LES CAHIERS DE L'ANALYSE DES DONNÉES

J.-P. BENZÉCRI
F. BENZÉCRI

Analyse du vocabulaire et recherche du thème dans les articles des volumes XII à XVII de CAD. (1) Le corpus et les résumés

Les cahiers de l'analyse des données, tome 18, n° 1 (1993), p. 47-60

<http://www.numdam.org/item?id=CAD_1993__18_1_47_0>

Les Cahiers de l'Analyse des Données
Vol XVIII - 1993 - n°1 - pp. 47-60.

ANALYSE DU VOCABULAIRE ET RECHERCHE DU THÈME DANS LES ARTICLES DES VOLUMES XII À XVII DE *CAD* (1) LE CORPUS ET LES RÉSUMÉS

[*CAD* XII-XVII (1)]

J.-P. & F. BENZÉCRI

0 Les étapes du présent exposé

Depuis le début de 1987, notre revue est composée sur un traitement de texte: ainsi, on dispose présentement du texte enregistré des six volumes XII à XVII, soit 24 cahiers comprenant, au total, quelque 200 articles. Ce corpus paraît offrir une matière suffisante pour des expériences d'analyse du vocabulaire.

Plusieurs études ont déjà paru portant, notamment, sur le texte grec du Nouveau Testament, puis sur des anthologies de textes littéraires en latin ou en espagnol. Dans ces études, l'agrégation des chapitres en classes peut être appréciée d'après la composition propre aux textes analysés et leur répartition admise en genres et en périodes. Comme l'étude du style passe avant celle du thème, on analyse d'abord la distribution des formes des mots outil.

Avec les articles de *CAD*, le point de vue change: ceux-ci ne sont pas répartis a priori en œuvres ayant chacune un genre et un thème propre; au contraire, le but des analyses de vocabulaire est d'élaborer ces notions plutôt que de retrouver la place de chaque article dans des catégories bien connues. Cependant, sans avoir des textes une vue d'ensemble préalable, on ne peut apprécier la pertinence des constructions multidimensionnelles: c'est pourquoi, dès le §1, on décrit minutieusement notre corpus.

Puis, à titre préliminaire, on analyse, au §2, les résumés en langues française et anglaise des articles. Les résumés en arabe n'ont pas été traités car notre choix n'est pas fait quant au découpage des formes (séparation éventuelle de l'article défini, de certaines prépositions, etc...).

On sait que, même dans la langue littéraire, il n'y a pas de limite nette entre mots pleins et mots outil; des locutions plus ou moins figées assumant volontiers les fonctions de ceux-ci. Dans la science et la technique, intervient, d'autre part, la terminologie spécialisée; qui est, nécessairement, à la base de

◊fbr [DISC.MORPH.CELL.]:XII-1: Discrimination morphologique de cultures de fibroblastes ...
§rét [NEO-VAISSEAUX]:XII-1: Facteurs de risque de l'apparition de néo-vais. sous-rétiniens...
£prl [PAROLE CONT.]:XII 1: Analyse factorielle de la parole continue: étude comparative...
@Src [FLANADES]:XII 1: Les Flan... enquête d'opinion ... un centre commerc... à Sarcelles
ßcib [VAL.MOB.]:XII-1: ... constitut. d'une cible pour l'identification des valeurs mobilières ...
@ngr [ENGRENAGES]:XII-1: L'applic. en t. réel de la discriminat. à des données industriellles

πpop [PROG. DISC. ALEAT.]:XII-2: Progr. pour la discrim. bary... étude par permut. aléatoire
$Lé1 [LEONTIEF]:XII 2: Décomposition d'une matrice de Léontief par l'analyse des corresp...
©Pét [4 PROD. PETROL.]:XII 2: ... consom. mensuelle par département de 4 produits pétr...
µcah [COMPLEXITE CAH]:XII2: Sur la complexité des algorithmes de CAH
µvoi [CAH. VOIS. CELL.]:XII2: ...rech. des plus proches voisins ... décomp cellul. de l'espace
§sST [INFARCTUS ST]:XII2: sous décalage du segment ST à la phase aiguë de l'Infarctus...
@man [MANUTENTION]:XII2: Implantation physique de testeurs dans l'atelier de validation
µVAC [CAH VAR CUM]:XII2: Qual. de la repr. d'une CAH sur un tab. cumulant des blocs...

µinh [INTERACTION INHIBITION]:XII3: Sur un modele math. d'interaction par inhibition
®rch [RECHERCHE C.E.E.]:XII3: ...la recherche dans ... la communauté économ. européenne
$Sé1 [SÉRIE MONNAIE FRANCE]:XII3: séries chron. décalées ... mon . France 1910 ... 1945
◊Oxy [OXYURES, PLATHYRRINIENS]:XII3: oxyures marq. évol. des primates platyrrhiniens
@stg [STAGE POST. RECR.]:XII3: attachés communaux: opinion des stagiaires sur la formation
µrec [REC. COR. BLOCS]:XII3: reconstitution d'un tabl corresp à partir du tabl cumulé par blocs
µSim [REPR. SIMPLEXE]:XII3: représentat. graphique de J par le nuage N(J) et le simplexe SJ

◊Loi [ESTUAIRE LOIRE]:XII4: évolut spatio-temp. meio et mixofaunes dans l'est. de la Loire
@prf [ORDRE PAIRE]:XII4: Classification par ordre et comparaisons en paires
πµdé [MODELE CHRON. DECAL.]:XII4: séries chron. décalées interprétation sur cas modèles
®Arb [CLIMAT ARABIE]:XII4: L'image climatique des mois et saisons de l'Arabie Séoudite
®§Qb [MALADIE QUÉBEC]:XII4: Ana spatio-temp par postes de l'assurance malad. du Québec
$Sé2 [SÉRIE MON. FR. 70 84]:XII4: séries chron. décalées ... hist mon .France 1970 ... 1984
ßvoi [VAL. MOB. 2]:XII4: Analyse du voisinage des valeurs mobilieres

Tableau *des titres des articles du* Volume XII (1987)

recherches documentaires comme les nôtres. On dira donc, au §3, suivant quels principes ont été choisis divers lexiques pour analyser les textes complets des articles.

Des recherches antérieures (cf., notamment, A. Aït HAMLAT, [IND. DOC.], in *CAD*, Vol. IX, n°2, 1984) ont appelé l'attention sur l'importance de la répartition en grappe des occurrences, pour la segmentation du lexique et l'interprétation des dénombrements: d'où l'objet du §4 qui complète le §3.

