

J. P. BENZÉCRI

**Élaboration statistique de données sur ordinateur
; application à l'analyse des textes ; contributions
attendues à l'étude de la Bible**

Les cahiers de l'analyse des données, tome 11, n° 1 (1986),
p. 7-30

http://www.numdam.org/item?id=CAD_1986__11_1_7_0

© Les cahiers de l'analyse des données, Dunod, 1986, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

ÉLABORATION STATISTIQUE DE DONNÉES SUR ORDINATEUR ; APPLICATION A L'ANALYSE DES TEXTES ; CONTRIBUTIONS ATTENDUES A L'ÉTUDE DE LA BIBLE

[TEXTE BIBLE]

par J.P. Benzécri*

Les 2 - 3 - 4 septembre 1985 se tint à Louvain la Neuve, sous les auspices de l'Association Internationale Bible et Informatique un Premier Colloque International sur le thème :

BIBLE ET INFORMATIQUE : LE TEXTE.

L'exposé qui suit a été rédigé d'après l'enregistrement d'une conférence prononcée au Colloque le mardi 3 septembre 1985. Outre des allusions aux autres communications entendues à L.-la-N., on trouvera ici quelques références à un autre colloque, tenu à Nice du 5 au 8 juin 1985 sur le thème : METHODES QUANTITATIVES ET INFORMATIQUES DANS L'ÉTUDE DES TEXTES avec le patronage conjoint du CNRS et de l'ALLC.

Je présenterai brièvement une série d'exemples, dont certains concernent explicitement la Bible ; et dont d'autres serviront de parallèles. Car, ainsi que je le répéterai en conclusion, la thèse que je défends est que les références externes sont très utiles dans l'étude de l'Écriture Sainte ; même si sont très rares pour l'Ancien Testament les références proches (textes contemporains écrits dans la même langue).

1 Le vocabulaire des évangiles : comparaison globale et chapitres de Jean : Le premier exemple est publié dans PRAT 3**. C'est une étude faite à l'Institut Catholique de Toulouse par l'Abbé J.M. Vacherot sous la direction du Recteur M^{OR} B. de Solaces. L'étude repose sur un tableau comprenant une ligne pour chaque mot et une colonne pour chaque chapitre, ou chaque texte. A la croisée d'une ligne *i* et d'une colonne *j* se trouve, dans le tableau, le nombre d'occurrences du mot *i* dans le texte *j*. On est parti du texte grec des Évangiles ; le texte a été lemmatisé, c'est-à-dire que chaque forme fléchie a été ramenée à une entrée du dictionnaire, ce qui, on le verra dans la suite, n'est aucunement indispensable et nous paraît même avoir plus d'inconvénients que d'avantages. Et quant au choix des données, le critère appliqué d'abord est que les mots retenus doivent se rencontrer 50 fois dans l'un au moins des quatre Évangiles. Ensuite, comme l'étude était concentrée sur l'Évangile de Jean, on a ajouté des mots dont la fréquence n'était pas aussi grande, mais dont on espérait qu'ils révéleraient la structure de cet évangile, objet propre de l'étude. Il s'agissait du problème du XXI-ème et dernier chapitre de l'Évangile selon St Jean, chapitre qui, comme vous le savez, vient après une suite de conclusions terminant le chapitre XX qui le précède.

(*) Professeur de statistique. Université Pierre et Marie Curie.

(**) Au cours de l'exposé, seront fréquemment cités d'une part le Volume Pratique de l'Analyse des Données en Linguistique et Lexicologie (cité PRAT 3), d'autre part la Revue : les Cahiers de l'Analyse des Données (cité CAD) ; livre et revue sont édités par Dunod à Paris.

Je commenterai trois graphiques issus de ce tableau par l'analyse des correspondances. Ces trois graphiques sont publiés dans PRAT 3 ; nous n'en reproduisons ici que le deuxième où l'on voit l'essentiel des résultats des trois analyses.

Le premier résultat de l'analyse du tableau restreint aux quatre colonnes relatives aux textes entiers des quatre Évangiles : ces quatre textes sont mis en rapport les uns avec les autres d'après le vocabulaire qu'ils utilisent. Sur le plan (1,2), on note d'abord suivant l'axe 1 une opposition entre Jean à droite et les trois autres Évangiles (dits synoptiques) à gauche ; ceux-ci s'étageant sur l'axe 2. Quant aux mots, les plus pleins, les plus évocateurs pour l'érudit, sont probablement ceux que j'ai soulignés ; les autres mots sont des mots vides. Vous voyez que πιστεύω (croire), Ἰουδαῖος (Juif ou Judéen), κόσμος (monde) Ἰησοῦ (Jésus), πατήρ (père) sont du côté de Jean ; βασιλεία (royaume) κύριος (seigneur), ἄνθρωπος (homme) du côté de Matthieu et des deux autres synoptiques. Remarquons encore λαλέω (parler) qui correspond à l'hébreu *DABER* : דַּבַּר et qui se trouve du côté de Jean ; tandis que λέγω et εἶπον (dire), qui correspondent à l'hébreu *AMAR* : אָמַר sont du côté des synoptiques (i.e. St Matthieu, St Marc, St Luc), au centre. Je fais cette référence à l'hébreu pour annoncer la comparaison qui se fera tout à l'heure entre les textes grecs et hébreux de deux chapitres de l'Ancien Testament. L'opposition de ces mots comme marquant des variantes du texte a d'ailleurs été affirmée au Colloque par le professeur Loimaranta.

La même analyse a été reprise en distinguant les chapitres de l'Évangile de St Jean. Le tableau de correspondance comprend cette fois 24 colonnes : 3 colonnes pour les Évangiles synoptiques considérés chacun dans son ensemble ; et 21 colonnes pour les 21 chapitres de Jean. Si l'auteur de l'étude avait disposé de dénombrements relatifs aux chapitres des synoptiques, on aurait eu un intéressant problème de comparaison et de classification ; on aurait vu, par exemple, si la classification mêlait, comme il est probable, un certain nombre de chapitres de récits des synoptiques avec le domaine des récits de Jean. Dans une certaine mesure, il y a sur le graphique (reproduit ici) du plan des axes 1 et 2, une réponse à la question posée : le dernier chapitre, XXI, de Jean se trouve, à l'analyse factorielle comme à la classification, bien au centre avec les récits, et peu séparé du chapitre XX. Les discours avant la Passion forment un groupe ; la Prière Sacerdotale qui est également dans ce groupe, s'en détache dans une troisième dimension.

Le déploiement des chapitres de Jean au-dessus de la bande du plan (1x2) où ils se projettent dans la présente analyse, apparaît dans une troisième analyse qui ne comprend, comme colonnes, que les chapitres de Jean ; le premier chapitre ayant été partagé en deux, le tableau de correspondance comprend 22 colonnes. L'opposition entre λαλέω (דַּבַּר parler) et εἶπον, λέγω (אָמַר dire) se trouve coïncider avec l'opposition entre discours (parler) et récit (dire), quand on est au sein de l'univers de Jean. La Prière Sacerdotale, très éloignée, naturellement associée au terme κόσμος. La première partie du chapitre I (qui a été, répétons-le, opportunément partagé en deux par l'abbé Vacherot) :

"Ἐν ἀρχῇ ἦν ὁ λόγος..." (Au commencement était le verbe) va du côté de la prière sacerdotale ; et le récit qui constitue la deuxième partie de ce premier chapitre rejoint les autres récits.

Ces résultats donnent une idée de la typologie qu'on peut obtenir sur des chapitres d'une oeuvre ou sur des oeuvres entières d'après le vocabulaire qui y est employé. Ici, l'étude est faite sur des mots lemmatisés (i.e. des entrées de dictionnaire). On peut se demander ce qu'il en serait si on comptait des formes : nous y reviendrons au § 4. J'ai laissé de côté le contenu théologique de l'interprétation.

2 Les variantes des manuscrits grecs de l'Épître de Jacques (d'après Ch. Amphoux in PRAT 3 pp 285 sqq) : Nous passons maintenant à un nouveau format de données avec une étude faite en collaboration avec M^r Amphoux ici présent : le classement de manuscrits de l'Épître de Jacques. Nous rencontrons le problème évoqué par M^r le Doyen Gryson à l'ouverture du Colloque : *que faut-il entendre par la multiplicité des états d'un texte ?* Faut-il comprendre que nécessairement il existe un archétype à partir duquel une action de division crée des types divergents, ou au contraire faut-il prendre les choses comme elles sont, c'est-à-dire accepter une multiplicité de types écrits, qui peut-être sont nés ensemble.

