LES CAHIERS DE L'ANALYSE DES DONNÉES

J. P. BENZÉCRI

Description des textes et analyse documentaire

Les cahiers de l'analyse des données, tome 9, n° 2 (1984), p. 205-211

<http://www.numdam.org/item?id=CAD_1984__9_2_205_0>

Les Cahiers d'Analyse des Données
Vol. IX - 1984 - n° 2 - 205-211

DESCRIPTION DES TEXTES ET ANALYSE DOCUMENTAIRE

[TEXT. DOC.]

par J. P. Benzécri (1)

1 *Etat de l'analyse multidimensionnelle des données linguistiques*

C'est d'après des considérations de linguistique que, sous le nom de distance distributionnelle, la distance du chi 2 a été introduite en analyse factorielle puis en classification automatique (cf. Histoire et Préhistoire de l'Analyse des Données V ; Dunod éd. 1981). On trouve dans les tomes I et II du Traité sur l'A. des D. deux exemples d'analyse de tableaux de correspondance croisant un ensemble I de textes et un ensemble J de vocables. (Les rôles des personnages de la Phèdre de Racine : TIA n° 2 § 3 ; et les Professions de foi des députés élus en 1881, TIIC n° 2). Depuis est paru un volume de la collection de "Pratique de l'Analyse des données (le Vol PRAT 3) entièrement consacré à des études de linguistique et de lexicologie ; et chaque année l'indice systématique des Cahiers de l'A. des D., donne sous le mot "linguistique", quelques références nouvelles, dont certaines à des compte-rendus de publications étrangères.

On peut conclure de l'ensemble de ces travaux que les tableaux de correspondance constitués de données verbales relevées dans des textes de toute nature, se prêtent à l'analyse multidimensionnelle autant que ceux issus d'autres compilations. Nous estimons cependant que le traitement des données linguistiques est à peine entamé, dans la mesure où des problèmes fondamentaux, qui ont été vus d'emblée, attendent depuis 20 ans d'être étudiés sur une base expérimentale assez large. Une langue en effet forme un tout organique ; on peut même dire que l'ensemble de toutes les langues est un domaine unique, où les modèles de pensée et d'expression se propagent de telle sorte que leur étude ne peut s'arrêter à aucune frontière.

Certains concluront de l'immensité même de l'objet, à l'impossibilité de toute étude adéquate. Nous concèderons sans réticence, qu'il y a de la témérité à tracer ne fût-ce que le programme d'une analyse d'ensemble de la prose française, ou de la prose scientifique internationale... Mais nous savons que des données - textes saisis et parfois élaborés - existent, qu'il est dès maintenant possible de confronter. Pourquoi à un tableau I × J croisant un ensemble I de textes de presse de la fin du XVIII-ième siècle et un ensemble J de vocables, ne pas adjoindre un ensemble J1 de lignes croisant avec le même vocabulaire des textes d'une autre classe ou d'une autre époque ? C'est nous dira-t-on parce que le vocabulaire est par trop différent de I à I1. Mais comment sinon par l'analyse apprécier cette différence ? On peut prévoir qu'outre des mots propres à I à l'exclusion de I1 (ou à I1 à l'exclusion de I) on trouvera un vocabulaire commun, avec des zones intermédiaires. L'illustre Nobel a bâti sa fabuleuse fortune en une suite d'expériences industrielles sans cesse plus audacieuses et plus fructueuses, mais toutes marquées par une réplique sonore et meurtrière de la dynamite. Nous pourrions

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

0339-3097/84/02 205 7/$ 2.70/

à un moindre risque tenter la gloire : jamais un ordinateur n'explosera si, par exemple, on lui soumet le tableau I×(J1 ∪ J2 ∪ J3) : où I est un ensemble de textes scientifiques ; et J1, J2, J3 le vocabulaire respectif des versions anglaises, arabes et françaises de ces textes...

Tant de projets de toute taille doivent nous encourager sans nous étourdir : de hautes murailles se sont élevées brique à brique, suivant l'ordre d'un plan. C'est pourquoi nous croyons utile de placer la thèse de Mademoiselle Akila Aït Hamlat (résumée dans l'article [*IND. DOC.*]), par rapport à ce qui est fait et ce qui doit immédiatement se faire en analyse documentaire.

