LES CAHIERS DE L'ANALYSE DES DONNÉES

J.-P. BENZÉCRI

Divertissement : décryptage d'un fichier de texte en grec moderne d'après la distribution des caractères

Les cahiers de l'analyse des données, tome 17, n° 1 (1992), p. 119-124

<http://www.numdam.org/item?id=CAD_1992__17_1_119_0>

Les Cahiers de l'Analyse des Données
Vol XVII 1992 n°2 pp. 119 124.

DIVERTISSEMENT: DÉCRYPTAGE D'UN FICHIER DE TEXTE EN GREC MODERNE D'APRÈS LA DISTRIBUTION DES CARACTÈRES

[DÉCRYPTAGE GREC]

J.-P. BENZÉCRI

1 Occasion de l'expérience de décryptage

Monsieur Gr. GIANNAROU poursuit d'intéressantes recherches stylistiques sur le discours politique en langue grecque. Accédant à notre demande, il a bien voulu nous communiquer une partie du corpus qu'il a constitué. Les textes étant saisis sur micro-ordinateur Macintosh par le logiciel "WORD" de Microsoft, la lecture ne nous paraissait pas susceptible d'offrir de grande difficulté. Nous pensions seulement, d'après notre expérience de plusieurs polices de caractères grecs utilisées en dehors de la Grèce, qu'il y aurait quelques ambiguités, d'ailleurs faciles à résoudre, quant à la représentation de lettres ou formes de lettres propres au grec: {Ψ, ψ, Ξ, ξ, ς, ...}.

```
□□□□□ □□□□□□□□□. H □□□□□□□□□ □□□ ¶A™OK
□□□□□□□□. O□ □□□□□□□ □□□ □□□□□□□□
□□□□□□□□□□□□ □□ N¢ □□□□□□□□□
□□□□□□□□□□ □□□□□□□□□. ¢□□□□□ □□□ ¶A™OK
□□ □□□□ □□□ □□□ □□□□□¿□□□. B¿□□□□ □□ ¶A™OK
```

Grande fut donc notre surprise en voyant paraître à l'écran, sous la police 'monaco' prise par défaut par le traitement de texte, une page dont presque tous les caractères étaient représentés comme manquants; passer à la police 'courrier' améliorant la présentation du texte, sans toutefois diminuer en rien notre perplexité!

```
Â›Ó·È Î˜,¤ÚÓËÛÈ. H ÂÚÈ¤ÙÂÈ ÙÔ‡ ¶A™OK
ÙÂÏÂ›˜ÛÂ. OÈ ÂÎÏÔÁ¤ Ù‹˜ Î˜ÚÈ·Î‹˜
·Ó· ÂÈÎÓ‡Ô˜Ó Ù‹ N¢ ·Ó›Û ˜ÚÉ
·˜ÙÔ ‡Ó·ÌÉ Î˜,¤ÚÓËÛÉ. ¢ ›ÓÔ˜Ó ÛÙ ¶A™OK
Ù Ú ÌÔ ÔÔ‡ ÙÔ‡ Ù·ÈÚÈ¿·ÂÈ. B¿ Ô˜Ó Ù ¶A™OK
```

Il était manifeste que le système et la police utilisés pour créer le fichier de texte se complaisaient dans les plages lointaines d'une extension du code ASCII. En toute innocence, Mr. GIANNAROU nous posait un problème de décryptage.

Les cahiers de l'analyse des données - 0339-3097/92/01 119 06/$ 2.60/ c Gauthier Villars

2 Indices et méthodes

En un sens, la police 'monaco' offre du texte une vue plus suggestive que 'courrier': en éliminant presque tous les caractères sortant de l'ASCII strict, elle montre clairement la part, d'ailleurs restreinte, que conservent dans le texte les caractères usuels. Les signes de ponctuation subsistent; on trouve fréquemment, après un '.', une capitale, telle {E, H, P, ...}, commune, au moins dans son dessin, aux deux alphabets grec et latin.

Quelques mots entiers se lisent, écrits en une langue occidentale; tel "Mirage", nom d'un objet aujourd'hui présent dans tous les cieux, même loin des sables. Le groupement ¶A™OK, trois fois répété dans le passage offert ici en exemple, évoque le sigle d'un parti politique 'PASOK'; et nous fournit ainsi le code des deux consonnes {Π, Σ} qui manquent à l'alphabet latin pour écrire ce sigle en grec.

Il est manifeste qu'à la différence de nos polices grecques usuelles qui substituent les formes grecques aux formes latines (mettant α pour 'a', β pour 'b'...), le système utilisé laisse à leur place les caractères latins, et loge ailleurs tout ce qui manque pour écrire également en grec; et, notamment, toutes les minuscules!