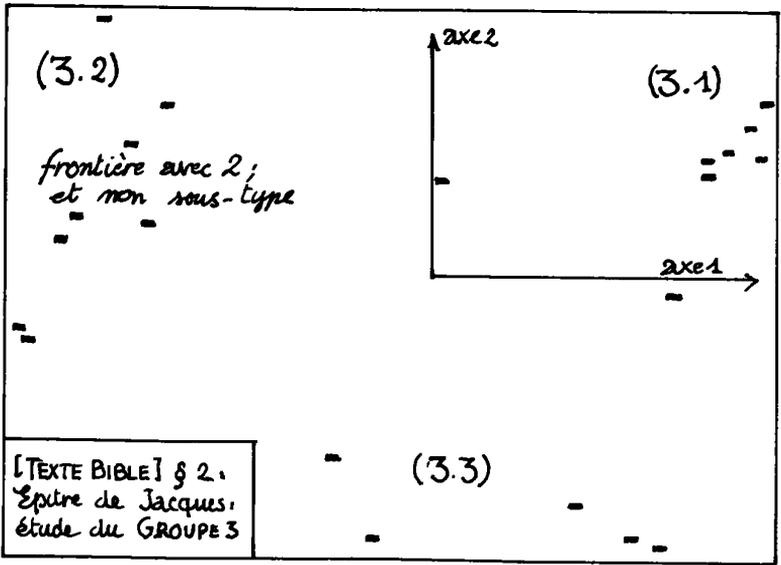
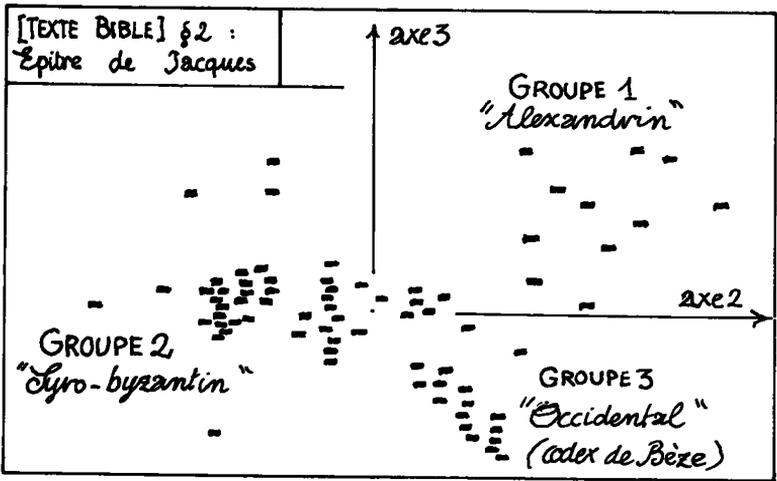
Dans l'étude précédente, on se bornait à dénombrer des mots dans certaines oeuvres ou dans leurs chapitres. Ici, les données sont différentes. Le tableau comprend 74 lignes correspondant chacune à un manuscrit et désignées par M1, M2, M3, M4 etc. et pour chaque lieu variant, un bloc de colonnes correspondant chacune à une variante attestée (tout au moins attestée avec une fréquence suffisante ; les variantes rares étant groupées sous le titre : "autres") ; il y a au total 101 colonnes. Par exemple, dans la 1-ère ligne, le manuscrit M1 comporte au 1-er lieu la variante B, au 2-ème lieu la variante B, au 3-ème lieu la variante C ; dans la 2-ème ligne, le manuscrit M2 a la variante D au 1-er lieu, la variante C au 2-ème lieu, la variante A au 3-ème lieu. Un tel tableau, du point de vue du format, ne diffère pas de ce qu'on obtient si l'on interroge toutes les personnes ici présentes et qu'on leur pose des questions closes. A la 1-ère question ; par exemple, il y a 4 réponses possibles étiquetées A, B, C, D ; à la 2-ème question, il y a 5 réponses possibles étiquetées A, B, C, D, E ; etc.. Dans le tableau de correspondance croisant les personnes interrogées (les lignes) et les réponses aux diverses questions, est affecté à chaque question un bloc de colonnes comportant autant de colonnes que cette question comporte de modalités de réponse. En un mot, le tableau de données de M^r Amphoux a le format de questionnaire.

La première analyse de ce tableau montre l'existence de trois groupes qui, selon les indications externes que possède M^r Amphoux, se laissent étiqueter comme alexandrin, syro-byzantin, occidental, ce troisième groupe s'apparentant au courant attesté par le codex de Bèze. Après cette première analyse, on s'interroge sur la structure du groupe 3, groupe occidental, auquel est consacrée une analyse partielle, montrant une subdivision de ce groupe en trois classes notées : 3.1, 3.2 et 3.3 : 3.1 et 3.3 correspondent à des versions distinctes ; mais 3.2 est un groupe limite entre occidental et syro-byzantin. La conclusion à laquelle est parvenu M^r Amphoux est que cette subdivision 3.2 n'existe pas en propre mais correspond à des contaminations.

Ceci pose encore une fois le problème de la structure d'un grand nombre d'états attestés d'un texte. On ne peut savoir s'il s'agit d'une divergence à partir d'un archétype unique, d'apports extérieurs qui ont diversifié un texte qui existait au départ ; ou d'écritures indépendantes : e.g. d'un texte oral préexistant au texte écrit. J'ose affirmer que les méthodes taxinomiques sont relativement bien au point ; les résultats qu'elles fournissent dépendent certes des informations traitées (toutefois des vérifications sont possibles quant aux données), mais la question des questions est l'interprétation. Ce qui est difficile, c'est ce qu'a fait M^r Amphoux : étiqueter e.g. les groupes selon des différences géographiques. Et derrière ces différences géographiques, il doit y en avoir d'autres plus profondes qui les expliquent.

Manuscripts	lieu 1: 4 variantes				lieu 2: 5 variantes					lieu 3: 3 var.		
	1a	1b	1c	1d	2a	2b	2c	2d	2e	3a	3b	3c
M1	0	1	0	0	0	1	0	0	0	0	0	1
M2	0	0	0	1	0	0	1	0	0	1	0	0
M3	1	0	0	0	0	0	0	0	1	0	1	0
M4	1	0	0	0	0	0	0	0	1	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

[TEXTE BIBLE] § 2 : D'après Ch. Amphoux : schéma du codage



La même analyse a été refaite en ajoutant aux chapitres, des groupes de chapitres ; on a procédé avec prudence, en s'assurant de la stabilité. On arrive ainsi à séparer à peu près parfaitement les chapitres 1 à 39 des chapitres 40 à 66, avec quelques écarts marqués par des flèches qui sont sensées reconduire les chapitres dans le demi-plan auquel ils appartiennent. Comme je l'ai dit, cette construction sort d'une façon plus agréable, avec la droite cloison coïncidant avec l'un des axes factoriels si on élimine les chapitres historiques, qui sont évidemment très à part.

On a fait également de multiples classifications sur les chapitres et groupes de chapitres pour confirmer la stabilité des résultats. Par exemple, dans une classification sur 13 tranches de 5 chapitres, les chapitres 1 à 35 sont groupés ; la partie finale subdivisée ; avec, détachés à un niveau hiérarchique assez élevé, les chapitres 36 à 40, c'est-à-dire les chapitres historiques. Reste la question : sur quelle structure sémantique, sur quelle composition de structures repose cette partition du texte entier ? Le verset 9 du ch. XLII ne donne-t-il pas le plan du livre ? "Les premiers événements, voici qu'ils sont arrivés ; j'en annonce aussi de nouveaux ; avant qu'ils arrivent je vous les ferai connaître".

4 Les journaux parisiens de l'été 1793 (d'après A. Salem ; Laboratoire d'études des textes politiques français ; Saint-Cloud ; in PRAT 3 pp 183-200) : Comme terme de comparaison avec les résultats obtenus sur le texte biblique, je présente maintenant des analyses provenant d'un tout autre domaine : il s'agit de 45 textes appartenant à 3 journaux différents publiés à Paris respectivement par Hébert, Roux et Leclerc pendant l'été brûlant de 1793. A la vérité, ces journaux se ressemblent autant qu'il est possible de se ressembler : ils s'adressent au même public, ils traitent des mêmes événements, et en même temps. Pourtant, des différences nettes ressortent comme nous le verrons.

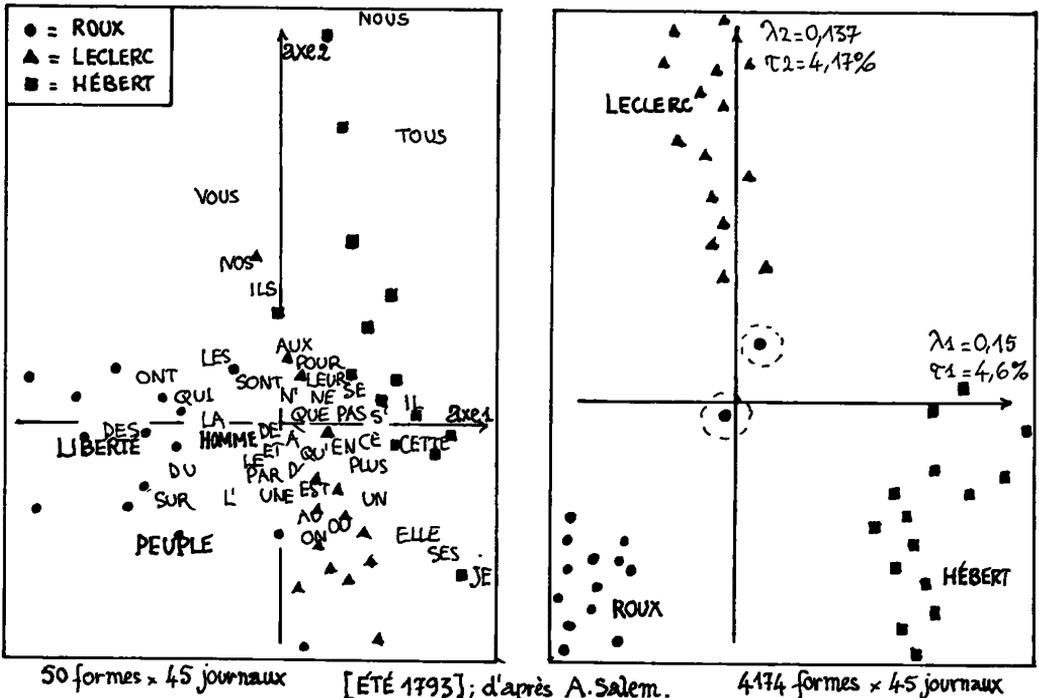
L'étude repose sur le vocabulaire, mais sans aucune élaboration préalable du texte : on a compté les formes ; en appelant *forme* l'intervalle compris entre deux blancs, sans chercher à lever les ambiguïtés (e.g. entre PORTE, nom, et PORTE du verbe PORTER). Les textes étant saisis, on construit automatiquement un tableau ayant 45 colonnes dont chacune concerne un texte (journal) ; chaque ligne concerne un mot, le nombre de lignes étant fixé librement selon e.g. un critère de fréquence ; on a à l'intersection de la ligne *i* et de la col. *j* le nombre des occurrences de la forme (mot) *i* dans le journal *j*.

