GILLES MALONEY

LE PROJET HIPPO: RECHERCHE EN AUTOMATISATION DE DONNÉES STYLOMÉTRIQUES À PARTIR DE L'OEUVRE D'HIPPOCRATE

Le Projet Hippo est une recherche sur l'oeuvre d'Hippocrate menée depuis trois ans par une équipe de quatre professeurs, d'assistants, d'un analyste et d'un programmeur, ainsi que d'une société commerciale de consultants en statistique. Poursuivi à l'université Laval de Québec, ce travail est subventionné par l'université elle-même, par le Gouvernement du Québec et par le Conseil des Arts du Canada.

Les hellénistes qui abordent la lecture du *Corpus hippocraticum* se trouvent en présence d'une multitude de traités divers, mais ils constatent bientôt que cet ensemble manque d'homogénéité et que, parmi ces quelques 70 oeuvres, à peine une vingtaine portent la marque de la médecine positive prônée par Hippocrate. Les philologues ont pu distinguer dans ce groupe des écrits relevant de trois écoles de pensée différente, mais n'ont pas encore réussi à attribuer avec une certitude suffisante un groupe précis d'oeuvres appartenant à Hippocrate et encore moins à attribuer d'autres traités à d'autres noms.

Le Projet Hippo vise donc à analyser et à classifier tous les traités du *Corpus* par des études quantitatives de style appuyant leurs conclusions sur la statistique et exécutées par ordinateur; nous espérons en arriver à une hypothèse générale capable de grouper les traités en sous-ensembles présentant des affinités suffisamment précises pour laisser croire qu'ils se rattachent à des auteurs différents.

Il faut souligner que nous n'avons aucunement la prétention d'attribuer telle ou telle oeuvre à tel personnage historique précis, même si le Projet, en tant que recherche littéraire et historique, a pour but ultime de classifier les pièces du *Corpus*. Nous ne pourrons pas non plus attribuer des oeuvres à Hippocrate lui-même, faute de repères extérieurs au *Corpus*. Beaucoup plus modestement, nous essayons de mettre en lumière un certain nombre de faits stylométriques, en accordant à des tests statistiques une importance à laquelle la philologie classique n'est

pas habituée lorsqu'il s'agit de comparer des données quantitatives avec des résultats d'analyse de contenu.

Nous avons estimé que le *Corpus* contient environ 350.000 mots soit 2.500.000 caractères, et qu'il faudrait ajouter un autre total de 4.200.000 caractères pour la codification grammaticale et les lexiques, sans compter la programmation. Nous prévoyons donc avoir besoin d'emmagasiner environ 7.000.000 de caractères. Pour le moment nous utilisons environ 700 K d'espace. Etant donné l'ampleur relative de l'entreprise et les hasards qui nous attendaient à coup sûr, nous avons d'abord décidé de tenter un Pré-projet portant sur trois traités, qui nous permettraient de sonder l'à-propos de cette recherche du point de vue de la philologie et d'en évaluer les chances de réalisation technique dans notre université. Cette phase a été parcourue en deux ans avec des résultats qui nous ont encouragé à réaliser l'ensemble du Projet, dont nous prévoyons une fin d'étape pour septembre 1975.

Le Projet Hippo étant conçu comme un travail expérimental dans tout son ensemble, et puisque nous ne sommes aucunement certains d'apporter une solution au problème d'authenticité que nous posons à propos du *Corpus*, nous prenons soin d'obtenir, tout en avançant, des résultats concrets qui seront valables quelle que puisse être l'issue de notre recherche.

Nous avons donc mis au point un système d'entrée-sortie sur APL qui nous permet de travailler directement en grec, en utilisant majuscules, minuscules, accentuation et ponctuation, par l'intermédiaire d'un terminal Selectric 2741. Nous en sommes en effet venus à la conclusion que l'APL était actuellement le système idéal pour produire des textes; sur disque, et de là sur ruban magnétique, vu sa grande flexibilité dans les manoeuvres de modifications de données. Dans le travail d'entrée, de correction ou de sortie nous utilisons un format de 4 ou 5 K, que nous appelons « page », reproduisant 40 lignes ou moins de texte qui ont la même longueur et la même disposition que dans le livre dont elles proviennent. Au moment de l'entrée nous pouvons corriger un mot mal écrit aussi souvent qu'on veut, modifier la ligne, supprimer celle-ci s'il le faut, grâce à la possibilité de marche arrière que les consoles 2741 permettent. A la sortie nous pouvons obtenir par une seule commande soit une ligne précise d'un traité, soit une page, soit plusieurs pages en séquence ou non, soit un traité au complet soit l'ensemble des traités. Lorsqu'il faut modifier le texte par suite d'erreurs constatées, il est possible d'ajouter, n'importe où et directement, des espaces ou des caractères ou même d'insérer des lignes oubliées, ou au contraire