Pour le décryptage des minuscules elles-mêmes, plusieurs voies s'ouvrent. En calculant les fréquences d'emploi des divers signes, on doit reconnaître sans peine celles des lettres dont les fréquences sont bien détachées. En fait, ce procédé classique ne nous a servi qu'à identifier avec une forme particulière de α le signe le plus fréquent. Certes, on peut créer une liste des fréquences d'après seulement quelques pages; mais nous n'avions sous la main que des textes en grec ancien, dont les désinences sont bouleversées par l'usage moderne, ce qui modifie grandement les fréquences. Aurions-nous saisi quelques colonnes d'un journal d'Athènes, que notre perplexité eût subsisté; car l'usage, que nous croyons instable aujourd'hui en démotique, des signes diacritiques (accents et esprits) influe sur la fréquence des lettres composées (telles les {é, è, ê, ...} du français); et c'est ce qui a placé un α au premier rang (cf. *supra*).

Le dénombrement des digrammes est un procédé presque aussi usuel que le dénombrement des signes. Nous savons qu'ordinairement, l'analyse des correspondances tire du tableau des k(ia, ib), nombre de fois que le signe 'ia' a été trouvé précédant immédiatement 'ib', un facteur prédominant séparant les voyelles des consonnes; car celles-ci sont plus souvent au contact de celles-là que d'elles-mêmes; et réciproquement. On verra qu'ici, des règles propres à l'écriture grecque ont permis de réussir la séparation sans analyse; et de reconnaître, d'emblée, la lettre ν entre toutes les consonnes...

Il faut enfin savoir que même si le début du décryptage est embarrassant, les indices, recueillis d'abord un à un, se confirment bientôt mutuellement, s'appellent et s'accumulent, pour précipiter vers la solution!

3 Fréquences des lettres et digrammes

```
239 224  32{166  65 170  79  75} 32  13 244 229 236 229 221 246 243 229  46  32
 79 233  32 229 235 236 239 231 219 247  32 244 220 247  32 235 249 242 233 225
235 220 247  32  13 225 238 225 228 229 233 235 238 224 239 249 238  32 244 220
 32  78 162  32 240 225 238 221 243 248 249 242 232  32  13 225 249 244 239 228
224 238 225 237 232  32 235 249 226 219 242 238 232 243 232  46  32 162 221 238
239 249 238  32 243 244 222  32{166  65 170  79  75} 32  13 244 222  32 242 222
```

Voici, transcrit comme la suite des octets qui sont les numéros de ses caractères, un passage du texte qui comprend, signalées par des accolades, les deux premières occurrences de ¶A™OK, déjà montrées plus haut. On voit, e.g., que la première occurrence est précédée d'un blanc, dont le code ASCII est '32'; on reconnaît, en position {2, 4, 5}, les codes {65, 79, 75} des capitales {A, O, K}. Le reste n'est qu'énigme... mais permet de constituer aisément un tableau de fréquence de tous les octets, de 0 à 255.

oct:	216	217	218	219	220	221	222	223
frq:	0	0	0	1333	1655	2166	1788	0
oct:	224	225	226	227	228	229	230	231
frq:	1128	5382	351	108	1032	3996	402	981
oct:	232	233	234	235	236	237	238	239
frq:	1657	3710	272	2272	1987	2338	3939	4195
oct:	240	241	242	243	244	245	246	247
frq:	2288	499	2352	2527	4593	912	702	2281
oct:	248	249	250	251	252	253	254	255
frq:	616	1692	209	15	8	0	0	1

La plupart de ces fréquences sont très basses. Si l'on met à part les quelques cas déjà cités – blanc, signes de ponctuation, capitales communes aux deux alphabets – il n'y a que des pointes isolées en dehors du bloc des octets {219, ..., 250}, octets presque tous très employés. C'est donc sur ce bloc (dont les fréquences sont seules publiées ici) que doit se concentrer notre étude distributionnelle, c'est-à-dire, l'analyse des fréquences des digrammes; sans oublier le blanc, dont nous avons déjà annoncé le rôle essentiel.

De façon précise, on a attribué au blanc le numéro '0'; et aux octets de 219 à 250, les numéros suivants, de 1 à 32. On a construit un tableau (33 × 33) donnant, à l'intersection de la ligne 'ma' et de la colonne 'nb' le nombre, k(ma, nb), des occurrences du couple d'octets (m, n). Simultanément, on a considéré le tableau transposé; car plutôt que d'étaler un tableau à 33 colonnes on a écrit sur un bloc de plusieurs lignes de texte les informations afférentes à chacun des 'ma'; ce qui rend difficile la consultation des fréquences afférentes à un 'nb' particulier (les notions graphiques de *ligne* et *colonne*, ne coïncidant pas ici avec celles de *ligne* et *colonne* d'un tableau de contingence).