Voici d'abord le plan (1,2) obtenu, si l'on se borne aux 50 formes les plus fréquentes. Ces 50 formes sont toutes des mots vides, excepté HOMME, LIBERTE, PEUPLE qui pourraient être pleins.

Mais l'abus qu'on en faisait ne les avait-il pas vidés de leur contenu ? Les journaux de Roux s'opposent à ceux de Hébert ; ceux de Leclerc occupent une position intermédiaire.

Considérons maintenant les résultats quand on utilise le maximum d'information : on prend toutes les formes sauf les hapax (plus précisément les mots qui ne se rencontrent que dans un seul journal). Les trois auteurs sont séparés, exceptés deux journaux de Roux qui se placent au centre.

La difficulté pour interpréter une telle analyse est que les formes (ici 4174) sont trop nombreuses pour être imprimées sur un graphique. Il faut choisir ; choix difficile parce qu'on ne sait pas pourquoi ces formes coexistent. Qu'est-ce qu'une forme telle que FAISONS ? C'est le verbe FAIRE ; mais c'est aussi NOUS, indication de personne.



Utiliser *FAISONS*, c'est s'identifier à ceux auxquels on s'adresse, ce qui peut être plus important que d'utiliser le verbe *FAIRE*. Un autre exemple montrant l'intérêt de l'usage des formes brutes plutôt que la lemmatisation s'est rencontré dans une analyse de textes faite par Melle A. Aït Hamlat (cf. CAD Vol IX n° 2 pp 173 sqq ; 1984) sur des comptes rendus de visite industrielle. L'entreprise avait fait les frais d'une lemmatisation, et le verbe *AGIR* sortait avec une fréquence élevée. Mais ce verbe ne figurait dans ces textes que sous la forme : *IL S'AGIT*. Il n'y avait donc pas une miette d'action dans l'affaire. (Ici, il eût fallu, avec A. Salem, faire un inventaire des segments répétés"; cf. CAD Vol IX, 1984, n° 4 pp 489-500 et Actes du Colloque ; mais à défaut, la forme *AGIT* est ici plus évocatrice que le mot *AGIR*).

Ainsi, dans bien des cas, utiliser des formes (ce qui est imprimé entre deux blancs) vaut mieux que de lemmatiser. Mais l'idéal serait de dénombrer (outre les segments répétés) des morphèmes : non ceux de la structure de surface, conçus comme des segments (*FAIS/ONS*) ; mais des morphèmes trouvés au niveau de la composition logique (verbe *FAIRE* + Indicatif Présent + 1-ère personne du pluriel).

5 Genre et style d'auteur dans la littérature chinoise du XX-ème siècle (d'après Peng Zheng-Ce, non publié) : Nous poursuivons notre voyage et à propos de la langue chinoise rencontrons ce problème : comment traiter une Schriftsprache. Un statisticien chinois, Mr Peng Zheng-Ce nous a soumis de Wuhan ses recherches en analyse factorielle, puis envoyé ses données sur lesquelles des classifications ont été faites par un statisticien de Hong-Kong : Mr Cheung Y.L.. Il s'agit de 12 oeuvres dont 4 romans et 8 pièces de théâtre. (les romans sont longs, et l'on gagnerait à utiliser comme unités textuelles, non

les romans entiers, mais des parties ou chapitres ou autres divisions).

La première analyse fut faite en utilisant 380 caractères chinois, ce qui paraît hasardeux si l'on sait que les caractères sont, plutôt que des mots, des éléments constitutifs des mots, parfois porteurs d'atomes de sens mal défini. Le résultat était assez mauvais du fait de certains de ces caractères qui intervenaient dans des noms propres.

On sait qu'il existe une très bonne définition de l'oeuvre de Corneille : un texte est de Corneille si on y emploie les mots : CHIMENE, POMPEE, CESAR etc. ; on pourra discriminer entre Corneille et Racine d'après la liste des personnages des pièces de ces deux auteurs, TITE et BERENICE étant seuls en commun. A quoi bon ? l'analyse discriminante devrait délimiter non l'oeuvre actuelle, écrite par l'auteur, mais son univers potentiel, fait d'idées et de style, plutôt que de noms de personnages.

Un premier nettoyage enlevant entre autres les constituants des noms propres réduit le nombre des caractères à 135. L'opposition entre le théâtre et les romans apparaît magnifiquement ; mais c'est un classique du genre (qui a touché la Chine par cette analyse) : le théâtre se distingue du roman parce qu'il contient plus de dialogue et que, de ce fait, les pronoms WO (je), NI (tu) y prédominent sur le pronom TA de la 3-ème personne.

On élimine donc les pronoms. Subsiste une opposition sur le 1-er axe entre théâtre et roman ; et cette opposition est confirmée à la classification automatique ; mais maintenant sur l'axe 2, les pièces de théâtre et le roman de l'auteur Lao She se trouvent groupés du même côté de l'origine, tout en étant séparés sur l'axe 1 et à la classification automatique (un seul roman de Lao She figure dans les données : Chameau le Bonheur ou Pousse-pousse).

On fait alors une épuration encore plus radicale : on supprime tous les mots qui contribuent à faire les axes, i.e. qui ont l'influence la plus grande sur l'analyse ; de façon à créer une situation *a priori* aussi difficile que possible, sans aucun indice net. L'interprétation de l'axe 1 subsiste pourtant ; et de plus, à la classification on a, enfin, un flot qui comprend à la fois les pièces de théâtre et le roman de Lao She, avec toutefois dans cette classe un roman d'un autre auteur, Zhao Shuli, qui dans toutes les analyses est très proche de celui de Lao She.

Reste à savoir sur quoi se fondent cette discrimination des genres et cette classification. L'embarras n'est pas si grand qu'avec les 4000 mots des journaux de la Révolution Française ; mais on a 100 caractères, ce qui est beaucoup pour un barbare d'Occident. On fait alors une classification automatique sur les caractères et l'on croise une partition des caractères en 10 classes avec les 12 oeuvres. Au lieu de dire : dans cette oeuvre il y a tant de fois ce caractère, on dit ici : dans cette oeuvre il y a tant d'occurrences de caractères de cette classe. La classification sur les oeuvres s'interprète dès lors à l'aide des classes de caractères : dans les oeuvres de telle classe prédominent telles classes de caractères. On trouve que l'opposition entre théâtre et roman est fondée sur deux groupes de caractères dans chacun desquels la majorité des caractères ont la clé de la parole (interjections etc.). Les caractères de l'un des groupes prédominent dans le théâtre ; ceux de l'autre groupe dans le roman.

哼 啊 嘿 嗎
THEATRE

啊 吧 呀
ROMAN

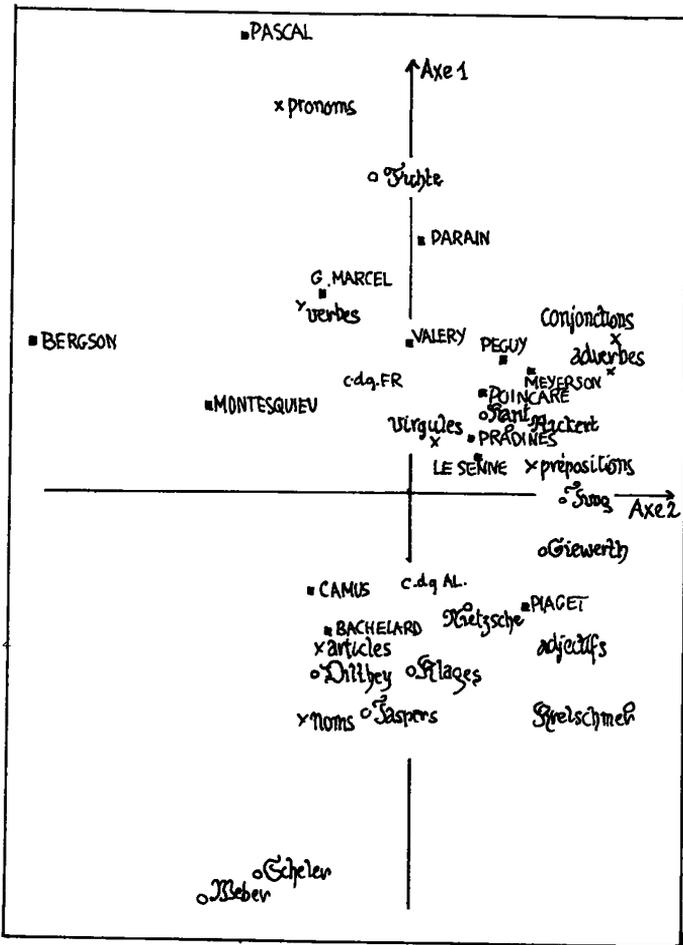
Mais la conclusion la plus importante de cette coopération avec la Chine est qu'il semble possible de passer au travers de la différence entre genres pour découvrir les auteurs, reconnaître qu'une pièce et un roman ont même auteur. C'est un problème très difficile qu'on doit aborder en étudiant des situations pour lesquelles on a des référents extérieurs certains. Il est vrai qu'un tel programme ne pique pas la curiosité. Personne ne se soucie de reconnaître que l'auteur des pièces de Voltaire, des lettres de Voltaire, de la *Fenriade* et du Dictionnaire Philosophique est le même homme : tout le monde le sait. Mais c'est seulement par l'analyse d'un corpus gigantesque renfermant avec les oeuvres de Voltaire celles de plusieurs contemporains de Voltaire, qu'on arrivera à savoir comment un seul et même auteur qui écrit sur divers sujets en divers genres, garde des traits distinctifs qui lui sont propres (comme en parlant on garde la même voix, quoi qu'on dise).

