Annexe A

Description sommaire de quatre logiciels utilisant la statistique textuelle multidimensionnelle

De nombreux logiciels permettent désormais de réaliser les opérations de segmentation, de comptage et de documentation (index, concordances, sélection de contextes) que l'on a définies au chapitre 2. Dans un proche avenir, on peut prévoir que leurs fonctionnalités communes seront de plus en plus intégrées au sein des logiciels de traitement de texte les plus courants, comme c'est déjà le cas pour le comptage des occurrences et la recherche des contextes d'une même forme graphique ou chaîne de caractères.

On ne tentera pas ici de dresser un inventaire des possibilités et des limites de chacun de ces logiciels ni de comparer leurs performances respectives. On se limitera au contraire à un survol rapide des seuls logiciels qui font appel à l'analyse statistique multidimensionnelle des données textuelles, objet du présent ouvrage.

Les deux premières sections seront consacrées à la description des deux logiciels qui ont servi à produire l'ensemble des analyses dont rendent compte les différents chapitres de ce livre : *SPAD.T* (section A.1), plus particulièrement orienté vers le traitement des réponses à des questions ouvertes fournies par des individus nombreux sur lesquels on possède par ailleurs des renseignements de type socio-économique, et *Lexico1* (section A.2) conçu pour le traitement lexicométrique de textes, moins nombreux, mais comportant chacun plusieurs milliers, voire centaines de milliers d'occurrences.
Les deux sections suivantes décrivent succinctement deux autres logiciels (également consacrés à l'analyse multidimensionnelle des données textuelles) auxquels on s'est référé dans le présent ouvrage. La section A.3 présente le logiciel *ALCESTE*, qui permet de mettre en oeuvre des méthodes voisines de celles que nous avons décrites dans le présent ouvrage, pour un contexte d'applications assez différent.
La section A.4 présente le logiciel *HYPERBASE.* Ce logiciel diffère peu, au plan proprement statistique, de ceux qui sont présentés dans les deux premières sections. Son originalité réside avant tout dans le recours aux méthodes de ce que l'on appelle l'*hypertexte*, qui permettent de *naviguer* aisément dans de grandes masses de textes.
Enfin, la section A.5 donnera les références de quelques autres logiciels, en général moins centrés sur la statistique textuelle.

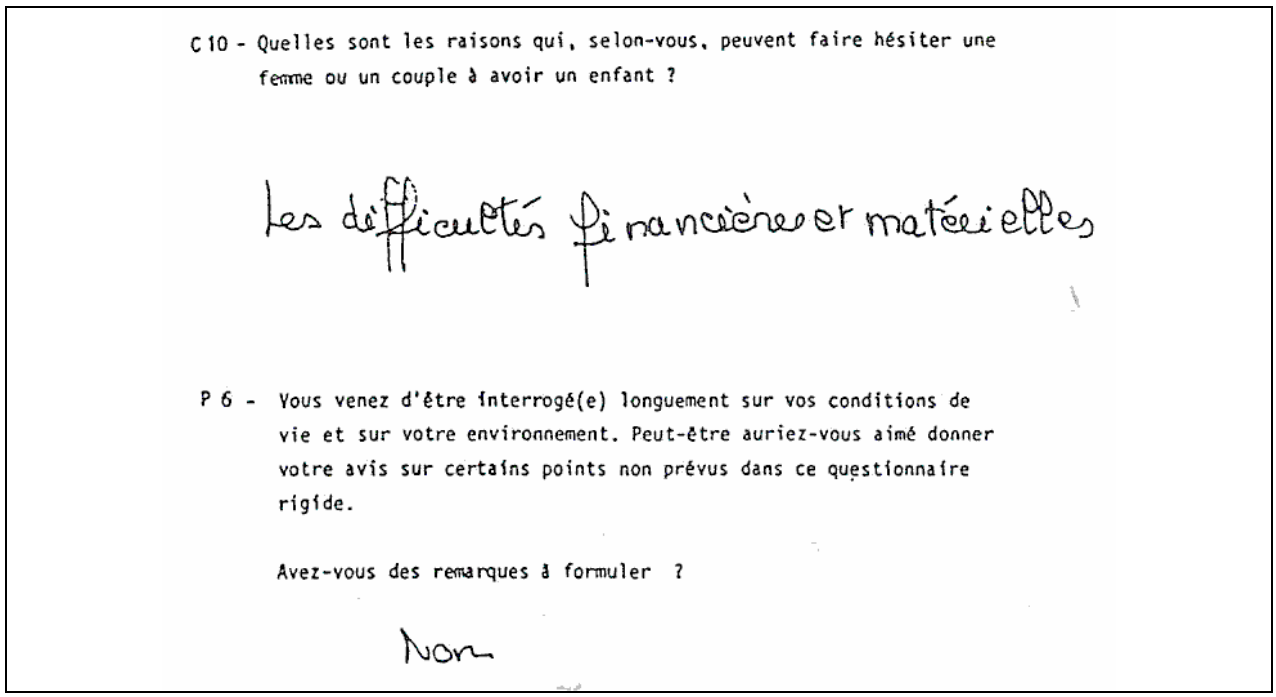
A.1 Le logiciel SPAD.T

On présente ici l'enchaînement des tâches, depuis la saisie de l'information, jusqu'à la production des listages et graphiques par SPAD.T.[1] Bien entendu, il ne s'agit que d'une illustration, et non d'un guide d'utilisation complet du logiciel.

a) Le document de base

La figure A-1 est un fac-similé des libellés de deux questions ouvertes du questionnaire de l'enquête précitée sur les conditions de vie et aspirations des Français, avec les réponses telles qu'elles ont été transcrites par un enquêteur.

Comme le montrent les numéros de ces questions, celles-ci ne sont pas consécutives dans le questionnaire. Elles sont en fait séparées par des questions fermées, qui concernent des thèmes distincts.


C 10 - Quelles sont les raisons qui, selon-vous, peuvent faire hésiter une femme ou un couple à avoir un enfant ?

les difficultés financières et matérielles

P 6 - Vous venez d'être interrogé(e) longuement sur vos conditions de vie et sur votre environnement. Peut-être auriez-vous aimé donner votre avis sur certains points non prévus dans ce questionnaire rigide.

Avez-vous des remarques à formuler ?

Non

Figure A1.1. Fac-similé de réponses à deux questions ouvertes

On reconnaît en la question C10, la question ***Enfants*** qui a illustré plusieurs exemples des chapitres 2 à 6. Les réponses à la question P6 ont, quant à elles, illustré la partition en noyaux factuels des chapitres 5 et 6.

b) Les fichiers de base

La figure A1.2 reproduit le listage du fichier de saisie des réponses à ces quatre questions pour les quatre premiers individus enquêtés. On reconnaît, pour l'individu 1 les réponses de la figure A.1.

[1] SPAD.T réalisé au départ dans un cadre universitaire, fonctionne sur PC et MacIntosh. Il est actuellement maintenu et distribué par le CISIA (1 avenue Herbillon, 94160, Saint-Mandé).

Le format de saisie est donc simple : le séparateur "----" annonce un nouvel individu, alors que le séparateur "++++" indique la fin d'une réponse, "====" indiquant la fin du fichier. Deux séparateurs de réponses consécutifs indiquent une réponse vide. Pour un individu donné, les réponses doivent toujours figurer dans le même ordre.

Dans le cas d'une seule question, il est possible d'introduire un séparateur "****" , dit séparateur de texte, qui permet de séparer des groupes d'individus consécutifs. Ainsi, les individus peuvent être des strophes et les textes des poèmes, les individus peuvent être des phrases et les textes des articles de journaux, etc...

```
----1
les difficultés financières et matérielles
++++
non
----2
l'avenir incertain,les problèmes financiers
++++
je trouve ce questionnaire intéressant
----3
les difficultés financières
++++
c'est un peu long
----4
les raisons matérielles et l'avenir qui les attend
++++
toutes les réponses ne sont pas formulées,on est obligé de choisir entre des
réponses qui ne correspondent pas toujours à ce qu'on pense
====
```

Figure A1.2. Listages du fichier textuel de base

SPAD.T suppose de plus qu'il existe un tableau numérique apparié au précédent décrivant, pour les mêmes individus, disposés dans le même ordre, leurs réponses à des questions fermées codées sous forme de variables nominales (variables dont les modalités s'excluent mutuellement). Ce fichier doit avoir dans la version actuelle, la forme d'un fichier standard SPAD.N, (logiciel de dépouillement d'enquête distribué par le CISIA) ou d'un fichier interfacé avec SPAD.N.

C'est la présence simultanée de ces deux fichiers qui permet de regrouper les réponses à une question ouverte selon les modalités de réponse à une question fermée, de positionner des formes graphiques dans des espaces factoriels calculés à partir de variables nominales, et de façon symétrique, de positionner des caractéristiques des individus dans les espaces calculés à partir des formes graphiques figurant dans leurs réponses libres.

c) Principes du logiciel

Ce logiciel admet donc comme entrée un fichier de données textuelles du type de celui de la figure A1.2 (fichier *texte*). Dans le cas (le plus fréquent) où il existe des variables nominales, il admet également des fichiers du type SPAD.N (fichier *données* et fichier *dictionnaire*).

Le logiciel comporte 15 procédures de base, qui seront désignées chacune par un mot-clé, suivi des valeurs d'un ou plusieurs paramètres. Divers enchaînements de ces procédures permettent de réaliser l'essentiel des traitements décrits précédemment.

Les procédures peuvent être réalisées une par une, car elles ne communiquent entre elles que par des fichiers externes qui peuvent être sauvegardés individuellement.

d) Les principales procédures

ARTEX Archivage du texte. Cette procédure lit le fichier de base (type figure A1.2) et le structure de façon à le rendre facilement accessible pour les procédures suivantes.

SELOX Sélection de la question à traiter dans le fichier archive, avec éventuellement filtre sur les individus.

NUMER Numérisation du texte de la question choisie. Les tableaux 5.1 et 5.2 du chapitre 5 font partie des sorties de cette procédure.

CORTE Cette procédure permet de supprimer des formes (par exemple : des mots-outil), et de déclarer équivalentes des listes de formes (par exemle : relatives à un même lemme)

MOTEX Construction d'un tableau lexical (après NUMER) ou segmental (après SEGME) pour une variable nominale à préciser.

APLUM Analyse des correspondances d'un tableau lexical issu de MOTEX, avec graphiques. Les figure 5-1, 8.3 sont issues des sorties de cette procédure.

MOCAR Calcul et impression des formes caractéristiques pour chacune des modalités de la variable nominale désignée dans MOTEX. Eventuellement, calcul des réponses modales pour le critère de la fréquence lexicale. Le tableau 6.6 est une des sorties de MOCAR.

RECAR Calcul et impression de réponses modales (pour chacune des modalités de la variable nominale sélectionnée dans MOTEX) selon le critère du chi-2.

ASPAR Analyse des correspondances directe d'un tableau de réponses numérisées issu de la procédure NUMER, sans regroupement. Cette procédure utilise des algorithmes particuliers adaptés aus grands tableaux clairsemés. La figure 5.4 est issue d'une des sorties graphiques de la procédure ASPAR.

SEGME Construction du tableau segmental donnant, pour chaque individu, les numéros des segments contenus dans sa réponse. La procédure SEGME doit suivre la procédure NUMER.

POLEX Cette procédure permet de positionner des formes ou des segments comme éléments supplémentaires sur une analyse factorielle concernant le même ensemble d'individus. La figure 3.4, où quelques formes lexicales sont positionnées sur le premier plan factoriel d'une analyse des correspondances multiples, a été construites avec les résultats d'une procédure POLEX.

TALEX Construction d'un juxtaposition de tableaux lexicaux (après NUMER) ou segmentaux (après SEGME) pour une liste de variables nominales à sélectionner par la procédure SELEC.

Les procédures suivantes sont communes aux logiciels SPAD.N et SPAD.T :

ARDIC Archivage d'un dictionnaire décrivant un fichier d'enquête (variables nominales et numériques continues).

ARDON Archivage du fichier de données décrit par ARDIC.

SELEC Sélection de groupes de variables nominales ou continues en vue d'une analyse factorielle, et en vue des étapes POSIT ou TALEX.

POSIT Positionnement de variables nominales illustratives sur des plans factoriels calculés par ailleurs. La figure 5.5 est un exemple de sortie de la procédure POSIT, qui fait suite à une procédure ASPAR.

e) Exemples d'enchaînements

On donnera ci-dessous quelques exemples d'enchaînements de procédures permettant d'effectuer les opérations les plus usuelles. Le fichier de variables nominales est supposé "archivé" (par les procédures ARDIC et ARDON, identiques à celles du logiciel SPAD.N)
Les instructions sont en format libre, et n'ont nul besoin d'être cadrées. Les titres (obligatoires sur la ligne qui suit l'instruction "PROC") sont à l'initiative de l'utilisateur. Les commentaires peuvent être insérés à droite du séparateur " : ". Lorsqu'un paramètre ne figure pas, sa valeur par défaut est utilisée. Dans la version 1994 pour PC, les commandes se font par menus déroulants.

1) Analyse des correspondances d'un tableau lexical : Séquence des instructions de commande

```
 PROC  ARTEX
 Saisie des questions ouvertes     (titre de la procédure)
 ITYP = 2,   NCOL= 68,  NBQT = 4
                  : Fichier type enquête (ITYP=2), enregistré sur 68 colonnes (NCOL),
                  : avec ici 4 questions ouvertes  (NBQT=4)
 PROC  SELOX
 Choix de la question à traiter   (titre de la procédure)
 NUMQ=2,  LSELI= 0
  : On choisit la question n°2 (NUMQ), sans filtre sur les individus (LSELI=0)
 PROC  NUMER
 Numérisation de la question "Enfants"      (titre de la procédure)
 NSEU= 13,   NXMAX= 160,  NXSIG= 60
 FAIBLE                    : Le nombre de formes retenues sera inférieur à 160
 ,   ;   '   -   (   )     : (NXMAX). De plus, celles-ci devront avoir une
 FORT                      : fréquence supérieure à 13 (NSEU), Il y a moins de
 .   !   ?                 : 60 caractères par ligne (NSIG). Suivent des listes de
 FIN                       : séparateurs faibles et forts.
 PROC MOTEX
 Croisement des formes avec la variable 341 ( âge-diplôme)  (titre )
 NVSEL= 341   : NVSEL est le numéro de la variable nominale.
PROC  APLUM
Analyse du tableau lexical et graphiques    (titre de la procédure)
NAXE=6,   LIMPR=1,   NGRAF=1
   : On calcule 6 axes factoriels, puis on liste lescoordonnées des lignes (formes)
```

: (LIMPR=1), on demande enfin un graphique, qui sera ici le plan factoriel (1,2).
STOP

Parmi les listages correspondant à cet enchaînement figurent les tableaux 5.1 et 5.2 (procédure NUMER), le tableau 5.3 (Procédure MOTEX) et les figure 5.1 (Procédure APLUM).

2) Formes caractéristiques et réponses modales

On ne répétera pas les commentaires relatifs aux procédures précédentes.

```
PROC  ARTEX
Saisie des questions ouvertes
ITYP = 2,   NCOL= 68,  NBQT = 4

PROC  SELOX
Choix de la question à traiter
NUMQ=2,  LSELI= 0

PROC  NUMER
Numérisation de la question "enfants"
NSEU= 13,   NXMAX= 160,  NXSIG= 60
FAIBLE
,   ;   '   -   (   )   :
FORT
.   !   ?
FIN

PROC MOTEX
Croisement des formes avec la variable 341 ( âge-diplôme)  (titre )
NVSEL= 341

PROC MOCAR
Formes caractéristiques et réponses modales   (titre de la procédure)
NOMOT = 10,   NOREP = 3
      : Il s'agit ici, pour les réponss modales, du critère de fréquence lexicale.
      : On demande 10 formes caractéristiques (NOMOT) et 3 réponses modales
      : (NOREP) par modalité de la variable nominale choisie dans MOTEX.

PROC RECAR
Réponses modales, Critère du Chi-2
NOREP = 4           : On demande 4 réponses modales par modalité.

STOP
```

3) Analyse directe, avec illustration par des variables nominales

On supposera que les étapes ARTEX, SELOX, NUMER ont déjà été effectuées, et que les fichiers correspondant sont sauvegardés.

```
PROC  ASPAR
Analyse directe du tableau de numérisation issu de NUMER   (titre)
```

```
NAXE= 5,  NGRAF= 2, NPAGE= 2,  NLIGN= 110
      : On demande 5 axes factoriels, 2 graphiques (plans 1,2  et 2,3), chacun
      : sur 2 pages en largeur et 110 lignes en longueur.
PROC  SELEC
Sélection des variables nominales supplémentaires  (titre )
NOPAR                          : Valeurs par défaut pour les paramètres
NOMI  ILL  313  326  17  34    : Numéros des 4 variables nominales illustratives
FIN

PROC  POSIT
Positionnement des nominales sélectionnées plus haut (titre de la procédure)
NAXED= 5,   NGRAF = 2
                    : On demande l'impression des coordonnées sur les
                    : 5 premiers axes factoriels, et 2 graphiques.
STOP
```

La figure 5.4 est un exemple de graphique obtenu à l'issue de la procédure ASPAR. La figure 5.5 est un exemple de graphique (il s'agit du même plan factoriel) obtenu à l'issue de l'étape POSIT.

