UN INSTRUMENT DE LECTURE ANALYTIQUE : PRÉSENTATION DE *CORPUTEX*

Pierre SADOULET
CIEREC, Université de Saint-Étienne

SOMMAIRE

Résumé : *Quelle que soit la qualité du texte obtenu à la suite de la numérisation, ce texte n'en devient pas pour autant, comme magiquement, le révélateur de ce qu'il est. Pour le décrire et l'analyser, comme dans un travail sur livre, il faut pouvoir le lire et le relire non pas par fragment mais comme un tout – le global conditionne le local – et comme un tout dont chaque détail peut être très important. De plus un travail analytique doit repérer des récurrences ou des localités qu'il faut savoir identifier, analyser et classer à partir de la seule chaîne de caractères qui le signifie.*

Le travail de linguistique, de lexicologie et de sémiotique textuelle qui est le nôtre nous a conduit à privilégier l'activité herméneutique du philologue aux automatismes mécaniques de la machine et aux comptages porteurs d'une objectivité qui peut être illusoire. La machine doit offrir toute sa puissance à accélérer des tâches répétitives nécessaires (création de balises, lemmatisation, distinction des homonymes, indexation) ; mais elle ne peut remplacer le talent herméneutique de l'analyste qui doit, dans tous les cas, examiner le passage et décider de ce qu'il perçoit de son effet de sens, même si ses pesées toujours trop rapides font nécessairement l'objet de constantes révisions.

L'élaboration par l'Université de Irvine d'un corpus presque complet de la littérature grecque, le TLG, qui existe depuis plus de 20 ans, puis le remarquable travail de l'Atilf sur le corpus de Frantext ont constitué pour nous des moyens considérables. Mais les applications qui servent à les consulter nous ont conduit très rapidement à regretter une insuffisance qui nous gênait beaucoup : les fragments que ces applications extraient restent toujours trop courts et lorsqu'on retravaille les passages ainsi conservés, il faut toujours aller chercher le livre ou le texte pour retrouver le cotexte, ce qui conduit à différer la réponse au besoin créé par le moment de lecture particulier, voire souvent à renoncer à cette enquête nécessaire donc à risquer de laisser passer l'effet de sens dans toute sa complexité.

Corputex est une application écrite dans le progiciel de base de données 4D qui utilise l'interface et les fonctions de traitement de chaîne propres à ce logiciel d'une grande puissance pour proposer un instrument de travail qui permet de répondre à tous les besoins d'étude et d'analyse d'un texte long numérisé, tant au fil du texte qu'à l'intérieur de dossiers d'extraits. Un système de signets, de notes diverses permet d'annoter le corpus à tous les niveaux, y compris par l'extraction de citations. Une fois certaines occurrences marquées par l'analyste ou retrouvées par le logiciel (recherche lemmatisée, distinction d'homonymes avec une interface très rapide, recherche de concordances) les passages considérés peuvent être classés et commentés selon les besoins. Mais à chaque moment, il suffit d'un clic pour pouvoir retrouver tout le cotexte, qui est toujours disponible à l'affichage dans la base, si possible selon la même linéation et la même pagination que l'édition originale.

Ce logiciel par certains côtés peut apparaître comme une usine à gaz, de par la multiplicité des fonctions qu'il offre. Mais il nous a rendu de nombreux service d'abord dans des études sur le grec ancien mais aussi dans des travaux de linguistique et de sémiotique françaises qui demandaient de

repérer des passages représentatifs et des extraits permettant de mieux identifier au fil du discours certaines spécificités du fonctionnement sémantique et linguistique du corpus. Nous examinerons un cas particulier qui montrera l'apport de cet environnement logiciel.

Depuis 1980, année où j'ai passé mon doctorat de troisième cycle en morphosyntaxe du grec, j'ai occupé sinon perdu des années de travail avec l'idée que la possibilité de numériser un texte allait entraîner d'énormes progrès pour les sciences du langage, dans la mesure où l'informatique permettrait d'accélérer considérablement les procédures de mise en fiche. En effet, travaillant à une étude morphosyntaxique synchronique du grec ancien, j'avais passé des mois entiers à extraire du texte des passages pertinents, que je fichais selon une certaine méthode pour me permettre de repérer d'un simple coup d'œil les cas les plus intéressants. Je mettais ainsi en évidence des relations de commutation entre plusieurs constructions syntaxiques qui confirmaient empiriquement la description morphosyntaxique que je finissais par proposer.

De premiers essais informatiques m'ont donc conduit à une perte de temps considérable puisque je rêvais alors – erreur de jeunesse – d'un système informatique qui ferait, grâce à un travail de manipulation considérable, une sorte de préanalyse des énoncés offrant une visualisation directe des constructions, à l'image des fiches que j'avais confectionnées pour la thèse.

Comme j'étais alors professeur du secondaire et que j'avais, à côté de mon métier, une grosse activité militante, le temps utilisé pour toutes ces recherches était celui des vacances et, faute des informations nécessaires, j'ai commis des erreurs fondamentales au niveau du choix des logiciels. Plus tard, sur les conseils de Richard Goulet, un collègue helléniste qui avait conçu un premier logiciel d'analyse lexicale et de lemmatisation (*Lexis*), écrit d'ailleurs avec un système Pascal, j'ai fini par adopter l'application 4D, un progiciel de gestion de bases de données liées, afin de profiter au mieux des facilités techniques de ce progiciel, notamment par la diversité des fonctions de traitement de chaîne et de recherche qu'il offre et par la possibilité qu'il donne, grâce à un traitement de texte interne, intitulé *Write*, de produire des documents mis en forme et lisibles par *Word*.

L'objectif avait évolué alors : il ne s'agissait plus d'avoir un analyseur plus ou moins automatique, qui risquait de mémoriser dès le départ un mode de formalisation syntaxique qui serait ensuite retrouvé quasi sûrement par la théorie. Je voulais une application qui ait le moins possible d'intelligence artificielle. Par contre, elle devait offrir une interface aussi rapide que possible pour formater et garder à disposition de l'utilisateur le corpus dans son entier.

C'est ce qui m'a conduit à écrire *Corputex*.

