Philologie et/ou herméneutique numérique : nouveaux concepts pour de nouvelles pratiques ?

Damon Mayaffre

▶ **To cite this version:**

Damon Mayaffre. Philologie et/ou herméneutique numérique : nouveaux concepts pour de nouvelles pratiques ?. XXVIIe Colloque d'Albi Langages et Signification , Jul 2006, Albi, France. pp.15-25. hal-00551477

HAL Id: hal-00551477
https://hal.science/hal-00551477

Submitted on 3 Jan 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Corpus en Lettres et Sciences sociales

Des documents numériques à l'interprétation

Sous la direction de François Rastier et Michel Ballabriga

Fac-similé

Actes du XXVIIe Colloque d'Albi
Langages et Signification

Publiés par Carine Duteil-Mougel et Baptiste Foulquié

PHILOLOGIE ET/OU HERMÉNEUTIQUE NUMÉRIQUE : NOUVEAUX CONCEPTS POUR DE NOUVELLES PRATIQUES ?

Damon MAYAFFRE
BCL (UMR 6039), Université de Nice

Fac-similé

SOMMAIRE

Résumé *: L'enjeu des sciences du texte est moins d'administrer la preuve que de contrôler l'interprétation. Hors de l'obscurantisme théologique, il faut admettre en effet que les textes, et les corpus qui en informent le sens, n'ont point de Vérité mais de multiples compréhensions. Selon une pensée attribuée à Foucault, la vérité d'un texte est d'abord et seulement ce qu'on dit de lui, et déjà Chladenius remarquait que, loin de la stricte intentionnalité des auteurs, « l'on peut, lorsqu'on cherche à comprendre leurs écrits, former des pensées qui n'étaient pas venues à l'esprit de l'auteur » (Chladenius cité par Szondi 1989 : 32).*
Seulement, sauf à verser dans un subjectivisme débridé et une interprétation divinatoire, « ce qu'on dit des textes » et ces « pensées » qu'il est permis d'avoir à propos d'eux, doivent être étayés, vérifiables, contrôlés. Cela passe par une composition/organisation ad hoc *des corpus, une prise en considération minutieuse du matériel linguistique qui les constitue, et une démarche heuristique rigoureuse. Dans les trois cas, la révolution numérique apporte des réponses adéquates.*

Note liminaire

Adoptons le parti pris dans cette contribution d'agglutiner philologie et herméneutique. Leur définition/spécification mériterait un article à part entière. Leur association – notée ici sous la forme relâchée et consensuelle : « philologie et/ou herméneutique » – témoigne simplement que nous ne considérons pas seulement la philologie, de manière réductrice, comme une technique d'établissement des textes, mais aussi, pour ce faire, comme l'art de leur appréhension, c'est-à-dire de leur compréhension ; c'est-à-dire de leur interprétation. De la même manière, on ne désignera pas uniquement par herméneutique, l'interprétation théologique, philosophique, allégorique, etc. des textes, mais l'art d'en établir non seulement le sens profond mais l'origine exacte, le fond supposé mais la forme attestée, le contenu mais l'expression ; l'esprit des textes donc, mais avant cela, nécessairement, la lettre.

En un mot, prises chacune dans une acception pleine, philologie et herméneutique sont indispensables l'une à l'autre ; partie prenante l'une de l'autre. Longtemps artificiellement séparées, pour des raisons historico-épistémologiques plurielles que certains auteurs ont décrites, elles peuvent se réconcilier à la faveur de la révolution numérique dont il sera question dans cette contribution : il s'agirait même d'une des conséquences les plus heureuses, au sein des sciences de la culture, de la révolution numérique en question.

Cette position liminaire nous est directement inspirée par la lecture de P. Szondi (avant propos de J. Bollack), *Introduction à l'Herméneutique Littéraire. De Chladenius à Schleirmacher* (Cerf, trad. 1989) où l'idée d'une herméneutique « critique » ou « matérielle » – à défaut, directement, d'une herméneutique philologique – est défendue. Et par celle de F. Rastier, *Arts et sciences du texte* (Puf, 2001) qui semble être le principal penseur contemporain à établir le « projet d'unifier l'herméneutique et la philologie » (p. 276 ; cf. aussi p. 2) quand bien même ce projet passerait par une reconsidération de l'obje(c)t(if) de la linguistique et une prise en considération novatrice des possibles du numérique en matière de textes, de corpus, de procédures heuristiques, d'outils de recherche, de formalisation des parcours interprétatifs.

Fac-similé

Introduction

Ce propos débute sur un constat empirique, d'ordre personnel, mais qui est, semble-t-il, suffisamment partagé aujourd'hui en SHS pour être généralisé.

Dans le cadre d'une étude linguistico-historique du langage politique français, nous avons étudié, au milieu des années 1990, des corpus textuels *papiers* – puisés par exemple dans l'œuvre de Maurice Thorez, éditée en plusieurs volumes par les Editions sociales. Nous poursuivons aujourd'hui notre travail par l'étude de corpus textuels *numériques* – puisés par exemple dans l'œuvre de Jacques Chirac éditée en plusieurs millions d'octets par le site officiel de l'Elysée.

Au terme de cette évolution, il apparaît que ce qui pouvait être considéré comme un simple changement du support de l'objet de recherche (des corpus textuels donc, ici composés d'une collection de textes politiques contemporains) entraîne un changement de la perception de sa nature, de la nature de ses composants (les textes) et, par là, un changement de leur compréhension-interprétation.

Pour cette raison, il faut, sans crainte d'apparaître naïvement moderne, affirmer, en France, avec (Rastier 2001) dès le début du siècle, plus modestement avec (Mayaffre 2002-a), récemment avec (Viprey 2005) et encore, cette année, avec (Adam 2006) que la philologie et/ou herméneutique numérique révolutionnent non seulement notre rapport aux textes et à la textualité, mais aussi nos pratiques heuristiques quotidiennes, mais encore, tout simplement, nos connaissances et notre appréhension de la culture (textuelle) humaine.

La question est aujourd'hui moins de savoir si la révolution numérique est aussi importante que celle de l'imprimerie dont on sait le rôle dans la propagation de l'humanisme, de la Réforme et des Lumières – d'évidence elle l'est ; aussi importante et plus rapide – que de savoir si une révolution, fût-elle scientifique ou culturelle, peut, au-delà de se vivre, se théoriser ?

La question est surtout de savoir si, comme toutes les révolutions, la révolution du tout numérique – ici des corpus textuels numériques – saura résister au double danger qui la menace sur sa gauche et sur sa droite par la surenchère ou la restauration.

