

J. P. BENZÉCRI

## **Annnonce de la parution d'un ouvrage sur la pratique de l'analyse des données en linguistique**

*Les cahiers de l'analyse des données*, tome 5, n° 4 (1980), p. 481-491

[http://www.numdam.org/item?id=CAD\\_1980\\_\\_5\\_4\\_481\\_0](http://www.numdam.org/item?id=CAD_1980__5_4_481_0)

© Les cahiers de l'analyse des données, Dunod, 1980, 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/>

ANNONCE DE LA PARUTION D'UN OUVRAGE  
SUR LA PRATIQUE DE L'ANALYSE DES DONNÉES  
EN LINGUISTIQUE  
[PRAT. LING.]

par J. P. Benzécri et coll

1 Thème du livre : Le présent volume (PRA3 = LING), troisième de la collection *Pratique de l'Analyse des Données* traite des applications à la linguistique, à la lexicologie et plus généralement aux disciplines qui étudient le langage.

L'objet de l'analyse des données est d'acquérir une vue globale simultanée d'un ensemble I d'objets et d'un ensemble J de propriétés par lesquelles sont décrits les éléments de  $\mathcal{T}$  ; ou encore de mettre sous forme géométrique le système des relations existant entre les éléments de deux ensembles I et J ; voire seulement au sein d'un seul ensemble I. Pour cela on constitue à partir des faits observés des tableaux rectangulaires de nombres, ou tableaux de correspondance ; l'élaboration mathématique de ces tableaux se faisant ensuite nécessairement par un outil de calcul électronique.

Le linguiste contemporain, familier avec le terme de structure, admet spontanément l'intérêt de notre projet. Le statisticien qui voit dans l'analyse multidimensionnelle le *Novius Organum* de l'honnête homme du XXI-ème siècle, est prêt à servir les lettres. Plusieurs dizaines d'études ont déjà été faites : et rien, sinon le labeur requis pour recueillir les données, ne nous semble freiner le progrès des recherches qui sont le thème de ce livre. On a donc voulu s'adresser tant au spécialiste du langage ou de la statistique qu'à tout esprit ambitieux à qui rien des sciences n'est étranger.

2 Forme de l'exposé : L'ouvrage commence par une partie d'introductions (désignée par le sigle LA), qui devrait suffire aux lecteurs de formation diverse pour découvrir l'ensemble des problèmes et des méthodes ; puis étudier les exemples qui font l'objet des parties suivantes. Dans ces parties numérotées de LC1 à LC4, de nombreux tableaux de données sont soumis à l'analyse factorielle, complétée éventuellement par la classification automatique.

Comme de multiples domaines sont abordés (mais, avouons-le, très inégalement couverts), on n'a pu se baser sur ceux-ci pour ordonner la table des matières : linguistique historique, dialectologie, phonétique, stylistique, sémantique ou syntaxe ne sont donc pas ici des titres de chapitres, mais des thèmes qu'on ne devrait jamais perdre e vue dans la succession des quatre parties LC ; celles-ci étant rangées suivant la complexité de la construction du tableau analysé : LC1 : tableau de contingence ; LC2 présence-absence ; LC3 : questionnaires ; LC4 : séquences. Ce plan (cf § 3) qui repose sur des considérations de forme et presque de technique a, à nos yeux, un double avantage. D'une part entre chercheur de formation littéraire et statisticien, le tableau des données est le moyen de communication obligé : au seuil de toute étude, le premier effort requis est un effort

d'abstraction logique et mathématique pour concevoir un tableau pertinent ; on pouvait donc graduer les exemples suivant cette difficulté majeure. D'autre part le pouvoir des mathématiques en général et particulièrement de la statistique multidimensionnelle se fonde sur l'analogie des formes. Ce qu'on a appris en traitant le tableau en présence-absence ( cf *infra*, § 3 LC2) construit par un écologiste sert non seulement en sédimentologie, mais aussi en linguistique comparée, à plus forte raison on se gardera de séparer l'analyse d'un questionnaire syntaxique de celle des variantes des états d'un texte puisque dans l'un et l'autre cas, il s'agit d'une suite de lieux, en chacun desquels un ensemble fini de modalités peuvent s'insérer ( cf *infra*, § 3 LC3). Plutôt que d'enfermer le chercheur dans un chapitre dont le titre (paléographie ou analyse du contenu...) lui est familier, on le contraindra donc à traverser de longues galeries d'exemples exotiques, propres pourtant à lui servir de modèle.

3 Le plan adopté : Ainsi qu'on l'a annoncé, le livre comprend cinq parties. Nous en donnons ci-dessous le sommaire détaillé .