de supprimer des espaces, des caractères, des mots, des lignes et même des pages, pour ainsi dire instantanément. Nous pouvons dire actuellement sans grande modestie que, pour ce qui regarde la préparation fidèle à l'original d'un texte grec sur ruban magnétique, nous nous sommes payés toutes les fantaisies dont nous avions besoin.

Cette fonction devient alors une addition aux ressources de notre Centre de Traitement de l'Information (CTI), qui, ne possédant que deux ordinateurs a jugé bon d'en distinguer les juridictions: un IBM 370-155 exécute les travaux en OS tandis qu'un IBM 370-145 est consacré uniquement à l'APL: quelques 105 consoles lui sont reliées par lignes téléphoniques et actuellement il en gère plus de 65 simultanément. En partageant ainsi le temps l'usager gagne en économie d'argent ce qu'il perd en espace de mémoire pour lui disponible. C'est justement pour cette raison que nous traitons nos propres données en OS par langage PL/1.

Ainsi nous produisons des documents en APL sur disques et lorsque nous les jugeons sans faute nous en faisons une copie sur ruban qui est alors traité par l'autre ordinateur. Quant à l'approche en APL, elle est suffisamment simple pour que même un non-initié au maniement des ordinateurs puisse apprendre à la maîtriser en quelques heures. Autre avantage: en plus de la console placée dans le local qui est attribué au Projet, nous pouvons utiliser toutes celles qui sont disséminées un peu partout dans l'université. Enfin, l'APL permet de travailler à plusieurs en même temps au Projet, tant sur les textes que sur la programmation.

En ce qui concerne l'aspect productif du Projet, on peut dire que les résultats obtenus se rangent en trois classes. D'abord, nous enregistrons des textes sur rubans: nous en sommes actuellement au septième traité du *Corpus*. De ce point de vue notre travail n'a pas été inutile, même pour les nombreux philologues intéressés à l'oeuvre d'Hippocrate, puisqu'ils peuvent disposer sur un ruban magnétique compatible avec la plupart des machines des textes que nous croyons sans fautes et des lexiques qui dépassent de loin par leur variété l'index d'Hippocrate qu'un groupe d'hellénistes allemands est en train de confectionner à Hambourg par des procédés traditionnels.

En effet nous obtenons aussi ce qui constitue généralement des routines dans les blocs de programmation: à savoir des lexiques. Nous imprimons pour chaque traité un lexique alphabétique, un autre selon les fréquences d'emploi, un selon les catégories morphologiques conventionnelles et un dernier qui classe les *hapax*. Une fois l'oeuvre complète enregistrée sur ruban nous produirons quatre index généraux du *Corpus*.

En troisième lieu, nous entretenons à propos des données numériques une crainte parmi nous: celle d'être inondés par une masse de chiffres que nous ne pourrons jamais analyser nous-mêmes et qui n'intéresseront personne. Aussi nous en tenons-nous à la connaissance des faits suivants:

1) le détail séquentiel du nombre de mots par phrase, accompagné du numéro d'ordre de la phrase et de la référence;

2) le nombre total de mots par traité;

3) la longueur moyenne des mots;

4) la distribution des fréquences de mots par phrase;

5) la comparaison de la position dans la phrase du premier substantif et du premier verbe à forme personnelle ou de tout autre paire de catégories morphologiques;

6) la répartition des mots selon les catégories morphologiques;

7) enfin, une double matrice de 22×22 compilant les voisinages primaires et secondaires des mots selon leur catégorie. Celles-ci indiquent combien de fois un substantif est suivi d'un autre substantif, combien de fois par un adjectif, par une conjonction, et le reste; puis, combien de fois un adjectif est suivi d'un substantif, d'un infinitif, et ainsi de suite. De plus, combien de fois on trouve un mot entre deux substantifs, entre un substantif et un adjectif, entre un participe et un infinitif, et le reste. Cette matrice est très riche en renseignements; elle fournit par exemple la répartition des mots en début et en fin de phrase.

