PERSPECTIVES ET PROBLEMES D'UTILISATION DU *THESAURUS LINGUAE GRAECAE**

Le département de Philologie classique et le Centre Informatique de Philosophie et Lettres de l'Université de Liège ont acquis, comme de nombreux instituts de langues anciennes les fichiers informatiques du *Thesaurus Linguae Graecae* (TLG).

Une fois en possession d'une masse de données aussi importante, nous avons souhaité l'exploiter dans les meilleurs délais. Encore devions-nous, pour y parvenir, connaître d'abord les informations figurant sur les bandes et développer ensuite des procédures d'utilisation adaptées au contenu. Notre démarche, nos observations, intéresseront et aideront peut-être ceux qui, comme nous, ont accès aux données du TLG. C'est la raison pour laquelle nous en proposons ici une synthèse.

C'est en 1972 que fut décidée, sous l'impulsion du professeur Th. F. Brunner de l'Université de Californie (Irvine), la constitution d'une banque de données informatisée rassemblant les œuvres littéraires grecques antérieures à l'an 600 de notre ère.

A la suite de cette décision, le projet californien, connu sous le nom de *Thesaurus Linguae Graecae*, n'a cessé de se développer pour aboutir en 1985 à un fichier qui regroupe 2884 auteurs, environ 8000 œuvres et quelque 57 millions de mots. A ce moment, bien que la tâche ne fût point achevée, les responsables du *Thesaurus* commencèrent à mettre à la disposition de la communauté scientifique internationale la masse documentaire déjà rassemblée.

Au point de vue matériel, le *Thesaurus Linguae Graecae* se présente sur deux supports différents, soit sur bandes magnétiques dans sa totalité, soit sur CD-ROM[1] pour un peu plus de deux tiers des

* Dans toutes les recherches de philologie classique, et aussi pour celles qui regardent de plus près les choix de cette revue, la disponibilité d'un instrument tel que le TLG est sans aucun doute d'une importance primordiale, mais les problèmes que son emploi comporte sont tels que, jusqu'à présent, il a été sous-utilisé. Pour cette raison, la rédaction de *Lexis* a demandé à M. Joseph Denooz de consacrer un article aux expériences qu'il mène sur le TLG, article qu'elle est très heureuse de présenter à présent à ses lecteurs.

[1] CD-ROM est l'acronyme de Compact Disk Read Only Memory. Il s'agit d'un disque compact sur lequel on peut conserver 600 millions de caractères. Une fois créé, le CD-ROM peut-être exploité par un matériel spécifique, mais il est impossible de le modifier. Dès lors, lorsque l'on utilise ce support, on n'a pas la possibilité de

textes (environ 41 millions de mots sur les 57 millions).

Dans sa version 'disque compact', le TLG est disponible sur deux machines, le système IBYCUS et l'ordinateur APPLE MACINTOSH fonctionnant sous le contrôle du logiciel PANDORA.

IBYCUS[2] est un matériel spécialement conçu pour le traitement du grec. Il a été construit par la firme Hewlett Packard à partir d'études faites par David Packard. Ce matériel semble poser actuellement des problèmes de maintenance et pourrait bien disparaître à brève échéance[3].

PANDORA, quant à lui, est probablement promis à un bel avenir: il s'agit d'un logiciel convivial aisément accessible à un non-spécialiste de l'informatique et fonctionnant sur le micro-ordinateur MACINTOSH. Le succès de PANDORA dépend toutefois de la mise au point de procédures de consultation complexes, souples et rapides qui sont en cours de développement.

Certains centres de recherches tentent de réaliser leurs propres logiciels d'exploitation du CD-ROM. C'est le cas, par exemple, de l'École normale supérieure de Pise où une équipe élabore, sous la direction de G. Nenci, le système SNS-List adapté aux données du TLG[4].

En raison des limites actuelles du CD-Rom et du manque de logiciels d'utilisation[5], les fichiers du TLG enregistrés sur bandes magnétiques présentent des possibilités plus étendues d'exploitation. Celle-ci se fait à l'aide d'un gros ordinateur et avec des programmes

corriger les inévitables erreurs que contiennent des fichiers de grande taille.

[2] Le lecteur désireux de mieux connaître IBYCUS se reportera à l'article d'A. Bozzi, *Archivio TLG e IBYCUS SC: nuove tecnologie per gli studi classici*, Materiali e discussioni per l'analisi dei testi classici, 17 (1986), 175-84.

[3] Au Congrès *Sciences historiques, sciences du passé et nouvelles technologies d'information* (Lille 16-18 mars 1989), J.F. Oates déclarait: «Quite frankly, the system is a dinosaur; it is no longer supported by the Ibycus company and it is now difficult to obtain replacement parts».

[4] D. Bouvier, *Lire et utiliser le "Thesaurus Linguae Graecae" avec MACINTOSH: présentation d'un projet de la "Scuola Normale Superiore" de Pise*, dans *Communications présentées au Colloque "Epigraphie et informatique" (Lausanne, 26-27 mai 1989)*, Lausanne 1989, 145-51.

[5] Il convient de noter que le CD-Rom, malgré les inconvénients qu'il présente, connaît un succès croissant: dans la dernière parution du bulletin d'informations du TLG (juillet 1989), Th.F. Brunner signale que près de quatre cents exemplaires du *TLG CD-Rom* sont actuellement diffusés.

que l'utilisateur élabore lui-même.

Avant de décrire le contenu des bandes du TLG, nous définirons l'expression «enregistrer un texte sur supports informatiques» et nous rappellerons quels sont les moyens physiques grâce auxquels sont conservées les données que l'on souhaite communiquer à la machine et quelle forme peuvent avoir ces données.

«Enregistrer une œuvre sur support informatique» consiste simplement à transcrire un texte sur un support physique et dans une forme logique que l'ordinateur pourra traiter. Ainsi, les spécialistes de la linguistique par ordinateur accomplissent-ils la même tâche que les moines copistes du Moyen Age: ils assurent la transmission et la diffusion des œuvres classiques. Notons d'ailleurs que certains d'entre eux commettent les mêmes erreurs que leurs prédécesseurs.

Prenant pour base l'édition d'une œuvre, si possible la meilleure, on la recopie, ponctuation comprise, sur une bande magnétique ou sur un disque. On dispose alors d'un fichier où se succèdent les mots qui appartiennent à un même ensemble. En ce sens, un fichier informatisé des *Géorgiques* est une séquence de données représentant la totalité du texte de Virgile.

Dans un fichier, l'unité logique d'information, appelée enregistrement, pourra être soit un vers, soit un mot, selon le choix de celui qui définit la structure de la banque de données.

A l'origine de l'informatique, les informations numériques ou linguistiques étaient perforées sur des cartes mécanographiques; celles-ci étaient encore largement employées à la fin des années soixante et même au début des années soixante-dix. On verra ensuite se répandre les bandes et les disques magnétiques qui restent à l'heure actuelle les moyens les plus communs de conservation des informations, tandis que se constituent maintenant de véritables bibliothèques électroniques sur vidéodisques, sur CD-Rom, etc.

Le fait que, pendant plusieurs décennies, la carte a été le support le plus utilisé, le plus fiable et le moins coûteux, a une influence déterminante et durable sur la manière dont les linguistes et les philologues ont conçu l'enregistrement de textes, c'est la raison pour laquelle nous décrirons brièvement les caractéristiques techniques du carton perforé.

Une carte mécanographique se divise verticalement en 80 colonnes. Chaque colonne peut recevoir des perforations qui symbolisent un caractère d'imprimerie, lettre, chiffre ou signe de ponctuation. Une

carte ne peut dès lors contenir une suite de caractères dépassant quatre-vingts signes.

Cette limite quantitative a eu une conséquence sur le plan linguistique: pour construire logiquement un fichier, il a fallu choisir des unités de traitement qui ne dépassent pas les quatre-vingts colonnes de la carte. Dès lors, l'unité d'enregistrement est soit le mot, soit la ligne pour la prose ou, ce qui est plus satisfaisant, le vers pour la poésie. Le mot est l'unité de base des fichiers du Laboratoire d'analyse statistique des langues anciennes (L.A.S.L.A.) de l'Université de Liège, tant pour le latin que pour le grec, tandis que la ligne d'édition et le vers ont été adoptés par le *Thesaurus Linguae Grecae*. Le vers est aussi l'unité dont ont fait choix Wilhelm Ott à Tübingen pour ses travaux sur la scansion de l'hexamètre latin et Etienne Evrard à Liège pour ses recherches sur la métrique grecque.