### 3.1 LA Introductions :

Sommaire : Cette partie comprend trois articles, dont le premier est une introduction générale aux problèmes et aux méthodes de tout le livre ; tandis que les deux autres concernent respectivement la linguistique générale et la statistique multidimensionnelle.

LA n° 1 [ANA. LING.] Analyse statistique des données linguistiques ; par J.P. Benzécri : Au printemps de 1973, le laboratoire de statistique, hôte à Grenoble du Groupe d'Etudes pour la Traduction Automatique que dirige B. Vauquois, tint un colloque sur l'analyse des données linguistiques . Grâce à Mademoiselle Françoise Gardes-Madray - de l'université de Montpellier - un écho magnétophonique de ce colloque nous a été conservé. Le présent article reprend en substance l'introduction enregistrée. Dans la rédaction pas plus que dans l'exposé oral, ne sont séparés en des parties distinctes, principes linguistiques, techniques statistiques et exemples illustratifs. On a tenté au contraire, d'initier simultanément comme par degrés, linguiste et statisticien à la discipline l'un de l'autre ; et de les convaincre tous qu'ils se devaient de collaborer.

LA n° 2 [LING. GEN.] Linguistique générale ; par J.P. Benzécri : Le terme même d'analyse des correspondances remonte à l'automne 1962, et le premier exposé de la méthode fut donné au Collège de France dans une Leçon du cours Peccot de l'hiver 1963. Une version de ce cours, rédigé comme une suite de leçons, fut multigraphiée à Rennes en 1963-1964. Nous en reproduisons ici la première leçon où l'on verra de quelles conceptions linguistiques est issu le projet initial de fonder sur la statistique une linguistique inductive. Le dernier paragraphe du texte original où s'introduit l'analyse statistique proprement dite, est ici remplacé par un exposé plus détaillé, extrait d'un article de la *Revue Philosophique* (Linguistique et Mathématique; pp 309-374 ; 1966) et rédigé principalement d'après des interventions au colloque de linguistique appliquée tenu à Nancy en octobre 1964.

LA n° 3 [INT. CORR. CLASS.] Introduction à l'analyse des correspondances et à la classification automatique ; par F. Benzécri (d'après un exposé de M. Morfin). Cet article, écrit en 1979-1980 comprend trois § : 1 Bases géométriques de l'a. des corr. ; 2 Lecture et interprétation des listages ; 3 Classification automatique. Les §§ 1 et 2 (à l'origine desquels est un exposé présenté par M. Morfin à Montpellier, en mai 1976) offrent une version simplifiée des parties I et III du livre ENSI (Analyse des Corr. : Exposé élémentaire ; Dunod 1980). Le §3, très élémentaire lui aussi, est à notre connaissance le premier exposé de la classification ascendante hiérarchique rédigé à ce niveau.

3.2 LC1 Tableaux de contingence :

Sommaire : Cette partie comprend 10 études qu'on rangera sous trois titres : A Textes et mots ; B Vers et phonèmes ; C Bases et morphèmes.

A Textes et mots : Dans 5 études on considère la correspondance entre un ensemble I de textes et un ensemble J de mots, le tableau de contingence analysé donnant à l'intersection de la ligne i et de la colonne j le nombre  $k(i, j)$  de fois que le mot j se rencontre dans le texte i. Avant de présenter ces études une par une, nous préciserons ce qu'on entend par mot et par texte, et pourquoi l'on effectue de telles analyses.

QUID : Que compte-t-on? Du point de vue matériel, on peut appeler *forme* une suite de lettres comprise entre deux blancs : e.g. *prennent* est une forme qu'on peut rencontrer dans un texte français ; *prennent*, *prenons* et *prirent* sont trois formes différentes qui relèvent toutes du même mot, le verbe *prendre*; de plus, la forme unique *prennent*, peut recevoir deux *étiquettes* différentes : 3-ème personne du pluriel de l'indicatif présent ; et 3-ème p. du plur. du subjonctif présent. Dans la plupart des études, on compte soit des *mots* soit des *formes* (non étiquetées). Des programmes assez simples permettent de dénombrer automatiquement les *formes*, une fois que le texte perforé a été introduit dans l'ordinateur. En revanche, pour reconnaître les *mots* sous toutes leurs formes (et *a fortiori* pour étiqueter les formes) il faudrait des programmes d'analyse morphologique (et même syntaxique) : ces programmes sont complexes et laissent subsister des ambiguïtés ; finalement ; il est indispensable qu'un expert lise le texte et en analyse les formes. Mais quel qu'en soit le coût, cet étiquetage nous paraît très intéressant (cf e.g. [INV. LEX. 1793] § 1 ; et ci-dessous titre C).