6 Parties du discours en allemand, en français et en grec (d'après

J.M. Zerb, in PRAT 3 pp 227 sqq ; et Ch. Rütten à paraître) :

Avant de repartir pour le Moyen-Orient, terre de la Bible, faisons une escale en Europe. Il y a 15 ou 20 ans, mon ami J.M. Zerb, actuellement professeur dans une université parisienne, m'a communiqué des décomptes de parties du discours (noms, verbes etc.) faits avec ses étudiants alors qu'il était assistant à l'Université de Fambourø. Ces décomptes portent sur 12 auteurs allemands et 14 auteurs français avec pour chacun une grande longueur de texte (quelque 10 000 mots). Il aurait été préférable de faire pour chaque auteur des dénombrements en plusieurs tranches, mais tout avait été fondu et nous avons dû prendre les données telles quelles. Le tableau analysé comporte en lignes les auteurs français et allemands ; et en colonnes les parties du discours : verbes, adjectifs etc.. Il en est résulté un premier facteur opposant le groupe nominal associé à une majorité d'auteurs allemands, au groupe verbal associé aux auteurs français (i.e. noms, adjectifs articles d'un côté, verbes et adverbes de l'autre) ; et un second facteur opposant les éléments non suppressibles (verbes, noms) aux éléments suppressibles (adjectifs, adverbes) : quand on dit CHAPEAU ROUGE ou COURIR VITE on peut dire CHAPEAU ou COURIR ; voilà en quoi les adjectifs et adverbes s'opposent à noms et verbes comme "suppressibles" à "non suppressibles".

Ces oppositions trouvées par J.M. Zerb semblent devoir être universelles. Pourtant, à partir de dénombrements des parties du discours des chapitres d'Aristote, le Professeur Rütten de Liège a obtenu un résultat différent. A une extrémité du premier axe : subordonnant, verbe et adverbe suivis d'adjectif ; à l'autre extrémité : préposition, article et nom. L'adjectif n'est pas associé au groupe nominal. C'est que souvent chez Aristote, l'adjectif ne qualifie pas un nom mais une proposition entière comme dans : "cela est possible" ou "cela est impossible" etc. et il y a beaucoup de phrases nominales ; il y en a tant que Boèce se vante dans son commentaire d'avoir corrigé Aristote qui prend un malin plaisir à changer les noms en verbes et les verbes en noms (*Si quid suo more Aristoteles nominum verborumque mutatione turbavit, nos intelligentiam servientes ad consuetum vocabulum reducamus* (PL LXIV Col 793)).



Correspondance entre auteurs et parties du discours :
(d'après J. M. Zemb) . [T B] § 6

7 Dénombrement automatique des mots arabes par classes morphosémantiques (d'après A. Abi Farah, Thèse Paris 1985) : La langue arabe est comme l'hébreu une langue sémitique. La grande différence entre l'hébreu biblique et l'arabe classique est qu'on a dans cette dernière langue une littérature infinie, tandis que la Bible est un corpus clos dont il n'est guère possible de corroborer l'étude par celle des textes contemporains écrits dans la même langue. C'est pourquoi quelque grand que soit l'intérêt des procédures d'analyse automatique de l'hébreu, il est bien clair que le texte mérite qu'on l'analyse complètement sans le secours de l'ordinateur, ce qui d'ailleurs a, dans une large mesure, été fait.

Pour comprendre la difficulté et l'intérêt du travail de M^r A.A.F., il faut avoir présentes à l'esprit quelques particularités de la langue arabe, que nous rappellerons au risque de lasser la patience des orientalistes ! D'une part les barrières entre mots ne se voient pas : à

l'intérieur d'un mot, il peut y avoir un blanc (car certaines lettres ne se lient jamais à celles qui les suivent) ; et au contraire, deux mots différents peuvent être écrits si proches l'un de l'autre que l'oeil ne les sépare pas. D'autre part en arabe comme en hébreu, il est de règle de n'écrire ou de n'imprimer que les consonnes et certaines voyelles longues (matres lectionis), en sorte que le même mot graphique MLKT où l'on reconnaît la base verbale MLK (posséder) peut, selon la vocalisation qu'on lui attribue parmi huit possibles, d'après le contexte, être à l'actif ou au passif et se rapporter à l'une ou l'autre des quatre personnes du singulier (1-ère ; 2-ème du masc. ou du fém. ; 3-ème du féminin). De 8, ce nombre doit même être porté à 16 car le verbe admet une forme factitive (rendre possesseur) caractérisée par le redoublement de la deuxième consonne radicale; redoublement dont le signe graphique est généralement mis à l'égal des voyelles. Certes, il existe une "scriptio plena" donnant toute précision au lecteur et quasi obligatoire dans la Bible hébraïque ; et aussi dans le Coran. Mais un texte usuel comporte au plus quelques voyelles destinées à lever les ambiguïtés sur lesquelles peut trébucher un lecteur même compétent.

Partir d'un texte arabe saisi tel qu'il est imprimé, puis en restituer par un programme de traitement linguistique automatique d'abord la segmentation en des mots, puis la vocalisation, c'est en 1985 une tâche qu'on ne sait même pas aborder. Ceci implique qu'une prédiction est indispensable avant de traiter un texte arabe sur ordinateur. Le moins qu'on puisse faire est de saisir le texte avec séparateurs, voyelles et lettres redoublées ; ce qui équivaut à créer un enregistrement isomorphe non à la graphie arabe usuelle, mais à celle des Orientalistes : MALAKTU (j'ai possédé) ; MULLIKAT (elle a été rendue possesseur) etc. . Transcription aisée s'il s'agit d'un texte moderne, mais épineuse avec un manuscrit ancien. On voit que le codage informatique des écritures non latines (présenté à Louvain-la-Neuve par J. Langton) et particulièrement de l'arabe (thème de P. Branca à Nice) est affaire de linguistique autant et plus que de techniques des écrans et claviers.

Puisque la prédiction confine à une préanalyse, on peut pousser celle-ci un peu au-delà d'une simple scriptio plena : ainsi M^r Abi Farah sépare des formes verbales leurs désinences personnelles (que la grammaire arabe appelle traditionnellement "pronoms" ; les regardant comme un sujet accolé à la base).

Pour aller au-delà, A. A.F. a établi une première liste de 80 catégories morphosémantiques, fondée sur une répartition des mots en classes qui, dans ses grandes lignes, est issue de la tradition grammaticale si riche de la langue arabe. Nous nous bornons à citer : formes des verbes, modèles des noms d'agents, des noms d'actions, diminutifs, mots d'exception (sauf...), réponses, verbes copules logiques (qu'on appelle poétiquement "être et ses soeurs"). Il a eu l'imprudence de s'y ajouter la notion si difficile à cerner de "sens propre" et de "sens figuré".

Un premier essai d'indexation de textes par une dizaine de lecteurs humains a abouti, après analyse factorielle, à une typologie où les différences entre experts analystes interféraient avec celles entre auteurs des textes saisis. M^r A. A.F. a donc eu l'audace de revoir sa liste de catégories et de créer un programme qui ventile automatiquement suivant ces catégories tous les mots d'un texte saisi comme nous l'avons expliqué. Cette analyse automatique n'est pas pleinement correcte : il y a 20% d'erreur, ce qui est beaucoup. Mais l'essentiel est que l'algorithme répond de manière uniforme à un texte quel qu'il soit, et l'analyse qui en résulte peut servir de base à une typologie objective des textes ; typologie pertinente dans la mesure où l'algorithme, même s'il n'a pas tout saisi, a recueilli de nombreuses informations. L'analyse morphologique est un réactif chimique en quelque

sorte : on y plonge le texte et l'on regarde la couleur qui en résulte. Et cette coloration est satisfaisante : si on analyse le tableau croisant un ensemble de textes indexés avec l'ensemble des catégories retenues, (analyse tout analogue aux travaux de M^r J.M. Zemb et de M^r Ch. Rütten, mais avec une grille descriptive plus fine), on trouve bien groupés les textes ou fragments afférents à un même auteur ; avec au sein des oeuvres une typologie dont l'étude mérite d'être poursuivie.

