

J.-P. BENZÉCRI

## **Philologie classique et taxinomie des textes**

*Les cahiers de l'analyse des données*, tome 18, n° 3 (1993),  
p. 377-382

[http://www.numdam.org/item?id=CAD\\_1993\\_\\_18\\_3\\_377\\_0](http://www.numdam.org/item?id=CAD_1993__18_3_377_0)

© Les cahiers de l'analyse des données, Dunod, 1993, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme  
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

## PHILOLOGIE CLASSIQUE ET TAXINOMIE DES TEXTES

### [TAX. TEXTES]

*J.-P. BENZÉCRI*

Dans le monde méditerranéen, la philologie classique a une tradition plus de deux fois millénaire: déjà, avant notre ère, à Alexandrie, on s'appliquait à éditer les grandes œuvres pour en fixer le texte, en corrigeant les fautes des copistes et écartant les contaminations et les mélanges. L'attribution de certains textes à des auteurs illustres a souvent été contestée; et, à défaut de témoignages historiques et biographiques suffisants, la charge de la preuve revenait à l'analyse du texte mis en cause et à sa comparaison avec d'autres textes.

Le terme même de stylométrie a été introduit par LESNIEWSKI, à la fin du XIX-ème siècle; depuis lors, de nombreux philologues admettent que des méthodes mathématiques peuvent aider à interpréter les dénombrements que leur science a toujours pratiqués; et l'avènement des ordinateurs a donné aux méthodes les plus classiques une puissance et une ampleur dont peu songeraient à faire fi.

Nous voulons ici, une fois de plus, défendre la thèse que toute étude de philologie ou de stylistique peut bénéficier de l'analyse de données multidimensionnelles recueillies dans un champ aussi vaste que possible.

Au §1, nous prenons pour exemple la longueur des phrases; indice stylistique qui requiert une définition préalable de la phrase; d'où un problème linguistique; qu'on se dispensera, d'abord, de résoudre en caractérisant la rhétorique d'un texte par un ensemble de données directement accessibles.

Au §2, on oppose l'étude générique, qui vise à définir les dimensions d'un espace comprenant tous les textes; aux études spécifiques, qui ne traitent qu'un ensemble particulier d'ouvrages sans demander à un contexte étendu d'en révéler les tenants et les aboutissants.

Au §3, toute bibliothèque est considérée comme une base de données; ou, plutôt comme un labyrinthe, dont seule l'analyse multidimensionnelle offre le fil

d'Ariane; pour un parcours qui n'a de charme que dans les jardins de la philologie classique (§4).

### 1 Longueur des phrases et complexité syntaxique

Entre autres caractères d'un style, il est d'usage de considérer la longueur des phrases. Cependant, pour la linguistique classique, la notion même de phrase n'est pas hors de conteste.

En tenant compte de la dépendance syntaxique on peut, dans une certaine mesure, constituer des suites de mots qui ne doivent certainement pas être partagées entre plusieurs phrases: ainsi une forme verbale est solidaire de son sujet; une proposition circonstancielle rentre dans la même phrase que la proposition principale à laquelle elle apporte des précisions quant au temps ou au lieu.

Mais la coordination, sous sa forme la plus simple, incite à considérer comme solidaires deux propositions dont chacune est cependant complète en elle-même. Ici, le critère de la dépendance devient aléatoire: ainsi, la suite de mots (ne disons pas *phrase!*):

elle tangué, mais ne coule pas ;

ne doit pas être coupée en deux phrases; car le pronom 'elle', accolé comme sujet au verbe 'tangué' manquera au verbe 'coule'; mais on pourrait scinder:

elle tangué, elle ne coule pas;

en partageant par un point plutôt que par une virgule:

Elle tangué. Elle ne coule pas.

il est, du moins, certain que, dans ce cas, la dépendance qu'il y a entre les deux membres n'est marquée par aucune réaction syntaxique proprement dite.

Encore n'avons-nous considéré jusqu'ici la coordination que comme juxtaposition de deux parties, dont chacun pourrait constituer une phrase complète (ou peu mutilée); avec, éventuellement, entre les membres un mot de conjonction. Autre exemple:

Pierre va à l'Opéra pour régaler son oreille; Jacques, pour le plaisir des yeux.

coordination réduite à un simple effet de parallélisme; qui, à lui seul, dispense de répéter le syntagme verbal: 'va à l'Opéra'. Peut-on faire de 'Jacques', un sujet de 'va', au même titre que l'est 'Pierre' ? ; ou encore de 'pour le plaisir des yeux', un circonstant de 'Jacques' ? La stricte syntaxe y répugne...