Le mot et la ligne posent problème: le premier n'a reçu à ce jour aucune définition unanimement acceptée par les linguistes et la seconde ne correspond à rien, sinon à une contrainte imposée par les techniques de l'imprimerie.

Quoi qu'il en soit, tous les fichiers de textes grecs et latins actuellement constitués ont des enregistrements de type mot, ligne ou vers.

La carte perforée qui se divise verticalement en 80 colonnes, se divise horizontalement en douze niveaux (ou lignes). Chacun de ceux-ci peut recevoir une perforation et une seule. Une perforation qui se situe au niveau 1 représente le chiffre 1, au niveau 2, le chiffre 2, etc. La codification des lettres résulte de la combinaison de deux perforations figurant à deux niveaux d'une même colonne.

Les premières perforatrices produisaient 48 combinaisons, c'est-à-dire 48 signes; ce nombre fut par la suite porté à 60. En raison de la pauvreté de ce système, les seuls signes que les machines connaissaient, étaient les chiffres (de 0 à 9), les principaux symboles mathématiques, les 26 lettres majuscules de notre alphabet, et quelques signes de ponctuation. Autrement dit, il n'était possible de coder sur carte ni les lettres minuscules, ni *a fortiori*, les caractères accentués des langues romanes et du grec.

Ces limites qualitatives n'ont que peu d'importance en latin où la liste des graphèmes est relativement réduite. Au reste, transcrire les textes latins en lettres capitales, n'est-ce pas renouer avec la tradition de l'écriture classique la plus ancienne?

Par contre, l'enregistrement du grec se heurtait à plusieurs difficultés. Comment symboliser les lettres que l'alphabet latin ne possède pas (êta, xi,...)? Comment perforer les esprit, les accents, le tréma, l'iota souscrit, etc.?

Nous verrons comment les responsables du TLG ont résolu ces questions. De manière générale, toutes les solutions qui ont été proposées et appliquées, que ce soit aux Etats-Unis, au Canada[6], en Angleterre[7] ou au L.A.S.L.A. de Liège, respectent autant que possible les normes de la tradition philologique.

De ces quelques observations techniques, nous retiendrons que les banques de données des littératures grecque et latine dont la plupart ont été conçues à une époque où le materiel et les systèmes logiques ne permettaient ni de produire, ni de reproduire les graphèmes que nous a transmis la tradition, reposent sur des conventions d'écriture qui perturbent le chercheur et rendent les traitements automatiques relativement complexes.

Un fichier qui ne contient que le texte brut d'une œuvre se prête déjà à quelques exploitations, il est toutefois peu utile s'il n'intègre des indications de référence qui aident à localiser précisément chaque mot, chaque ligne ou chaque vers. On trouve en matière de référenciacion deux solutions: ou bien la référence est calculée à chaque emploi du fichier grâce à des codes intégrés aux données textuelles, ou bien le calcul est effectué lors de l'enregistrement d'une œuvre et le résultat conservé avec le texte sur le support magnétique. La première solution qui répond à un critère d'économie d'occupation du support a été adoptée pour le TLG, la seconde, qui obéit à un critère d'économie de temps de travail, est préconisée par le L.A.S.L.A.

L'expérience a montré qu'il est préférable d'opter pour la deuxième solution en intégrant aux fichiers la référence de chacune des lignes d'un texte, voire, ce qui est plus précis encore, la référence de chacun des mots. Nous constaterons, dans la suite de cet article, que le système adopté par le TLG présente de nombreuses imprécisions.

La référence se compose d'une suite d'indications numériques ou alphabétiques qui permettent de situer chacun des mots d'une œuvre. Elle doit être conforme à l'*ars citandi* de la tradition philologique. En outre, dans un fichier informatique, on veillera à ce que la référence

[6] A l'Université Laval à Québec, pour le projet HIPPO que dirige Gilles Maloney.

[7] Voir, par exemple, l'Index de Platon publié par Léonard Brandwood.

soit aussi précise que possible. Ainsi, pour le *De bello gallico*, elle comportera les éléments suivants:

1. - une indication qui identifie l'auteur;
2. - une indication qui permet de désigner l'œuvre;
3. - le numéro du livre;
4. - le numéro du chapitre;
5. - le numéro du paragraphe;
6. - le numéro de la ligne ou du mot dans le paragraphe;
7. - éventuellement, le numéro de la ligne ou du mot dans l'œuvre.

En résumé, un fichier informatisé standard contient le texte et des éléments de référenciation. A ces données de base, s'ajoutent éventuellement des informations lexicales, morphologiques, syntaxiques et plus rarement sémantiques. Dans les fichiers latins et grecs élaborés au L.A.S.L.A., chaque forme d'un texte est rattachée à un 'lemme' et suivie de son analyse morpho-syntaxique.

Revenons aux bandes magnétiques du *Thesaurus Linguae Graecae*. Lorsqu'on les achète, elles sont accompagnées des trois documents. Le premier est une liste des œuvres acquises, le deuxième est un fascicule intitulé *Thesaurus Linguae Graecae, Beta Manual* dans lequel sont énumérés et explicités - trop peu - les principes de la codification des données. Enfin, le troisième est le *Thesaurus Linguae Graecae, Canon of Greek Authors and Works*[8], ouvrage de près de 400 pages qui est indispensable non seulement à l'utilisateur du TLG mais aussi aux hellénistes et aux historiens de la littérature grecque.

La lecture de l'introduction du *Canon* fournit de précieuses indications sur l'organisation et le contenu des bandes magnétiques du TLG.

En plus de l'introduction, le volume se compose de deux parties. La première est un répertoire des auteurs et des œuvres étudiés (pp. 1-326). La deuxième partie, beaucoup moins étendue, s'intitule *Index of TLG Author Numbers* (pp. 327-341), elle présente une simple liste des auteurs classés selon l'ordre croissant du numéro qui leur a été attribué dans la banque du TLG.

Dans la première partie, les auteurs sont rangés par ordre alphabétique. Le nom de l'auteur est précédé d'un numéro qui l'identifie

[8] Nous nous référons à la deuxième édition du *Canon* qui a été publiée en 1986 aux Presse de l'Université d'Oxford par Luci Berkowitz et Karl A. Squitier.

dans tous les fichiers du TLG. Ainsi, à la page 267, *PLUTARCHUS* est affecté du numéro 0007. Le nom de l'auteur est accompagné de l'abreviation de son épithète littéraire. Plutarque est dit *Biogr.* et *Phil.*

La deuxième ligne de la notice réservée à un auteur présente, d'une part, une indication chronologique et, d'autre part, l'épithète géographique. Pour Plutarque, nous trouvons *A.D. 1-2* et *Chaeronensis*.

L'épithète du genre littéraire sert aussi à distinguer les auteurs homonymes. Nous voyons, par exemple, que le TLG recense trois auteurs nommés *PLATO* (auxquels s'ajoute *Plato Iunior*). Le premier est un auteur de comédie dont nous n'avons conservé que des fragments, le deuxième est dit *Med.* et le troisième est le philosophe.

Sous le nom de l'auteur, viennent ensuite les titres de ses œuvres. Chacun d'eux est précédé d'un numéro de trois chiffres qui distingue les uns des autres les écrits d'un même auteur: 007 désigne *Lysistrata* dans la série des comédies d'Aristophane et le *De coloribus* dans les écrits d'Aristote. Par conséquent, pour trouver dans le corpus de Californie telle œuvre de tel auteur, il faut et il suffit d'indiquer les numéros qui identifient l'auteur et l'œuvre.

Chaque titre est suivi de la référence complète de l'édition à partir de laquelle les enregistrements ont été constitués (la page XXIII de l'introduction au *Canon* précise les critères qui ont présidé au choix des éditions).

Les informations bibliographiques sont suivies de deux renseignements qui se trouvent sur une même ligne. Le premier concerne le mode de transmission du texte. Celui-ci sera *Cod.* pour un texte transmis per la tradition manuscrite, *Pap.* si le texte nous est parvenu sur papyrus, etc.

La deuxième indication est le nombre de mots que contient l'œuvre. Cette valeur est en principe calculée par l'ordinateur, mais il s'agit parfois d'une estimation; dans ce cas, le nombre de mots est entouré de crochets droits. Cette information doit être utilisée avec la plus grande prudence, ainsi que l'exposent les auteurs du *Canon* aux pages XXVII et XXVIII de leur introduction en reconnaissant que les dénombrements sont très approximatifs.

Une fois repérés dans le *Canon* le numéro de l'auteur et celui de l'œuvre que l'on souhaite étudier, on pourra utiliser les bandes de Californie dont nous préciserons maintenant les particularités techni-

ques, tout en tenant compte des problèmes linguistiques qu'elles présentent.