À sa gauche, le passage du papier à l'électronique a entraîné le développement de l'Analyse de Données Textuelles (ADT) et, de manière plus désincarnée, du Traitement Automatique des Langues (TAL). Or ces pratiques, si elles ne devaient être que techniques ou algorithmiques, et devaient toujours surenchérir vers l'automatisme, souffriraient d'un déficit philologique pour la première, et d'un déni philologique pour la seconde. Il y aurait là, autour des textes, un divorce désastreux entre elles et les humanités.

À sa droite, les tenants de l'ancien régime papier continuent une longue tradition qui n'a aucune raison de s'éteindre. En dépit d'une évolution que l'on peut juger comme inéluctable, la lecture empathique ou intuitive des textes – lecture pré-saussurienne d'une part qui fait fi des apports des sciences du langage, lecture anté-numérique d'autre part qui ignore les possibilités des nouveaux supports, des nouveaux médias, des nouveaux outils –, demeure encore aujourd'hui majoritaire en SHS et en appelle seulement, comme suprême argument, à la sensibilité et l'érudition de l'analyste. Les logiciels d'analyse de données textuelles par exemple restent au mieux des gadgets d'appoint dans l'art d'interpréter les textes ; au pire totalement ignorés. Le divorce serait alors à la fois social et scientifique : aux internautes d'un côté et aux linguistes spécialisés de l'autre le loisir de manipuler, télécharger, formaliser et disséquer les textes, aux lettrés érudits le privilège supposé de les goûter et de les comprendre.

1. Visions sur les corpus textuels numériques

Les textes sont des artefacts, les corpus des construits. Ces deux postulats de la linguistique textuelle et de la linguistique de corpus, difficilement contestables, et aux conséquences épistémologiques multiples, ne sont pas strictement liés à la révolution numérique. Mais il n'est pas un hasard si (Viprey 2005) les rappelle à l'occasion de son article « Philologie numérique et herméneutique intégrative » dans lequel il décrit les apports décisifs du support digital dans les sciences et arts du texte.

Tout se passe en effet comme si la transition vers le numérique avait rendu incontournables et impérieuses quelques évidences philologiques et/ou herméneutiques oubliées.

1.1. Le texte est un artefact

Le texte est un artefact (*artis factum* : fait de l'art), *phénomène d'origine humaine, artificielle*, comme l'indique la définition. Le passage du papier au numérique, le travail technique et quotidien

de saisie par exemple[1], la simple lecture du texte sur *son* écran *via* l'ascenseur de *son* traitement de texte[2], sans parler de la réflexion théorique et pratique sur l'édition numérique, les options de codage, de balisage, d'étiquetage, tout cela nous fait rompre avec l'idée qu'il existerait un texte naturel, dont la forme intangible serait le folio ou le livre avec sa couverture et sa pagination. Bien sûr, la philologie traditionnelle, en insistant sur les différentes éditions et en développant le comparatisme non hiérarchisé (voir récemment Heidmann 2005 ; Adam 2005), avait prévenu contre la naturalisation abusive d'un texte source et réifié. Mais la philologie numérique expérimente cette réalité tous les jours en relativisant la forme textuelle.

Cette relativisation peut aller loin dans l'Analyse de données textuelles puisque les logiciels permettent de faire apparaître, à l'écran, le texte sous différentes formes conventionnelles. La convention la mieux établie est la surface graphique ; et la stabilité relative de l'apparence graphique ne devra pas nous faire perdre de vue qu'il ne s'agit là que d'une convention. Mais à côté du texte graphique, nu ou brut, le texte lemmatisé et étiqueté peut aussi se laisser voir à l'écran. *HYPERBASE*, articulé au lemmatiseur *CORDIAL*, permet ainsi de juxtaposer, dans un même mouvement, plusieurs conventions (*illustration 1 : Texte brut et texte lemmatisé de Jacques Chirac (14 juillet 1995, conférence de presse)*). Dans la fenêtre de gauche le texte brut ; dans la fenêtre de droite le texte lemmatisé où tous les mots graphiques ont été ramenés à leur lemme d'origine et où chaque lemme est suivi d'un code de 0 à 9 pour les grandes catégories grammaticales (1 = verbe, 2 = substantif, etc.).

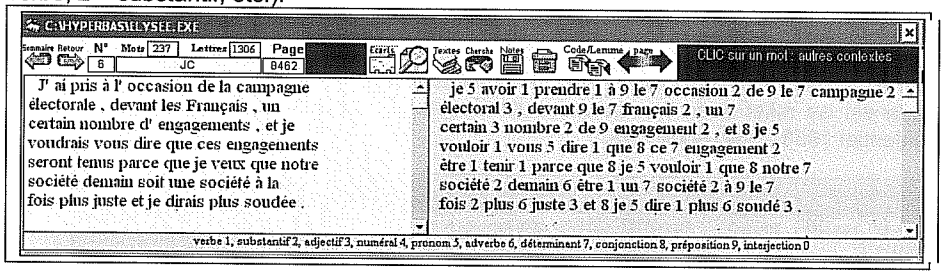

Illustration 1 : Texte brut et texte lemmatisé d'un discours de J. Chirac (14 juillet 1995)

Ce que nous voulons montrer, par cet exemple, c'est que le numérique en multipliant les mises en forme des textes propose une autre vision du texte. Un texte anti-naturel donc, dématérialisé – virtuel pourrait-on dire commodément –, dont les contours physiques tels que perçus depuis des siècles sont abolis, et la structure et le contenu – entendons, pour faire simple : la textualité – reconsidérés.

Il faut insister, ici, sur l'aspect le plus novateur de ces visions alternatives du texte que peut entraîner le numérique : *le dépassement/complément de la linéarité*.

La plupart des définitions du texte insistent en effet sur l'unité dynamique qu'il représente. La plus significative, dans ce sens, est celle que donnent (Détrie, Siblot, Vérine 2001) dont on souligne les éléments saillants :

> *Un texte est une* ***suite*** *d'énoncés oraux ou écrits posés par leur producteur – et destinés à être reconnus par leur(s) destinataire(s) – comme un ensemble cohérent* ***progressant vers une fin*** *et parvenant à constituer une complétude de sens. (Détrie, Siblot, Vérine 2001 : 349)*

---

1 Que saisit-on exactement ? Le corps du texte seulement ? La couverture et les en-têtes ? Et quelle édition choisir ? Quel format de restitution demander au logiciel de reconnaissance de caractères ? Même lorsqu'ils sont tirés de documents papiers, les documents électroniques ne peuvent être la reproduction exacte d'originaux, à moins de seulement photographier les textes. Mais précisément, nous aurions alors affaire à des images et non plus à des textes. Les derniers développements du format PDF sont intéressants à ce sujet. Pendant longtemps le PDF était la reproduction fidèle et intangible du format papier. Seulement, la manipulation de ces fichiers images a très vite paru rigide pour l'utilisateur. Aussi, il est désormais possible de transformer avec *PDF Converter* l'image en texte,... et le caractère intangible du contenu se trouve remis en cause.

