Etablissement d'un programme d'analyse automatique des textes papyrologiques documentaires grecs

Le L.A.S.L.A. a présenté au XIIe Congrès international de Papyrologie, tenu à Ann Arbor en 1968, un programme de traitement automatique des textes papyrologiques grecs. Ce programme a fait l'objet de trois communications : une de M. Evrard, sous le titre : *Exploitation automatique de la documentation papyrologique : description d'un traitement par ordinateur*; une de M. Bodson, traitant de l'application de ce programme à un *Corpus*; j'ai moi-même parlé du *Travail papyrologique par ordinateur et des papyrologues*, j'ai montré les avantages que ceux-ci pouvaient en attendre pour la poursuite de leurs recherches et d'autre part de l'intervention réduite qu'on leur demanderait pour constituer les fichiers conformes aux exigences de la mécanographie.

Un *Choix de Papyrus*,1 réalisé automatiquement suivant le traitement exposé, fut présenté aux Congressistes pour leur fournir un exemple des résultats que l'on pouvait atteindre. Outre MM. Evrard, Bodson et moi-même, M. Bingen, professeur à l'U.L.B., MM. Denooz et Dupont, de l'équipe du L.A.S.L.A., avaient collaboré à cette publication. Elle comportait, outre le texte émendé de vingt-huit documents, un index général des mots et un index des noms propres, une liste de fréquence des mots, des spécimens de reproduction automatique de l'édition princeps

qui avait servi de base à l'édition émendée, des exemples de transcription diplomatique des textes, d'index inverse des lemmes, de concordances à contexte variable en partant d'un groupe de lettres ou d'un mot, de concordances sélectives donnant le contexte de deux mots, le tout obtenu automatiquement en partant d'un seul fichier.

Cette intervention du L.A.S.L.A. au Congrès suscita un vif intérêt; il convenait désormais pour convaincre entièrement les papyrologues de la valeur de ce procédé de travail, de le mettre en oeuvre dans un cadre plus large que celui représenté par le *Choix de Papyrus* et de réaliser automatiquement non seulement l'édition, mais l'analyse d'un groupe important de textes pour en dégager la matière d'une étude. C'est le programme auquel je me suis attaché.

Je songeai d'abord à mettre sur cartes perforées un recueil ancien, épuisé depuis longtemps et dont la réédition aurait sans doute été accueillie avec plaisir, mais la diversité des sujets traités dans les textes n'aurait guère permis de réunir sur chacun d'eux une documentation suffisante pour faire une exploitation vraiment significative; des problèmes d'autorisation et de droit d'auteur auraient d'autre part exigé des démarches qui auraient retardé le début du travail; pour un choix de textes ou pour des extraits, la question ne se posait pas. J'en suis ainsi venu à opter plutôt pour la constitution d'un dossier réunissant tous les textes relatifs à un même sujet : *les ousiai de l'Egypte romaine*, ces grands domaines créés par Auguste pour aider à la mise en valeur de cette nouvelle province dont il s'était personnellement attribué la gestion.

La raison qui justifie mon choix est double : d'abord l'intérêt du travail. Il n'existe sur le sujet aucune étude d'ensemble récente et complète. Celle

que Rostovtzeff lui a consacrée en 1926 dans sa *Social and Economic History of the Roman Empire* ne constitue qu'une partie de chapitre, la dernière mise à jour du texte date de la réédition de 1957;[2] des éditeurs de textes relatifs aux *ousiai* ont bien traité également, dans leur commentaire, de cette catégorie de biens privés,[3] mais cette documentation reste fragmentaire et d'un accès peu aisé. Refaire systématiquement ce travail était donc oeuvre utile.

La seconde justification de mon choix est que depuis plus de dix ans j'ai consacré, dans une large mesure, mon activité à l'étude des *ousiai*[4] et que j'ai déjà constitué et tenu à jour le dossier des textes qui les concernent; les documents destinés à établir le fichier mécanographique pour le travail envisagé étaient ainsi à pied d'oeuvre.

Une seule critique pourrait m'être faite, c'est que, de la sorte, le rassemblement des matériaux n'a pas été le fait de la machine. Le reproche n'est pas grave, car je n'ai fait qu'anticiper de peu la réalité, me plaçant artificiellement dans la situation du papyrologue qui d'ici quelques années, j'espère, quand tous les recueils auront été mis sur bandes ou sur cartes perforées, pourra demander à l'ordinateur de constituer *lui-même* le dossier que je mets aujourd'hui à sa disposition. Cette anticipation n'est d'autre part pas du domaine de l'utopie : pour réaliser les concordances présentées dans le *Choix de Papyrus*, l'ordinateur a déjà effectué une sélection et un regroupement semblables de documents, en puisant dans les textes à sujets variés mis à sa disposition sur cartes perforées.