UBI : Où compte-t-on? Certaines études portent sur un ouvrage unique qu'on divise en un ensemble I de chapitres ou de fragments : soit qu'on accepte la division usuelle de l'ouvrage, soit qu'on le sectionne, e.g. systématiquement en tranches successives de 500 mots. Parfois les dénombremments portent sur un ensemble I de textes indépendants les uns des autres. Eventuellement on a un corpus dont chacun des textes est subdivisé en fragments. La définition des textes et fragments devrait dépendre du but de l'étude ; mais souvent, le statisticien ne peut qu'accepter les données disponibles...

CUR : Pourquoi compte-t-on? Souvent l'intérêt du chercheur porte d'abord sur un seul des deux ensembles : soit I (textes ou fragments) ; soit J (mots ou formes). Là, on se demande si un ouvrage reçu comme un tout, ne doit pas être partagé entre plusieurs auteurs ; ou encore si pour un ensemble de textes dont les auteurs sont connus avec certitude, les fréquences relatives d'emploi des mots fournissent une signature, reconnaissable malgré la diversité des thèmes. Ici, particulièrement s'il s'agit de textes anciens, on veut préciser le sens des mots d'après les affinités mutuelles que révèlent leurs associations au sein des textes. Mais en définitive, quel que soit le problème initial, l'analyse factorielle offre des deux ensembles en correspondance, une représentation simultanée où chaque élément, mot ou texte, est illustré par tous ses voisins. C'est ce qu'on verra sur les exemples.

LC1 n° 1 [ISAIE] : Le livre d'Isaïe et l'A. critique des sources textuelles par G.E. Weil, A. Salem et M. Serfati. J est l'ensemble des 89 mots les plus fréquents au sein du texte hébraïque du livre d'Isaïe ; I est l'ensemble des 66 chapitres en lesquels ce livre a été découpé. On s'interroge sur l'homogénéité du livre. Pourquoi 89 mots ? Parce

qu'initialement on en avait pris 100, dont on a préféré ôter 11 noms propres ; pourquoi 66 chapitres ? Parce que cette division introduite seulement au moyen âge par un érudit latin (Etienne Langton, 1205) est aujourd'hui dans toutes les éditions. Avec le tableau I x J (66x89), on a analysé d'autres tableaux construits en groupant les chapitres par trois ou par cinq ; en éliminant certains mots... Il est apparu de façon stable que le livre peut être divisé en quatre parties. Entre ces parties, les différences de thème sont manifestes ; mais pour décider si l'on doit ou non les rapporter à un seul homme, il faudrait au moins savoir dans quelle mesure le style d'un même auteur varie avec le sujet qu'il traite et quels en sont les traits invariants ; ce que seules d'amples expériences statistiques nous apprendront. Une expérience relativement facile à réaliser (parce que d'ampleur limitée) et pourtant instructive serait de soumettre aux mêmes analyses que celles rapportées ici, non le texte hébraïque du livre d'Isaïe, mais ses versions antiques ou modernes (grecques, latines, françaises) dont la plupart ont certainement été rédigées d'une seule pièce ; en sorte que si l'on y retrouve la quadripartition du livre, cela résultera de la multiplicité des thèmes traités (et dans une moindre mesure les hétérogénéités du style de l'original), non d'un changement de rédacteur.

LC1 n° 2 [ESCLAV. ATT.] : Le champ lexical de l'esclavage chez les orateurs attiques par M.M. Mactoux. I est un ensemble de V plaidoyers qui constituent un échantillon préliminaire du corpus des orateurs attiques ; J est un ensemble de 110 mots grecs employés dans ces plaidoyers et rentrant tous dans le champ lexical de l'esclavage. Grâce à sa connaissance des textes l'auteur parvient à interpréter les facteurs comme autant de dimensions thématiques d'après lesquelles elle peut notamment distinguer entre les acceptions techniques de mots, qui pour l'usage commun qu'attestent les lexiques, sont dans l'ombre.

LC1 n° 3 [EVANGILES] : Le vocabulaire des évangiles ; analyse des similitudes entre chapitres de Jean par B. de Solages et J.M. Vacherot. Les auteurs ont constitué un premier tableau : J ensemble des 70 mots utilisés plus de 50 fois dans le texte grec de l'un au moins des 4 évangiles ; I ensemble des 4 évangiles ; puis un second tableau croisant avec un vocabulaire J' un peu modifié, l'ensemble I' constitué par les 21 chapitres de Saint Jean et les 3 autres évangiles non subdivisés. Ainsi on peut apprécier l'homogénéité du texte de Jean après l'avoir confronté globalement aux autres textes. Comme à propos d'Isaïe on peut suggérer ici entre autres recherches ultérieures des comparaisons de traductions ; et aussi la comparaison des chapitres de l'évangile avec les chapitres de la version grecque de l'ancien testament ; notamment pour déceler les sémitismes des évangiles.