2 La multiplicité et la personnalisation des écrans d'ordinateurs (taille, forme, résolution) et des traitements de texte (quelle police par défaut ? Mode page ou mode normal ?) font qu'aucun texte n'apparaît désormais au lecteur sous la même forme.

« Suite » (cf. aussi Rastier 2001 : 21), « plan » (Adam 1999 : 5) : la linguistique textuelle insiste, non sans argument, sur la linéarité, le déroulement séquentiel, l'enchaînement, la progression, la *cohésion*[1] d'un texte.
Pourtant le support et l'outillage électroniques permettent à moindre coup de doubler le point de vue de la linéarité par d'autres points de vue que proposent d'autres types de lecture.
Ont été relevées, dans (Mayaffre 2002-a), trois lectures électroniques complémentaires à la lecture oculaire linéaire traditionnelle : lecture quantitative (complémentaire de la lecture qualitative), lecture paradigmatique (complémentaire de la lecture syntagmatique), lecture hypertextuelle (complémentaire de la lecture textuelle). Et si l'on insiste sur la dimension complémentaire de ces approches, c'est que l'opposition entre numérique et oculaire n'a pas lieu d'être : la philologie et/ou herméneutique numérique entend prolonger, mais aucunement abolir, l'analyse de texte habituelle. *HYPERBASE* par exemple s'applique à croiser l'approche quantitative du texte et l'approche qualitative. Aux fonctions statistiques, caractéristiques du logiciel (« spécificités », « accroissement lexical », « distance intertextuelle », « corrélation chronologique », « richesse du vocabulaire », etc.), se combinent des fonctions d'exploration qualitative (« lecture », « concordance », « contexte »). Surtout, ces fonctions tentent de se féconder, de se juxtaposer, de se superposer dans l'ergonomie même du logiciel. Le bouton « Lecture », par exemple, donne accès au texte tel qu'il a été saisi et invite à une lecture linéaire, qualitative, intuitive, ordinaire en faisant défiler le texte, dans sa continuité, comme on tourne les pages d'un ouvrage. Pourtant si l'on actionne le bouton « Ecarts », alors le texte « naturel » s'anime et met en relief les mots qui sont caractéristiques statistiquement de la partie du corpus concernée. (*Illustration 2 : Lecture assistée d'un discours J. Chirac (3 avril 2002, interview télévisée)*). À gauche, le texte est lisse. À droite le texte est en relief avec les mots sur-utilisés par Chirac (par rapport à l'ensemble du corpus présidentiel 1958-2002) soulignés.

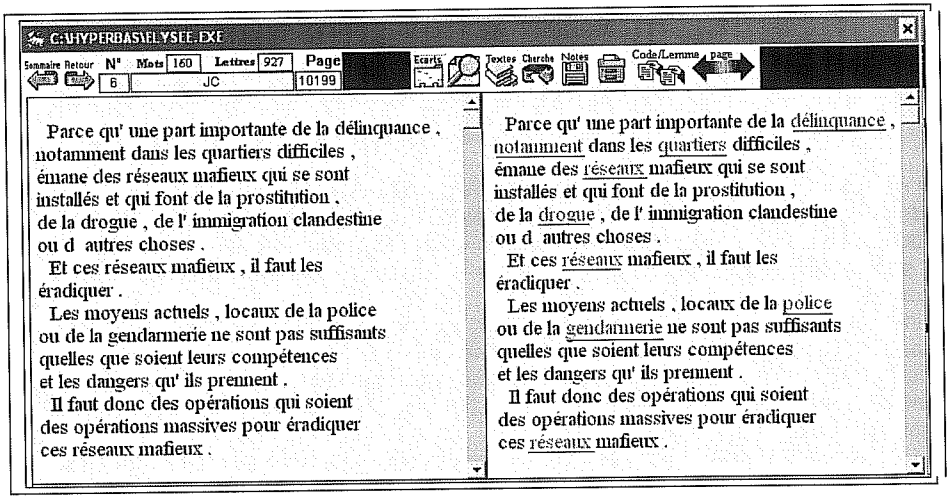

Illustration 2 : Lecture assistée d'un discours de J. Chirac (3 avril 2002, interview télévisée)

Le lecteur pourra donc lire et compter dans un seul élan. Sa lecture intuitive sera *assistée* par la statistique selon le mot d'Étienne Brunet, et l'esprit mis en alerte sur les mots quantitativement discriminants de telle ou telle partie du corpus. Loin d'être un gadget, la fonction « Ecarts » fond, en espérant les réconcilier, deux approches désormais bien établies du texte, deux traditions longtemps séparées : le scriptural et la métrique.
Lecture quantitative, lecture paradigmatique (par le biais d'index notamment), lecture hypertextuelle (par le jeu des liens et des renvois), disions-nous, en complément de la lecture

[1] À nous de montrer avec (Viprey 2005 : 66 et ss) que la *cohésion* d'un texte ne désigne pas seulement « sa **continuité** sémantique » (Détrie, Siblot, Vérine 2001 : 57) ou sa « **progression** thématique » (Charaudeau et Maingueneau 2002 : 99). En attendant, le concept s'inscrit bien dans la vision linéaire du texte.

linéaire usuelle : les mots étaient peut-être maladroits et (Viprey 2005) résume le changement en des termes plus percutants. Il fixe comme objectif à la philologie et/ou herméneutique numérique de combiner la lecture linéaire à des lectures *tabulaire et réticulaire*.

De fait, les logiciels d'Analyse de données textuelles, notamment ceux qui privilégient l'approche quantitative, commencent par faire exploser la linéarité du texte pour présenter leurs données en tableaux : tableaux alphabétiques, tableaux de fréquences, tableaux de distances, etc. Ces tableaux ne prétendent certes pas être le texte, mais ils sont une vision systématique et organisée – après l'explosion, le rangement – de la matière textuelle et deviennent les matrices sur lesquelles nos interprétations seront fondées.

