Traitement des textes grecs anciens par la mécanographie et les ordinateurs

Un fichier de cartes perforées joue le rôle d'une mémoire à laquelle peut être confiée une quantité illimitée de renseignements. Ceux-ci sont représentés par des perforations, petits trous rectangulaires dont un système conventionnel détermine la signification en fonction de leur place sur la carte.

A l'intention de ceux de nos lecteurs qui ne le connaîtraient pas, je commencerai par un résumé très bref du système employé sur les machines IBM et sur celles de la plupart des autres firmes.

La carte mécanographique a des dimensions standard. Les machines sont capables d'y discerner quatre-vingts colonnes et, dans chacune de ces colonnes, douze niveaux auxquels on attribue conventionnellement les valeurs suivantes, en allant du bas vers le haut: 9, 8, 7 . . . 1, 0, puis 11 et 12. Chacun des douze niveaux de chacune des quatre-vingts colonnes peut être perforé, et les machines peuvent discerner les 960 (12x80) positions possibles d'une perforation.

On a attribué à ces perforations des significations symboliques de telle sorte que chaque colonne peut recevoir la représentation codée d'un élément, et d'un seul. Pour les chiffres, il n'y a aucune difficulté: les perforations des niveaux zéro à 9 sont les symboles des chiffres correspondants et elles en jouent le rôle dans les divers travaux effectués en machines (opérations arithmétiques, impression, tri, etc).

Extrait de la Revue (R.E.L.O.)
II, 1 à 4, 1966. C.I.P.L. - Université de Liège - Tous droits réservés.

Pour l'alphabet, on use d'un code à deux perforations par colonne. Les perforations 12, 11 et zéro, jouant le rôle d'un zoning, délimitent dans l'alphabet trois groupes ou zones: A à I pour le zoning 12, J à R pour le zoning 11 et S à Z pour le zoning zéro. Dans chaque groupe, l'élément distinctif est fourni par l'une des perforations 1 à 9 (digits). Ainsi le zoning 12 joint au digit 1 forme le code de A; joint au digit 2, il forme celui de B et ainsi de suite. Mais les mêmes digits 1 et 2 fournissent, avec le zoning 11, les combinaisons représentatives de J et de K, et, avec le zoning zéro, celles qui correspondent à un signe spécial (cf. plus loin) et à S.

Enfin, le système permet encore de représenter un certain nombre de signes spéciaux. "Plus" et "Moins" correspondent aux perforations 12 et 11. Un trait oblique (/) correspond au code formé des perforations zéro et 1. Enfin, des codes formés de l'une des combinaisons 3/8 ou 4/8, avec ou sans l'un des trois zoning, fournissent huit possibilités supplémentaires, utilisées pour les principaux signes de ponctuation (.,) ainsi que pour des caractères intéressants dans les opérations commerciales (%, $, etc).

Le code standard ne contient aucune représentation pour les signes de ponctuation autres que le point et la virgule, non plus que pour les accents, le tréma, et les autres signes orthographiques non littéraux. Les systèmes d'impression courants ne connaissent que la capitale: tout au plus certains d'entre eux permettent-ils d'opposer une grande et une petite capitale.

C'est avec ce système relativement pauvre que le LASLA a

entrepris de traiter les textes latins. Il ne sera pas inutile de rappeler très brièvement ici les grandes étapes de ce traitement. On comprendra mieux ainsi les conditions dans lesquelles fut conçu et mis au point le traitement des textes grecs. (Pour plus de détails sur le traitement des textes latins, on pourra voir un article du Professeur L. Delatte paru dans le Bulletin de l'Association des Amis de l'Université de Liège, 36e année, 1, 1964, pp. 5 à 34.)

Les cartes perforées reçoivent d'abord, en perforation manuelle une forme (mot tel qu'il apparaît dans le texte) avec, éventuellement, sa ponctuation puis la référence de cette forme, calculée et perforée automatiquement par une 602A ou par l'ordinateur 1620. Ensuite, chaque forme est pourvue de son lemme (forme sous laquelle le mot apparaît dans le dictionnaire) et d'une analyse morphologique et syntaxique exprimée au moyen d'un code conventionnel (le code a été longuement décrit dans un article de la Revue de l'Organisation Internationale pour l'Etude des Langues anciennes par Ordinateur, 1966, 1, pp. 1-50). Lors des premiers travaux du LASLA, le lemme et l'analyse étaient écrits et perforés manuellement. Maintenant, nous les obtenons grâce à un programme d'analyse automatique qui fonctionne sur l'ordinateur 1620, avec deux armoires à disques connectées (on trouvera la description de ce programme dans un article de la Revue de l'Organisation Internationale pour l'Etude des Langues anciennes par Ordinateur, 1966, 2, pp. 17-46).

