

CAD

Classification des états d'un texte, mathématiques et informatique

Les cahiers de l'analyse des données, tome 3, n° 2 (1978), p. 243-249

http://www.numdam.org/item?id=CAD_1978__3_2_243_0

© Les cahiers de l'analyse des données, Dunod, 1978, 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/>

MÉMOIRE REÇU

Classification des états d'un texte,
mathématiques et informatique

[TEXTES ANCIENS]

d'après J. Duplacy

Analyse par J.-P. Benzécri (1)

Jean Duplacy : Classification des états d'un texte, Mathématiques et Informatique : repères historiques et recherches méthodologiques ; in *Revue d'histoire des textes* Tome 5 (1975) ; pp 249-309 ; CNRS Paris-1977.

On ne prétend pas faire ici une analyse exhaustive du travail de Jean Duplacy : car d'une part le texte concis et ordonné ne pourrait être résumé fidèlement en moins de mots que n'en a écrit l'auteur ; d'autre part ces mots sont assez clairs pour être compris de tout statisticien qui s'appliquera à les lire. Mais un aperçu même rapide sur l'histoire des textes intéressera tout homme cultivé ; et le spécialiste d'analyse multidimensionnelle sera heureux de voir qu'après d'autres savants, les philologues du XX^e siècle finissant apportent à nos algorithmes avides leur tribut de données ! A la vérité nombreuses sont les disciplines où le travail s'organise en trois étapes (dont le cycle doit encore être parcouru plusieurs fois) : connaissance du terrain, choix et collecte des données, élaboration de celles-ci et recherches des structures. Le statisticien qui ne touche que des chiffres, et n'a aux mains qu'une ombre du réel, doit vénérer l'homme de terrain, qui décrypte les parchemins, panse les plaies, ou cueille et reconnaît les plantes. Mais celui-ci doit savoir que l'analyse, étape ultime et indispensable de son étude, n'appartient pas en propre à la discipline dont il est spécialiste : elle doit se régler suivant des principes communs à tous. Or cette communauté de principes est encore si peu reconnue que J. Duplacy, exposant les travaux du philologue d'Oxford J.J. Griffith (1968, 1969) inspirés par un article de Sokal (1966), s'exclame : "Ce fut une assez belle performance que de percevoir une analogie entre le classement des [nématodes] et celui des manuscrits de Juvénal et, plus généralement, entre la taxinomie des être vivants et ce que nous appelons, inspirés par Griffith, la taxinomie textuelle. Dans les lignes qui suivent, comme partout dans ces cahiers, notre propos est d'apprendre à percevoir ces fécondes analogies, non par une performance, mais sans effort.

Le profane qui sur la reproduction d'un parchemin enluminé tente de déchiffrer un verset de psaume s'arrête bientôt déconcerté par la calligraphie dont les u ressemblent aux n voire les c aux i ; *a fortiori* abandonne-t-il au spécialiste la lecture des traits à l'encre qui se mêlent aux fibres d'un papyrus déchiqueté ; pour ne rien dire des *palimpsestes* (peaux grattées pour être réécrites) qui portent sous une copie de tel chapitre de la vulgate le vestige unique d'un texte classique qu'on croyait perdu mais dont le laboratoire de physique restituera au philologue une photographie utilisable. Ce n'est toutefois pas des sciences de la lecture que parle J. Duplacy, mais, comme son titre le précise, de la confrontation et de l'élaboration des *états d'un texte* ; ou, si l'on préfère, de manuscrits déjà lus et retranscrits en clair ; et s'il ne considère que les "textes qui furent publiés pour la première fois au temps où les livres n'étaient pas imprimés, mais écrits et reproduits à la main" c'est sans doute principalement parce que ces textes anciens sont susceptibles d'avoir reçu des mains des copistes un très grand nombre d'*états* divers, d'ailleurs généralement fautifs : , tout copiste est un gâcheur, dit l'adage arabe.