2 *L'automatisme dans le dénombrement des faits linguistiques*

Nous avons admis une fois pour toutes que la donnée de base universelle est le tableau de correspondance I × J ; et que nous savons comment traiter un tel tableau. Reste un problème unique, mais multiforme : quel tableau analyser. Mettons que l'ensemble I soit un ensemble de textes (dont on accepte pour l'instant le découpage), reste à choisir J et définir les k(i,j).

Il est commun de prendre pour J un ensemble de mots : mais même si cet ensemble est fixé, plusieurs questions se posent. Prendra-t-on ces mots tels qu'ils sont dans le dictionnaire : " aimer, beau"...; il faudra alors de quelque manière rattacher "aimons" à "aimer" et "belles" à "beau" et que fera-t-on d' "aimant" qui est participe, ou objet magnétique ? Beaucoup de ces menus détails sont mal perçus par un automate : aucun programme d'analyse morphologique et de levée des homonymies n'est encore parvenu à tout trancher sans erreur. Si l'on craint, à juste titre, le coût d'une élaboration des textes par un expert humain, on peut comme le fait L. Lebart (cf. PRAT3 ; LC3 n° 4 [*REP. LIB.*] et CAD Vol VI n° 2 pp. 229-243 [*ANA. LEX.*], 1981) dénombrer purement et simplement des formes (ou suite de lettres comprises entre deux séparateurs) en acceptant que soient comptés ensemble les deux "aimant", ou le "un" article avec le pronom et le chiffre. Entre les mains de L. Lebart et d'autres chercheurs cette méthode expéditive a fait ses preuves. D'ailleurs l'analyse des données est même capable de reconstruire des informations non codées voire mutilées, comme celles relatives aux locutions composées (cf. [*IND. DOC.*] § 2.1). Cependant, puisque tout texte est à quelque moment frappé sur un clavier, que les résultats de cette frappe sont aujourd'hui souvent enregistrés sur un support magnétique, que d'ailleurs c'est sur un tel enregistrement que repose toute élaboration statistique.... il est légitime de considérer que l'objet des études linguistiques et documentaires est un ensemble de textes déjà saisi. Sur ces textes on peut effectuer une analyse des formes imparfaite , laissant quelques ambiguïtés éventuellement résolues par interrogation sur écran d'un expert humain répondant au clavier. Dans le cas de la thèse de Mademoiselle A.H. A., les données ont été soumises à une élaboration avancée; justifiée par le fait qu'il s'agit de comptes rendus de voyages d'étude ; voyages coûteux dont on peut, sans regarder au surcroît de dépense, chercher à retrouver les fruits en interrogeant un fichier bien ordonné.

3 *Choix des mots pertinents*

Qu'il s'agisse de mots ou de formes, que les ambiguïtés soient ou non levées, la nature et l'étendue de l'ensemble J doivent être arrêtées. Une solution extrême est de prendre toutes les formes se rencontrant au moins dans deux textes du corpus étudié. Pour absurde qu'elle puisse paraître à qui désire rester maître des ingrédients de son mélange, cette méthode libérale peut aboutir à des résultats

excellents dont A. Salem donne un exemple (cf. PRAT3 ; LC1 n° 5 [*INV. LEX.* 1793]). D'ailleurs il est facile de nuancer, sans sortir de l'automatisme : on peut imposer aux formes retenues ; un seuil inférieur de fréquence ("se rencontrer plus de x fois, dans plus de y textes du corpus") ou de longueur : ainsi T. Behrakis écarte de son analyse des observations médicales des épidémies d'Hippocrate, les formes de moins de 3 lettres : ce qui revient à écarter une bonne part des mots outils (articles, prépositions) ; cf. [HIPPOCRATE] in CAD Vol VIII n° 4 ; 1983). Il reste possible de soumettre la liste automatiquement constituée, à un juge humain qui tente d'apporter diverses modifications dont l'issue des analyses permettra d'apprécier l'intérêt. Ici on doit souligner que les critères *a priori* les plus simples peuvent se trouver difficilement applicables : A. Salem (op. laud.) dit que les chercheurs avec lesquels il collabore ont finalement renoncé à tracer une frontière entre mots fonctionnels (ou outils : le, sur, avons...) et mots pleins : car dans la langue tout se fige et s'anime : un nom s'estompe dans une locution prépositionnelle à la mode ; et le verbe "avoir", dans le parler familier, dit la victoire du filou ou le dépit de la victime (je l'ai eu ; il m'a eu...).

