TECHNIQUES ET METHODES

par
J. DENOOZ,
Chef de Travaux, L.A.S.L.A.
Université de Liège

Extrait de la Revue (R.E.L.O.)
XIV, 1 à 4, 1978. C.I.P.L. - Université de Liège - Tous droits réservés.

PRESENTATION DU L.A.S.L.A.

Le Laboratoire d'Analyse statistique des Langues anciennes a été créé pour apporter une solution à des problèmes que la philologie traditionnelle est impuissante à résoudre : je veux parler des problèmes d'authenticité et de chronologie relative des oeuvres d'un auteur. En outre, nous croyons que les éléments recueillis par des analyses formelles permettront de déceler les mécanismes profonds de la création littéraire à tous les niveaux, même les plus inconscients.

Pour atteindre ces objectifs, il faut se fonder sur l'étude statistique des mots d'un texte aux points de vue phonétique, lexicologique, syntaxique et stylistique. Nous avons donc résolu de faire une analyse morphologique et syntaxique complète de chaque forme ainsi qu'une lemmatisation, c'est-à-dire le renvoi du mot-forme au mot tel qu'il figure au dictionnaire. Une telle recherche suppose l'établissement de fichiers extrêmement complets dont la réalisation et l'exploitation exigent le recours aux machines mécanographiques et à l'ordinateur. L'emploi de ces machines entraîne l'utilisation

Extrait de la Revue (R.E.L.O.)
XIV, 1 à 4, 1978. C.I.P.L. - Université de Liège - Tous droits réservés.

d'un langage simplifié qui permet de condenser certaines informations sous forme de codes.

Je commencerai par décrire brièvement la carte mécanographique (annexe 1).

Dans une carte mécanographique, les informations sont représentées sous forme de perforations. La carte se divise verticalement en 80 colonnes numérotées de 1 à 80 et horizontalement en 12 niveaux. Dans chaque colonne, on peut perforer un caractère d'imprimerie numérique ou alphabétique. La signification d'une perforation dépend du niveau où elle se situe. La position des colonnes et des niveaux est indiquée, sur les cartes de modèle standard, par des chiffres imprimés, sauf pour les niveaux 12 et 11 situés dans la partie supérieure de la carte : cette partie est souvent utilisée pour l'impression d'autres indications, notamment pour la traduction en clair du contenu des colonnes.

Une perforation à l'un des trois premiers niveaux (12, 11, zéro) est appelée perforation de zoning, une perforation à l'un des niveaux de 0 (zéro) à 9 est appelée perforation numérique ou de digit. La perforation zéro est mixte puisqu'elle peut jouer le rôle de zoning pour les signes alphabétiques et le rôle de digit pour les signes numériques.

Les chiffres sont représentés par une des dix perforations numériques. Quant aux caractères alphabétiques, ils exigent la présence dans une même colonne d'une perforation de zoning (12, 11 ou 0) et d'une perforation numérique (1 à 9).

Les perforations numériques et alphabétiques peuvent être utilisées soit avec leur valeur naturelle, soit avec des valeurs symboliques. Ainsi, AD peut être la préposition écrite en clair mais dans d'autres colonnes, AD sera le code de subordination de l'ablatif absolu.

Cela dit, nous établissons, à propos de chaque mot du texte, une carte qui comporte : la forme, le lemme, la référence, l'analyse morphologique et syntaxique, éventuellement, des indications relatives à la tradition manuscrite et un code de ponctuation.
Ces renseignements occupent une place fixe dans la carte.

Extrait de la Revue (R.E.L.O.)
XIV, 1 à 4, 1978. C.I.P.L. - Université de Liège - Tous droits réservés.

Voici un aperçu général du dessin de la carte (Annexe 3 - 2e carte).

1	Code-carte
2-17	Lemme
18	Indice de lemme
19-38	Forme du texte
39-40	Code d'oeuvre Il s'agit de deux perforations permettant d'identifier l'oeuvre à laquelle le mot appartient.
41-53, 64,68-75	Référence du mot
41-43	Numéro du chapitre (001, 002 etc.)
44-47	Numéro du paragraphe ou du vers (0001, 0002 etc.)
48-50	Numéro d'ordre du mot dans le paragraphe ou dans le vers (001, 002 etc.)
51-53	Numéro d'ordre du mot dans la phrase (001, 002 etc.)
64	Code du rôle pour les oeuvres dramatiques
68-70	Indication éventuelle de la subdivision en livres
71-75	Numéro d'ordre du mot dans l'oeuvre (00001, 00002 etc.)
54-63	Analyse complète de la forme
65	Indications relatives à la tradition manuscrite
66	Ponctuation du texte
67	Colonne libre
76-80	Numéro d'ordre du mot dans l'index.

Il me faut maintenant, à propos d'un certain nombre d'informations contenues dans la carte, préciser les conventions que nous avons adoptées.

1. Nous appelons forme le mot tel qu'il figure dans le texte de l'édition que nous avons choisie. Nous reproduisons le texte de cette édition aussi fidèlement que possible, à quelques rares exceptions près sur lesquelles je n'insisterai pas ici.

2. Par lemme, nous entendons le mot tel qu'il figure au dictionnaire. Nous utilisons comme ouvrage de référence le *Lexicon Totius Latinitatis* de FORCELLINI. Ce dictionnaire nous a paru être l'instrument de travail le mieux adapté à nos recherches.

En ce qui concerne le lemme, nous avons mis au point un certain nombre de conventions qui sont indispensables pour lire un index et que le lecteur trouvera en annexe 4.

Je décrirai maintenant notre système d'analyse morphologique.
Chaque forme fait l'objet d'une analyse qui est perforée dans les colonnes 54-63 (annexes 5 et 6).

Colonne 54

On note, dans cette colonne, la catégorie grammaticale à laquelle le mot appartient. 1 substantif, 2 adjectif ...

Colonne 55

Selon la manière dont le mot a été analysé en colonne 54, on précise en colonne 55 :
a) la sous-catégorie (classe, conjugaison, etc.);
b) le degré;
c) la voix.

Pour les substantifs, un chiffre indique la déclinaison à laquelle le substantif appartient.

Pour les adjectifs et les verbes nous avons utilisé, pour économiser de la place, un tableau à deux entrées.

Pour les adjectifs, la codification indique à la fois la classe de l'adjectif et son degré.

1 adjectif en -Vs (1ère classe)
A adjectif 1ère classe au comparatif
J adjectif 1ère classe au superlatif

Pour les verbes on indique en colonne 55 à la fois la conjugaison et la voix.

1 1ère conjugaison voix active
A 1ère conjugaison voix passive
J 1ère conjugaison déponent

La colonne 56 est réservée aux indications suivantes :

1) le cas et le nombre pour les formes déclinées;
2) la personne et le nombre pour les formes conjuguées;
3) le cas régi pour les prépositions.

Ex. A pour un subst. représente le nominatif singulier
A pour un verbe représente la 1ère pers. du singulier

Colonne 57

1) Pour les verbes, cette colonne est réservée à l'indication du mode.
2) Pour les mots introduisant un verbe de proposition subordonnée (adjectifs-pronoms, adverbes relatifs et interrogatifs, conjonctions de subordination), on note le mode du verbe régi.