Les résultats principaux, relatifs à la typologie des articles et de leur vocabulaire, sont donnés au §5; puis, au §6, on utilise l'analyse discriminante pour chercher, dans un corpus, l'article le plus proche d'un nouvel article, extérieur au corpus. Du fait de sa longueur, le présent exposé a été fractionné en trois articles, comprenant respectivement les §§ 1-2, 3-4 et 5-6: dans les §§3-4, les références renvoient au §1; dans les §§5-6, elles revoient , principalement, aux §§1 et 3.

1 Description du corpus des articles

Tout lecteur pourra juger par lui-même de l'intérêt des classes de mots qui résulteront des analyses factorielles et classifications; mais pour apprécier une

£hxL [HEXA. DACTYL. LATIN]:XIII1: composition métrique de l'hexamètre dactylique latin
£frL [FRÉQ. CAT. LATIN]:XIII1: ana. de la fréquence des catégories grammaticales en Latin
£Mtl [MÉT. ARISTOTE]:XIII1: chap. de la métaphysique d'A: fréqu des parties du discours
£Arb [STYL. ARAB.]:XIII1: essais pour une stylométrie appliquée aux textes arabes
£Chn [TYP. CHIN.]:XIII1: Essai de typologie des Écritures manuscrites chinoises
£Gns [GENÈSE]:XIII1: Stylométrie et théorie des sources pour les versets du livre de la genèse
£etr [LETTRES]:XIII1: Reconnaissance de l'auteur d'un texte d'après les caractères utilisés
£sty [PROGRAMME]:XIII1: Programme de recherches en stylomÉtrie
£vox [SPECTR. STAT. VOIX]:XIII1: Analyse spectrale et analyse statistique de la voix parlée
ƒqal [QUAL. QUANT.]:XIII1: Qualité et quantité dans la trad des philosophes et en a. des donn

@Chr [CHRON . INSERT. PROF.]:XIII2: chronologie de l'insertion professionnelle
@Erg [ERGO. RÉGLAGES]:XIII2: questionnaire en ergonomie: réglages d'un poste de travail
μSta [STABILITÉ]:XIII2: stabilité des sous-esp princip d'inertie sous changement de métrique
πVPr [ALG. V. V. PROPRES]:XIII12: algorithme de calcul des valeurs pr et vect pr en a des corr
@πrc [CODE MICROPROC.]:XIII2: Identification du microprocess destinataire d'un code-objet
@pGr [POLIT. GREC.]:XIII2: thermomètre de sympathie vis-à-vis de personnalités polit grecques
@Pré [SOND. PRÉS.]:XIII2: image des candidats à la présidence auprès de leurs partisans
§EsL [ESS. LIBRE]:XIII3: Essais thérapeutiques ouverts en libre choix: notes de lecture
@∫sd [SOND. SIDA]:XIII2: Attitudes des adultes de 11 pays vis à-vis des malades du SIDA

ƒArj [ACT. ARISTOTE]:XIII3: Actualité de la pensÉe d'Aristote
ƒAÉc [ÉCO. ARIST.]:XIII3: L'Économie chez Aristote
ƒPrê [ARIST. PRÊT]:XIII3: Aristote et le prêt à intéret
ƒusu [USURE XII-XIV]:XIII3: doctrine de l'usure et de l'intérêt du XII-ème au XIV ème siècle
ƒLgS [LOG. STAT.]:XIII3: d'Aristote à l'A des D: logique ontologie dans l'induction statistique
ßMrt [MARTELL INTERNATION.]:XIII3: compar internat entre ind de places et c d'un titre: Martell
©Pnt[CARTE VIE FRANCE]:XIII3: où vit-on le mieux en France? carte de l'a de 33 variables
§$Nr [CONSOM. PHAR.]:XIII3: consom produits pharm prescrits ou non dans le Nord-P.deC.
©1Pr [PREMIERS TOURS 88,81]:XIII3:cart par dép des votes aux élections présidentielles

∫tk1 [STADES INF. VIH]:XIII4: l'antigénémie dans la défin des stades de l'infection par le VIH
μΔel [TRIANG. DELAUNAY]:XIII4: Géométrie anallagmatique et triangulation de Delaunay
$@Bq [IMAGE BANQUES]:XIII4: Image des banques auprès des grandes entreprises
μbl1 [REC. BLOC. CAH]:XIII4: Reconnaiss de la structure de blocs d'un tab de corr par la CAH
$Arb [MARCH. COMM. ARABE]:XIII4: comm int et ext du marché commun arabe: 1972-1982
@μCt [POND. CONTRI.]:XIII4: pondération des contrib.des modalités des var: ex en Écologie
§hm1 [VAR. HÉMO.]:XIII4: variations de l'état hémodynamique et du taux sérique d'un produit
ßpôl [MOUV. VAL. PÔLE]:XIII4: Mouvement des valeurs d'une cible relativem. à un titre-pôle
μdéc [CORRÉL. CHRON. DÉCAL.]:XIII4: corrél entre séries chronolog: méth des séries décal

Tableau *des titres des articles du* Volume XIII (1988)

typologie des documents, il faut avoir présent à l'esprit le genre et le contenu de ceux-ci. Certes les articles eux-mêmes pourront être consultés dans certains cas, mais il s'impose de caractériser avec concision chaque article et de faire choix de sigles rendant compte explicitement des caractères retenus, et évoquant, autant que possible, l'article individuel lui-même (au moins pour un lecteur qui connaît celui-ci). Assurément, le choix des sigles ne procède pas de principes scientifiques originaux: mais la lecture des résultats sera facile ou quasi impossible selon que le lecteur saisira, ou non, dans les sigles, les caractères des articles que ceux-ci représentent.