Il est plus difficile encore de reconnaître, à première vue, au sein d'un ensemble de mots, ceux propres à caractériser l'intention ou le thème d'un texte. Aussi peut-on demander à la statistique de faire ce choix ou, au moins de le préparer. Dans l'étude déjà citée des professions de foi des députés en 1881, A. Prost, visant à préciser l'opposition entre "droite" et "gauche", dépouille d'abord exhaustivement un petit nombre de textes issus des deux extrémités de l'hémicycle et ne retient que 53 mots pour lesquels le contraste de fréquence est maximum. Cette méthode pourrait être appliquée en documentation automatique : un ensemble I de textes étant étiqueté par des experts suivant un ensemble C de classes (e.g. chimie ; mécanique ; économie ; biologie ...etc.), on construit un tableau $C \times J$, où $k(c,j)$ est le nombre de fois que le mot j a été trouvé dans le texte i ayant reçu l'étiquette c ; puis on ne conserve de J que le sous-ensemble de mots apportant les plus fortes contributions à l'inertie globale, ou encore aux premiers axes interprétables. Il est même apparu que la simple considération du tableau $I \times J$ (k_{ij} = nombre d'occurences du mot j dans le texte i) pouvait distinguer suivant ce critère de l'inertie un sous-ensemble de J pertinent pour le documentaliste (les calculs de INR se faisant, éventuellemnt, sans analyse factorielle).

4 *Modèle d'urne et répétitions des mots*

Cependant c'est dans une autre voie que Mademoiselle A.H. A. a obtenu des résultats à la fois prometteurs pour les praticiens de la classification des documents et suggestifs pour les théoriciens du style. On sait que selon un modèle simpliste, tout texte a pu être considéré comme un ensemble de mots issus d'une suite de tirages aléatoires indépendants à partir d'une urne dont le contenu dépendrait du sujet traité et des modes d'expression qu'affectionne l'auteur (cf. e.g. PRAT3 ; LA n° 1 § 3.2). Nul ne prétend que ce modèle soit adéquat : pourtant on l'a utilisé abusivement pour décider suivant des épreuves de statistique inférentielle, si tel texte peut être d'un auteur, dont l'*urne* est supposée connue d'après le vocabulaire de ses oeuvres dûment attribuées. Et plus raisonnablement, on se base sur le modèle pour cerner un sous-vocabulaire dans l'emploi duquel deux ou plusieurs auteurs diffèrent le plus nettement. Or il y a à la base du modèle, outre l'hypothèse tout à fait grossière qu'il suffit de tirer un tas de mots pour en faire un texte, l'idée plus générale que l'insertion de chaque mot fait l'objet d'un acte indépendant, et constitue à ce titre une unité statistique. Même s'il est difficile de préciser cette idée, et donc de la critiquer, on voit qu'elle n'est pas sans rapport avec un phénomène observable qui a depuis longtemps

attiré l'attention des pédagogues et des critiques du style : la répétition. Disons tout de suite que c'est en tenant compte des répétitions que A.H. A. a pu distinguer un sous-ensemble intéressant du vocabulaire de son corpus.

