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.

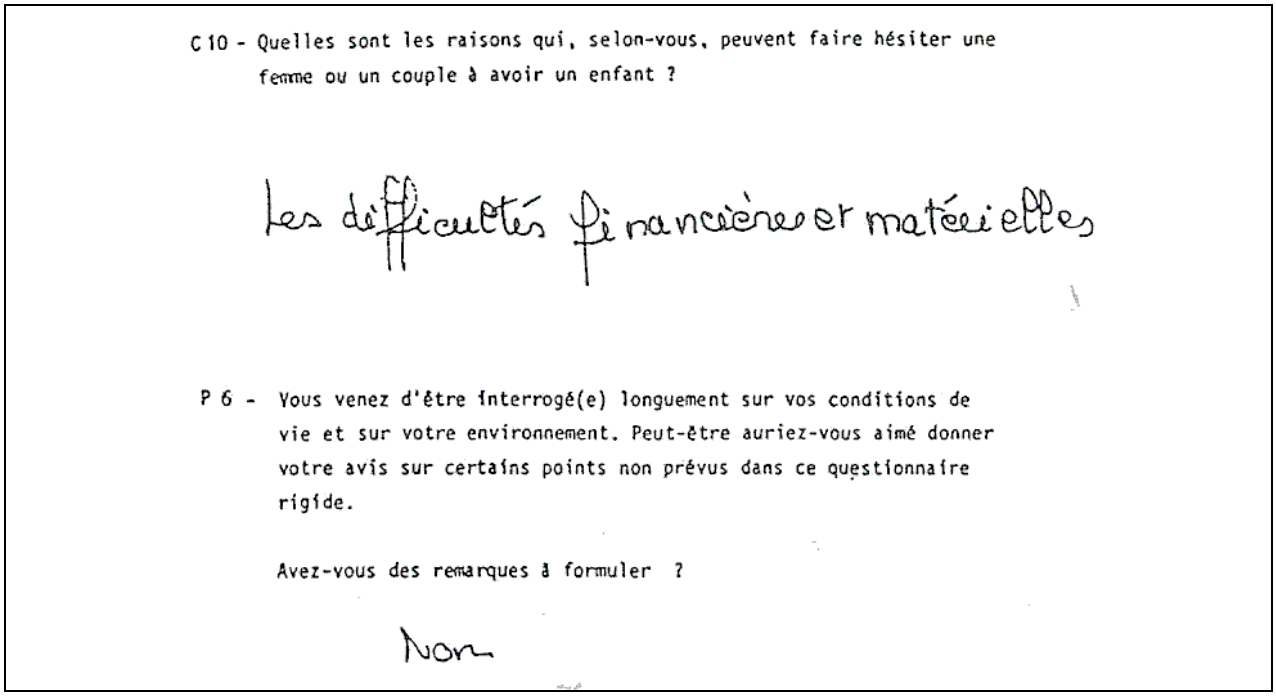
A.1 Le logiciel SPAD.T

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.

---

[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".

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).

```
<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.