Corputex est d'abord conçu comme un logiciel de recherche de chaînes semblable à *Frantext*. Des moyens de recherche par chaîne, par lemme ou, tout simplement, par balisage direct au fil de la lecture permettent de retrouver les passages pertinents dans le cadre d'une étude donnée : ces « extraits » réunis par le logiciel dans des « dossiers » peuvent être ensuite analysés un à un, par la mise en place de classements et de commentaires à partir d'ensembles de propriétés hiérarchisées établies par le chercheur. Cet étiquetage par clé des extraits, toujours à revoir, peut se faire, au départ, très rapidement, grâce aux possibilités de l'interface. Il est ainsi fait appel à l'intuition immédiate de l'analyste qui, bien sûr, devra revoir ensuite plusieurs fois les extraits les plus caractéristiques, pour affiner les distinctions. Il finira enfin par sélectionner les plus intéressants qui lui serviront pour un travail approfondi avec, à ce moment-là, surtout, une évaluation attentive du *cotexte*. La tâche sera d'assurer définitivement les effets de sens des extraits dans leur contexte grâce au recours qu'il est possible de faire à l'ensemble du corpus et aux aides-mémoire qui y auront été aménagés.

Même si *Corputex* sait compter tout ce qu'il trouve, il ne fait jamais de statistiques compliquées. Car tout ce qui est fait sur *Corputex* est le résultat d'un jugement posé par son utilisateur. Chaque base de donnée, consacrée à un texte et à une étude particulière, est donc conçue strictement comme un instrument de travail individuel. Ce qu'il contient, c'est un corpus lu et annoté par un individu.

Il n'est pas question ici de faire une présentation ou une démonstration du logiciel mais d'en décrire les fonctions essentielles, tout en indiquant les bases théoriques qui ont conduit à son élaboration. Nous verrons donc d'abord son cahier des charges. Puis nous examinerons quelques fondements qui justifient les services qu'il peut rendre au professionnel des sciences du langage. Pour conclure, nous évoquerons trois expériences d'études faites avec l'aide de *Corputex*.

1. La conception du cahier des charges

1.1. Les insuffisances de l'existant

* Le TLG : le logiciel *Pandora*

Lorsque, après mon troisième cycle, dans les années 80, je m'orientais vers ce travail d'élaboration d'interface logiciel, il commençait à être diffusé, pour les hellénistes, un instrument de travail très précieux, le *TLG* (*Thésaurus linguae graecae*), un CD élaboré par l'Université d'Irvine[1], qui contenait l'ensemble des textes connus de la littérature du grec ancien. Ce corpus – critiquable du fait qu'il avait dû, pour des raisons de droits, reprendre des éditions anciennes des auteurs – n'en représentait pas moins un instrument d'investigation très puissant.

D'abord réservé au système d'ordinateur Ibycus, le corpus devint très vite accessible à tous les micro-ordinateurs *Apple*, grâce au logiciel *SNS Greek* de l'Université de Pise, et au système *Pandora*, développé par l'université de Chicago[2]. D'autres ont été développés depuis pour le système *Windows*[3].

Ces logiciels offraient la possibilité d'extraire des œuvres entières. Mais la recherche d'occurrences à partir d'un traitement de texte s'avérait difficile, en raison du codage particulier des jeux de caractère grecs accentués. Il fallait donc utiliser les logiciels eux-mêmes pour extraire les passages contenant telle ou telle chaîne de caractère. Un système de codage très complexe permet, dans ces systèmes, la recherche des séries de formes correspondant à une déclinaison ou à une conjugaison. Avec une bonne technique, il était donc possible de retrouver toutes les variantes morphologiques d'un lemme. La série d'occurrences ainsi retrouvées était exportée sous la forme d'une suite d'extraits du corpus contenant au plus quatre lignes avant et quatre lignes après.

Il était possible ensuite de traiter ces données soit dans un traitement de texte soit, après importation, dans des bases de données spécifiques[4].

* *Frantext*

Près de dix ans après, j'ai pu avoir accès à un autre système largement utilisé par tous les spécialistes de langue française : il s'agit de *Frantext,* la base de donnée de l'ATILF, disponible sur *Internet*[5].

Il faut dire que nous retrouvions le même défaut : même si nous augmentions au maximum le nombre de lignes copiées pour l'export, les extraits restaient parfois insuffisants pour permettre certaines identifications nécessaires à la bonne interprétation du passage. Or c'était justement ces données qui semblaient les plus décisives : c'est quand une occurrence résiste à l'interprétation immédiate qu'on peut penser avoir un cas qui permette d'améliorer le système descriptif utilisé. Que ce soit en morphosyntaxe ou en lexicologie, et même en sémiotique, ce sont les contre-exemples apparents qui peuvent faire avancer le modèle d'analyse[6].

1.2. De nouveaux objectifs pour un outil commode

Partant du constat de ces insuffisances très pratiques, j'ai donc essayé de les dépasser pour écrire l'application *Corputex*. Son cahier des charges, d'abord élaboré à partir de ces expériences, s'est complété peu à peu au fil des besoins.

1.2.1. Retrouver l'oiseau rare qui permettra de comprendre enfin

L'objectif de départ du projet est d'abord de faciliter un travail en linguistique. L'axiome premier de la démarche scientifique était qu'il fallait trouver un moyen de renouveler la description linguistique dans les nombreux cas où elle s'avère insuffisante en cherchant « l'oiseau rare », l'exemple

[1] réf. : [s.a.], 1991 – *Thesaurus grec de Irvine*, Californie. CD ROM.

[2] Il s'agit d'un pile Hyperbase pour le système Macintosh. Dernière version connue : 2.5.2. Voir l'aide au logiciel http://www.lib.uchicago.edu/e/ets/TLG.html .

[3] Si nous arrivons à résoudre certains problèmes de transcodage et d'identification de l'organisation du CD, il est techniquement possible d'avoir une version de *Corputex* qui gère directement les données du CD.

[4] Voir P. Sadoulet, 1996

[5] Adresse actuelle : http://atilf.atilf.fr/frantext.htm. Nous rappelons que cet accès n'est disponible que sur abonnement.