Sur notre présentation des k(ma, nb) qui détache nettement les k(ma, 0b), l'élément '29a', i.e., l'octet 218+29=247, se signale par le fait qu'il n'est quasiment jamais suivi d'un autre octet que '0b', le blanc ou octet 32. Si la fréquence, 2185, des paires (247, 32), est inférieure à celle, 2281 (qu'on lit au tableau de fréquence ci-dessus), de l'octet 247 lui-même, c'est essentiellement

```
 dénombrement des paires d'octets
33     0b
  1b    2b    3b    4b    5b    6b    7b    8b    9b   10b   11b   12b   13b   14b   15b   16b
 17b   18b   19b   20b   21b   22b   23b   24b   25b   26b   27b   28b   29b   30b   31b   32b
0a    2932
 165    29    17   231     0     4   761    71    22   378   607    78   217   211    55    38
 796   125   686   423   265   792    20    13   604  1542   249     7     0   110    88    24
29a    2185
   0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
   0     0     0     0     0     1     0     0     3     1     0     0     0     0     0     0
```

parce que 247 peut être suivi de signes de ponctuation (octets non dénombrés dans notre tableau des paires); les 3 rencontres (29a, 25b), i.e. (247, 243) doivent provenir d'erreurs de frappe.

Pour qui a tant soit peu lu de grec, il est clair que l'octet 247 représente la forme particulière, ς, que prend, en fin de mot, la lettre 'sigma', σ.

```
T.w: dénombrement des paires d'octets
33     0a    1a    2a    3a    4a    5a    6a    7a    8a    9a   10a   11a   12a   13a   14a
      15a   16a   17a   18a   19a   20a   21a   22a   23a   24a   25a   26a   27a   28a   29a
      30a   31a   32a
 0b  2932   238   525   739   694     0   427   961     1     0     0   581     0     0   666
     1034     0     1     2     2  1372   387     3    45     3     9     3     3    99  2185
        0   161     0
29b     0   107   315   174   108     0   133   373     0     0     0   269     0     0   112
      111     0     3     0     1     0   159     0    45     0     0     0     0   109     0
        0   154     0
```

Consultons maintenant le tableau transposé. Une grande similitude apparaît entre les fréquences, k(ma, 0b), d'occurrence des divers octets 'ma' avant l'octet 32, 0b=blanc, et celles et k(ma, 29b), d'occurrence avant l'octet 29b=247=ς. Dans les deux blocs les mêmes cases sont quasi nulles, ou nettement différentes de zéro; à quelques exceptions près, toutefois:

$$k(\,0a, 0b) = 2932 \quad ; \; k(\,0a, 29b) = 0 \quad ;$$

on dénombre 2932 fois un blanc suivi d'un blanc; mais le ς, qui termine un mot, ne peut être précédé d'un blanc;

$$k(20a, 0b) = 1372 \; ; \; k(20a, 29b) = 0 \quad ;$$

ce qui incite à découvrir le caractère que représente l'octet 238, '20a'.

Il faut se souvenir que, même en grec ancien, il est rare qu'en fin de mot, le 's' soit précédé d'une autre consonne: le mot du grec ancien ἅλς, sel, est usuel; mais, dans le Nouveau Testament, déjà, ἅλας le remplace; le dictionnaire ancien donne, pour le mot agneau, un nominatif ἄρς, mais c'est en signalant qu'à ce cas le mot est inusité; et la forme moderne est ἀρνί... On peut donc postuler qu'en grec moderne ς est toujours précédé d'une voyelle.

Devons-nous maintenant conclure de la similitude remarquée entre les k(ma, 29b) et k(ma, 0b) que, dans l'usage moderne, une lettre qui précède le blanc, c'est-à-dire la dernière lettre d'un mot, ne peut être qu'une voyelle aux seules exceptions près de ς et de la lettre énigmatique que représente l'octet 238?

Une telle conclusion surprend d'abord. Certes, qui a seulement tenté d'apprendre la grammaire grecque a pu remarquer que, dans les déclinaisons et conjugaisons, les suffixes propres aux cas, genres, nombres et personnes se terminent en général par l'une des consonnes {ς, ν} (voici décrypté, au passage, l'octet 238!); ou par une voyelle. Mais il y a des noms dont le nominatif se termine autrement: ainsi 'père' se dit πατήρ, 'air' se dit ἀήρ; cependant, un dictionnaire moderne donne: πατέρας, ἀέρας; c'est-à-dire le mot ancien, reconstruit avec la racine au degré bref (ε pour η), et une désinence.