Et que faire des paroles rapportées, des discours directs ou indirects; des phrases nominales, exprimant un rapport principal entre des termes qui ne sont

pas liés par un verbe; toutes difficultés qu'on s'est appliqué à accumuler dans la présentation de l'exemple précédent!

Difficultés insolubles, à moins qu'on ne regarde le texte, non comme un ensemble actuel de phrases distinctement exprimées, mais comme un ensemble potentiel de phrases dont certaines ne sont que suggérées.

Or quel que soit, pour nous, l'agrément d'une telle perspective stylistique indéfinie, on peut, par certains dénombrements, recenser des faits de structure qui caractérisent l'ampleur syntaxique de la composition d'un discours sans qu'on ait, au préalable, tranché celui-ci en phrases.

D'abord, sans mise en forme du texte, dénombrement de mots outil, principalement des divers types de conjonctions, ainsi que des relatifs; ensuite, si l'on a étiqueté les mots, dénombrements des formes verbales par mode; et des noms par cas; ou, dans les langues sans flexion nominale, par ce qui peut en tenir lieu: noms sujets de verbes, noms compléments directs, noms régis par une préposition. On reprend ici la notion de rection, mais sans lui demander un découpage; lequel n'en résulte pas toujours univoquement.

## 2 L'étude générique des textes

Il nous paraît qu'il y a, entre la philologie classique et l'analyse des données linguistiques, telle que nous la pratiquons, une différence de point de vue qu'il convient de préciser afin de dissiper des malentendus. Peut-être, est-il permis de caractériser, d'abord, cette différence en termes juridiques: pour le philologue, le texte est un corps certain, pour nous, c'est un chose de genre.

Nous confessons qu'il y a de l'irrévérence à parler des dialogues de PLATON comme d'un ensemble non délimité de textes, occupant une aire stylistique où des événements majeurs ou mineurs ont fait lever le Phédon ou le Lachès...

Mais, d'une part, ce que l'on sait de la création littéraire atteste que les plus grandes œuvres sont comme des éclairs se propageant suivant une ligne contingente entre le haut potentiel d'un esprit depuis longtemps chargé dans les hauteurs et la terre qu'il vient brûler; et, d'autre part, les théories et les théorèmes qui sont les chefs-d'œuvre des mathématiciens, échappent d'autant plus à leur auteur, que la certitude des découvertes leur assigne, au sein de la science, une place nécessaire.

La lecture d'ARCHIMÈDE, de LAPLACE ou de POINCARÉ apporte au mathématicien, avec le plaisir ineffable d'approcher ces grands hommes, des leçons irremplaçables sur ce qu'une simple remarque peut renfermer potentiellement de vérités que nous croyions dépendre d'une exposition ordonnée qui n'est venue qu'ensuite. Mais, finalement, des regards par lesquels deux esprits aussi différents que NEWTON et LEIBNITZ ont discerné les

linéaments du calcul différentiel, il ne survit, dans la mathématique que tous pratiquent, que des nuances dans le symbolisme imparfait des notations.

C'est la gloire des textes philosophiques que les générations se les transmettent, dans leur langue originale. Cette gloire en marque aussi la fragilité: si l'on doit relire HEGEL, n'est-ce pas parce qu'il faut, en quelque manière, mettre sa tête où il avait la sienne, afin de voir, comme par ses yeux, ce qu'il nomme Geist? Objets aperçus, comme par des alignements entre plusieurs plans; mais non objets solides, dont quelques vues permettent de tracer une épure exhaustive; d'où se déduit tout ce qu'on peut voir, de quelque point de vue et sous quelque angle que ce soit.

Ainsi, au géomètre, l'ensemble des textes apparaît-il comme un nuage de points dans un espace dont il cherche à préciser les dimensions en y traçant des axes; tandis que, pour le philologue, chaque texte est un objet distinct qu'il comparera à d'autres objets semblables; sans prétendre édifier une taxinomie générale analogue à celle où, depuis plusieurs siècles, on s'applique à distribuer l'ensemble des espèces animales ou végétales.

### 3 Le catalogue actif d'une très grande bibliothèque

Cependant, le progrès rapide des techniques d'enregistrement magnétique ou optique, requiert, de façon urgente, qu'on élabore une telle taxinomie. Une disquette magnétique, du format commun utilisé pour les micro-ordinateurs, peut renfermer, dans un faible volume, le texte d'un gros livre. Des cartouches optiques, seulement deux fois plus épaisses, contiennent le texte de quelque cent livres. Il existe déjà des supports dix fois plus denses.