[6] De plus, dans *Frantext*, avant la mise au point d'un corpus catégorisé permettant de différencier les homonymes, les recherches donnaient des résultats trop importants, ce qui obligeait ensuite à un fastidieux travail d'élimination.

inattendu qui serait l'occasion d'identifier une propriété particulière, le cas qui nous révèlerait, pour ainsi dire, qui nous ferait trouver le fonctionnement caché rendant compte de toutes les virtualités d'emplois de telle ou telle unité linguistique[1].

1.2.2. Extraire tout ce qui est pertinent par lecture directe ou par recherche de chaînes du signifiant

La recherche de ces « oiseaux rares » supposait d'abord des lectures et des sélections depuis le corpus entier. *Corputex* nous permettait de lire directement l'ensemble du texte. Les extractions pouvaient se faire à l'ancienne, en marquant scrupuleusement les passages intéressants du corpus par des balises qui permettaient à *Corputex* de les retrouver pour les copier ensuite dans un dossier.

Mais, souvent, il s'avérait qu'il était plus rapide de réunir les extraits à partir de recherches sur le signifiant, quitte à éliminer par un simple clic les occurrences non pertinentes. Il a fallu donc d'abord concevoir un système de recherche par chaîne qui puisse exiger une correspondance des codes ascii, pour tenir compte des caractères accentués et des majuscules. Puis j'ai pu mettre au point des algorithmes spécifiques, qui existent aussi dans *Frantext*, pour des recherches par mots ou même pour identifier des concordances et autres collocations.

1.2.3. Lemmatisation et distinction rapide des homonymes

Une indexation automatique permet de mémoriser toutes les occurrences des mots du corpus, y compris en traitant les mots composés avec tirets et en réunissant les formes ayant une césure en fin de ligne. On peut ainsi visualiser toutes les formes d'un même lemme. Des procédures rapides par simple clic permettent d'affecter à chaque lemme toutes ses formes attestées dans le corpus et de réunir leurs occurrences dans un fichier d'index spécifique.

Il apparaît alors de « vrais » homonymes : après indexation automatique, il est possible de distinguer très rapidement ces homonymes par « *glisser-déposer* » pour éviter ensuite toute confusion[2].

1.2.4. Autres recherches

Au fur et à mesure des utilisations, d'autres recherches plus complexes sont apparues comme nécessaires. Elles ont pu être insérées à l'application.

* Synthèmes ou lexies

L'indexation automatique repose sur la reconnaissance des séparateurs de mots ou l'identification des mots composés à partir de leur tiret. Mais il existe des groupes de mots figés ou des expressions qui ne pouvaient être identifiés comme tels à partir des séparateurs de mots. Ces expressions peuvent être recherchées puis *Corputex* en mémorise les références dans un fichier d'index.

* Recherche de passages à partir de citations

Parfois on peut avoir rencontré une citation du texte sans savoir où elle se trouve dans le corpus : *Corputex* sait la retrouver, pour peu qu'elle soit incluse dans une seule ligne[3].

* Indexation thématique

Un des derniers outils créé dans *Corputex* est la constitution possible, à partir d'une série de chaînes de caractères, d'une fiche d'index réunissant toutes les occurrences synonymes ou coréférentielles d'un thème particulier ou d'un personnage. Cela permet en particulier de pouvoir mémoriser manuellement les occurrences de certaines désignations pronominales d'un

[1] Un instrument de recherche sur corpus comme *Frantext* avait déjà, bien sûr, cette fonction.

[2] Le rythme moyen du traitement de ces homonymes est de l'ordre de 500 occurrences à l'heure. Autrement dit, seul le traitement qui différencie l'article défini « le » du pronom personnel homonyme représente un travail un peu trop long. Mais il est possible de le faire en plusieurs fois. Et à tout moment une erreur peut être corrigée.

La mémorisation d'une catégorisation grammaticale pour chaque lemme permettrait en fait de préparer la répartition en utilisant les tests distributionnels connus : « le » article avec un nom à droite, « le» particule préverbale devant le verbe ou d'autre particules.

[3] En fait il peut aussi parfois reconnaître une « suite » située en début de ligne suivante, mais l'algorithme n'est pas encore sûr à 100%.

personnage ou de certaines périphrases le désignant. L'indexation de toutes ces occurrences dans une fiche spécifique permet de pouvoir retrouver, avec un peu de patience, l'ensemble du parcours d'un personnage dans le corpus. On peut aussi réunir toutes les occurrences rendant compte du parcours thématique d'une notion ou d'un thème, quelles que soient ses différentes désignations[1].

1.2.5. Classer les trouvailles pour pouvoir les différencier et les retrouver facilement

Une fois ces recherches faites, il faut enregistrer le *sous-corpus* considéré, composé des passages intéressant le problème à étudier. *Corputex* permet de classer dans un dossier spécifique les trouvailles ainsi faites, éditées sous formes de fiches qui peuvent, elles-mêmes, recevoir un classement et tous les commentaires utiles. Si, par la suite, le texte du corpus est amélioré, chaque extrait reprendra le nouveau texte pour la bonne raison que les fiches d'extraits ne mémorisent que la référence du passage : elles recopient systématiquement le texte du corpus d'origine lors de l'affichage de la fiche[2].

Le travail d'étude des extraits dans les dossiers consiste d'abord à les classer selon des critères hiérarchisés posés par l'utilisateur. Ce travail de classement est assez rapide, du fait de l'interface de travail[3]. Chaque dossier peut être dupliqué pour être soumis à plusieurs classements successifs qui peuvent, d'ailleurs, relever d'analystes différents. Il est possible aussi de déplacer certains extraits d'un dossier à l'autre soit de façon individuelle, soit de façon collective. Un export de toutes les fiches se fait en format texte.

Chaque extrait, chaque dossier peuvent recevoir des commentaires.