Quant au genre, les articles peuvent relever des mathématiques, de la programmation, des études de cas, voire de la philosophie et de la méthodologie

πCor [NOT. CORR. CAH]:XIV1: Programmes d'a. des correspondances et de CAH: Notice
πCré [NOT. CRÉ. TAB.]:XIV1: Programmes de création de tableaux: Notice d'utilisation
π©rt [NOT. PROG. CART.]:XIV1: Prog de cartographie ... d'une analyse multidim: Notice
£mot [NOTES MOTS]:XIV1: Analyse des notes attribuées par des sujets aux mots d'une liste
£Vox [FOND. REC. PAROLE]:XIV1: fondements scientif pour la reconnaissance de la parole
®Vin [EXPORT. VIN]:XIV1: Exportations françaises de vins par crus et pays destinataires
®Imm [PRIX IMMOB.]:XIV1: prix de l'immobilier neuf d'habitation en France depuis 1970

ßHKg [BOURSE H. K.]:XIV2: Bourse de Hong Kong: cours en févr-mars 87: séries décalées
©§SF [ÉQUITÉ SÉCU.]:XIV2: disparité géograph et équité du syst franç de sécurité sociale
©Cal [CALÉDONIEN]:XIV2: référendum caléd & 1-ers t des présidentielles de 88 et 81 : cartes
§d&d [ASS. ANTALGIQUE]:XIV2: comp de l'efficacité de deux antalgiques et de l'association
µEnt [ENTROPIE]:XIV2: Croissance de l'entropie et réversibilitÉ des lois de la mécanique
£exp [EXP. SIDA]:XIV2: Modèle exponentiel et épidémiologie du SIDA
@str [ASTROLABE]:XIV2: observateurs et observations: mesures à l'observatoire de Paris
πlin [CODAGE LIN.]:XIV2: codage linéaire par morceaux: réalisation et applications
©Grc [AGRI. GREC.]:XIV2: agriculture grecque: étude chron et régionale des cultures 1970 1981
$cdr [RECRUT. CADRES]:XIV2: recrut de cadres en France par branche et formation en 1988
©pub [CARTE AGENTS]:XIV2: carte des performances par départ des agents d'une entreprise

µbl2 [REC. BLOC. II]:XIV3: Reconnaiss de la structure de blocs d'un tab de corr par la CAH
©vgn [VIGNETTES]:XIV3: carte des départements français d'après les vignettes sur les voitures
§éhX [ÉCHELLES]:XIV3: Échelles d'anxiété et plaintes somatiques dans les essais thérapeutiques
§hm2 [VAR. HÉMO.]:XIV3: variations de l'état hémodynamique et du taux sérique d'un produit

πprs [ÉQ. PERS.]:XIV3: Codage linéaire par morceaux et Équation personnelle
@rch [QUEST. MÉM. RECH.]:XIV3: quest sur le mémoire de recherche en admin des affaires
µméd [MÉDIANE]:XIV3: médiane généralisée d'une distribution de masse multidimensionnelle
@chê [CHÊNE]:XIV3: Critères pour l'aspect des placages de bois de chêne
µblð [REC. BLOC. IIbis]:XIV27: Rec. de la structure de blocs d'un tab de corr par la CAH: ex.

◊vir [VIRUS]:XIV4: typologie de virus de plantes d'après leur protéine de capside
©Par [IMMOB. PROV. PARIS]:XIV4: marché immob de 50 villes de province et locations à Paris
§ECG [ECG SPORTIFS]:XIV4: ECG numérisé chez des sujets de 20 ans sportifs ou sédentaires
$usi [USINES FRANCE]:XIV4: hommes espace dépenses pour les premières usines de France
§eff [COMPAR. EFFORT]:XIV4: comparaison des sujets après l'effort sous divers traitements
©stt [STAT. CART.]:XIV4: état de la cartographie automatique des données statistiques
§hm3 [COMPAR. CARD.]:XIV4: prod et placebo chez insuff cardiaques résistants au trait usuel

Tableau *des titres des articles du* Volume XIV (1989)

générale. Les thèmes les plus fréquemment rencontrés sont la médecine, l'économie, la linguistique, la représentation cartographique de données ventilées par départements...

Le premier caratère du sigle évoque le genre ou le thème suivant un code mnémonique que nous expliquons sur des exemples:

µ: Mathématique ou Physique: ex. µSta, stabilité des sous-espaces principaux; µRel, relativisation du continuum spatio-temporel;...

π: Programmation: ex. πCum, programme d'extension d'un tableau par des cumuls de lignes et colonnes;...

f: Philosophie: ex. *f*qal, qualité et quantité en philosophie et en a. des données; *f*Ord, l'ordinateur au service du logicien;...

£Dia [DIAFWTISMOS]:XV1: prologues des livres grecs de sciences édités de 1730 à 1820
£Aur [CLASS. CORPUS]:XV1: méthode de classification des énoncés d'un corpus et application
£Mt2 [MÉTAPHYSIQUE 2]:XV1: Méta d'Aristote et Théophraste: empl des parties du discours
£Tr1 [LING. TRI]: XV1: progr de statistique linguistique et tri par fusion des mots du texte
£NT1 [NOUV. TEST. GREC.]:XV1: chap du texte grec du Nouveau Testament d'après mots outil
πCum [CUM. LI. COL.]:XV1: extension d'un tableau par des cumuls de lignes et colonnes
£Par [PARMÉNIDE]: XV1: le Parménide de Platon : répartition des vocables

µfus [ASS. FUS. TRI.]:XV2: associativité de la fusion et parallélisme dans les algorithmes de tri
πPCc [COMP. MICRO.]:XV2: performances comparées de 64 micro-ordinateurs
§gel [DOUL. SPORT.]:XV112: un gel pour les douleurs, contractures et œdèmes du sportif
§Imp [IMPATIENCE]:XV2: La révolte des patients
§2sé [DEUX SÉDATIFS]:XV2: efficacité de 2 sédatifs: troubles du sommeil et états nerveux
©Elz [POLITIQUE ITALIE]:XV2: Italie, géographie politique d'après 4 scrutins de 1983 à 1989
µRel [REL. CONT. SPAT.]:XV2: moment et position: relativisation du continuum spatio-temporel
ßJap [BOURSE TOKYO]:XV2: cours de 19 titres àTokyo du 30.12.1986 au 30.3.1990
©Ent [FLUX INTERDÉP.]:XV2: flux interdépartementaux de prestations d'une entreprise