Ainsi, on peut se représenter une bibliothèque d'un million de volumes tenant à l'aise dans une petite pièce; et le précieux contenu d'une très grande bibliothèque, reproduit à quelques dizaines ou centaines d'exemplaires pour être physiquement présent partout. La seule difficulté étant d'offrir, aux lecteurs intéressés, un canal d'accès aux textes; et d'abord un catalogue actif qui leur suggère des titres répondant à leurs demandes.

De ce point de vue, l'accès aux textes d'une bibliothèque peut être assimilé à l'accès au contenu d'une base de données: sans une élaboration linguistique préalable fondée sur l'analyse statistique multidimensionnelle, les cribles logiques qui viennent à l'esprit sont d'un si médiocre rendement que l'abondance des informations accable l'utilisateur plus qu'elle ne le sert.

Aux recherches taxinomiques, les textes classiques offrent une matière particulièrement agréable. Ces textes sont nombreux; mais le corpus en est à peu près clos; et il a le mérite d'être composé d'une majorité d'ouvrages dignes de retenir l'attention; en sorte que, pour le nouveau philologue, statisticien et informaticien, des tâches, en elles-mêmes fastidieuses, revêtent le charme qu'on trouve à la fréquentation des grands esprits.

#### 4 Utilité d'une taxinomie générale des textes pour la philologie classique

Nous sommes d'ailleurs convaincu que, faute d'une taxinomie générale, la philologie classique a peu de chances de répondre aux questions particulières mêmes qu'elle se pose.

Comment dire que tel texte, attribué à un seul auteur et reçu sous un titre unique, doit être découpé en plusieurs parties, dont chacune constitue une œuvre distincte; alors qu'on n'a pas une vue d'ensemble des productions de cet auteur; une appréciation précise et, s'il se peut, mathématique, de la propension qu'il montre à changer de genre, non seulement en passant d'une œuvre à une autre, mais au sein d'une même œuvre, et comme sans cesser un instant d'écrire.

Et, même sans mettre en doute les attributions des œuvres aux auteurs dont elles portent le nom, la diversité des productions d'un seul homme ne doit-elle pas être comprise par référence à l'étendue des œuvres qu'il a connues; dont il a assimilé une partie pour en pratiquer le genre ou remodeler celui-ci; qu'il a quittées, peut-être, pour innover et ouvrir à la postérité des voies; encombrées depuis d'imitations dont l'homogénéité témoigne la puissance du premier modèle.

Peut-on, sans se faire illusion, recourir à la statistique quand le sens du philologue est en défaut; alors qu'à des problèmes simples, qu'on ne doute pas de résoudre selon l'intuition, on ne cherche même pas à donner un énoncé, puis une solution, mathématiques? On calculerait pour apprécier l'hétérogénéité d'un corpus, avant de savoir, par le calcul, distinguer entre plaidoyer, récit historique et discours philosophique? N'est-ce pas comme de prétendre le *plus*, avant d'avoir tenté le *moins*?

Il est nécessaire de prendre pour objet des corpus de textes de plus en plus étendus, de plus en plus hétérogènes.

Il ne nous échappe pas que, de l'analyse de tels corpus, on tirera d'abord des conclusions de forme banale; et qui, sans doute, ne seront même pas toutes vraies.

Mais l'expérience montre que, des multiples énoncés de forme banale qu'on pourrait conjecturer par intuition et appuyer par des arguments du sens commun, l'analyse des données élimine la plupart; et n'en suggère qu'un petit nombre dignes d'être reçus comme des hypothèses et mis à l'épreuve d'analyses ultérieures, fondées sur la collecte de données complémentaires appropriées.

Ce qu'on nomme en français *Ordinateur* fut d'abord appelé *Machine à calculer*; *Computer* a toujours cours en anglais. Mais l'on n'a jamais cessé de dire avec amphase: *Cerveau Électronique*.

Certes, la première fonction dévolue à ces outils fut d'effectuer, à une vitesse qui parut, d'emblée, prodigieuse, les opérations mêmes qu'aurait effectuées un calculateur humain, si le temps lui en avait été donné.

Cependant, taille de mémoire et rythme de calcul deviennent tels, qu'il semblerait parfois qu'un système d'équations aux dérivées partielles confié à l'ordinateur ne fût pas véritablement intégré, comme il pourrait l'être par le génie d'un géomètre, mais traduit en un processus physique équivalent qui se déroule dans la machine jusqu'à l'affichage du résultat. L'agilité de l'outil frustrerait l'homme des raccourcis sublimes de la pensée.

D'aucuns attendent de la volubilité du traitement des informations non numériques, une *intelligence artificielle*.

L'analyse multidimensionnelle prend le même outil dans un autre sens, qui est celui du terme italien d'*Elaboratore*. Aux informations pesamment accumulées, conférer une légèreté qui les rende propres aux considérations de la pensée.