1.2.6. Pouvoir lire tout le cotexte *nécessaire et surtout pouvoir le lire immédiatement !*

Avant d'analyser le sous-corpus, l'utilisateur choisit le nombre de phrases ou de lignes demandées avant et après la ligne référencée. S'il a besoin d'en voir plus, le passage concerné peut être élargi par un simple clic. S'il veut avoir accès à tout le corpus, cela est possible de façon immédiate, en pouvant retrouver, directement sur le corpus, le passage considéré.

Lors de ces lectures visant à mieux identifier le *cotexte* et son *contexte référentiel*, *Corputex* permet de faire toute une série d'investigations sur l'ensemble du corpus, en particulier, d'installer des index thématiques pour mieux se représenter tous les détails utiles de l'intrigue, comme par exemple, ce que vient de vivre le personnage.

Nous remarquerons ici un inconvénient pratique de *Corputex* dont nous reparlerons en fin d'exposé : quand l'utilisateur se met à annoter le corpus, la tâche devient une sorte de tonneau sans fond. Sa curiosité le conduit à faire de nombreuses recherches et comme le corpus est lu comme un rouleau continu et non page par page, le pauvre lecteur finit par ne plus savoir où il est, où il en est et ce dont il s'occupait au départ. S'il utilise l'application pour préparer des cours, surtout s'il se lance à l'improviste dans des recherches non prévues, il finit vite par ne plus avoir le temps de faire le travail projeté au départ.

2. Les conditions théoriques de cette élaboration

Il est sûr que ce serait bien plus simple d'avoir un instrument de recherche ou d'évaluation qui sélectionnerait, sur des critères quantitatifs, ou par analyse automatique, ce qui serait plus pertinent.

Cela réduirait d'autant le fastidieux travail d'analyse extrait après extrait[4]. La productivité du travail de recherche en serait grandement améliorée.

Il faut donc examiner maintenant pourquoi il faut tant tenir à un outil qui oblige à cet énorme travail personnel du chercheur qui, comme le Saint Thomas des Évangiles, ne voudra prendre comme

[1] Il est bien sûr toujours possible de supprimer toutes les références qui s'avéreraient erronées après vérification.

[2] Ce principe a été un peu réduit dans la mesure où il s'est avéré nécessaire de conserver un certain balisage qui met entre chevrons l'empan de texte qui semble pertinent pour attester du fonctionnement de l'extrait par rapport au problème posé. Le texte de la ligne convoquée ou du passage délimité par les chevrons est conservé dans la sous-fiche. Mais une procédure simple – un effacement du champ contenant ce texte – permet de retrouver le texte modifié du corpus.

[3] Si la lecture n'est pas trop difficile on peut atteindre les 100 extraits à l'heure.

[4] Pour permettre un gain de temps, il est cependant possible de faire des sous-sélections au hasard dans le dossier d'extraits. Chaque dossier est pratiquement limité à 200 occurrences.

fait empirique que ce qu'il aura vu et qu'il aura pu interpréter et évaluer. Cette méthodologie s'appuie sur certains principes avancés par la *sémantique interprétative*. Mais nous verrons que le point de vue *praxématique* que j'adopte dans ma démarche en sciences du langage l'exige plus encore.

2.1. Herméneutique philologique

Nous ne reprendrons pas ici tout ce que François Rastier expose très clairement dans *Arts et science du texte,* notamment dans son chapitre III « Philologie numérique »[1]. Ce sont des bases méthodologiques sur lesquelles il n'est pas nécessaire de revenir ici.
Mais pour bien caractériser la méthodologie scientifique qui sous-tend l'écriture de l'application, il faut quand même expliciter quelques éléments que nous formulerons à notre façon, marquant ainsi certaines originalités par rapport au point de vue de la sémantique interprétative.

2.1.1. L'établissement du signifiant

Un helléniste ne peut déroger à un principe philologique fondamental : il faut essayer d'avoir le meilleur texte possible, c'est-à-dire celui qui a le plus de chances de correspondre à celui qui a été établi par l'auteur. La paléographie ancienne puis les règles de l'édition critique moderne ont établi des critères précis pour atteindre au moins mal cet objectif. Et beaucoup d'éditions modernes montrent de grandes qualités dans ce domaine.
Pour des raisons de temps, ou tout simplement de droits, il n'est pas toujours possible de disposer de la meilleure version numérique du texte étudié. En effet les documents dont on peut disposer sur *Internet* sont, on le sait, des versions souvent fautives dans le détail, du fait qu'elles ont été établies par des logiciels de reconnaissance de caractères et que les relectures qui ont suivi ne peuvent pas ne pas laisser échapper certaines des erreurs de détail que commettent les logiciels. C'est pourquoi on peut découvrir des coquilles. Il faut bien sûr les corriger dès qu'on les repère. Cette possibilité est prévue par *Corputex.*
Si l'on fait le travail de numérisation soi-même à partir d'une édition reconnue (ne serait-ce que celle qui est préconisée par le programme d'agrégation), il faut utiliser une technique qui reproduise exactement la pagination et les retours lignes, autrement dit la *linéation* de l'original afin de pouvoir retrouver très facilement le passage sur l'édition papier[2]. *Corputex* mémorise toutes ces références, y compris les numéros de page, pour permettre toutes les vérifications nécessaires sur l'édition originale. Un système de notes de bas de page permet même de montrer un apparat critique, quand il a été numérisé à partir de l'édition[3].

2.1.2. Le signifiant comme interprétant

À l'inverse de toute une tradition idéaliste qui tend à valoriser le sémantisme et le contenu intellectuel, nous poserons comme axiome d'une herméneutique matérielle qu'un des interprétants fondamentaux pour toute interprétation reste les différences et les réseaux explicités par le signifiant[4]. Il faut donc que celui-ci soit établi avec précision. De plus, l'objectif de fonder des sélections d'énoncés comparables dans une étude en sciences du langage sur des recherches de

---

[1] Ouvr. cit. pp. 73 ss

[2] Peut-on garder l'espoir que la possession de l'édition papier autorise le fac-similé privé que constitue la numérisation pour satisfaire aux droits des auteurs ? Il y a lieu de penser en tout cas que ce serait une civilité indispensable pour rétribuer le travail des éditeurs.