Le fichier ainsi complété est soumis à une série d'opérations: tri grammatical, dénombrements grammaticaux, tri alphabétique du lemme, comptage de la fréquence des lemmes et constitution du tableau de distribution du vocabulaire (ceci au moyen d'un programme de 1620); traitement du fichier-mots et du fichier- fréquence d'abord en trieuse, puis en ordinateur, en vue de l'impression de l'index et de la liste de fréquence. Ce n'est pas ici le lieu de décrire la technique utilisée pour l'impression. Disons seulement qu'elle se fait

sur un système 870, machine dont il sera longuement question plus loin et où l'impression se fait sur une machine à écrire automatique commandée par la lecture des cartes. Le fichier qui sert à l'impression est perforé par l'ordinateur grâce à un programme dont les effets sont essentiellement ceux-ci. Du fichier-mots et du fichier-fréquences, il extrait les éléments qui doivent apparaître dans le document imprimé (p. ex. l'index). Il intercale aux endroits voulus les codes de service commandant les retours chariot, les blancs, les tabulations, le passage en majuscules, etc. De plus, en fonction des indications fournies au départ, il prépare la mise en page (longueur des lignes, nombre de lignes par colonne, etc). Les feuilles imprimées en 870 à partir du fichier ainsi obtenu sont ensuite reproduites en off set.

Comme la plupart des machines à écrire automatiques, la machine connectée à la 870 ne distingue que des grandes capitales et des petites capitales. Pour améliorer la présentation de ses travaux, le LASLA a obtenu des services IBM une série de modifications qui nous permettent d'imprimer, soit à partir du clavier, soit à partir de la lecture des cartes, non seulement les majuscules et les minuscules, mais encore tous les signes de ponctuation et, en outre, les lettres accentuées (é, è, etc). Ces dernières nous sont utiles non pas sans doute pour les indices d'auteurs latins, mais bien pour l'impression automatique de textes français, spécialement pour les articles de cette Revue. Du point de vue du code, il a suffi de sacrifier quelques signes spéciaux inutiles dans les domaines non commerciaux et de renoncer à la possibilité, usuelle dans les machines standard, d'imprimer les chiffres aussi bien à partir des grandes que des petites capitales. Il a fallu en outre, pour l'accent circonflexe et pour le tréma, prévoir la possibilité de barres d'impression n'entraînant pas l'avance du chariot (caractères morts).

La Revue elle-même est une illustration du procédé qui vient

d'être décrit. Quant aux indices latins, on en trouvera un bon exemple dans l'Index du Corpus Tibullianum, qui constitue le 5e fascicule des Travaux du LASLA.

*

* *

Après ces préliminaires, venons-en maintenant à l'objet propre de cet article. Le traitement des textes grecs anciens par les machines mécanographiques et par les ordinateurs présente des difficultés spécifiques: l'extrême diversité des arrangements possibles des esprits, des accents et du iota souscrit entre eux et avec les voyelles fait monter le nombre des caractères à un total qui dépasse de loin la capacité des codes conventionnels des perforatrices alphanumériques. Pour α, par exemple, ce nombre s'élève à 24 (ἀ, ἁ, ᾶ, ὰ; ἄ, ἅ, ᾷ, ᾳ; etc). Or, le système en usage sur les machines IBM et sur celles de plusieurs autres firmes comporte un total de 48 codes à une, deux ou trois perforations par colonne.

Par ailleurs, les accents, le iota souscrit et, dans une certaine mesure, les esprits doivent rester ignorés dans diverses opérations, par exemple, pour le classement alphabétique. Dès lors, la représentation symbolique de ces signes devrait être indépendante de celle des voyelles qu'ils affectent et l'on devrait avoir la possibilité de les prendre en considération ou de les négliger, selon ce que demandent les traitements à exécuter. Ceci suppose, en outre, une possibilité de surimpression: en effet, l'impression d'une lettre et celle de son accent, étant commandées par deux codes différents, entraînent deux frappes distinctes qui doivent porter sur un même emplacement dans la ligne. C'est là un problème qui, en 870, peut être résolu par les caractères morts, comme on le verra plus loin.

Pour éviter toutes ces complications, on pourrait sans doute penser à une solution simple et radicale, qui consisterait à enregistrer les mots grecs sans accent ni esprit ni iota souscrit. C'est ainsi que, dans l'usage courant, l'on procède pour le grec moderne. Il en va de même aussi pour le français, pour lequel les machines mécanographiques normales ne disposent d'aucun caractère accentué.

Cependant, quels que soient les avantages du traitement automatique, ce serait les payer trop cher que de leur sacrifier toute une catégorie de données. Au reste, une méthode qui reposerait sur un tel appauvrissement serait ipso facto disqualifiée aux yeux des philologues. Sans doute pourrait-on faire observer que la notation des accents grecs est une création de l'époque hellénistique et que, sur plusieurs points, elle est incertaine et flottante. Il n'empêche qu'elle nous donne une masse considérable de renseignements utiles sur la langue. De plus, elle résout un nombre appréciable d'homographies: c'est une raison suffisante de ne pas la négliger.

C'est pourquoi, dès le début de son existence, le LASLA a étudié avec des ingénieurs de la firme IBM les modifications qu'il faudrait apporter aux machines existantes pour le traitement du grec accentué, et il a mis au point un ensemble de programmes destinés à ce traitement. Nous tenons à souligner la patience et la complaisance avec lesquelles les ingénieurs et les services techniques IBM se sont appliqués à la solution de problèmes d'un intérêt pourtant bien faible sur le plan du commerce et des affaires.