Colonne 58

1) Pour les verbes, cette colonne contient l'indication du temps. Il va de soi que l'adjectif verbal, le gérondif et le supin ne portent aucune indication dans cette colonne.
2) Pour les mots introduisant un verbe de proposition subordonnée, on note

le temps du verbe régi.

Colonne 59

Les cartes de verbes reçoivent l'indication de la fonction :

Perforation 12 verbe de proposition indépendante ou principale
Perforation 11 verbe de proposition subordonnée

La colonne 61 est réservée à l'indication du genre pour les adjectifs, les numéraux, les adjectifs-pronoms et les formes déclinées du verbe.

Nous prenons en considération non pas le genre que le mot a effectivement dans le texte, mais le genre formel en dehors de tout contexte. Cette façon de procéder nous permet de réaliser aisément la mise en ordre des formes dans l'index.

Dès lors, nous distinguons les six catégories suivantes :

1 genre commun (masculin, féminin et neutre)
2 féminin
3 masculin et féminin
4 masculin
5 masculin et neutre
6 neutre

Ainsi, *bonum*, à l'accusatif, sera codifié 5, tandis que le nominatif *bonum* recevra le code 6.

Colonnes 62-63

Ces colonnes reçoivent une indication uniquement dans le cas des verbes subordonnés. Nous avons symbolisé chaque subordonnant par un code formé de 2 caractères alphabétiques. Ainsi, LN indique que l'on a affaire à un verbe dépendant du pronom relatif QVI.

La liste des subordonnants accompagnés de leur code se trouve en annexe 6. Cette liste n'est certes pas exhaustive. Lors de chaque étude de texte, nous sommes amenés à y faire des additions. C'est pourquoi, nous n'avons pas utilisé jusqu'à présent toutes les lettres de l'alphabet dans la seconde colonne.

Colonne 65

Les indications relatives à la tradition manuscrite sont codifiées en colonne 65 : par exemple, 5 représente une crux.

Colonne 66

La colonne 66 contient des renseignements qui permettent à l'ordinateur de calculer automatiquement la référence de chaque forme du texte.

Le code utilisé est alphanumérique; les lettres correspondent aux cas où deux renseignements relatifs à la ponctuation doivent être donnés simultanément.

11 fin de chapitre et du §
K fin du chapitre, du § et de la phrase etc.

Après avoir exposé les conventions relatives à l'enregistrement des textes latins sur cartes mécanographiques, je voudrais donner un bref aperçu historique qui montrera comment, progressivement, le Laboratoire s'est développé.

Au début, nous n'avions que des machines classiques : une perforatrice, une trieuse, une petite calculatrice et un système d'impression automatique.

Lorsque la dactylo avait perforé dans la carte la forme du mot telle qu'elle apparaît dans le texte (annexe 2, carte 1) et lorsque les cartes avaient été référenciées sur la calculatrice (annexe 2, carte 2), chaque membre de l'équipe prenait une partie du fichier à sa table de travail et, à la main, écrivait sur chaque carte, dans des emplacements prévus à l'avance, le lemme (annexe 3, carte 3) et l'analyse complète du mot (annexe 3, carte 4).

On se rend compte aisément de la somme de travail que représentaient ces analyses. Elles exigeaient une attention sans défaillance jusque dans les plus petits détails; elles ne souffraient pas la moindre approximation. C'est la raison pour laquelle un philologue ne pouvait guère analyser plus de deux cents à deux cent cinquante mots par jour. Au surplus, nous nous étions astreints à de nombreuses vérifications, vérifications de la codification de l'analyse et de la perforation de cette analyse sur les cartes.

De 1961 à 1965, soit dans un espace de quatre ans, nous avons publié cinq volumes consacrés aux Dialogues de Sénèque. Pendant ces années, désespérés par la lenteur, la lourdeur de l'analyse, nous avons cherché des moyens d'accélérer le travail en l'automatisant au maximum. Une telle automatisation

supposait l'acquisition de moyens techniques plus puissants que les simples machines classiques, c'est-à-dire l'emploi d'un ordinateur. Nous avons obtenu du F.N.R.S. un petit ordinateur scientifique IBM 1620. Cet ordinateur est arrivé en janvier 1965. Après avoir suivi des cours de programmation et nous être familiarisés avec le maniement de la machine, nous avons réalisé en huit mois un programme d'analyse automatique de la morphologie du latin, et, plus récemment, un programme de lemmatisation du grec.

Peut-être faut-il rappeler brièvement ce qu'est un ordinateur.

Un ordinateur est, à l'origine, une simple machine comptable capable, d'une part, d'effectuer les opérations arithmétiques traditionnelles et, d'autre part, de procéder à des comparaisons entre des valeurs qu'elle a emmagasinées dans une mémoire.

A partir de cette conception de base, tout le progrès a été consacré à augmenter la puissance de la machine dans le domaine de la mémoire et dans le domaine de la rapidité des opérations. Pour augmenter la rapidité, il a été nécessaire de réduire la taille des éléments constitutifs de la machine où la seule longueur des câbles de connexion joue un rôle primordial dans le déroulement des opérations.

Dans ce but, l'effort des constructeurs a porté sur la miniaturisation. En effet, les premières machines, vers les années 50, étaient des monstres à lampes qui consommaient autant d'électricité pour fonctionner que pour se refroidir (240 kw par heure pour le Gamma Bull 60). La deuxième

génération, qui est apparue vers 1960 était équipée de transistors et avait une taille nettement moindre. Actuellement la troisième génération d'ordinateurs est équipée de circuits intégrés analogues à ceux qui équipent les montres digitales, où, sur un centimètre carré, sont groupés 30 à 40.000 transistors.

L'ordinateur du Centre de Calcul de l'Université de Liège est une machine de cette 3e génération (IBM 370/158). Il est très puissant; en fait, c'est une des machines les plus puissantes du monde. Sa mémoire centrale a une capacité de plus d'un million cinq cent mille caractères. Mais cette mémoire ne constitue qu'une faible partie de sa capacité totale. En effet, le 370 est doté de mémoires périphériques gigantesques sous forme de disques et de bandes magnétiques qui représentent plus de quarante milliards de caractères. Pour faire comprendre la puissance de cet instrument, on peut dire que les mémoires globales du 370 équivalent à un bibliothèque de 20.000 ouvrages de 1.000 pages chacun. En d'autres termes, un lecteur qui consacrerait 10 heures par jour, 365 jours par an à lire cette bibliothèque devrait vivre 456 ans et encore 200 jours.
L'ordinateur, quant à lui, lit cette bibliothèque en moins de 10 minutes.

Cette machine ne travaille pas uniquement dans le lieu où elle est installée. Elle possède des ramifications à travers tous les bâtiments universitaires. Ces ramifications sont en fait des terminaux travaillant simultanément. Un terminal est un appareil relié à un ordinateur central par une ligne téléphonique spéciale qui permet le transfert d'informations à grande vitesse et avec un maximum de sécurité. On apprendra avec étonnement qu'à l'Université de Liège, il y a 55 terminaux qui fonctionnent simultanément.

Le L.A.S.L.A. possède actuellement 4 terminaux. Il s'agit tout d'abord d'un terminal lourd, c'est-à-dire d'une machine qui possède un lecteur de cartes qui traite les cartes à la cadence de 600 à la minute, un perforateur de cartes (9.600 perforations à la minute ou 120 cartes) et une imprimante semi-rapide à 600 lignes/minute. On ne se rend pas compte de ce que signifie "600 lignes à la minute". Cela signifie qu'un volume de 1.000 pages, à raison de 60 lignes par page, peut être imprimé en moins de 2 heures. Par ailleurs, ce terminal possède une machine à écrire qui permet de dialoguer avec l'ordinateur lui-même.

