de l'expression*

*Figure 5 : Le passage et ses contextes*

Les corrélats d'un fragment sont d'autres fragments, les cooccurrents d'un extrait, d'autres extraits ; les relations entre passages intéressent ainsi tant la textualité que l'intertextualité.

Notons que le passage n'a pas de bornes fixes et son empan dépend évidemment du point de vue qui a déterminé sa sélection. Sa définition s'écarte donc de l'objectivisme traditionnel de la tradition logico-grammaticale. Redéfinir le signe comme un passage conduit à s'éloigner de la logique des « idées » et des représentations, pour en élaborer une conception purement relationnelle et donc contextuelle. La relative clôture organisationnelle du passage se traduit par le fait que les relations au sein du passage sont plus denses et sémiotiquement plus fortes que les relations entre passages. Le rapport entre global et local va du texte au passage : le passage est une *zone de localité*, définie par une sémiosis propre (mode d'appariement entre contenu et expression) et, sur chacun de ses plans (fragment et extrait) par des relations contextuelles internes fortes.

Ainsi les « données textuelles » sont-elles qualifiées comme des passages, fussent-ils de petite taille, comme les lexies.

### *2.5. Pour un modèle intertextuel*

Dans la perspective non ontologique qui est la nôtre, le sens est fait de différences (non de références) ; et d'autre part toute catégorisation véritable est contrastive. Si donc le sens d'un texte s'étudie en contrastant ses différentes parties et passages, il s'étudie aussi en le contrastant avec d'autres textes. Il ne s'agit pas là de deux degrés de complexité, la macrosémantique traitant du texte et une « mégasémantique » des corpus. En effet, dès lors que l'on entend restituer des parcours interprétatifs appuyés sur les métamorphismes, pour passer d'un mot à un autre, d'un passage à un autre, on a souvent, voire toujours besoin de passer par un ou plusieurs autres textes. Dans cette mesure, l'intertexte se trouve inclus et convoqué par l'interprétation du texte ; c'est pourquoi les associations sémantiques imputables à la doxa se voient souvent vite confirmées par l'interrogation de concordanciers (cf. Louw, 2007).

Pour la perspective grammaticale étendue adoptée par la linguistique textuelle traditionnelle, le texte est l'unité linguistique maximale. Cette position doit être nuancée, car pour la problématique herméneutique, c'est l'unité *minimale* (bien que non élémentaire). Un texte ne peut se lire que dans un corpus, qu'il soit implicite, comme en général dans les études littéraires, ou explicite, comme en linguistique de corpus. L'intertexte n'est donc pas cette nébuleuse où se meuvent avec agilité les herméneutiques déconstructionnistes : il est structuré *a minima* par (i) des degrés de proximité (entre aires d'auteur, de genre et de discours), (ii) des degrés de connectivité, des cycles (deux textes distants renvoient au même texte), (iii) des relations contextuelles *in absentia*. Si la topographie textuelle se développe heureusement, la topographie intertextuelle reste pour l'essentiel à concevoir (cf. l'auteur, 2007a, § Parcours interprétatifs dans l'intertexte).

Dans la mesure où un genre est une lignée historique, on peut retracer des stemmas sémantiques, sachant toutefois que les textes de même génération se réécrivent mutuellement, et que là encore le modèle arborescent le cède au modèle réticulaire – pour peu que l'on puisse représenter les dynamiques du réseau intertextuel.

Par ailleurs, en général le corpus n'est pas donné et varie méthodologiquement (corpus d'étude, de travail, de référence, archive, etc. ; cf. l'auteur 2005). Chaque corpus peut être considéré comme une réponse anticipée à une classe de questions, mais son exploration en fait naître d'autres : un corpus bien constitué peut répondre aux questions imprévues qui se posent lors de son exploration et le mode de constitution du corpus, en tant qu'il témoigne d'une précompréhension, peut garantir pour une part l'interprétation ultérieure.