LC1 n° 4 [JOIE ANC. FR.] : Le thème de la joie en ancien français : analyse des cooccurrences de 30 vocables dans 15 romans par Ch. Tronc. Deux tendances se manifestent : aspect social et ludique de la joie, d'une part ; appréhension intime et individuelle de la joie, de l'autre.

LC1 n° 5 [INV. LEX. 1793] : Signalement et inventaire lexical : textes politiques français de 1793 par A. Salem, analyse de données recueillies par le laboratoire d'Etude des textes politiques français de l'E.N.S. de Saint-Cloud. Il s'agit de dénombrement de formes (i.e. avaient mais non le verbe avoir) effectué dans des discours politiques ou des journaux : l'analyse factorielle montre ici son aptitude à discriminer des textes d'origine différente d'après le seul inventaire des formes.

B Vers et phonèmes : On range sous ce titre deux études.

LC1 n° 6 [PHON. IND.] : Analyse de quelques données statistiques phonétiques sur les langues de l'Inde : d'après des tableaux donnant les fréquences de quelques dizaines de sons, (ou des modèles de structure de syllabe) dans des langues indo-européennes ou dravidiennes de l'Inde, ainsi qu'en anglais, l'analyse dispose sur un axe les langues et les faits dénombrés.

LC1 n° 7 [PROFILS POEMES] : Recherches sur les profils sonores des textes poétiques par A. Hathout, M. Reinert et coll.. En dénombrant les sons dans des vers ou fragments poétiques de Racine, Dannunzio, Rimbaud... les auteurs cherchent au-delà des règles de versification d'autres structures par lesquelles on reconnaît la voix d'un poète.

C Bases et morphèmes : On range sous ce titre trois études.

LC n° 8 [ALL. FRANC.] : Analyse des fréquences d'emploi des parties du discours en allemand et en français : d'après des données recueillies par J.M. Zemb, on caractérise ici un aspect du style de 20 auteurs des deux langues.

LC1 n° 9 [NOMS VERBES RUSSES] : Analyse statistique des emplois de mots en langue russe : cas des noms et rection des verbes par J.J. Castot. D'après le dictionnaire d'Evy Steinfeldt, on analyse d'une part la correspondance entre noms et cas, d'autre part celle entre verbes et rections : ces études (dont la première est la plus approfondie) permettent d'extraire de données morpho-syntaxiques, des facteurs dont l'interprétation est nettement sémantique.

LC1 n° 10 [VERBES FR.] : Les emplois par mode, temps et personne de cent verbes les plus fréquents dans un corpus de textes littéraires français par Fr. Gardès-Madrav, analyses de S. Stépan : ici encore l'analyse de faits morpho-syntaxiques débouche sur le sens, et suggère une vue d'ensemble des verbes ainsi qu'une conception nouvelle des rôles des personnes et du système des modes et temps. Du point de vue formel, il s'agit ici d'une correspondance entre 4 ensembles, qui offre matière à de multiples analyses, ainsi qu'on l'explique dans l'article.

### 3.3 LC2 Tableaux en présence-absence :

Sommaire : Cette partie comprend 4 études (n°s 1 à 4) accompagnées de deux notes méthodologiques (n°s 1' et 2'). Le thème général en est l'analyse des tableaux en (0,1) dont le type est le tableau  $I \times J$  croisant un ensemble  $I$  d'individus avec un ensemble  $J$  de caractères et où  $k(i,j) = 1$  si l'individu  $i$  possède le caractère  $j$  et zéro sinon. Ainsi qu'on l'a expliqué dans l'avant-propos, rentre dans ce type le cas particulier des tableaux sous forme disjonctive complète, ou tableaux de questionnaires ; l'analyse de ces derniers tableaux qui offre des propriétés particulières, fait l'objet de la partie suivante LC3.

LC2 n° 1 [FEUILLETONS] : Les feuillets du Petit Journal de 1890 à 1894 : contenu des présentations et typologie des personnages ; par H. Steinberg (Mme H. Castro) et M. Lazare. L'étude repose sur deux tableaux en présence-absence. Le premier croise 27 textes de présentation avec 65 mots ou locutions. Le second croise 214 personnages avec un ensemble  $J$  de traits de signalement, ou de modalités de variables sociologiques. On conçoit que la description des personnages, fondée sur des informations relevées au fil de la lecture, offre matière à des variantes dans le codage et le choix des éléments principaux. Ce considérable travail d'analyse du contenu, aboutit à une vue d'ensemble de la société du feuilleton : échelle d'appréciation ; établissement social ; beauté et laideur ; ... sont autant d'axes où il vaut la peine de considérer en détail la place de chacun des attributs.