8 Distribution des noms dans les sourates du Coran (d'après I. Kharchaf ; documents en langue arabe publiés par Dāru n našri' l mağribiyyū, Casablanca ; et thèse en préparation) : Nous rendons compte brièvement de cette étude, parce que d'une part elle offre à la critique textuelle un intéressant terme de comparaison et que, d'autre part, elle corrobore les belles recherches dont Mme R. Gabr a rendu compte à Nice.

Au non-spécialiste il suffira de dire que le Coran est divisé en 114 chapitres de très inégales longueurs appelés en arabe sura : on dit, en français, "une sourate". En conformité avec l'histoire de la prédication de Muhammad, le texte du Coran comprend une partie mecquoise et une partie médinoise : il y a des sourates de La Mecque, des sourates de Médine (et quelques sourates dont des versets sont rapportés aux deux villes). La différence est, en bref, qu'à La Mecque on a la proclamation, face à un public hostile, d'un enseignement sur les origines et les fins dernières dont beaucoup d'éléments ne diffèrent pas de l'enseignement de la Bible, tandis qu'à Médine, après la hijra, point de départ du calendrier islamique, c'est la fondation d'une communauté nouvelle avec ses lois.

En analysant le tableau 155 × 114 (croisant avec les 114 sourates l'ensemble des 155 noms les plus fréquents), I. Kharchaf obtient un premier axe sur lequel sourates de La Mecque et sourates de Médine s'opposent ; chaque groupe étant associé à son vocabulaire caractéristique.

Qui étudie le texte de la Bible s'intéressera particulièrement à la place de deux noms qui servent à désigner Dieu : RABB (Seigneur et ALLAH. Le premier est associé aux sourates de La Mecque (enseignement général sur Dieu comme créateur, sur les patriarches et les prophètes.) ; le second, ALLAH, va avec les sourates de Médine (prescriptions religieuses). Voilà dans un texte propre à un milieu religieux sémitique l'usage simultané de deux noms pour Dieu ; attirés comme le sont YAHWEH et ELOIM dans le Pentateuque, par des contextes différents. Et cela alors que l'histoire assez bien connue du texte du Coran ne permet pas, comme on l'a fait pour le Pentateuque de postuler l'assemblage de fragments longtemps après leur composition.

Nous signalerons encore que 'INSĀN (homme ; de l'espèce humaine, sans distinction de sexe) est du côté des sourates de La Mecque. Alors que les noms qui servent à distinguer homme de femme (tels que RIJĀL et NISA' ou DAKAR et 'UNTA, noms de faible fréquence et que, donc, I. Kh. n'a pas retenus) sont, ainsi que l'a montré Mme R. Gabr, du côté de Médine : ce que laissait attendre la brève explication donnée ci-dessus du partage du texte.

9 Distribution des vocables dans le texte hébraïque et dans la traduction des Septante (analyses de Ch. Arbache) : Je me suis demandé si les analyses faites par le P^r G.E. Weil sur le texte hébreu d'Isaïe passeraient au travers du filtre de la traduction.

Posons le problème en toute généralité. On traduit un texte. On a fait dans la langue d'origine un tableau croisant chapitres et mots fréquents ; et l'on fait un second tableau croisant les chapitres traduits et les mots de la traduction. L'analyse produira-t-elle une même typologie des chapitres d'après l'un et l'autre tableau ? Si tel est le cas, on pourra supposer ou bien que la typologie des chapitres d'origine est fondée sur le contenu plutôt que sur le style (ou d'éventuelles différences d'auteurs) ; ou bien que le traducteur varie son style en parfaite harmonie avec l'original) ; à moins que la traduction ne soit tout simplement un calque mot pour mot !

Bien que la question se pose particulièrement à propos des traductions de la Bible en langue moderne (traductions dont le style est généralement l'objet d'une attention particulière) elle nous intéresse encore plus pour les textes anciens dans la mesure où l'on a pu supposer que le texte hébreu que nous possédons serait une mise à jour d'un texte antérieur, selon un processus peut-être analogue à celui de la traduction des Septante dont je parlerai maintenant.

J'ai appris au Colloque que le P^r Tov et ses collaborateurs avaient achevé cette oeuvre admirable d'un alignement du texte massorétique entier de la Bible hébraïque avec la version des Septante. Si j'ose encore rendre compte d'une expérience qui n'a porté que sur deux chapitres, c'est seulement parce que l'analyse des données me paraît ouvrir la voie à des essais en grande échelle d'alignement automatique.

JOSUE I 1,2

I, καὶ ἐγένετο μετὰ τὴν τελευταίην μωνοῆ
εἶπε κύριος τῷ Ἰησοῦ νῦν κατὰ τὴν ὑπουργίαν
μωνοῆ, λέγων μωνοῆς ὁ θεράπων μου
τετελευτήκει· νῦν σὺν ἀναστάσι διαβήθη τὸν
ἰορδάνην σὺ καὶ πᾶς ὁ λαὸς οὗτος εἰς τὴν
γῆν, ἣν ἐγὼ δίδωμι αὐτοῖς.

17... μωνοῆς ὁ παῖς μου.

11... κατὰ μέσον τῆς παρεμβολῆς τοῦ λαοῦ..

II REGUM V 14,15

καὶ κατέβη αμαμάν,
καὶ ἐβαπτίσαστο ἐν τῷ ἰορδάνῃ, ἐπιτάκας
κατὰ τὸ ῥῆμα ἑλισαίᾳ, καὶ ἐπέστρεψεν
ἡ σὰρξ αὐτοῦ ὡς σὰρξ παιδαρίου μικροῦ
καὶ ἑκαθάρισθη.
καὶ ἐπέστρεψεν πρὸς ἑλισαίαν αὐτὸς καὶ
πᾶσα ἡ παρεμβολὴ αὐτοῦ, καὶ ἦλθε
καὶ ἔστη ἐνώπιον αὐτοῦ, καὶ εἶπεν· ἰδοὺ
ἐγνωκα ὅτι σὺ εἶσι θεὸς ἐν πνεύματι καὶ γῆ,
ὅτι ἀλλ' ἡ ἐν τῷ ἰσραὴλ καὶ νῦν λάβε
τὴν εὐλογίαν παρὰ τοῦ δούλου σου.

IV 43 καὶ εἶπεν ὁ λειτουργὸς αὐτοῦ·

[TEXTE BIBLE] § 9

יְהִי אַחֲרַי מוֹת מֹשֶׁה עַבְד יְהוָה
וַיֹּאמֶר יְהוָה אֵל יְהוֹשֻׁעַ בֶּן־נוּן מִשְׁרַת
מֹשֶׁה לֵאמֹר; מֹשֶׁה עַבְדִּי מֵת וְעַתָּה
קוּם עִבֵר אֶת־יְהוּדָה הַזֹּאת אִתָּה וְכָל־
הָעָם הַזֶּה אֶל הָאָרֶץ אֲשֶׁר אָנֹכִי נֹתֵן
לָהֶם לְבְנֵי יִשְׂרָאֵל

... משה עבדי
... בקרב המחנה

... וירד ויטבל בירדן
שבע פעמים כדבר איש האלהים
וישב בשרו כבשר נער קמן
ויטהר

וישב אל-איש האלהים הוא וכל-
מהנהו ויבא ויעמד לפניו ויאמר
הנהינא ידעתי כי אין אלהים בכל-
הארץ כי אם-בישראל ועתה קחנא
ברכה מאת עבדך

וַיֹּאמֶר בְּמִשְׁרַתוֹ...

Les chapitres analysés ont été d'abord 2 Rois V 1- 19 (guérison de Naaman par Elisée) puis Josué I (préparation à l'entrée en Terre Promise). J'ai pris les textes dans le tome II de la Polyglotte de Vigouroux, et indiqué le vocabulaire d'après le dictionnaire grec de Bailly et *The Analytical Hebrew and Chaldee Lexicon* de Davidson (éd.).

Bagster & Sons) ; auquel je me suis fié absolument pour rattacher les formes à leur lemme.

Après ce que je vous en ai dit, je dois justifier ici le recours à une lemmatisation. Le grec et l'hébreu diffèrent quant à la structure ; bien plus, les différences de structure sont accentuées par des conceptions très différentes de l'écriture et surtout du découpage de la chaîne parlée en mots. L'écriture sémitique est agglutinante, liant à un nom préposition, article, pronom, suffixe d'appartenance... et de même pour le verbe. Pour accéder à une correspondance biunivoque entre unités de texte et unités de la traduction, il aurait fallu aux deux langues des découpages fondés sur un même principe : une telle réforme linguistique était certes au-dessus de nos forces !

Les données que j'ai recueillies se prêtent à de multiples traitements statistiques. Je me bornerai à rendre compte d'une analyse due à M^r Ch. Arbache (qu'on ne confondra pas avec son cousin, le bibliste S. Arbache, présent au Colloque...).

Précisons les critères adoptés pour coder les textes. On a retenu les mots pleins grecs ou hébreux (noms ou verbes quasi exclusivement) qui se rencontreraient au moins dans deux versets différents parmi les 37 considérés ici (de JOSUE I ou de 2 ROIS). On a construit un tableau croisant l'ensemble des versets avec l'ensemble des mots ; avec e.g. à l'intersection de la ligne m et de la colonne v le nombre de fois que le mot m se rencontre dans le verset v . Par exemple $k(\text{POTA}, J4) = 2$, parce que le mot $\pi\omicron\tau\alpha\mu\acute{o}\varsigma$ se trouve 2 fois dans le verset Jos I 4 du texte des Septante ; et de même $k(\text{NHK}, J4) = 2$ parce que $\eta\eta$ est deux fois employé dans le texte hébraïque du même verset. On peut dire que j'ai considéré chaque verset comme composé d'un texte hébreu et d'un texte grec mis bout à bout ; et procédé sur cet objet hybride aux mêmes décomptes que ceux faits dans les §§ précédents pour des unités textuelles plus longues (chap.) écrites en une seule langue.