Le document 1 reproduit les premières lignes d'un fichier[9] imprimé avec un simple programme de lecture-écriture. Il nous aidera à aborder les questions techniques.

```
~a"0086"b"035"c"Pol"y"1252a"z"t"
~@@@@@@@ 1$20"P"O"L"1"T"1"K"W"N "A$@1
~z1
@")EPEIDH\ PA=SAN PO/LIN O(RW=MEN KOINWNI/AN TINA\ OU)=SAN KAI\
PA=SAN KOINWNI/AN A)GAQOU= TINOS E(/NEKEN SUNESTHKUI=AN [1TOU= GA\R
EI)=NAI DOKOU=NTOS A)GAQOU= XA/RIN PA/NTA PRA/TTOUSI PA/NTES]1, DH=
LON W(S PA=SAI ME\N A)GAQOU= TINOS STOXA/ZONTAI, MA/LISTA DE\
```

Document 1

1.- Format des données

Les fichiers du TLG sont constitués d'une séquence de lignes; chaque ligne a une longueur, occupe un nombre de positions, qui varie en fonction du nombre de caractères qu'elle contient. Ainsi, la ligne 3 occupe 3 positions et sa longueur utile est 3 tandis que la ligne 1 occupe 32 positions.

Les lignes 4, 5 et 6 du document 1 représentent chacune une ligne de l'édition à partir de laquelle a été constitué le fichier, leur longueur respective est de 59, 63 et 64 positions.

2.- La référence

Dans le TLG, ni les lignes ni les mots ne sont accompagnés de leur référence. La localisation précise de chaque donnée doit être calculée à partir d'un codage intégré aux données textuelles. La brochure *Thesaurus Linguae Graecae, Beta Manual* contient malheureusement peu d'explications au sujet du système de référenciation des œuvres et elle laisse sans réponse la plupart des questions que se pose l'utilisateur. Celui-ci, s'il veut comprendre le système, est obligé de procéder, après bien des tâtonnements, à un examen minutieux des fichiers tout en se reportant aux éditions qui ont servi à les constituer.

[9] Il s'agit des premières lignes du fichier de la *Politique* d'Aristote.

Le document 1 illustre partiellement ce que nous venons de dire. La première ligne contient des informations qui serviront au calcul automatique de la référence. On y trouve dans l'ordre:

- l'accent circonflexe; il marque toutes les lignes qui portent des codes destinés au calcul de la référence.

- la deuxième information de la première ligne est la lettre minuscule *a*. Elle annonce la présence d'un numéro de référence qui se trouve en principe entre guillemets et qui représente un auteur déterminé. Ici 0086 est le numéro attribué à Aristote dans le système du TLG. On vérifiera qu'il s'agit bien du Stagirite en se reportant à la page 53 du *Canon*.

- vient ensuite un *b* minuscule qui introduit aussi un numéro figurant entre guillemets. Dans le document 1, il s'agit de 035 qui selon les conventions de l'Université de Californie est le numéro attribué à la *Politique* (cf. le *Canon* à la page 54, colonne 2).

- toujours sur la première ligne, se trouve ensuite une information introduite par la lettre minuscule *c* et placée entre guillemets. Il s'agit du début du titre de l'œuvre: *Pol* rappelle que l'on a affaire à la *Politique*. Cette information facultative n'apporte aucun renseignement nouveau, elle ne fait que répéter, sous une autre forme, ce qui est précisé à la rubrique *b*.

En résumé, les lettres *a*, *b* et *c* introduisent les éléments généraux de la référence, c'est-à-dire le numéro qui symbolise l'auteur, le numéro attribué à une œuvre et enfin les premières lettres du titre.

A la suite de ces éléments, la première ligne du document 1 porte encore deux codes de référence qui sont respectivement les lettres minuscules *y* et *z*. Ces deux lettres ont toujours pour fonction d'introduire des éléments qui permettent de localiser précisément chacune des lignes du texte conformément à l'*ars citandi*. Dans le cas d'Aristote, la lettre *y* introduit le numéro de la page et la lettre de la colonne de l'édition de Bekker, la lettre *z* étant destinée au calcul du numéro de la ligne dans la colonne.

Prenons comme deuxième exemple le fichier des *Argonautica* d'Apollonius de Rhodes, dont la première ligne est:

~a"0001"b"001"c"Arg"y1

Le numéro 0001 qui succède au *a* minuscule symbolise Apollo-

nius et 001 placé après le *b* minuscule représente les *Argonautica*; *c* minuscule introduit l'abréviation courante de l'œuvre *Arg.*; vient ensuite *y* qui annonce le numéro du livre (ici 1).

On notera que le 1, placé à la suite de *y*, n'est pas entre guillemets, contrairement à ce que nous avons vu pour les informations introduites par *a*, *b* et *c*. La raison en est que les éléments annoncés par les lettres *y* et *z* apparaissent tantôt entre guillemets, tantôt immédiatement à la suite des lettres, ce qui complique singulièrement les traitements et oblige à réaliser des programmes relativement complexes qui analysent les données caractère par caractère.

Pour *Argonautica*, il n'y a pas d'élément de référence introduit par *z*. Cela signifie implicitement que la première ligne du texte porte le numéro 1 et que celui-ci croîtra d'une unité à chaque nouvelle ligne.

En plus des codes de localisation *y* et *z*, le TLG a encore les codes *w* et *x* pour introduire des niveaux supplémentaires de référence. Le *Beta Manual* donne en exemple, à la rubrique *Citation System*, la référence *x12y12z3* qui, appliquée au *Nouveau Testament*, signifie chapitre 12, verset 12, ligne 3. On trouvera le même système de référenciation pour Plutarque où *x2y3z1* désigne le chapitre 2, paragraphe 3, ligne 1 de chacune des *Vitae*.

En fait, les éléments référencés par les codes *w*, *x* et *y* ne sont pas univoques, leur signification varie selon les auteurs et les œuvres. Ainsi, nous avons vu que *y* introduit pour Aristote, la page et la colonne de l'édition Bekker, pour le *Nouveau Testament*, le verset et pour Plutarque, le paragraphe. Pour Homère, *y* annonce le livre, ainsi, la suite y2z44 identifie le vers 44 du livre 2 d'une des épopées homériques.

La première ligne d'un fichier contient la référence initiale du texte; dans la suite du fichier, figurent des lignes qui sont destinées à l'ajustement de cette référence. Ici encore le système apparaît relativement complexe.

Supposons une ligne qui contient simplement *y*. Elle peut indiquer qu'il faut augmenter de 1 le contenu de l'élément régi par *z*. Dans l'*Iliade*, par exemple, les caractères ~*y* signifient que le numéro du livre doit être augmenté de 1 et que le numéro du vers doit être ramené à 1. Mais un enregistrement contenant seulement *y* peut aussi marquer un autre type de changement. Reprenons une fois encore Aristote pour lequel nous avons vu que l'élément *y* introduit le numéro de page et, simultanément, la lettre qui désigne la colonne. Dans les fichiers d'Aristote, une ligne portant *y* annonce que l'on

passe à la colonne b de Bekker.

Une ligne *~b*, située à un endroit quelconque du fichier, aura pour effet, quant à elle, de réinitialiser à 1 les variables de référence introduites par *w*, *x*, *y* et *z*, puisque *b* annonce une nouvelle œuvre.