LC2 n° 1' [RECH. SENS] : La recherche du sens ; par J.P. Benzécri. A propos de la typologie des personnages du feuilleton, on considère suivant quelles voies celui qui dénombre des mots, des sons, des morphèmes... peut atteindre le sens. Sans éliminer la part subjective du jugement, le traitement mécanique d'un tableau qu'on avait empli sans le voir, réserve des surprises, qu'on doit accepter pour échapper à soi-même.

LC2 n° 2 [EPITRE N. TEST.] : L'analyse factorielle au service de l'édition de textes anciens : application à un texte grec du nouveau testament, l'épître de Jacques ; par Ch. Amphoux. L'auteur considère d'abord un tableau en présence-absence, 74 x 101, croisant un ensemble de 74 témoins (i.e. états attestés par un manuscrit etc.) du texte de l'épître, avec un ensemble de 101 variantes (i.e. formes possibles du texte en un certain nombre de lieux, dits lieux variants où les témoins ne s'accordent pas unanimement) ; mais sans retenir toutes les variantes de chaque lieu (ce qui fait sortir de la forme disjonctive complète objet de la partie LC3). Puis l'analyse ayant permis de cerner un nouveau groupe de témoins (en plus des deux déjà bien connus), on consacre à ce groupe une étude plus approfondie.

LC2 n° 2' [TEXTES ANCIENS] : Classification des états d'un texte, mathématique et informatique, d'après J. Duplacy ; analyse de J.P. Benzécri. En suivant dans leur développement historique les progrès de la science des textes on voit s'imposer l'idée que les données de philologie doivent être conçues comme un questionnaire appliqué aux états d'un texte. C'est pourquoi J. Duplacy suivi bientôt par Ch. Amphoux (cf supra LC2 n° 2) a ouvert aux philologues les voies de l'analyse factorielle.

LC2 n° 3 [CHAM. SEM.] : Contribution à la taxinomie des langues chamito-sémitiques : analyse des données rassemblées par Marcel Cohen dans son essai comparatif sur le vocabulaire et la phonétique du chamito-sémitique ; par J.P. Benzécri. On a construit un tableau en 0,1 donnant pour quelque 500 racines des indications de présence ou d'absence dans une trentaine de langues, dialectes ou groupe de dialectes. Les langues étant inégalement inventoriées, le tableau se présente sous diverses variantes selon que certaines informations sont conservées séparément, cumulées ou mises en éléments supplémentaires : d'où de multiples analyses factorielles, corroborées par des classifications automatiques. La conclusion linguistique est que quant au vocabulaire retenu par M. Cohen, on doit distinguer deux grandes sous-familles sémitique (y compris le berbère ; et l'égyptien ancien, particulièrement proche du sémitique proprement dit) ; et couchitique (auquel s'adjoint le haoussa, seul représentant des langues tchadiennes dans nos données).

LC2 n° 4 [ATLAS] : L'atlas linguistique et ethnographique de la Bretagne romane, de l'Anjou et du Maine ; par M. Baka et J.P. Chauveau. Les données analysées proviennent d'un ensemble Q de 21 cartes donnant chacune pour un mot (e.g. orge, moisson...) la forme attestée en chaque lieu i d'un ensemble I de 126 points d'enquête. Au total on a retenu un ensemble J de 153 modalités ou formes de mots (4 pour orge ; 9 pour moisson...) ; d'où un tableau 126 x 153, avec  $k(i,j) = 1$  si la forme j est attestée au lieu i ; et zéro sinon. Du point de vue formel, ce tableau est non seulement un tableau en présence-absence, mais un tableau sous forme disjonctive complète : chaque carte q étant comme une question admettant un ensemble Jq de modalités de réponse : les formes du mot ; il conviendrait donc de placer la présente étude dans la partie LC3. Toutefois le thème linguistique étant analogue à celui de [CHAM. SEM.] on a préféré ne pas séparer les deux articles ; qui rentrent donc dans la partie LC2.

3.4 LC3 Questionnaires :

Sommaire : Cette partie comprend 7 études précédées d'une introduction à l'a. des questionnaires.

LC3 n° 0 [INT. QUEST. LING.] : Introduction à l'analyse des questionnaires d'après un exemple de données linguistiques ; par J.P. Benzécri. Ainsi qu'on l'a expliqué dès l'avant propos du présent volume, toute analyse de données rentrant dans le format des questionnaires repose sur un tableau en présence-absence (cf LC2), croisant un ensemble I d'individus et un ensemble J de modalités descriptives ; avec cette particularité (caractéristique de la forme disjonctive complète) que les modalités (colonnes) vont par blocs et que la ligne afférente à chaque individu présente dans chaque bloc un seul chiffre 1, avec ailleurs des zéros. Sur un exemple schématique de données extraites de l'étude de J.C. Sergeant (LC3 n° 6) on présente ici au linguiste tous les traitements auxquels se prêtent de telles données.