Outre cet équipement, le L.A.S.L.A. possède 3 terminaux légers. L'un est constitué par une simple machine à écrire qui sert essentiellement au traitement du grec, le deuxième est un écran cathodique et enfin, le troisième est un terminal acoustique qui permet d'entrer directement en communication avec l'ordinateur, par l'intermédiaire du téléphone.
Ce dernier appareil préfigure l'avenir puisque dans un délai de trois ans maximum, on pourra parler à l'ordinateur et il répondra oralement.

Je vais décrire maintenant le programme d'analyse automatique du latin : le but du système est de faire produire par l'ordinateur la lemmatisation et l'analyse codée de toutes les formes d'un texte et, pour les formes ambiguës, tous les lemmes et toutes les analyses qui sont formellement possibles : soit, par exemple, la forme QUAE qui reçoit douze analyses différentes :

QUAE	QUI pronom relatif	nom. fém. sing.
		nom. fém. plur.
		nom. neutre plur.
		acc. neutre plur.
	QUIS adjectif-pronom interrogatif	nom. fém. sing.
		nom. fém. plur.
		nom. neutre plur.
		acc. neutre plur.
	QUIS adjectif-pronom indéfini	nom. fém. sing.
		nom. fém. plur.
		nom. neutre plur.
		acc. neutre plur.

Pour effectuer ces opérations d'analyse et de lemmatisation, l'ordinateur fait subir au texte cinq traitements successifs.

1. Dans une première phase, le texte entier, perforé sur cartes mécanographiques, est enregistré sur disques magnétiques. Chaque mot reçoit à ce

moment un numéro d'ordre séquentiel qui permettra de le reclasser dans l'ordre du texte à n'importe quel moment. En outre, chaque forme est éventuellement suivie d'un code de ponctuation qui servira au calcul automatique de la référence.

2. La deuxième phase du travail consiste à classer les formes du texte en ordre alphabétique. Cette opération a pour but de diminuer la durée de la phase suivante. En effet, partant de l'ordre alphabétique intégral, l'ordinateur analyse la première forme qu'il rencontre et reproduit ensuite les analyses autant de fois qu'il y a de formes identiques dans le texte. Ainsi, si la forme EST se trouve cent fois dans un texte, l'ordinateur procède une seule fois à la recherche de ses analyses puis il les recopie cent fois à très grande vitesse sans consultation nouvelle du lexique. L'économie de temps que l'on réalise ainsi est considérable puisque, pour un texte de 10.000 mots, ce procédé réduit presque de moitié le nombre de formes à analyser.

3. Au cours de la troisième phase s'effectuent la lemmatisation et l'analyse.

4. La quatrième phase consiste à reclasser les formes en ordre du texte à l'aide du numéro séquentiel qui leur a été attribué lors de la première phase.

5. Finalement, dans une cinquième phase, chaque analyse est imprimée et est enregistrée sur bande magnétique. C'est également au cours de cette phase que chaque mot du texte reçoit sa référence complète.

Les deux premières phases du système ainsi que les deux dernières ont un rôle secondaire puisqu'elles consistent, d'une part, à classer les mots d'un texte dans l'ordre le mieux adapté au traitement et, d'autre part, à présenter les résultats d'une manière telle que le philologue puisse facilement les examiner. En revanche, la troisième phase occupe une place prépondérante puisque c'est à ce moment que chaque forme est lemmatisée et analysée. Pour effectuer ces opérations, l'ordinateur doit au préalable, avoir reçu tous les éléments qui lui permettront d'agencer ses réponses.

Notre système d'analyse se fonde sur la distinction entre la partie du mot qui reste constante à travers toute la flexion et la lettre ou le groupe de lettres susceptibles d'affecter tous les lemmes appartenant à une même flexion. Nous désignons ces deux parties respectivement par les termes *radical* et *désinence* en donnant à ces mots une signification purement conventionnelle sans aucun lien avec la grammaire. Si l'on prend, par exemple, la forme AMABAT, on dira que AM- est le radical de AMO, AMARE et que -ABAT est la désinence de la troisième personne du singulier de l'indicatif imparfait actif des verbes de la première conjugaison. C'est pourquoi il est nécessaire d'enregistrer dans les mémoires de l'ordinateur, deux Thesaurus, c'est-à-dire un lexique et une table des désinences. Le programme les consultera constamment pour identifier les formes du texte. En plus des données linguistiques, ces fichiers contiennent des informations d'ordre technique qui guident le travail de l'ordinateur. On trouvera en annexe 7 et 8, des échantillons du lexique et de la table des désinences. En voici une brève description.

Le lexique doit contenir les éléments qui permettront à l'ordinateur de

reconnaître un mot quelconque de la langue, de le rattacher à un lemme et enfin de déterminer pour ce mot, la ou les analyses possibles.

Dans le dictionnaire de la machine, une unité lexicale est composée du radical d'un mot et de son lemme Elle intervient dans l'analyse de toutes les formes de ce mot. Dès lors, avec un lexique relativement restreint (environ 16.000 unités), l'ordinateur peut lemmatiser et analyser plus de 90 pourcents des formes de n'importe quel texte latin.

Le principal intérêt du processus de décomposition réside dans l'économie appréciable d'unités lexicales qu'il permet de réaliser. Il existe toutefois certains mots pour lesquels cette méthode est soit inutile, soit inapplicable.

Ce sont, par exemple, les mots invariables qui sont représentés dans la langue par une forme unique : les conjonctions, les prépositions et certains adverbes.

Ce sont aussi toutes les formes anomales. Elles sont traitées de la même manière que les mots invariables : chaque unité lexicale de forme irrégulière conduit à une seule analyse : c'est le cas des diverses formes du verbe ESSE.

Enfin, pour terminer la description du lexique, il convient de noter que pour des raisons d'économie, les verbes composés ne font pas l'objet d'unités lexicales séparées. En effet, leur analyse et leur lemmatisation sont réalisées à partir des verbes simples et d'une table des préverbes mise en mémoire.

Un mot maintenant sur *la table des désinences.*

Après avoir décomposé une forme en radical et terminaison et après avoir découvert une unité lexicale compatible avec le radical, l'ordinateur consulte la table des désinences pour y découvrir l'analyse du mot.

Dès lors, la table des désinences contient obligatoirement toutes les terminaisons des mots à flexion et, parallèlement, toutes les analyses morphologiques possibles; c'est pour cette raison que les désinences formellement identiques figurent dans la table autant de fois qu'elles donnent lieu à des analyses différentes. En procédant de la sorte, nous avons dénombré quelque 3.100 désinences différentes. (cf. annexe 8).

Les trois premières colonnes sont rigoureusement identiques aux colonnes 3 à 5 des unités lexicales. Cette identité est fondamentale dans le processus d'analyse puisque c'est elle qui permettra de préciser la série de désinences qui est compatible avec chaque radical.

La désinence occupe les positions 6 à 14, les colonnes 17 à 22 contiennent l'analyse morphologique complète de la désinence. Ces colonnes correspondent aux colonnes 54 à 58 de notre code d'analyse.