Les indications qui précèdent exposent les principes généraux du système de référenciation, il nous est malheureusement impossible de détailler ici les nombreux cas particuliers que l'on rencontre dans les bandes du TLG et qui sont parfois difficiles à comprendre. En voici quelques exemples.

a.- Les codes de référenciation se situent toujours entre deux lignes de textes et non au milieu d'une ligne. Il en résulte que, dans bien des cas, une partie de la ligne ne sera pas référencée correctement. Ainsi, dans Démosthène, la marque de changement de paragraphe affecte toute la ligne alors que le nouveau paragraphe commence généralement après une ponctuation: à la fin du paragraphe 4 de la *Première Olynthienne*, le signe *y* marque le début d'un paragraphe devant la ligne πρὸς 'Ολυνθίους, ἐναντίως ἔχει. δῆλον γάρ ἐστι τοῖς 'Ολυνθίοις. Dès lors, toute la ligne sera considérée comme faisant partie du deuxième paragraphe alors que celui-ci commence seulement à δῆλον.

b.- Pour certaines œuvres, le système de référenciation du TLG ne respecte pas les usages des philologues. Ainsi, alors que la plupart des œuvres du Stagirite sont normalement référencées selon l'édition de Bekker, pour *Magna Moralia*, la référence se fait par paragraphe, en suivant l'édition de F. Susemihl[10].

c.- Le *De falsa legatione* est référencé, comme la plupart des œuvres de Démosthène par paragraphe. Dans l'édition de Butcher à laquelle se réfère le TLG, il y a une rupture de la numérotation séquentielle pour les paragraphes 104 à 109 que l'éditeur note en marge *104 ad 109*. Cette présentation est fidèlement reproduite dans le fichier ce qui, on s'en doute, pose des problèmes au moment de la référenciation automatique.

d.- Dans les *Antiquitates Romanae*, Jacoby, l'éditeur de Denys d'Ha-

[10] Voir à ce propos le *Canon* à la page 54, colonne 1.

licarnasse, note le numéro des chapitres en chiffres romains et rappelle par des chiffres arabes placés entre parenthèses une numérotation alternative. Or, au livre XX, entre le chapitre X, on trouve dans l'édition de Jacoby un chapitre numéroté (*10*) qui, dans les données du TLG est mis sur le même pied que les chiffres romains. Dès lors, au moment de la référenciation automatique, ce chapitre (*10*) devient X, le vrai chapitre X porte le numéro XI, etc.

Pour terminer ces remarques sur la référenciation, il faut noter que les fichiers du TLG prévoient un second système de référence. En effet, en plus des codes relatifs à l'art de citer habituel, les bandes contiennent des codes (@1) qui marquent les fins de page de l'édition utilisée pour la saisie du texte, des codes @2 qui spécifient les fins de colonnes,... En ce qui nous concerne, nous n'utilisons jamais ce second système de référenciation.

3. L'alphabet

Code	Lettre grecque	Code	Lettre grecque
A	Alpha	O	Omicron
B	Bêta	P	Pi
C	Xi	Q	Thêta
D	Delta	R	Rho
E	Epsilon	S	Sigma
F	Phi	T	Tau
G	Gamma	U	Upsilon
H	Eta	V	Digamma
I	Iota	W	Omega
K	Kappa	X	Chi
L	Lambda	Y	Psi
M	Mu	Z	Zêta
N	Nu		

Document 2

La rubrique *The Alphabet* du *Beta Manual* donne le système de codage de l'alphabet grec que nous reproduisons au document 2.

Les options prises pour la représentation des caractères grecs visent avant tout à faciliter la lecture de documents imprimés en caractères latins, c'est la raison pour laquelle W représente omega, G, gamma, H, êta et Z, zêta. La suite de caractères TON BARBARON (τον βαρβαρον) se lit aisément. De même, on reconnaîtra, parfois peut-être avec un effort d'imagination WS (ως), AEQLWN (αεθλων) ou encore AMFRUSSOIO (αμφρυσσοιο)&[11].

Le fait qu'un même signe symbolise le sigma interne et le sigma final n'entraîne guère de difficultés[12]; il suffit, si l'on veut imprimer en caractères grecs, de commander par programme la production du sigma interne et du sigma final.

Le TLG pose quelques problèmes au moment du traitement: dans la mesure où le codage des lettres ne respecte pas l'ordre de l'alphabet grec, il faut développer des procédures de préparation au tri alphabétique qui restitueront la séquence habituelle des caractères.

4. Les esprits et les accents

Il n'est pas possible d'enregistrer les esprits, les accents, le tréma ou l'iota souscrit en même temps que la lettre sur laquelle ou sous laquelle ils figurent: la codification binaire propre à l'ordinateur ne dispose pas d'un nombre de signes suffisamment élevé pour symboliser par un seul code toutes les combinaisons possibles. Prenons par exemple la lettre alpha, elle peut apparaître soit seule, soit avec un des signes suivants: esprit doux, esprit rude, accent aigu, accent grave, accent circonflexe, iota souscrit. En outre, ces signes se combinent parfois: esprit doux et accent aigu, esprit doux et accent grave, esprit rude et accent aigu, esprit rude et accent grave, etc.

Les limitations de la codification binaire s'ajoutent à la difficulté de disposer de matériels capables de dactylographier et d'imprimer ces différents signes. Au moment où l'enregistrement des textes a débuté en Californie, peu d'imprimantes d'ordinateur produisaient

[11] C'est volontairement que nous omettons dans ce paragraphe les esprits et les accents.

[12] L'emploi d'un signe unique pour le sigma interne et pour le sigma final pose aussi des problèmes sur CD-Rom comme le montre l'article de David Bouvier cité précédemment.

l'alphabet grec, les accents et les esprits.

Pour la représentation des diverses combinaisons de graphèmes, les concepteurs du TLG ont décidé de scinder les informations et de traiter comme des signes distincts des lettres, les accents, les esprits, l'iota souscrit et le tréma. Ils ont en outre choisi pour représenter ces signes des caractères qui existaient sur les imprimantes. Les conventions que nous reprenons ici sont extraites du *Beta Manual*.

|	iota souscrit	/	accent aigu
)	esprit doux	\	accent grave
(	esprit dur	=	accent circonflexe
+	tréma		

En règle générale, l'accent, l'esprit, le tréma et l'iota souscrit se placent à la suite de la lettre qu'ils affectent comme le montrent les mots A)RXO/MENOS (ἀρχόμενος) - KATA/ (κατά) ou KATA\ (κατὰ) - O((ὁ) - OI((οἱ) - R(A|DI/AN (ῥᾳδίαν) - PROI+E/MENOI (προϊέμενοι).

Lorsqu'un esprit et un accent portent sur la même lettre, les conventions du TLG sont de placer d'abord l'esprit et ensuite l'accent comme on le voit dans les mots A)/LLOS (ἄλλος) - E(\N (ἓν) - W)=(ὦ) - OU)=N (οὖν) - A)E/QLWN (ἀέθλων) - W(=N (ὧν).

Dans les cas où esprit et/ou accent, ainsi que l'iota souscrit affectent une même lettre, iota souscrit vient en dernier lieu, c'est le cas pour des formes telles que TH=| (τῇ) - POLLW=| (πολλῷ) - W(=| (ᾧ).

Les normes de codification des différents signes se résument comme suit:

- en premier lieu vient la lettre qui est affectée d'un esprit, d'un accent, etc.
- apparaissent ensuite dans l'ordre, l'esprit, l'accent, le tréma et l'iota souscrit.

La postposition des esprit et des accents par rapport à la lettre sur laquelle ils portent, n'est pas une règle absolue. En effet, il faut tenir compte d'une exception notable: l'esprit et l'accent précèdent la lettre lorsque celle-ci est un caractère majuscule. On aura donc, par exemple,)AQHNAI=OI (Ἀθηναῖοι) et)/ARGON (Ἄργον).

Cette dernière convention s'applique indifféremment aux noms propres et aux mots dont l'initiale est une capitale parce qu'ils sont

placés en début de phrase. Ainsi, on aura à l'intérieur d'une phrase A)RXO/MENOS (ἀρχόμενος) mais au premier vers des *Argonautica* d'Apollonius de Rhodes, on trouvera la forme *)ARXO/MENOS (Ἀρχόμενος), de même on rencontrera tantôt *)/ETI (Ἔτι), tantôt E)/TI (ἔτι) selon la position de ce mot dans la phrase.

Nous ajouterons aux différents signes qui symbolisent les esprits et les accents, le signe de l'apostrophe (') qui marque l'élision.

5. Lettres majuscules et minuscules

Les lettres majuscules et minuscules sont représentées par un code identique de telle sorte que les unes et les autres apparaissent toujours en majuscule même si la majorité d'entre elles sont en réalité des minuscules.

Dans les fichiers du TLG, toute lettre est une minuscule sauf si elle est précédée immédiatement d'un astérisque. Ce signe indique que la lettre qui le suit est une capitale. Dès lors, la suite de caractères *KALLIO/PH doit se lire 'Kallio/ph' (Καλλιόπη).

Ce principe souffre une exception qui est de nature à entraver la recherce et qui complique quelque peu les programmes d'exploitation: entre l'astérisque et la lettre sur laquelle il porte, on trouve les accents et les esprits situés normalement avant l'initiale: *)/ARGON (Ἄργον) doit être interprété comme esprit doux et accent aigu suivis d'un alpha majuscule.

Un rappel au sujet des capitales: le souci de respecter scrupuleusement le texte de l'édition fait que toutes les majuscules sont notées de la même manière sans distinguer celles qui marquent le début d'une phrase. Ce choix est regrettable dans la mesure où il empêche le développement de logiciels capables d'extraire automatiquement tous les noms propres employés dans un texte littéraire, dans des citations ou encore dans un document papyrologique ou épigraphique.