Comme le font apparaître les considérations qui précèdent, l'accentuation grecque pose des problèmes de deux types distincts. D'une part, il faut adapter le système de codes perforés à une représentation adéquate de l'écriture grecque, et conditionner une machine imprimante de façon qu'à partir de cartes perforées dans ce système, elle puisse écrire le

grec selon les normes traditionnelles.

D'autre part, il convient d'associer à chaque forme accentuée la forme non accentuée correspondante, de manière qu'au cours des opérations, le passage de l'une à l'autre soit constamment possible. Ainsi, le tri alphabétique qui se fait sur la forme non accentuée, doit entraîner simultanément les formes accentuées, puisque ce sont ces dernières qui sont utilisées pour une impression en ordre alphabétique (index).

Nous traiterons successivement les deux types de problèmes qui viennent d'être distingués.

I. L E C O D E G R E C.

Pour des raisons qui apparaîtront au cours de l'exposé, l'établissement du code est lié au choix d'une machine de base. Celle qui nous a paru répondre le mieux à nos besoins est un système IBM 870 modifié. On notera que la 870 grecque du LASLA est actuellement unique au monde. Mais il serait évidemment possible d'en construire d'autres exemplaires à partir des plans et des schémas existants.

Le système 870 comporte une unité centrale et une ou plusieurs unités périphériques. L'unité centrale (IBM 836) est un lecteur-perforateur de cartes qui peut fonctionner comme une simple perforatrice (IBM 026) mais qui est en outre capable d'une série de fonctions commandées par l'action conjointe d'un programme affiché et d'une carte-programme fixée sur un petit tambour. En fait d'unités périphériques, notre système 870 comporte une machine à écrire de type 866, qui se prête à l'utilisation manuelle et au fonctionnement auto-

matique.

L'ensemble constitué par la 866 et la 836 peut perforer des cartes et imprimer à la machine à écrire à partir du clavier de la perforatrice (opération manuelle) ou à partir de la lecture de cartes (opération automatique). Perforation et impression sont indépendantes l'une de l'autre, et il est possible de les associer ou de les isoler selon les résultats que l'on désire obtenir. De plus, toutes les opérations peuvent, indépendamment l'une de l'autre, être limitées à certaines zones: c'est ainsi que, des données transmises par le clavier, les unes peuvent être perforées, les autres imprimées, d'autres encore, perforées et imprimées. Enfin, au cours d'un même travail, il est possible de faire entrer alternativement les données par le clavier et par le lecteur de cartes: ceci permet, en particulier, d'obtenir la perforation automatique de certaines constantes (p. ex. le code d'oeuvre) et la perforation manuelle des données particulières (p. ex. les mots d'un texte).

Le clavier de la 866 est susceptible de deux positions: la position alphabétique et la position numérique. Le passage d'une position à l'autre peut être commandé soit manuellement à partir d'une touche, soit automatiquement par le programme. En position alphabétique, toutes les touches sont disponibles et elles produisent les perforations correspondant aux lettres de l'alphabet ainsi qu'à quelques signes spéciaux. En position numérique, une bonne partie des touches sont verrouillées. Les autres produisent les perforations numériques et les signes spéciaux restants.

En cas d'impression à la machine à écrire, chaque code de perforation commande une barre d'impression. Celle-ci porte deux caractères, l'un correspondant au niveau majuscule de la corbeille (Upper Case Shift), l'autre au niveau minuscule (Lower Case Shift). Le caractère imprimé dépend donc du niveau où se trouve la corbeille. Celui-ci est commandé à la

main ou programmé.

De plus, il existe un dispositif qui permet de programmer la mise au niveau majuscule pour l'impression d'un seul caractère et le retour à la minuscule immédiatement après (Singel Upper Case Shift): c'est là une possibilité qui intéresse particulièrement notre problème, comme la suite de l'exposé le montrera.

La machine standard que je viens de décrire permet de perforer et d'imprimer le grec au prix de modifications dont je voudrais maintenant exposer l'économie générale et les détails. Le lecteur en suivra plus facilement l'exposé s'il se reporte au tableau et au schéma des pp.33-35: il y trouvera au fur et à mesure les exemples adéquats.

Pour l'alphabet, il n'y a guère de problèmes quant aux codes perforés. Nous avons repris à peu près sans changement ceux qui servent pour l'alphabet "latin". Cependant, comme le grec attique n'a que 24 lettres, tandis que nous en avons 26, nous récupérons deux codes, utilisés l'un pour le digamma (perf. 0/3), l'autre pour la représentation de certains accents (perf. 11/7). Quant à l'impression, il a fallu, sur chacune des barres, remplacer le bloc de caractères latins par le bloc de caractères grecs adéquats.

La mise en majuscule pour un caractère est programmée, dans notre système, au moyen du code du caractère spécial qui, dans les machines standard, correspond à l'astérisque (perf. 11/4/8). Ce code perforé s'obtient à partir du clavier en position alphabétique. Au cours des opérations d'impression, sa présence dans une colonne a pour effet de mettre la machine à écrire au niveau majuscule pour l'impression du caractère correspondant au code de la colonne suivante. Ainsi, tout caractère appartenant au niveau majuscule de la corbeille occupe nécessairement deux colonnes dans la carte.