Tels sont les deux thesaurus qui permettent d'organiser le programme proprement dit de l'analyse automatique du latin. C'est ce programme que je voudrais décrire maintenant.

J'ai dit précédemment que, pour produire le lemme et l'analyse d'un mot latin, le programme exécute essentiellement deux opérations :

1- la décomposition de la forme,
2- la consultation du lexique et de la table des désinences.

La décomposition de la forme s'effectue à partir de la fin du mot en remontant progressivement lettre par lettre vers le début. Chaque nouvelle décomposition est suivie d'une consultation de la table des désinences.

	natura	-e	rien	} de NATURA
naturae	natur-	ae	Génitif singulier	
			Datif singulier	
			Nominatif pluriel	
			Vocatif pluriel	
	natu	-rae	rien	
	nat-	urae	nascor -participe futur de nasci	
	nat-	urae	no -participe futur de nare	

Le lecteur trouvera en annexe 9 un échantillon d'analyse corrigée par le philologue. A ce stade, en effet, le philologue reprend les listings d'analyse et il indique par un signe quelle est la bonne analyse. Comme les différentes analyses sont numérotées, il suffit à la dactylo d'indiquer à la machine le numéro de l'analyse correcte.

On remarquera l'extraordinaire degré d'amphibologie de la langue puisque chaque forme est susceptible de recevoir en moyenne trois analyses. On le comprend d'autant mieux si l'on se rappelle que les formes QVAE et NATURAE, peuvent recevoir chacune 12 analyses.

Pour terminer cette partie de l'exposé, je voudrais dire un mot de l'efficacité du programme. Notre lexique comporte environ 16.000 unités, ce qui représente quelque 8.500 mots différents. Malgré cette faible étendue du dictionnaire, la machine parvient à analyser plus de 90 pourcents de n'importe quel texte latin, qu'il s'agisse de prose ou de poésie, de latin classique, de latin de la Renaissance ou de l'époque moderne : c'est ainsi que nous avons analysé les Regulae de Descartes et le Tractatus Politicus de Spinoza avec le même succès.

En ce qui concerne la rapidité d'exécution, un ordinateur de la 2e génération analysait 10.000 mots en une nuit. L'ordinateur que nous utilisons actuellement traite ces mêmes 10.000 mots en moins de 12 minutes.

Par ailleurs, il est très important de noter que, lorsque l'analyse a été corrigée par le philologue, tout le reste du travail, c'est-à-dire la constitution de l'Index, des listes de fréquence, des concordances, des relevés statistiques, est entièrement automatique.

Un Laboratoire qui s'intitulait des Langues anciennes se devait de traiter aussi les textes grecs.

Le traitement du grec par machines mécanographiques et par ordinateur présente certaines difficultés liées, d'une part, à l'alphabet et, d'autre part, à l'existence d'accents, d'esprits et d'iota souscrit etc... En ce qui concerne l'alphabet, les codes de perforations du latin pouvaient difficilement répondre à nos besoins en raison du fait qu'ils ne suivaient pas l'ordre alphabétique de la langue grecque. C'est ainsi que γ se plaçait non pas entre β et δ mais avant la lettre η. C'est pourquoi nous avons adopté une codification des lettres grecques qui répond parfaitement à leur succession dans l'alphabet.

Le problème des esprits et des accents était plus difficile à résoudre. Cependant, il était impossible de les ignorer, comme le font les philologues américains, parce qu'ils sont porteurs d'informations linguistiques.
Nous avons donc établi un code de perforation capable de représenter tous les graphèmes du grec ancien. Ce code comporte, outre les caractères alphabétiques, les divers accents et esprits, l'iota souscrit, le tréma et éventuellement le digamma.

Parallèlement à ces difficultés de perforation liées à l'enregistrement, nous avons dû rechercher une solution au problème de l'impression des résultats : les imprimantes d'ordinateur ne possèdent pas de chaîne de caractères pour le grec ancien; nous avons donc demandé à la firme IBM d'adapter pour nos besoins deux systèmes 870 qui se composent chacun d'une perforatrice connectée à une machine à écrire automatique, pour perforer et imprimer le grec.

Les méthodes que nous utilisons pour le grec sont très simples. Nous nous sommes limités, jusqu' à présent, à lemmatiser les formes sans procéder à la

décomposition en radical et terminaison, en indiquant uniquement la catégorie grammaticale des formes (cf. annexe 10). Nous avons procédé de façon purement empirique : voici comment.
Au départ, nous disposions de deux textes lemmatisés de façon artisanale : une tragédie d'Euripide, Electre et Saint Jean Chrysostome, la Lettre d'exil. Ce sont ces deux textes que nous avons mis en mémoire pour constituer le premier dictionnaire automatique, avec lequel nous avons ensuite traité les lettres à Olympias de Saint Jean Chrysostome. Toutes les formes de cette oeuvre qui ne se trouvaient pas dans le lexique initial ont été lemmatisées par un philologue et ajoutées au dictionnaire de l'ordinateur.

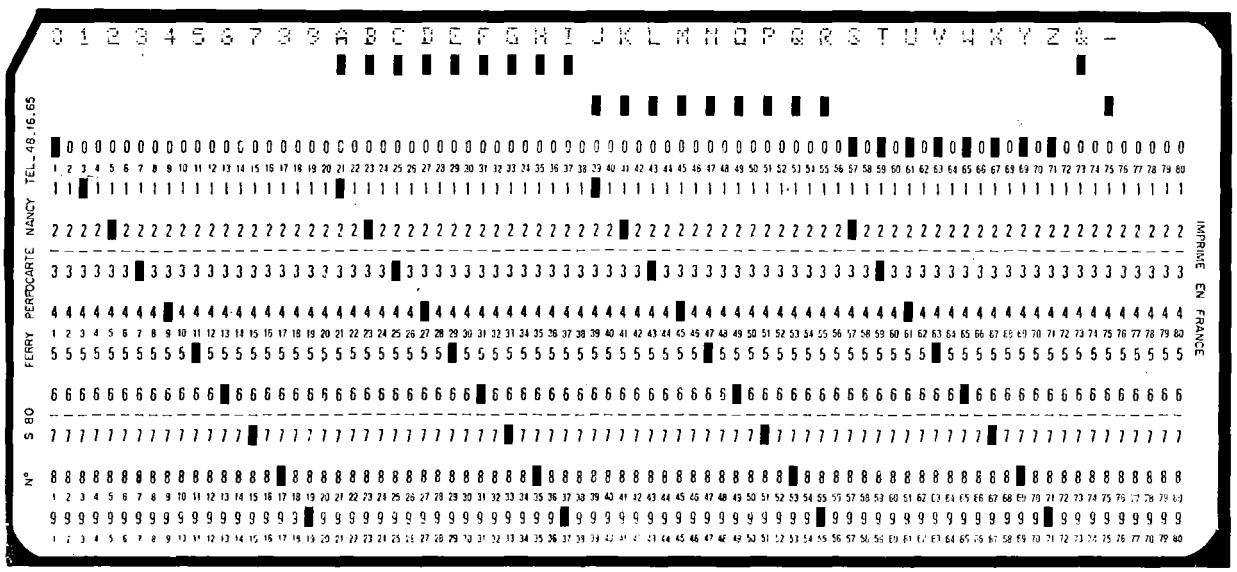
Cette même technique a été utilisée pour le Corpus Hermeticum que nous venons de publier et pour la Métaphysique d'Aristote qui est en cours de réalisation ainsi que pour les orateurs attiques que nous étudions en collaboration avec l'Université de Besançon.