En effet, quiconque a copié ou relu une copie, sait que *destruction* peut se changer en *distraktion* ; que le texte compris entre deux mentions d'une expression X répétée à quelques lignes d'intervalle (suivant le schéma X t X), disparaît si la copie interrompue au premier X ne reprend qu'après le second ; cette faute est si fréquente qu'on a créé pour elle le
(1) Professeur de statistique. Université Pierre et Marie Curie. Paris.

terme d'*homototeleute* : qui a le même tout); etc. Aussi la philologie est-elle dominée par la considération des fautes, avec pour objectif pratique premier la restauration du texte même en son état original correct. Cependant on a une vue trop simple de la prolifération des états d'un texte, si on se borne à en schématiser l'histoire suivant une arborescence orientée ; chaque état engendrant par copie quelques descendants ne différant de leur unique parent que par l'introduction de fautes ou leur correction plus ou moins heureuse. D'une part, dans un louable souci de correction un copiste peut avoir collationné plusieurs sources : d'où un hybride issu de plusieurs parents (les états collationnés). D'autre part *erreurs* et *corrections* ne sont pas les causes uniques de variations entre états : *résumer* ou *expurger*, et au contraire *développer* ou *compléter* ont des effets, beaucoup plus amples que ceux des innocentes erreurs et corrections : on le peut voir de nos jours sur les éditions imprimées. L'auteur lui-même a pu reprendre son oeuvre. On lit par exemple dans l'*Avertissement* à l'édition française (Plon Paris 1930) des mémoires du Chancelier Prince de Bulow que, sur le manuscrit qu'il avait dicté, "l'auteur procéda à des suppressions et à des additions, toutes certifiées exactes et signées de sa main" : or tout écrivain n'a pas la conscience d'un Prince, ni l'exactitude d'un chancelier ! Et que dire de l'Iliade et de l'Odyssee, épopées qui courent depuis plus de 2000 ans sous le nom d'Homère, mais où l'on a pu voir des anthologies plutôt qu'une oeuvre fondue d'une seule coulée.

Il est assez facile d'imaginer une grande variété d'épisodes susceptibles d'engendrer de multiples états d'un texte attestés dans les manuscrits. Certains de ces épisodes peuvent être bien connus dans leurs conditions historiques : travail d'une école de copistes, mobiles idéologiques d'une falsification... ; beaucoup ne sont que soupçonnés et ne seront jamais prouvés, encore moins décrits avec précision. Fait plus frappant pour le statisticien, le système des relations - similitude, contiguïté ou succession - entre les états connus du texte, semble ne pouvoir être décrit adéquatement par aucune structure mathématique simple. L'idéal de l'analyse des données sera donc seulement de figurer l'ensemble des états et de leurs caractéristiques comme des points sur des cartes, tout en suggérant des classifications et des lignes de parcours. A cet idéal, au terme d'une tradition qui remonte aux éditeurs alexandrins, se rallient des philologues contemporains armés de l'outil informatique : grâce à Jean Duplacy, nous marquerons quelques jalons sur leur route, qui converge avec la nôtre.

En 1734, J.A. Bengel (philologue spécialiste du nouveau testament), suggéra de grouper les manuscrits d'après leur texte "par paires... par petits ou grands groupes, par familles, tribus et nations..." Ainsi vingt ans avant le naturaliste Adanson, Bengel formulait le principe de la classification ascendante hiérarchique. Mais de ses *schematismi* représentant "proximités et distances", J.A.B., espérait édifier "une sorte de tableau quasi-généalogique". Le vers était dans le fruit ! tout au long de cette note nous répétons notre défiance vis-à-vis des arbres généalogiques : non que la structure d'arbre soit à proscrire : elle sert grandement l'analyse des données ; ni les rapports généalogiques hors de propos : il est clair que les états successifs s'engendrent. Mais, dans un domaine où existent des rapports très divers et mal connus, il est suprêmement imprudent d'identifier un arbre qui est au mieux le résultat impartial de l'analyse de certaines données, avec une généalogie qui d'une part n'explique que partiellement les états et d'autre part, du fait des hybrides (cf *supra*) n'a sûrement pas une structure arborescente.