Le code de mise en majuscules pour un caractère est particulièrement utile pour la représentation des accents. Mais avant d'y venir, il nous reste à dire un mot encore à propos de l'alphabet. L'existence d'un signe spécial pour le sigma final, en effet, pose un petit problème. Nous l'avons résolu de la manière suivante: la barre d'impression commandée par la perforation 1 (obtenue à partir du clavier en position numérique) porte, à son niveau majuscule, le signe ς . Dès lors, l'enregistrement de ce dernier sur la carte exige les opérations que voici: frappe du code 11/4/8, mise en numérique, frappe du code 1, et elle entraîne l'occupation de deux colonnes.

Venons-en maintenant aux codes des accents et des esprits. Le premier système auquel on pense consisterait à symboliser par des codes distincts chaque esprit et chaque accent. Mais cela signifierait que, pour une voyelle affectée d'un esprit et d'un accent, il faudrait trois frappes venant de trois barres différentes. Un tel procédé, même après des réglages minutieux, entraînerait sans doute souvent des chevauchements (esprit et accent) qui rendraient l'impression peu lisible et lui donnerait un aspect désagréable. Aussi avons-nous préféré attribuer un code distinct à chaque combinaison possible d'un esprit avec un accent. Ces combinaisons sont au nombre de onze (les trois accents isolés; les deux esprits isolés; chacun des deux esprits associés à l'un des trois accents). Quant au tréma et au iota souscrit, nous les avons traités tous deux isolément.

Les signes et les combinaisons de signes sont groupés par deux. A chaque groupe est associé un code perforé unique. La barre d'impression correspondante porte les deux combinaisons, l'une au niveau minuscule et l'autre au niveau majuscule. Dès lors, le code complet qui permet de distinguer et d'imprimer un signe d'accentuation est constitué par le code du groupe et par la présence ou l'absence du code de majuscule dans la colonne précédente. La seule exception est cel-

le du iota souscrit, que l'on peut obtenir aux deux niveaux. Donnons ici un exemple: le code perforé 11/7 donne l'accent circonflexe accompagné de l'esprit rude si la colonne précédente contient le code 11/4/8, et de l'esprit doux dans le cas opposé.

Les barres d'impression correspondant aux esprits, aux accents, au iota souscrit et au tréma ont évidemment dû recevoir les caractères voulus, qui ont été spécialement fondus pour nos besoins. De plus, ces barres sont mortes, c'est-à-dire qu'elles ne provoquent aucun avancement du chariot. Il en résulte que, pour une impression correcte, les accents etc. doivent être perforés avant la lettre qu'ils affectent.

Les codes utilisés pour les accents sont, à une exception près (code 11/7, déjà signalé), des codes propres aux caractères spéciaux. Ils se perforent les uns en position numérique, les autres en position alphabétique.

Une conséquence du changement de valeur de ces codes est la perte des signes auxquels ils correspondent dans les machines standard. Parmi ceux-ci, les seuls qui soient intéressants sont les parenthèses (au point et à la virgule, en effet, sont affectés des codes spéciaux non utilisés pour les accents). Mais en outre, on devrait aussi pouvoir représenter le point et virgule ainsi que le point en haut. Pour l'impression de ces signes, il a fallu apporter à la machine un dernier type de modifications. Sur les machines normales, les chiffres peuvent s'obtenir indifféremment aux deux niveaux. Il y a donc là une surabondance souvent utile, mais jamais indispensable. Dès lors, il nous suffisait de faire remplacer certains chiffres en position minuscule par les signes qui nous étaient nécessaires. De la sorte, les chiffres 1 à 5 restent disponibles aux deux niveaux tandis que 6, 7, 8 et 9 ne le sont plus qu'au niveau majuscule. En pratique, il suffit de considérer que les données numériques

appartiennent toutes au niveau majuscule pour éviter les risques d'erreur.

On observera que quelques codes de caractères spéciaux restent encore disponibles. Ils sont utiles pour la programmation de différentes opérations de la machine, tels les retours de chariot, les tabulations, etc.

Le tableau et le schéma qui suivent présentent la synthèse des codifications dont on vient de lire les détails. Les colonnes du tableau donnent successivement les codes de perforation, la valeur standard correspondante, les deux valeurs dans le système grec (majuscule et minuscule) et le code correspondant en mémoire centrale lors des traitements par l'ordinateur 1620 (sur ce code, on trouvera des indications plus loin).

Quant au schéma, il représente le clavier de la perforatrice. Sur les touches disponibles en numérique et en alphabétique, les caractères du bas correspondent à la position alphabétique et ceux du haut à la position numérique; en outre, quand il y a lieu de distinguer, ceux qui se trouvent à droite appartiennent au niveau minuscule et ceux de gauche au niveau majuscule.

TABLEAU DE LA CODIFICATION GRECQUE.