Actuellement, notre dictionnaire grec automatique comprend approximativement 36.000 formes différentes, accompagnées de leur lemme, ce qui nous permet de lemmatiser entre 80 et 85 pourcents des occurrences de n'importe quel texte.

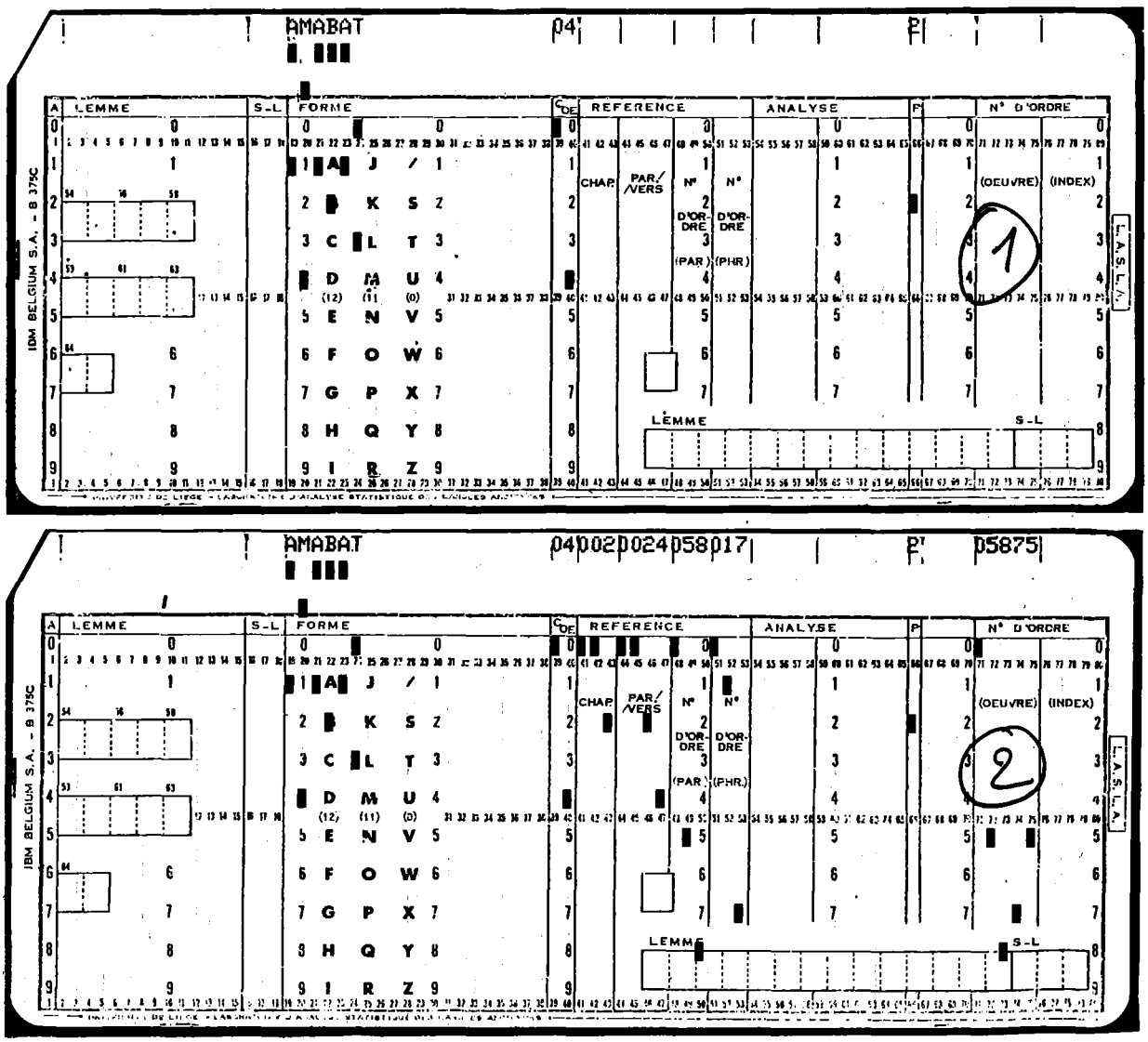
Encore faut-il ajouter que, pour une oeuvre donnée, une fois les premières pages analysées et les mots inconnus introduits en machine, le pourcentage de formes analysées automatiquement, pour cette oeuvre, atteint 95 %.

Enfin, pour terminer cet exposé, nous avons repris dans les annexes 10, 11, 12 et 13, la liste des textes latins et grecs que le L.A.S.L.A. a étudiés en utilisant les méthodes et les programmes que nous avons décrits dans cet article.