Plus subtilement, l'enjeu le plus complexe de l'Analyse de données textuelles est de déceler les relations – relations autres que syntaxiques – que les items linguistiques entretiennent entre eux, non dans la phrase mais dans le texte en sa globalité. Texte, textualité, texture : l'objectif est de renouer avec l'étymologie même de ces mots et de démêler les trames et les entrelacs sous-jacents. Vision réticulaire donc des textes et des corpus qui met à jour les réseaux lexicaux pour (re)construire les thématiques, les isotopies ou isotropies récurrentes. De manière magistrale, (Viprey 2005), outillé par l'AFC, illustre le propos par l'étude de « l'organisation micro-distributionnelle » (*ibid.* : 61) des vocables dans le *Monde Diplomatique* grâce à l'étude du « système de collocation » (*ibid.* : 62). Et la fonction « Thème » d'*HYPERBASE* appliquée au corpus présidentiel français (1958-2002), permet de repérer les mots attirés par un mot pôle et de reconstituer ainsi dans une approche micro d'un macro corpus (la fenêtre d'étude étant le simple paragraphe et le corpus embrassé comptant plus de 500 discours) le système des co-occurrences qui font *nombre* c'est-à-dire *sens* (*illustration 3 : Environnement lexical du mot « mondialisation » dans le discours de J. Chirac*). Le tableau fait apparaître par ordre hiérarchique les mots qui sont le plus attirés par « mondialisation ». Trois traits isotopiques du discours peuvent ainsi être distingués. Dans un propos assez proche de l'altermondialisme, Chirac (i) dénonce les « dangers » de la mondialisation. Seulement, (ii) il juge le mouvement « inéluctable » et, pourquoi pas, porteur de certains « avantages ». Aussi (iii) milite-t-il pour une mondialisation « maitrisée » (voir Mayaffre 2004 : 133-140).

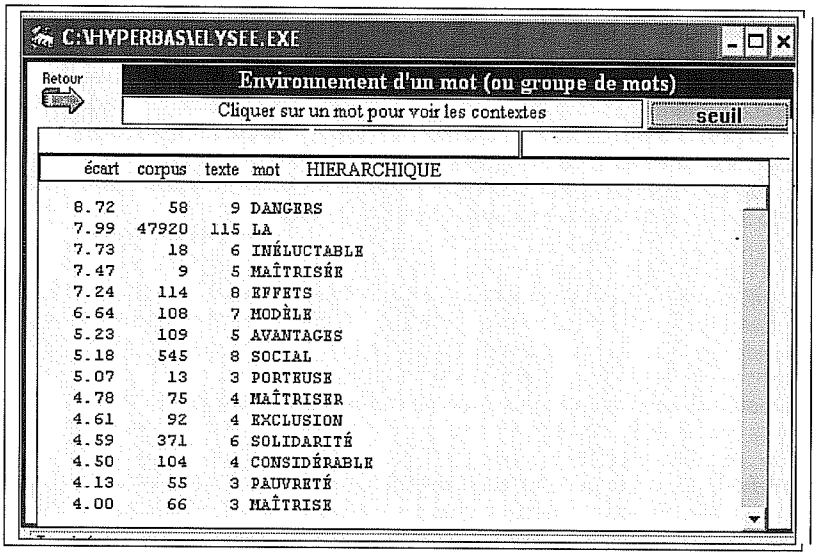

Illustration 3 : Environnement lexical du mot « mondialisation » dans le discours de J. Chirac

Bref, dans une concession décisive, (Adam 2006), un des meilleurs représentants de la linguistique textuelle traditionnelle, peut ainsi déclarer récemment devant les chercheurs en ADT :

> ***...la textualité doit résolument être pensée comme la combinaison de parcours linéaires et réticulaires.*** *(Adam 2006 : 5, souligné par l'auteur)*

Comme l'on sait que l'organisation du parcours linéaire a été le fait de la linguistique textuelle et des lectures oculaires depuis plusieurs lustres, l'on comprend que l'organisation du parcours réticulaire, désormais partie intégrante de la compréhension d'un texte, est laissée à la charge de l'approche assistée par ordinateur seule à même de formaliser des réseaux trans-phrastiques et a-séquentiels, à partir du moment où le texte est long et qu'il s'inscrit dans de gros corpus dont on prétend rendre compte[1].

Fac-similé

1.2. Le corpus est un construit… qui construit

Les corpus ne sont pas des objets donnés mais des objets construits. Cette affirmation, qui n'est plus, espérons-le, à démontrer[2], n'est pas, elle non plus, le fait du tournant numérique. Elle prend cependant un tour particulier avec lui.

Si l'ordinateur dématérialise le texte en l'arrachant de son support physique habituel, il matérialise, délimite, organise – en un mot : construit – les corpus plus strictement qu'ils ne l'étaient auparavant. Dans les SHS, les corpus avaient parfois cessé, en effet, d'être des réalités pour devenir des potentialités. Selon l'exemple personnel cité, notre corpus papier était composé des discours de Maurice Thorez, que l'on savait exister dans les bibliothèques-archives les mieux documentées et que l'on pouvait lire à l'occasion ici ou là. Mais jamais il n'a pris la forme d'un objet autre qu'intellectuel. D'autre part, et conséquemment, son organisation était pratiquement nulle. Au mieux pouvait-on se prévaloir d'une hiérarchie chronologique dans la pile partielle de photocopies que l'on envisageait de faire et de quelques fiches de renvoi d'un texte à l'autre susceptibles de suppléer l'organisation informelle – l'anarchie ? – de notre mémoire.

Le numérique est jusqu'à nouvel ordre plus contraignant en matière de constitution et définitivement plus performant en matière d'organisation.

— *Constitution*. L'on ne pourra considérer, de droit comme de fait, comme appartenant aux corpus que les textes que l'on aura fait l'effort de *saisir* (dans son acception pleine mais d'abord physique) et que l'on pourra soumettre, effectivement, aux logiciels d'exploitation. Si l'esprit humain peut se satisfaire de potentialités, avantageusement ajustables au fil de la recherche, le système binaire des logiciels (oui/non) ne supporte que les choix définitifs et les traitements ne pourront s'opérer que sur des objets réellement constitués. Par là, *la clôture* des corpus est toujours contraignante dans le travail numérique, lorsque les chercheurs avaient tendance à élargir ou rétrécir au cours de leur étude, au gré de leur humeur, leur corpus d'étude. Dans les termes de (Pincemin et Rastier 1999), une certaine confusion, au moins une porosité, était souvent maintenue entre *corpus de travail* et *corpus de référence*. Aujourd'hui, de manière implacable, un texte fera partie ou non du corpus de travail. Le simple décompte par l'ordinateur des unités linguistiques du corpus, par exemple, ne peut supporter aucune ambiguïté quant à l'appartenance ou non d'un texte au corpus de l'analyse ; plus généralement, le traitement statistique de la lexicométrie opère nécessairement, selon les lois de la norme endogène, sur des corpus clos et réels.