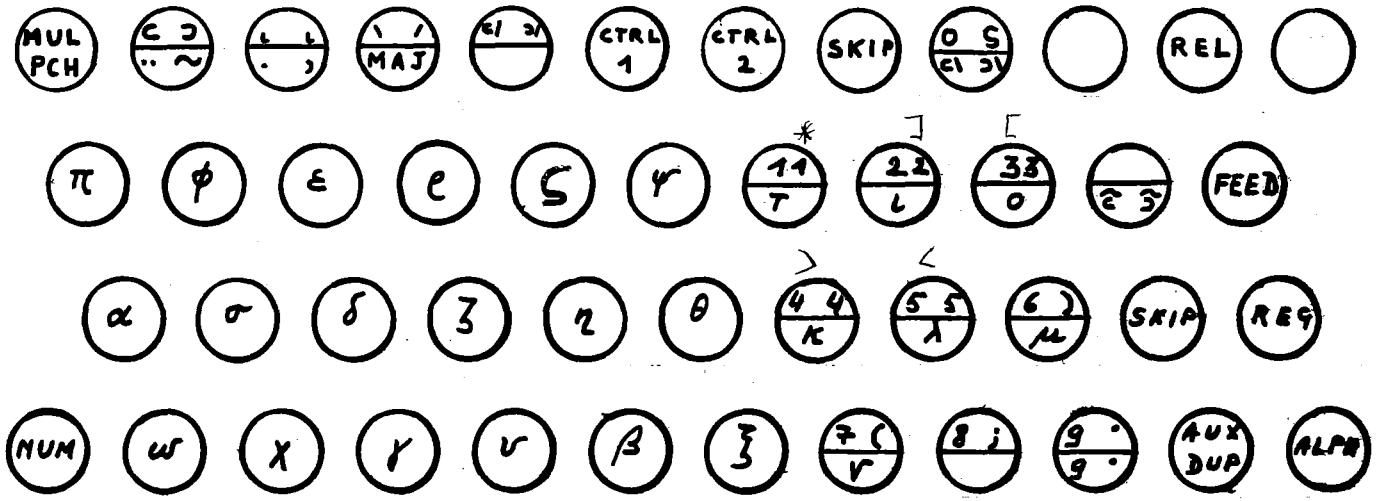
Code perforé (cartes)	Valeur standard	Valeur code grec majuscule	minuscule	Code alphanumérique (1620)
I. Clavier alphabétique				
12.1	a	A	α	41
12.2	b	B	β	42
12.3	c	Γ	γ	43
12.4	d	Δ	δ	44
12.5	e	E	ϵ	45
12.6	f	Z	ζ	46
12.7	g	H	η	47
12.8	h	Θ	θ	48
12.9	i	I	ι	49
11.1	j	K	κ	51
11.2	k	Λ	λ	52
11.3	l	M	μ	53
11.4	m	N	ν	54
11.5	n	Ξ	ξ	55
11.6	o	O	o	56
11.8	q	Π	π	58
11.9	r	P	ρ	59
0.2	s	Σ	σ	62
0.3	t	F	F	63
0.4	u	T	τ	64
0.5	v	T	υ	65
0.6	w	Φ	ϕ	66
0.7	x	X	χ	67
0.8	y	Ψ	ψ	68
0.9	z	Ω	ω	69

Code perforé (cartes)	Valeur standard	Valeur code grec majuscule	Valeur code grec minuscule	Code alphanumérique (1620)
11.7	p	Ψ	ψ	57
0.1	/	Π	π	21
4.8	@	¨	~	34
11.4.8	*	mise en majuscule pour un caractère		14
0.4.8	(	•	,	24

II. Clavier numérique

3.8	=	'	'	33
0.3.8	,	,	,	23
11.3.8	$	`	´	13
12.3.8	.	"	"	03
0	0	0	ς	70
1	1	1	ι	71
2	2	2	2	72
3	3	3	3	73
4	4	4	4	74
5	5	5	5	75
6	6	6	)	76
7	7	7	(	77
8	8	8	;	78
9	9	9	:	79

CLAVIER DE LA 870 GRECQUE.



Le système dont on vient de lire la description paraîtra sans doute lourd et compliqué. De plus, on craindra peut-être qu'il n'augmente exagérément le nombre de colonnes nécessaires à l'enregistrement des mots. Certes, il exige une attention soutenue. Frappe des accents avant celle des voyelles; frappe du code de majuscule dans les cas opportuns; passage à la position numérique pour les caractères qui relèvent de celle-ci et retour à la position alphabétique si d'autres caractères restent à enregistrer: tels sont quelques-uns des détails pour lesquels l'opérateur doit montrer une vigilance sans défaut. Quant au nombre de colonnes, il arrive qu'il s'accroisse de cinq unités pour un seul mot.