4) Analyse des correspondances d'un tableau segmental

On supposera encore que les étapes ARTEX, SELOX, NUMER ont déjà été effectuées, et que les fichiers correspondant sont sauvegardés.

```
PROC  SEGME
Recherche des segments répétés   (titre de la procédure)
NXLON= 6,   NXSEG= 600,  NFRE1= 8,   NFRE2 = 12

NSPA= 'NSPB'      : On précise que MOTEX ne doit pas fonctionner
                  : avec le tableau issu de NUMER, mais avec le
                  : tableau de même type issu de SEGME.

PROC MOTEX
Croisement des segments avec la var. 341 ( âge-diplôme)  (titre)
NVSEL= 341        : NVSEL est le numéro de la variable nominale.

PROC  APLUM
Analyse du tableau lexical et graphiques    (titre de la procédure)
NAXE=6,   LIMPR=1,   NGRAF=1
                  : On calcule 6 axes factoriels, puis on liste les
                  : coordonnées des lignes (formes) (LIMPR=1), on
                  : demande enfin un graphique, qui sera ici le plan
                  : factoriel (1,2).

STOP
```

Ces quelques exemples n'épuisent pas les possibilités du logiciel. Ils doivent montrer que le langage de commande permet des enchaînements assez variés.

A.2 Le logiciel Lexico1

Lexico1 est un ensemble de programmes lexicométriques fonctionnant sur micro-ordinateur.[1] L'ensemble est pour l'instant composé de cinq modules distincts :

SEGMENTATION crée une base de données numérisées à partir d'un fichier texte fourni par l'utilisateur. Cette base est constituée d'un dictionnaire des formes rencontrées dans le texte qui leur affecte également un numéro d'ordre, et d'une version numérisée du texte.

DOCUMENTATION permet de lancer plusieurs types de requêtes documentaires dont les résultats seront, selon le désir de l'utilisateur, affichés à l'écran et/ou stockés dans un fichier éditable par la suite.

STAT1 (Statistiques - module 1) Calcule les segments répétés du texte (cf. chapitre 2), construit des tableaux lexicaux à partir des partitions du corpus décidées par l'utilisateur, opère des calculs statistiques portant à la fois sur les formes et les segments répétés du corpus.

AFC réalise l'analyse des correspondances du tableau lexical constitué à partir d'une partition du texte.

CHRON2 calcule, pour une série textuelle chronologique (cf. chapitre 7) les spécificités chronologiques du corpus ainsi que les accroissements spécifiques pour chacune des parties.

A.2.1 Le texte en entrée

Lexico1 accepte en entrée tout texte, saisi sur traitement de texte[2] mais sauvegardé avec les options : "texte seulement avec ruptures de lignes"[3].

Les deux caractères < et > sont des caractères réservés à l'encodage des clefs. Ils ne doivent figurer dans l'enregistrement d'entrée que pour introduire des informations péritextuelles.

ex : <Epg=238> introduit la page 238
<AD=25> assigne la valeur 25 à la clef AD

Le type des clefs (i.e. la zone située entre le signe < et le signe =) peut être quelconque. Le contenu des clefs (i.e. la zone située entre le signe = et le signe >) est obligatoirement numérique dans cette version.
Dans l'exemple ci-dessous, la clef AD permet d'affecter un code âge-diplôme à chacun des répondants d'une enquête socio-économique (le chiffre des dizaines correspond à une catégorie d'âge, celui des unités à une catégorie de diplôme).

---

[1] Lexico1 est développé par A. Salem au laboratoire *Lexicométrie & textes politiques* de l'E.N.S. de Fontenay-Saint-Cloud. Une version de ce logiciel fonctionne actuellement sur les micro-ordinateurs de type MacIntosh. Elle permet de traiter des corpus comptant jusqu'à 700 000 occurrences environ.

[2] Tels les traitements de texte WORD, MacWrite, QUED, Edit, etc. que l'on trouve actuellement sur le marché.

[3] Sur les traitements de textes anglo-saxons option "text only".

```
<AD=23>   *les difficultes financieres et materielles
<AD=13>   *les problemes materiels,une certaine angoisse vis a vis
          de l'avenir
<AD=23>   *la peur du futur,la souffrance,la mort,le manque
          d'argent
<AD=23>   *l'avenir incertain,les problemes financiers
<AD=23>   *les difficultes financieres
<AD=12>   *les raisons materielles et l'avenir qui les attend
<AD=13>   *des problemes financiers
<AD=31>   *l'avenir difficile qui se prepare,la peur du chomage
<AD=13>   *l'insecurite de l'avenir
<AD=12>   *le manque d'argent
<AD=33>   *la guerre eventuelle
<AD=23>   *la charge financiere que ca represente,la
          responsabilite morale aussi
<AD=13>   *la situation economique,quand le couple ou la femme
          n'est pas psychologiquement pret pour accueillir un
          enfant
```

Figure A2.1. Lexico1 : exemple de texte en entrée

A.2.2 La segmentation du texte

A partir d'un tel fichier texte, en s'appuyant sur la liste des caractères délimiteurs fournie par l'utilisateur, le premier module opère la segmentation automatique du texte, calcule le nombre des occurrences et l'ordre alphabétique pour chacune des formes graphiques contenues dans le corpus.

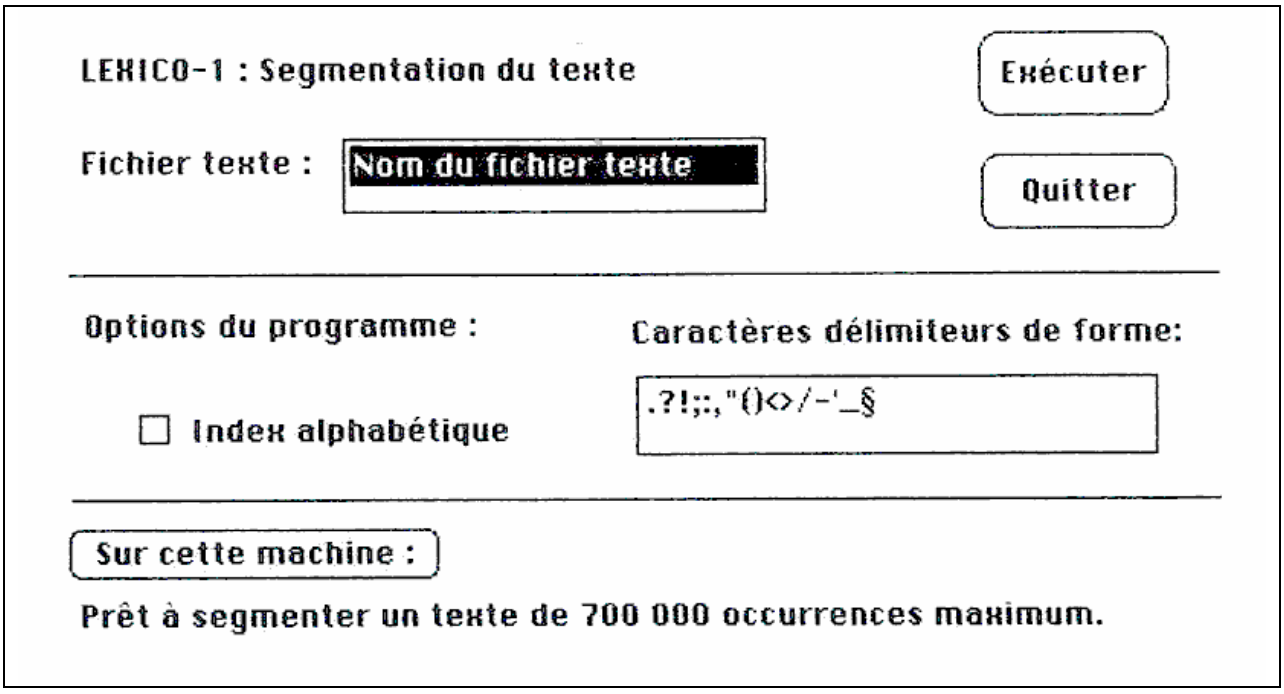


Figure A2.2. Réglage des options de segmentation du texte

Le programme crée ensuite une base de données numérisées qui servira de point de départ pour les travaux documentaires et pour l'étude statistique du texte. Cette base est constituée d'un dictionnaire des formes rencontrées dans le texte lequel contient

également pour chacune d'entre elles un numéro d'ordre lexicométrique (ordre par fréquence décroissante, l'ordre alphabétique départageant les formes en concurrence) et un numéro d'ordre alphabétique, ainsi que d'une version numérisée du texte. En cochant la case "index alpha", on obtient en plus un fichier des formes graphiques classées par ordre lexicographique

A.2.3 La phase documentaire

Le module de documentation permet de retrouver l'ensemble des contextes d'une forme sélectionnée par l'utilisateur. Pour une forme-pôle donnée cet ensemble (un contexte par occurrence de la forme pôle) peut être ordonné au gré de l'utilisateur :

— en fonction de la forme qui précède la forme-pôle (tri avant)
— en fonction de la forme qui suit la forme-pôle (tri après)
— en fonction de l'ordre d'apparition dans le texte.

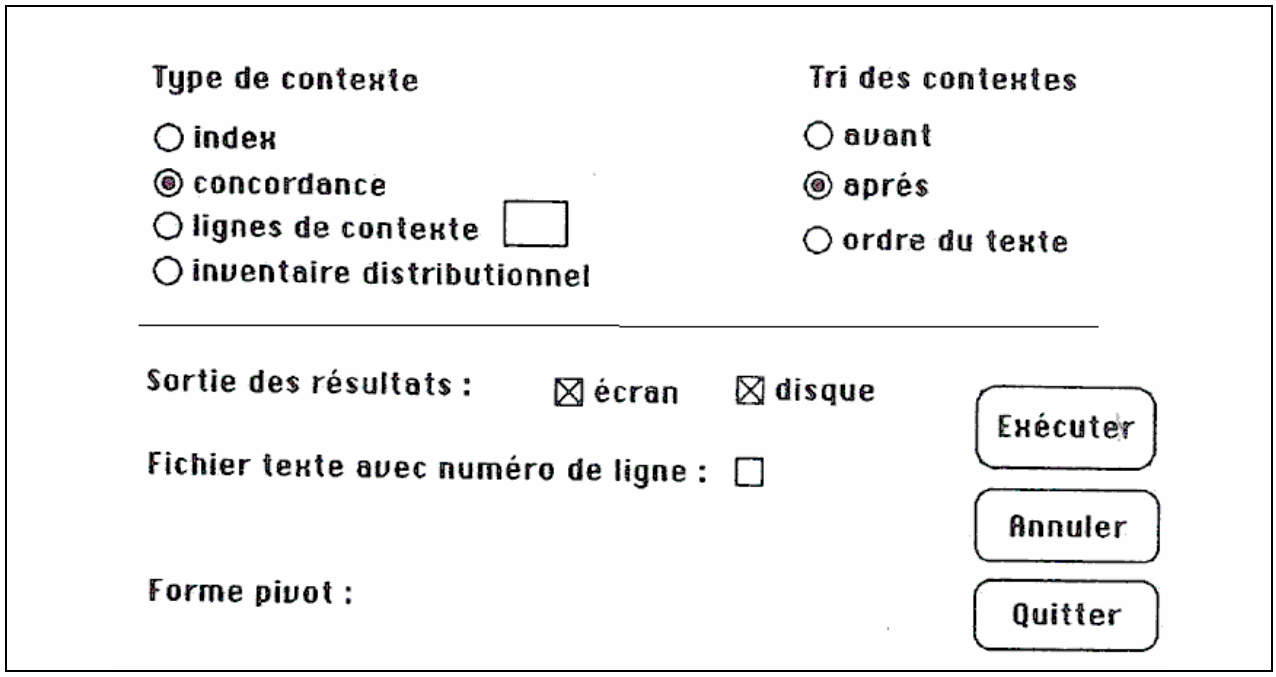


Figure A2.3

Lexico1 : Réglage des options du module de documentation

Les types de contextes disponibles dans cette version sont :

— Index : aucun contexte, mention des lignes ou la forme est attestée.
— concordance : une ligne de contexte centrée sur la forme pivot comportant la mention du numéro de ligne de l'occurrence de la forme pivot.
— contexte: un nombre, définit par l'utilisateur, de lignes de contexte avant et après chaque occurrence de la forme pivot comportant la mention du numéro de ligne de cette occurrence.

Dans ce programme, index, concordances, lignes de contextes renvoient un numéro de ligne qui est le numéro de la ligne dans le fichier d'entrée. Il est donc recommandé d'établir une édition de référence dans laquelle les lignes du texte sont numérotées

A.2.4 Partition du corpus et segments répétés

Ce module permet d'opérer une partition du corpus d'après les différentes valeurs d'une des clefs introduites avant l'étape de segmentation. Les différentes parties du corpus permettent ensuite de construire le tableau lexical qui servira de base aux différentes analyses statistiques. Ce même module calcule également les segments répétés du texte, dont la fréquence dépasse un seuil minimal fixé par l'utilisateur, ainsi que leur ventilation dans les parties du corpus.

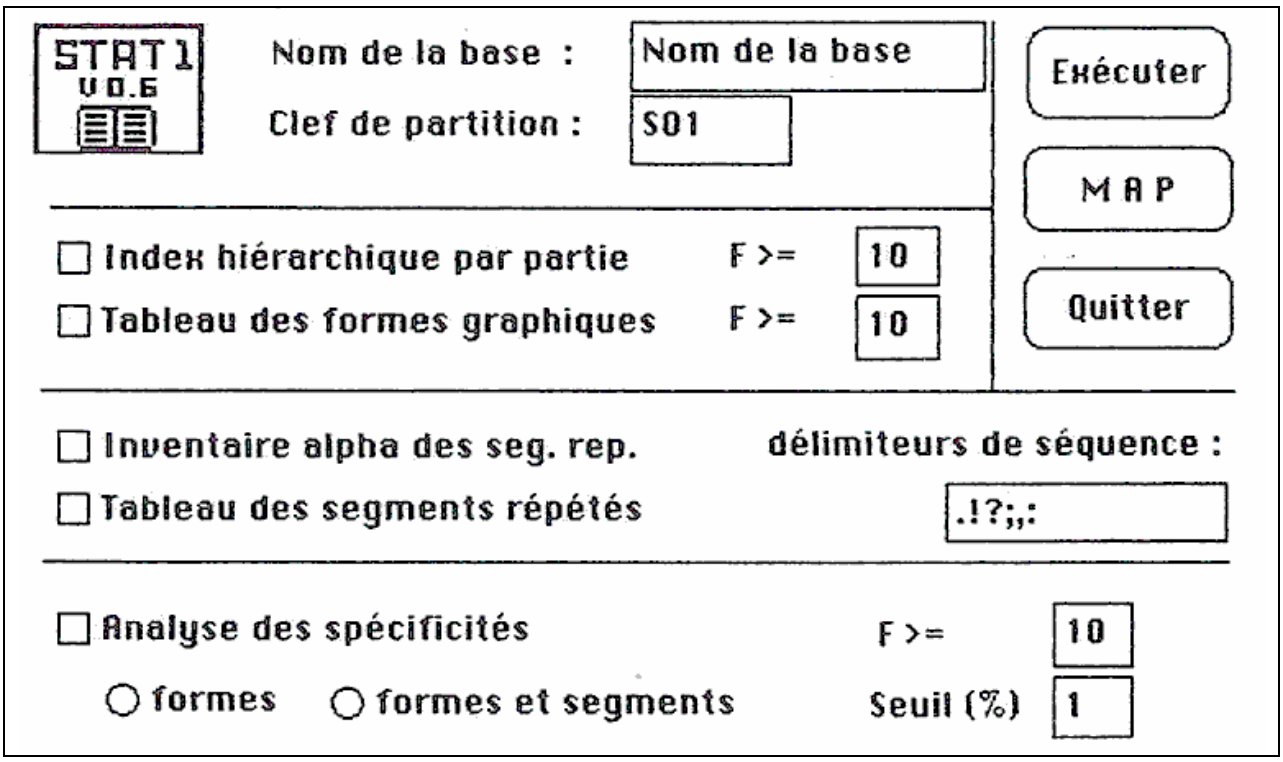


Figure A2.4 . Lexico1 : réglage des options du module *statistiques -1*

On obtient, un *index hiérarchique par partie*, pour chacune des parties du corpus, en cochant la case correspondante et en indiquant une fréquence minimale. Chaque index est classé par ordre de fréquence des formes dans la partie sélectionnée. Chaque index est suivi du rappel des principales caractéristiques lexicométriques de la partie.