Le tableau ainsi construit croise 37 versets et 133 mots (hébreux ou grecs) ; parmi ceux-ci j'ai constaté qu'environ 100 allaient par paires formées chacune d'un mot hébreu et de sa traduction grecque ; tel : *fleuve* = $\eta\eta$ = $\pi\omicron\tau\alpha\mu\acute{o}\varsigma$; avec pour les deux mots des distributions soit identiques, soit peu modifiées. C'est pourquoi la classification automatique des mots a fait précisément apparaître ces paires au niveau inférieur de la hiérarchie, réalisant ainsi, en partie, ce que le Pr Tov appellerait un alignement automatique du texte et de sa traduction.

Avant d'exposer les résultats de l'analyse factorielle, il vaut la peine de s'arrêter à quelques détails des concordances et des discordances. On a *parler* = דַּבֵּר = $\lambda\alpha\lambda\acute{\epsilon}\omega$; *parole* = דְּבָר = $\rho\acute{\eta}\nu\alpha$; et *dire* (verbe transitif à la différence de *parler*) = אָמַר = $\{\lambda\acute{\epsilon}\gamma\omega, \epsilon\dot{\iota}\pi\omicron\nu\}$; ... Le mot grec $\kappa\upsilon\rho\iota\omicron\varsigma$, *seigneur* correspond au nom hébreu דָּן et aussi au tétragramme divin יהוה : ce qui suggère que les Septante lisaient déjà דָּן (*adonai*) comme on l'a fait dans la suite : c'est pourquoi dans les relevés, on s'est fondé sur cette lecture.

Comme on peut le vérifier sur les quelques fragments de textes joints à mon exposé, là où le texte hébreu de JOSUE dit "*Moise l'esclave de Dieu*", "*l'esclave de Dieu*" a disparu du texte grec. De même dans 2 ROIS, "*l'homme de Dieu*" אִישׁ הָאֱלֹהִים est plusieurs fois traduit Elisée : $\epsilon\lambda\iota\sigma\alpha\dot{\iota}\acute{\epsilon}$. Pourtant la traduction littérale $\delta\ \acute{\alpha}\nu\theta\rho\omega\pi\omicron\varsigma\ \tau\omicron\upsilon\ \theta\epsilon\omicron\upsilon$ se trouve dans 2 ROIS IV.

Le mot hébreu עַבְד serviteur (ou esclave) est traduit $\delta\omicron\upsilon\lambda\omicron\varsigma$ dans 2 ROIS V et παῖς dans JOSUE I. Au contraire, l'hébreu צְבָא est traduit par παρεμβολή dans les deux textes, bien qu'en JOSUE I il désigne l'armée du peuple hébreu et en 2 ROIS V, la suite de Naaman.

Il est remarquable que dans deux versets consécutifs de 2 ROIS V, בשרו soit traduit par le même $\epsilon\pi\epsilon\sigma\tau\rho\acute{\epsilon}\psi\epsilon\nu$; bien qu'entre les deux la différence de sens soit celle qu'il y a en français entre "redevenir" (la chair du lépreux redevenue saine) et "revenir" (Naaman revient vers le prophète qui l'a guéri).

Somme toute, la traduction des Septante apparaît assez littérale pour qu'on reproduise d'après elle un "urtext" hébraïque ainsi que l'a accompli le P^r Tov. Les différences avec le texte massorétique impliquent soit l'existence d'un autre état du texte hébreu chez les Septante ; soit comme on l'a parfois conjecturé dans le passé, un parti pris de dissimulation ("*Sententias dogmatum septuaginta interpretes ne arcana fidei regi revelarent, interdum occultaverunt*" dit en substance Saint Jérôme, cité d'après R. Cornely *Cursus Sanctae Scripturae* T I p. 347 Lethelleux 1925). D'autre part l'utilisateur de traductions différentes pour un même mot (cf. *supra* $\pi\alpha\iota\varsigma$ et $\delta\omicron\upsilon\lambda\omicron\varsigma$) suggère une analyse statistique ultérieure destinée à classer les livres ou parties de livres d'après leur mode de traduction : autre problème familier (dont les données consistent d'ailleurs plutôt en mots outils - אך traduit par ouv! - qu'en mots pleins).

Il est temps de revenir au graphique issu de l'analyse factorielle. Celui-ci nous réserve une agréable surprise (confirmée par la CAH) ; dans le plan (1,2), non seulement beaucoup de mots (écrits sur le graphique dans la mesure de la place disponible) vont, ainsi qu'on l'attendait, par paires, le grec se superposant à l'hébreu, ou s'en écartant peu, mais les versets des deux textes se séparent : on a dans le demi-plan de gauche ($F_1 < 0$) les versets de JOSUE I (notés J suivi d'un chiffre et ceux de 2 ROIS (notés N, comme Naaman, suivi d'un chiffre) à droite ($F_1 > 0$). Seuls font exception J1 et N17. Discrimination d'autant plus remarquable qu'ont été écartés les noms propres qui l'auraient facilitée (Josué, Naaman).

Au lecteur attentif et averti je dois encore une précision qui m'a été demandée à Louvain-la-Neuve après l'exposé. Des taux aussi faibles que ceux de 8% et 7,3% inscrits sur les axes ne sont-ils pas suspects ? Je répondrai d'abord en des termes accessibles à tous : une discrimination quasi parfaite entre les versets des deux chapitres ne peut être fortuite : elle valide donc le facteur 1. En termes plus techniques, j'ajouterais qu'il est certain que deux versets, unités textuelles courtes, ne peuvent que différer grandement quant au profil lexical, à moins de se répéter. Rendre compte de toutes ces différences requiert l'ensemble des 36 facteurs que produit l'analyse et ne peut s'achever même à 60% dans le plan (1,2). Le 1-er axe toutefois a l'intérêt de révéler une structure générale (JOSUE I \neq 2 ROIS V), tandis que dès le 2-ème axe apparaissent des effets locaux (e.g. groupes de versets N5, N6, N7, N8, i.e. 2 ROIS V 5 - 8 dans le quadrant : $F_1 > 0$, $F_2 > 0$).

10 Analyse de la Table Pastorale de la Bible (d'après I. Kharchaf, thèse à paraître) : La table est issue de l'abbaye de Maredsous à laquelle appartient l'éminent organisateur du présent Colloque. Nous rappellerons qu'on y trouve dans l'ordre alphabétique, de Aaron à Zacharie, une présentation méthodique des mots, des idées, des sujets et des noms principaux du Livre Saint avec des références par livres et versets. Voici par exemple ce qu'on lit sous le mot *Abaissement* :

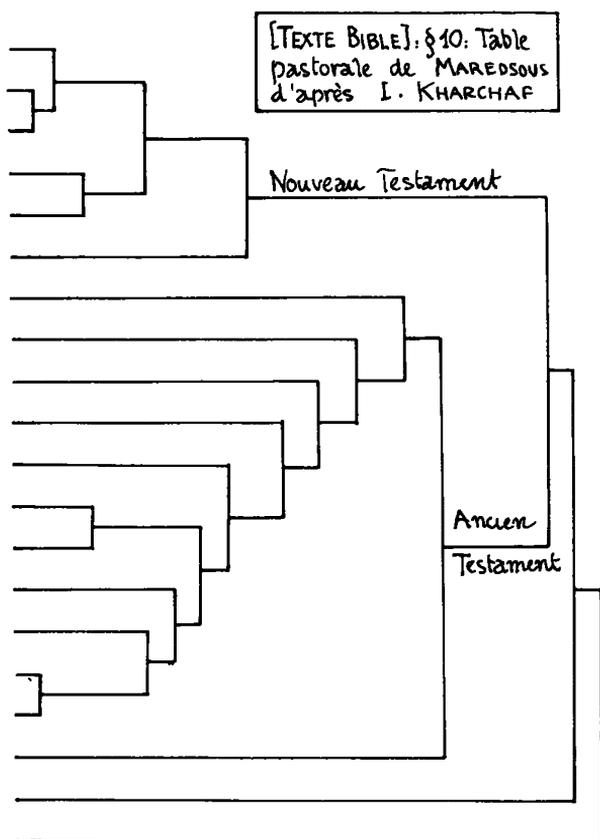
Jud VI - 19 : prends en pitié l'abaissement de notre nation
Pr XXIX- 23 : l'orgueil d'un homme le mène à l'abaissement
Sir II - 4 : dans l'abaissement aie patience
VII - 19 : humilie profondément ton esprit
XX - 11 : tel relèvera la tête après son abaissement...

A partir de ces données, I. Kharchaf a construit un tableau de correspondance croisant avec l'ensemble des livres de la Bible les 520 mots ou locutions faisant l'objet du plus grand nombre de références. A l'intersection de la ligne m et de la colonne l, on a le nombre des références au livre l trouvées sous le mot m. Ce tableau a été soumis à l'analyse factorielle et à la classification automatique. Nous nous bornons ici à présenter brièvement la classification obtenue sur l'ensemble des livres.