Prenons l'exemple d'une forme telle que ῥᾳδιουργὸς . Le iota souscrit occupe à lui seul une colonne; de plus, l'esprit rude et l'accent grave exigent chacun le code de majuscule avant leur code propre. On arrive bien ainsi au total de cinq colonnes additionnelles. En outre, le code de l'esprit rude, celui du iota souscrit et celui de l'accent grave appartenant au clavier numérique, il faudra passer en numérique avant de les frapper, puis retourner immédiatement en alphabétique. Enfin, le sigma final exige lui aussi le passage au numérique mais, comme il est le dernier signe du mot, on peut se dispenser ici du retour à l'alphabétique. Si l'on ajoute à toutes ces manoeuvres la frappe des lettres proprement dites, on constate que l'enregistrement d'un tel mot demande vingt-deux frappes. Mais c'est là un cas exceptionnel. Le plus souvent, l'accentuation d'un mot occupe une ou deux colonnes et les passages en numérique ne sont nécessaires que dans une proportion raisonable de cas.

En fait, l'expérience nous a montré qu'avec un peu d'exercice, on arrive facilement à perforer 700 à 800 mots par heure, avec une quantité d'erreurs relativement faible. C'est là un chiffre satisfaisant, si l'on veut bien considérer qu'en tout état de cause, la frappe du grec avec l'accentua-

tion est toujours plus lente que celle de langues telles que le français ou le latin.

II. L E T R A I T E M E N T.

Ici encore, quelques préliminaires sont indispensables. Nous avons établi pour le traitement des textes grecs un processus semblable à celui que nous utilisons pour le latin. La seule différence importante (mais que nous espérons temporaire) est que, pour le grec, nous ne disposons pas de programmes de lemmatisation et d'analyse automatiques.

Par ailleurs, la longueur de certains mots grecs et la place exigée par les accents nous ont contraints à adopter un dessin de carte particulier. C'est d'autant moins grave qu'en tout état de cause, la nécessité, pour certaines opérations, de réduire les mots à leur forme non accentuée nous obligeait à d'autres aménagements.

D'après les conventions que nous avons adoptées, les cartes-mots accentués ont le dessin que voici

Colonne	Contenu
1	12 (code de carte accentuée)
2-26	Lemme accentué
27-54	Forme accentuée
55	"Ponctuation"
56-57	Code d'oeuvre
58-80	Référence
58-60	Chapitre
61-64	Paragraphe

65-67	no d'ordre dans le §
68-70	no d'ordre dans la phrase
71-75	no d'ordre dans l'oeuvre
76-80	no d'ordre dans l'index

Pour l'enregistrement des mots du texte, un programme de 870 assure la perforation automatique du code-carte et du code d'oeuvre. En outre, il positionne la carte d'abord en colonne 27 pour la perforation manuelle de la forme, puis en colonne 56 pour la "ponctuation", et, quand une carte est terminée, il provoque automatiquement l'éjection de cette carte et l'introduction de la suivante. Enfin, pour faciliter le contrôle, le programme commande, en même temps que la perforation, l'impression automatique des formes et de la ponctuation à la machine à écrire. De la sorte, l'opérateur peut à tout moment vérifier son travail.

Pour la "ponctuation", nous avons gardé le même code perforé que dans les textes latins. En tenant compte des valeurs grecques des perforations, on a donc:

fin de chapitre	K
fin de §	ς
fin de phrase	2
fin de phrase et de §	9

La seconde opération est la référenciatiation, qui se fait automatiquement à partir de la "ponctuation", soit au moyen

de la calculatrice 602, soit en ordinateur.

Ensuite vient la lemmatisation, pour laquelle un programme de 870 assure adéquatement le positionnement et les éjections de cartes nécessaires.

C'est ici qu'intervient le traitement qui a pour rôle de créer automatiquement et d'enregistrer en perforations les formes non accentuées à partir des formes pourvues de leur accent.

Nous avons dit déjà pour quelle raison ce traitement est indispensable. Mais il convient d'ajouter encore quelques remarques.

Pour les opérations en machine, les données doivent avoir leurs éléments homologues dans des positions identiques. C'est ainsi que, pour un tri alphabétique, la lettre initiale du mot doit toujours se trouver dans la même colonne, et ainsi de suite pour les lettres suivantes. Dès lors, il ne suffit pas d'éliminer les signes des accents et des esprits. Il faut surtout serrer sur la gauche les caractères alphabétiques, sans laisser aucune position libre entre eux.

Quelques exemples ne seront pas inutiles. Dans une forme telle que γάρ, le code du γ se trouve dans la première colonne de la zone réservée à la forme. En revanche, le α initial des mots ἄλλος, ἄμιλλα et ἀδῆς est perforé respectivement dans la 2e, la 3e et la 4e colonne de cette zone, puisque le groupe esprit doux-accent aigu (niveau minuscule) exige une colonne, tandis que l'esprit rude, avec ou sans accent, en exige deux (niveau majuscule) et que le iota souscrit en demande une supplémentaire. La remarque vaut évidemment aussi pour les cas où l'accent frappe une voyelle non initiale: pour βαλλόμενος, par exemple, l'accent aigu décale toute la finale ομενος d'un caractère vers la droite. On voit dès lors comment se pose le pro-

blème: les codes d'accent doivent être éliminés et leur place, occupée par les caractères qui se trouvent initialement à leur droite; cette progression vers la gauche doit se poursuivre de manière à ne laisser aucun vide intérieur.