### *2.6. Rétrospection*

Comme ailleurs dans les sciences de la culture, les modèles du texte que nous venons de présenter ne sont pas des modèles au sens de la théorie des modèles, mais des schèmes substrats de l'élaboration conceptuelle. Sacs de mots, merveilleux nuages, arbres ou réseaux, chaque type de théorie projette sur le texte ses prédilections, les modèles dynamiques de la lexicométrie, liés aux mathématiques, contrastant avec les modèles logico-grammaticaux de la linguistique textuelle. Dans la mesure où elle pose le problème crucial des inégalités qualitatives, où elle ne dépend pas nécessairement des catégories grammaticales (des parties du discours à la phrase) ni de l'ontologie implicite qui s'y attache, la lexicométrie reste sans doute plus facile à concilier avec la problématique interprétative que la linguistique textuelle.

Peu importe au demeurant qu'une application, fût-elle didactique, privilégie tel ou tel modèle : l'essentiel demeure la juste appréciation des transformations textuelles. La conception praxéologique de la textualité insiste sur le régime génétique des réécritures de passages et le régime herméneutique des parcours interprétatifs entre passages. Elle permet ainsi, tout en respectant la forte connectivité du texte, de s'étendre à l'intertexte, qu'il s'agisse du corpus d'élaboration, du corpus d'interprétation, ou des domaines complexes des lignées textuelles, de la traduction et du commentaire.

## 3. Le texte comme document

### *3.1. La reconquête de l'expression et le mystère des métadonnées*

Consommée de fait depuis un demi-siècle, la séparation entre linguistique et philologie a beau être récente, ses raisons mêmes furent oubliées avec l'oubli des textes par les linguistiques universelles. Cependant, depuis sa formation disciplinaire à Alexandrie, la grammaire avait toujours été considérée comme une discipline auxiliaire pour la lecture et l'analyse critique des textes. La sémantique (*Semasiologie*), lors de sa création par Reisig dans les années 1820, était une sorte de lexicologie des textes classiques. Alors qu'il était ordinaire de combiner l'analyse de textes et la linguistique historique et comparée (Steinthal, Bréal, Saussure en sont des exemples bien connus), à partir des années 1950, l'adoption unilatérale de perspectives synchroniques, le privilège exclusif donné à la modélisation de la morphosyntaxe, l'image prégnante des langages formels, et complémentairement les études sur l'oral privilégiées par une pragmatique du *hic et nunc*, ont accompagné voire causé la quasi-disparition de la philologie en linguistique générale.

Les grammaires de texte ont confirmé ou consommé l'abstraction du concept de texte. Il fut généralement réduit à son plan sémantique : la *Sémantique structurale* de Greimas (1966)

proposait ainsi sa formalisation en une série de propositions dans un format inspiré explicitement de la logique de Reichenbach. Il reste la source principale du modèle de Van Dijk (dont Greimas dirigea les premières recherches), puis des divers modèles propositionnels du cognitivisme orthodoxe (la Forme Logique chomskyenne des années 1980, par exemple). On est revenu ainsi au programme des grammaires générales, justement dites *philosophiques*, qui à l'âge classique, avant donc la formation de la linguistique, représentaient les phrases comme des propositions logiques et le discours comme un raisonnement (sur le modèle de la logique des classes). L'appauvrissement qui accompagnait cette réduction « sémantico-logique » conduisit à faire du texte une série de chaînes de caractères considérées comme des données. Si en revanche, à la suite de Saussure, on rapatrie le signifié dans les langues et dans les textes, on doit donner corrélativement toute sa place au plan de l'expression, puisque la corrélation entre expression et contenu détermine la sémiosis textuelle.