∫tki [RECH. PRON. SIDA]:XV3: facteurs pronostiques dans la pathogénèse du SIDA
∫tk2 [STADES VIH]:XV3: estimation du stade de l'infection par le VIH chez les séro positifs
∫Imm [IMM. CLIN. VIH]:XV3: État du syst immun. et hist clinique des patients infectés par VIH
∫ana [ANA. CLIN.]:XV3: Analyse des données biologiques et pathologie clinique
∫∂ia[DISCUSSION PASTEUR]:XV3: informaticiens et spéc. du SIDA, après exposé d'A des D
ƒOrd [ORD. LOG.]:XV3: L'ordinateur: un outil au service du logicien
ƒOmn[OMNISCIENCE.]:XV3: limites de l'omniscience: théologie et intelligence artificielle
ƒ∂ia [MATH. LOG. OMN.]:XV3: connaissance mathém., c. logique, omniscience: un dialogue
ßFrn [BOURSE PARIS]:XV3: cours de 26 titres à Paris de Déc 1986 à Juin 1990
πInd [MacSAIF]: XV3: indice des programmes du logiciel 'MacSAIF': descr. de leurs fonctions

µCv1 [MÉD. CONV.]:XV4: médiane et c de g pour une distrib de densité constante sur un convexe
µCv2 [MÉD. DIS.]:XV4: médiane et c de g pour une distrib de densité convexe sur un convexe
µQs1 [MOD. VAR. CLASS.]:XV4: modèle d'un ensemble redondant de var découpées en classes
£HxG[HEXA. DACTYL. GREC]:XV4: composition métrique de l'hexamètre dactylique grec
§HLT [PRESS. ART. NYCTH.]:XV4: trait antihypert et enregistr de la press artér du nycthémère
®§Vi [ESP. VIE MONDE]:XV4: espérance de vie, équipement sanitaire, nutrition dans le monde
§Sev [SEVRAGE HYPN. ANX.]:XV4: sevrage trait hypnotique ou anxiolyt et thérap substitutive
§MzT [RÉAC. THÉRAP.]:XV4: réactions de 2000 sujets à une thérapeutique
®Exm [RECR. MATH.]:XV4: stat par centres des concours de recrut des prof de math en 1989

Tableau *des titres des articles du* Volume XV (1990)

£: Linguistique: ex. £prl, analyse factorielle de la parole; £etr, reconnaissance de l'auteur d'un texte d'après les caractères utilisés; £ñOr, typologie de textes espagnols du siècle d'or;...

$: Économie: ex. $Lé1, décomposition d'une matrice de Léontief (1-er article); $Sé2, séries chronologiques de l'histoire monétaire (2-ème article);...

ß: Bourse (subdivision de $): ex. ßHKg, la bourse de Hong Kong en février-mars 1987; ßFrn, cours de 26 titres à Paris de Déc 1986 à Juin 1990;...

§: Médecine (§ à l'imitation du caducée): §hm1, variations de l'état hémodynamique et du taux sérique d'un produit (1-er article); §2sé, efficacité de 2 sédatifs;...

§$âg [PHARM. AGE SEXE]:XVI1: consomm des classes de prod pharmac selon âge et sexe
µbar [MOD. CODE BARY.]:XVI1: modèle d'une var continue unique codée barycentriquement
®Cao [CACAO]:XVI1: marché mondial du cacao de 1976/1977 à 1986/1987
@Mor [QUEST. PRÉ-MORT]:XVI1: questionnaire sur l'expérience de pré-mort
£NT2 [TEXTES GRECS]:XVI1: typologie de textes grecs d'après les occurrences des mots outil
µbrd [BORD BOULES]:XVI1: bord et volume pour une réunion de boules de même rayon
ßSIC[SICAV]:XVI1: valeur liquidative des SICAV et indice économique
§$pr [PRAT. FAIBLE REC.]:XVI1: omnipraticiens à faibles recettes en 1989
§gên [COMP. VAR.]:XVI1: variation comparée sous deux traitements de la gêne des patients
@Goo [MOD. GRAPH.]:XVI1: mémoire reçu: measures, models and graphical displays

£Tr2 [LING. TRI 2]:XVI2: prog de stat ling et application au contenu de textes bibliques en grec
@Ths [THESS. PIER. POL.]:XVI2: outils tranchants thessaliens en pierre polie: typologie
§ADN [HÉMATOPROTECT.]:XVI2: effet hématopr de l'ADN-HP et chimiothér anticancéreuse
µGut [MOD. DÉC. VAR.]:XVI2: modèle d'un ensemble de découpages d'une variable unique
µPrt [MOD. CLASS. PART.]:XVI2: mod gén de classes et partitions par déc d'une var unique
§@Br [QUEST. BORTNER]:XVI2: questionnaire de bortner: réponses de 97 cardiaques
µBl3 [REC. BLOC. III]:XVI2: Reconnaiss de la structure de blocs d'un tab de corr par la CAH
©Alg [ÉLEC. ALGÉRIE]:XVI2: Carte de pronostics électoraux pour l'Algérie

©$Tr [PRIX TERRES]:XVI3: variation du prix des terres agricoles en France par dép: 1972 89
®Caf [CAFÉ]:XVI3: Marché mondial du café de 1983/1984 à 1988/1989
@Jrd [ENQ. SCOL. JORD.]:XVI3: Enquête scolaire effectuée en Jordanie
µprs [DOUB. REC. PERS.]:XVI3: modèle de double recadrage des notes suiv équ personnelle
@Alg [QUEST. ALGÉRIE]:XVI3: questionnaire sur l'avenir politique de l'Algérie
π©dg [CARTE DIGIT.]:XVI3: utilisation de fonds de carte digitalisés créés par divers logiciels
©Mrt [MARTINIQUE]:XVI3: carte du recensement de 1982 à la Martinique
©Mrf [MARTINIQUE FLUX]:XVI3: carte des flux de population à la Mart de 1975 à1982
©Chô [CHÔMAGE 1980 1989]:XVI3: chômage en France par départements de 1980 à 1989

§Sev [COMP. SEVRAGE]:XVI4: comparaison entre 4 méthodes de sevrage après anxiolytique
$Lé2 [LÉONTIEF GRÈCE]:XVI4: matrice de Léontief de la Grèce: 1958-1977
$Esp [BANQUE ESP.]:XVI4: secteur bancaire espagnol et secteurs bancaires de la C.E.E.
µMlt [SOM. COD. MULT.]:XVI4: som directe d'ens ordonnés et codages mult d'une var unique
@Hst [MOD. HISTOG.]:XVI4: inadéquation de la descript d'un histogram par moy et écart type
£Lat [TEXTES LATINS]:XVI4: typologie de textes latins d'après les occurrences de mots outil
§Tnd [TENDINITES]:XVI4: efficacité comparée de deux gels dans les tendinites superficielles
@Prd [MÉTH. PRÉD.]:XVI4: méthodologie de la régression et de la prédiction fondée sur la CAH

Tableau *des titres des articles du* Volume XVI (1991)

∫: SIDA (subdivision de §): ∫Imm, État du syst immununitaire et histoire clinique des patients infectés par VIH;...