Voici comment se pratique cette opération. Comme la carte-mot accentuée est complètement remplie (cf. le dessin, p. 37), la forme non accentuée ne peut apparaître que sur une seconde carte. Ce dédoublement du fichier était, en tout état de cause, nécessaire pour la perforation de l'analyse, qui ne peut non plus trouver place sur la première carte. Mais elle complique les opérations et oblige à des précautions particulières.

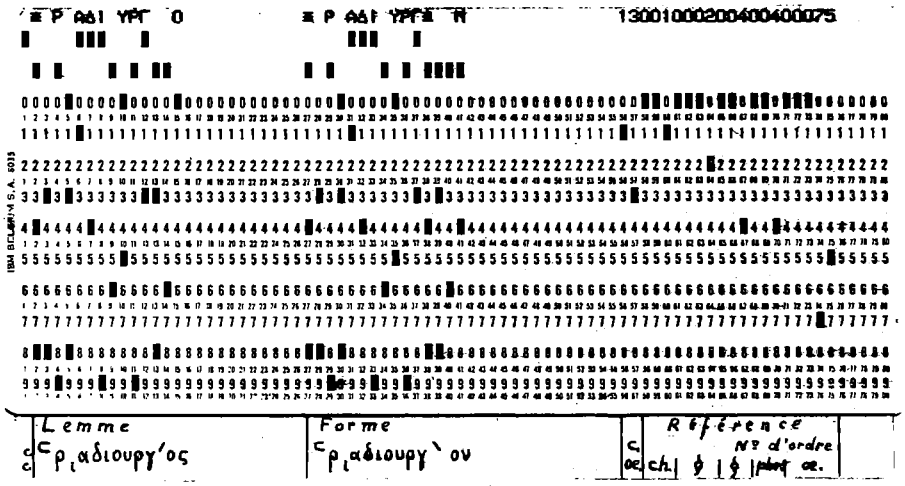
L'élément commun qui permet l'interclassement du fichier accentué et du fichier non accentué et le regroupement des deux cartes d'un même mot est le numéro d'ordre dans l'oeuvre. Quant à la distinction des cartes accentuées et des cartes non accentuées, elle est obtenue par le code-carte de la colonne 1. Nous sommes ainsi arrivés au dessin que voici:

Colonnes	Contenu
1	9 (code de carte non accentuée)
2-21	Lemme non accentué
22-44	Forme non accentuée
52-53	Code d'oeuvre
66	Ponctuation
71-75	No d'ordre dans l'oeuvre
autres colonnes (54-65; 67-70; 76-80)	disponibles pour l'analyse

A part l'analyse, la carte non accentuée est produite d'une manière entièrement automatique par l'ordinateur IBM 1620. Nous avons en effet mis au point un programme qui opère de la manière que voici. L'ordinateur lit en mode alphanumérique une carte accentuée. La lecture alphanumérique, en 1620, a pour effet de représenter chaque caractère alphabétique ou numérique (et donc chaque code perforé) par un code de deux chiffres dont la position en mémoire centrale obéit à des impératifs qu'il n'est pas nécessaire de préciser ici (la liste de ces codes se trouve dans la dernière colonne du tableau de la p. 33). Une fois la carte lue et enregistrée de cette manière, l'ordinateur transfère, en vue de la perforation, le numéro d'ordre dans l'oeuvre, la ponctuation et le code d'oeuvre. Ensuite, il compare un à un les codes de deux chiffres de chacun des caractères qui composent la forme et le lemme avec les codes propres aux esprits, aux accents et au iota souscrit. Chaque fois qu'il repère une identité, il avance de deux positions vers la gauche la partie du mot qui suit le code incriminé: ainsi se réalise à la fois la suppression des accents et le serrage des caractères alphabétiques. Quand le lemme et la forme ont été complètement traités, l'ordinateur perfore la carte non accentuée, avec un code 9 en colonne 1, et il passe à la lecture suivante.

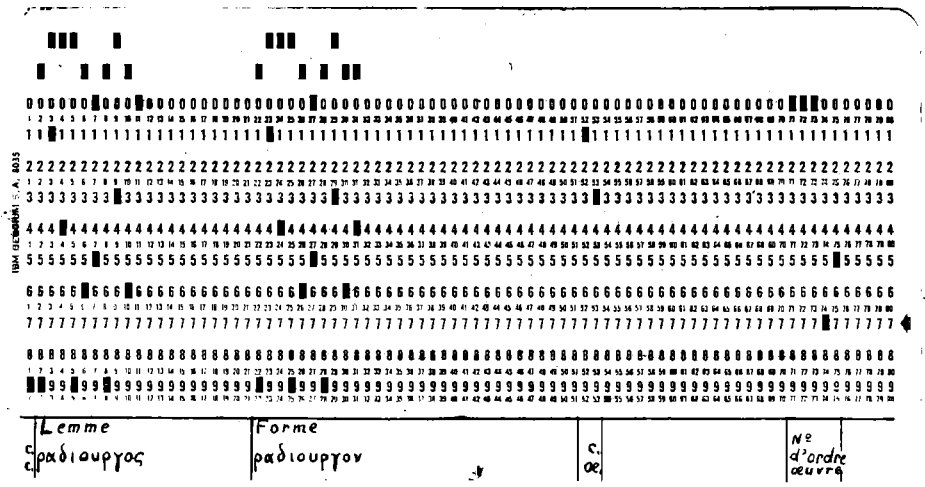
Les deux cartes reproduites ci-dessous illustrent ce qui vient d'être dit. Ce sont les deux cartes (accentuée et non accentuée) de la forme ῥᾳδιουργὸν , lemme ῥᾳδιουργός.