D'abord, une particularité : le 1-er livre des Maccabées et celui d'Esther se séparent de tout le reste de la Bible. C'est que dans la Table Pastorale, on a multiplié les références à certains personnages qui n'apparaissent que dans l'un ou l'autre de ces livres.

Ensuite, le Nouveau et l'Ancien Testament sont parfaitement séparés : on ne s'en étonnera pas non plus ; car chacun des mots *Apostolat*,

PHILé ; 2COR, COLOs, 2JEAN
 EPHE ; 2TIM, 1COR, 2THE, JUDE
 JEAN ; GALAT, PHILIP ;
 ROMA ; 1TIM, TITE, LUC, MATT, MARC
 ACTE, APOC, HEBR, 2PIER ;
 JACQ ; 1THEs, 1PIER, 1JEAN ;
 1SAM ; 1CHR, 2SAM
 2ROI ; 2CHR ; 1ROI
 GENE
 EXOD | NOMB
 2MACe, JOSUé, JUGEs, ABDAas ;
 LEVI, NAHUM ;
 ISAÏe ; NEHEm, JOEL ; JERE m, EZECH
 CANT, RUTH ; SIRAe ;
 JOB, TOBI, LAMEN ; BARU, MALA, PROV, SAC.
 PSAU, OSEE, DAN, ESDR, ZACH, AGG, HAB, SOPH.
 AMOS, MICH, DEUT, JUDITH
 QOHElet
 ESTHer
 1MACcabées



Evangile, Foi, Jésus Christ... renvoie à la quasi-totalité des livres du N. T.. C'est sur le détail des subdivisions des deux parties de la Bible que nous appellerons l'attention des exégètes. Pourquoi, tandis que les Evangiles synoptiques (Luc, Matthieu, Marc) vont avec trois épîtres de St Paul (Romains, 1-ère à Timothée, Tite), l'Evangile selon Saint Jean va-t-il avec deux autres épîtres (Galates et Philippins) ? Et dans l'A. T., comment expliquer la place du Deutéronome, (dont le sigle DEUT est encadré comme ceux des autres livres du Pentateuque et

des Evangiles) à proximité de nombreux livres prophétiques (dont les sigles sont surlignés) ainsi que de Judith, de Daniel, d'Esdras et du Psautier ?

L'analyse des données peut expliquer comment ont été agrégées les classes de la CAH : il suffit en bref de croiser une partition des livres avec une partition des mots pour savoir (par des calculs précis) entre quelles classes de livres et quelles classes de mots les références de la *Table Pastorale* ont établi les liens les plus étroits qui sont aussi la cause de la cohésion de ces classes.

Mais au fond, les références elles-mêmes résultent d'un choix plus ou moins conscient de ceux qui ont conçu la Table Pastorale. Certes au niveau de chaque livre et de chaque mot les exégètes ont, pas à pas, relevé ce qui leur semblait significatif. Mais le résultat, considéré dans son ensemble a une structure que révèle l'analyse des données. En résulte une vision du contenu de la Bible. Par delà les erreurs de pondération (telles que celles qui ont donné au livre d'Esther un caractère unique qu'il n'a sans doute pas dans l'esprit des auteurs de la Table) n'est-ce pas le conception du message biblique sous-jacente à toute l'entreprise que nous découvrons ? Les exégètes peuvent maintenant regarder leur propre oeuvre et faire un examen de conscience.

11 Prosopographie des personnages de feuilleton (d'après H. Castro,

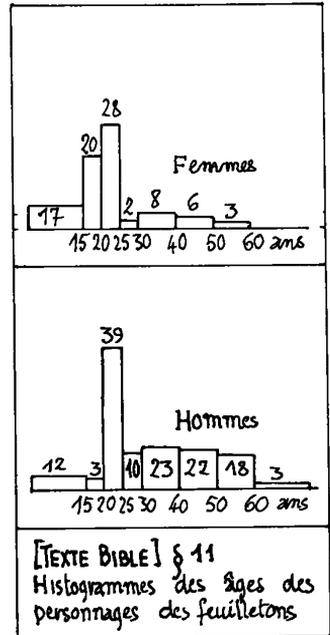
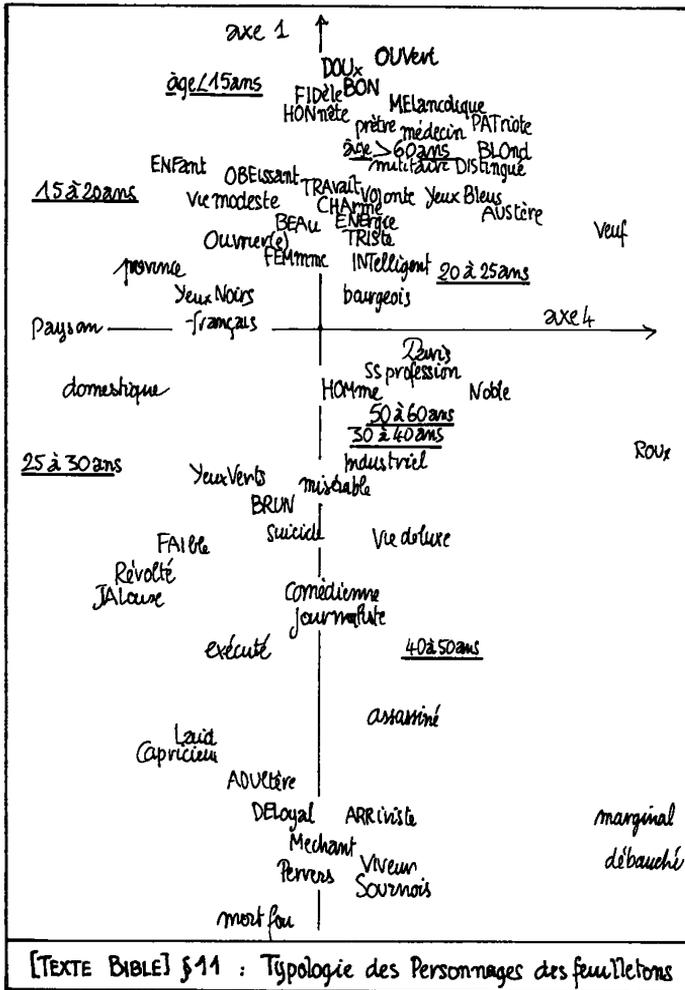
PRAT 3 pp 261 sqq) : Par les yeux des collaborateurs au *Petit Journal* portons maintenant sur l'humanité un regard profane. Le *Petit Journal* est le premier quotidien français qui ait atteint le tirage d'un million d'exemplaires : dès 1863, il est diffusé dans tout le pays, grâce à un réseau de distribution sans rival à l'époque. Mais, ébranlé par la première guerre mondiale, le *Petit Journal* disparut dans la seconde, et n'est pas revenu depuis. A ses lecteurs, le P. J. offrait des nouvelles fraîches et surtout des faits divers et des feuilletons.

Bien que le mot de feuilleton évoque une histoire centrée sur un héros aux multiples aventures, on trouve plutôt dans le P. J. des enchevêtrements de péripéties réquérant une multitude de personnages. Certains ne font que traverser la scène ; mais il y en a dans chaque feuilleton cinq ou dix dont le 10^{ème} est essentiel. Des 31 feuilletons publiés de 1890 à 1894, on a extrait les signalements de 214 personnages : âge, sexe, traits physiques et moraux, comme si le romancier répondait aux questions de l'analyste (ou parfois, comme il arrive dans toute enquête! refusait de répondre..).

Filles ou garçons, hommes ou femmes, quel âge ont-ils quand ils sont présentés au public ? Mis à part quelques enfants, les filles commencent d'exister à 15 ans, les garçons à 10. Après 25 ans on accepte encore bon nombre d'hommes, mais peu de femmes...

A l'analyse factorielle, le premier axe montre une échelle de valeur morale illustrée de traits physiques dont on découvre ou vérifie le sens symbolique. Et d'abord l'âge : avant 25 ans, après 60, l'on est bon, honnête... ; l'âge ingrat dure ici de 25 à 60 ans ; avec entre 40 et 50 ans le paroxysme des troubles. Pourquoi est-il acceptable (F1 faiblement positif) d'avoir les Yeux Noirs ; mais non d'être Brun (F1 < 0) ? Il est rassurant que la beauté du visage aille avec celle de l'âme ; que le Charme ne soit point à craindre. Pour conclure, la Femme est préférée à l'Homme. Tel est l'univers du feuilleton.

Qu'en est-il dans l'Evangile ? Les peintres peuvent à leur goût donner couleur aux cheveux et aux yeux : ni Jean ni les synoptiques ne les contrediront. Mais des foules de la Galilée et de Judée se détachent des figures qui ont âge, sexe, profession, taille parfois... et dont l'âme nous est rendue proche avec une sobriété de style



[TEXTE BIBLE] § 11
Histogrammes des âges des personnages des feuilletons

[TEXTE BIBLE] § 11 : Typologie des Personnages des feuilletons

inimitable. Il vaut la peine de recenser ces personnages, de noter tout ce qui en est dit, de les reconnaître s'ils reviennent... N'est-ce pas de l'Evangile que Jules Romain a pris le titre du long roman de sa génération : "Les hommes de bonne volonté".

12 Sémitologie (d'après PRAT 3) : Avant de conclure, je soumettrai aux participants à ce Colloque, sémitologues pour la plupart, les résultats de deux analyses ; dont la première concerne les similitudes de vocabulaire entre langues chamito-sémitiques et la seconde l'harmonie des phonèmes au sein des racines hébraïques.