En cochant la case "Tableau des formes graphiques", et en indiquant une fréquence minimale on obtient un tableau lexical (formes x parties). On peut également obtenir le tableau qui donne la ventilation des segments dans les mêmes parties en cochant la case correspondante.

On effectue l'Analyse des Spécificités du corpus sur les formes, ou sur les formes et les segments dont la fréquence dépasse un seuil sélectionné, en indiquant une fréquence minimale et un seuil en probabilité.

A.2.4 Analyse des correspondances

Le module AFC effectue l'analyse des correspondances des tableaux (formes x parties) ou (formes + segments x parties)[1].

Tous les paramètres du programme peuvent être modifiés par l'utilisateur en éditant le fichier "afc.param" qui contient les options par défaut. Sans modification explicite de la part de l'utilisateur, ce programme effectue une analyse des correspondances du tableau des formes (à partir de la fréquence minimale sélectionnée dans le module précédent), les segments répétés au-delà de la même fréquence intervenant en qualité d'éléments supplémentaires.

A.2.5 Spécificités chronologiques

A partir d'une série textuelle chronologique divisée en périodes, ce dernier module calcule les spécificités chronologiques du corpus ainsi que les accroissements spécifiques. Les diagnostics de spécificités sont chaque fois triés par probabilité croissante (i. e. des plus remarquables aux moins remarquables) et par période.

A.3 Le logiciel ALCESTE

Le logiciel *ALCESTE*[2] est le résultat d'un approche d'analyse des données textuelles plus spécifiquement orientée vers l'analyse de corpus de textes homogènes (par exemple, un roman, un recueil de poèmes, un corpus d'entretiens, un recueil d'articles sur un même thème, un ensemble de réponses à une question ouverte, etc…). Il est conçu pour être utilisé en liaison avec une analyse de contenu.

L'objectif est d'obtenir un premier classement des "phrases" (i.e. unités de contexte) du corpus étudié en fonction de la répartition des mots dans ces "phrases" (deux phrases se "ressemblent" d'autant plus que leur vocabulaire est semblable, les mots grammaticaux étant exclus) ceci afin d'en dégager les principaux "*mondes lexicaux"*.

L'auteur du logiciel fait l'hypothèse que l'étude statistique de la distribution de ce vocabulaire peut permettre de retrouver la trace des "espaces référentiels" investis par

[1] Ce module recouvre en fait un interface du programme ANCORR du logiciel LADDAD mis à notre disposition par l'association ADDAD.

[2] ALCESTE est développé par M. Reinert (CNRS, Université de Toulouse-Le Mirail ; 5, allée Antonio Machado, 31058 Toulouse-cedex) pour micro-ordinateurs de type MacIntosh. Capacité de la version actuelle : Corpus traité : environ 20 000 lignes de 70 caractères (environ 1 MO). Nombre d'unités de contexte élémentaires (u.c.e.) : 10 000 ; Longueur maximum d'une u.c.e : environ 240 caractères. Nombre maximum d'unités de contexte naturelles : 4 000.

l'énonciateur lors de l'élaboration du discours, trace perceptible sous forme de "mondes lexicaux" (ensemble des mots plus spécifiquement associés à telle classe caractéristique de phrases). Chaque "monde" est ensuite décrit de différentes manières ; notamment à l'aide d'un ensemble de *phrases caractéristiques*, à l'aide de *segments de texte répétés*, à l'aide aussi des lignes du corpus contenant tel ou tel mot caractéristique du monde lexical étudié.
Au niveau informatique, *la version 2.0* du logiciel ALCESTE fonctionne pratiquement sur tout Macintosh muni d'un coprocesseur arithmétique et d'une mémoire centrale égale ou supérieure à 5 MO. Cette version comprend deux modules :

- *Un module de calcul,*
- *Un module de documentation et d'aide à la préparation du plan d'analyse* écrit en HyperTalk sous HyperCard.[1] Ce module se présente sous la forme d'une pile HyperCard d'une centaine de cartes environ.

1) La procédure de mise en oeuvre du logiciel :

a) Crée un dossier d'analyse ne contenant au départ que le seul fichier texte à analyser (en format ascii).
b) Prépare ce dossier à l'aide de la pile HyperCard en le complétant d'un certain nombre de fichiers auxiliaires dont *un plan d'analyse.*
c) Exécute le logiciel en "batch" qui se contente d'appliquer le plan d'analyse au traitement du corpus.
d) Lit les résultats à l'aide d'un éditeur quelconque, la description des principaux fichiers étant présentée dans la pile ou la notice du logiciel.

2) Cette exécution "batch" du logiciel enchaîne 3 étapes, chacune d'elles comprenant plusieurs opérations (programmables) dont voici la liste :

L'étape A permet la définition des "unités de contexte" (sensiblement des phrases), la recherche et la réduction du vocabulaire et le calcul des tableaux de données ; le calcul des couples et segments répétés…

A1 : Lecture et calcul des unités de contexte élémentaires (environ les phrases).
A2 : Recherche et réduction des formes (distinction des mots pleins et des mots outils ; réduction des désinences de conjugaison, des pluriels, etc…).
A3 : Calcul du tableau des données croisant unités de contexte par formes.
A4 : Calcul des couples de formes successives et des segments de texte répétés.

L'étape B effectue une classification des unités de contexte en fonction de la distribution du vocabulaire, classification simple ou double selon que l'on veut tester ou non la stabilité des résultats, en fonction d'une variation de longueur de l'unité de contexte ;

[1] Hypertalk et Hypercard sont des produits Apple.

B1 : Classification simple par la technique de classification descendante hiérarchique mise au point par l'auteur pour le traitement de tableaux clairsemés.

B2 : Classification double : deux classifications successives sur des tableaux ayant en lignes des unités de contexte de longueur différentes afin d'apprécier la stabilité des classes en fonction d'une variation du découpage en unités de contexte.

L'étape C permet plusieurs calculs auxiliaires pour aider à l'interprétation des classes ;

C1 : Choix des classes d'unités de contexte retenues.

C2 : Vocabulaire spécifique de chaque classe.

C3 : Analyse des correspondances sur le tableau de cooccurrences croisant classes par vocabulaire.

C4 : Choix des unités de contexte les plus représentatives de chaque classe.

C5 : Liste des segments répétés par classe.

C6 : Liste des formes d'origine par classe.

C7 : Calcul du concordancier pour les formes les plus spécifiques des classes.

A.4 Le logiciel Hyperbase

Ce logiciel[1] est construit à partir d'un langage à objets[2] et intègre la notion d'Hypertexte. Il en résulte pour l'utilisateur une commodité incomparable pour toute la partie qui concerne la "navigation" entre le texte et les outils documentaires, c'est à dire entre une forme de dictionnaire et ses différents contextes, concordances, ventilation dans les parties d'un corpus etc.
Les commandes sont d'accès facile, en général lancées par la simple sollicitation d'un "bouton". Les aides en ligne sont faciles d'accès et très explicites.

La partie statistique du logiciel, rançon inévitable du langage choisi pour l'écriture du logiciel, est moins développée que la partie documentaire. Elle fournit cependant l'accès aux principales méthodes lexicométriques, spécificités, analyse des correspondances, ainsi qu'une partie permettant de produire des histogrammes à partir des ventilations de formes sélectionnées.

---

[1] Hyperbase à été conçu et développé par E. Brunet, professeur à l'Université de Nice, pour micro-ordinateurs de type MacIntosh. Il permet de traiter des corpus qui comptent plusieurs millions d'occurrences. Cf. la publication : *Hyperbase*, CUMFID n° 17, URL 9, INaLF (CNRS), Faculté des Lettres, 98 Bd Herriot, 06007, Nice. Ce logiciel est interfacé avec le logiciel ADDAD pour la partie *analyses multidimensionnelles*.

[2] Le logiciel Hypercard est distribué par la firme Apple.

Figure A.4.1 Exemple d'écran *Hyperbase*
Choix des mots dans la version complète du logiciel

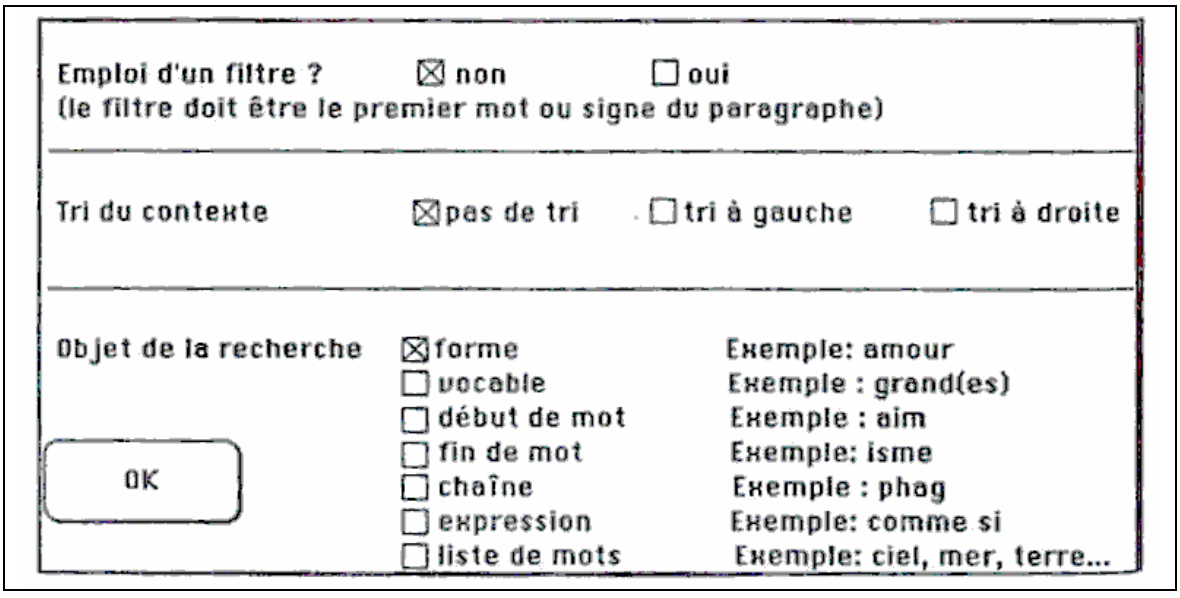


Figure A.4.2. Exemple d'écran *Hyperbase*
Dialogue de la commande *concordance*

Une dernière particularité du logiciel est de fournir, pour tout corpus entré par l'utilisateur, une comparaison statistique avec les données du trésor de la langue française (TLF) pour un corpus comportant plusieurs millions d'occurrences.

A.5 Autres Logiciels

En sus des quatre logiciels qui viennent d'être très brièvement présentés, on mentionnera, sans prétendre à l'exhaustivité, quelques autres produits, en général moins centrés sur la statistique textuelle.

On se limitera aux produits disponibles sur micro-ordinateurs.

SATO[1] qui remplit toutes les fonctions d'indexation du texte que nous avons citées plus haut et permet en outre d'affecter à chacune des occurrences du texte un certain nombre de propriétés (grammaticales, sémantiques ou autres) laissées au choix de l'utilisateur.

SAINT-CHEF[2] logiciel tout particulièrement consacré à la réalisation et à l'édition de concordances sur microordinateur.

PISTES[3] consacré à l'indexation et à l'analyse des spécificités d'un texte découpé en parties.

PHRASEA[4] consacré à la recherche documentaire au sein de vastes corpus de textes.

Le SPHINX[5], logiciel de dépouillement d'enquête doté d'une interface-utilisateur élaborée et comportant des modules de traitements de données textuelles.

LEXIS[6], un des modules de la série de logiciels de dépouillement d'enquêtes développée et distribuée par la société Eole.

---

[1] Sur PC, F Daoust, Centre d'analyse de textes par ordinateur, (ATO), UQAM, Case postale 8888, succursale A, Montréal, Québec, Canada H3C 3P8.
[2] Sur PC, M Sekhraoui, Lexicométrie & textes politiques ENS de Fontenay-St.Cloud.
[3] Sur PC, P. Muller, diffusé par le Centre National de Documentation Pédagogique, Paris.
[4] Sur MacIntosh, C. Poveda et J. Y. Jourdain, B&L Parenthèses, 79 av. Guynemer, 59 700, Marc en Bareuil
[5] Sur PC ou MacIntosh, J. Moscarola, Le SPHINX Développement, 13 Chemin des Amarantes, 74600, Seynod
[6] Sur PC, EOLE, 6 rue du Quatre Septembre, 92130, Issy-les-Moulineaux.

Annexe B

Esquisse des algorithmes et structures de données pour la statistique textuelle

Les logiciels décrits ou évoqués dans l'annexe A permettent la segmentation automatique d'un texte, l'indexation des unités textuelles, la recherche des contextes, etc. Les similitudes que l'on note entre ces différents logiciels montrent que leurs auteurs ont trouvé des solutions souvent très proches aux problèmes rencontrés ; les différences qui existent entre ces logiciels témoignent par ailleurs de leurs objectifs distincts.

Il nous a semblé utile de décrire sommairement dans cette annexe les principaux algorithmes qui sont à la base du logiciel *Lexico1*. Pour ce logiciel, les méthodes relatives à la segmentation automatique et à l'indexation des occurrences d'une forme sont fondées sur une structure de données particulièrement simple.[1] La numérisation du texte proposée permet, dans la plupart des cas, de stocker à la fois en mémoire vive le texte numérisé et le dictionnaire des formes. Cette structure s'est révélée très efficace pour la réalisation des principales tâches au sein d'un logiciel dont l'ambition se limite au domaine de la recherche lexicométrique.

Classification des items textuels

Dans ce qui suit, on appellera *item* toutes les occurrences des unités que l'on peut rencontrer lors du dépouillement d'un texte (occurrences de formes graphiques, occurrences de ponctuations diverses, etc.), on appellera ces unités elles-mêmes des *articles*.

L'algorithme de segmentation automatique repose sur une classification des items (et articles) du fichier-texte. Cette classification implique qu'au moment de soumettre le texte au processus de segmentation un certain nombre d'ambiguïtés aient été levées et que l'on puisse considérer comme réalisée, par rapport au processus de segmentation, la condition :

un signe de l'enregistrement = un statut

Les items que l'on s'attend à rencontrer appartiennent à deux grandes catégories : les *occurrences de forme* et les *jalons textuels*. Les items de la première catégorie sont les unités dont le décompte motive l'entreprise de la statistique textuelle telle qu'elle a été définie aux deux premiers chapitres.

---

[1] Cette structure a pu être améliorée à la suite de discussions avec des spécialistes du domaine de l'informatique textuelle et avec J. Dendien en particulier.

L'ensemble des jalons textuels se subdivise en deux catégories les *ponctuations* et les *clés*.[1]

Les clés

Les clés permettent d'introduire dans le texte des informations péritextuelles de toutes sortes. Elles sont du type :

<Type1=va11>

Les clés sont introduites entre deux caractères réservés < et >. Le signe "=" sépare le type et le contenu de la clé.

Type1 — indique le *type* de la clé.

val1 — qui peut être précédé par des caractères "blancs" ne faisant pas partie de ce contenu, indique un *contenu* de clé, c'est-à-dire une valeur que l'on assigne à cette clé.

Les ponctuations

La classe "ponctuation" rassemble une classe de jalons textuels qui dépasse largement l'ensemble des signes que l'on regroupe sous cette appellation dans le langage courant.[2]

La liste des signes qui seront considérés comme signes délimiteurs de forme par le programme de segmentation est déterminée par l'utilisateur.[3]

A ces signes délimiteurs de formes s'ajoutent obligatoirement[4] :

— les caractères < et > qui servent à introduire les clés
— le caractère "blanc"
— le caractère "RC" ou retour chariot qui sert à découper le flot d'entrée en lignes.