En français, la répétition est généralement regardée comme un vice de style : on doit s'en affranchir par le jeu des synonymes, des périphrases ou des pronoms. Pourtant Pascal a proclamé dans une de ses pensées qu'il ne convenait pas d'effacer les répétitions d'un mot qui marque le caractère du texte ; et, en 1923, un Traité de Psychologie (cf. A. Lalande ; in G. Dumas ; TI p. 51) avertissait les lecteurs de formation littéraire qu' "on ne doit pas, pour respecter la règle française qui interdit de répéter un même mot, remplacer le terme propre par un équivalent approximatif". D'ailleurs, certaines langues européennes semblent faire moindre cas de la répétition ; tandis que les langues sémitiques recourent avec prédilection à la répétition des racines comme à une sorte de superlatif emphatique. La répétition, non seulement des mots, mais des tours de phrases et particulièrement des mots usuels, tend à conduire la main de celui qui écrit sans y prendre garde. On peut encore prétendre, par fidélité au modèle des urnes que la répétition affecte nécessairement les mots dont la fréquence est très élevée dans celle des urnes qui convient au sujet traité ! Finalement, ici comme ailleurs, nous pensons qu'il s'agit d'un phénomène qu'il faut observer, en réservant les suggestions et hypothèses contradictoires pour en broder le commentaire de quelques graphiques issus d'une analyse factorielle !

5 *Choix des mots d'après leur taux de répétition*

Dans sa thèse, A.H. A., construit le tableau J × N croisant un ensemble J de mots avec l'ensemble N des entiers, avec pour k(j,n) le nombre des documents où le mot j figure exactement n fois : inutile de dire que N n'est pas pris jusqu'à l'infini : on a seulement N = {0, 1, 2, 3, 4, 5, 6, {7,8}, {9,10} {11....36}} ; i.e. les nombres (7,8), sont cumulés sur une seule colonne ainsi que (9,10) et les nombres supérieurs ou égaux à 11 (ainsi k(j,{7,8}) est le nombre des doc. où le mot est employé 7 ou 8 fois etc.).

En revanche le zéro joue un rôle essentiel. Nous renvoyons le lecteur à la thèse, ou à l'article qui en est extrait : il y verra dans l'un des quadrants du plan 1 × 2, des mots apparaissant volontiers avec une fréquence relativement élevée (mais absents chacun de la majorité des documents) et susceptibles à première vue d'être ceux qui signent le thème du document. Hypothèse qu'on a pu vérifier en analysant le tableau I × J1, croisant l'ensemble des documents avec l'ensemble des mots de ce quadrant, puis soumettant I et J1 à la classification ascendante hiérarchique. On a aussi obtenu des partitions satisfaisantes (notées D pour I ; et V pour J1) : le tableau D × V (k(d,v) = nombre total des occurrences des mots de la classe v dans les doc. de la classe d) a lui-même été analysé, offrant une vue d'ensemble des rapports des classes. Enfin tout document nouveau is peut être décrit par un profil sur V dont toute composante k(i,sv) est la somme des k(is,j) pour j dans v . Il faut noter que la conclusion générale étant que les mots qui apparaissent une seule fois dans un document en caractérisent aucunement le contenu, on a avantage à ne tenir compte dans la construction du tableau I × J1, ou des k(i,v), que des k(i,j) strictement supérieurs à 1.

Ces résultats obtenus sur un petit corpus particulier (264 compte-rendus de voyages d'étude ; tenant pour la plupart dans une page) sont-ils généralisables ? Certainement *non* , si on s'en tient à la lettre des conclusions que l'on peut formuler ; mais *oui* affirmons-nous, si l'on retient seulement que l'analyse des répétitions est un outil efficace pour segmenter le vocabulaire (ou plus

généralement tout ensemble d'unités linguistiques : morphèmes, "constructions-type", etc.). L'obstacle immédiatement visible à la généralisation est que le nombre n des occurrences des mots dans un texte ne prend sens que relativement à la longueur de ce texte : l'analyse de la thèse de A.H. A. n'a réussi que parce que les textes (à une ou deux exceptions près qui ne sont d'ailleurs pas passées inaperçues) sont du même ordre de grandeur. Que faudrait-il faire autrement ? Découper chaque texte en pages ou mieux en paragraphe ? C'est là un palliatif simple et sans doute efficace. Mais le plus juste nous paraît être de caractériser chaque emploi d'un mot j (dans un texte i) par le nombre et la proximité des répétitions qu'en offre le contexte.