[3] Toute modification personnelle d'importance peut être saisie dans la même note.
Signalons ici qu'un système de macros pour le logiciel Microsoft Word permet de préparer assez facilement le balisage des références pour le corpus, pour peu que l'on ait eu soin de bien numériser les numéros de page.

[4] La perception de ce signifiant ne peut être indépendante de la reconnaissance de ce signifiant comme produit d'une production de sens. Comme l'écrit François Rastier, « les relations constituantes du sens vont de signifiés en signifiés, mais aussi des signifiés vers les signifiants : ainsi, la *sémiosis* se définit comme un réseau des relations entre signifiés au sein du texte, en considérant les signifiants comme des *interprétants* qui permettent de construire certaines de ces relations (cf. *supra,* chap. I). (…) En d'autres termes, le sens n'est pas donné par un codage préalable qui associerait strictement un signifiant et un signifié ou une classe de signifiés (car la langue n'est pas une nomenclature) : il est produit dans des parcours qui discrétisent et unissent des signifiés entre eux, en passant par des signifiants. » *ouvr. cit*. pp. 103-104. Mais dans cette dialectique, nous insisterons sur le caractère décisif de l'ancrage dans le signifiant concret de tout ce qui contribue à en enrichir la signifiance.

chaînes signifiantes, comme le font *Frantext* et d'autres logiciels comme *Lexis*, *Pandora* et *Corputex,* n'est pas un bricolage provisoire en attendant que l'intelligence artificielle nous donne les instruments pour faire mieux. Il s'agit d'un outil essentiel.

2.1.3. L'activité du linguistique est une démarche herméneutique

François Rastier remarque que toute analyse de propriétés morphosyntaxiques, de significations ou de représentations par le langage est une interprétation et non la simple description de faits de langues, indépendants des cultures qui les ont créées historiquement. Les sciences du langage ne peuvent élaborer que des constructions herméneutiques.
Comment donc est-il possible d'assurer philologiquement de telles constructions, si le linguiste lui-même ne s'emploie pas à vérifier chaque cas avec son propre jugement ? C'est lui qui doit s'assurer de l'adéquation de l'interprétation. Il ne peut laisser cette vérification à la machine. Si donc l'activité herméneutique veut être philologique – c'est à dire aussi soucieuse que possible de garantir l'altérité du texte dans l'intégrité de son signifiant, comme dans la richesse de sa portée signifiante – il faut que l'interprète mette son nez partout pour assumer et vérifier toutes les analyses que le logiciel a mémorisées.

2.2. Un point de vue praxématique

La mention faite ici de la *sémantique interprétative* ne doit pas dissimuler que, dans tous mes travaux en sciences du langage, j'ai toujours pris en compte, y compris comme sémioticien, le point de vue *psychomécanique* et *praxématique*.
Car pendant les dix ans que j'ai passés à Montpellier, j'ai eu l'occasion de travailler avec Robert Lafont et son équipe. Et j'ai trouvé, dans cette théorie matérialiste, une base très solide pour décrire le fonctionnement concret de la langue[1].

* Robert Lafont

Robert Lafont propose une théorie du signe beaucoup plus radicale encore que la sémantique interprétative : pour lui, il n'existe plus de signe biface mais seulement des outils de production de sens, les *praxèmes,* qui sont de simples signifiants liés à une *praxis*, c'est à dire à un programme de sens expérimenté par chaque individu. Ce programme relève plus d'une sorte de mode d'emploi pour un parcours symbolique à travers un lexique virtuellement hiérarchisé que des contenus notionnels à proprement parler.
Il me semble d'ailleurs que l'on devrait remarquer, pour mieux la comprendre, que cette théorie particulière doit plus qu'elle ne le dit à la tradition du refus de la prise en compte du sens défendu par *l'antimentalisme* de la tradition distributionnaliste. Celui-ci refuse de parler du sens et ne s'occupe que de l'analyse des distributions des signifiants : le *praxème,* lui aussi, n'est qu'un signifiant qui sert *d'outil de production de sens* dans le cadre du réglage taxinomique qui définit ses conditions d'emploi. Mais, au-delà de cette tradition, bien sûr, la *praxématique* ne se contente pas de faire confiance au dispositif syntaxique pour expliciter une signifiance, elle ajoute que le sens lui-même est produit par la sélection des signifiants concrètement effectuée par l'énonciateur. Celle-ci consiste en une *visée lexicale*, en un parcours présupposé du locuteur à travers l'organisation hiérarchisée des signifiants du lexique, la *logosphère*. Cette *visée* conduit, par pesée des différences entre les potentialités de glose autrement dit les *traits* affectés à chacun des signifiants, au choix, à une *saisie* du praxème le plus pertinent.
On voit bien ici, qu'au-delà de ce refus de la correspondance biunivoque signifiant/signifié, réifiée dans l'oukase d'une mise en miroir essentialisante, la *praxématique* récupère à sa façon les théories guillaumiennes.

* La psychomécanique comme théorisation d'une herméneutique généralisée

Elle reprend, en particulier, la notion de *visée* et de *saisie* propres à la *psychomécanique guillaumienne*, pour interpréter certains constituants de l'énoncé comme *l'ancrage*, la manifestation des opérations qui ont conduit à leur choix. Dans ces cas, le signifiant mémorise, en quelque sorte,

[1] Ce point de vue spécifique nous écartera parfois de la sémantique interprétative qui tend à valoriser le signifié, ce qui est attendu d'une sémantique.
Mais il y a toujours eu une certaine connivence entre François Rastier et la praxématique. Je n'en veux pour preuve que le fait que nous ayons pu organiser ensemble, avec l'équipe praxiling, une série de conférences de François à Montpellier en 1993.