6. Les ponctuations

Les conventions adoptées pour les ponctuations sont simples:

, virgule	. point
: point en haut	; point d'interrogation

7. Les signes critiques

Les signes critiques usuels tels que (,), <, >, [,], etc. que l'on rencontre dans nos éditions de textes classiques sont codés dans le TLG selon les conventions reprises dans le tableau ci-dessous:

Code	Signification	TLG	Editions
[	crochet droit ouvert		
]	crochet droit fermé	[KAI\]	[ ]
[1	parenthèse ouverte		
]1	parenthèse fermée	[1KAI\1]	( )
[2	crochet oblique ouvert		
]2	crochet oblique fermé	[2KAI\2]	< >
[3	accolade ouverte		
]3	accolade fermé	[3KAI\3]	{ }
[4	double crochet droit ouvert		
]4	double crochet fermé	[4KAI\4]	⟦ ⟧

Un exemple de l'emploi des parenthèse figure aux lignes 5 et 6 du document 1.

Le système de codification du TLG présente d'autres signes dont on tiendra compte lors de la mise au point des procédures d'exploitation. La liste complète des convention en matière de signe critiques figure dans le *Beta Manual* à la rubrique *Brackets*.

8. La présentation du texte

La connaissance des conventions aide à comprendre comment se présente un texte enregistré sur les bandes du *Thesaurus* de Californie. Il reste néanmoins quelques points sur lesquels notre attention doit se porter. Il s'agit notamment de l'emploi du tiret, du traitement réservé aux titres et aux noms des personnages dans le théâtre, de la

division en paragraphes, ainsi que de la disposition des lignes de texte incomplètes et alignées à droite par l'éditeur.

a.- L'emploi du tiret

Puisque les bandes du TLG respectent la présentation des éditions, elles conservent fidèlement la répartition des mots en lignes; cela implique que lorsqu'un mot situé en fin de ligne est coupé dans l'édition de référence, il est aussi coupé dans le fichier informatique. On trouve un cas de césure à la sixième ligne du document 1 qui se termine par DH=-, première syllabe de DH=LON (δῆλον).

Cette disposition présente des inconvénients certains. La recherche d'une forme quelconque dans un texte suppose la réunion préalable des mots coupés. Imaginons que l'on essaie de localiser chez un auteur tous les emplois de δῆλον en explorant le fichier ligne par ligne, il est clair que le logiciel ne repérera jamais l'occurrence du début de la *Politique* si l'on n'a pris la précaution de regrouper la fin de la ligne 6 et le début de la ligne 7.

Il n'est pas difficile de programmer l'ordinateur pour qu'il repère lignes se terminant par un tiret, puis, lorsqu'il en a trouvé une, d'en extraire les caractères qui se situent entre le tiret et le premier espace rencontré en remontant vers le début de la ligne et enfin d'ajouter les caractères au début de la ligne suivante.

Toutefois, dans le cas du TLG, pour réunir correctement les parties d'un mot coupé, il faut éliminer au préalable les signes 'fin de page - @1', 'fin de colonne - @2', etc. qui se situent normalement à la fin d'une ligne et qui peuvent dès lors s'intercaler entre les deux parties d'un mot.

b.- Les noms de personnages et les titres

Les titres, les sous-titres, les subdivisions d'œuvre et, dans les œuvres dramatiques ou dans les dialogues de Platon notamment, les noms des personnages apparaissent dans les fichiers informatiques. Ces données sont placées entre accolades, ce qui signifie qu'elles ne font pas partie intégrante du texte et qu'elles doivent être ignorées pour certaines exploitations.

Il est regrettable que le même signe (l'accolade) ait été choisi

pour marquer les titres et les noms des personnages, en effet, les premiers peuvent être laissés de côté dans la plupart des cas, mais il n'en va pas de même pour les seconds qui doivent être conservés pour certaines études stylistiques et thématiques.

c.- Conventions typographiques

Le lecteur trouvera ici quelques remarques sur des difficultés de traitement liées à la typographie des éditions.

Sans doute, les normes d'impression adoptées par les éditeurs ne devraient-elles pas influencer l'exploitation des données textuelles, mais ici encore la volonté des responsables du TLG de produire des fichiers qui reflètent fidèlement, scrupuleusement, les modèles de base, est de nature à rendre plus complexe l'obtention de résultats corrects. Nous donnerons trois exemples.

Les imprimeurs ont l'habitude de signaler le début d'un paragraphe par un retrait de la ligne vers la droite (*indentation*). Dans les fichiers du TLG; l'alinéa est marqué par le signe @ qu'il conviendra d'éliminer.

En principe, une ligne incomplète commence, comme les autres, à la marge de gauche. Toutefois, sous certaines conditions, les éditeurs alignent le texte à partir de la droite, laissant ainsi en blanc le début de la ligne. Dans les fichiers du TLG, le signe @ apparaît autant de fois qu'il y a d'espaces avant le texte.

Dans une édition de texte, on recourt à différentes polices de caractères de manière à mettre en évidence un mot, une phrase, un paragraphe, une citation, etc. Les fichiers du TLG enregistrent tous les changements de type ou de police de caractères. Ainsi, pour marquer le passage du caractère grec normal à des caractères droits, on insère les signes *$8*. Ces suites de signes (il y en a quelques dizaines) seront exclues des données en vue des exploitations linguistiques.

Nous ne prétendons pas décrire dans cet exposé toutes les conventions du TLG: les exemples retenus illustrent suffisamment les problèmes que posent les données aux points de vue technique et linguistique. Nous espérons qu'ils donnent au futur utilisateur les clefs d'accès les plus nécessaires.

Nous avons jugé utile d'insister assez longuement sur certaines difficultés parce qu'elles doivent être résolues par la programmation

avant même que l'on tente d'extraire des fichiers la moindre information. On constatera que dès l'instant où l'on a bien assimilé les principes qui ont présidé à la constitution de la banque de données, l'exploitation du TLG devient possible mais elle n'est jamais ni simple, ni exempte de problèmes.

Cette constatation nous amène à une question: tous ces auteurs, tous ces textes recopiés sur support informatique, à qui et à quoi cela peut-il servir?

Personne n'a apporté à ce jour une réponse vraiment satisfaisante à cette question. Très souvent, ceux qui constituent des banques de données se contentent de dire qu'il convient de faire preuve d'imagination pour utiliser des moyens qui favorisent un nouveau type d'approches de la littérature, de la civilisation ou de l'histoire de la Grèce et de Rome. Ils affirment que c'est au chercheur à trouver lui-même quelles enquêtes il mènera à partir des fichiers informatiques. C'est là une attitude un peu facile qui montre combien il est malaisé de définir avec précision en quoi les nouvelles technologies de traitement de l'information sont utiles à nos disciplines.

En proposant, à titre d'exemples, quelques exploitations, je ne prétends en aucun cas énumérer toutes les possibilités d'études qu'offrent les banques de données linguistiques et littéraires.

L'ordinateur, le *computer* comme disent les Anglais et les Américains, est parfois considéré comme une machine à calculer. Cette conception pourrait conduire à penser que les fichiers informatiques servent uniquement à des recherches quantitatives. Or, si l'ordinateur est souvent utilisé à des fins de calcul, il rend aussi service à celui qui tente une approche qualitative de l'œuvre littéraire. En réalité, l'informatique trouve sa raison d'être chaque fois qu'il faut extraire d'une œuvre tel phénomène, soit de manière exhaustive, soit à titre exemplatif.

Parmi les documents que produit la machine, nous connaissons bien les index, les concordances, les listes de fréquence du vocabulaire ou encore les dénombrements relatifs à la morphologie et à la syntaxe. Mais l'ordinateur fournit, s'il est programmé de manière adéquate, des données plus spécifiques: le philologue qui prépare une recherche portant sur une question précise rassemblera plus facilement la documentation dont il a besoin s'il sait interroger les fichiers informatiques. C'est notamment pour cette raison qu'il importe,

préalablement à toute enquête, de définir avec soin les questions que l'on soumet à la machine.

Mais quels faits linguistiques peut-on rechercher à partir des fichiers du TLG? Nous envisagerons les relevés sur les graphèmes, sur les mots et sur la syntaxe.

1. Les graphèmes

On réalise aisément plusieurs types de relevés si l'on choisit simplement comme critère les lettres grecques.