6 *Le décompte des cooccurrences dans un contexte*

I : ensemble de textes ;
J : ensemble de mots ;
L(i) : longueur du texte i comptée en mots (il serait facile de repérer la place d'un mot de i non par le nombre des mots précédents ; mais par un nombre de lettres.

t(i,x): le x-ème mot du texte i : défini pour $x \leq L(i)$.
L : un seuil de longueur, d'ailleurs ajustable.

$$C(i,jx,L) = \text{Card}\{y \mid y \leq L(i) \ ; \ |x - y| \leq L \ ; \ t(i,y) = j\} \ ;$$

la fonction C donne le nombre d'occurrences du mot j dans l'intervalle de texte i, centré au x-ème mot, et de longueur 2L-1 ;

$$\text{Cum}(i,j,n,L) = \text{Card}\{x \mid x \leq L(i) \ ; \ C(i,j,x,L) = n\} \ ;$$

la fonction Cum donne le nombre des points x du texte i qui se trouvent au centre d'un intervalle de longueur 2L-1 contenant exactement n occurrences du mot j. Si par exemple le mot j est absent du texte i, on a Cum(i,j,0,L) = L(i) ; et pour $n \neq 0$, les Cum (i,j,n,L) sont nuls ; en général on a, quels que soient i, j, L

$$\Sigma\{\text{Cum}(i,j,n,L) \ n \in N\} = L(i).$$

Si le texte i est de longueur inférieure à L, et contient n occurrences du mot j, alors la fonction C(i,j,x,L) est constamment égale à n quand x varie de 1 à L(i) ; et Cum(i,j,n,L) = L(i), tandis que les autres Cum(i,j,n',L) sont nuls. En ce sens le profil sur N du vecteur (de total L(i)) :

$$\{\text{Cum}(i,j,n,L) \mid n \in N\},$$

généralise le nombre d'occurrences de j dans i, quand le décompte est fait dans un contexte de longueur 2L-1. On verra que le calcul de la fonction Cum est aisé, une fois reconnues les occurrences du mot j dans le texte i.

On construit maintenant un tableau $J \times N$ qui, en bref recense sur l'ensemble du corpus les grappes de n-occurrences de j survenant dans un intervalle d'amplitude 2L-1 :

$$\text{rep}(j,n,L) = \Sigma\{\text{Cum}(i,j,n,L) \mid i \in I\}$$

On notera que, comme dans la thèse de A.H. A., chaque ligne j du tableau rep a même total ; dans la thèse ce total est Card I ; ici on a ,

$$\Sigma_n\{rep\ (j,n,L)\} = \Sigma_{i,n}\{Cum(i,j,n,L)\} = \Sigma_i\{L(i)\} = LTOT$$

où LTOT est le nombre total des occurrences du corpus.

Il semble bon d'autre part, de considérer le tableau binaire grap(I × N), qui donne en bref, pour chaque document i le nombre des grappes de mots, de quelque mot qu'il s'agisse :

$$grap(i,n,L) = \Sigma_j\{Cum(i,j,n,L)\}$$

La somme sur j pouvant être étendue à un sous-ensemble de J choisi *ad libitum* : si on prend $j \in JT$, où JT désigne le vocabulaire technique (que l'analyse de rep(J × N) a pour objet de cerner) la présence de grappes signalera le caractère technique du document. Avec des mots non techniques on attendrait plutôt un aspect du style de l'auteur.

Manifestement la construction proposée dépend du seuil L choisi : mais il est facile de paramétrer le programme et de confronter les résultats obtenus avec plusieurs valeurs de L, voire d'analyser des tableaux juxtaposés correspondant à plusieurs valeurs de L.

7 *Méthode de décompte des cooccurrences* : Dans ce § pour simplifier l'écriture, nous omettrons la variable L.

Dans les construction du § 6, la fonction Cum(i,j,n,) joue un rôle central. On peut se proposer de calculer cette fonction pour (i,j) fixés : les résultats obtenus étant ajoutés immédiatement au tableau rep(J × N) et grap(I × N), si, (comme ce sera généralement le cas) on ne peut garder le tableau ternaire Cum(I × J × N).