On montre en particulier que l'analyse du tableau descriptif (en 0,1) équivaut à celle du tableau de Burt (ou tableau de cooccurrences des modalités :  $k(j, j')$  = nombre des individus possédant à la fois les modalités  $j$  et  $j'$ ) qui est un véritable tableau de contingence. Ainsi on légitime en général l'extension aux tableaux en (0,1) des calculs de l'a. des correspondances, conçus d'abord pour les t. de contingence.

LC3 n° 1 [OUI NON ABS.] : Sens et valeur des réponses à un questionnaire clos ; par Mesdames F. Pétaïpermal et M. Rigaud. L'analyse des réponses à une enquête produit en général un ou plusieurs facteurs parmi les premiers qui se rapportent non au contenu des questions posées mais à la forme "oui" "non" "abstention" des attitudes permises. Sur ce fait de parole, on attire ici l'attention des linguistes.

LC3 n° 2 [MOYEN] : Actif et moyen dans le verbe : questionnaire sur le contenu d'une opposition linguistique ; par J.P. Benzécri et S. Stépan. Se fondant sur le sens d'une vingtaine de verbes indoeuropéens qui sont soit seulement actifs, soit seulement moyens, E. Benveniste propose une définition de l'opposition entre actif et moyen. Cette définition a été soumise par nous à un certain nombre de linguistes et de statisticiens, en leur demandant de répartir d'après eux en deux classes les vingt verbes dont est parti E.B. . On analyse ici les réponses reçues.

LC3 n° 3 [MOD. ESP.] : L'usage de l'opposition modale en espagnol ; analyse d'une enquête ; par C. Salaün ; (traitements statistiques de S. Stépan). Pour explorer le choix modal de la langue espagnole, on a proposé à quelque 350 sujets un ensemble de 36 phrases, où le verbe donné à l'infinitif devait être mis à une forme personnelle choisie d'après le contexte. On étudie ici d'après cette enquête l'usage moyen puis ses variantes.

LC3 n° 4 [REP. LIBR.] : Vers l'analyse automatique des textes : le traitement des réponses libres aux questions ouvertes d'une enquête ; par L. Lebart. On présente ici une chaîne de programmes, permettant d'une part de classer et d'éditer les réponses libres d'après les réponses fournies par les mêmes sujets à des questions closes ; et d'autre part de dénombrer les mots les plus fréquents usités dans les réponses libres pour les mettre en correspondance avec les attitudes des sujets.





ici une ligne  $i$  recense à la fois les éléments précédant  $i$  et ceux qui le suivent : i.e.  $k(i, i^-) = k^1(i', i) =$  nombre de fois que  $i'$  précède  $i$  ; et  $k(i, i^+) =$  nombre de fois que  $i'$  suit  $i$ . Plus généralement, on peut mettre quatre blocs de colonnes  $I--$ ,  $I-$ ,  $I+$ ,  $I++$  avec  $k(i, i^{--}) = k^2(i', i)$  (nombre de fois que  $i'$  précède  $i$  avec un intermédiaire) ;  $k(i, i^-) = k^1(i', i)$  ;  $k(i, i^+) = k^1(i, i')$  ; et

$k(i, i^{++}) = k^2(i, i')$ . C'est ainsi qu'on a analysé la suite des lettres de textes écrits dans diverses langues ; le résultat étant toujours sur le 1-er axe une nette séparation entre consonnes et voyelles : voici par exemple le résultat obtenu avec l'espagnol (langue où l'écriture suit presque exactement la prononciation), en comptant le blanc (#) comme une lettre.

E	A	O	I	U	#	P	K	J	D	X							
						Y	B	C	L	F	N	J	K	D	Q		
												H	S	R	T	V	M

La présente partie LC4 offre des variantes de telles analyses. Dans [ANA. MUS.] on considère l'étude simultanée de plusieurs séquences (partition à plusieurs voix : la construction des tableaux est expliquée au § 2). Dans [EPINOCHÉ], il s'agit d'un dialogue de gestes, dont chacun peut être précédé ou suivi d'un geste émis soit par le même animal, soit par son adversaire (cf. LC4 n° 4§ 5).

C Position au sein de la séquence : Une telle position peut être caractérisée par sa distance aux extrémités et par la longueur de la séquence elle-même ; ainsi apparaît l'affinité d'un élément avec les séquences courtes plutôt qu'avec les séquences longues ; ou encore avec le début des séquences plutôt que la fin de celles-ci : cf. e.g. [EPINOCHÉ] § 4, ou LA n° 1 § 3.1. Éventuellement (cf. [DIST. CHIN.] les séquences ont toutes même longueur.