Cette clôture du corpus – essentielle pour la rigueur de la démarche scientifique – va de pair avec la prétention de l'exhaustivité du traitement. Clos, délimité, le corpus numérique sera soumis *dans sa totalité* au même traitement systématique et exhaustif. Là encore, il en va de l'entêtement algorithmique des machines. La recherche d'un mot par exemple, puis de ses co-occurrents, dès lors qu'elle pourra se faire ici, pourra s'effectuer partout dans le corpus, rompant ainsi avec le caractère aléatoire, partiel et partial de l'attention humaine.

Enfin, cette exhaustivité du traitement prendra sa valeur seulement lorsqu'on aura indiqué que la taille des corpus numériques semble ne pas avoir de limite là où la mémoire humaine ne peut embrasser que des ensembles de quelques dizaines de textes. Sans assistanat numérique (moteur de recherche, indexation lexicale, navigation hypertextuelle, traitement quantitatif, tri alphabétique ou hiérarchique, concordanciers), il paraît difficile de prétendre rendre compte d'un corpus de 100 discours politiques ; avec assistanat, il devient aisé de fouiller des corpus qui en comptent plusieurs milliers. Ce changement d'échelle de la taille des corpus, qui rend difficilement

---

1 Explicitement : « Nous avons, de toute évidence, besoin les uns des autres : tandis [...] que nous mettons l'accent sur la définition des unités élémentaires, sur le traitement de la linéarité des textes, sur les enchaînements transphrastiques et sur la combinatoire d'unités de rangs de complexité supérieurs à la phrase, vos travaux insistent sur la structure non-séquentielle et réticulaire des textes. » (Adam 2006 : 4 ; propos tenu à la communauté ADT le 19 avril 2006, à Besançon, à l'occasion des 8ème JADT).

2 Voir, par exemple, la philosophie de la revue *Corpus*, notamment, dès le premier numéro (Mellet 2002), puis (Scheer 2004), (Mayaffre 2005-b) etc.

contournable les descriptions quantifiées, est en lui-même déterminant, d'autant que les traitements d'ADT se fixent comme objectif de combiner analyse globale [décompte systématique des unités, typologies des textes, classifications automatiques ; ceci sur de grands corpus] et analyse locale [retour au cœur des textes, pointages hypertextuels et repérages spatiaux des unités dans leurs contextes (le mot, la syllabe, la lettre dans la partie, le paragraphe, la phrase)]. Conçus pour cela, les logiciels défrichent et déchiffrent ; imposent au corpus un traitement synthétique et un traitement analytique, articulent, pour reprendre la terminologie de l'herméneutique, l'analyse du *tout* (en général grâce aux fonctions d'exploitations statistiques) et l'analyse des *passages* (en général grâce aux fonctions d'explorations documentaires)[1].

— *Organisation*. Si le numérique apporte une rigueur appréciable dans la constitution (entendons donc pleinement : *la saisie*) de gros corpus, il offre surtout une possibilité sans précédent de les organiser afin de mieux les interpréter. C'est ici que se trouve l'enjeu épistémologique le plus important de la philologie et/ou herméneutique numérique.

Le sens naît en/du contexte. La linguistique textuelle pose que celui-ci est minimalement le texte. Sans ignorer la rupture que cela constitue avec la tradition saussurienne orthodoxe, il apparaît aujourd'hui que cet élargissement de l'objet de la linguistique de la phrase au texte, pour être subversif, n'est pas suffisant. Car dans la recherche ou la construction du sens, aucun texte ne se suffit à lui-même.

Il s'agit-là de thèses inutiles à plaider sauf à remettre en cause les notions établies de co-texte, d'intertextualité ou de dialogisme et à ignorer quelques grands auteurs tel Bakthine.

Précisément, la linguistique de corpus telle que nous la concevons se propose de formaliser, autant que possible, cet au-delà du texte. Elle considère les corpus bien conçus comme des lieux nécessaires qui permettent d'objectiver le co-texte des textes qui les composent, c'est-à-dire, comme des réseaux sémantiques auto-suffisants (ce que ne sont pas les textes seuls). Mieux : elle considère avec (Rastier 1998 : 17) que « le corpus est la seule forme possible d'objectivation de l'intertexte » immédiatement nécessaire à l'interprétation des textes constituants. En un mot, les corpus numériques – par leur taille et leur organisation – doivent être élaborés et perçus comme des architextes sémantiques qui comprennent, en leur sein, les ressources textuelles nécessaires à leur compréhension/interprétation[2].

Nous avons effectivement pointé ailleurs (Mayaffre 2002-b) l'injustifiable inégalité de traitement entre les textes analysés (le corpus) et les textes mobilisés comme ressources interprétatives (l'intertexte ou, pour restreindre le propos, le co-texte). Quoique de même nature textuelle, les premiers font l'objet d'une approche scientifique (sélection, regroupement, traitement linguistique), les seconds interviennent, à discrétion dans l'analyse, sans autre précaution. C'est pour palier cette anomalie épistémologique que le numérique et les possibilités qu'il donne, doivent permettre de fondre autant que possible source et ressources textuelles au sein même du corpus.

Pour ne pas manquer la vocation que nous lui assignons, à savoir celle de matrice du sens, le corpus doit donc tendre vers la mise en forme de parcours sémantiques ou interprétatifs valides et fertiles ; parcours endogènes au corpus donc, dans lesquels, répétons-le, texte et co-texte ne sont pas discriminés et où les ressources interprétatives se trouvent internalisées.

Pour cela, nous avons insisté sur la dimension *réflexive* que les corpus gagnent à avoir. En miroir, les textes du corpus doivent s'éclairer mutuellement ; se *réfléchir* les uns les autres ; chacun d'entre eux constituant le co-texte immédiat de tous, et l'ensemble, l'intertexte de chacun. Ainsi par exemple, l'étude du discours de Jacques Chirac qui a été entreprise (Mayaffre 2004) est passée par un corpus qui comprenait outre les textes du président actuel, ceux de ses prédécesseurs à l'Elysée. Les discours de de Gaulle, Pompidou, Giscard et Mitterrand constituaient à nos yeux l'intertexte générique et l'intertexte historique du discours chiraquien. Le corpus comprenait aussi les discours de Lionel Jospin car ils semblaient constituer, pendant la période de la cohabitation, le co-texte politique, immédiat et incontournable, des propos du président. Dès lors, les mots de

---

1 On ne saurait trop insister ici sur la puissance et la souplesse des ordinateurs pour décomposer/recomposer un tout en parties. Dans une mise en abîme impressionnante, l'utilisateur paramétrera dans sa recherche (i) l'unité à rechercher (le segment, le syntagme, les mots ou co-occurrents, les lemmes ou les codes grammaticaux, la chaîne de caractères, la syllabe ou la lettre) et (ii) la largeur de la fenêtre textuelle qui lui semblera pertinente pour la contextualisation (le corpus dans sa totalité, les textes, les parties, le paragraphe, la phrase, le début de ligne, la fin du vers, etc.).