Le calcul de $\{Cum(i,j,n) \mid n \in N\}$ repose sur la fonction $\{C(i,j,x) \mid x \leq L(i)\}$: cette fonction est constante sur des intervalles successifs dont les bornes ne sont autres que celles des intervalles d'amplitude (2L-1) centrées sur les occurrences du mot j dans le texte i . Il suffit de classer ces bornes pour avoir les intervalles de niveau de la fonction C ; et de ventiler sur Cum les longueurs de ces intervalles.

De façon précise soit KIJ le nombre des occurrences du mot j dans le texte i ; les abscisses de ces mots sont données par un tableau AB[1:KIJ] ; les abscisses des bornes des intervalles de rayon L centrés sur ces occurrences forment deux suites ABP[1:KIJ] et ABM[1:KIJ] avec :

$$\forall\ R \in [1:KIJ] : ABP[R] + L = 1 + ABM[R] - L = AB[R]$$

(On verra dans la suite pourquoi nous avons posé ABM = AB + L - 1). En interclassant les abscisses contenues dans les tableaux ABP et ABM, on a un tableau DAB [1:2KIJ] qui donne les bornes gauches et droites des intervalles, il faut seulement noter sur un autre tableau SIG[1:1KIJ] s'il s'agit d'une borne gauche (SIG = +1) ou droite (SIG = -1). Par exemple si L = 5 ; KIJ = 3 ; on pourra avoir :

AB=	2	10	19

; ABP=

-3	5	14

; ABM=

6	14	23

DAB=	-3	5	6	14	14	23
SIG=	+	+	-	+	-	-

(Si deux valeurs du tableau DAB sont égales, l'une a signe +, l'autre signe - ; et elles peuvent en fait être supprimées pour la suite).

Il est maintenant facile de calculer de proche en proche la fonction C et donc de remplir le tableau CUM. On a (en omettant désormais les indices i, j), quel que soit X :

$$C[X] = \Sigma\{SIG[R] \mid R \in [1{:}2KIJ] \ ; \ DAB[R] <_s X\} \ ;$$

ceci permet de calculer C[1] comme le nombre des indices R pour lesquels $DAB[R] \leq_s 0$. Dans l'exemple C[1] = 1) ; ensuite la valeur de C change chaque fois qu'on *dépasse* une valeur de DAB (la convention ABM = AB + L - 1 a été posée pour cela) chaque intervalle de niveau N de la fonction C (depuis X = 1 jusqu'à X = LI longueur du texte i) est à ajouter dans CUM[N], d'où en bref l'algorithme suivant.

```
      C:=0 ; B:=0
      pour R:=1 pas 1 tant que DAB[R]≤0 faire
           C:= C+1 ;
Commentaire : on a calculé C = C[1].
      pour R:=C+1 pas 1 tant qur DAB[R]≤LI - 1
      faire début
          CUM[C]:=CUM[C]+DAB[R]-B ;
          B:=DAB[R] ; C:=C+SIG[R] fin
```

Commentaire : chaque valeur DAB[R] tient lieu de borne incluse dans l'intervalle qu'elle limite à droite ; et non incluse dans l'intervalle qu'elle limite à gauche ; la fonction C est modifiée de SIG[R] à la traversée de la borne DAB[R].

```
      CUM[C]:=CUM[C] + LI - B.
```

Commentaire : le dernier intervalle est celui qui se termine par le dernier mot du texte (abscisse LI). A titre de vérification, on notera que si le tableau DAB est vide (KIJ = 0) seules jouent les instructions C:=0 ; B:=0 ; CUM[0]:=LI. On notera que nous n'avons précisé ni la mise à zéro de CUM, ni les bornes [1:2KIJ] sur lesquelles il est licite de lire le tableau DAB.

8 *Conclusion*

Nous répéterons qu'il ne fait pas de doute pour nous que l'étude des répétitions servira non seulement en documentation mais dans les analyses de cooccurrences des traits linguistiques en prose ou en poésie.