C'est pourtant sous forme d'arbres généalogiques de manuscrits ou *stemmata codicum* que le rêve de Bengel vit le jour en Allemagne après quelque cent ans de gestation : à partir de 1830 les *stemmata* se répandirent jusqu'à devenir, dit Duplacy, "une habitude sinon une obligation". Sur quelles données fondait-on ces classifications certes non automatiques, ni même algorithmiques ? *A priori* toutes les informations : historiques (date...), géographiques (provenance...), graphiques (écriture...) peuvent servir à ordonner les manuscrits. Mais "on se basa de plus en plus sur des critères [purements] textuels" ; passant ainsi selon le titre adopté

par J.D., de la classification des manuscrits à celle des *états d'un texte*. Plus précisément, l'attention se porte sur les *lieux variants*, passages du texte où les différents états ne s'accordent pas et dont on recense les variantes. Après avoir accepté comme critère de proximité l'accord des manuscrits sur toute variante quelle que fût sa qualité critique... "on s'avisa vers la fin du [XIX^e] siècle que ce type d'accord s'il pouvait servir à répartir les manuscrits en groupes, ne pouvait permettre d'établir leur généalogie. Pour ce faire, seuls les accords sur des variantes fautes, ..., pouvaient entrer en ligne de compte". A cette méthode des *fautes* communes fut à tort selon J.D., attaché le nom de Lachmann.

On s'avisa... Etait-ce bien avisé? Avant les critiques qu'on voit sourdre sous l'encre de J. Duplacy, la justification : seules les fautes d'inadvertance, recopiées telles quelles avec le respect dû au texte même, se propagent par contiguïté aussi longtemps qu'on ne les a corrigées ; les variantes idéologiques, au contraire, ont comme les doctrines dont elles procèdent, un cours tantôt visible, tantôt souterrain. Mais qu'est-ce qu'une faute d'inadvertance? Et comment même parler de faute sinon par référence à une norme - le texte correct - qui est justement l'objet pratique de la recherche (ne l'oublions pas, même si la taxinomie des états d'un texte est en elle-même une passionnante investigation historique). Les fautes vulgaires, les simples variantes orthographiques apprennent peu : les autres variantes doivent être recensées sans jugement de valeur.

Cependant que des philologues anglo-saxons signalaient les "*completions of genealogies by mixtures*" (hybrides), tandis que d'autres s'engageaient dans la voie statistique en comptant les accords textuels (ou nombre des lieux variants où s'accordent deux états) "on mettait [parfois] en question l'intérêt même de la classification des manuscrits" "Remarquant que les arbres généalogiques... proposés par les éditeurs de textes médiévaux étaient presque toujours bifides [nous disons binaires], chaque manuscrit n'en engendrant que deux autres, J. Bédier, à tort... ou à raison, attribua ce phénomène curieux à la classification par les fautes communes". Sans nous prononcer plus que J.D. sur la culpabilité des fautes communes, nous rattacherons le paradoxe décelé par J. Bédier à la marche naturelle de tout algorithme de classification ascendante hiérarchique qui procédant par comparaison entre paires engendre un arbre binaire. Assurément l'illustre éditeur de la Chanson de Roland avait raison de souligner qu'une véritable généalogie de manuscrits (plus exactement des manuscrits conservés) n'avait que peu de chance d'être binaire ; la postérité immédiate de chaque texte (plus exactement l'ensemble des descendants connus reliés à celui-ci soit directement, soit par une chaîne d'intermédiaires disparus) pouvant aussi bien comprendre un seul individu que plusieurs. Nous l'avons déjà dit : le résultat d'une procédure (fût-elle conduite avec rigueur et prudence) ne saurait être interprété dans le cadre d'un modèle (ici la généalogie à un seul parent) suggestif, mais certainement inadéquat.

De par la multiplicité des états du texte et de ses traductions anciennes la Bible pose aux philologues de formidables problèmes dont certains intéressent la doctrine et la foi. C'est chez un savant bénédictin, Dom Henri Quentin, responsable de la préparation d'une grande édition du texte latin de la Bible (Vulgate) que nous trouvons pour la première fois une présentation complète des données exactement adaptées à nos méthodes d'analyses. Aussi préparant des généralisations ultérieures exposerons-nous les principes de Dom Quentin avec les notations mêmes familières à l'analyse des correspondances.