[1] Le système de codage des clés présenté ici s'inspire très largement de celui présenté dans Lafon et al. (1985).

[2] Bien que ce signe ait une fonction complètement distincte au plan de l'encodage du texte, de celle des signes usuels de ponctuation, le caractère "retour chariot" peut être considéré du point de vue de la segmentation automatique comme un caractère de type "ponctuation".

[3] Le programme Segmentation propose à l'utilisateur une liste contenant les signes de ponctuations les plus courants [. ? ! ; : , " ( ) / ' _ §]. Cette liste est entièrement modifiable par l'utilisateur. Le caractère § peut être utilisé pour marquer le début de chacun des paragraphes du texte de référence.

[4] Ces signes délimiteurs de forme sont rajoutés automatiquement par le programme de segmentation à la liste des caractères délimiteurs choisis par l'utilistateur.

Avec cette convention on peut regrouper en trois classes l'ensemble des jalons textuels que l'on trouve dans un texte :

ponctuation	—	caractère délimiteur de forme (à l'exclusion des signes réservés < et >)
type de clé	—	suite de caractères non-délimiteurs, précédée par le signe < et terminée par le signe "="
contenu de clé	—	suite de caractères non-délimiteurs terminée par le signe >

La segmentation automatique du texte

Le but de l'étape de segmentation automatique du texte est de permettre la construction, à partir d'un fichier texte que l'on appellera *text1*, d'une base textuelle numérisée qui servira de point de départ à l'ensemble des algorithmes de documentation et d'analyse statistique. Cette base est constituée par deux fichiers complémentaires :

text1.dicnum ou *dictionnaire*
et *text1*.textnum ou *fichier texte numérisé*

Le dictionnaire

Le fichier dicnum contient un enregistrement pour chacun des articles du texte à l'exception des articles du type "ponctuation". Les enregistrements correspondant aux articles sont classés en fonction de leur type dans l'ordre suivant :

formes	—	interclassées par ordre lexicométrique[1] à l'intérieur de cette catégorie (i. e. par ordre de fréquence décroissante, l'ordre lexicographique départageant les formes de même fréquence).
types de clés	—	classés par ordre lexicographique
contenus de clés	—	classés dans le même ordre

Chacun de ces enregistrements contient les renseignements suivants :

lexicog.	—	ordre lexicographique de l'article dans la liste ci-dessus (l'ordre lexicographique pour les formes).
fréq.	—	la fréquence de l'article
fgraph.	—	la forme graphique de l'article : forme dans le cas d'une occurrence textuelle, suite de caractères donnant le type ou le contenu d'une clé dans les autres cas.

Le nombre des formes différentes est noté *nbform,* celui des articles *nbarticles*.

[1] Cf. les définitions du chapitre 2.

Le texte numérisé

Le fichier dicnum se présente sous la forme d'une suite de **nbitem** (= nombres des items du texte). Chacun de ces nombres correspond à un des items du texte.

B.1 Tâche n°1 : La numérisation du texte

En fonction de la place mémoire disponible[1] le programme commence par fixer la taille maximale du problème que l'on pourra traiter dans l'environnement considéré.

Les deux principaux paramètres du problème sont *ItemMax* et *ArtMax* respectivement égaux au nombre maximal des items que l'on se propose de traiter et au nombre maximal d'articles que l'on s'attend à trouver dans un texte ItemMax items.

On réserve ensuite les tableaux à une dimension **T**(*ItemMax*), **V**(*ArtMax*) ainsi qu'une série de tableaux de travail. L'un de ces tableaux est muni d'une structure d'arbre binaire dont le noeud élémentaire est structuré de la manière suivante :

structure ***tdic*** *{*

mot :	pointeur sur une zone texte qui contient la forme graphique de l'article;
freq	variable entière contenant le nombre des occurrences de l'article rencontrées depuis le début du processus;
lexicog	rang lexicographique de l'article (qui sera calculé après la fin de la lecture du texte);
lexicom	rang lexicometrique de l'article (idem);
numorg	numéro de l'article dans la liste des articles classés par ordre d'apparition dans le texte;
suivant	pointeur qui pointe sur le noeud suivant;

}

Le tableau **Ftex** est un tableau de caractères dans lequel sont mis bout à bout les suites de caractères qui composent tant les forme graphiques que les types et contenus des clés.

L'adresse du premier de ces caractères est stockée dans la variable *mot* de la structure **tdic**. Le dernier caractère de la forme est suivi d'un caractère spécifique qui permet de le repérer en tant que tel.

La deuxième phase de l'algorithme voit une lecture du fichier texte et une première numérisation item par item. Les items qui se présentent dans le flot d'entrée sont isolés et analysés par la procédure **LireItem** qui calcule pour chacun d'eux :

[1] A titre indicatif, cette structuration des données permet de traiter environ 700 000 occurrences sur une machine disposant de 4MO de mémoire vive.

— le numéro d'item dans le texte depuis le début du texte
— le statut de l'item par rapport au 4 catégories d'articles (occurrence, ponctuation, type de clé, contenu de clé)
— la forme graphique de l'item

Chaque item est ensuite présenté à la racine de l'arbre binaire (ou d'un B-arbre) pour être comparé aux articles déjà stockés dans l'arbre. Les articles sont comparés sur une base lexicographique (aa < ab).

Ce processus permet de calculer un code numérique *CodNum* pour chacun des items entrés. Si l'item considéré correspond à l'occurrence d'un article qui est déjà apparu lors de l'exploitation en cours, sa présentation à l'arbre de stockage permet de retrouver et d'incrémenter la variable qui comptabilise le nombre de ses occurrences. Dans le cas contraire, il s'agit d'un nouvel article dont le numéro de stockage est calculé par incrémentation de la variable qui comptabilise le nombre des objets stockés dans l'arbre.

A la fin de la lecture du fichier texte, l'arbre contient tous les articles présents dans le texte. Pour chaque article on connaît en outre le nombre de ses occurrences dans le texte. Les articles sont numérotés d'après l'ordre de leur apparition dans le texte. On en profite pour affecter ce numéro d'ordre à la variable *numorg*.

Les tris

Le calcul des rangs lexicographiques et lexicométriques de chacun des articles est réalisé à partir du fichier des articles.[1] On commence par remplir *List (*)* un tableau de pointeurs de longueur *nbarticles* dont chacun pointe sur un des noeuds de l'arbre de stockage des articles.

Ce tableau de pointeurs est trié une première fois d'après l'ordre lexicographique de chacun des articles.[2] Les articles correspondant aux formes graphiques se trouvent placés en tête de cette liste du fait de l'adjonction de caractères de poids très élevé devant les chaînes de caractères correspondant aux autres catégories. Les articles du type "type de clé" sont suivis eux-même par les articles du type "contenu de clé". Après cette opération, on peut affecter à la variable *lexicog* une valeur égale au rang de l'article après tri de l'ensemble selon l'ordre lexicographique.

Un second tri par ordre décroissant des pointeurs du tableau *List (*)* d'après la valeur de la variable *freq* de l'article sur lequel ils pointent, les formes de même fréquence étant départagées d'après les valeurs ascendantes de la variable *lexicog* que nous venons de calculer, permet de classer les articles correspondant aux formes graphiques d'après l'ordre lexicométrique. Cette opération terminée, on peut assigner à la variable *lexicog* le numéro d'ordre lexicométrique.

---

[1] Cette manière de procéder constitue un progrès important par rapport aux algorithmes de segmentation qui effectuent les tris sur le fichier des items beaucoup plus volumineux que celui des articles.

[2] Dans la pratique, ce tri est notablement accéléré par une technique de "tri par morceaux". Les articles sont d'abord regroupés en sous-groupes en fonction du premier caractère. On effectue ensuite un tri à l'intérieur de chacun des sous-groupes.

Il est possible alors de passer à l'écriture sur support externe du dictionnaire. Cette écriture se fait en suivant l'ordre lexicométrique dans lequel sont actuellement triés les pointeurs contenus dans le tableau list. Pour chaque article on écrit successivement :

freq	le nombre des occurrences de l'article dans le texte;
lexicog	rang lexicographique de l'article;
mot	suite de caractères composant l'article terminée par un retour chariot.

Un dernier tri de la liste de pointeurs contenue dans le tableau list, effectué d'après les valeurs de la variable *numorg,* replace cette liste dans l'ordre initial.

La numérisation finale

La dernière phase de l'algorithme de segmentation permet de stocker sur support externe le texte numérisé d'après l'ordre lexicométrique des articles du texte. Pour effectuer cette opération il suffit de reprendre le tableau T dans lequel les articles sont numérisés d'après l'ordre de leur apparition dans le texte et de substituer le numéro lexicométrique de chaque article à ce numéro. On trouve le numéro lexicométrique correspondant à l'article dont le numéro d'apparition est égal à i dans la variable *lexicom* du noeud de l'arbre de stockage pointé par l'élément i.

B.2 Tâche n°2 : La recherche de contextes

On regroupe sous cette appellation la recherche de différents sites textuels pour un ensemble d'occurrences correspondant à une *forme-pôle* donnée ou à une liste de ces formes. On construit une telle liste en introduisant l'une des sélections suivantes :

- une forme graphique
- une liste de formes graphiques
- un groupe de caractères constituant le début d'une ou de plusieurs formes présentes dans le texte.
- un groupe de caractères constituant la fin d'une ou de plusieurs formes présentes dans le texte.
- un groupe de caractères présent dans le corps d'une ou de plusieurs formes présentes dans le texte.

Les unités de contexte retenues pour l'ensemble des formes-pôle sélectionnées peuvent être :

- une ligne de contexte

— un nombre fixe de lignes de contexte pour chacune des occurrences de la liste des formes-pôle.

— un inventaire distributionnel des segments répétés après la forme-pôle.

Les contextes peuvent-être triés selon deux types d'arguments dont chacun peut être choisi comme argument majeur du tri ou comme argument mineur :

— en fonction des formes graphiques

— en fonction des valeurs d'une clé sélectionnée dans le texte

Pour une forme graphique donnée, les contextes peuvent être également triés, par rapport à la forme-pôle :

— en fonction de l'ordre lexicographique des formes qui précèdent chacune des occurrences de cette forme.

— en fonction de l'ordre lexicographique des formes qui suivent chacune des occurrences de cette forme.

— en fonction de l'ordre d'apparition des occurrences de cette forme dans le texte.

B.3 Tâche n°3 : Le calcul de la gamme des fréquences.

Le calcul de la gamme des fréquences peut être réalisé à partir du seul dictionnaire dans lequel les formes qui ont une même fréquence sont classées côte à côte. En commençant par la fréquence la plus élevée (ou au contraire par la fréquence 1) on calcule par simple cumul les effectifs V_i qui correspondent à chacune des fréquences pour i variant de 1 à fmax.

B.4 Tâche n°4 : La construction des Tableaux Lexicaux.

Le tableau lexical est déterminé par deux paramètres fixés par l'utilisateur :

— *fmin* : le seuil minimal retenu pour la sélection des formes qui correspondront aux lignes du tableau lexical (on ne retiendra que les formes dont la fréquence est égale ou supérieure à ce seuil).

— *Sxx* : un type correspondant à une clé (le tableau lexical comptera une colonne pour chacune des valeurs différentes prises par le contenu de ce type de clé dans le fichier texte).

Une fois ces valeurs fixées, on peut calculer *x1*, le nombre des lignes du tableau lexical qui est égal au nombre des formes dont la fréquence est au moins égale à *fmin* dans le corpus. Remarquons que, par suite de la définition de l'ordre lexicométrique que nous

avons adoptée, les numéros lexicométriques de toutes les formes dont la fréquence est inférieure à ce seuil sont supérieurs au numéro lexicométrique de la dernière forme de fréquence est égale au seuil retenu. Appelons ce numéro *fder*.

Une exploration des contenus existants dans le texte pour la clé sélectionnée permet de fixer *x2,* le nombre des colonnes du tableau égal au nombre des valeurs différentes prise par la variable "contenu de la clé sélectionnée". Les différentes valeurs prises par cette variable sont triées par ordre lexicographique ascendant et numérotées dans cet ordre. Cette numérotation permet d'affecter à chaque code un numéro de partie.

On peut alors réserver une zone mémoire de taille (*x1* x *x2*) qui contiendra le tableau que l'on se propose de calculer. Le calcul de ce tableau s'effectue en une seule lecture du fichier texte numérisé. En parcourant ce tableau à partir de la première de ses cases on commence par trouver la première occurrence de la clé sélectionnée et le premier contenu de cette clé. Ce contenu nous renvoie à un numéro de partie d'après la table établie précédemment. Ce numéro s'appellera le "numéro de partie courant".

La poursuite de ce traitement pour chacune des cases du tableau "tnum" amène à trois situations différentes :

— le code tnum[k] est :

  soit le code d'une forme dont la fréquence est inférieure au seuil retenu (i.e. un code supérieur à *fder*) ;

  soit le code d'une ponctuation ; soit le code d'un type ou d'un contenu de clé correspondant à une autre clé que la clé sélectionnée et dans ce cas on ignore la case tnum[k] pour passer au traitement du code contenu dans la case tnum[k+1].

— le code tnum[k] est le code de la clé sélectionnée. On lit alors dans la case tnum[k+1] la valeur du contenu de la clé qui permet de mettre à jour le code de partie courant (puisque l'on passe au traitement des occurrences situées sans une autre partie du texte).

— le code tnum[k] est le code d'une occurrence de la forme *forme* dont la fréquence est supérieure ou égale au seuil retenu (i.e. un code inférieur à *fder*). Si ce code est égal à i, on ajoute une unité à la case [i, numéro de partie courant] du tableau lexical que l'on est en train de calculer.

0	de	31	44	52	112	34	72	136	36	48
1	l	35	36	57	82	29	51	82	27	30
2	la	22	28	42	82	27	43	118	25	24
3	les	24	24	32	48	21	33	65	13	26
4	le	36	20	33	61	13	28	71	8	14
5	d	18	20	25	47	14	26	55	7	16
6	pas	15	11	21	43	10	18	72	9	11
7	avenir	21	22	31	34	12	30	28	12	8
8	chômage	21	18	13	35	5	9	58	6	10

Tableau B.1
Exemple de tableau lexical

Lorsque ce processus est achevé on a calculé le tableau lexical des formes de fréquence supérieure ou égale au seuil fixé. On peut éditer ce tableau en faisant précéder la ventilation de chacune des formes sélectionnées par son numéro lexicométrique et son libellé, que l'on trouvera dans le dictionnaire, comme dans le tableau B.1 ci-dessus.

B.5 Tâche n°5 : Tableaux des segments répétés

Les calculs portant sur la ventilation des segments répétés dans les parties d'un corpus de textes permettent de compléter les analyses effectuées à partir des tableaux de formes graphiques (tableaux lexicaux).[1]

Cependant les segments répétés d'un texte sont avant tout caractérisés par leur énorme redondance.

Pour réduire le volume des segments répétés, on écarte de la liste des segments ceux d'entre eux qui sont des segments contraints (i.e. qui sont toujours précédés ou suivis par une même forme et qui entrent donc dans la composition de segments plus longs).

Pour certaines applications statistiques, il est en outre possible de ne considérer que les segments dont la fréquence dépasse un certain seuil de répétition (on abaissera ce seuil à 2 occurrences si l'on désire obtenir la liste de tous les segments répétés dans le texte).

On peut distinguer pour la réalisation de la tâche n°5 deux sous-tâches, qui interviennent alternativement lors de la construction du TSR d'un corpus.

Ces sous-tâches sont :

5-1 — le repérage des segments non contraints dans le corpus qui commencent par une forme graphique donnée et dont la fréquence dépasse un seuil fixé.

5-2 — le calcul de la ventilation dans les parties du corpus de l'ensemble des occurrences d'un segment ainsi repéré.

Le repérage de l'ensemble des segments non-contraints débutant par une forme donnée, pour chacune des formes graphiques dont la fréquence est égale ou supérieure à un seuil fixé réalise l'ensemble de la tâche n°5.

B.5.1 Sous-tâche 5.1 : Repérage des segments répétés non contraints

Pour une forme graphique donnée *Form1* et un seuil de fréquence *Sfr* fixé, le but de la sous tâche 5.1 consiste donc dans le repérage de l'ensemble des segments non-contraints débutant par *Form1* dont la fréquence est au moins égale à *Sfr* occurrences.