Mon dossier sur les *ousiai* comprend quelque trois cent cinquante textes, documents entiers ou extraits, comptant environ trente-cinq mille mots. Le travail mécanographique de mise sur cartes perforées a commencé au

début d'avril de cette année, avec la collaboration de M. Evrard pour l'établissement des programmes et de M. Denooz pour la conduite technique du travail. Deux de mes anciens élèves : M. Bolland, licencié en philologie classique, et M. Straus, licencié en histoire ancienne, m'assistent dans les différentes opérations faites sur machines; leur adaptation aux aspects variés de cette activité a été aisée et rapide, et si leur virtuosité n'est pas encore celle d'un professionnel, la vitesse avec laquelle ils perforent dépasse pourtant déjà le rythme de cent cinquante mots à l'heure.

Le travail n'est donc encore qu'à ses débuts; si j'ai pris l'initiative de vous en parler, c'est pour faire le point, du moins en ce qui concerne le L.A.S.L.A., des recherches qui doivent être menées actuellement dans plusieurs centres, si on suit les vues exposées au Congrès d'Ann Arbor; c'est aussi pour vous entretenir de deux problèmes qui intéressent particulièrement les papyrologues : celui de la lemmatisation automatique parce que la solution proposée conditionnera l'importance de leur participation dans l'élaboration du fichier, celui de la codification sémantique parce que la position que nous avons prise dans le *Choix de Papyrus* semble ne pas avoir recueilli leur accord.

En ce qui concerne l'état d'avancement de notre travail, 59 textes, le sixième du dossier, ont été mis sur cartes perforées, il s'agit de 3 textes des *P. Aberdeen* et de 56 textes ou extraits de textes des tomes publiés des *Berliner Griechische Urkunden (B.G.U.)*; le fichier actuel comprend 4.961 cartes à raison d'une carte par forme.

La perforation de la forme en version non accentuée a été effectuée mécaniquement; elle est nécessaire pour le traitement mécanographique classique que nous comptons appliquer pour l'étude de ce premier groupe de

textes, mais le L.A.S.L.A. a un programme d'ordinateur qui permet le traitement direct des cartes accentuées, nous l'utiliserons par la suite pour éviter l'alourdissement du fichier que crée le doublement des cartes. La référenciation, elle aussi automatique, est terminée; la lemmatisation est en cours, elle se fait suivant un programme établi par M. Evrard qui comporte trois opérations, une manuelle, deux automatiques.

Un tri automatique classe d'abord toutes les formes dans l'ordre alphabétique et regroupe ainsi dans des séries homogènes toutes les formes semblables de fréquence 2 au moins.

On traite ensuite les formes de fréquence 1 et 2 dont le lemme est distinct de la forme, en perforant ce lemme manuellement, dans l'ordre du fichier, sur toutes les cartes portant des formes de fréquence 1 et sur la première carte *seulement* de chaque série de formes de fréquence 2 au moins; pour les autres cartes de la série ce lemme est en effet immédiatement répété, automatiquement, par duplication.

Il ne reste plus à lemmatiser que les formes dont le lemme est identique à la forme ou n'en diffère que par un accent grave de mot oxyton qu'il faut rectifier en accent aigu. L'ordinateur s'en charge : il découvre ces formes dans le fichier qui lui est confié, en raison de l'absence de perforations, sur les cartes, dans la zone réservée au lemme; il y perfore automatiquement un lemme identique à la forme avec rectification éventuelle de l'accent.

Pour préciser l'importance de la réduction de l'intervention manuelle dans la lemmatisation ainsi réalisée, j'ai analysé en détail un échantillon de 2.076 cartes représentant 41,8 % du dossier traité. L'échantillon correspond à l'ensemble des formes commençant par les 5 premières lettres de

l'alphabet grec. Ces formes se répartissent comme suit :

Lemmatisation manuelle

Formes dont le lemme n'est pas identique à la forme

Nombre de formes de fréquence 1 :	155	soit 7,7% du nombre des formes
Nombre de lemmes à perforer :	155	
Nombre de formes de fréquence 2 ou plus :	1.378	soit 66,3% du nombre de formes
Nombre de groupes de formes identiques :	194	soit 9% du nombre de formes
Nombre de lemmes à perforer :	194	

Lemmatisation automatique

Formes dont le lemme est identique à la forme

Nombre de formes de fréquences diverses :	543	soit 26% du nombre de formes

Répartition

Mots indéclinables et en quantité moindre mots déclinables et verbes :	266	soit 13% du nombre de formes
Groupes de lettres provenant de mots incomplets fréquents dans les papyrus en raison des lacunes et chiffres figurés par des lettres :	277	soit 13% du nombre de formes