Le traitement statistique de ces données par rapport au problème que nous définissons à propos du *Corpus* s'est révélé au cours du Pré-Projet trop complexe et hasardeux pour nos forces. Nous avons donc recours aux services de professionnels dans le domaine, la Société de Mathématiques Appliquées (SMA) de Montréal, compagnie de consultants qui est unique au Québec. Comme nous n'avons pas affaire à un véritable problème d'authenticité mais à une discussion sur des groupements hypothétiques, nous avons résolu de n'appliquer pour le moment à nos données que deux types de tests: l'un portant sur l'entropie des oeuvres, l'autre étant le test d'analyse factorielle dit ANAFACO, fourni par Metra International. Jusqu'à maintenant les résultats de ces essais coïncident avec les hypothèses communément admises par les philologues; nous continuerons donc à les appliquer dans les mêmes conditions pour tous les traités, quitte à ajouter en cours de route quelques tentatives différentes. Là encore nous essayons de garder un juste milieu entre l'emploi systématique du χ^2, par exemple, et des tests si complexes

qu'ils risquent de ne plus être compatibles avec l'aspect littéraire d'un texte.

Pour revenir au traitement du texte, il faut dire que les lexiques que nous obtenons ne sont pas lemmatisés, et que l'analyse morphologique que l'ordinateur en connaît est très élémentaire.

En effet, dans les trois premiers traités que nous avons enregistrés, nous faisons suivre chaque mot d'une cote allant de 1 à 22 et correspondant à une catégorie morphologique. De là nous avons tiré un lexique alphabétique contenant la fréquence, la forme, son code morphologique et ses références. Nous avons reporté sur disque le lexique du premier traité d'Hippocrate (*L'ancienne médecine*), amputé de ses références, lui attribuant une cinquantaine de pages en APL. Il est alors devenu malléable comme les autres textes: nous demandons donc une ligne qui imprime disons, «5 ἀγαθούς», c'est-à-dire: adjectif ἀγαθούς. Comme le format des lignes est variable, nous écrivons à une certaine distance à la suite du mot: ἀγαθός. Puis nous faisons suivre une analyse complète qui détaille le « 5 » qui caractérisait le mot jusqu'à ce moment: elle consiste en 10 éléments numériques qui indiquent si le mot est un lemme ou non, puis s'il est un homonyme ou non, les 8 autres éléments codifiant l'analyse morphologique détaillée de la forme rencontrée. Bien entendu, cette analyse manuelle peut se faire en étapes successives. C'est à construire ce lexique au complet que nous travaillons actuellement, les sections qui en sont terminées nous permettant de roder la programmation subséquente.

Car ce lexique va constituer la base de l'analyse semi-automatique de tout le *Corpus*: par comparaison avec lui le lexique non-lemmatisé du traité *Des Airs, des eaux et des lieux* recevra l'analyse complète de tous les mots non-homonymes. Nous complèterons sur console APL l'analyse des formes nouvelles et celle des homonymes, et par un fondu des lexiques 1 et 2 nous ferons un petit index général lemmatisé d'Hippocrate; celui-ci sera comparé au lexique non-lemmatisé du traité des *Epidémies* et ainsi de suite. Il n'est pas difficile à ce moment-là de bâtir un index définitif basé sur les lemmes en faisant reporter automatiquement chaque forme à la suite de son lemme et d'obtenir ainsi un lexique lemmatisé pour chaque traité et de même un lexique général du *Corpus*.

Une fois ce bloc de programmation terminé, il nous faudra transporter ces données sur l'ordinateur 370-155 et là commencera la dernière étape du Projet, qui consistera à rendre possible le traitement des

données numériques et lexicologiques en TSO, c'est-à-dire en direct, par l'intermédiaire là aussi d'une console.

Comme les personnes surtout engagées dans le Projet Hippo font partie d'un département des littératures, nous espérons que cette recherche continuera à avoir des effets utiles pour nos collègues: il n'est pas difficile en effet de passer d'un caractère d'imprimerie à l'autre sur console 2741, et tout ce que nous faisons devient immédiatement applicable à des textes en d'autres langues. En même temps les études littéraires automatisées feront chez nous des progrès, en s'installant parmi les activités du Centre de Traitement de l'Information, à côté de la gérance de la bibliothèque, de la consultation par télé-information des Statuts du Québec offerte aux juristes, de la Banque d'Information en Bibliographie Patristique, et de quelques recherches importantes en linguistique.

C'est ainsi qu'Hippocrate serait sans doute étonné de voir que ses carnets aident maintenant un groupe d'universitaires à demeurer dans leur époque.