Les indications philologiques élémentaires, quand elles sont retenues, sont aujourd'hui classées comme des « métadonnées ». Cette conception prévaut aujourd'hui avec le Web sémantique. La plus grande confusion règne dans ce domaine, puisqu'on classe dans les métadonnées toutes sortes de données incompatibles avec la théorie appauvrie du texte qui prévaut généralement : on juxtapose des indications simplement bibliographiques, comme l'auteur, l'éditeur, l'ISBN, le lieu d'édition ; des indications documentaires, comme le résumé ou les mots-clé ; des caractérisation textuelles globales, comme le genre. Les théories linguistiques du péritexte, qui limitent le texte à l'intratexte, en séparent les titres, voire les notes, etc. n'ont fait qu'ajouter à la confusion. En règle générale, les données relèvent de la linguistique interne, les métadonnées de la linguistique externe et, faute de réfléchir leur dualité, on ne peut théoriser le rapport entre données et métadonnées. Les problèmes négligés reviennent alors, réifiés, sous la forme de métadonnées. Par exemple, dans le domaine du multimédia, les textes eux-mêmes deviennent les métadonnées des images.

Sans trop croire à l'efficacité d'un moratoire sur les métadonnées, retenons que les métadonnées sont des critères globaux et les données des grandeurs locales qui en dépendent : au lieu de les séparer *a priori*, c'est à une théorie élaborée de la textualité qu'il revient d'établir systématiquement les corrélations entre métadonnées et données, pour restituer la complexité des textes.

La notion de métadonnée doit ainsi être critiquée et refondue. Le succès de Google s'explique ainsi par l'introduction d'un nouveau type de métadonnées (les liens qui pointent vers le document) et par une perspective praxéologique implicite qui représente le document en fonction d'un *point de vue* (de qui met le lien) et d'un *garant* (l'évaluation positive qu'il apporte).

### *3.2. Stratification ouverte et extension du domaine des données*

Formations historiques, les langues sont des artéfacts : le nombre et la nature de leurs niveaux de sémiotisation ne sont pas fixés *a priori*, comme on le sait depuis l'invention de l'écriture[8]. Les codes des formats numériques poursuivent cette évolution, même s'ils ne sont pas réservés exclusivement aux textes. Par exemple, le HTML et le XML qui codent un texte ou des éléments de ce texte peuvent être considérés comme des niveaux sémiotiques supplémentaires. En effet, le texte est en quelque sorte une idéalisation linguistique, et son

[8] Cela vaut d'ailleurs sur les deux plans, puisque les « recodages » sémantiques sont attestés par différentes herméneutiques, en général ésotériques.

support documentaire introduit inévitablement d'autres sémiotiques : le rouge n'est pas réservé aux textes, mais, signe d'excellence dans l'Antiquité, il a donné lieu au signalement des rubriques, si bien nommées, dans les manuscrits, puis aux titrages en rouge dans l'imprimerie renaissance, etc.

Il faut cependant dans la notion de document distinguer deux choses : l'expression du texte et la configuration matérielle du support. La structure typographique de la mise en forme relève d'un niveau de l'expression textuelle ; en revanche, la mise en page relève de la norme du document – même si elle n'est pas sans effet sur l'appréhension du texte. Ainsi les paginations comme les titres courants font-ils partie du document, non du texte. À la philologie numérique répond ainsi une *diplomatique* numérique, qui ne traite point du texte, mais seulement de caractères spécifiques au document qui le véhicule.

Enfin, la stratification même du langage doit encore beaucoup à des simplifications théoriques : ainsi la séparation entre contenu et expression a-t-elle été creusée, sinon imposée, par le dualisme métaphysique qui oppose matière et pensée ; mais aussi à des simplifications méthodologiques : on a maintenu séparés les niveaux linguistiques de manière à réduire la complexité et pouvoir formuler des règles dont l'application ne soit pas conditionnée par l'incidence de phénomènes situés à d'autres niveaux.