la *praxis* interne, le processus psychique concret qui en a construit le sens. Pour reprendre ce que nous venons d'exposer à propos de la sélection lexicale, la *psychomécanique* du choix d'un lexème le conçoit comme un parcours à travers des signifiants possibles qui inclurait progressivement des traits sémantiques de plus en plus spécificateurs, du fait des différences qu'ils potentialisent, tout en excluant les autres. Si un locuteur dit « vache », cela présuppose qu'il a parcouru l'ensemble du *taxème* des animaux de la ferme pour ne pas choisir, par exemple, le signifiant « animal » ni le signifiant « cheval » mais le signifiant « vache»[1]. Selon le point de vue *praxématique,* le fait que l'on puisse analyser le sémantisme de « vache » en traits successifs repose sur une mémoire des possibilités du parcours de visée ; il ne suppose pas, encore une fois, un *signifié* essentiel à ce *praxème*, ce qui permet de comprendre les variations de sens possibles[2].
De ce fait, on peut analyser que l'acte psychomécanique qui conduit à l'emploi de tel *lexème* plutôt que tel autre est une *opération herméneutique*, en ceci qu'il relève d'une *praxis d'interprétation du monde*. C'est d'ailleurs pourquoi, sans nier qu'il y a bien un réel distinct de nous, nous ne pourrons admettre que la *signification* d'un énoncé puisse correspondre de façon totale au référent de réalité qu'il prétend reproduire. Toute mise en référence, toute construction d'une *impression référentielle* est inévitablement abstraction et interprétation. C'est un produit culturel imaginé en autonomie, en distance par rapport à cette réalité. Dès que, petit bébé, nous nous trouvons en train de montrer du doigt quelque chose que nous désirons, ce geste symbolique que nous faisons interprète ce qu'il montre comme désirable. L'objet désigné n'est déjà donc plus une simple réalité. Il s'agit d'une symbolisation de celle-ci qui explicite le désir qu'elle suscite.

* Le X– (lire « X tiret »)

L'interprétation consciente de ces énoncés suppose deux étapes : nous percevons d'abord intuitivement un effet de sens, puis nous verbalisons celui-ci à partir des réglages présupposés de la *praxis* qui ont conduit à le produire.
Après l'avoir soumise à l'équipe *Praxiling* en 1992, j'avais présenté, il y a plus de 10 ans, en ce même lieu, lors d'un colloque de sémiotique[3], une formulation que je continue à défendre, même si elle ne semble pas encore avoir vraiment trouvé d'échos. Nous pouvons observer que tout énoncé perçu ne peut être identifié dans son contenu que par une glose qui l'analyse de la même façon que tout acte langagier encyclopédique analyse une référence au monde. Autrement dit, si nous nous trouvons devant le texte entier de la *Princesse de Clèves* de Madame de Lafayette, tout ce que nous pouvons dire de ce que nous en avons lu est globalement un X– (lire X tiret). C'est un X parce que c'est une inconnue, un *in posse* dont nous ne pouvons prendre conscience qu'en paraphrasant le contenu. Mais c'est un X qui est nécessairement lié à un tiret. Ce tiret est, en fait, une forme de pesée de nécessité, une loi de signifiance qui nous donne l'aptitude à percevoir si ce que nous disons de l'énoncé considéré est conforme ou pas à la signification de celui-ci.

* Les pesées herméneutiques

Au-delà de cette nécessité d'interpréter pour comprendre ce qu'on a compris, il faut constater aussi que cette interprétation repose, comme toutes les autres, sur un complexe d'évaluations intuitives. Pour juger si notre glose élucidant l'inconnue X est conforme au *texte*, que nous définissons comme l'équivalent du tiret, c'est à dire comme ce *poids* complexe de contraintes diverses qui nous conduisent à dire la signification de façon fidèle, nous recourrons à une autre évaluation tout aussi complexe, qui ne peut jamais relever d'une certitude objectivable. Nous posons donc alors toute une série de jugements qui sont du même type que les *jugements de grammaticalité* et *d'acceptabilité* supposés par la *grammaire générative*. Comme l'a montré François Rastier, ces jugements relèvent d'évaluations, je dirai de « *pesées* » purement esthétiques, aussi (peu) sûres que l'acte par lequel nous attribuons une bonne note à une

[1] Nous ne reprenons pas l'analyse classique en praxématique qui fait parcourir tout le vocabulaire scientifique pour faire descendre au praxème « vache » à partir du générique « animal » en passant par le praxème « ruminant », car l'emploi du mot « vache » est plus souvent lié à l'expérience de la ferme qu'à la connaissance zoologique des animaux. Cet autre parcours est bien sûr tout aussi possible. Pour comprendre toutes les potentialités de *signifiants praxèmes* qui pourraient être choisis à la place de « vache » il faut regarder qu'elles sont toutes les gloses possibles du *signifiant* « vache » dans l'énoncé considéré, autrement dit toutes les paraphrases qu'on peut en faire pour dire sa *signification*.
[2] C'est une originalité de la *praxématique* par rapport à la *psychomécanique* guillaumienne qui intègre sans broncher le *signe saussurien*.
[3] Voir P. Sadoulet, 1998.

dissertation qui nous est apparue très brillante, ou que le jugement qui nous conduit au choix d'une pièce de vêtement. Comme dit le proverbe : « Des goûts et des couleurs »…
Pourtant toutes les sciences du langage reposent empiriquement sur ce type de jugement, car elles ne peuvent produire que des gloses, elles-mêmes contrôlées par ces pesées de type esthétique : les « faits de langue » n'existent pas indépendamment de leurs interprétations et d'un jugement sur l'adéquation de ces interprétations.

* Les sciences du langage pour leur vertu heuristique

En outre, lorsqu'on fait de la linguistique, tout changement dans la description généralement admise n'a aucun intérêt, s'il ne permet pas de faire découvrir de nouveaux réglages du signifiant. Si l'on veut que les sciences du langage dépassent quelque peu les œillères de la vieille grammaire, il faut qu'elles confirment leur vertu heuristique. La description la plus utile sera celle qui mettra en conscience une interprétation évidente, un trait du X– , mais un trait que personne jusqu'ici ne savait décrire, c'est à dire expliciter par des mots. C'est le fond de notre métier.

* Les analyses de statistiques lexicales

Si nous laissons aux méthodes quantitatives la fonction d'évaluer ce qui serait un élément saillant, du fait de sa plus grande fréquence, nous risquons prendre en compte un fait différentiel qui ne pourrait relever, par lui-même, d'aucune *praxis* langagière. Du point de vue *praxématique*, une plus grande fréquence ne veut donc rien dire *a priori*, tant qu'elle n'est pas rapportée à une opération *psychomécanique* : elle décrit des phénomènes qui peuvent tout aussi bien relever du hasard, surtout si l'on mesure bien la petitesse des différences que la quantification peut identifier dans le domaine langagier[1].