2 Se risquera-t-on ainsi à préciser que le corpus devient alors l'objet nécessaire et maximal – comment imaginer un objet constitué ou empirique plus important ? – d'une linguistique aboutie ?

l'insécurité de Chirac, par exemple, n'ont pris sens qu'en considérant ceux que prononçait, en contrepoint, le Premier ministre durant la même période.
D'un point de vue technique, précisons simplement que la *réflexivité du corpus*, c'est-à-dire, au fond, la mise en dialogue des textes constituants, est assurée avant tout par les vertus de l'hypertextualité. Celle-ci semble être une solution puissante pour formaliser la notion d'architextualité de (Genette 1979) et rendre possible cette *sémantique de l'intertexte* que réclamaient les auteurs du (*Cahier de praxématique* 1999). Les logiciels d'analyse textuelle sur le marché considèrent en effet les corpus comme de vastes hypertextes : toutes les unités sont indexées et liées les unes aux autres. Le mot « délinquance », par exemple, trouvé dans la bouche de Chirac le 14 juillet 1996, renvoie non seulement à toutes les occurrences du mot dans le discours du président, mais dans ceux de Jospin. Et cette mise en résonance des textes, impossible à imaginer manuellement, est systématisée, grâce à l'hypertextualité et au traitement statistique contrastif, pour *toutes* les unités dans *l'ensemble* du corpus.
Quoiqu'utopique, l'idée de corpus réflexifs, c'est-à-dire la prétention d'internaliser les ressources textuelles interprétatives au sein de gros corpus dûment constitués, a été reprise par (Guilhaumou 2002 : 40), (Rastier 2005 : 31-32) et (Adam 2006 : 16), chacun insistant sur la dimension philologique et/ou herméneutique de la proposition. Car cette propriété des corpus, comme d'autres propriétés sur lesquelles il est impossible de revenir, revient à admettre qu'un « **moment philologique** » (Adam 2005 : 83 souligné par l'auteur) doit présider à leur constitution : l'acte interprétatif devant être pressenti au moment de la sélection et de l'organisation des textes en corpus. Au-delà de l'inévitable circularité de la démarche (cf. *infra 2.2.*, la question du cercle herméneutique), il s'agit de circonscrire le problème épineux du « point de vue » au seul geste inaugural de la recherche (le corpus comme un « point de vue ») pour mieux objectiver ensuite, dans le reste de l'analyse, l'interprétation.

2. Vers un contrôle de l'interprétation

L'enjeu des sciences du texte est moins d'administrer la preuve que de contrôler l'interprétation. Hors de l'obscurantisme théologique, il faut admettre en effet que les textes, et les corpus qui en informent le sens, n'ont point de Vérité mais de multiples compréhensions. Selon une pensée attribuée à Foucault, la vérité d'un texte est d'abord et seulement ce qu'on dit de lui, et déjà Chladenius remarquait que, loin de la stricte intentionnalité des auteurs, « l'on peut, lorsqu'on cherche à comprendre leurs écrits, former des pensées qui n'étaient pas venues à l'esprit de l'auteur » (Chladenius cité par Szondi 1989 : 32).
Seulement, sauf à verser dans un subjectivisme débridé et une interprétation divinatoire, « ce qu'on dit des textes » et ces « pensées » qu'il est permis d'avoir à propos d'eux, doivent être étayés, vérifiables, contrôlés. Cela passe, comme indiqué, par une composition/organisation *ad hoc* des corpus, cela passe aussi par la prise en considération rigoureuse du matériel linguistique qui les constitue.

2.1. Une herméneutique matérielle

L'herméneutique numérique est une herméneutique matérielle ; pas seulement par conviction mais par nécessité. Ou plutôt : avec le numérique l'évidence devient nécessité.
Pour les raisons techniques indiquées plus haut, la machine en effet ne saurait embrasser le texte autrement que par sa matière. Sauf à renverser le procès de la démarche et s'illusionner sur les possibilités de l'intelligence artificielle, l'ordinateur ne peut donner accès au sens d'un texte sans appréhender sa lettre ; il ne saurait aborder son esprit sans traiter (« saisir », « implémenter », « digitaliser », « numériser ») sa matière.
Concrètement, du côté de la linguistique quantitative, se retrouve ici l'objection la plus pertinente de (Tournier 1985 et 1987) contre les lemmatisations aveugles et toute forme d'analyse lexicométrique reposant sur un traitement linguistique liminaire du corpus. Lemmatiser un texte, c'est ramener son vocabulaire (particulier, historique, idéologique) à un lexique (universel, intemporel) : c'est plaquer sur des textes historiques un sens préalable (celui canonique du dictionnaire), là où l'analyse prétendait justement déconstruire/reconstruire froidement les textes pour faire percer le sens sous la surface matérielle, graphique – supposée neutre – du corpus[1].

[1] On objectera néanmoins à Maurice Tournier sa propre critique : la forme graphique qu'il réifie et semble considérer comme neutre est elle-même arbitraire, historique, conventionnelle. Le texte graphique

Se retrouve, aussi, ici, la critique la plus forte de la linguistique textuelle adressée à l'analyse du discours, qui, si elle n'y prend garde, « manque le texte en tant que tel » (Sarfati 2003 : 432) ; critique que l'on peut généraliser.
Manquer le texte pour l'analyse du discours, c'est prendre en considération les conditions de production des textes et négliger les productions elles-mêmes. Manquer le texte, pour l'herméneutique traditionnelle, c'est prétendre toucher l'âme des textes en négligeant leur chair. Manquer le texte pour la rhétorique, par exemple, c'est s'éblouir sur quelques fleurs de langage ou figures de style remarquables, lorsque la matérialité du texte, dans son ensemble, participe de l'éloquence du discours.
Chevillée donc à la matière textuelle, sûre de la description formelle des corpus, l'herméneutique numérique ne prétend certes pas produire des interprétations infalsifiables mais entend toujours s'appuyer sur des unités linguistiques attestées de textes établis. C'est en ce sens qu'elle peut se revendiquer de Peter Szondi et de son herméneutique critique ; c'est en ce sens que l'on parle d'une herméneutique philologique. Les parcours interprétatifs sont toujours sujets à caution[1], mais la trajectoire de ceux de la philologie et/ou herméneutique numérique a l'avantage d'être solidement inscrite dans la bonne direction grâce à son décisif et premier mouvement : par la prise en compte nécessaire, systématique et exhaustive, des matériaux linguistiques (lettres et syllabes, formes graphiques et lemmes, code grammaticaux et enchaînements syntaxiques, segments répétés, expressions, cooccurrents, collocations micro-distributionnelles, réseaux lexicaux, concordances phrastiques, contextes paragraphiques, etc.) des textes.