De fait, la linguistique de corpus a permis de produire de nouveaux observables qui associent des éléments de niveaux d'analyse ordinairement séparés (ainsi de la corrélation entre des noms de sentiments ou des temps verbaux avec certaines ponctuations, cf. l'auteur, 2005) : elle témoigne ainsi de la solidarité sémiotique entre plans du langage. Plus impressionnante encore, la solidarité entre paliers de complexité permet de caractériser des textes en corrélant des indices d'expression locaux (comme la ponctuation ou la longueur moyenne des mots) à des « métadonnées » sémantiques globales comme le genre et le discours : les résultats présentés dans Malrieu et Rastier (2001) établissent ainsi la concordance complète entre la classification « manuelle » en genres et discours sur un corpus de 2 600 textes et la classification automatique à partir de moyennes établies pour chaque texte sur 251 variables locales d'expression.

Des critères ordinairement réputés non textuels, comme le code de la police de caractères peuvent se révéler fort discriminants : par exemple, dans une application de détection de sites racistes, le code des polices gothiques obligeamment téléchargeables sur les sites néo-nazis peuvent suffire à caractériser un texte à la volée, sans grand risque d'erreur. Bref, à partir de la sémantique des textes, la « reconquête » de l'expression a d'autant plus de conséquences que les systèmes informatiques ne traitent aisément que les indices de cette strate linguistique. Cela engage une reconception sémiotique du texte, plus précise que celle qui a été élaborée naguère dans les études littéraires. Pour les applications, cela permet de développer des caractérisations multicritériales, les coalitions d'indices permettant de former des conjectures éprouvées, dans des systèmes multi-agents par exemple. Ces coalitions sont extraites par « des patrons complexes d'expressions régulières multiniveaux combinant différents items de natures formelles différentes (chaînes de caractères, POS, annotations, positions, etc.) » (Valette et Slodzian, à paraître).

### *3.3. Enjeux actuels*

On sait que le programme du Web Sémantique consiste à passer du « Web des documents » au « Web des données » sur le mode de ce que l'on nommait naguère la représentation des connaissances, en extrayant des textes des données, et en les organisant en ontologies. Les

enjeux politiques et économiques ne peuvent cacher les conséquences épistémologiques de ce programme. En effet, l'adoption de standards « de bas niveau » comme HTML, ou Unicode voire XML n'entraîne aucunement que l'on doive ériger en standard des langages de représentation comme RDF ou OWL, sauf à céder benoîtement à la tentative de coup de force du W3C en faveur du « Web sémantique ».

Un texte n'est pas un réservoir de connaissances qui pourraient être extraites par indexation et condensées en données résumant son contenu informationnel ; l'indexation donc n'a qu'une relative valeur de recherche et de classement. Prenons un exemple : à l'heure actuelle, dans les services de renseignement militaire d'un grand pays européen, des personnels extraient de documents Word des mots et expressions qu'ils transfèrent dans des feuilles Excel où ils sont classés en « ontologies ». Ces feuilles sont ensuite transmises à des analystes qui en font la synthèse sous la forme de Powerpoints présentés à l'état-major[9]. L'éloquence militaire prise certes le laconisme, mais toute modification systématique d'un texte en change le genre et donc l'interprétation. La sélection de ces passages minimaux que sont les mots et expressions reste de fait incontrôlable, puisque le recouvrement de deux indexations du même texte par la même personne s'établit en moyenne à 40%. La délinéarisation, la « compression » augmentent l'équivoque et créent l'ambiguïté.

Au « Web sémantique », il faudra inévitablement substituer une sémantique du Web (cf. l'auteur, à paraître), car les besoins sociaux pour la recherche d'information, l'amélioration des moteurs de recherche, le *data-mining*, ne pourront être satisfaits que par une linguistique et une sémiotique de corpus permettant l'analyse des données textuelles.

### *3.4. Perspectives pratiques et théoriques*

La communauté française de lexicométrie s'est formée dans les années 1960, exploitant notamment les recherches statistiques de Benzécri et de son école ; depuis elle a étendu et élargi son champ de réflexion et perfectionné ses outils, dans une remarquable continuité incarnée par des figures indéfectiblement tutélaires comme Charles Muller et Etienne Brunet. Ce n'est qu'au milieu des années 1990 qu'elle s'est rebaptisée *textométrie*, et, chose encourageante, plusieurs auteurs se disputent amicalement la paternité de ce mot.