On constituera d'abord un tableau dans lequel chaque graphème sera accompagné de sa fréquence d'emploi dans une œuvre ou dans un ensemble d'œuvres.

Cette application élémentaire et purement statistique débouche sur des enquêtes stylistiques dans la mesure où elle aide à déceler chez un auteur des passages où la fréquence des graphèmes diffère significativement du pourcentage moyen.

Partant des graphèmes, il est encore possible, avec un programme de syllabation, de réunir les matériaux nécessaires á des recherches sur la syllabe ou sur l'allitération.

Enfin, on développera, pour la poésie, des programmes de scansion automatique qui conduiront à des relevés et à des travaux relatifs à la métrique.

2. Recherche sur les mots

Le mot est l'unité la plus fréquemment retenue pour les études faites à partir des fichiers informatiques. Ce choix semble normal dans la mesure où le mot est la plus petite entité qui soit porteuse de sens.

Parler du 'mot', c'est poser le problème fondamental de sa définition. Qu'est-ce qu'un 'mot'? Ni les linguistes, ni les philologues, ni les lexicologues n'en ont donné une définition satisfaisante. Les spécialistes de la linguistique informatique acceptent dès lors une définition approximative qui sera applicable dans la majorité des cas[13]: le mot est la suite de caractères comprise entre deux espaces, étant entendu que nous appelons caractères les lettres de l'alphabet, les accents, les esprits, le tréma, l'iota souscrit, les signes de numération,

[13] Voir à ce sujet Ét. Brunet, *Les noms propres chez Zola*, dans *Studies in Honour of Roberto Busa S.J.*, Linguistica computazionale, 4-5 (1987), 30.

etc.

Dans le TLG, pour distinguer les mots les uns des autres, on éliminera d'abord tous les signes critiques, les ponctuations, les codes de référenciation, etc. Au surplus, pour certaines exploitations, on supprimera le signe 'astérisque' qui précède toutes les lettres majuscules.

A l'aide d'un programme qui réalise la découpe du texte en mots, il sera possible de produire plusieurs relevés.

Le premier est un index de toutes les formes employées dans une ou plusieurs œuvres.

L'ordinateur imprime une liste dans laquelle les mots sont rangés en ordre alphabétique et accompagnés de leur référence. Ce document porte éventuellement des indications de fréquence de telle manière que l'on connaisse le nombre d'occurrences de chacune des formes et aussi le nombre total de formes de l'ensemble étudié.

Le document 3 donne un bref extrait de l'index de la *Politique* d'Aristote où les différentes formes du mot ἄνθρωπος sont accompagnées de leur référence. Les formes identiques sont regroupées et classées selon l'ordre croissant des pages et des lignes de l'édition de Bekker. Enfin à la suite d'une 'rubrique' apparaît la fréquence d'emploi de chaque forme.

A)/NQRWPOI	0086	035	1252b	27
1				
A)NQRW/POIS	0086	035	1253a	16
1				
A)/NQRWPO/S	0086	035	1253a	32
1				
A)/NQRWPOS	0086	035	1253a	2
A)/NQRWPOS	0086	035	1253a	4
A)/NQRWPOS	0086	035	1253a	7
A)/NQRWPOS	0086	035	1253a	10
A)/NQRWPOS	0086	035	1253a	34
5				
A)NQRW/POU	0086	035	1252b	34

Document 3

Lors de la consultation d'un index, le chercheur portera son attention sur les quelques points qui concernent la présentation générale des relevés et la répartition des mots en rubriques.

a.- l'ordre alphabétique de nos documents n'est pas celui de l'alphabet latin. Cela signifie que les mots qui commencent par le caractère G sont rangés entre B et D parce que la lettre G représente en fait gamma. De même les mots dont l'initiale est Z sont classés à la suite de ceux dont la première lettre est E. Le classement que l'on obtient pour l'ensemble de l'alphabet est donné dans le document 4.

A	alpha	N	nu
B	bêta	C	xi
G	gamma	O	omicron
D	delta	P	pi
E	epsilon	R	rho
Z	zêta	S	sigma
H	êta	T	tau
Q	thêta	U	upsilon
I	iota	F	phi
K	kappa	X	chi
L	lambda	Y	psi
M	mu	W	oméga

Document 4

En regardant ce tableau, on comprendra qu'il faut chercher un mot à la place qu'il occupe normalement dans l'ordre alphabétique du grec puisque, par exemple, tous les mots qui commencent par thêta (lettre Q) se trouvent à la suite des mots dont l'initiale est H (êta) et non entre P et R comme en latin.

b.- Afin d'effectuer le meilleur regroupement possible des formes identiques, nous avons décidé de normaliser les données du TLG dans deux cas précis. Le premier concerne le traitement des lettres majuscules et le deuxième celui de l'accent grave.

Etant donné que les mots commençant par une majuscule sont tous marqués par un astérisque, nous avons décidé d'éliminer ce signe

dont la présence provoquait des distinctions erronées dans nos relevés. En effet, l'ordinateur est incapable de réunir sous une même rubrique des formes telles que *PRO/S (Πρός) ou PRO/S (πρός), simplement en raison du fait que les unes sont situées au début et les autres à l'intérieur de la phrase.

Notre choix présente sans doute l'inconvénient de faire disparaître les majuscules initiales de noms propres mais cette suppression ne nous paraît pas constituer un handicap grave pour le chercheur.

c.- Nous avons procédé à une deuxième transformation des données: elle concerne l'accent grave. Reprenons la préposition πρός et constatons qu'elle peut apparaître dans un texte soit avec l'accent aigu (PRO/S - πρός), soit avec l'accent grave (PRO\S - πρὸς). L'ordinateur range ces deux formes dans des rubriques différentes et calcule distinctement la fréquence d'emploi de l'une et de l'autre. Cette façon de procéder n'a aucune justification sur le plan linguistique, c'est pourquoi la substitution systématique de l'accent grave permet de ranger dans un même ensemble les formes que différencient les deux accents.

d.- Problèmes liés à la consultation d'un index de formes

Les listes alphabétiques de formes, quels que soient les services qu'elles rendent, doivent être consultées avec prudence car leur utilisation présente des risques certains de confusion. Nous prendrons deux exemples.

Celui qui s'intéresse à un mot quelconque veillera à rechercher toutes les variantes sous lesquelles ce mot est susceptible d'apparaître. Ainsi, une enquête sur la préposition META/ (μετά) serait incomplète si l'on ne prenait soin de vérifier aussi la présence des formes syncopées MET'(μετ') et MEQ'(μεθ').

A fortiori une recherche sur un mot à flexion nominale ou verbale demande-t-elle une consultation plus longue. Pensons que γίγνομαι apparaît dans la *Métaphysique* d'Aristote sous plus de quarante formes différentes qui dans un index non lemmatisé formeront plus de quarante rubriques.

A l'opposé du problème de dispersion des formes issues d'un même vocable, on trouvera réunis sous une même rubrique des mots

qui appartiennent à des unités lexicales différentes. L'exemple le plus fréquent est sans doute celui de formes de substantifs en -σις qui peuvent se confondre avec des verbes employés au futur. Ainsi, dans la *Métaphysique*, la forme κινήσει est usitée 15 fois. Cette observation étant faite, on constate en se reportant au contexte que κινήσει est 3 fois une forme de κινέω et 12 fois le substantif κίνησις. L'ordinateur, quant à lui, n'opère pas la distinction entre les deux vocables, il regroupe les quinze emplois sous une même rubrique.

L'index oblige le chercheur qui souhaite avoir sous les yeux le contexte d'un mot, à se reporter constamment à une édition. Or, dans certaines enquêtes, il est possible de se contenter d'un contexte limité à quelques mots. On choisit dans cette perspective de faire produire par l'ordinateur une concordance complète d'une œuvre, c'est-à-dire une liste dans laquelle les mots sont rangés en ordre alphabétique, accompagnés de leur contexte immédiat. Ce dernier sera plus ou moins étendu selon les desiderata de l'utilisateur (Document 5).

AUTEUR: 0086, TEXTE: 035

1253a 33	XALEPWTS/TH GA\R	*** A)DIKI/A	*** E)/XOUSA O(/PLA O(
1253a 15	DI/KAION KA\I TO\	*** A)/DIKON	*** TOU=TO GA\R PRO\S
1253a 17	KAI\ DIKAI/OU KAI\	*** A)DI/KOU	*** KAI\ TW=N A)/LLWN
1253a 27	TA\ EI/DH E(AUTOI=S	*** A)POMOIOU=SIN	*** OI(A)/NQRWPOI, OU(/TW
1253a 5	U(F'(OMH/ROU LOIDORHQEI\S	*** A)FRH/TWR	*** A)QE/MISTOS A)NE/STIOS A(/MA
1253a 36	A)RETH=S, KAI\ PRO\S	*** A)FRODI/SIA	*** KAI\ E)DWDH\N XEI\RISTON.