®: données par Régions ou par pays: ®Cao, marché mondial du cacao de 1976/1977 à 1986/1987; ®étu, Étudiants à Paris VI par cycle d'étude et nationalité: 1975-90;...

©: Cartographie (subdivision de ®): ©Alg, Carte de pronostics électoraux pour l'Algérie; ©Sal, salaires du secteur privé dans les dép français: 1976-87;...

@: Enquêtes, Sociologie, Psychologie, Ergonomie, descriptions diverses: @Jrd, Enquête scolaire effectuée en Jordanie; @&Ps, validité des échelles en psychologie et psychiatrie et corrélations psychosociales;...

$for [COMPT. FORM.]:XVII1: salaire et formation de 1973 à 1988 en France, dans 34 secteurs
@&Ps [VALID. PSY.]:XVII1: validité des échelles en psychologie et psychiatrie et corrélations psychosociales
@&ps[NOTE VALID. PSY.]:XVII1: pratique de la validation des éch en psychopharmacologie
µnqT [TRAC. MANQ.]:XVII1: critère de la trace minima pour les données manquantes
µEPR [CORRÉL. QUANT.]:XVII1: corrélat à distance entre phén simultanés en méca quantique
£cry [DÉCRYPTAGE GREC]:XVII1: décr d'un fichier en grec moderne d'après la distrib des carac
§not [NOTE MÉD.]:XVII1: à travers la presse médicale

£SMF [S.M.F. – LIOUVILLE]:XVII2: vocab des titres de 2 périodiques mathématiques en 100ans
$@ct [LOISIRS ET PATRIMOINE]:XVII2: analyse des pourcentages calculés d'après 2 enquêtes
$Cmc [COMECON]:XVII2: commerce extérieur des pays du COMECON: 1950-1984
µInM [INF. MANQ.]:XVII2: trace min et information mutuelle pour un tab où manque une donnée
µInB [INF. BLOCS.]:XVII2: trace et information mutuelle pour un tableau décomposé en blocs
©Sal [SALAIRES DÉP.]:XVII2: salaires du secteur privé dans les dép français: 1976-87
@µCh [POLYCHROME]:XVII2: images polychromes et sensibilité au contraste chromatique

©µar [TRAC. MANQ. FLUX]:XVII3: estim de la diagonale d'après la trace min: flux à la Martinique
®com [COMMERCE 65-89]:XVII3: commerce par types de marchandises pour 95 pays: 1965-89
®étu [ÉTUDIANTS P&M]:XVII3: Étudiants à Paris VI par cycle d'étude et nationalité: 1975 90
§&nx [ÉCH. ANX. QUOT.]:XVII3: rép de 98 sujets à une échelle d'anx présentée quotidiennement
£Rep [REPAS IDÉAL]:XVII3: le repas idéal: analyse de réponses libres en trois langues
£ibr [RÉP. LIBRE]:XVII3: sur l'analyse des réponses libres dans une enquête internationale
@fum [FUMEROLLES]:XVII3: composition chimique des fumerolles et origine géologique
®Thr [THERMAL]:XVII3: analyse d'un tableau décrivant 74 stations thermales françaises
@µch [NOTE CHROM.]:XVII3: Note sur la sensibilité au contraste chromatique
§Pra présentation du livre:XVII3: pratique de l'analyse des données en médecine, pharmacologie...

$cmp [COMPL. SALAIRE]:XVII4: compléments du salaire, en 1984, en France, dans 31 secteurs
µniv [DOUBL. DIFF. VAR.]:XVII4: rép de sujets, de niveau différent, à des quest de difficulté variable
§&DA [4 ÉCH. DÉPR. ANX.]:XVII4: réponses hebdomadaires à 4 Échelles de dépress et d'anxiété
£ñOr [SIÈCLE D'OR]:XVII4: typologie de textes espagnols du siècle d'or d'après les mots outil
µnq0 [TRAC. VAL. NUL.]:XVII4: trace minima pour un tableau où certaines valeurs sont nulles
£©F©[TITR. C.F.C.]:XVII4: vocabulaire des titres des articles du Com. Français de Cartographie

Tableau *des titres des articles du* Volume XVII (1992)

◊: Sciences naturelles (dans peu de cas): ◊vir, typologie de virus de plantes d'après leur protéine de capside;

Certains articles participent à la fois de plusieurs genres dont les symboles sont alors réunis dans leur sigle; voici quelques cas:

π©rt, Programme de représentation cartographique des résultats d'une analyse multidimensionnelle: notice d'utilisation;

§$Nr, Consommation de produits pharmaceutiques prescrits ou non, dans la région Nord-Pas-de-Calais;

@∫sd, Attitudes des adultes de 11 pays vis-à-vis des malades du SIDA.

Nous publions une table complète des sigles, chacun accompagné du titre de l'article, abrégé, au besoin, pour tenir dans une ligne. Il faut toutefois prendre garde qu'un titre peut indiquer le thème (e.g. "économie") sans faire

connaître le genre; notamment, le niveau mathématique, ou, s'il s'agit de programmes, la présence d'algorithmes...

Reste à préciser qu'avant toute élaboration des textes, on a ôté formules, légendes de graphiques et tableaux, et bibliographie finale. Assurément, les signes, les symboles, les mots techniques des titres cités, indiquent mieux que toute autre partie du texte le thème et le genre; mais nous avons choisi de borner la présente étude à la forme générale d'expression qui se rencontre dans tous les textes, qu'il s'agisse ou non d'articles scientifiques ou techniques.

2 Analyse des sommaires, en deux langues

2.1 Des résumés aux analyses

On a retenu les résumés, en langues anglaise et française, d'un ensemble J de 201 articles parus dans les volumes XII-XVII: ont seulement été éliminés quelques résumés réduits à un titre. À partir d'un lexique f, ou ensemble de formes de mots français, on construit, comme d'usage, d'après le texte des résumés en français, un tableau de correspondance $f \times J$; On procède de même avec un lexique de mots anglais, &, et les résumés traduits en anglais, d'où un tableau $\& \times J$. On peut effectuer des analyses factorielles et classifications séparément, d'après chacun des tableaux $f \times J$ ou $\& \times J$; mais on peut aussi considérer le tableau $(f \cup \&) \times J$, obtenu en superposant les tableaux $f \times J$ et $\& \times J$: la CAH de l'ensemble $(f \cup \&)$ fournit alors des classes mixtes comportant des formes des deux langues, les mots de même sens (ou de même fonction) s'agrégeant, généralement, à un niveau très bas.