Chaque lieu variant est comme une question q qui se pose à propos de chaque état i du texte : l'ensemble des réponses possibles J_q n'est autre que l'ensemble des variantes qu'offre le lieu q . D'où en notant I l'ensemble des états du texte et Q l'ensemble des lieux variants un tableau v_{IQ} donnant à l'intersection de la ligne i et de la colonne q l'indication (par un numéro ou tout autre sigle), de la variante $v(i,q)$ que le texte i présente dans le lieu q . Dom Quentin construit ce tableau (exactement : il met les textes en colonnes et les lieux en lignes...) et le

découpe en lanières pour effectuer commodément les comparaisons entre états du texte. D'abord il construit le tableau carré symétrique a_{II} , donnant à l'intersection de la ligne i et de la colonne i' , le nombre $a(i, i')$ des accords entre les états i et i' (i.e. le nombre des lieux q où i et i' présentent la même variante : $a(i, i') = \text{Card}\{q | q \in Q; v(i, q) = v(i', q)\}$.) Puis afin de faciliter l'exploitation de ce tableau a_{II} (que nous appellerions un tableau de similitude), il construit pour chaque état i une fiche colonne donnant de haut en bas la suite des états i' dans l'ordre de leur proximité décroissante avec i ; avec dans chaque case simultanément le numéro (ou sigle) i' et le nombre $a(i, i')$ de ses accords avec i . Enfin il entreprend une comparaison des manuscrits trois à trois: notons $z(i, i', -i'')$ le nombre des lieux q où i et i' concordent entre eux mais diffèrent de i'' :

$$z(i, i', -i'') = \text{Card}\{q | q \in Q; v(i, q) = v(i', q) \neq v(i'', q)\} ;$$

Dom Quentin s'intéresse particulièrement aux triplets (i, i', i'') pour lesquels : $z(i, i', -i'') = 0$; autrement dit i et i' ne s'accordent jamais contre i'' : dans ce cas le savant bénédictin croit pouvoir affirmer que i'' est un intermédiaire entre i et i' (e.g. i' copie i qui copie i ; ou l'inverse ; en tout cas, toutes les variantes que i et i' ont en commun ont été transmises de l'un à l'autre par i'').

De cette méthode taxinomique (clairement exposée, 1922, 1926, par son auteur) on soulignera trois principes :

- 1) D.Q. aborde l'étude des manuscrits sans référence aux informations de temps, de lieux, d'écriture, etc... : il ne considère que des états du texte.
- 2) Au niveau du classement réciproque de ces états il ne connaît "ni erreurs ni fautes communes... mais seulement des formes diverses du texte".
- 3) Il distingue nettement la découverte de l'enchaînement formel des manuscrits (proximités, chiffrées par $a(i, i')$) ; et aussi alignements recherchés par le calcul de $z(i, i' - i'')$ de l'orientation chronologique ou généalogique de cet enchaînement.