Document 5

La plupart des concordances présentent le mot-vedette - celui dont on donne le contexte - au centre de la ligne entouré d'un nombre identique de mots à gauche et à droite, mais on peut dans certains cas imposer à l'ordinateur de répartir différemment le contexte en précisant, par exemple, que l'on désire le mot-vedette en début de ligne.

Pour l'échantillon de concordance que reproduit le document 5, nous avons choisi de centrer le mot-vedette et de donner un contexte de 3 mots avant et de 3 mots après celui-ci. On remarque que dans la première ligne, il n'y a que deux mots avant le mot-vedette en raison du fait que XALEPWTA/TH (χαλεπωτάτη) est en début de phrase.

L'ordinateur peut, en même temps qu'il réalise un index général des formes ou une concordance, préparer un index hiérarchisé, c'est-à-dire une liste où figure en regard de chaque forme une valeur numérique qui indique le nombre d'emplois de la forme dans le corpus étudié. Dans un tel relevé, les formes sont présentées soit en ordre alphabétique, soit selon l'ordre croissant ou decroissant de leur fréquence. L'intérêt de ce type de listes est évident lorsque l'on travaille sur des lexèmes et non sur des formes. Néanmoin un index hiérarchisé de formes est utile, si on l'exploite avec discernement, pour des études de vocabulaire, pour des études thématiques ou encore pour des enquêtes portant sur la morphologie ou la syntaxe. Ainsi, à partir d'un index ou d'une concordance, on étudiera les emplois de ἄν, les relatives ou les propositions subordonnées introduites par tel ou tel type de conjonction.

Dans l'échantillon d'index hiérarchisé que nous donnons au document 6, les formes sont classées en ordre décroissant de leur fréquence, et, pour les fréquences identiques, en ordre alphabétique.

KAI/	73	EI)=NAI	7	OI(=ON 5
DE/	28	E)N	7	O(/TI 5
GA/R	25	POLITIKO/N	7	OU)K 5
H(	23	DIO/	6	TA/ 5
TO/	22	E)K	6	TH=S 5
ME/N	16	KOINWNI/A	6	TOU/TWN 5
FU/SEI	14	MH/	6	W(S 5
TOU=	11	OU(/TW	6	A)GAQOU= 4
O(	10	A)LL'	5	BASILIKO/N 4
D'	9	A)/LLOIS	5	DH=LON 4

H)/	9	A)/N	5	DOU=LON	4
TW=N	9	A)/NQRWPOS	5	E(/KASTOS	4
W(/SPER	9	DIA/	5	E(/NEKEN	4
OU)=N	8	E)C	5	E)STI/N	4
PO/LIS	8	E)STIN	5	E)/TI	4
TH/N	8	KATA/	5	OI)KI/A	4
TOI=S	8	OI(	5	PLEIO/NWN	4

Document 6

Le quatrième type de documents que la machine produit sans difficulté est un index inverse, c'est-à-dire une liste des formes rangées en ordre alphabétique non pas à partir du début du mot mais à partir de la finale (Document 7).

En fonction de l'état du texte, l'index inverse rendra service à celui qui édite un papyrus, une inscription ou même une œuvre littéraire. Il est utile aussi pour des recherches sur la morphologie en général et sur les mots suffixés en particulier. Il fournira, par exemple, la documentation nécessaire à une étude sur la formation et l'emploi des substantifs neutres en -μα, à la condition, bien entendu, que l'on sélectionne toutes les finales déclinées -ματος,-ματι,-ματα, etc. et que l'on élimine les termes qui, bien que se terminant en -μα, ne sont pas formés à partir du suffixe auquel on s'intéresse.

AUTEUR: 0086, TEXTE: 035

DIA	DIA/	1252a 31
		1252b 21
		1253a 3
		1253a 3
AUTARKEIA	AU)TA/RKEIA	1253a 1
ADIKIA	A)DIKI/A	1253a 33

OIKIA	OI)KI/A	1252b 10
		1252b 20
		1253a 19
		1253b 3
APOIKIA	A)POIKI/A	1252b 17
KOINWNIA	KOINWNI/A	1252a 7
		1252b 7
		1252b 13
		1252b 15
		1252b 28
		1253a 18

Document 7

Dans l'index inverse, nous imprimons deux fois chaque mot, d'abord sous une forme normalisée, sans esprit et sans accent et ensuite avec esprit et accent: la forme normalisée est la rubrique de classement. Cette présentation nous paraît avantageuse parce qu'elle permet de regrouper sous une même entité des mots qui ne diffèrent que par les accents et les esprits: A)LLA/ (ἀλλά) et A)/LLA (ἄλλα) seront rangés sous ALLA.

Parmi les possibilités d'exploitation systématique des bandes de Californie, nous citerons encore la préparation automatique d'un *Index nominum*. Pour réaliser ce relevé, il suffit d'extraire de la banque de données toutes les formes dont l'initiale est un caractère majuscule, c'est-à-dire, dans le cas du TLG, toutes les formes dont le premier signe est l'astérisque. Bien entendu, ce procédé quelque peu brutal fournit des informations partiellement inadéquates puisque l'on isolera en même temps que les noms propres tous les mots qui sont en début de phrase, voire en début de vers. Il appartiendra alors au chercheur d'éliminer, en recourant à un progiciel d'édition de fichiers, les mots qui ne sont pas des noms propres.

L'ordinateur peut aider davantage le chercheur dans sa tâche d'identification et de sélection des noms propres. En effet, on peut programmer la machine pour lui faire constituer deux fichiers, le premier contiendra tous les mots dont l'initiale est une majuscule et

qui ne sont situés ni en tête de phrase ni en tête de vers. Quant au second fichier, il recevra toutes les formes commençant par une capitale qui ne sont pas conservées dans le premier. Le chercheur aura donc à corriger plus particulièrement cette deuxième liste.

Indépendamment des relevés globaux que nous venons de passer en revue, le TLG se prête aussi à des consultations ponctuelles dont l'objectif est de retrouver dans le corpus de la littérature grecque un ou plusieurs mots. Nous citerons à titre d'illustration quelques exploitations possibles.

Le balayage intégral des bandes du TLG facilite certaines études thématiques. Il rend possible le repérage de tous les emplois d'un terme chez un ou plusieurs auteurs. Ainsi, l'ordinateur établit le relevé des passages où apparaît le nom d'une divinité dans toute la littérature grecque ou bien il dresse la liste de toutes les occurences d'un ou de plusieurs mots chez un auteur. On isolera, par exemple, tous les passages où Aristote emploie le mot ἰδέα.

Les avantages d'une telle exploitation sont évidents: non seulement la documentation nécessaire à une enquête est rapidement rassemblée, mais l'utilisateur a en outre la certitude de disposer d'un relevé exhaustif. Encore faut-il pour cela tenir compte, au moment d'interroger l'ordinateur, des problèmes liés à la polymorphie. Ainsi, dans le cas de ἰδέα, il conviendra d'extraire toutes les formes déclinées du mot, sans oublier les éventuelles variantes dialectales possibles.

Le deuxième exploitation consiste à retrouver une citation d'un texte grec faite par un auteur ancien ou par un humaniste. La recherche s'effectue à partir d'un ou de plusieurs mots.

Récemment, une étudiante préparant un travail sur Juste Lipse s'efforçait d'identifier certaines citations faites par le célèbre *humaniste, et notamment celle-ci:*

μεγίστη πρᾶξις ἡ ἀπραξία.

Nous avons donc interrogé la totalité de la banque du TLG. Nous aurions pu programmer la machine à partir d'un des mots cités par Juste Lipse, par exemple à partir de πρᾶξις, mais nous craignions d'obtenir un trop grand nombre de références et par conséquent beaucoup de contextes à examiner; c'est pourquoi nous avons décidé

de localiser l'ensemble de l'expression. En vain!

Supposant que Juste Lipse citait de mémoire et n'hésitait pas à adapter le texte à ses propos, nous avons pensé qu'il avait peut-être introduit lui-même l'adjectif μεγίστη.

Nous avons donc recommencé la consultation mais en bornant cette fois le balayage des bandes à l'expression πρᾶξις ἡ ἀπραξία. Et comme précédemment, l'ordinateur n'a rien trouvé.