Une première expérience semblable, due à Ch. ARBACHE, est rapportée dans [TEXTE BIBLE] §9, in CAD, Vol XI, n°1; où l'on traite parallèlement les textes, découpés en versets, de deux chapitres de l'Ancien Testament, considérés dans l'original hébraïque et dans la version grecque des Septantes.

Afin de rendre compte de quelques uns des nombreux traitements effectués ici, nous publions un graphique plan et plusieurs CAH.

2.2 Croisement des résumés avec la réunion de deux lexiques, anglais et français, comprenant chacun environ 40 mots outil

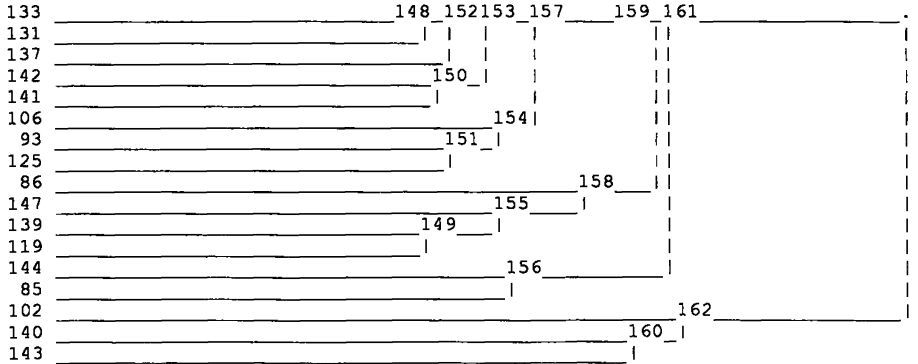
Pour chacune des deux langues, on a extrait, du corpus des résumés, les mots outil dont la fréquence dépasse 18 (donc sans veiller à ce que, si un mot est pris dans une langue, sa traduction le soit dans l'autre); soit:

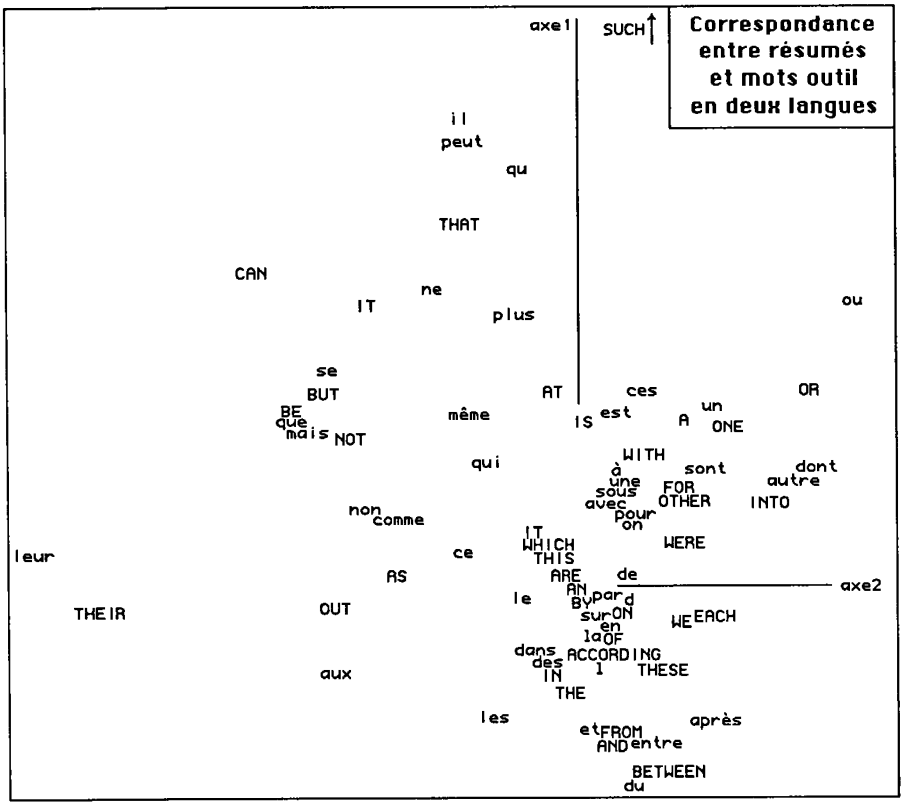
f : lexique de 45 mots français
& : lexique de 37 mots anglais

et on a analysé le tableau $(f \cup \&) \times J$. Afin d'éviter toute confusion, les mots anglais sont écrits en capitales; et les mots français en minuscules.

```
 c | Partition en  17 classes : mots outil de la classe c
---+-------------------------------------------------------------------------
133| (a les) (AND et)
131| (ACCORDING (FROM après)) du
137| (dans IN) (INTO en)
---+-------------------------------------------------------------------------
142| ((THE OF) d) ((de l) (AN ((des TO) à)))
141| (OUT AT) ((BY  par) la)
---+-------------------------------------------------------------------------
106| (comme AS)
 93| (sur ON)
125| (aux (WITH avec))
---+-------------------------------------------------------------------------
 86| (entre BETWEEN)
147| (EACH(même(THIS ce))) (sous le) ((qui WHICH)(IS est)) (dont(sont ARE))
139| (on WE) (pour FOR)
119| (un (A une))
---+-------------------------------------------------------------------------
144| (WERE (THESE ces)) ((ONE au) (OTHER autre))
 85| (OR ou)
---+-------------------------------------------------------------------------
102| (leur THEIR)
140| (ne NOT) (non (mais BUT))
143| (se ((que (il IT)) plus)) (BE (peut CAN)) (SUCH (qu THAT))
---+-------------------------------------------------------------------------
```