[1] Cf. chapitres 2 et 5.

Ainsi définie, cette sous tâche devient très simple à réaliser. On commence par définir une liste de pointeurs list1[k] qui pointent chacun sur une des occurrences de la forme *Form1*.[1]

Dans un deuxième temps ces pointeurs sont triés en fonction de l'ordre lexicographique des formes qui suivent la forme *Form1*.[2]

A la fin de ce tri les segments répétés éventuels qui commencent par la forme *Form1* se trouvent placés les uns à la suite des autres dans l'ordre lexicométrique des segments répétés (i.e. les segments sont classés en fonction de l'ordre lexicographique de la première forme, départagés en cas d'égalité par l'ordre lexicographique de la forme qui suit et ainsi de suite) comme c'est le cas dans l'exemple ci-dessous.[3]

A B
A B
A B C
A B C A C
A B C A C D
A B C A C D
A C B A

Dans le processus de comparaison des segments en vue de compter les occurrences de segments répétés, on procède par comparaison des formes graphiques situées à chacune des positions du segment.

Si le tri effectué nous assure que les segments répétés identiques se trouvent bien côte à côte, il reste à repérer, en parcourant la liste des segments commençant par la forme *Form1*, les segments qui marquent la fin d'un groupe d'occurrences relatives à un même segment de longueur *L1*.

Remarquons qu'à l'intérieur des occurrences des segments commençant par *L* formes identiques, les occurrences d'un segment répété comportant *L+1* formes identiques constituent un ensemble de segments consécutifs après le tri lexicographique des segments.

Comparaison de segments

Les cases des tableaux SEGn[k] sont remplies par des valeurs relatives à des occurrences de la forme *Form1*. Pour le repérage des segments répétés dont la fréquence est supérieure ou égale à *Sfr,* on ne s'intéresse qu'aux séquences ne comprenant aucune forme de fréquence inférieure à ce seuil ne chevauchant pas en outre de délimiteurs de séquences.

[1] La taille de cette liste est égale, au maximum, à *fmax* la fréquence de la forme la plus fréquente.
[2] Remarquons que cet ordre correspond exactement à celui dans lequel apparaîssent les contextes-lignes lors de la réalisation d'une concordance triée sur le contexte qui suit.
[3] Comme dans les exemples du chapitre 2 relatifs au corpus **P**, les lettres capitales symbolisent ici chacune une forme graphique.

Les séquences à comparer sont donc constituées par des suites de codes correspondant au numéro lexicométrique de formes de fréquence supérieure au seuil retenu. On sélectionne un code particulier noté *VIDE* pour remplir les autres cases du tableau SEGn[k].

Par rapport à une occurrence de forme graphique, on appellera dans ce qui suit **forme suivante**, soit la forme *VIDE,* soit la première forme graphique située après cette forme, en négligeant les jalons textuels et signes de ponctuation non-délimiteurs de séquence situés entre les deux formes. Les mêmes conventions s'appliquent pour la définition de la **forme précédente**.

Les séquences débutant par la forme *Form1* qui vont être soumises à la comparaison sont rangées dans les tableaux SEGn[k] de la manière suivante: on range dans la case SEGn[0] le numéro lexicométrique de la forme précédant *Form1*, en respectant les conventions ci-dessus.

La case SEGn[1] contient le numéro lexicométrique de la forme *Form1* qui sert de pivot à la recherche des segments.

Les cases suivantes SEGn[2], SEGn[3],..., SEGn[longmax] contiennent soit les numéros lexicométriques des formes qui suivent soit le code *VIDE* (à partir d'une certaine position).

Recherche de l'accident

La comparaison entre deux segments successifs permet de déterminer la longueur du sous-segment répété le plus long qu'ils ont en commun, c'est-à-dire le nombre des premières occurrences qui sont identiques pour les deux segments.

Nous appellerons *accident* la position pour laquelle une occurrence du second segment ne correspond pas, pour la première fois à l'occurrence du segment précédent. Ainsi, par exemple, dans la comparaison des deux segments

```
    1 2 3 4 5 6 7
    A B C D E F G
et  A B C D E G H
```

nous dirons que l'accident se situe en position numéro 6.

Recherche des segments répétés

Le tableau SEGc[n], qui contient le "segment courant", est initialisé par les valeurs du segment placé en tête par le tri lexicographique des segments. On gère parallèlement un tableau FreqSeg[n] qui contient pour n variant de 2 à *longmax* la fréquence courante des segments de longueur 2 à *longmax*.

On considère ensuite le segment placé immédiatement après par le tri lexicographique. La détermination d'un accident en position *x* par rapport au segment précédent indique que l'on en a terminé avec les occurrences du sous-segment précédent (dont la longueur est égale à *x*). Si la fréquence des segments répétés ainsi répertoriés dépasse le seuil *Sfr,* nous avons repéré un segment répété dont il nous reste à calculer la ventilation dans les parties du corpus.

B.5.2 Sous-tâche 5.2 : Calcul de la ventilation des segments répétés

Le calcul de la ventilation des occurrences du segment répété s'opère à partir du numéro de chacune des occurrences de la forme *Form1*. Ce problème suppose que l'on ait un moyen de déterminer le numéro de la partie à laquelle appartient une occurrence du texte. Plusieurs possibilités s'offrent pour réaliser cette tâche dont la plus simple consiste à construire un tableau récapitulatif indiquant les cases du tableau *tnum* qui correspondent à des changements de partie.

Glossaire pour la statistique textuelle

NB : Les astérisques renvoient à une entrée de ce même glossaire. Les abréviations qui suivent entre parenthèses précisent le domaine auquel s'applique plus particulièrement la définition.

Abréviations :

ac	Analyse factorielle des correspondances
acm	Analyse des correspondances multiples
cla	Classification
sp	Méthode des Spécificités
sr	Analyse des segments répétés
ling	Linguistique
stat	Statistique
sa	Segmentation automatique

accroissement spécifique - (sp) spécificité* calculée pour une partie d'un corpus par rapport à une partie antérieure

algorithme - ensemble des règles opératoires propres à un calcul.

analyse factorielle (stat) - famille de méthodes statistiques d'analyse multidimensionnelle, s'appliquant à des tableaux de nombres, qui visent à extraire des "facteurs" résumant approximativement par quelques séries de nombres l'ensemble des informations contenues dans le tableau de départ.

analyse des correspondances (stat)- méthode d'analyse factorielle s'appliquant à l'étude de tableaux à double entrée composés de nombres positifs. L'AC est caractérisée par l'emploi d'une distance (ou métrique) particulière dite distance du chi-2 (ou χ^2).

analyse des correspondances multiples (stat) **-** méthode d'analyse des correspondances s'appliquant à l'étude de tableaux disjonctifs complets. Tableaux binaires dont les lignes sont des individus ou observations et les colonnes la juxtaposition des modalités de réponse à des questions (les modalités de réponse à une question s'excluant mutuellement).

caractère (sa) - signe typographique utilisé pour l'encodage du texte sur un support lisible par l'ordinateur.

caractères délimiteurs / **non-délimiteurs** (sa) - distinction opérée sur l'ensemble des caractères, qui entrent dans la composition du texte permettant aux procédures informatisées de segmenter le texte en occurrences* (suite de caractères non-délimiteurs bornée à ses extrémités par des caractères délimiteurs).

On distingue parmi les caractères délimiteurs:

- les caractères délimiteurs d'occurrence (encore appelés "**délimiteurs de forme**") qui sont en général : le blanc, les signes de ponctuation usuels, les signes de préanalyse éventuellement contenus dans le texte.

- les caractères **délimiteurs de séquence** : sous-ensemble des délimiteurs d'occurrence correspondant, en général, aux ponctuations faibles et fortes contenues dans la police des caractères.

- les caractères **séparateurs de phrase** : (sous-ensemble des délimiteurs de séquence) qui correspondent, en général, aux seules ponctuations fortes.

classification (stat) - technique statistique permettant de regrouper des individus ou observations entre lesquels a été définie une distance.

classification hiérarchique (cla) - technique particulière de classification produisant par agglomération progressive des classes ayant la propriété d'être, pour deux quelconques d'entre-elles, soit disjointes, soit incluses.

concordance (sa) - l'ensemble de lignes de contexte se rapportant à une même forme-pôle.

contribution absolue (ou contribution) - (ac) contribution apportée par un élément au facteur . Pour un facteur donné, la somme des contributions sur les éléments de chacun des ensembles mis en correspondance est égale à 100.

contribution relative (ou cosinus carré) - (ac) contribution apportée par le facteur à un élément. Pour un élément donné, la somme des contributions relatives sur l'ensemble des facteurs est égale à 1.

cooccurrence (sa) - (une c.) - présence simultanée, mais non forcément contiguë, dans un fragment de texte (séquence, phrase, paragraphe, voisinage d'une occurrence, partie du corpus etc.) des occurrences de deux formes données.

corpus (ling) - ensemble limité des éléments (énoncés) sur lesquels se base l'étude d'un phénomène linguistique.

(lexicométrie) ensemble de textes réunis à des fins de comparaison; servant de base à une étude quantitative.

délimiteurs de séquence - (sa) sous-ensemble des caractères délimiteurs* de forme* correspondant aux ponctuations faibles et fortes (en général - le point, le point d'interrogation, le point d'exclamation, la virgule, le point-virgule, les deux points, les guillemets, les tirets et les parenthèses).

dendrogramme - (cla) représentation graphique d'un arbre de classification hiérarchique, mettant en évidence l'inclusion progressive des classes.

discours/langue - La langue est un ensemble virtuel qui ne peut être appréhendé que dans son actualisation orale ou écrite; "discours" est un terme commode qui recouvre les deux domaines de cette actualisation.

distance du chi-2 - distance entre profils* de fréquence utilisée en analyse des correspondances* et dans certains algorithmes* de classification*.

éditions de contextes (sa) - éditions de type concordanciel dans lesquelles les occurrences d'une forme sont accompagnées d'un fragment de contexte pouvant contenir plusieurs lignes de texte autour de la forme-pôle. La longueur de ce

contexte est définie en nombre d'occurrences avant et après chaque occurrence de la forme-pôle.

éléments d'un segment (sr) - chacune des formes correspondant aux occurrences qui entrent dans sa composition. ex : A, B, C sont respectivement les premier, deuxième et troisième éléments du segment ABC.

éléments actifs- (ac ou acm) ensemble des éléments servant de base au calcul des axes factoriels, des valeurs propres relatives à ces axes et des coordonnées factorielles.

éléments supplémentaires (ou illustratifs)- (ac ou acm) ensemble des éléments ne participant pas aux calculs des axes factoriels, pour lesquels on calcule des coordonnées factorielles qui auraient été affectées à une forme ayant la même répartition dans le corpus mais participant à l'analyse avec un poids négligeable.

énoncé/énonciation - (ling) à l'intérieur du texte un ensemble de traces qui manifestent l'acte par lequel un auteur a produit ce texte.

expansion contrainte -(sr) terme dont les occurrences constituent chaque fois les expansions (du même côté) d'un même terme ayant plusieurs occurrences dans le corpus.

expansion d'un segment - (sr) segment situé immédiatement avant (expansion gauche) ou après (expansion droite) d'un segment donné, non séparé de ce segment par un délimiteur de séquence.

expansion récurrente d'un terme - terme dont les occurrences constituent plusieurs fois l'expansion des occurrences d'un terme donné.

facteur- (ac ou acm) variables artificielles construites par les techniques d'analyse factorielle permettant de résumer (de décrire brièvement) les variables actives initiales.

forme- (sa) ou "**forme graphique**" archétype correspondant aux occurrences* identiques dans un corpus de textes, c'est-à-dire aux occurrences composées strictement des mêmes caractères non-délimiteurs d'occurrence.

forme banale - (sp) pour une partie du corpus donnée, forme ne présentant aucune spécificité (ni positive ni négative) dans cette partie .

forme caractéristique - (d'une partie) synonyme de spécificité positive*.

forme commune - forme attestée dans chacune des parties du corpus.

forme originale- (pour une partie du corpus) forme trouvant toutes ses occurrences dans cette seule partie.

fréquence (sa) - (d'une unité textuelle) le nombre de ses occurrences dans le corpus.

fréquence d'un segment (sr) - (ou d'une polyforme) le nombre des occurrences de ce segment, dans l'ensemble du corpus.

fréquence maximale (sa) - fréquence de la forme la plus fréquente du corpus (en français, le plus souvent, la préposition "de").

fréquence relative (sa) - la fréquence d'une unité textuelle dans le corpus ou dans l'une de ses parties, rapportée à la taille du corpus (resp. de cette partie).

gamme des fréquences (sa) - suite notée V_k, des effectifs correspondant aux formes de fréquence k, lorsque k varie de 1 à la fréquence maximale.

hapax - gr. hapax (legomenon), "chose dite une seule fois".

(sa) forme dont la fréquence est égale à un dans le corpus (hapax du corpus) ou dans une de ses parties (hapax de la partie).

identification - (stat, ling, sa) reconnaissance d'un seul et même élément à travers ses multiples emplois dans des contextes et dans des situations différentes.

index - (sa) liste imprimée constituée à partir d'une réorganisation des formes et des occurrences d'un texte, ayant pour base la forme graphique et permettant de regouper les références* relatives à l'ensemble des occurrences d'une même forme.

index alphabétique (sa) - index* dans lequel les formes-pôles* sont classées selon l'ordre lexicographique* (celui des dictionnaires).

index hiérarchique (sa) - index* dans lequel les formes-pôles* sont classées selon l'ordre lexicométrique*.

index par parties - ensemble d'index (hiérarchiques ou alphabétiques) réalisés séparément pour chaque partie d'un corpus.

item de réponse - (ou modalité de réponse) élément de réponse préétabli dans une question fermée.

lemmatisation - regroupement sous une forme canonique (en général à partir d'un dictionnaire) des occurrences du texte. En français, ce regroupement se pratique en général de la manière suivante :

- les formes verbales à l'infinitif,
- les substantifs au singulier,
- les adjectifs au masculin singulier,
- les formes élidées à la forme sans élision.

lexical - (ling) qui concerne le lexique* ou le vocabulaire*.

lexicométrie ensemble de méthodes permettant d'opérer des réorganisations formelles de la séquence textuelle et des analyses statistiques portant sur le vocabulaire* d'un corpus de textes.

lexique - (ling) ensemble virtuel des mots d'une langue.

longueur (sa) - (d'un corpus, d'une partie de ce corpus, d'un fragment de texte, d'une tranche, d'un segment, etc.) le nombre des occurrences contenues dans ce corpus (resp. : partie, fragment, etc.). Synonyme de taille.

On note: T la longueur du corpus; t_j celle de la partie (ou tranche) numéro j du corpus.

longueur d'un segment (sr) - le nombre des occurrences entrant dans la composition de ce segment.

noyaux factuels - (cla) classes d'une partition* d'un ensemble d'observations synthétisant une batterie de descripteurs objectifs (sexe, âge, profession, etc.).

occurrence (sa) **-** suite de caractères non-délimiteurs bornée à ses extrémités par deux caractères délimiteurs* de forme.

ordre lexicographique -

_ pour les formes graphiques :

l'ordre selon lequel les formes sont classées dans un dictionnaire.

NB : Les lettres comportant des signes diacrisés sont classées au même niveau que les mêmes caractères non diacrisés, le signe diacritique n'intervenant que dans les cas d'homographie complète. Dans les dictionnaires, on trouve par exemple, rangées dans cet ordre, les formes : *mais, maïs, maison, maître* .

_ pour les polyformes:

ordre résultant d'un tri des polyformes par ordre lexicographique sur la première composante, les polyformes commençant par une même forme graphique sont départagées par un tri lexicographique sur la seconde, etc.

ordre lexicométrique (sa) **-**

_ pour les formes graphiques :

ordre résultant d'un tri des formes du corpus par ordre de fréquences décroissantes; les formes de même fréquence sont classées par ordre lexicographique.

_ pour les polyformes:

ordre résultant d'un tri par ordre de longueur décroissante des segments, les segments de même longueur sont départagés par leur fréquence, les segments ayant même longueur et même fréquence par l'ordre lexicographique.

paradigme- (ling) ensemble des termes qui peuvent figurer en un point de la chaîne parlée.

paradigmatique- (sa) qui concerne le regroupement en série des unités textuelles, indépendamment de leur ordre de succession dans la chaîne écrite.

partie - (d'un corpus de textes) fragment de texte correspondant aux divisions naturelles de ce corpus ou à un regroupement de ces dernières.

partition - (d'un corpus de textes) division d'un corpus en *parties* constituées par des fragments de texte consécutifs, n'ayant pas d'intersection commune et dont la réunion est égale au corpus.