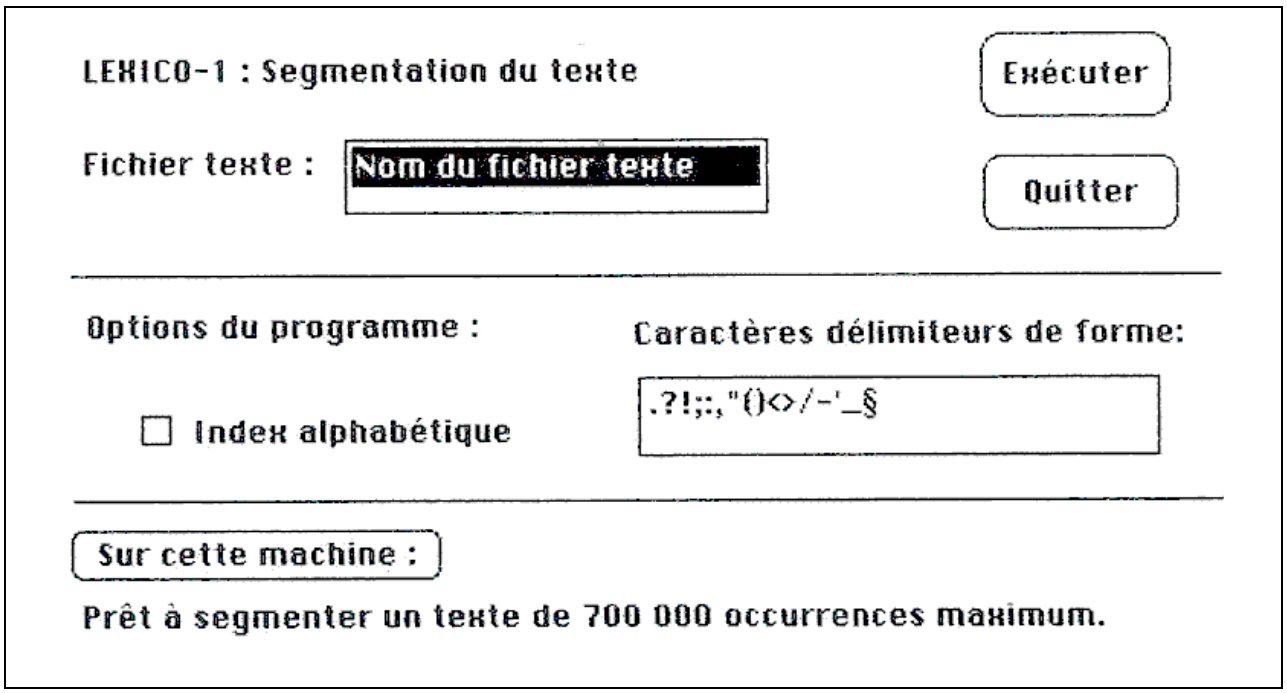


Figure A2.2. Réglage des options de segmentation du texte

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.

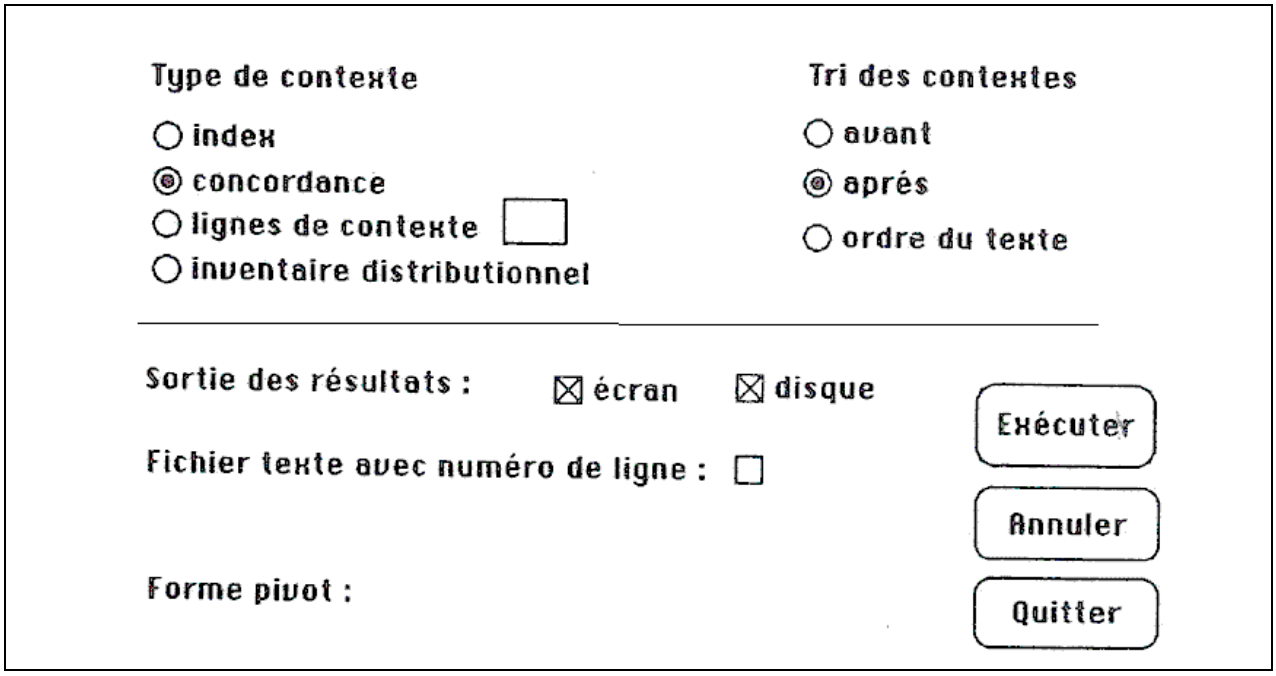


Figure A2.3

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.