Par ailleurs, les textes se sont considérablement enrichis depuis les années 1960 ; au format *.txt* se sont ajoutés d'autres formats agrémentés par diverses balises, de structuration notamment. D'autre part, les conceptions linguistiques du texte se sont approfondies en réfléchissant les traditions herméneutiques et philologiques, transposées aux documents numériques. Enfin, les besoins sociaux se sont notoirement accrus.

Les théories antérieures à la linguistique de corpus peinent à s'adapter aux nouveaux moyens descriptifs élaborés par la linguistique de corpus. Par exemple, les théories de Halliday et Hasan (1976) ne s'appuyaient pas sur des observations méthodiques et instrumentées ; aussi leurs thèses sur la texture, la cohésion lexicale et la méronymie ont-elles été infirmées en linguistique de corpus par des auteurs aussi prestigieux que Sinclair.

On peut certes commenter la linguistique de corpus du point du point de vue de la linguistique textuelle, comme l'a fait talentueusement Adam (2006), mais la synthèse entre linguistique textuelle, analyse de discours et linguistique de corpus ne se fera pas nécessairement sur le

[9] Je n'invente rien : partir de Word, passer par Excel pour arriver à Powerpoint, tel est aujourd'hui le cycle d'extraction et d'exploitation des « connaissances ».

mode éclectique si prisé par la tradition universitaire. Pour peu qu'elles prétendent à une cohérence, les théories inadéquates vont au-devant de désillusions. Dans bien des cas, les données de prédilection et les phénomènes attendus sont simplement absents ou n'ont ni l'importance ni la portée qu'on leur attribue ; certains observables attestés restent tout simplement inconcevables pour des théories trop puissantes et sans capacité descriptive - d'où l'hostilité des courants générativistes, Chomsky en tête, pour la linguistique de corpus. L'observation instrumentée peut revêtir une fonction critériale. Ainsi, Louw (2007) montre-t-il empiriquement, par simple convocation de concordances, l'invalidité de la fonction poétique jakobsonnienne, comme il récuse le simplisme de la stylistique cognitive.

La sémantique interprétative, dans les années 1980, a reformulé et adapté certains modèles partiels venus de l'IA comme celui de Wilks ou de Sowa ; mais depuis, avec les travaux sur des corpus étendus (cf. l'auteur, éd., 1995), le travail instrumenté sur corpus est devenu un critère heuristique et de validation (cf. Malrieu et Rastier, 2001, sur l'objectivation des genres). L'élaboration de modèles de l'intertexte et du document va dans le même sens.

Pour ce qui concerne les recherches récentes, les études sur la composante tactique des textes tirent profit des travaux sur la topographie textuelle (Salem, Mellet et Barthélémy) ; le repérage de passages pertinents peut s'appuyer sur les techniques de détection d'*isotropies* (selon Viprey), ou de groupements thématiques (selon Brunet). Sans pouvoir les détailler, je mentionnerai des projets qui ont mis ou mettent en œuvre – et à l'épreuve – la conception du texte que nous présentons ici à grands traits : le projet européen *Princip.net* était consacré à la détection automatique de sites racistes en trois langues (cf. l'auteur, 2006) ; le projet *Minos* (2008-2009, soutenu par l'ISCC) explore la détection de concepts considérés comme des formes sémiotiques dans les textes théoriques, en utilisant des coalitions de critères sémantiques et expressifs ; *C-mantic*, projet ANR conduit par l'ERTIM (Inalco) compare avec les mêmes méthodes des corpus pro- et antitabac en français, anglais et chinois ; enfin le projet *GenaText* utilise des algorithmes issus de la génomique pour détecter des passages.

Derrière les données textuelles se cachent ainsi de nouveaux observables, et tant d'autres que nous ne savons encore discerner, car les instruments et les hypothèses théoriques qui le permettraient restent encore à élaborer.