Il n'y a rien à attendre ici d'une classification des articles d'après le style de leurs résumés (observé sur le profil des mots outil). On se bornera donc à la CAH des formes; avec le nuage de celles-ci dans le plan (1, 2) issu de l'analyse des correspondances. Sans tracer l'arbre complet, on a, sur une partition en 17 classes, marqué, par des parenthèses les agrégations depuis le plus bas niveau. L'ajustement des deux lexiques est généralement bon. Des irrégularités valent d'être observées. Dans (ACCORDING (FROM après)) du), on reconnaît que le français "après" se trouve deux fois sur trois dans la locution "d'après" qui peut, selon le contexte se traduire par "FROM" ou "ACCORDING TO"; l'agrégat (SUCH (qu THAT)) s'explique parce que les formes de "tel" (tel, telle, tels, telles) n'atteignent pas le seuil de fréquence: on a donc "SUCH THAT" sans "tel qu(e,')..."; etc.






```
45 mots françaisf + 37 mots anglais& = 82 outils x 201 résumés bilingues;
trace : 1.405e+0
rang   :   1     2     3     4     5 ... 10  .  15  .  25  .  50 ... 75
lambda: 733   597   544   520   477   381   311   214    88    22 e-4
taux   : 522   425   387   370   339   272   222   152    63    16 e-4
cumul  : 522   947  1334  1704  2043  3501  4692  6505  9026  9929 e-4
```

On retrouve, sur le plan (1, 2), l'ajustement des deux lexiques de mots outil, montré, dans tous ses détails, par la CAH.

2.3 Croisement des résumés avec la réunion de deux lexiques, anglais et français, comprenant chacun environ cent mots pleins

Pour chacune des deux langues, on a extrait, du corpus des résumés, les mots pleins dont la fréquence est au moins égale à 8; soit:

π : lexique de 99 mots français
$\prod$: lexique de 94 mots anglais

puis on a analysé le tableau $(\pi \cup \prod) \times J$; et, après analyse factorielle, on a classé les résumés et les formes de mots.

Dans la partition des articles que nous avons retenue, la plupart des classes sont interprétables; et, une fois cette interprétation adoptée, on peut dire (tenant compte à la fois des classes non inteprétées et des affectations aberrantes aux autres classes) que les deux tiers des documents sont rangés de façon satisfaisante.

c	Partition : Sigles des articles de la classe numéro c
373	Ofbr £Gns µfus ©Elz ßJap ƒOrd Ovir ∫tk1 ∫Imm ∫tk2
375	@µch $gên $rét $hm2 $ECG $Se4 $d&d $sST $MzT $HLT $SeH $2sé $Imp $Tnd
	$ADN $hm3 $xrc analyse d'essais cliniques
303	ƒAÉc ƒPrê ƒAri philosophie en économie
370	£Mt2 £Arb £NT2 £NT1 £Lat £ñOr analyse des textes
330	∫tki ∫∂ia ∫ana ∫@XI ∫atu ∫exp sur le ∫ida
315	ßHKg $Sé2 $Sé1 πµdé µdéc analyse des séries décalées
359	@stg $cmp $for stagiaires et salariés
325	©Pét $$Nr $$âg consommation: du pétrole, des médicaments
345	µnq0 ©µar µnqT µInM estimation de données manquantes
385	πlin @&ps @pGr @πrc πprs @&Ps µniv @chê @Erg @Alg @rch @Mor $@Br $&nx
	questionnaires avec échelles de réponse
368	®Cao ®Caf π©dg ©Ent ©Mrƒ international et cartes
361	£HxG £cry ƒOmn £etr πpop πCum £Tr1
371	@Src @Jrd µRel $@Bq µSta @µCh @Pré ƒqal ƒLqS $&DA $gel @Ths $hm1 @fum
	$&hX µΔel µSim £Par ƒusu µprs $EsL πPCc £Tr2 £ibr £Chn £mot $Lé1 $Lé2
	£Aur $eff @prf £pr1 ©$SF ®Vin $cdr £©F© ®Thr ©pub £SMF $@ct £Rep
360	£hxL £frL ®Cé® πVPr µEnt @ngr
362	µQs1 ®Arb £Dia £vox @Goo ®Exm µEPR @Chr ©Grc π©rt ©Mrt
	©Alg ©Cal ©1Pr ©Chô ƒ∂ia ©$Tr ©Pnt $usi ◊Loi @man ®µVi µVAC ©Sal $$pr
	©vgn prédominance des cartes (©)
366	@µCt @Prd πjxt πCré ©stt πInd πCor … des programmes (π)
380	µinh µGut @Hst µPrt £Mt1 µbar µMlt µrec µInB … des mathématiques (µ)
36	£sty très brève note sur la stylistique
344	£pr1 ®rch µvoi µcah
339	$Arb $Esp $Cmc ®com ®étu économie et échanges internationaux
223	µB13 µbl∂ µbl2 µbl1 reconnaissance de la structure en blocs diagonaux
383	◊Oxy ®$Qb ßSIC ßcib ßvoi ßMrt ßpôl ßFrn @str ®Imm ©Par surtout ßourse
379	µbrd µCv1 µCv2 µméd géométrie: convexité, bord, médiane

```
  c | Partition : Sigles des formes de la classe numéro c
----+-------------------------------------------------------------------------
 359| (analyse ANLYSIS) (correspondances CORRESPONDENCES)
    | (propre (exemple EXAMPLES))
    | (TYPE facteurs) (statistique STATISTICAL) ((données DATA)(années YEARS))
    | temps (COMPARISON (factorielle FACTOR)) forme (methodes METHODS) LATIN
    | (multidimensionnelle MULTIDIMENSIONAL) ((compte INFORMATION) recherche)
    | (espace (COLLECTED CONSIDERED))
-------------------------------------------
 331| (ORDER ordre) ((modalité MODALITY)(CATEGORY RESPONSE))(VARIABLE variable)
-----------------------------------------------
 335| (classes CLASSES) ((représentation REPRESENTATION) (variables VARIABLES))
 313| (tableaux TABLES) ((ANALYSE (croisant CROSSING)) (ensemble SET))
 309| (MATRIX économie) ((base colonne)ANALYSED)((tableau TABLE)correspondance)
-------------------------------------------
 254| notice (PROGRAMS programmes)
-------------------------------------------------
 358| (note (ARTICLE article)) (SURVEY enquête)
    | (résultats RESULTS) (auteur AUTHOR)
-------------------------------------------------
 340| (objet titre) (groupes ((PRESENT PROGRAM) programme))
 336| (évolution EVOLUTION) fonction (période PERIOD) (étude STUDY)
 320| (valeurs VALUES) (PRICES SHARES)
-----------------------------------------------------
 341| (vih virus)
-----------------------------------------------------------
 350| (FRENCH((france FRANCE)(departements DEPARTEMENTS))) (région(MAP carte))
 280| (PRODUCTS produits) (consommation CONSUMPTION)
-------------------------------------------------------------
 279| (critère CRITERION) (MINIMUM (trace TRACE))
-------------------------------------------------------
 370| RESEARCH    (((population flux) (MARKET marché)) (COUNTRIES pays))
-------------------------------------------------------------
 216| PARIS paris
-----------------------------------------------------------------------
 326| (dépenses SALARY) (TRAINING formation)
-------------------------------------------------------------
 360| TEST      (CODING codage)
-----------------------------------------------------------
 303| (LEVEL niveau) (réponse (QUESTION question))
 323| (SUBJECTS sujets) ((questionnaire QUESTIONNAIRE) (réponses RESPONSES))
-----------------------------------------------------------------------------
 226| (modèle MODEL)
 300| (METHOD méthode) (TIME (SERIES (SHIFTED (décalées séries))))
-------------------------------------------------------------
 241| (AIDS sida)
-----------------------------------------------------------------------------
 295| (CLASSIFICATION classification) (HIERARCHICAL (CLUSTERING cah))
    | (ascendante hiérarchique)
 334| (ensembles SETS) ((structure STRUCTURE) (BLOCK (blocs BLOCKS)))
-----------------------------------------------------------------------------
 228| (ARISTOTLE aristote)
 367| ((TEXT texte) (GREEK grec)) (CHAPTERS chapitres)
    | ((TYPOLOGY typologie) (WORD mots)) ((TEXTS textes) (FORMS formes))
-----------------------------------------------------------------------------
 357| (TREATMENT traitement) ((COMPARATIVE traitements) (patient PATIENT))
    | ((placebo PLACEBO) (PRODUCT produit))
-----------------------------------------------------------------------------
 361| (mesure (DIMENSION dimension)) ((DISTANCE distance) (masse MASS))
```