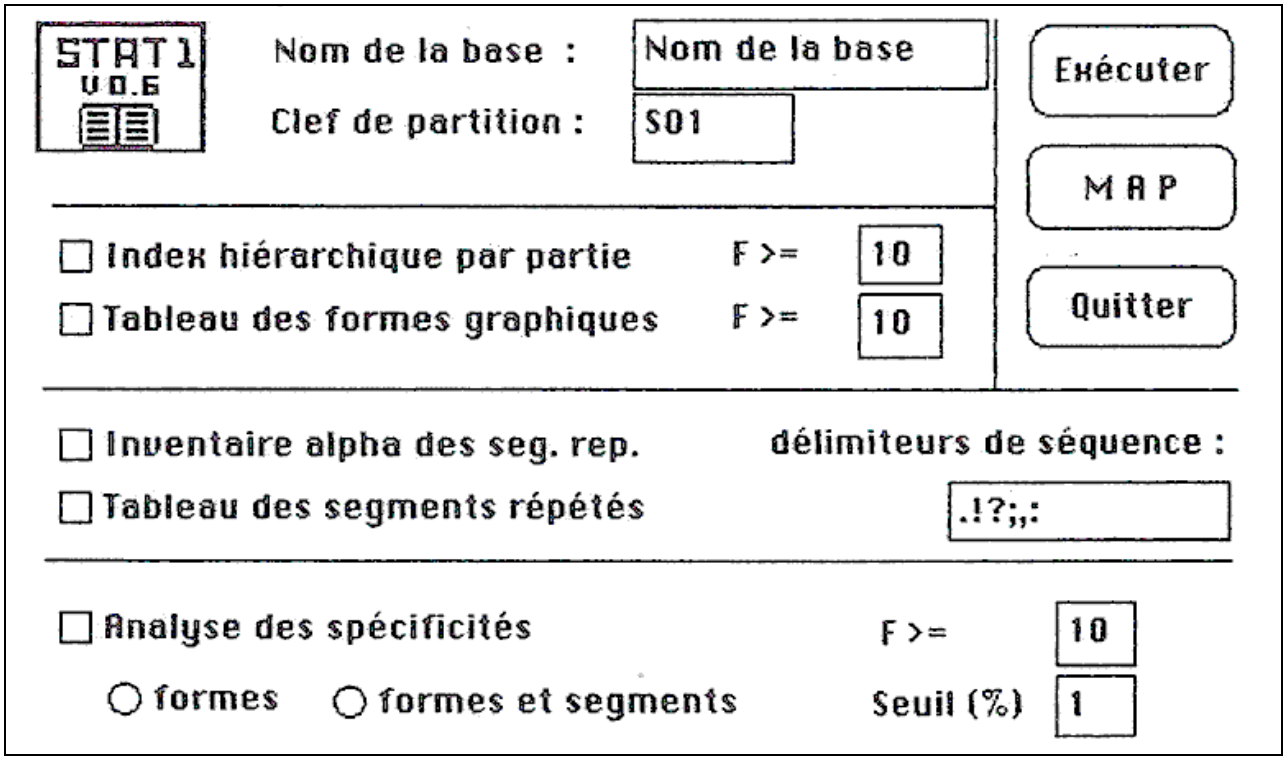


Figure A2.4 . Lexico1 : réglage des options du module *statistiques -1*

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

[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.