C'était être bien proche de l'a. des données, telle que nous la pratiquons. Toutefois, outre qu'il manque à Dom Quentin un ordinateur pour passer aux actes..., sa méthode même laisse à désirer divers éclaircissements. En réalité, singulièrement dans le cas de bible vulgate latine, l'ensemble I des états connus du texte, comme l'ensemble Q des lieux variants est formidable. Dans son exposé de 1926, D.Q. se borne à un exemple : au lieu du texte entier de la bible, 8 chapitres choisis parmi les 235 que comporte les huit premiers livres de l'Ancien Testament (Octateuque) ; et pour ces chapitres, seulement les variantes attestées dans les textes de 70 manuscrits (choisis parmi des centaines de manuscrits anciens). C'est si l'on veut un échantillonnage : mais fait dit Jean Duplacy, comme à *la sauvette*. Or, pour faire mieux, nous ne voyons d'autre méthode qu'une suite hiérarchisée d'analyse de données : e.g. travailler d'abord sur un chapitre avec plusieurs centaines de manuscrits, afin d'obtenir une typologie de l'ensemble Q des lieux variants (plus exactement de l'ensemble J de toutes les variantes : $J = \cup \{J_q | q \in Q\}$) ; ce qui permettra ultérieurement de résumer par quelques caractères la description de l'état du texte de ce chapitre (e.g., si les variantes sont partagées en 10 classes d'après leurs tendances à se retrouver dans les mêmes états du texte : $J = J_1 \cup J_2 \dots \cup J_{10}$; on comptera pour chaque état du texte le nombre des variantes de chaque classe qu'il possède : soit 10 nombres seulement ; alors qu'un chapitre peut offrir 36 lieux variants totalisant 600 variantes) ; et de même pour tous les chapitres d'un livre ; puis livre par livre etc. Mais il y a plus : la notion de lieux variants et de variante n'est elle-même pas absolument claire. Quelques heures de travail en compagnie de - Amphoux, docte élève de J. Duplacy, nous ont montré que si l'on se tient à la lettre, les variantes se multiplient à l'infini ; il faut donc assimiler des formes voisines ; considérer dans tel

verset, l'ordre des mots plutôt que la variante de chaque mot ; dissocier le temps du verbe lui-même, etc. Ainsi bien plus que le schéma classique des lieux variants c'est la formulation en questionnaire qui rend véritablement compte du travail du philologue : devant un grand nombre de variantes d'un même texte (épisode, paragraphe, ou seulement verset), celui-ci doit d'abord établir le questionnaire clos qui rendra compte adéquatement de la diversité des variantes. Problème difficile, et dont la solution semble largement en butte à l'arbitraire : mais ici comme ailleurs l'analyse des données peut, dans le cadre des principes du domaine (ici la philologie), aboutir à une solution stable, à une représentation spatiale de l'ensemble des données insensible aux choix faits dans le détail. Finalement l'analyse matérielle stricte des états du texte (en lieux variants et variantes) ; la critique interne (correction des formes et du sens : celle-ci aboutissant à l'analyse du contenu) ; l'épigraphie *lato sensu* (date et origine du manuscrit ; caractère du support et de l'écriture), aboutissent toutes à des informations d'un même format : le questionnaire. Format universel, aujourd'hui très familier au statisticien : qui construit le tableau de Burt (ou tableau de cooccurrence des modalités de réponse aux diverses questions de toute nature) et en analyse divers sous-tableaux ; non sans mettre en éléments supplémentaires les informations extérieures à ce sous-tableau. Par exemple sur une analyse du tableau de cooccurrence des modalités de réponses aux questions textuelles (variantes), figureront en éléments supplémentaires les modalités de nature épigraphique ; ou encore on analysera le tableau rectangulaire croisant modalités textuelles et modalités épigraphiques (ce qui révélera la distribution spatio-temporelle des variantes ; et en retour donnera des présomptions sur les coordonnées d'un état du texte attesté dans un manuscrit d'origine problématique ; etc.) ; tandis que les modalités (variantes) du contenu seront portées par celles du texte, comme l'esprit par la lettre.

On ne s'étonnera pas non plus, que nous formulions des réserves quant à l'ingénieux principe de l'alignement des états du texte suivant le critère du zéro quentinien ($z(i, i', -i)$) ; *a priori* on était assuré que les textes hybrides en rendraient l'application difficile. On peut parler de "quasizéros" si $z(i, i', -i)$ est petit vis-à-vis de $z(i, i', -i')$ et de $z(i', i', -i)$... : au fond, selon-nous c'est l'objectif même assigné à la recherche qui ne convient pas. Répétons-le encore : il faut chercher non un *stemma* (arbre) généalogique ou non, mais une vue d'ensemble, un déploiement spatial des informations retenues.