(d'un ensemble, d'un échantillon) division d'un ensemble d'individus ou d'observations en *classes* disjointes dont la réunion est égale à l'ensemble tout entier.

partition longitudinale - (sa) partition d'un corpus en fonction d'une variable qui définit un ordre sur l'ensemble des parties

périodisation (sa) - regroupement des parties naturelles du corpus respectant l'ordre chronologique d'écriture, d'édition ou de parution des textes réunis dans le corpus.

phrase - (sa) fragment de texte compris entre deux séparateurs* de phrase.

places - (sa) pour un texte comptant T occurrences, suite de T objets correspondant chacun à une des occurrences du texte, éventuellement séparés par des délimiteurs* de séquence correspondant aux ponctuations du texte de départ.

polyforme (sr) - archétype des occurrences d'un segment; suite de formes non séparées par un séparateur de séquence, qui n'est pas obligatoirement attestée dans le corpus.

ponctuation - Système de signes servant à indiquer les divisions d'un texte et à noter certains rapports syntaxiques et/ou conditions d'énonciation.

(*sa*) caractère (ou suite de caractères) correspondant à un signe de ponctuation.

post-codage - opération manuelle qui consiste à repérer les principales catégories de réponses libres sur un sous-échantillon de réponses, puis à fermer la question ouverte correspondante, en affectant toutes les réponses à ces catégories.

pourcentages d'inertie - (ac ou acm) quantités proportionnelles aux valeurs propres* dont la somme est égale à 100. Notées τ_α.

profil - (stat et ac) (d'une ligne ou d'une colonne d'un tableau à double entrée) vecteur constitué par le rapport des effectifs contenus sur cette ligne (resp. colonne) à la somme des effectifs que contient la ligne (resp. la colonne).

question fermée - question dont les seules réponses possibles sont proposées explicitement à la personne interrogée.

question ouverte - question posée sans grille de réponse préétablie, dont la réponse peut être numérique (ex: *Quelles doivent être, selon vous, les ressources minimum d'une famille ayant trois enfants de moins de 16 ans?*), ou textuelle (ex: *pouvez-vous justifier votre choix?*).

références (sa) - système de coordonnées numériques permettant de repérer dans le texte d'origine chacune des occurrences issues de la segmentation (ex : le tome, la page, la ligne, la position de l'occurrence dans la ligne) ou de situer rapidement cette occurrence parmi des catégories prédéfinies (auteur, année de parution, citation, mise en valeur, etc.).

répartition (sa) - (des occurrences d'une forme dans les parties du corpus) nombre des parties du corpus dans lesquelles cette forme est attestée.

réponse modale - (d'une classe d'individus, d'une partie* de corpus*) réponse sélectionnée en fonction de son caractère représentatif d'une classe ou d'une partie en général à partir des formes* caractéristiques qu'elle contient.

segment - (sr) toute suite d'occurrences consécutives dans le corpus et non séparées par un séparateur* de séquence est un segment du texte.

segment répété (sr) - (ou polyforme répétée) suite de forme dont la fréquence est supérieure ou égale à 2 dans le corpus.

segmentaire - (sr) ensemble des termes* attestés dans le corpus.

segmentation - opération qui consiste à délimiter des unités minimales* dans un texte.

segmentation automatique - ensemble d'opérations réalisées au moyen de procédures informatisées qui aboutissent à découper, selon des règles prédéfinies, un texte

stocké sur un support lisible par un ordinateur en unités distinctes que l'on appelle des unités minimales*.

séparateurs de phrases - (sa) sous-ensemble des caractères délimiteurs* de séquence* correspondant aux seules ponctuations fortes (en général : le point, le point d'interrogation, le point d'exclamation).

séquence - (sa) suite d'occurrences du texte non séparées par un délimiteur* de séquence.

seuil - (stat) quantité arbitrairement fixée au début d'une expérience visant à sélectionner parmi un grand nombre de résultats, ceux pour lesquels les valeurs d'un indice numérique dépassent ce seuil (de fréquence, en probabilité, etc.).

seuil d'absence spécifique - (sp) pour un seuil fixé, pour une partie du corpus fixée, fréquence A(j) pour laquelle toute forme de fréquence supérieure à A(j) dans le corpus et absente dans la partie j est spécifique (négative) pour la partie j.

seuil de présence spécifique - (sp) pour un seuil fixé, pour une partie du corpus fixée, fréquence B(j) pour laquelle toute forme de fréquence inférieure à A(j) dans le corpus et présente dans la partie j est spécifique (positive) pour la partie j.

sous-fréquence (sa) - (d'une unité textuelle dans une partie, tranche, etc.) nombre des occurrences de cette unité dans la seule partie (resp. tranche, etc.) du corpus.

sous-segments (sr) - pour un segment donné, tous les segments de longueur inférieure et compris dans ce segment sont des sous-segments. ex : AB et BC sont deux sous-segments du segment ABC.

spécificité chronologique - (sp) spécificité* portant sur un groupe connexe de parties d'un corpus muni d'une partition longitudinale*.

spécificité positive - (sp) pour un seuil de spécificité fixé, une forme i et une partie j données, la forme i est dite spécifique positive de la partie j (ou forme caractéristique* de cette partie) si sa sous-fréquence est "anormalement élevée" dans cette partie. De façon plus précise, si la somme des probabilités calculées à partir du modèle hypergéométrique pour les valeurs égales ou supérieures à la sous-fréquence constatée est inférieure au seuil fixé au départ.

spécificité négative - (sp) pour un seuil de spécificité fixé, une forme i et une partie j données, la forme i est dite spécifique négative de la partie j si sa sous-fréquence est anormalement faible dans cette partie. De façon plus précise, si la somme des probabilités calculées à partir du modèle hypergéométrique pour les valeurs égales ou inférieures à la sous-fréquence constatée est inférieure au seuil fixé au départ.

stock distributionnel du vocabulaire - (d'un fragment de texte) le vocabulaire* de ce fragment assorti de comptages de fréquence pour chacune des formes entrant dans sa composition.

syntagmatique- (sa) qui concerne le regroupement des unités textuelles, selon leur ordre de succession dans la chaîne écrite.

syntagme- (ling) groupe de mots en séquence formant une unité à l'intérieur de la phrase.

tableau de contingence (stat) - synonyme de tableau de fréquences ou de tableau croisé: tableau dont les lignes et les colonnes représentent respectivement les modalités

de deux questions (ou deux variables nominales) , et dont le terme général représente le nombre d'individus correspondant à chaque couple de modalités.

tableau lexical entier (TLE) - tableau à double entrée dont les lignes sont constituées par les ventilations* des différentes formes dans les parties du corpus. Le terme générique *k(i,j)* du TLE est égal au nombre de fois que la forme *i* est attestée dans la partie *j* du corpus. Les lignes du TLE sont triées selon l'ordre lexicométrique* des formes correspondantes.

tableau des segments répétés (TSR) - tableau à double entrée dont les lignes sont constituées par les ventilations* des segments répétés dans les parties du corpus. Les lignes du TSR sont triées selon l'ordre lexicométrique* des segments. (i.e. longueur décroissante, fréquence décroissante, ordre lexicographique).

tableau lexical- tableau à double entrée résultant du TLE par suppression de certaines lignes (par exemple celles qui correspondent à des formes dont la fréquence est inférieure à un seuil donné).

taille- (sa) (d'un corpus) sa longueur* mesurée en occurrences (de formes simples).

terme - (sr) nom générique s'appliquant à la fois aux formes* et aux polyformes*. Dans le premier cas on parlera de termes de longueur 1. Les polyformes sont des termes de longueur 2,3, etc.

termes contraints / termes libres - Un terme S1 est contraint dans un autre terme S2 de longueur supérieure si toutes ses occurrences* sont des sous-segments* de segments correspondant à des occurrences du segment S2. Si au contraire un terme possède plusieurs expansions distinctes, qui ne sont pas forcément récurrentes, c'est un terme libre.

unités minimales (pour un type de segmentation) - unités que l'on ne décompose pas en unités plus petites pouvant entrer dans leur composition (ex : dans la segmentation en formes graphiques les formes ne sont pas décomposées en fonction des caractères qui les composent).

valeur modale - (stat) valeur pour laquelle une distribution atteint son maximum.

valeurs propres - (ac ou acm) quantités permettant de juger de l'importance des facteurs successifs de la décomposition factorielle. La valeur propre notée λ_α. mesure la dispersion des éléments sur l' axe.α.

valeurs-tests - (ac ou acm) quantités permettant d'apprécier la signification de la position d'un élément supplémentaire* (ou illustratif) sur une axe factoriel. Brièvement, si une valeur test dépasse 2 en valeur absolue, il y a 95 chances sur 100 que la position de l'élément correspondant ne puisse être due au hasard.

variables actives - variables utilisées pour dresser une typologie, soit par analyse factorielle, soit par classification. Les typologies dépendent du choix et des poids des variables actives, qui doivent de ce fait constituer un ensemble homogène.

variables supplémentaires (ou illustratives) - variables utilisées a posteriori pour illustrer des plans factoriels ou des classes. Une variable supplémentaire peut-être considérée comme une variable active munie d'un poids nul.

variables de type T - variable dont la fréquence est à peu près proportionnelle à l'allongement du texte. (ex : la fréquence maximale)

variables de type V- variable dont l'accroissement a tendance à diminuer avec l'allongement du texte (ex : le nombre des formes, le nombre des hapax).

ventilation (sa) - (des occurrences d'une unité dans les parties du corpus) La suite des n nombres (n = nombre de parties du corpus) constituée par la succession des sous-fréquences* de cette unité dans chacune des parties, prises dans l'ordre des parties.

vocabulaire (sa) - ensemble des formes* attestées dans un corpus de textes.

vocabulaire commun - (sa) l'ensemble des formes attestées dans chacune des parties du corpus.

vocabulaire de base - (sp) ensemble des formes du corpus ne présentant, pour un seuil fixé, aucune spécificité (négative ou positive) dans aucune des parties , (i.e. l'ensemble des formes qui sont "banales" pour chacune des parties du corpus).

vocabulaire original- (sa) (pour une partie du corpus) l'ensemble des formes* originales* pour cette partie.

voisinage d'une occurrence - (sa) pour une occurrence donnée du texte, tout segment (suite d'occurrences consécutives, non séparées par un délimiteur de séquence) contenant cette occurrence.

voisinages d'une forme - (sa) ensemble constitué par les voisinages de chacune des occurrences correspondant à la forme donnée.

Références bibliographiques

Abi Farah A. (1988) - Reconnaissance de l'auteur d'un texte d'après les caractères utilisés, *Les Cahiers de l'Analyse des Données*, XIII, n°1, p 95-96.

Achard P. (1993) - *La sociologie du langage*, Que-sais-je ? PUF, Paris.

Adams L. L.(1975) - *A statistical analysis of the book of Isaiah in relation to the Isaiah problem,* Brigham Young University, Provo.

Aitchison J., Aitken C. G. G. (1976) - Multivariate binary discrimination by the kernel method, *Biometrika,* 63, p 413-420.

Akuto H. (Ed.) (1992) - *International Comparison of Dietary Cultures*, Nihon Keizai Shimbun, Tokyo.

Akuto H., Lebart L. (1992) - Le repas idéal. Analyse de réponses libres en anglais, français, japonais, *Les Cahiers de l'Analyse des Données*, vol XVII, n°3, Dunod, Paris, p 327-352.

Aluja Banet T., Lebart L. (1984) - Local and partial principal component analysis and correspondence analysis, *COMPSTAT, Proceedings in Computational Statistics*, Physica Verlag, Vienna, p 113-118.

ASU (1992) - *La qualité de l'information dans les enquêtes*, Dunod, Paris.

Babeau A., Lebart L. (1984) - Les conditions de vie et les aspirations des Français, *Futuribles,* Avril, p 37-51.

Bardin L. (1989) - *L'analyse de contenu,* PUF, Paris.

Bartell B.T., Cottrell G.W., Belew R.K. (1992) - Latent semantic indexing is an optimal special case of multidimensional scaling, *Proceedings of the 15th Int. ACM-SIGIR Conf. on Res. and Dev. in Information Retrieval*, Belkin N and al. Ed., p 161-167, ACM Press, New York.

Barthélémy J.P., Luong X. (1987) - Sur la typologie d'un arbre phylogénétique : Aspects théoriques, algorithmes et applications à l'analyse de données textuelles, *Math. Sci. Hum.* n°100, p 57-80.

Baudelot C. (1988) - Confiance dans l'avenir et vie réussie (Réponse à un questionnaire), *Recherches économiques : études en hommage à Edmond Malinvaud*, Economica et EHESS, Paris.

Bécue M. (1988) - Characteristic repeated segments and chains in textual data analysis, *COMPSTAT, 8th Symposium on Computational Statistics*, Physica Verlag, Vienna.

Bécue M. (1989) - *Un sistema informatico para el analisis de datos textuales*, Tesis. Facult. d'Informatica, Univ. Politecnica de Catalunya, Barcelona.

Becue M., Peiro R. (1993) - Les quasi-segments pour une classification automatique des réponses ouvertes, in *Actes des 2ndes Journées Internationales d'analyse des données textuelles*, (Montpellier), ENST, Paris, p 310-325.

Behrakis T., Nicolaidis E. (1990) - Typologie des prologues des livres grecs de sciences édités de 1730 à 1820 : Humanisme et esprit des Lumières, *Les Cahiers de l'Analyse des Données*, p 9-20.

Belson W.A., Duncan J.A.(1962) - A Comparison of the check-list and the open response questioning system, *Applied Statistics* n°2, p 120-132.

Benveniste E. (1966) - *Problèmes de linguistique générale*, Gallimard, Paris.

Benzécri J.-P.(1964) - *Cours de linguistique mathématique*, Document mimeographié, Faculté des Sciences de Rennes (cf. aussi Benzécri et al., 1981a).

Benzecri J.-P.(1977) - Analyse discriminante et analyse factorielle, *Les Cahiers de l'Analyse des Données*, II, n °4, p 369-406.

Benzécri J.-P. & coll. (1973) - *La taxinomie*, Vol. I ; *L'analyse des correspondances,* Vol. II, Dunod, Paris.

Benzécri J.-P. (1982) - *Histoire et préhistoire de l'analyse des données,* Dunod, Paris.

Benzécri J.-P.& coll. (1981a) - *Pratique de l'analyse des données,* tome 3, Linguistique & Lexicologie, Dunod , Paris.

Benzécri J.-P. (1991a) - Typologies de textes grecs d'après les occurrences des formes des mots-outil, *Les Cahiers de l'Analyse des Données*, XVI, n°1, p 61-86.

Benzécri J.-P. (1991b) - Typologies de textes latins d'après les occurrences des formes des mots-outil, *Les Cahiers de l'Analyse des Données*, XVI, n°4, p 439-465.

Benzécri J.-P. (1992a) - Note de lecture : sur l'analyse des données dans une enquête internationale, *Les Cahiers de l'Analyse des Données*, vol XVII, n°3, Dunod, Paris, p 353-358.

Benzécri J.-P. et F. (1992b) - Typologie de textes espagnols de la littérature du Siècle d'Or d'après les occurrences des formes des mots outil, *Les Cahiers de l'Analyse des Données*, vol XVII, n°4, Dunod, Paris, p 425-464.

Benzécri J.-P. (1992c) - *Correspondence Analysis Handbook*, (Transl : T.K. Gopalan) Marcel Dekker, New York.

Bergounioux A., Launay M.-F., Mouriaux R., Sueur J-P., Tournier M. (1982) - *La parole syndicale*, P.U.F., Paris.

Bernet C. (1983) - *Le vocabulaire des tragédies de Jean Racine*, *Analyse statistique*, Slatkine-Champion, Genève 1983.

Blosseville M.J., Hébrail G., Monteil M.G., Pénot N. (1992) - Automatic document classification : natural language processing, statistical analysis and expert system techniques used together, *Proceeding of the ACM-SIGIR*, Copenhagen, p 51-58.

Bochi S., Celeux G., Mkhadri A. (1993) - Le modèle d'indépendance conditionnelle : le programme DISIND, *La revue de MODULAD (INRIA),* Juin, p 1-5.