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 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 :

[1] Hypertalk et Hypercard sont des produits Apple.

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 ;

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[1] et intègre la notion d'Hypertexte. Il en résulte pour l'utilisateur une commodité incomparable pour toute la partie qui

[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*.

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.

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

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

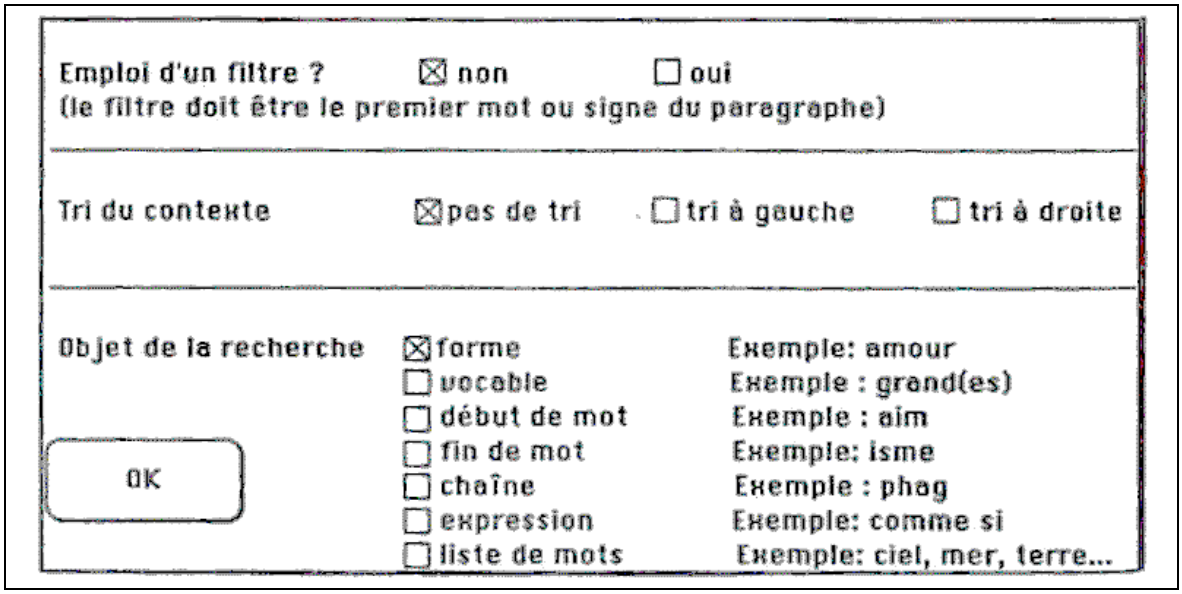


Figure A.4.2. Exemple d'écran *Hyperbase*
Dialogue de la commande *concordance*

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.

---

[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.

PHRASEA[1] consacré à la recherche documentaire au sein de vastes corpus de textes.

Le SPHINX[2], logiciel de dépouillement d'enquête doté d'une interface-utilisateur élaborée et comportant des modules de traitements de données textuelles.

LEXIS[3], un des modules de la série de logiciels de dépouillement d'enquêtes développée et distribuée par la société Eole.

[1] Sur MacIntosh, C. Poveda et J. Y. Jourdain, B&L Parenthèses, 79 av. Guynemer, 59 700, Marc en Bareuil

[2] Sur PC ou MacIntosh, J. Moscarola, Le SPHINX Développement, 13 Chemin des Amarantes, 74600, Seynod

[3] Sur PC, EOLE, 6 rue du Quatre Septembre, 92130, Issy-les-Moulineaux.