4 Opposition consonne et voyelle et décryptage des lettres

Nous pouvons, avec quelque confiance, poser:

voyelles : {1a, 2a, 3a, 6a, 7a, 11a, 14a, 15a, 21a, 23a, 28a, 31a} ;

il n'y a rien là qui laisse espérer qu'on ait suivi l'ordre alphabétique, la présence d'une voyelle, 21a, après 20a, le ν, laissant, tout au plus conjecturer que cette voyelle puisse être un ο. D'autre part, le nombre des voyelles est très élevé: mais il faut se souvenir que les accents multiplient les signes distincts.

Pourtant le décryptage des minuscules sera rapide! On trouve d'abord, des phrases commençant par la capitale 'K', débutant un mot de trois lettres; laquelle (comme le confirme un regard sur un exemplaire du quotidien 'BHMA', 'la Tribune') ne peut être que la conjonction de coordination Καὶ, 'et'; écrite plutôt Καί, aujourd'hui. La lecture de α est d'ailleurs confirmée par la fréquence maxima de son octet. Connaissant la forme de 'et' débutant par une capitale, on reconnaît bientôt celle commençant par une minuscule: d'où la consonne κ.

Plusieurs phrases commencent par la capitale 'T', débutant, elle aussi, un mot de trois lettres, dont la dernière peut être ς; il s'agit donc de l'une des deux formes τας ou της, l'accentuation choisie restant incertaine... (et variant d'ailleurs, en toute rigueur, selon qu'il s'agit de l'article ou d'un pronom moderne, forme abrégée de αὐτή); ceci nous met sur la voie de τ et η... ainsi que de μ, à partir de la forme μας, de nous; τίς offrant une confirmation.

Se signale à notre attention, dans le cours même d'une phrase, la capitale 'E', suivie d'une lettre répétée: le nom de la Grèce, Ἑλλας, est connu de tous; ainsi que les adjectifs qui en dérivent. L'attaque de phrase transcrite ici:

```
Αθηναιεσ και Αθηναιοι , Ελληνιδεσ , Ελληνεσ
```

"Athéniennes et Athéniens, Grecques et Grecs", ne surprendra pas dans un discours politique. Nous tenons maintenant la lettre λ!

Il faut ici heureusement tempérer une remarque antérieure: même si l'ordre alphabétique est manifestement bouleversé, il en subsiste des vestiges: l'atteste la succession des octets {235, 236, 237, 238} pour les lettres {κ, λ, μ, ν}. Dès lors, on peut tenter de décrypter les mots l'un après l'autre: de multiples conjectures s'imposent; dont la vérification devient de plus en plus facile.

```
συγκεντρωσεισ του.Και ο λαοσ δεν
παιρνει πισω την υπογραφη του . Ο λαοσ
```

Ci-dessus, est transcrit sans signe diacritique, sans distinction entre les deux formes du σ, un de ces débuts en Και, qui nous a bien servi: on y reconnaît le mot peuple, λαοσ.

```
ειναι κυβερνηση. Η περιπετει του ΠΑΣΟΚ
τελειωσε. Οι εκλογεσ τησ κυριακησ
αναδεικνυουν τη ΝΔ πανισχυρη
αυτοδυναμη κυβερνηση. Δινουν στο ΠΑΣΟΚ
το ρολο που του ταιριαζει. Βαζουν το ΠΑΣΟΚ
στην αντιπολιτευση. Ισωσ εκει ωριμασει.
```

Et nous terminons sur le passage donné d'abord en exemple; transcrit, maintenant, à la seule exception des trois occurrences du sigle du PASOK: on y rencontre aussi le sigle, ΝΔ, du parti de la Nouvelle Démocratie. L'auteur affirme le dynamisme de la ΝΔ, face au PASOK, dont le temps est fini... Le mot de 'Démocratie', en plein ou en initiale, nous a servi à reconnaître le Δ, capitale codée par l'octet 162. Ainsi, au milieu de l'extension du code ASCII, dans la séquence {Γ,Δ,Θ,Λ,Ξ,Π}, de 161 à 166, s'offre un autre vestige de l'alphabet; l'ordre de succession des capitales non latines étant conservé.

La preuve est donc faite que des connaissances imprécises en langue ancienne, conjuguées avec quelque curiosité de la langue moderne, peuvent suffire pour décrypter un code, brouillé sans malice, du grec démotique.