Boeswillwald E. (1992) - L'expérience du CESP en matière de qualité des mesures d'audience, in : *La qualité de l'information dans les enquêtes*, ASU, Dunod, Paris.

Bolasco S. (1992) - Sur différentes stratégie dans une analyse des formes textuelles : Une expérimentation à partir de données d'enquête, *Jornades Internacionals d'Analisi de Dades Textuals*, UPC, Barcelona, p 69-88.

Bonnafous S. (1991) - *L'immigration prise aux mots. Les immigrés dans la presse au tournant des années quatre-vingt,* Kimé, Paris.

Bonnet A. (1984) - *L'intelligence artificielle, promesses et réalités*, InterEditions, Paris.

Boscher F. (1984) - *Le système d'enquêtes sur les conditions de vie et aspirations des Français*, Phase 5, Thème Transport, Rapport Credoc.

Bouchaffra D., Rouault J. (1992) - Solving the morphological ambiguities using a nonstationnary hidden Markov model, *AAAI Fall Symposium on Probabilistic Approach to Natural Language processing*, Cambridge, Massachusetts.

Bouroche J.M., Curvalle B. (1974) - La recherche documentaire par voisinage, R.A.I.R.O., V-1, p 65-96.

Bradburn N., Sudman S., and Associates (1979) - *Improving Interview Method and Questionnaire Design,* Jossey Bass. San Francisco.

Brainerd B. (1974) - *Weighing Evidence in Language and Literature. A Statistical Approach*, University of Toronto Press.

Brunet E. (1981) - *Le vocabulaire français de 1789 à nos jours, d'après les données du Trésor de la langue française*, Slatkine-Champion, Genève-Paris.

Burt C. (1950) - The factorial analysis of qualitative data, *British J. of Stat. Psychol.,* vol 3, n°3, p 166-185.

Burtschy B., Lebart L. (1991) - Contiguity analysis and projection pursuit, in : *Applied Stochastic Models and Data Analysis,* R. Gutierrez and M.J.M. Valderrama, Eds, World scientific, Singapore, p 117-128.

Callon M., Courtial J.-P., Turner W. (1991) - La méthode Leximappe, in : *Gestion de la recherche, nouveaux problèmes, nouveaux outils*. Vinck D. (Ed.), De Boeck-Wesmael, Bruxelles.

Carré R., Dégremont J.F., Gross M., Pierrel J.M., Sabah G. (1991) - *Langage humain et machine*, Presses du CNRS, Paris.

Celeux G., Hébrail G., Mkhadri A., Suchard M. (1991) - Reduction of a large scale and ill-conditioned statistical problem on textual data, in : *Applied Stochastic Models and Data Analysis, Proceedings of the 5th Symposium in ASMDA*, Gutierrez R. and Valderrama M.J. Eds,World Scientific, p 129-137.

Chartron G. (1988) - *Analyse des corpus de données textuelles, sondage de flux d'Information*, Thèse. Univ. Paris 7.

Church K.W., Hanks P. (1990) - Word association norms, mutual information and lexicography, *Computational Linguistics,* Vol 16, p 22-29.

Cibois P. (1992) - Eclairer le vocabulaire des questions ouvertes par les questions fermées : le tableau lexical des questions, *Bull. de Method. Sociol.*, 26, p 24-54.

Cohen M. (1950) - Sur la statistique linguistique, *Conférence de l'institut de linguistique de l'université de Paris*, Vol IX, Année 1949, Klincksieck, Paris.

Cooper W., Gey F.C., Dabney D.P. (1992) - Probabilistic retrieval based on staged logistic regression, *Proceedings of the 15th Int. ACM-SIGIR Conf. on Res. and Dev. in Information Retrieval*, Belkin N and al., Eds, ACM Press, New York, p 198-209.

Coulon D., Kayser D. (1986) - Informatique et langage naturel : Présentation générale des méthodes d'interprétation des textes écrits, *Techniques et sciences informatiques*, Vol.5, n°2, p 103-128.

Cutting D. R., Karger D.R., Pedersen J.O., Tukey J.W. (1992) - Scatter / Gather : A cluster based approach to browsing large document collections. *Proceedings of the 15th Int. ACM-SIGIR*

Conf. on Res. and Dev., in Information Retrieval, Belkin N. and al., Ed, ACM Press, New York, p 318-329.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshman R. (1990) - Indexing by latent semantic analysis, *J. of the Amer. Soc. for Information Science,* 41 (6), p 391-407.

Demonet M., Geffroy A., Gouaze J., Lafon P., Mouillaud M., Tournier M. (1975) - *Des tracts en Mai 68. Mesures de vocabulaire et de contenu*, Armand Colin et Presses de la Fondation Nat. des Sc. Pol., Paris.

Dendien J. (1986) - La Base de données de l'Institut National de la Langue Française, *Actes du colloque international CNRS,* Nice, juin 1985, 2 vol., Slatkine-Champion Genève, Paris.

Desval H., Dou H. (1992) - *La veille technologique. L'information scientifique, technique et industrielle*, Dunod, Paris.

Diday E. (1971) - La méthode des nuées dynamiques, *Revue Stat. Appl.*,vol.19, n°2, p 19-34.

Diday E. (1992) - From data to knowledge : Probabilist objects for a symbolic data analysis, in : *Computational Statistics*, Dodge Y., Whittaker J. (Eds), Physica Verlag, Heidelberg, p 193-214.

Efron B. (1982) - *The Jacknife, the Bootstrap and other Resampling Plans*, SIAM, Philadelphia.

Efron B., Thisted R. (1976) - Estimating the number of unseen species : how many words did Shakespeare know ?, *Biometrika*, 63, p 435-437.

Ellegard A. (1962) - A statistical method for determining authorship : the Junius Letters, 1769-1772, *Gothenburg Studies in English*, n°13, University of Gothenburg.

Escofier B. (1978) - Analyse factorielle et distances répondant au principe d'équivalence distributionnelle, *Revue de Statist. Appl.*, vol. 26, n°4, p 29-37.

Escofier B. [Cordier] (1965) - *Analyse des correspondances*, Thèse, Faculté des Sciences de Rennes.

Escoufier Y. (1985) - L'Analyse des correspondances, ses propriétés, ses extensions. *Bull. of the Int. Stat. Inst.*, 4, p 28-2.

Estoup J. B. (1916) - *Gammes sténographiques,* 4ème Edition, Paris. (cf. aussi Mandelbrot, 1961).

Fiala P., Habert B., Lafon P., Pineira C. (1987) - Des mots aux syntagmes. Figements et variations dans la résolution générale du congrès de la CGT de 1978, *Mots*, N° 14, p 47-87.

Fisher R. A. (1936) - The use of multiple measurements in taxonomic problems, *Annals of Eugen*ics, 7, p 179-188.

Fowler R.H., Fowler W.A.L., Wilson B.A. (1991) - Integrating query, thesaurus, and documents through a common visual representation, *Proceedings of the 14th Int. ACM Conf. on Res. and Dev. in Information Retrieval*, Bookstein A. and al., Ed, ACM Press, New York, p 142-151.

Friedman J.H. (1989) - Regularized discriminant analysis, *Journal of the American Statistical Association*, 84, p 165-175.

Froeschl K.A. (1992) - Semantic metadata : query processing and data aggregation, in : *Computational Statistics*, Dodge Y., Whittaker J. (Eds), Physica Verlag, Heidelberg, p 357-362.

Fuchs W. (1952) - On the mathematical analysis of style, *Biometrika,* 39, p 122-129.

Fuhr N., Pfeifer U. (1991) - Combining model-oriented and description-oriented approaches for probabilistics indexing, *Proceedings of the 14th Int. ACM Conf. on Res. and Dev. in Information Retrieval*, Bookstein A. and al., Ed, ACM Press, New York, p 46-56.

Furnas G. W., Deerwester S., Dumais S.T., Landauer T.K., Harshman R. A., Streeter L.A., Lochbaum K.E. (1988) - Information retrieval using a singular value decomposition model of latent semantic structure, *Proceedings of the 14th Int. ACM Conf. on Res. and Dev. in Information Retrieval*, p 465-480.

Geary R.C. (1954) - The contiguity ratio and statistical mapping, *The Incorporated Statistician,* Volume 5, N° 3, p 115-145.

Geffroy A., Lafon P., Tournier M. (1974) - L'indexation minimale, Plaidoyer pour une non-lemmatisation, Colloque sur l'analyse des corpus linguistiques : *"Problèmes et méthodes de l'indexation minimale",* Strasbourg 21-23 mai 1973.

Geisser S. (1975) - The predictive sample reuse method with applications, *J. Amer. Statist. Assoc.* 70, p 320-328.

Gifi A. (1990) - *Non Linear Multivariate Analysis*, Wiley, Chichester.

Gobin C., Deroubaix J. C. (1987) - Du progrès, de la réforme de l'Etat, de l'austérité. Déclarations gouvernementales en Belgique, *Mots,* n°15, p 137-170.

Goldstein M., Dillon W. R. (1978) - *Discrete Discriminant Analysis,* Wiley, Chichester.

Good I. J., Toulmin G.H. (1956) - The number of new species and the increase in population coverage when a sample is increased, *Biometrika*, 43, p 45-63.

Greenacre M.(1984) - *Theory and Applications of Correspondence Analysis*, Academic Press, London.

Greenacre M. (1993) - *Correspondence Analysis in Practice*, Academic Press, London.

Gruaz C. (1987) - *Le mot français, cet inconnu. Précis de morphographémologie,* Publ. de l'Univ. de Rouen, n°138, Rouen.

Guilbaud G.-Th. (1980) - Zipf et les fréquences, *Mots N° 1*, p 97-126.

Guilhaumou J. (1986) - L'historien du discours et la lexicométrie. Etude d'une série chronologique : Le père Duchesne de Hébert, juillet 1793- mars 1794, *Histoire & Mesure ,* Vol. I, n° 3-4.

Guiraud P. (1954) - *Les caractères statistiques du vocabulaire*, P.U.F., Paris.

Guiraud P. (1960) - *Problèmes et méthodes de la statistique linguistique*, P.U.F., Paris.

Guttman L. (1941) - The quantification of a class of attributes: a theory and method of a scale construction, in *The prediction of personal adjustment* (P. Horst, ed.), SSCR New York, p 251 -264.

Habbema. D. F., Hermans J., van Den Broek K. (1974) - A stepwise discriminant analysis program using density estimation, in *COMPSTAT 1974*, Bruckman G.(ed), Physica Verlag, Vienna.

Habert B., Tournier M. (1987) - La tradition chrétienne du syndicalisme français aux prises avec le temps. Evolution comparée des résolutions confédérales (1945 - 1985), *Mots*, n°14.

Hand D. J. (1981 and 1986) - *Discrimination and Classification*, Wiley, New-York.

Hand D. J. (1992) - Microdata, macrodata, and metadata, *Computational Statistics*, Dodge Y., Whittaker J. (Eds), Physica Verlag, Heidelberg, p 325-340.

Haton J. P. (1985) - Intelligence artificielle en compréhension automatique de la parole : état des recherches et comparaison avec la vision par ordinateur, *Techniques et sciences informatiques*, 4 (3), p 265-287.

Hayashi C. (1956) - Theory and examples of quantification (II), *Proc. of the Institute of Stat. Math.* 4 (2) p 19-30.

Hayashi C. (1987) - Statistical Study of Japanese National Character, *J. Japan Statistical Soc.*, Special Issue, p 71-95.

Hayashi C., Suzuki T., Sasaki M. (1992) - *Data Analysis for Social Comparative research : International Perspective*. North-Holland, Amsterdam.

Hébrail G., Marsais J. (1993) - Experiment in textual data analysis, *EDF, Centre de Recherche*, Clamart, France.

Hébrail G., Suchard M. (1990) - Classifying documents : a discriminant analysis and an expert system work together, *COMPSTAT 90*, (Momirovic K. and Midner, eds), Physica Verlag, p 63-68.

Herdan G. (1964) - *Quantitative Linguistics*, Londres, Butterworths.

Hirschfeld H.D. (1935) - A Connection between correlation and contingency, *Proc. Camb. Phil. Soc.* 31, p 520-524.

Holmes D.I. (1985) - The analysis of literary style - A Review, *J. R. Statist. Soc.,* 148, Part 4, p 328-341.

Holmes D.I. (1992) - A Stylometric analysis of mormon scripture and related texts. *J. R. Statist.Soc.,* 155, Part 1, p 91-120.

Jambu M. (1978) - *Classification automatique pour l'analyse des données, Tome 1 : méthodes et algorithmes*, Dunod, Paris.

Juan S. (1986) - L'ouvert et le fermé dans la pratique du questionnaire. *R. Fr. Socio.* XXVII, Ed. du CNRS, Paris, p 301-306.

Kasher A. (1972) - The book of Isaiah : Caracterization of Authors by Morphological Data Processing, *Revue (R.E.L.O) LASLA N°3,* Liège, p 1- 62.

Labbé D. (1990) - *Le vocabulaire de François Mitterand,* Presses de la Fond. Nat. des Sciences Politiques, Paris.

Labbé D. (1983) - *François Mitterrand - Essai sur le discours*, La pensée sauvage, Grenoble.

Labbé D. (1990) - *Normes de dépouillement et procédures d'analyse des textes politiques*, CERAT, Grenoble.

Labbé D., Thoiron P., Serant D. (Ed.) (1988) - *Etudes sur la richesse et la structure lexicales*, Slatkine-Champion, Paris-Genève.

Lachenbruch P.A., Mickey M.R. (1968) - Estimation of error rate in discriminant analysis, *Technometrics*, 10, p 1-11.

Lafon P. (1980) - Sur la variabilité de la fréquence des formes dans un corpus, *Mots N°1* , p 127-165.

Lafon P. (1981) - Analyse lexicométrique et recherche des cooccurrences, *Mots N°3* , p 95-148.

Lafon P. (1981) - *Dépouillements et statistiques en lexicométrie*, Slatkine-Champion, 1984, Paris.

Lafon P., Salem A. (1983) - L'Inventaire des segments répétés d'un texte, *Mots N°6*, p 161-177.

Lafon P., Salem A., Tournier M. (1985) - Lexicométrie et associations syntagmatiques (Analyse des segments répétés et des cooccurrences appliquée à un corpus de textes syndicaux). *Colloque de l'ALLC*, Metz -1983, Slatkine-Champion, Genève, Paris, p 59-72.

Lazarsfeld P.E. (1944) - The controversy over detailed interviews - an offer for negotiation, *Public Opinion Quat.* n°8, p 38-60.

Lebart L. (1969) - L'Analyse statistique de la contiguïté, *Publications de l'ISUP*, XVIII- p 81 - 112.

Lebart L., Houzel van Effenterre Y. (1980) - Le système d'enquête sur les aspirations des français, une brève présentation, *Consommation* n°1, 1980, Dunod, p 3-25.

Lebart L. (1987) - Conditions de vie et aspirations des français, évolution et structure des opinions de 1978 à 1986, *Futuribles ,* sept 1987, p 25-56.

Lebart L. (1982a) - Exploratory analysis of large sparse matrices, with application to textual data, *COMPSTAT,* Physica Verlag, p 67-76.

Lebart L. (1982b) - L'Analyse statistique des réponses libres dans les enquêtes socio-économiques, *Consommation,* n°1, Dunod, p 39-62.

Lebart L. (1992a) - Discrimination through the regularized nearest cluster method, *in : Computational Statistics*, (Y. Dodge, J. Whittaker, Eds) Physica Verlag, Heidelberg, p 103-118.

Lebart L. (1992b) - Assessing and comparing patterns in multivariate analysis, *Second Japanese French Seminar on Data Science*, in Data Science and application, Hayashi et al. Eds, HBJ, Tokyo, Japan..

Lebart L., Memmi D. (1984) - Analisi dei Dati Testuali : Applicazione al Discorso Politico. Atti delle XXXII Riunione Scient. (Soc. Ital. Statis.), p 27-41.

Lebart L., Morineau A., Bécue M. (1989) - *SPAD.T, Système Portable pour l'Analyse des Données Textuelles*, *Manuel de Référence,* CISIA , Paris.

Lebart L., Morineau A., Warwick K. (1984) *Multivariate Descriptive Statistical Analysis,* Wiley, New York.

Lebart L., Salem A. (1988) - *Analyse statistique des données textuelles*, Dunod, Paris.

Lebart L., Salem A., Berry E. (1991) - Recent development in the statistical processing of textual data, *Applied Stoch. Model and Data Analysis,* 7, p 47-62.