J. DENOOZ

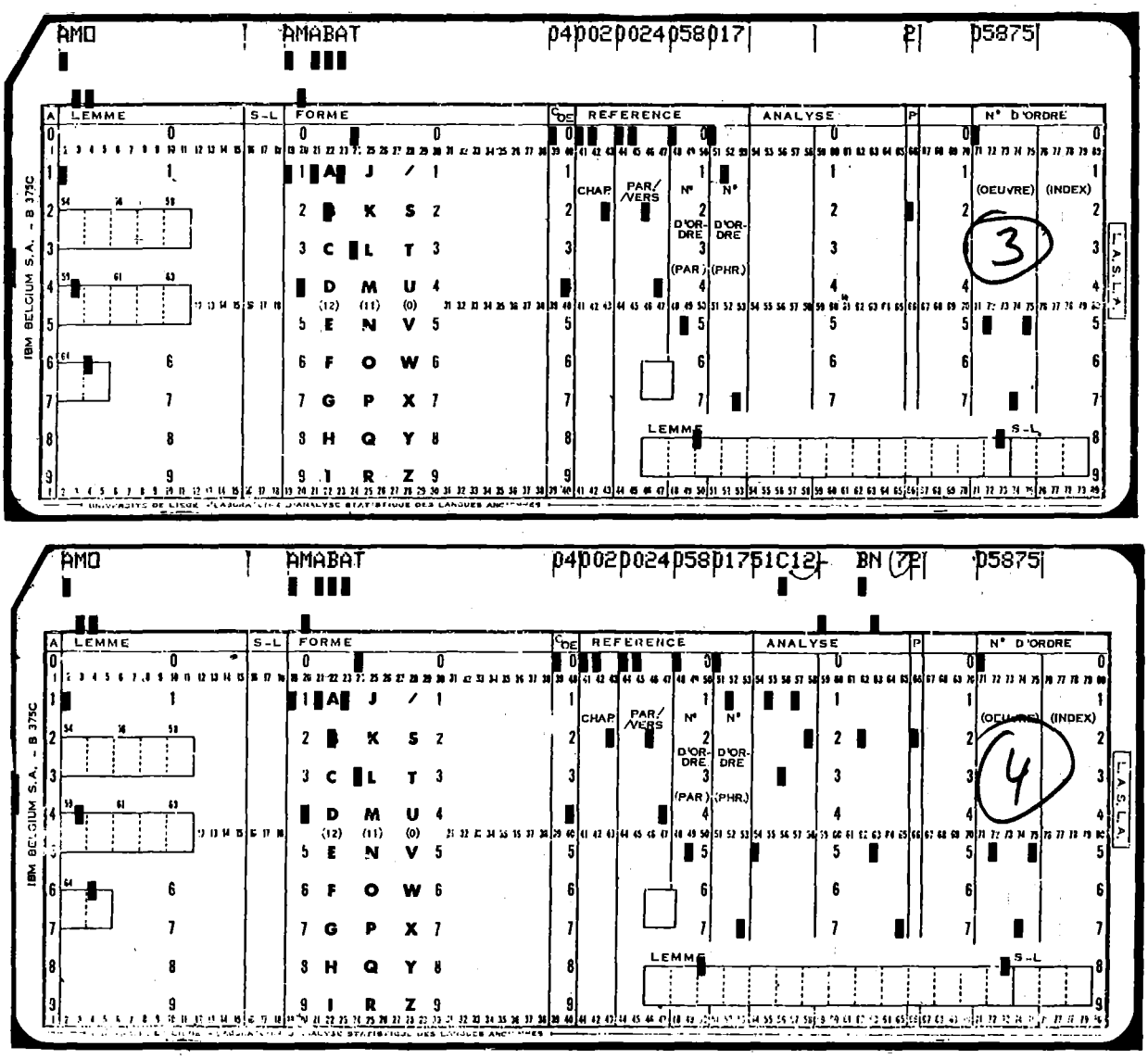


ANNEXE 1



ANNEXE 2

Extrait de la Revue (R.E.L.O.)
XIV, 1 à 4, 1978. C.I.P.L. - Université de Liège - Tous droits réservés.



ANNEXE 3

CONVENTIONS DU L.A.S.L.A.

Les lemmes homographes sont distingués par un indice numérique selon un système de classement qui nous est propre. La connaissance de ces conventions est nécessaire pour lire avec fruit les index des auteurs du programme.

1. D'une manière générale, l'ordre adopté pour l'attribution des indices de lemme correspond à la succession numérique des catégories et des sous-catégories grammaticales dans notre codification. Nous attribuons l'indice 1 au lemme qui se range le premier dans cette succession.
   EXEMPLES: TAMEN, taminis recevra l'indice 1 parce que les substantifs sont codifiés 1;
   TAMEN conjonction recevra l'indice 2 parce que les conjonctions sont codifiées 8.

   DICO, -are recevra l'indice 1 parce que les verbes de la première conjugaison sont codifiés 51,
   DICO,-ere recevra l'indice 2 parce que les verbes de la troisième conjugaison sont codifiés 53.

2. Lorsque deux lemmes appartiennent à la même catégorie grammaticale, nous les distinguons d'abord en fonction de l'ordre alphabétique des radicaux ou, si les radicaux sont identiques, en fonction de la quantité.
   EXEMPLES: OS, oris sera affecté de l'indice 1,
   OS, ossis sera affecté de l'indice 2,
   les composés en -CIDO de CADERE recevront l'indice 1,
   les composés en -CIDO de CAEDERE recevront l'indice 2.

3. En l'absence de tout autre élement distinctif, c'est le lemme le plus fréquent qui reçoit l'indice 1.
   EXEMPLES: TEMPVS 1, le temps; TEMPVS 2, la tempe;
   MOROR 1, s'attarder; MOROR 2, être fou.

4. Nous avons également distingué par un indice, les emplois de ESSE et de IRE comme auxiliaires. L'indice 1 est réservé aux cas où ces verbes sont employés en tant que verbes pleins, l'indice 2 aux cas où ils sont auxiliaires.

5. Tous les noms propres sont suivis d'un sigle alphabétique, en général la lettre N, tandis que l'adjectif reçoit le sigle A, signifiant adjectif dérivé de nom propre.
   EXEMPLES: LATINVS N désigne le roi Latinus,
   LATINVS A, l'adjectif LATINVS, -A, -VM.

   Lorsque deux homographes appatiennent à la même catégorie grammaticale, nous les distinguons en faisant précéder le sigle N d'un indice numérique choisi en fonction de leur fréquence.
   EXEMPLE: GALLI 1N, les Gaulois; GALLI 2N, les Galles.

ANNEXE 4

54 Catégorie grammaticale

55 Sous-catégorie Degré Voix

56 Cas Personne Nombre

57 Mode

58 Temps

59 Fonctions

65 Tradition

66 Ponctuation

54	55	56	57	58	59	65	66
Substantif 1	1e décl. 1 2e décl. 2 3e décl. 3 4e décl. 4 5e décl. 5 Anomal 6 Décl. gr. 7				Verbe principal 12 Verbe subord. 11	Leçon contestée 1 Mot absent dans certains mss. et adopté par édit. 2 Conjecture reçue 3 Lacune comblée 4 CRUX 5 Leçon rejetée 6 Mot présent dans certains mss. et rejeté par édit. 7 Mot impossible à analyser en fonction d'une lacune 0	K dernier mot du chap., du paragraphe ou du vers et de la phrase S dernier mot du paragraphe ou du vers et de la phrase 2 dernier mot de la phrase 11 dernier mot du chapitre et du paragraphe ou du vers 0 dernier mot du paragraphe ou du vers
Adjectif 2	P C S 1e classe 1 A J 2e cl.cons. 2 B K -er 3 C L -is 4 D M imp. 5 E N Anomal 6 F O Décl. gr. 7				**60** Emplois		
Numéral 3	P C S Cardinal 1 Ordinal 2 B K Distributif 3 Multiplic 4 Adv. ord. 5 E N Adv. mult. 6	S P Nominatif A J Vocatif B K Accusatif C L Génitif D M Datif E N Ablatif F O Locatif G P Indéclinable Z					
Adjectif-pronom 4	Personnel 1 Possessif 2 Réfléchi 3 Possessif réfléchi 4 Démonstratif 5 Relatif 6 Interrogatif 7 Indéfini 8		Mode Temps du verbe subordonné régi par un pronom relatif ou interrogatif (cf. verbe)		Emploi subst. 1 adject. 2 adverb. 6		
Verbe 5	A P D S-D 1e conj. 1 A J 2e conj. 2 B K S 3e conj. 3 C L T 4e conj. 4 D M 4e bis 5 E N Anomal 6 F O	S P 1e A J 2e B K 3e C L	Indicatif 1 Impératif 2 Subjonctif 3 Participe 4 Adj. verbal 5 Gérondif 6 Infinitif 7 Supin en -UM 8 Supin en -U 9	Présent 1 Imparfait 2 Futur simple 3 Parfait 4 Plus-que-parfait 5 Futur antérieur 6 -us fui/fuerim/fuisse 7 -us fueram/fuissem 8 -us fuero/urus fuisse 9	**61** Genre - Tmèse	CODE ALPHABETIQUE 12 11 Zéro 1 A J / 2 B K S 3 C L T 4 D M U 5 E N V 6 F O W 7 G P X 8 H Q Y 9 I R Z La lettre O sera notée Ø Le zéro sera noté O La conjonction du digit 1 et du zoning zéro est représentée non par une lettre mais par le signe /	
Adverbe 6	Relatif 6 Interrogatif 7 Négatif 8 Interrogatif-négatif 9 Comparatif 12 Superlatif 11		Mode Temps du verbe subordonné régi par un adverbe relatif ou interrogatif (cf. verbe)		a) Genre commun 1 féminin 2 masc. et fém. 3 masculin 4 masc. et neutre 5 neutre 6 b) Tmèse 12		
Préposition 7	Type MECUM 1	Cas régi 3, 4 ou 6					
Conjonction 8	Coordination 1 Subordination 2		Mode Temps du verbe subordonné régi par une conjonction de subordination (cf. verbe)				
Interjection 9							