2.2. Cercle herméneutique et démarche inductive

La conséquence la plus directe de l'approche matérielle de la philologie et/ou herméneutique numérique est, nous semble-t-il, le caractère à dominante inductive de la démarche.
Les procédures de l'herméneutique, et peut-être celles de l'acquisition de la connaissance, sont prisonnières d'un *cercle* [cf. la plupart des auteurs qui ont traité le sujet et particulièrement Schleiermacher]. La première façon d'évoquer ce cercle a été rappelée dès la note liminaire : philologie et herméneutique apparaissent attachées dans une relation sans commencement ni fin : l'établissement d'un texte passe par sa compréhension profonde c'est-à-dire son interprétation, et l'interprétation ne peut reposer que sur un texte solidement établi. La seconde façon d'évoquer ce cercle est plus classique dans la littérature sur l'herméneutique. Elle souligne comment l'analyse du tout et celle des passages se trouvent liées dans un rapport sans issue. Le passage ne pouvant être compris que dans/par l'ensemble, et l'ensemble ne pouvant être construit qu'à partir de la compréhension des parties.
Précisément, Peter Szondi citant plusieurs fois Heidegger, nous invite à renoncer à l'idée d'échapper à cette circularité de la compréhension, de l'interprétation et de la connaissance, et pose que « l'essentiel [...] n'est pas de sortir du cercle, mais d'y entrer de la bonne manière » (Heidegger cité sous des formes différentes par Szondi 1989 : 10 et 105).
Pour notre part, l'on entre dans le cercle herméneutique, ou en tout cas dans le corpus, *par le bas*. Dans les termes que (Williams 2005) reprend à (Tognini-Bonnelli 2001), nos études sont *corpus-driven* (versus *corpus-based*) et la démarche *bottom–up* (versus *top-down*). La linguistique de corpus pose que le corpus n'est pas l'outil de la recherche (un corpus-ressources documentaires, un corpus-base de données, un corpus-échantillon représentatif de la langue) mais son objet vivant et dynamique ; il est non pas le réceptacle d'un sens déjà là mais sa matrice ; non pas une chose que l'on interroge, mais une chose qui nous interroge.
Dès lors, si le corpus construit un sens que l'on cherche à appréhender, si c'est lui qui, une fois constitué, conduit objectivement l'analyse, la meilleure démarche est celle qui permet de faire remonter l'information du tréfonds, afin de nourrir le plus objectivement possible nos interprétations.
Cette démarche à dominante inductive est rendue cohérente par les contraintes du numérique telles qu'évoquées précédemment. L'ordinateur décompose ses objets en plus petites unités sémiotiques. Et un corpus est pour lui d'abord constitué de lettres concaténées, de blancs et de

---

n'est, guère plus qu'un texte lemmatisé, un texte objectif ou naturel. (cf. Mayaffre 2005-a). Grâce à la performance des lemmatiseurs/étiqueteurs, et malgré leurs erreurs résiduelles, la surface lemmatisée ou grammaticalisée du texte n'est, aujourd'hui, guère moins contestable ou arbitraire que celle d'un texte brut.

[1] « Il est certain que l'on ne peut pas simplement biffer la part de subjectivité, et même d'affinité dans la démarche de la compréhension » (Szondi 1989 : 117-118).

ponctuation, d'octets et de bits. Si ces unités sont ensuite combinées, reliées, contextualisées (voire interprétées comme dans le cadre de la lemmatisation), la description comme l'interrogation numériques du texte s'appuieront sur ces signaux informatiques premiers et minimaux.
Cette posture inductive est non seulement essentielle dans l'acte descriptif qui précède l'interprétation, mais dans la démarche interprétative elle-même. De fait, en partant d'en bas, les logiciels peuvent décrire des gros corpus avec la minutie, la systématicité et l'exhaustivité mentionnées. Ils outillent donc le chercheur dans sa recherche d'indices objectifs. Mais, par les lectures complémentaires qu'ils proposent et les visions alternatives qu'ils donnent des textes (cf. *supra*), les logiciels d'ADT doivent surtout interroger différemment l'herméneute, loin d'hypothèses de travail, imposées par en haut et trop contraignantes. Car la plus-value attendue de l'ADT et de la philologie et/ou l'herméneutique numérique est avant tout une plus-value heuristique. Il s'agit de retourner la démarche hypothético-déductive dont l'usage apparaît trop dangereux dans les sciences humaines (cf. Mayaffre 2002-a), en faisant émerger, du corpus même, des hypothèses objectives de travail sur lesquelles on se met à réfléchir. Il s'agit de refuser le risque de projeter dans les textes un questionnement surplombant et un sens préalable, pour se laisser interroger par eux sans *a priori* et sans tabou.
Mieux que contrôler les parcours interprétatifs en se donnant les moyens de décrire puis de vérifier nos (hypo)thèses sémantiques, l'ADT se propose d'objectiver l'élaboration desdites hypothèses. Contrôler les conditions d'émergence des hypothèses apparaît ainsi comme le suprême objectif de l'herméneutique numérique, pour une meilleure maïeutique du sens.