Tel quel le travail de Dom Quentin nous est grandement sympathique. Il est d'un demi-siècle en avance sinon sur les idées (n'est-il pas possible de trouver à toute idée humaine quelque témoin en tout temps ?) au moins sur les techniques de sa génération. Un savant médiéviste d'Oxford, Walter Greg publie en 1927 "*The Calculus of Variants: an Essay on Textuel Criticism*" ouvrage théorique que J.D. juge "encore plus rébarbatif que les pages correspondantes de Dom Quentin",. Divers autres philologues proposent de rechercher les structures par des procédés confinant aux mathématiques, et certains vont jusqu'à constituer d'assez grands tableaux de nombres ! A ces pionniers dont les travaux ne peuvent être analysés dans cette brève note, J. Duplacy rend un hommage mérité. Avec l'avènement de l'ordinateur une aube nouvelle se lève : les premiers rayons (Parvis 1952, Ellison 1954) touchent l'Amérique : mais l'hymne au soleil éclate en France : Dom Froger émule de Dom Quentin, publie en 1968 : La critique des textes et son automatisation (Paris ; Dunod).

Dès 1961 Dom Froger présente une méthode quentinienne de taxinomie textuelle qu'il a mise au point avec des non-philologues, Mesdames Poyen et Renaud et M. Poré. S'en inspirent un philologue P. Carivet et un mathématicien P. Malvaux, qui sans le secours de l'ordinateur s'appliquent à "classer quinze états d'un texte de Théodoret de Cyr sur la base de 585 lieux variants". Au spécialiste de l'analyse des données, nous signalerons que "Malvaux dispose, en une seule colonne (et en valeurs croissantes), la série de chiffres exprimant les distances avec l'indication, en face de chacun d'eux de la paire ou des paires de manuscrits que sépare la distance en question" : ainsi tandis que le psychologue R.N. Shepard publiait

ses remarquables recherches sur l'analyse des proximités (1962), Malvaux (1964) considérerait l'ordonnance (ou système d'inégalités entre distances) de ses quinze manuscrits.

Du mémoire de 1968, Jean Duplacy croit pouvoir affirmer que par lui "les philologues sont pour la première fois avertis que, même dans ses aspects non numériques, la taxinomie textuelle qu'ils pratiquent depuis longtemps recourt bon gré mal gré à des aspects très élaborés de la mathématique, très proches d'ailleurs, au moins initialement, de la logique formelle classique. Toutes les dimensions mathématiques, du travail philologique sont désormais explicites. La voie est ainsi ouverte à l'*automatisation*..." Mais tout en appréciant hautement la clarté de vue de Dom Froger, J.D. reste sur sa faim quant aux problèmes taxinomiques concrets qui alimentent ses propres recherches : des centaines voire des milliers d'états du texte du nouveau testament à ordonner ; l'obstacle majeur étant constitué par les cas de tradition contaminée (états hybrides). L'océan des variantes - même si la plupart n'importent pas au sens - doit être affronté pour dominer l'histoire du texte.

De ce point de vue les travaux de Griffith (1968, 1969 ; cités dès le début de cette note) ouvrirent à J.D. des perspectives nouvelles. Classification hiérarchique ; mais aussi matrice de similitude avec lignes et colonnes réordonnées pour faire apparaître des blocs (cf Czekanowski) ; et la rencontre avec des disciplines diverses classant toutes sortes de réalités. De "mystérieux calculs statistiques" et des problèmes philosophiques chers au philologue : "la tension, par exemple, [en biologie], entre une classification "phénétique", purement [fondée sur les formes], et une classification "phylétique", c'est-à-dire généalogique ; ..." les hybrides défilant le schéma arborescent. "Décidément, conclut J.D., les philologues ne pourraient certainement pas s'occuper seuls de la taxinomie des traditions contaminées".