Lelu A., Rozenblatt D. (1986) - Représentation et parcours d'un espace documentaire. Analyse des données, réseaux neuronaux et banques d'images, *Les Cahiers de l'Analyse des Données,* Vol XI - n° 4, p 453-470.

Lelu A. (1991) - From data analysis to neural networks : new prospects for efficient browsing through databases, *Journal of Information Science,* vol. 17, p 1-12.

Lewis D. D., Croft W. B. (1990) - Term clustering of syntactic phrases, *Proceedings of the 13th Int. ACM Conf. on Res. and Dev. in Information Retrieval*, Vidick J.L., Ed, A.C.M.Press, New York, p 385-395.

Lewis D.D. (1992) - An evaluation of phrasal and clustered representation on a text categorization task, *Proceedings of the 15th Int. ACM-SIGIR Conf. on Res. and Dev. in Information Retrieval*, Belkin N. and al., Eds, ACM Press, New York, p 37-50.

Luong X. (Ed.) (1989) - *Analyse arborée des données textuelles, INALF, CUMFID/CNRS*, Nice.

Mahalanobis P.C. (1936) - On the generalized distance in statistics, *Proc. Nat. Inst. Sci. India*, 12, p 49-55.

Mandelbrot B. (1961) - On the theory of words frequency and on related Markovian models of discourses, *Structure of Language and its Mathematical Aspects*, Am. Math. Soc., Providence, R.I., p 190-219.

Mandelbrot B. (1968) - Les constantes chiffrées du discours, *Le Langage*, Encyclopédie de la Pléiade, vol XXV, Gallimard, Paris.

Margulis E.L. (1992) - N-Poisson document modelling, *Proceedings of the 15th Int. ACM-SIGIR Conf. on Res. and Dev. in Information Retrieval*, Belkin N and al., Ed, ACM Press, New York, p 177-189.

McKevitt P., Partridge D., Wilks Y. (1992) - Approaches to natural language discourse processing, *Artificial Intelligence Review,* 6, p 333-364.

McKinnon A., Webster R. (1971) - A method of "author" identification, *The Comput. in Liter. and Linguist. Res*., R.A. Wisbey, Ed., Cambridge Univ. Press.

McLachlan G.J. (1992) - *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New-York.

Menard N. (1983) - *Mesure de la richesse lexicale, théorie et vérifications expérimentales*, Slatkine-Champion, Paris.

Michelet B. (1988) - *La logique d'association,* Thèse, Université Paris 7.

Moran P.A.P. (1954) Notes on continuous stochastic phenomena, *Biometrika,* 37, p 17-23.

Morton A.Q. (1963) - The authorship of the Pauline corpus, *The New Testament and Contemporary Perspective*, W. Barclay, ed., Oxford.

Mosteller F. , Wallace D. (1964) - *Inference and disputed Authorship : The Federalists.* Addison-Wesley, Reading, Mass.

Mosteller F., Wallace D.L. (1984) - *Applied Bayesian and Classical Inference, the Case of the Federalist Papers*, Springer Verlag, New York.

Muller C. (1964) - *Essai de statistique lexicale : L'illusion comique de P. Corneille,* Klincksieck, Paris.

Muller C. (1968) - *Initiation à la statistique linguistique*, Larousse, Paris.

Muller C. (1977) - *Principes et méthodes de statistique lexicale,* Hachette, Paris.

Muller C.(1967) - *Etude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille*, Paris, Larousse.

Nishisato S.(1980) - *Analysis of Categorical Data. Dual Scaling and its Application,* Univ. of Toronto Press.

Palermo D., Jenkins J. (1964) - *Word Association Norm*, University of Minnesota Press, Minneapolis.

Pêcheux M. (1969) - *Analyse automatique du discours*, Dunod, Paris.

Peschanski D. (1988) - *Et pourtant, ils tournent. Vocabulaire et stratégie du PCF (1934 - 1936)*, Klincksieck, Paris.

Petruszewycz M. (1973) - L'histoire de la loi d'Estoup-Zipf, *Math. Sciences Hum.*, n°44.

Pitrat J. (1985) - *Textes, ordinateurs, et compréhension*, Eyrolles, Paris.

Plante P.(1985) - La structure des données et des algorithmes en Dérédec, *Revue Québecoise de Linguistique,* 14:2.

Radday Y. T.(1974) - *"And* " in Isaiah, *Revue (R.E.L.O) LASLA N°2* , Liége, p 25-41.

Rao C.R. (1989) - *Statistics and Truth*, International Cooperative Publishing House, Fairland, USA.

Reinert M. (1983) - Une méthode de classification descendante hiérarchique : Application à l'analyse lexicale par contexte, *Les Cahiers de l'Analyse des Données*, 3, Dunod, p 187-198.

Reinert M. (1986) - Un Logiciel d'analyse lexicale, *Les Cahiers de l'Analyse des Données*, 4, Dunod, p 471-484.

Reinert M. (1990) - Alceste, Une méthodologie d'analyse des données textuelles et une Application : Aurélia de Gérard de Nerval, *Bull. de Méthod. Sociol.* n°26, p 24-54.

Romeu L. (1992) - *Approche du discours éditorial de Ya et Arriba (1939 - 1945)*, Thèse Paris 3.

Rugg D. (1941) - Experiments in wording questions, *Public Opinion Quarterly*, 5, p 91-92.

Salem A. (1979) - *Contribution à une méthodologie de la validation en analyse des données textuelles*, Thèse, Université Paris 6.

Salem A. (1982) - Analyse factorielle et lexicométrie. Synthèse de quelques expériences, *Mots N°4* , p 147-168.

Salem A. (1984) - La typologie des segments répétés dans un corpus, fondée sur l'analyse d'un tableau croisant mots et textes, *Les Cahiers de l'Analyse des Données,* Vol IX, n° 4, p 489-500.

Salem A. (1986) - Segments répétés et analyse statistique des données textuelles, Etude quantitative à propos du père Duchesne de Hébert, *Histoire & Mesure, Vol. I- n° 2*, Paris, Ed. du CNRS.

Salem A. (1987) - *Pratique des segments répétés, Essai de statistique textuelle*, Klincksieck, Paris.

Salem A. (1993) - *Méthodes de la statistique textuelle*, Thèse d'Etat, Université Sorbonne Nouvelle (Paris 3).

Salton G. (1988) - *Automatic Text Processing : the Transformation, Analysis and Retrieval of Information by Computer,* Addison-Wesley, New York.

Salton G., Mc Gill M.J. (1983) - *Introduction to Modern Information Retrieval,* International Student Edition.

Sasaki M., Suzuki T. (1989) - New directions in the study of general social attitudes : trends and cross-national perspectives, *Behaviormetrika*, 26, p 9-30.

Saussure F. de (1915) - *Cours de linguistique générale*, Payot, Genève (rééd. 1986).

Schank, R. C., (Ed.) - *Conceptual Information Processing*, North-Holland, Amsterdam, 1975.

Schuman H. (1966) - The random probe : a technique for evaluating the validity of closed questions, *Amer. Socio. Rev.* n°21, p 218-222.

Schuman H., Presser F. (1981) - *Question and Answers in Attitude Surveys*, Academic Press, New York.

Sekhraoui M. (1981) - *La saisie des textes et le traitement des mots : Problèmes posés, essai de solution,* Mémoire, Ecole des hautes études en sciences sociales, Paris.

Shannon C.E. (1948) - A mathematical theory of communication, *B.S.T.J.*, n° 27.

Simpson E.H. (1949) - Measurement of diversity, *Nature,* 163, p 688.

Smith M. W. A. (1983) - Recent experience and new developments of methods for the determination of authorship, *Ass. for Lit. and Linguist. Comput. Bull.*, 11, p 73-82.

Somers H. H. (1966) - Statistical methods in literary analysis, *The Computer and Literary Style,* (J. Leed, Eds), Kent State University Press, Kent, Ohio.

Spevack M. (1968) - *A Complete and Systematic Concordance of the Work of Shakespeare*, George Olms, Hildesheim.

Stone C.J. (1974) - Cross-validatory choice and assessment of statistical predictions. *J.R. Statist. Soc., B 36, p 111-147.*

Sudman S., Bradburn N. (1974) - *Response Effects in Survey,* Aldine, Chicago.

Suzuki T. (1989) - Cultural link analysis : its application to social attitudes - A study among five nations, *Bull. of the Int. Stat. Inst.*, 47° Session, vol.1, p 363-379.

Tabard (1974) - *Besoins et aspirations des familles et des jeunes*, Coll. Etudes CAF, n° 16, CNAF, Paris.

Tabard N. (1975) - Refus et approbations systématiques dans les enquêtes par sondage, *Consommation,* n°4, Dunod, p 59-76.

Tenenhaus M., Young F. W. (1985) - An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data, *Psychometrika*, vol 50, p 91-119.

Thisted R., Efron B. (1987) - Did Shakespeare write a newly discovered poem?, *Biometrika*, 74, p 445-455.

Tournier M. (1985a) - Sur quoi pouvons-nous compter ? Hommage à Hélène Nais, *Verbum*.

Tournier M. (1985b) - Texte propagandiste et cooccurrences. Hypothèses et méthodes pour l'étude de la sloganisation, *Mots N°11*, p 155-187.

Tournier M. (1980) - D'ou viennent les fréquences de vocabulaire?, *Mots N°1*, p 189-212.

Van Rijckevorsel J. (1987) - *The application of fuzzy coding and horseshoes in multiple correspondances analysis,* DSWO Press, Leyde.

Van Rijsbergen C.J. (1980) - *Information Retrieval*, 2nd Ed., Butterworths, London.

Von Neumann J. (1941) - Distribution of the ratio of the mean square successive difference to the variance, *Annals of Math. Stat.*, vol.12.

Warnesson I., Parisot P., Bedecarrax C., Huot C. (1993) - Traitements linguistiques et analyse des données pour une exploitation systématique des banques de données, *Revue Française de bibliométrie*, i 21.

Weber R.P. (1985) - *Basic Content Analysis*, Sage, Beverly Hills.

Weil G.E., Salem A., Serfaty M. (1976) - Le livre d'Isaïe et l'analyse critique des sources textuelles, *Revue (R.E.L.O) LASLA , N°2* , Liège.

Wilks Y., Fass D., Guo C., McDonald J. E., Plate T., Slator B. M. (1991) - Providing machine tractable dictionnary tools, *Machine translation,* p 99-154.

Wold S. (1976) - Pattern recognition by means of disjoint principal component models, *Pattern Recognition,* 8, p 127-139.

Yule G.U. (1944) - *The Statistical Study of Literary Vocabulary*, Cambridge University Press, Reprinted in 1968 by Archon Books, Hamden, Connecticut.

Zipf G. K. (1935) - *The Psychobiology of Language, an Introduction to Dynamic Philology*, Boston, Houghton-Mifflin.

Bibliographie complémentaire

On trouvera ci-dessous une liste (non-exhaustive) d'***ouvrages généraux*** ou de ***thèses,*** pour la plupart en langue française, qui complètent les références précédente du point de vue des domaines de recherche abordés et des domaines connexes (Statistique et analyse des données, linguistique et analyse de discours, méthodes d'enquêtes, applications).

Andreewsky A., Fluhr C. (1973) - *Apprentissage, analyse automatique du langage, application à la documentation,* Documents de linguistique quantitative, n° 21, Dunod, Paris.

Antoine J. (1982) - *Le Sondage, outil du marketing*, Dunod, Paris.

Balpe J. P. (1990) - *Hyperdocuments, hypertextes, hypermedias,* Eyrolles, Paris.

Bar-Hillel Y.(1964) - *Language and Information*, Addison-Wesley,

Baudelot C., Establet R. (1989) - *Le niveau monte,* Seuil, Paris.

Bourdieu P. (1982) - *Ce que parler veut dire*, *l'économie des echanges linguistiques,* Fayard, Paris.

Bouroche J.M., Saporta G. (1980) - *L'analyse des données*, coll."Que sais-je", n°1854, PUF, Paris .

Bourques G., Duchastel J. (1988) - *Restons traditionnels et progressifs, pour une nouvelle analyse du discours politique*, Boréal, Montréal.

Brian E. (1984) - *Analyse des données lexicométriques,* Rapport Credoc / D.G.T.

Caillez F., Pagès J.P. (1976) - *Introduction à l'analyse des données*, S.M.A.S.H., Paris.

Cibois P. (1984) - *Analyse des données en sociologie*, P.U.F. Paris.

Cliff A.D. and Ord J.K. (1981) - *Spatial Processes : Models and Applications*, Pion, London.

Daoust F. (1990) - *SATO: Système de base d'analyse de textes par ordinateurs*, ATO-UQAM, Montréal.

Deroo M., Dussaix A. M. (1980) - *Pratique et analyse des enquêtes par sondage*, P.U.F., Paris.

Dodge Y. (1993) - *Statistique. Dictionnaire encyclopédique*, Dunod, Paris.

Ducrot O., Todorov T. (1972) - *Dictionnaire encyclopédique des sciences du langage,* Ed. du Seuil, Paris.

Eissfeldt O. (1965) - *The Old Testament. An introduction*, transl. by P.R. Ackroyd, Oxford.

Fénelon J.P. (1981) - *Qu'est-ce-que l'analyse des données?*, Lefonen, Paris.

Fuchs C. (1993) - *Linguistique et traitements automatiques des langues*, Hachette, Paris.

Gardin J-C. et coll. (1981) - *La logique du plausible : essai d'epistémologie pratique*, Ed. de la Maison des sciences de l'homme, Paris.

Gardin J-C. (1971) - *Les analyses de discours*, Delachaux & Niestlé, Neuchâtel.

Georgel A. (1992) - Classification statistique et réseaux de neurones formels pour la représentation des banques de données documentaires,Thèse - Université Paris-7.

Ghiglione R. (1986) - *L'homme communicant,* Armand Colin, Paris.

Ghiglione R., Matalon B. (1991) - *Les enquêtes sociologiques,* Armand Colin, Paris.

Grangé D., Lebart L. (Ed.) (1993) - *Traitements statistiques des enquêtes,* Dunod, Paris.

Groves R.M. (1989) - *Survey Errors and Survey Costs*, J. Wiley, New-York.

Lafon P., Lefevre J., Salem A., Tournier M. (1985) - *Le Machinal*, *Principes d'enregistrement informatique des textes*, Publications de l'INaLF, Klincksieck, Paris.

Lebart L., Morineau A., Fénelon J.P. (1979) - *Traitement des données statistiques*. Dunod, Paris.

Lebart L., Morineau A., Tabard N. (1977) - *Technique de la description statistique,* Dunod, Paris.

Lelu A. (1993) - *Modèles neuronaux pour l'analyse documentaire et textuelle*, Thèse, Univ. Paris 6.

Maingueneau D. (1976) - *Initiation aux méthodes de l'analyse du discours*, Hachette, Paris.

Marchand P. (1993) - *Engagement politique et rationalisation - Analyse psychosociale du discours militant,* Thèse de l'Université de Toulouse II.

Meynaud H., Duclos D. (1985) - *Les Sondages d'opinions,* La Découverte, Paris.

Muller P., (1991) - *Vocabulaire et rhétorique dans Etudes socialistes de Jean Jaurès*, Thèse, Paris 3.

Prost A. (1974) - *Vocabulaire des proclamations électorales de 1881,1885 et 1889*, PUF, Paris.

Rey A. (1970) - *La Lexicologie*, Klincksieck, Paris.

Robin R. (1973) - *Histoire et Linguistique,* Armand Colin, Paris.

Rouanet H., Le Roux B. (1993) - *Analyse des données multidimensionnelles,* Dunod, Paris.

Rouault J. (1986) - *Linguistique automatique : Applications documentaires,* Berne P. Lang.

Roux M. (1985) - *Algorithmes de classification*, Masson, Paris.

Saporta G. (1990) - *Probabilité, analyse des données et statistique*, Technip, Paris.

Searle, J.R. (1979) - *Sens et expression : étude de la théorie des actes de langages.* Editions de Minuit, Paris.

Tournier M. (1975) - *Un Vocabulaire ouvrier en 1848, essai de lexicométrie*, Thèse d'état, multigr., ENS de Saint-Cloud.

Volle M. (1980) - *Analyse des données,* Economica, Paris.

Wagner R.-L. (1970) - *Les vocabulaires français*, tomes 1 et 2, Didier, Paris.

Wisbey R. A. (Ed) (1971) - *The computer in Literary and Linguistic Research*, Cambridge Univ. Press, Cambridge.

Index des auteurs