Admettons quand même qu'il y ait bien, dans ces particularités retrouvées du signifiant textuel, des phénomènes quantitatifs incontestables en eux-mêmes. Quelle peut être leur pertinence ?
Car donner directement foi à ces données quantitatives, ce serait attribuer aux phénomènes statistiques la capacité de manifester des différences significatives sans l'intervention du moindre sujet humain.
C'est là tout le problème. Ces données quantitatives ne peuvent être pertinentes que si l'on retrouve l'activité de production de sens qui les ont produites. Si ces écarts quantitatifs peuvent être corrélés à un réglage plausible de la *praxis*, la démarche devient acceptable. Tous ces constats quantitatifs doivent donc être suspectés *a priori* comme illusoires, tant qu'ils ne sont pas rapportés, grâce à une interprétation, à une pratique culturelle de production de sens.
Nous constaterons ainsi que les analyses stylistiques ont souvent recours à l'observation de la fréquence particulière d'un thème, d'un sème ou de tel ou tel morphème grammatical dans l'extrait considéré. Mais ce constat est fait dans une séquence restreinte et en prenant en compte le relevé précis des occurrences et les effets de sens créés par leur récurrence. Rien n'interdit alors d'interpréter ces phénomènes du signifiant, pour peu qu'ils présupposent la *praxis énonciative* qui a convoqué plus souvent ces traits sémantiques, à ce moment de l'énoncé. L'écart de fréquence, qu'on peut identifier comme un phénomène de rythme, devient alors, dans ce cas, et seulement dans ce type de cas, une présomption d'isotopie[2].

* Vive les cas rares sources d'inventions heuristiques

Peu enclins donc à nous laisser séduire par des données quantitatives, nous préférerons jouer plutôt les pêcheurs de perles pour trouver l'occurrence qui nous montrera une production de sens

[1] Les classements fréquentiels, quels qu'il soient, montrent que la plupart des lemmes ne dépassent pas un effectif conséquent dans le corpus donné – le nombre d'occurrences de la plupart des lemmes constitue moins de 0,5% de la masse des occurrences. Les fonctions statistiques qui calculent comment le décompte des attestations d'un lemme dépasse son effectif attendu est une amplification qui dissimule que la valeur attendue est minime donc peu valable sur le plan statistique.

[2] Ce sont d'ailleurs ces phénomènes d'isotopie soit de sèmes soit de rythmes qui sont présentés par François Rastier dans le chapitre « Philologie numérique » d'*Arts et sciences du texte*. En fait, dans les exemples qu'il donne, il semble que les phénomènes quantitatifs ont servi d'indicateurs pour conduire le linguiste à s'interroger sur les réglages et poser une interprétation qui cherche à retrouver l'explication sémantique, le fait de praxis derrière le phénomène statistique.

particulièrement éclairante, celle qui proposera un renouvellement interprétatif, celle qui fera deviner d'autres inventions de langue[1].
Cette position qui peut paraître assez conservatrice, applique assez systématiquement, comme nous l'avons dit, l'attitude d'un Saint Thomas. Comme praticien d'une herméneutique philologique, nous préférerons examiner, autant que possible, chaque occurrence : nous ne croyons que ce que nous avons vu ou plutôt soupesé et interprété dans la recherche de la glose la plus pertinente et surtout la plus éclairante pour comprendre toute la richesse intersémiotique de la production de sens.

Conclusion : un tonneau sans fond ?

Faute de place, il faut renoncer à raconter en détail comment *Corputex* m'a permis de mener des études relevant des domaines différents des sciences du langage dont j'ai eu à m'occuper : la *morphosyntaxe*, la *lexicologie sémantique* et la *sémiostylistique littéraire* en lien très étroit avec la *sémantique interprétative*. Je ne mentionnerai rapidement qu'un exemple pour chacun de ces domaines.
Une double étude lexicale[2] sur l'adjectif du grec ancien « *axiologos* » et sur le français « *considérable* » repose sur des extraits analysés avec *Corputex*. Le logiciel ne m'a sûrement pas permis de poser le modèle descriptif que j'ai pu proposer, qui repose sur des spéculations sémiotiques et praxématiques, mais il m'a aidé à rassembler les données grecques et françaises. Prenant une posture qui voulait identifier systématiquement les contre-exemples, j'ai été amené à confirmer l'adéquation d'une proposition descriptive qui imaginait un *motif* premier pouvant expliquer la diversité des acceptions[3].
J'ai utilisé aussi *Corputex* pour établir le fonctionnement tensif de toutes les constructions consécutives dans le grec de Strabon et certains textes de romans français du XIXe siècle[4]. Si aucune des constructions corrélatives trouvées ne pouvait mettre en question l'analyse par un sème tensif, absolument évident dans ces cas, je me suis aperçu très vite, par une familiarisation due aux diverses relectures que nous impose l'usage du logiciel, qu'il existait aussi un poids créé par le cotexte précédent qui servait de base d'explication pour maintenir un sème tensif aux adverbes de liaison « si bien que » placés en début de phrase. Ce que représentait le quantificateur « si » dans la locution, c'était tout le poids argumentatif ou affectif de ce qui venait d'être dit.
Enfin *Corputex* m'a servi pour une étude intersémiotique de type thématique, dans un travail sur la conception de la beauté dans *Le Songe de Poliphile* attribué à Francesco Colonna et traduit en français par l'humaniste Jean Martin[5]. Menant une recherche de vocabulaire et une sélection par lecture directe du texte, j'ai pu extraire un sous-corpus que j'ai retravaillé pour choisir finalement les exemples insérés dans l'article. Il faut signaler ici que le travail d'insertion des extraits dans l'article s'avère très rapide, car *Corputex* formate directement les exemples qu'il suffit de recopier via le presse-papier de l'ordinateur.
Tous les textes que j'étudie en cours sont, maintenant, systématiquement importés dans une base *Corputex*. Cela me facilite grandement toutes les tâches matérielles nécessaires pour produire les documents pédagogiques, mais surtout, grâce à cet outil, je peux annoter mon corpus et en constituer une fiche de lecture réutilisable par la suite. J'en profite pour « *peser* » dans le texte le fonctionnement des « faits de langue » que j'enseigne, sans chercher à avoir des relevés complets.
Mais cette puissance accrue exige toujours des travaux supplémentaires pour que les résultats soient sûrs, donc exploitables… alors que les rythmes universitaires ne nous laissent plus vraiment le temps de préparer sérieusement nos cours. Il m'est arrivé souvent d'hésiter à me lancer dans la préparation de la version numérique d'une œuvre au programme, car cette préparation du corpus demande un gros travail, qu'il faille numériser le texte à partir de l'édition imprimée ou réaménager