Et après s'être appliqué, comme il le déclare avec une généreuse modestie, "à réduire un peu [son] ignorance statistique" Jean Duplacy rencontre un tableau de données (Fee, 1971) apte à l'analyse : 27 états du texte du ch IV de l'Evangile selon Saint Jean, décrits d'après 97 lieux variants totalisant 231 variantes. Sans relater tous les calculs plus ou moins classiques (distances, pourcentages...) auxquels après Fee s'est appliqué J.D., disons que celui-ci aboutit, grâce à la collaboration du statisticien E. Huret, à deux essais de taxinomie textuelle automatique (1972-1973) : application de l'analyse des correspondances et de l'agrégation autour des centres variables (*nuées dynamiques* de E. Diday) au tableau logique 27×231 : 27 états \times 231 variantes ; $k(i, j) = 1$ si l'état i comporte la variante j , et zéro sinon. De cet essai J.D. conclut que "l'analyse topologique [E. Diday] serait capable de dégrossir au moins le classement d'une tradition manuscrite très abondante et de mettre en tous cas en lumière les pôles taxinomiques... L'analyse factorielle, par contre conserve un double avantage : mieux traduire l'équilibre dynamique interne des classements qu'elle fournit ; pouvoir mettre en évidence les variantes (et par là les groupes de textes covariants) dont dépend ce classement" [représentation simultanée des états et variantes]. Quant à la classification ascendante hiérarchique, J.D. estimait, à tort, que son tableau dépassait les capacités de l'algorithme.

Au long de cette note, nous avons mêlé à l'exposé des travaux que signale J.D., bien des critiques et des suggestions : notre conclusion n'aura donc rien d'innatendu pour le lecteur.

Répétons que les données de philologie doivent être conçues comme un questionnaire appliqué aux états du texte : ce format universel épouse bien les préoccupations du chercheur ; il se prête à des traitements statistiques maintenant classiques (pour des références, cf indice systématique du Vol II des Cahiers ; in Vol III n° 1) ; et permet de confronter aux données strictement textuelles toute autre information supplémentaire, répondant ainsi aux exigences de "certains philologues spontanément réfractaires à [une] formalisation" des "états d'un texte que l'on considère *en soi* hors de l'histoire, de la géographie et de leurs conditions de production".

Comme tout modèle expression mathématique d'une schématisation excessive du réel, le schéma généalogique arborescent, en butte aux états hybrides, doit être regardé avec réserve. Le philologue ne peut demander au statisticien une solution achevée ; ni même espérer donner lui-même de ses propres recherches une synthèse totalement algébrique. Au mieux les mathématiques fournissent un support géométrique, où s'inscrira dans son ordre naturel, tout ce qu'on sait des textes : c'est l'objet de l'analyse factorielle.

Avec l'analyse factorielle qui représente dans un même espace les états de texte et les modalités servant à décrire ceux-ci on conjuguera la classification automatique, notamment la classification ascendante hiérarchique. La méthode de E. Diday a l'inconvénient d'être sensible au choix initial des centres d'agrégation, mais d'une part elle réalise efficacement la réduction de grands ensembles de données ; d'autre part elle fournit des partitions intéressantes quand les centres initiaux autour desquels s'agrègent les individus i sont choisis. (dans l'espace des facteurs) d'après une classification hiérarchique effectuée sur les modalités j , ou *vice versa* (ainsi procède J. Bordet). L'analogie avec les grands tableaux de données écologiques analysés par Briane et coll (Cahiers Vol III n° 2) nous paraît nette : aux relevés correspondent les états du texte ; les variants (ou autres modalités descriptives) prenant le rôle des espèces végétales. Ainsi (en se fondant sur un premier groupement des états du texte en quelques dizaines de classes ; cf *op. laud.*) l'analyse d'un tableau 1000 x 1500 cesse d'être utopique. Dès lors le choix des données (états de référence du texte ; variantes conservées) sur lequel J.D. revient plusieurs fois avec inquiétude, n'est plus une contrainte. D'ailleurs, quelle que soit la capacité des algorithmes, le philologue aura d'abord besoin d'analyses partielles pour assurer ses pas : entre les premières analyses de G. et M. Roux portant sur quelques dizaines de relevés et la compilation de très grands tableaux traités par Briane, Lazare et Salanon, dix ans ont passé, marqués par la publication de *La Phytosociologie* (Masson ; Paris 1973) où M. Guinochet communique magistralement son expérience d'écologiste dans le langage à la fois rigoureux et simple de l'analyse des données : Souhaitons à Jean Duplacy et à ses disciples une aussi féconde collaboration avec la statistique!