Conclusion

Revenons pour conclure à l'élémentaire. Le moyen le plus simple de souligner l'importance de la révolution numérique dans la perception des textes, des corpus et dans les pratiques interprétatives est sans doute de montrer l'amplification décisive qu'elle représente avec le meilleur de la philologie et/ou herméneutique traditionnelle.
N'a-t-on pas souligné dans les humanités, l'apport scientifique de l'invention de la glose et des notes infrapaginales. Il s'agissait, tout à coup, d'enrichir le texte d'un paratexte et de renvoyer le lecteur à des références bibliographiques ou à des sources, bref à d'autres textes. La révolution numérique permet de rendre effective ce système d'enrichissement et de références et de transporter, sur-le-champ, le lecteur au-delà de son document d'origine. Dans l'édition électronique des *Fleurs du mal* qu'a entreprise (Viprey 2002), par exemple, les différentes éditions de l'œuvre de Baudelaire sont instantanément consultables ainsi que des dictionnaires, des graphes ou des index. Bien sûr, en cela, les pratiques numériques redéfinissent la conception du texte, en violent les frontières physiques (classiquement le livre), et renoncent définitivement à le percevoir comme une monade.
N'a-t-on jamais apprécié de pouvoir entrer dans une œuvre par l'index des noms de personnes, de lieux ou de notions ? Les corpus numériques et les logiciels d'ADT, dans leur plus simple appareil, se proposent, comme première étape, d'indexer l'ensemble des mots qui seront autant d'entrées au cours de la recherche. Outre l'amplification du travail d'indexation, cette généralisation signifie d'un point de vue épistémologique que les *a priori* quant aux mots jugés pertinents disparaissent. Par là, c'est donc une interrogation non bornée que l'on s'autorise, jusqu'à un renversement du système hypothético-déductif qui implique, toujours, une lecture orientée pour des interprétations convenues.
Enfin, n'a-t-on jamais navigué, à la recherche du sens, dans un dictionnaire ou une encyclopédie entre plusieurs articles *via* le système de renvois thématiques ? La conception numérique des corpus multiplie ces passerelles ; elle n'est que passerelles. Elle rend industrielle et, pour tout dire, enfin opératoire, l'artisanat dérisoire des renvois manuels. Par simple clic, la navigation hypertextuelle fait passer le lecteur, au sein du corpus, d'un mot à l'autre (de tous les mots à tous les autres), d'un texte à l'autre, d'un thème à l'autre. Les corpus textuels par leur taille et leur structure peuvent être perçus comme de gros architextes : leur lecture hypertextuelle et leur traitement statistique sont la condition de leur exploitation en tant que tels. L'enjeu apparaît alors aussi simple qu'évident : faciliter, organiser, contrôler, mieux qu'auparavant, la contextualisation des mots, des phrases et des textes constituants, sans laquelle aucune interprétation scientifique n'est envisageable.

Fac-similé

BIBLIOGRAPHIE

ADAM, J.-M. 1999. *Linguistique textuelle. Des genres de discours aux textes*, Paris, Nathan.
ADAM, J.-M. 2005. Les sciences de l'établissement des textes et la question de la variation, in J.-M. Adam et U. Heidmann (éds.), *Sciences du texte et analyse de discours. Enjeux d'une interdisciplinarité*, Genève, Slatkine, pp. 69-92.
ADAM, J.-M. 2006. Autour du concept de *texte*. Pour un dialogue des disciplines de l'analyse de données textuelles, in *JADT 2006* [texte en ligne sur *Lexicométrica* (http://www.cavi.univ-paris3.fr/lexicometrica/jadt/JADT2006-PLENIERE/JADT2006_JMA.pdf)].
Cahiers de praxématique 1999. « Sémantique de l'intertexte », n°33.
CHARAUDEAU, P. et MAINGUENEAU, D. (sous la dir.) 2002. *Dictionnaire d'analyse du discours*, Paris, Seuil.
DÉTRIE, C., SIBLOT, P., VÉRINE, B. 2001. *Termes et concepts pour l'analyse du discours. Une approche praxématique*, Paris, Champion.
GENETTE, G. 1979. *Introduction à l'architexte*, Paris, Seuil.
GUILHAUMOU, J. 2002. Le corpus en analyse de discours. Perspective historique, *Corpus,* 1, pp. 21-49.
HEIDMANN, U . 2005. Comparatisme et analyse de discours. La comparaison différentielle comme méthode, in J.-M. Adam et U. Heidmann (éds.), *Sciences du texte et analyse de discours. Enjeux d'une interdisciplinarité*, Genève, Slatkine, pp. 99-116.
MAYAFFRE, D. 2002-a. L'Herméneutique numérique, *L'Astrolabe. Recherche littéraire et Informatique* (http://www.uottawa.ca/academic/arts/astrolabe/).
MAYAFFRE, D. 2002-b. Les corpus réflexifs : entre architextualité et intertextualité, *Corpus*, 1, pp. 51-70 (http://revel.unice.fr/corpus/document.html?id=11).
MAYAFFRE, D. 2004. *Paroles de président. Jacques Chirac (1995-2003) et le discours présidentiel sous la V*ème *République*, Paris, Champion.
MAYAFFRE, D. 2005-a. De la lexicométrie à la logométrie, *L'Astrolabe. Recherche littéraire et Informatique* (http://www.uottawa.ca/academic/arts/astrolabe/).
MAYAFFRE, D. 2005-b. Les corpus politiques : objet, méthode et contenu. Introduction, *Corpus,* 4, pp. 5-19.
MELLET, S. 2001. Corpus et recherches linguistiques : introduction, *Corpus*, 1, pp. 5-13.
PINCEMIN, B. et RASTIER, F. 1999. Des genres à l'intertexte, *Cahiers de Praxématique*, 33, pp. 83-111.
RASTIER, F. 1998. Le problème épistémologique du contexte et le statut de l'interprétation dans les sciences du langage, *Langages*, 129, pp. 97-111.
RASTIER, F. 2001. *Arts et sciences du texte*, Paris, Puf.
RASTIER, F. 2005. Enjeux épistémologiques de la linguistique de corpus, in G. Williams (éd.), *La linguistique de corpus*, Rennes, Pur, pp. 31-45. [En ligne sur *Texto !* (http://www.revue-texto.net/Inedits /Rastier/ Rastier_Enjeux.html)]
SCHEER, T. 2004. Le corpus heuristique : un outil qui montre mais ne démontre pas, *Corpus*, 3, pp. 153-193.
SZONDI, P. (trad. 1989). *Introduction à l'Herméneutique Littéraire. De Chladenius à Schleirmacher*, Paris, Cerf.
TOGNINI-BONNELLI, E. 2001. *Corpus Linguistics at Work*, Amsterdam, John Benjamin's Publishing.
TOURNIER, M. 1985. Sur quoi pouvons-nous compter ? Réponse à Charles Muller, in *Études de philologie et de linguistique offertes à Hélène NAIS*, *Verbum* (numéro spécial), Presses universitaires de Nancy.
TOURNIER, M. 1987. *La réduction : principe de lexicométrie politique*, brochure de l'URL "Lexicométrie et textes politiques", 14 pages.
VIPREY, J.-M. 2002. *Analyses textuelles et hypertextuelles des* Fleurs du mal, Paris, Champion.
VIPREY, J.-M. 2005. Philologie numérique et herméneutique intégrative, in J.-M. Adam et U. Heidmann (éds.), *Sciences du texte et analyse de discours*, Genève, Slatkine, pp. 51-68.
WILLIAMS, G. 2005. Introduction, in G. Williams (éd.), *La linguistique de corpus,* Rennes, Pur, pp. 13-18.