D À la recherche d'une syntaxe : En bref du point de vue de la syntaxe, un texte est considéré comme une suite d'éléments (mots ou syntagmes) dont on ne retient que l'appartenance à une classe de parties du discours (nom, verbe, adjectif, adverbe par exemple), en faisant abstraction du sens. En présence d'un corpus de séquences, on cherchera donc à en répartir les éléments en un petit nombre de classes ; de telle sorte que dans chaque classe les éléments soient à peu près substituables entre eux (on dit : soient équivalents distributionnellement cf. LA n° 2 § 4.2 ; LA n° 1 § 2.1) ; puis les séquences étant ainsi schématisées (comme le serait un texte dont on remplacerait les mots par la seule indication de la partie du discours dont chacun de ceux-ci relève) on cherchera les règles de succession de ces classes (comme on a nom-verbe-nom pour schéma d'une phrase etc.). À de telles recherches - où en fait la syntaxe ne peut être séparée de la sémantique - se prêtent remarquablement les distiques chinois, grâce à la simplicité de leur structure ; et aussi au parallélisme rigoureux qui en fournit la clef : cf. LA n° 2 [DIST. CHIN.].

LC4 n° 1 [MON. CHIN.] : Analyse des monosyllabes chinois, par J.P. Benzécri. Le terme de syllabe, évoque d'abord les deux modèles les plus fréquents : syllabe ouverte CV (consonne-voyelle ; e.g. ba) et CVC syllabe fermée, e.g. bac. En chinois il convient plutôt de couper la syllabe en deux parties : d'une part un préphonème ou consonne initiale (éventuellement absente, comme dans toute langue e.g. en français a, ac ; on parle alors de préphonème vide) ; d'autre part une finale qui peut être complexe, centrée sur une voyelle et appelée ici postphonème. Il s'en faut de beaucoup que soient attestées dans la langue toutes les associations d'un préphonème avec un postphonème : d'où un tableau de compatibilité entre ces éléments qui est le premier qu'en 1963 nous ayons soumis à l'a. des c. . L'étude se complique parce que l'inventaire des pré - et postphonèmes admet des variantes ; qu'on peut tenir compte non seulement de la simple possibilité mais de la fréquence ; et que la syllabe chinoise est affectée d'une ligne mélodique ou ton. Ainsi nous rencontrons plusieurs problèmes fondamentaux :

non-unicité du système phonémique : points de vue phonétique et phonologique sur l'analyse de la chaîne parlée;

explication du système des tons ; états antérieurs, transformations des postphonèmes et des tons, dont l'analyse statistique trouve des vestiges, et que le dépouillement des anciens dictionnaires phonétiques éclairerait.

structure antique de la racine dans toutes les langues...

Les principaux résultats sont que : a) le point d'articulation des préphonèmes est un facteur issu du tableau de compatibilité ; b) le ton est plus corrélé au préphonème qu'au postphonème (on eût attendu le contraire) ; c) les facteurs issus de la correspondance préphonème x ton et postphonème x ton s'interprètent bien par référence au système ancien des tons .

LC4 n° 2 [DIST. CHIN.] : Le parallélisme dans les distiques chinois : recherche inductive de la syntaxe et du sens, par H. Ko. Ces distiques sont en eux-mêmes une langue dont la structure est révélée par le parallélisme. La correspondance entre caractères de même rang (i.e.  $k(j, j')$  = nombre de fois que le caractère  $j$  se trouve dans un vers de l'un des distiques de notre corpus, en face du caractère  $j'$  figurant au même rang dans l'autre vers) permet d'abord de faire une partition du lexique (plus exactement, on s'est borné aux cent mots-caractères les plus fréquents) en classes qui ne sont autres que les parties du discours : i.e. nom, verbe, préposition... Et cette partition est construite par le seul calcul, sans reposer sur une connaissance de la langue. En poursuivant l'analyse on trouve d'une part une représentation des verbes et des noms liée au sens ; et d'autre part, d'après la succession des parties du discours on accède à la syntaxe.

LC4 n° 3 [ANA. MUS.] : L'analyse statistique des partitions de musique, par B. Morando. Plaçons-nous en un temps déterminé de l'exécution d'une partition à plusieurs voix : les notes émises simultanément par ces voix, celles qu'elles viennent d'émettre, celles qu'elles vont émettre peuvent être considérées comme des réponses à quelques questions dont chacune admet un ensemble fini de modalités qui n'est autre que l'ensemble des notes que fait entendre la voix considérée, sur l'ensemble de la partition. Ainsi une partition offre matière à construire maints tableaux de Burt (cf. LC3 n° 0) ; et selon le tableau ou le sous-tableau qu'on choisit d'analyser, divers aspects apparaissent de la mélodie ou de l'harmonie ; offrant sur un dessin les affinités, les attirances et les oppositions qui sont le propre de l'écriture musicale. On se borne ici à une seule partition - un choral de Bach ; mais l'auteur en a déjà analysé beaucoup d'autres et il attend de telles recherches après une traduction géométrique des formes, des vues neuves sur la diversité des styles et leur évolution.