## Références

Adam J.-M. (1992). *Les textes : types et prototypes*. Bruxelles, Mardaga.

Adam J.-M. (2005). *La linguistique textuelle. Introduction à l'analyse textuelle des discours*. Paris, Nathan.

Adam J.-M. (2006). Autour du concept de *texte*. Pour un dialogue des disciplines de l'analyse des données textuelles. In *Actes des 8es JADT*, Besançon.

Berners-Lee T. (1998). *Weaving the Web*, Harper, San Francisco.

Berners-Lee T. (2007). Le web va changer de dimension. *La Recherche*, 413, pp.34-38. Propos recueillis par Marie-Laure Théodule.

Bird S. & Liberman M. (2001). A formal framework for linguistic annotation. *Speech Communication*, 1/2-33, pp.23-60.

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

Eco U. (1974). *Trattato di semiotica generale*. Milan, Bompiani.

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

Halliday M. A. K. et Hasan R. (1976). *Cohesion in English*. Londres, Longman.

Loiseau S. (2006). *Sémantique du discours philosophique : du corpus aux normes. Autour de G. Deleuze et des années 60*. Thèse de doctorat, Université Paris X-Nanterre.

Loiseau S. (2007). CorpusReader. un dispositif de codage pour articuler plusieurs interprétations. *Corpus*, 6.

Louw B. (2007). Collocation as the determinant of verbal art. In *Language and Verbal Art Revisited*, Donna R. Miller et Monica Turci (eds). Londres, Equinox, pp.149-180.

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

Mauceri C. (2007). *Indexation et isotopie : vers une analyse interprétative des données textuelles*. Thèse de doctorat, ENST-Bretagne (Telecom Bretagne) et Université de Bretagne Sud. Rééd. : http://www.revue-texto.net.

Pédauque R. T. (2006). *Le document à la lumière du numérique*. Caen, C&F éditions.

Poudat C. (2006). *Etude contrastive de l'article scientifique de revue linguistique dans une perspective d'analyse des genres*. Thèse de doctorat, Université d'Orléans, http://www.revue-texto.net.

Rastier F. éd. (1995). *L'analyse thématique des données textuelles*. Paris, Didier.

Rastier F. (2001a). L'action et le sens. — Pour une sémiotique des cultures. *Journal des Anthropologues,* 85-86, pp.183-219.

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

Rastier F. (2002). Anthropologie linguistique et sémiotique des cultures. *Une introduction aux sciences de la culture*, ch. 14, pp.243-267.

Rastier F. (2004). Doxa et lexique en corpus - Pour une sémantique des « idéologies ». In *Actes des Journées scientifiques en linguistique 2002-2003 du CIRLLEP*. Reims : Presses Universitaires de Reims.

Rastier F. (2005). Enjeux épistémologiques de la linguistique de corpus. In G. Williams (éd.). *La Linguistique de corpus*. Rennes : Presses Universitaires de Rennes, 31-46.

Rastier F. (2006). Sémiotique des sites racistes. *Mots*, 80, pp.73-85.

Rastier F. (2007a). Indices et parcours interprétatifs. In Denis Thouard, éd. *L'interprétation des indices*, Lille, Presses du Septentrion, pp.123-152.

Rastier F. (2007b). Passages. *Corpus*, 6, pp.127-162.

Rastier F. (à paraître). Web semantics v. Semantic Web. *International Journal of Corpus Linguistics*.

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

Valette M. (2004). Sémantique interprétative appliquée à la détection automatique de documents racistes et xénophobes sur Internet. *Approches Sémantiques du Document Numérique, Actes du 7e Colloque International sur le Document Electronique*, 22-25 juin 2004, Patrice Enjalbert et Mauro Gaio, eds, 2004, pp.215-230.

Valette M. et Slodzian M. (à paraître). Sémantique des textes et recherche d'information. *Revue française de linguistique appliquée* (soumis).