Dans la CAH des mots, environ les deux tiers sont correctement agrégés en paire bilingues. D'autre part, plusieurs classes de formes s'interprètent bien comme relevant soit d'un thème soit d'une méthode. Par exemple, 357 concerne les essais thérapeutiques; 367, les données linguistiques; les deux classes 295 et 334 renferment le vocabulaire des articles consacrés à la reconnaissance, par la CAH, de la structure de blocs d'un tableau de correspondance; dans 226 et 300 on retrouve la méthode des séries décalées; dans 303 et 323, l'acquisition d'information par questionnaire; etc.

2.4 Croisement des résumés avec un lexique français de 109 mots pleins

c	Partition en 20 classes : Sigles des formes de la classe numéro c
163	âge sexe consommation
187	classes modèle variable modalités
204	image enquête point banques marché
211	formation temps ((français système) mathématique)
	(dépenses compte) (correspondances travail formation)
	(cohérence population aide corpus) (années exemple correspondance texte)
	(temporelle période France (monétaire série cours))
	(son titre valeurs) (départements régions carte)
210	effet groupes discriminante objet entreprises programme codage
208	cartographie programmes macsaif
217	latin grec outil chapitres formes mots textes
200	flux trace critere
215	bourse modèles croissance décalées séries corrélation
5	Aristote
197	(espace cah classification ascendante hiérarchique)
	(structure reconnaissance ensembles blocs)
141	secteurs économie matrice branches
198	centre masse distribution dimension mesure
224	sida pays
180	Paris prix
227	(vie politique) Algérie
202	infection vih virus
218	thérapeutiques essais stabilité
	(échelles notes états typologie sujets)
	(questionnaire réponses échelle question sujet réponse)
219	sevrage efficacité placebo thérapeutique traitement traitements patients
113	tri

À titre complémentaire, on présente une classification portant sur un lexique ∂ de 109 mots français. Le lexique diffère de π, en ce que, d'une part, on a abaissé de 8 à 6 le seuil de fréquence; mais, d'autre part, ont été éliminés plusieurs mots pleins génériques, ou peu susceptibles d'évoquer, sans ambiguïté, un contenu.

Par exemple, des 19 mots de fréquence 8 retenus dans π, ne sont pas dans ∂ ceux précédés ici du caractère $\int$.

{$\int$colonnes dépenses économie $\int$facteurs $\int$fonction formation forme groupes $\int$intérêt objet placebo population $\int$propre régions réponse $\int$représentation titre traitements VIH}.

Comme au § précédent, on trouve plusieurs classes ou groupes de classes dont l'interprétation est claire: {202, 218, 219} renferment presque tout ce qui concerne la médecine; la reconnaissance de la structure de blocs par la CAH est dans 197; la linguistique dans 217; etc.

```
99 mots français π + 94 mots anglais Π = 193 mots pleins × 201 résumés bilingues;
trace   : 1.763e+1
rang    :    1     2     3     4     5    10 ..  15 ..  25 .  50    75
lambda  : 5619  5422  5298  4815  4638  4093  3308  2352  1325   644 e-4
taux    :  319   308   300   273   263   232   188   133    75    37 e-4
cumul   :  319   626   927  1200  1463  2669  3683  5229  7728  9074 e-4
```

```
119 mots français de ∂ x 201 résumés de CAD
trace   : 2.961e+1
rang    :    1     2     3     4     5 ..  10 ..  15 ..  25 ..  50 ..  75
lambda  : 7535  7428  6897  6780  6609  6014  5395  4486  2489  1217 e-4
taux    :  254   251   233   229   223   203   182   151    84    41 e-4
cumul   :  254   505   738   967  1190  2247  3215  4858  7633  9143 e-4
```

L'occasion s'offre de comparer les tableaux de valeurs propres issues des analyses qui font l'objet des §§2.2, 2.3 et 2.4. Au §2.2, avec des mots outil pour variables, la trace et les premières valeurs propres sont dix fois plus faibles que dans les analyses prenant en compte des mots pleins; la trace est maxima au §2.4, où le seuil de fréquence est le plus bas.

Références bibliographiques

A. Aït HAMLAT: "Analyse des répétitions et indexation automatique des documents", [IND. DOC.], in *CAD*, Vol. IX, n°2, pp. 173-204; (1984).

Ch. ARBACHE: "Distribution des vocables dans le texte hébraïque de la Bible et dans la traduction grecque des Septante", [TEXTE BIBLE] §9, in *CAD*, Vol. XI, n°1, pp. 19-23; (1986).