LC4 n° 4 [EPINOCHÉ] : Un dialogue des gestes : les combats d'épinoches, par J.P. Benzécri, d'après B. Chauvin et Y. Grelet-Püterflam. Dans un petit aquarium, deux épinoches mâles ne tardent pas à s'affronter ; B. Chauvin a minutieusement observé ces combats, et en a décrit 40 comme des suites de gestes émis successivement par l'un ou l'autre des protagonistes, et répertoriés suivant une nomenclature (ou *éthogramme*) de 20 gestes-type. En analysant ce remarquable corpus de données, Y. Grelet considère d'abord la diversité des profils d'emplois des gestes par les antagonistes des combats ; puis la place des gestes au sein des répliques (ou séquences émises par un même animal sans intervention de son antagoniste) ; enfin l'enchaînement de ces gestes au sein d'une réplique et leur rôle de stimulus et de réponse dans le passage d'une réplique à l'autre. Par ces analyses l'éthologiste précise la valeur des gestes et en perfectionne le codage. Et le linguiste sensible à l'analogie des formes, se laissera porter à traiter de même d'autres discours.

4 Actualité de la méthode inductive en linguistique : Histoire et Pré-histoire de l'Analyse des Données, ont fait l'objet d'un exposé particulier qu'on ne reprendra pas ici. Expliquons seulement comment l'analyse des correspondances a été initialement proposée comme une méthode inductive d'analyse des données linguistiques.

Vers 1960, la traduction automatique semblait un objectif assez rapidement accessible (sont apparus depuis des obstacles qu'on n'a pas surmontés...). L'Association pour la Traduction Automatique A.T.A.L.A., fondée à l'initiative de *E. Delavenay* aidait efficacement aux rencontres des chercheurs français ; auxquels le regretté *Pr. J. Favard* ouvrit bientôt un séminaire spécialisé. La linguistique mathématique (à laquelle *Y. Lecerf* avait invité le rédacteur du présent recueil), était alors dominée par le renom de *N. Chomsky* dont le petit volume *Syntactic Structures* s'imposait à tous.

Entre autres thèses, *N. Chomsky* affirme là, qu'il ne peut exister de procédure systématique pour déterminer la grammaire d'une langue ou plus généralement les structures linguistiques, à partir d'un ensemble de données tel qu'un recueil de textes que les linguistes nomment corpus (cf *Supra* Combattre pour la linguistique). Contre cette thèse (idéaliste ; en ce qu'elle tendait à séparer le jeu de l'esprit des faits qui en sont l'inspiration et l'objet) on a voulu proposer, à défaut d'un algorithme universel pour passer de 10.000 pages de texte d'une langue à une syntaxe doublée d'une sémantique, une méthode inductive efficace pour traiter utilement des tableaux de données qu'on pouvait immédiatement recueillir ; avec à l'horizon l'ambitieux étagement des recherches successives ne laissant rien dans l'ombre des formes, du sens et du style.

En 1973 se tint à Grenoble dans le cadre accueillant du laboratoire de *B. Vauquois*, un premier colloque consacré à l'analyse des données linguistiques. Du rapport de ce colloque est issu le texte [ANA. LING.] LA n° 1, qui passe en revue les analyses de correspondances effectuées dans ce domaine de 1963 à 1974. Après une journée à Saint-Cloud dans le Centre dirigé par *L. Wagner*, on eut à Montpellier, en 1976, un deuxième colloque : sous un aréopage de maîtres bienveillants, le zèle de deux néophytes - *F. Gardes-Madray* et *Ch. Tronc* - fit le succès de cette rencontre.

Les travaux des chercheurs de Nancy, Saint-Cloud, Vincennes, Montpellier, Besançon..., commencent à porter leur fruits. N'était-il pas temps de paraître ?

Et c'est ainsi qu'encouragés par une subvention du Centre National de la Recherche Scientifique (auprès duquel *B. Vauquois* se fit notre avocat), les éditions Dunod publient le présent volume ; oeuvre collective de plus de trente auteurs dont certains tel *M. Clay* ont relu et revu de grandes parties de l'ouvrage ; sans oublier Madame *O. Laraise* qui aidée de plusieurs collaboratrices occasionnelles, en a assuré la composition au Laboratoire de Statistique de l'Université P. et M. Curie ; ni Madame *Ch. Perrichon* qui sait seule entre combien de mains chaque page doit passer avant qu'elle la remette à l'éditeur ; lequel a pour nous le visage amical d'*A. Cardona*.