Récapitulation

du nombre de lemmes à perforer manuellement pour les 2.076 formes :

194 + 155 = 349 soit 16,7% du nombre de formes

Gain de travail

sur une perforation des lemmes entièrement manuelle : 83,3% du nombre de formes

sur une lemmatisation automatique limitée aux formes semblables aux lemmes : 57,3% du nombre de formes

Ce gain de travail n'est pas le seul profit tiré de ce programme de lemmatisation; il faut y ajouter la garantie que la perforation se fera sans erreur, de façon certaine là où elle est entièrement automatique, et pour toutes les séries de formes de fréquence 2 pour autant que le lemme initial ait été perforé correctement.

Cette étude comparative du nombre de lemmes et de formes souligne la pauvreté relative de la langue et de la grammaire des papyrus documentaires, elle résulte en grande part d'une rédaction des textes figée dans les formules d'une diplomatique imposée.[5]

Cette constatation m'a amené à envisager une lemmatisation automatique des formes non semblables au lemme en mettant en mémoire en ordinateur ces formes et les lemmes correspondants.

En admettant que le pourcentage de 16,7% seulement de types de formes différents, relevé dans notre échantillon de 2.076 formes, se maintienne pour l'ensemble de notre travail, il suffirait de mettre en mémoire 5.880 formes et 5.880 lemmes pour lemmatiser automatiquement les 35.000 formes de l'ensemble de nos textes, les lemmes semblables aux formes, perforés automatiquement, n'entrant pas en ligne de compte pour la mise en mémoire.

Ce programme n'a pourtant pas été retenu car il ne conviendrait que pour les papyrus documentaires et une expérimentation assez longue serait nécessaire pour s'assurer que les chiffres avancés pour notre série de textes ne seraient pas dépassés pour des papyrus documentaires d'un autre type. En plus, ce programme risquait, avant sa mise au point, d'être largement dépassé, car le L.A.S.L.A. étudie dès maintenant l'adaptation au grec du

programme d'analyse automatique du latin qu'il possède déjà comme on le sait.[6] Cette adaptation est d'autant plus aisée que les deux langues appartiennent au même type morphologique. Ce programme couvrira évidemment l'analyse de tous les textes grecs et, partant, leur lemmatisation, si bien que non seulement les papyrologues, mais tous les hellénistes, quelle que soit leur spécialité, pourront en bénéficier.

La papyrologie devra avoir sa version propre de ce programme, car la présence dans les papyrus de nombreux noms propres particuliers à l'Egypte, exigera un dictionnaire spécial, par contre son programme d'analyse des désinences n'épuisera certainement pas toutes les possibilités prévues. L'examen de l'usage fait des cas et des modes pour les 1.533 mots variables figurant dans l'échantillon de 2.076 formes qui a servi à nos autres investigations, montre en effet une prédominance remarquable de deux cas : le nominatif et le génitif, fait naturel pour ce dernier cas en raison de l'abondance des patronymiques au génitif.

Ces emplois se répartissent de la façon suivante pour 1.394 mots déclinables :

Nominatif :	645 emplois,	soit 46% des formes analysées
Génitif :	638 emplois,	soit 45,5% des formes analysées
Accusatif :	66 emplois,	soit 5% des formes analysées
Datif :	45 emplois,	soit 3,5% des formes analysées

Pour les formes verbales, au nombre de 139, la distribution montre également une prédominance de deux modes : l'indicatif et le participe.

Indicatif :	64 emplois,	soit 46% des formes analysées

Participe : 54 emplois, soit 36,5% des formes analysées
Infinitif : 15 emplois, soit 11,5% des formes analysées
Impératif : 6 emplois, soit 5% des formes analysées
Les autres modes ne sont pas apparus.

La lemmatisation, dans le *Choix de Papyrus*, était accompagnée d'une codification sémantique destinée uniquement à distinguer les noms propres; dans notre travail actuel nous hésitons à l'étendre davantage et notamment à reprendre toutes les catégories de mots des index spéciaux de nos recueils de papyrus, car, en fait, pour des textes mis sur cartes perforées et destinés à être traités mécanographiquement, l'index va être utilisé d'une façon toute nouvelle qui n'implique plus une spécialisation compliquée. Il peut être réduit à un index général des mots où le papyrologue relèvera ceux qui sont en rapport avec le sujet qu'il traite et pour lequel il veut réunir une documentation; la demande à l'ordinateur de concordances sélectives portant sur deux ou plusieurs de ces mots se substituera alors à l'investigation, à travers les rubriques des index spécialisés, pour obtenir des références, et les textes ainsi repérés pourront même être fournis, automatiquement, recopiés et classés, avec leur référence.