I. CARTE ACCENTUEE.



N.B. La transcription en clair sur le bord supérieur de la carte est à la fois incomplète et approximative. Telle qu'elle est, elle aide cependant aux opérations.

II. CARTE NON ACCENTUEE.



A ce moment, le seul élément qui doive encore s'ajouter est l'analyse. Pour celle-ci, M. J. De Bie a établi un code symbolique qui s'inspire du code latin auquel nous avons fait allusion au début de cet article. Nous espérons le publier prochainement. Comme nous l'avons déjà dit, il n'existe pas encore de programme d'analyse automatique du grec. Il faut donc ici procéder manuellement, avec, pour la perforation, l'aide d'un programme de 870 qui assure le positionnement correct des cartes.

Le fichier ainsi complété peut être soumis aux diverses exploitations que nous pratiquons pour les textes latins et que nous avons brièvement décrites plus haut. Pour certaines de ces exploitations, les méthodes mises au point pour le latin peuvent être appliquées sans autres changements que ceux qu'entraîne la modification du dessin de carte. Tel est le cas, par exemple, pour les tris et les comptages grammaticaux, pour lesquels le fichier non accentué fournit tous les éléments.

En revanche, les traitements qui reposent sur un tri alphabétique exigent l'emploi d'une technique particulière, qui permette d'ordonner le fichier accentué en fonction du fichier non accentué. Il est clair, en effet, que si le fichier accentué était soumis à un tri alphabétique du type habituel, les colonnes occupées par les codes des esprits et des accents provoqueraient des perturbations dans l'ordre attendu. C'est au moyen d'une trieuse 108 que nous évitons cet inconvénient. La première opération consiste à interclasser les deux fichiers, de manière que chaque carte non accentuée soit suivie de la carte accentuée correspondante. Le traitement ultérieur utilise un dispositif de la 108 appelé Group Hold. Ce dispositif permet de trier non pas les cartes prises isolément, mais bien des groupes de cartes en

fonction des perforations lues sur la première carte de chaque groupe. Dans notre cas particulier, le programme de la 108 a été conçu de manière que le tri se réalise d'après les cartes non accentuées, mais que celles-ci entraînent chaque fois la carte accentuée qui leur correspond.

Une fois le tri alphabétique terminé selon cette méthode, il reste à extraire automatiquement du fichier les cartes non accentuées qui, ayant dirigé le tri, ont rempli leur rôle et sont désormais inutiles. Le fichier résultant est un fichier alphabétique des formes accentuées. On peut d'abord le passer en 1620 pour en tirer les données nécessaires à la constitution du fichier-lemmes (qui donne la fréquence de chaque mot dans le texte étudié), et du fichier de distribution du vocabulaire (qui indique le nombre de mots employés x fois). Ensuite, un nouveau passage en 1620 doit produire le fichier qui, confié à la 870 avec un programme approprié, permet l'impression de l'index.

Notons, avant de conclure, qu'il était possible de concevoir le tri alphabétique d'une manière apparemment plus élégante, grâce aux disques magnétiques connectés à l'ordinateur 1620. Ceux-ci constituent, on le sait, des mémoires périphériques à grande capacité. On pourrait donc y enregistrer les formes accentuées, puis opérer un tri alphabétique en ordinateur, par l'intermédiaire de formes non accentuées qui seraient extraites au fur et à mesure des besoins et ne devraient donc pas être conservées.

Malgré les apparences, ce procédé n'entraînerait cependant aucune économie de cartes. En effet, ce que trie l'ordinateur, ce sont des enregistrements en mémoire et non pas des cartes. Dès lors, les effets d'un tri en ordinateur ne deviennent disponibles que par une sortie de ses résultats. Dans notre cas, la sortie serait nécessairement la perforation d'un nouveau fichier, qui servirait à la recherche des fréquences et à la préparation du fichier d'impression. Par

ailleurs, on se souvient que l'analyse grammaticale ne peut trouver place sur la carte accentuée. On ne voit donc pas comment on l'enregistrerait si l'on renonçait à constituer, pour chaque mot, une seconde carte.

Au terme de cet exposé, notre lecteur estimera peut-être que le traitement automatique des textes grecs présente une complication excessive. Sans doute est-il plus difficile et plus lent que celui des textes latins. Mais il ne l'est que dans l'exacte mesure où l'exigeait l'enregistrement de données plus riches. Nous tenons d'ailleurs à dire qu'à l'usage, il est nettement moins malaisé qu'on ne pourrait le croire.

Pour l'équipe du LASLA,
Et. EVRARD.