ANNEXE 5

62 - 63

CODES DE SUBORDINATION

subjonctif seul complément ..... AA
ablatif absolu ..... AD
proposition infinitive ..... AG

ac ..... AK
proinde ac si ..... AN
perinde ac si ..... AS
... an ..... AX
an ... an ..... BA
... anne ..... BD
anteaquam ..... BG
antequam ..... BK

cum ..... BN
cumcumque ..... BS
cur ..... BX

donec ..... CA
dum ..... CD
dummodo ..... CG
dumtaxat ..... CK

etiamsi ..... CN
etsi ..... CS

licet CX

modo ..... DA

ne C-S ..... DD
- ne ..... DG
-ne ... an ..... DK
-ne ... anne ..... DN
-necne ..... DS
nedum ..... DX
neve, neu ..... EA
ni ..... ED
nisi ..... EG
nonne ..... EK
num ..... EN

postquam ..... ES
posteaquam ..... EX
priusquam ..... FA
proquam ..... FD
prout ..... FG

qua (rel.) ..... FK
qua (int.) ..... FN
quacumque ..... FS
qualis (rel.) ..... FX
qualis (int.) ..... GA
qualiscumque ..... GD
quam (rel.) ..... GG
quam (int.) ..... GK
quamdiu (adv. int.) ..... GN
quamdiu C-S ..... GS
quamdudum ..... GX
quamlibet ..... HA
quamobrem (rel.) ..... HD
quamobrem (int.) ..... HG
quamquam ..... HK
quamvis ..... HN
quando ADV ..... HS
quando C-S ..... HX
quandocumque ..... JA
quandoque ..... JD
quandoquidem ..... JG
quantopere (rel.) ..... JK
quantopere (int.) ..... JN
quantulus (rel.) ..... JS
quantulus (int.) ..... JX
quantuluscumque ..... KA
quantus (rel.) ..... KD
quantus (int.) ..... KG
quantuscumque ..... KK
quare (rel.) ..... KN
quare (int.) ..... KS
quasi ..... KX
quatenus (rel.) ..... LA
quatenus (int.) ..... LD
quemadmodum (rel.) ..... LG
quemadmodum (int.) ..... LK
qui (rel.) ..... LN
qui (int.) ..... LS
qui (abl. rel.) LX
qui (abl. int) ..... MA
quia ..... MD
quicumque ..... MG
quin C-S ..... MK
quippe ..... MN
quis ..... MS
quisnam ..... MX
quisquis ..... NA
quo (rel.) ..... ND
quo (int.) ..... NG
quo ADV ..... NK
quo C-S ..... NN
(non) quo(d) ..... NS
quoad (adv. int.) ..... NX
quoad ..... PA
quocumque ..... PD
quod ..... PG
quominus ..... PK
quomodo (rel.) ..... PN
quomodo (int.) ..... PS
quoniam ..... PX
quoquo ..... RA
quot (rel.) ..... RD
quot (int.) ..... RG
quotcumque ..... RK
quotiens (rel.) ..... RN
quotiens (int.) ..... RS
quotienscumque ..... RX
quotquot ..... SA
quotus ..... SD
quotuscumque (rel.) ..... SG
quotusquisque ..... SK
quousque ..... SN

seu ..... SS
si ..... SX
sicut ..... TA
simulac ..... TD
sin ..... TG
siquidem ..... TK
sive ..... TN

tametsi ..... TS
tamquam ..... TX
tamquam si ..... WA

ubi (rel.) ..... WD
ubi (int.) ..... WG
ubi C-S ..... WK
ubicumque ..... WN
unde (rel.) ..... WS
unde (int.) ..... WX
undecumque ..... XA
ut (adv. rel.) ..... XD
ut (adv. int.) ..... XG
ut C-S ..... XK
ut si ..... XN
utcumque ..... XS
uter (rel.) XX
uter (int.) ..... YA
utercumque ..... YD
uti (adv. rel.) ..... YG
uti (adv. int.) ..... YK
uti C-S ..... YN
utqui ..... YS
utrum ...an ..... YX
utrum ...necne ..... ZA

velut ..... ZD
velut si ..... ZG

ANNEXE 6

9CVSTODIO	CVSTODIT	01548 0- -
9CVSTODIO	CVSTODIV	01544 0- -
9CVSTOS	CVSTOS	0113 2- -
9CVTIS	CVT	01133 0-
9CVTIO	CVT	0255 00 -
9CVTIS	CVTIS	0113 2- -
9CYCNVS	CYCN	01123 0- -
9CYCNVS	CYGN	01123 0- -
9CYMA	ICYM	0111 0- -
9CYMA	2CYMA	0113 3- -
9CYMA	2CYMAT	01136 0- -
9CYMBA	CYMB	0111 0- -
9CYMBALVM	CYMBAL	01126 0- -
9CYNICI	NCYNIC	01123-0-
9CYNICVS	NCYNIC	0221 0- -
9CYPRVS	NCYPR	01123 0- -
9CYRVS	NCYR	01123 0- -
9DEVS	D	01123-0-
9DO	D	0251 0‡
9DO	D	0353 00 -
9DAEDALVS	DAEDAL	0121 0- -
9DEMIVRGVS	DAMIVRG	01123 0- -
9DAMNVM	DAMN	01126 0-
9DAMNO	DAMN	0251 0- -
9DAMNO	DAMNAT	01518 0- -
9DAMNATIO	DAMNATI	0113 0- -
9DAMNO	DAMNAV	01514 0- -
9DAMNOSVS	DAMNOS	0121 0- -
9DAPS	DAP	01133 0- -
9DAPS	DAPS	0113 2- -
9DAREVS	NDARE	01123 0- -
9DAREVS	NDARI	01123 0- -
9DO	DAT	01518 0‡ -
9DEA	DE	0111 0-
9DEVS	DE	02123 0-
DE	DE	0370600-1 -
DEA	DEABVS	0111000-
DEA	DEABVS	0211N00-1 -
9DEBEO	DEBT	0152 0- -
9DEBILIS	DEBIL	0124 0- -
9DEBILITAS	DEBILIT	0113 0-
9DEBILITO	DEBILIT	0251 0- -
9DEBILITO	DEBILITAT	01518 0- -
9DEBILITO	DEBILITAV	01514 0- -
9DEBITVM	DEBIT	01126 0-
9DEBITOR	DEBIT	0213 0-
9DEBEO	DEBIT	03528 0- -
9DEBEO	DEBV	01524 0- -
9DECOR	DEC	0113 0- -
9DECET	DECE	0152 0- -
DECEM	DECEM	0131Z00-1 -
9DECEMBER	NDECEMBER	0113 2-
9DECEMBER	ADECEMBER	0223 2- -
9DECEMBER	NDECEMBR	01133 0-
9DECEMBER	ADECEMBR	0223 0- -
9DECENS	DECENS	0125 3- -
9DECENS	DECENT	0125 0- -
9DECEPTORIVS	DECEPTORI	0121 0- -
DECIENS	DECIENS	0136000-1 -
DECIENS	DECIES	0136000-1 -
9DECIMVS	DECIM	0132 0- -
9DECLIVVS	DECLIV	0121 0-