Ce n'est que si les papyrologues désirent – et ils devront le dire – que, même pour une édition fondée sur un fichier mécanographique, il y ait des index spécialisés pour qu'on puisse continuer à travailler manuellement, qu'une codification sémantique spécialisée se justifierait. Techniquement la perforation du code ne présentera pas de difficultés, elle pourra accompagner la perforation du lemme et bénéficier du même traitement que lui, mais elle devra être préparée manuellement par un classement préalable de tous les lemmes dont on décidera la codification, car là aussi il y aura une décision à prendre étant donné la variété de composition des index spécialisés qu'on trouve dans les recueils.[7]

Après la codification sémantique réduite, la dernière opération à effectuer avant l'exploitation du fichier pour une édition des textes ou pour une analyse, est la transcription sur cartes perforées des corrections et des restitutions nouvelles introduites dans les textes ou proposées depuis l'édition princeps; l'ordinateur reprendra celles qui auront été retenues nous permettant de travailler toujours sur l'état le plus récent des documents.

Arrivés à cet état du fichier pour notre premier groupe de 59 textes, nous comptons, avant de poursuivre la perforation, procéder à un essai d'analyse, de façon à vérifier si notre programme de perforation et nos procédés mécaniques d'investigation sont suffisants et bien adaptés à notre projet d'exploitation.

Outre l'édition des textes émendés, nous comptons entreprendre l'établissement d'un tableau général des sources, par recueils, puis par *ousiai*, d'une liste alphabétique des bénéficiaires, puis selon l'ordre chronologique des dotations; de simples tris effectués sur les lemmes, les références, les dates y suffiront sans doute.

A l'aide d'autres tris et de concordances, nous poursuivrons l'analyse statistique des données en nous plaçant à différents points de vue : répartition géographique et composition des domaines, étude de la nature des parcelles, des surfaces, des cultures, des modes de gestion et d'exploitation, de l'évolution du statut, enfin établissement d'une prosopographie des personnes mentionnées dans les textes, dans l'ordre alphabétique, puis par fonctions.

Des problèmes surgiront sans doute, nous imposant une mise au point des procédés envisagés ou le recours à des techniques nouvelles. Ces difficultés

sont précisément la justification et l'intérêt de la recherche entreprise; j'espère que les moyens mis à ma disposition par le *Laboratoire d'Analyse Statistique des Langues Anciennes* et par le *Fonds National de la Recherche Scientifique* me permettront, avec le concours de mes collaborateurs, de la mener à bien et dans un temps dont la longueur ne sera pas excessive.

Université de Liège
(L.A.S.L.A.)

Alfred TOMSIN

NOTES

1. *J. Bingen, A. Tomsin, A. Bodson, J. Denooz, J. C. Dupont, Et. Evrard, Choix de papyrus grecs*, Université de Liège, Faculté de Philosophie et Lettres, Laboratoire d'Analyse Statistique des Langues Anciennes sous la direction du Professeur Louis Delatte, Liège, 1968.

2. *M. Rostovtzeff, The Social and Economic History of the Roman Empire*, 2 éd. par *P. M. Fraser,* Oxford, 1957, t. 2, pp. 669 et ss.

3. par ex. *P. Collart, Les papyrus Bouriant*, Paris, 1926, pp. 160 et ss.; *Ture Kalén, Berliner Leihgabe Griechische Papyri*, Upsala, 1932, pp. 67 et ss.

4. par ex. *Un document de comptabilité fiscale, P. Lond. III, 900,* p. 89, Bulletin de l'Académie royale de Belgique, (Classe des Lettres), Bruxelles, 1954, t. 40, pp. 91 et ss.; *Notes sur les ousiai de l'époque romaine*, Studi in onore di Aristide Calderini e Roberto Paribeni, Milan, 1957, t. II, pp. 210 et ss.

5. sur la diplomatique des reçus de taxe cf. *U. Wilcken, Griechische Ostraka aus Aegypten und Nubien*, Leipzig, 1899, t. 1, pp. 94 et ss.; *C. Préaux, L'emploi des temps de διαγράφω dans les reçus de banque thébains du Ier siècle après J.-C., Chronique d'Egypte*, t. 30, Bruxelles, 1955, pp. 375 et ss.

6. *L. Delatte, Programme d'analyse morphologique et syntaxique du latin, Revue*, t. 1, Liège, 1965, pp. 57 et ss.

7. On comparera par ex. *P. Aberd, P. Berl. Leihg., P. Leit, P. Strasb. (Papyrus grecs*, 1963).