[1] Voir G. Deleuze, 1993, *Critique et clinique*, Paris, Minuit. chapitre 1.
[2] Pour l'étude sur le grec voir P. Sadoulet, 2003.
[3] Sur la notion de *motif* voir P. Cadiot et Y.-M. Visetti, 2001, *Pour une théorie des formes sémantiques : motifs, profils, thèmes*, Paris, PUF.
[4] Voir P. Sadoulet, 2005.
[5] F. Colonna, J. Martin tr., 1546 éd. texte italien. 1499, *Le songe de Poliphile*, Paris, Pour Jacques Kerver aux deux Cochets, Rue St Jaques. Voir l'étude dans P. Sadoulet, 2003.

une version numérique pour retrouver la pagination voire la linéation de l'édition originale. Sans compter qu'il faut annoter l'œuvre ensuite etc…
Je me suis souvent demandé si, finalement, je n'avais pas conçu, comme je l'ai déjà dit, une application qui fonctionne comme un tonneau des Danaïdes et qui deviendrait, de ce fait, une vraie torture pour l'utilisateur. Car quand on utilise *Corputex*, rien n'est jamais fini, rien n'est définitif et il y a toujours un ouvrage à remettre sur le métier. Mais la richesse de perception sémantique que j'obtiens, je crois, dans mes études, laissent l'impression que l'obligation de relecture créée par *Corputex* permet d'aller beaucoup plus loin dans l'enrichissement de l'interprétation qu'on ne pourrait le faire avec de simples fichiers sur papier qui ne feraient jamais lire autant d'extraits, vu la longueur du travail de copie à la main qu'ils exigent. Ce travail à la main était, lui aussi, finalement, un tonneau sans fond, mais un tout petit tonneau. Avec *Corputex* nous gagnons nettement en puissance et en rapidité.

BIBLIOGRAPHIE

LAFONT, R. 1978. *Le travail et la langue*, Paris, Flammarion.
LAFONT, R. & GARDES-MADRAY, F. 1988. *Introduction à l'analyse textuelle réédition 88*, Montpellier, Langue et Praxis, Université Paul Valéry.
LAFONT, R. 1994. *Il y a quelqu'un : La parole et le corps*, Montpellier, Praxiling.
RASTIER, F. 1987. *Sémantique interprétative*, Paris, PUF pr. éd 1986, éd. 1991, éd., Paris, PUF, 1996.
RASTIER, F. 1989. *Sens et textualité*, Paris, Hachette.
RASTIER, F. 1994. Le problème du style pour une sémantique du texte, in P. Cahné & G. Molinié (éds.), *Qu'est-ce que le style ?*, Paris, PUF, pp. 263-28.
RASTIER, F. (éd.). 1996. *Textes et Sens*, Paris, Didier érudition.
RASTIER, F. 2001. *Arts et sciences du texte*, Paris, PUF.
SADOULET, P. 2005. *Corputex : logiciel d'analyse textuelle et de constitution de dossiers d'extraits* (base de données en 4 D), Saint-Étienne. Version 18.
Études de sémiotiques sur des livres d'artistes :
SADOULET, P. 1998. Rhétorique et épaisseur sémantique, in *Actes du colloque d'Albi (GDR de sémiotique) Sémantique et rhétorique* juillet 1995, Toulouse, Editions Universitaires du Sud, pp. 81-103.
Etudes de linguistiques et sémiotiques sur corpus :
SADOULET, P. 1980. *Le principe d'économie dans l'expression* (2 tomes), Thèse de troisième cycle (dir. Michel Casevitz), Université de Lyon II.
SADOULET, P. 1996. Un jeu original sur le signifié du Paon : étude de dix estampes de "Pavo, fragment sur le paon bleu", de François Righi, in M.-L. Honeste, R. Sauter (éds.), *Animots*, Université de St Etienne, CIEREC, pp. 159-179.
SADOULET, P. 1998. Du global au local : effets de sens et corrections d'auteur. Travail à partir du manuscrit de « À chaque pas prenant congé », in J.-Y. Debreuille (dir.), *Un poète dans la classe, Jean-Vincent Verdonnet*, Lyon, PUL, pp. 145-164.
Etudes publiée menées à l'aide du logiciel Corputex sur corpus numérisés :
SADOULET, P. 2003. Axiólogos chez Strabon. Essai d'apport sémiotique à l'étude d'une polysémie lexicale, in S. Remi-Giraud & L. Panier (dir.), *La polysémie ou l'empire des sens. Lexique, discours, représentations*, Lyon, PUL, p. 65 ss.
SADOULET, P. 2003. L'émotion esthétique et sa représentation verbale dans le Songe de Poliphile (livre I), in C. Ziberberg & F. Parouty (dir.), *Sémiotique et esthétique*, Limoges, PULIM.
SADOULET, P. 2002. Le corps du voyageur dans la description géographique /Traveller's body in geographic description, in *Symposium* organisé par le bureau de l'association internationale de sémiotique, Université de Lyon II septembre 2002.
SADOULET, P. 2005. Le morphème intensif "hôste" dans la géographie de Strabon : entre corrélation et coordination, Communication au colloque "Subordination et corrélation", Bordeaux 26 et 27 septembre 2002, Saint-Étienne.