ANNEXE 7

5T802	VROS	-5TL734
5T802	VRVM	5TB436
5T803	A	5TF442
5T803	I	-5TK444
5T803	AE	5TJ442
5T803	VM	5TC445
5T803	VRA	5TF432
5T803	VRI	-5TK434
5T803	VRAE	5TJ432
5T803	VRVM	5TC435
5T804	A	5TJ446
5T804	AE	-5TK442
5T804	VM	-5T080
5T804	VRA	5TJ436
5T804	VRAE	-5TK432
5T804	VRVM	-5TC735
5T805	A	5TK446
5T805	VRA	5TK436
5T806	A	-5TL446
5T806	VRA	5TL436
5T807	VRA	-5TL736
51 01	A	-51B21
51 01	O	-51A11
51 01	AS	-51B11
51 01	AT	-51C11
51 01	EM	-51A31
51 01	ER	-5AA31
51 01	ES	-51B31
51 01	ET	-51C31
51 01	OR	-5AA11
51 01	ABO	-51A13
51 01	ANS	51A411
51 01	ANT	-51L11
51 01	ARE	51071
51 01	ARI	-5A071
51 01	ATE	-51K21
51 01	ATO	51B23
51 01	ENT	-51L31
51 01	ABAM	-51A12
51 01	ABAR	-5AA12
51 01	ABAS	-51B12
51 01	ABAT	-51C12
51 01	ABIS	-51B13
51 01	ABIT	-51C13
51 01	ABOR	-5AA13
51 01	AMVR	-5AJ11
51 01	AMVS	-51J11
51 01	ANDA	5AA502
51 01	ANDE	-5AB504
51 01	ANDI	51D60
51 01	ANDO	51E60
51 01	ANTE	-51F411
51 01	ANTI	-51E411
51 01	ANTO	-51L23
51 01	AREM	-51A32
51 01	ARER	-5AA32

ANNEXE 8

Tusculanes, V, 1: Quintus hic dies, Brute, finem faciet Tusculanarum disputationum, quo die est a nobis ea de re quam tu ex omnibus maxime probas disputatum.

5	1	1	1	QUINTUS	QVINTVS			32A00	4	1
5	1	1	2	HIC	HIC	1		45A00	4	1
					HIC	2		60000		2
5	1	1	3	DIES	DIES			15A00		1
					DIES			15J00		2
					DIES			15L00		3
5	1	1	4	BRUTE	~~BRVTE~~ BRVTVS N			~~60300~~ 12B00		1
5	1	1	5	FINEM	FINIS			13C00		1
5	1	1	6	FACIET	FACIO			55C13	0	1
5	1	2	1	TUSCULANARUM	TVSCVLANVS N			21M00	2	1
5	1	2	2	DISPUTATIONUM	DISPVTATIO			13M00		1
5	1	2	3	QUO	QVI	1		46F14	5	1
					QVIS	1		47F	5	2
					QVIS	2		48F00	5	3
					QVO	1		660		4
					QVO	2		660		5
					QVO	3		670		6
					QVO	4		60300		7
					QVO	5		820		8
5	1	2	4	DIE	DIE			60000		1
					DIES			15F00		2
				DISPUTATUM						
5	1	2	5	EST	EDO	2		56C11		1
					SVM	1		56C11		2
					SVM	2		E6C11		3
5	1	2	6	A	AB			70600		1
					AH			90000		2
5	1	2	7	NOBIS	NOS			41N00		1
					NOS			41000		2
5	1	2	8	EA	EA			60000		1
					IS			45A00	2	2
					IS			45F00	2	3
					IS			45J00	6	4
					IS			45L00	6	5
5	1	2	9	DE	DE			70600		1
5	1	2	10	RE	RES			15F00	0	1
5	1	3	1	QUAM	QVAM	1		660		1
					QVAM	2		670		2
					QVI	1		46C11	2	3
					QVIS	1		47C	2	4
					QVIS	2		48C00	2	5
5	1	3	2	TU	TV			41A00		1
					TV			41B00		2

TEXTES LATINS

Apulée, Florides
Le traité de l'Asclepius dans le Corpus Hermeticum (8748 mots)
Catulle, Carmina (13108 mots)
César, Bellum Gallicum (45748 mots)
Cicéron, Catilinaires (12488 mots)
Pro Caecina (11980 mots)
Pro Rabirio (3552 mots)
Pro Lege Manilia (6698 mots)
De Natura deorum (± 25000 mots)
Cornelius-Nepos (21000 mots)
Descartes, Compendium Musicae
Regulae (20707 mots)
Horace, Art Poétique (300 mots)
Chant Séculaire (339 mots)
Epitres (10450 mots)
Epodes (3093 mots)
Odes (13660 mots)
Satires (14694 mots)
Juvénal, Satires (26402 mots)
Ovide, Métamorphoses (extraits-10259 mots)
Perse, Satires (4646 mots)
Plaute, Amphitruo (9447 mots)
Asinaria (7807 mots)
Aulularia (6662 mots)
Bacchides (9551 mots)

ANNEXE 10

Casina (7328 mots)
Captivi (8514 mots)
Quinte-Curce (72497 mots)
Saint Cyprien, Quod Idola
Ad Donatum
De Immortalitate (± 7000 mots)
Salluste, Catilina (10950 mots)
Bellum Jugurthinum (21453 mots)
Sénèque, La Consolation à Helvia (6775 mots)
La Consolation à Marcia (8384 mots)
La Consolation à Polybe (5688 mots)
De la Clémence (8283 mots)
De la Constance (5354 mots)
Des Bienfaits (45591 mots)
De la Brièveté de la Vie (6213 mots)
De la Colère (22612 mots)
Des Loisirs (1962 mots)
De la Providence de Dieu (4142 mots)
De la Tranquillité de l'âme (7629 mots)
De la Vie Heureuse (7313 mots)
Lettres à Lucilius (119698 mots)
Agamemnon (5719 mots)
Hercule Furieux (7818 mots)
Hercule sur l'Oeta (11617 mots)
Médée (5809 mots)
Oedipe (6088 mots)
Phèdre (7355 mots).

ANNEXE 11

Les Phéniciennes (4226 mots)
Thyeste (6448 mots)
Les Troyennes (7073 mots)
Pseudo-Sénèque, Octavie (5326 mots)
Spinoza, Tractatus Politicus (21081 mots)
Tacite, Agricola (6880 mots)
La Germanie (5867 mots)
Le Dialogue des Orateurs (6391 mots)
Les Annales (90682 mots)
Tertullien, De Pallio (3463 mots)
Tibulle, (12635 mots)
Tite-Live, (extraits-17598 mots)
Virgile, Bucoliques (5713 mots)
Enéide (1 à 6 - 32210 mots)
Géorgiques (14834 mots)
Vitruve, Livres 5 (6700 mots)
7 (5716 mots)
8 (5896 mots)
10 (± 8000 mots)

ANNEXE 12

TEXTES GRECS

Saint Jean Chrysostome : La lettre d'exil (8523 mots)
Les lettres à Olympias (29485 mots)
La Providence de Dieu (20535 mots)

Hesychius de Jerusalem : (37753 mots)

Corpus Hermeticum : (36566 mots)

Aristote : Métaphysique (3 livres - 20054 mots)

ANNEXE 13