Dans son "Essai comparatif sur le vocabulaire et la phonétique du chamito-sémitique", Marcel Cohen donne, rangée dans un ordre phonétique, une liste de racines attestées chacune dans deux au moins des quatre domaines suivants : sémitique ; couchitique ; berbère ; égyptien ancien (et copte) avec pour ces racines des mots qui éventuellement s'y rattachent dans une trentaine de langues de ces

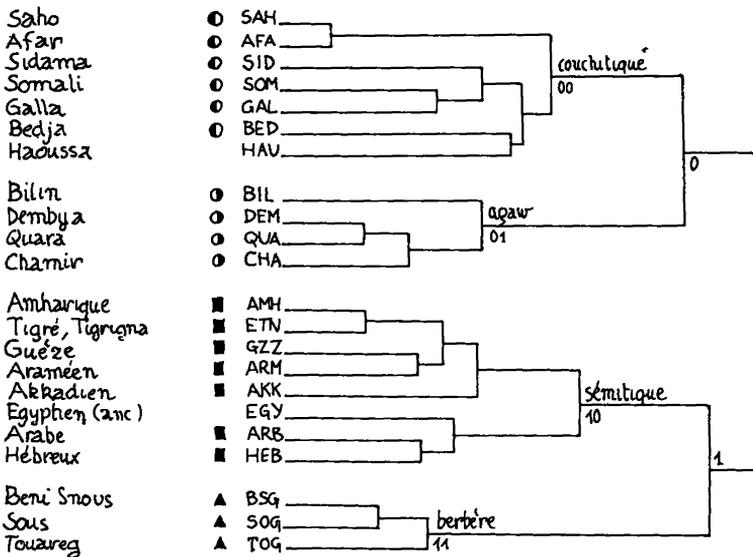
domaines, ainsi que parfois en haoussa. Les racines sont numérotées de 1 à 510 (à quelques suppressions ou additions près, dans l'état définitif de la liste). Voici un exemple :

381 bis. SEM. h. *pōl*, ar. *fūl* "fève"
 BERB. *baw*, ibiw "fève" (sans *l* ; rapport avec le latin *faba* ?)
 COUCH. ag. *bil*; *bal-dangūa* "haricot" (sans *w*) ;

on déchiffre sans peine les initiales h. hébreux, ar. = arabe ; dans le berbère ne sont pas distinguées les langues particulières (dont les plus étudiées par M. Cohen sont le touareg et, au Maroc, la langue du Sous (Agadir) et des Bni snous) ; pour le COUCHitique (dont l'aire géographique débordé peu les frontières de l'Ethiopie) on trouve la mention *agaw* avec spécification d'un dialecte "bilin".

Il est naturel de regarder ces données comme la matière d'un tableau rectangulaire 510 × 30 (510 mots ou racines × 30 langues ou dialectes) avec à l'intersection de la ligne *m* et de la colonne *l* le chiffre 1 si le mot *m* est attesté dans la langue *l* ; zéro sinon. La difficulté est d'une part d'arrêter une liste des langues suffisamment représentées ; d'autre part d'interpréter des mentions générales "SEM" ou "BERB" censées valoir pour plusieurs langues.

Sans préciser nos tâtonnements (dont le détail est dans PRAT 3 pp 303-333) nous donnerons la classification issue du tableau auquel nous nous sommes arrêté (p. 321). L'ensemble des langues et dialectes se divise en deux : d'une part le sémito-berbère, d'autre part le couchitique. Le couchitique se subdivise en deux : les dialectes *agaw* s'opposant au reste, auquel s'associe le haoussa. Dans le sémito-berbère, on a d'une part les langues et dialectes proprement berbères ; et d'autre part, sous le nom de sémitique, outre les langues antiques ou modernes du Proche-Orient asiatique (akkadien, arabe, araméen, hébreu...) les langues sémitiques de l'Ethiopie (guèze, anharique...) et l'égyptien ancien (que la géographie rapprocherait plutôt du domaine berbéro-lybique...). Bref, la conclusion tient en un mot : "couchito-sémitique", à substituer à un autre : "chamito-semitique".



L'essentiel de mes conclusions (que le graphique plan issu de l'analyse factorielle permet de nuancer opportunément : cf. PRAT 4 p. 39) tient dans les petits tableaux suivants (*ibid* p. 40) :

1 \ 2	G	D	L	S	1 \ 2	G	D	L	S
G	66	190	182	191	G	0,42	1,24	1,09	1,30
D	164	48	194	133	D	1,24	0,37	1,32	1,06
L	70	87	33	68	L	1,09	1,38	0,47	1,13
S	148	114	85	30	S	1,58	1,25	0,82	0,34

on a noté :

Dentales = D = {d, z, t, s, s', t} ;

Gutturales = G = {>, g, h, h, k, c, q} ;

Labiales = L = {b, u, m, p} ;

Sonnantes = S = {i; l, n, r} ;

et 1 est mis pour 1-ère radicale ; 2 pour 2-ème radicale ; avec dans le tableau de gauche, e.g. à l'intersector de la ligne G et de la colonne D le nombre (190) des racines dont la 1-ère radicale est une gutturale et la 2-ème une dentale. Tandis que le tableau de droite, par les quotients :

$$k(C1,C2) \cdot k / (k(C1) k(C2))$$

(où C1 désigne une ligne, C2 une colonne du tableau de gauche $k(C1,C2)$ le nombre inscrit à la croisée de C1 et de C2, $k(C1)$ le total de la ligne C1, $k(C2)$ le total de la colonne C2, k le total du tableau) familiers à l'analyse des correspondances, exprime les affinités (quotient > 1) ou répulsions (quotient < 1) entre classes de consonnes. La fréquence des couples DL (puis LD) sera expliquée par une articulation antérieure ; pour les couples GS (puis SG), on a une articulation postérieure. Il s'agit donc dans les deux cas d'une harmonie entre consonne et sonnante.

13 Du texte au livre

Pour conclure, peut-être sera-t-il permis à l'invité que je suis parmi les bibliistes de s'interroger sur le thème du présent Colloque : "Bible et Informatique : le texte".

La plupart des participants sont convaincus que la Bible est, selon l'étymologie, le Livre par excellence ; au message du Livre beaucoup ont répondu en lui consacrant leur vie. Pourtant, c'est à la dure épreuve de l'automate qu'ils ont résolu d'exposer le Livre ; ou du moins, le texte. Une curiosité que rien sur terre ne peut satisfaire, les porte à rechercher dans l'ordre matériel du texte quelque chose de la genèse du Livre.

Certes, ainsi que l'a écrit G.E. Weil, "il semble... qu'aucun outil statistique ne pourra mettre en évidence les raisons des aramaismes qui apparaissent dans tel ou tel livre de la Bible, ni la raison de la présence du nom de Cyrus (-VI-ème siècle) dans le livre d'Isaïe, ouvrage attribué à un auteur ayant vécu au VIII-ème siècle avant notre ère".

Mais toute réflexion critique sur un texte requiert une conception globale de sa structure ; et une structure se révèle à nous par des rapports entre parties. Ne pouvant renfermer en moins qu'une leçon la substance d'un livre, j'ai voulu vous convaincre ; vous montrer, sur une suite d'exemples, que l'analyse multidimensionnelle est, par excellence, un outil efficace pour acquérir une vue d'ensemble de rapports et redistribuer des parties.

Nous pouvons non seulement réaliser des typologies (ce qui a toujours été fait) ; mais les réaliser de manière inductive, sans introduire aucune hypothèse, en centrant simplement notre attention sur des informations d'une certaine classe. Nous pouvons vérifier nos résultats en analysant des informations d'une autre classe, et nous savons au niveau interne de quoi sont faites ces typologies. (Par exemple, dans le cas des textes chinois, nous savions finalement, après élimination des pronoms personnels, que l'opposition entre textes de théâtre et textes de roman reposait sur deux groupes de caractères, liés surtout à l'expression orale rendue différemment au théâtre et dans le roman).

Oui, nous savons, au niveau interne, rendre compte d'une typologie. Mais le problème des problèmes est de donner en terme d'interprétation externe la signification des oppositions, partitions, pentes et gradients que l'on découvre. Un même auteur parle-t-il de deux choses différentes ? Deux parties d'un même texte ont-elles été diversement altérées ? Ou s'agit-il de l'oeuvre de deux auteurs différents qui peut-être n'ont pas vécu dans le même siècle ? A de telles questions, nous ne pourrions répondre valablement que d'après l'expérience acquise en dehors du domaine biblique.

Car il n'y a de science que du général. La science du texte de la Bible ne sera fondée que sur des lois d'interprétation de la structure des textes découverts d'après l'analyse de textes très divers dont la genèse et l'histoire sont parfaitement connues ; ce qui n'est pas le cas pour la Bible.

A la différence du chercheur intéressé à ce qu'il poursuit, l'ordinateur n'a pas d'inclination propre. Les structures taxinomiques qu'il produit résultent de l'application d'une méthode universelle à des données qui lui ont été expressément soumises.

Il est vrai que le choix du tableau à analyser ne s'impose jamais absolument ; qu'une partialité inconsciente peut biaiser ce choix. Mais en multipliant les analyses ; en attendant pour laisser courir sa pensée, que le calcul ait fait son oeuvre ; en acceptant les surprises qu'offre le traitement d'un tableau qu'on avait empli case après case, sans en voir la totalité... on pourra si l'on s'y applique échapper à soi-même.