En dernier ressort, nous avons décidé d'isoler tous les contextes où apparaissait simplement ἡ ἀπραξία, tout en nous jurant bien que c'était l'ultime tentative. La banque de données a enfin reconnu un passage de Grégoire de Naziance où se trouve μεγίστη πρᾶξις ἐστίν ἡ ἀπραξία. Juste Lipse avait, dans sa citation, omis la forme ἐστίν.

Les trois essais réalisés pour Juste Lipse montrent sans doute que notre programme était trop restrictif mais ils ont l'avantage de susciter une réflexion méthodologique: pour qu'un traitement par ordinateur ait quelque chance d'aboutir, il faut qu'il y ait identité parfaite entre la suite de caractères recherchés et ce que l'on trouve sur les bandes magnétiques. En effet, l'ordinateur travaillant selon une logique binaire (oui ou non, vrai ou faux) ne connaît pas l'approximation. Pour l'ordinateur, deux informations sont identiques lorsqu'il n'y a entre elles aucune différence si minime soit elle. Oserais-je dire, m'adressant à des hellénistes, que si les deux données varient d'un iota, l'ordinateur les considérera différentes l'une de l'autre?

La rigidité de la logique informatique et plus souvent l'insuffisance de la programmation contraignent le chercheur à interpréter avec prudence les résultats - et les non-résultats - que produit l'ordinateur. En effet, les variantes graphiques, les accents, les esprits et les divers signes critiques qui figurent dans nos éditions risquent d'empêcher certaines identifications. Prenons deux exemples.

Si je m'intéresse aux occurences de θάλασσα dans plusieurs œuvres et chez divers auteurs, je devrai être attentif au fait que l'éditeur a peut-être opté pour la graphie θάλαττα que l'ordinateur ne reconnaîtra pas si je l'interroge uniquement à partir de θάλασσα.

Les esprits et les accents peuvent fausser les relevés. Ainsi, la recherche du substantif ἄνθρωπος dans la *Politique* d'Aristote conduira à des résultats partiels si l'on demande à la machine de repérer toutes les occurences de A)/NQRWPOS (ἄνθρωπος), A)/NQRWPE (ἄνθρωπε), A)/NQRWPON (ἄνθρωπον), A)N-

QRW/POU (ἀνθρώπου),... car, à deux reprise, Ross, l'éditeur du Stagirite, a estimé qu'il y avait une crase et a écrit le mot avec une coronis. Nous verrons bientôt quelles solutions nous avons adoptées pour résoudre ce problème.

Le programme d'identification de citations est utilisé aussi en papyrologie. En ce domaine, les perspectives qu'offre le TLG avaient été mises en évidence par Th.F. Brunner, dans un article intitulé *Two Papyri of Appian from Dura-Europus*[14].

A l'Université de Liège, le professeur P. Mertens, titulaire du cours de papyrologie, a ressemblé des reproductions et des photos des papyrus littéraires. Or, est-il utile de dire, pour la plupart de ces documents, on ignore quel auteur ou quelle œuvre y est transcrit. La consultation systématique du TLG a déjà permis de reconnaître le texte de plusieurs papyrus.

Les premiers essais d'identification de papyrus littéraires n'ont, à de rares exceptions près, abouti à aucun résultat probant. Les raisons de ces échecs sont évidentes. Aux problèmes liés aux variantes, aux esprits et aux accents s'ajoutent des difficultés spécifiques à la documentation papyrologique: documents fragmentaires, lectures douteuses, voire impossibles, coupes en mots incertaines...

N'ayant ni la capacité, ni la naïveté de tout résoudre à l'aide de l'ordinateur, nous avons tenté de contourner et même de supprimer les principaux obstacles à une meilleure reconnaissance des papyrus. Pour atteindre ce but, nous avons adopté des méthodes de recherches qui font abstraction de certains éléments jugés perturbateurs.

Nous avons d'abord développé une procédure qui élimine du TLG les accents, les esprits, les trémas, les iotas souscrits, les ponctuations et tous les signes critiques, de manière à produire un fichier de travail dans lequel sont repris les seuls caractères alphabétiques et les espaces entre mots. Dans un tel fichier, l'ordinateur repère toutes les occurrences de la suite de lettres ANQRWPOS (ανθρωπος) sans tenir compte des esprits et des accents.

Même si cette approche, moins nuancée, est à l'origine de confusions formelles (l'ordinateur ne distinguera plus la préposition εἰς et le numéral εἷς), elle rend possible le repérage de contextes que la

[14] Dans cet article [GRBS 25 (1984), 171-75], Th.F. Brunner montre comment il a pu identifier deux papyrus littéraires. L'un avait été mal identifié en 1939 et l'autre n'avait jamais été localisé.

logique binaire de l'ordinateur aurait éliminés.

La deuxième méthode de consultation repose aussi sur les fichiers 'alphabétiques' que nous venons de décrire. Elle permet de rechercher non des mots entiers mais bien tous les mots commençant par telle ou telle séquence de lettre. Ainsi, pour repérer tous les emplois du vocable ἄνθρωπος, il suffit d'imposer à l'ordinateur de retenir toutes les formes dont les premières signes sont ANQRWP (ανθρωπ) ou même ANQRW (ανθρω).

Dans ce cas encore, les relevés produits par la machine méritent un examen minutieux; ils comportent parfois nombre de références qui n'ont rien à voir avec l'objet de la recherche. Supposons que l'on s'intéresse à Κύπρις, la déesse de Chypre, et que, pour éviter d'énumérer toutes les formes fléchies de ce mot, on demande à la machine d'extraire toutes les références de termes commençant par KUPRI (κυπρι). Les résultats seront sans doute complets mais ils reprendront des références à des mots tels que Κυπρία, Κυπριακός, κυπρίζω, κυπρινέλαιον, Κύπρινος, ou κυπρῖνος, Κύπριος et κυπρισμός.

La troisième procédure est plus réductrice encore que les précédentes. Dans les fichiers qu'elle exploite, sont éliminés non seulement les signes critiques mais aussi les espaces entre mots. Le texte apparaît alors en *scriptio continua*, longue séquence de caractères *alphabétiques dans laquelle ni les lignes, ni les mots* ne se distinguent les uns des autres, de sorte que le début de l'Iliade, par exemple, sera:

MHNINAEIDEQEAPHLIADEWAXILHOSOULOMENHN.

Le logiciel d'exploitation permettra notamment de connaître l'origine de papyrus pour lesquels la découpe en mots est incertaine.

L'identification de certains papyrus nécessite des logiciels plus élaborés capables d'opérer des recherches multicritères: il s'agit alors de retrouver le ou les passages où apparaissent, dans un contexte plus ou moins étendu, deux ou plusieurs suites de caractères. Cette démarche permet de découvrir l'origine d'un texte dont on ne possède que quelques fragments épars.

Au terme de cette brève description des premiers logiciels que nous avons développés pour exploiter les bandes magnétiques du *Thesaurus Linguae Graecae*, nous voudrions faire encore deux remarques.

Les résultats obtenus sont sans doute loin d'être parfaits puisqu'ils

sont parfois inadéquats. Lors de l'analyse et de l'interprétation des données, il convient de se souvenir que l'ordinateur est une machine servile qui ne fait rien d'autre qu'exécuter d'une manière rigoureuse, objective et logique, mais non intelligente, les tâches qui lui sont confiées. Si le chercheur est convaincu que les limites des méthodes informatiques sont liées à la richesse du contenu des fichiers, il saura que la seule chance d'obtenir des résultats corrects est de bien formuler ses questions, ce qu'il ne pourra faire s'il n'apprend au préalable à bien analyser l'objet de son enquête.

Par ailleurs, toute tentative de consultation d'une banque de données échouera si le chercheur n'a pas une connaissance préalable et précise du contenu exact des fichiers, de la manière dont il sont construits et des principes qui ont présidé à leur constitution. C'est pour cette raison notamment que nous avons décidé, avant même d'exploiter systématiquement le corpus californien, d'intégrer aux données la référence complète de chaque ligne de texte. Cette tâche nous a évidemment demandé beaucoup de temps, mais elle améliore de manière très significative la qualité des résultats que nous obtenons.

En résumé, le TLG tel qu'il existe aujourd'hui mérite notre considération et peut rendre de grands services, mais il devrait être accompagné de programme qui en faciliteraient l'exploitation et rendraient les données immédiatement accessibles aux hellénistes. Dans cette perspective, nous sommes prêts à diffuser nos programmes, malgré leurs limites et leurs imperfections, auprès des chercheurs qui souhaiteraient en disposer.

L.A.S.L.A
Université de Liège

Joseph Denooz