SPAD-N AU SERVICE D'UNE MÉTHODOLOGIE POUR L'ANALYSE DES DONNÉES TEXTUELLES

Javier SANCHEZ

I. PRÉSENTATION DE LA MÉTHODOLOGIE D'ANALYSE TEXTUELLE INFORMATIQUE

Pour progresser dans l'étude du texte, il est indispensable d'admettre en linguistique la nécessité d'introduire des facteurs externes tels que sont l'ordinateur et ses composantes. C'est ainsi que, depuis 1986, nous avons élaboré (et nous continuons) une méthodologie adaptée à la recherche linguistique sur les textes de grande dimension, qui a nécessité l'alternance de différents outils informatiques. Jusqu'en 1991, nous avions un vide méthodologique, dans la mesure où nos explorations ne disposaient pas encore de techniques conviviales pour l'interprétation des grandes masses de données textuelles. En effet, avant l'apparition de SPAD-N (Système Portable pour l'Analyse des Données Numériques) [1] il n'existait pas de logiciels accessibles par nos étudiants et par nos chercheurs en langues étrangères et linguistique. SPAD-N est un logiciel de statistique évoluée qui propose différentes techniques factorielles dont nous avons besoin pour nos recherches. Ainsi, nous appliquons, sur des grands tableaux lexicaux, les techniques d'analyse factorielles de correspondances simples et de classification automatique, dont nous ne rappellerons pas ici les principes (voir références bibliographiques).

SPAD-N, dans sa dernière version, propose un environnement convivial adapté à l'enseignement et à la recherche en sciences humaines, pour des utilisateurs non avertis, grâce à la hiérarchisation des écrans et des sous-écrans et à des commandes intégrées dans des menus déroulants. Il propose également un module assisté pour la programmation des différents traitements statistiques qu'il est capable de réaliser, et dont il faut noter la grande rapidité.

1 SPAD-N (Système Portable pour l'Analyse des Données Numériques), Centre international de Statistiques et d'Informatique Appliquées (C.I.S.I.A.), Saint-Mandé (France).

Cette méthodologie d'analyse textuelle comporte donc un certain nombre de procédures que nous allons passer en revue avant de présenter plus en détail le logiciel SPAD-N qui constitue le dernier maillon de cette chaîne. Les techniques informatiques que nous utilisons répondent aux besoins spécifiques de l'analyse informatique des textes. Elles sont, dans l'ordre, l'enregistrement du corpus, le repérage des alternances typologiques, la génération de listes, la création des bases de données textuelles et, comme nous l'avons déjà précisé, l'analyse statistique à partir de SPAD-N (analyse des correspondances). Ces procédures ont pour but de rendre possible, évolutive et dynamique la description des textes de très grande dimension, tout en respectant leurs références internes et externes.

Pour la première phase de la méthode, nous avons exploré la plupart des logiciels de reconnaissance optique de caractères, et nous avons adopté ceux qui nous semblaient les plus appropriés à notre problématique : notamment Omnipage. C'est ainsi qu'il nous a été possible d'enregistrer les oeuvres suivantes de Miguel de Cervantes : les 12 Nouvelles Exemplaires, Don Quichotte et les Intermèdes, puisque nous travaillons surtout en espagnol. Ces textes constituent une banque de données textuelle de plus de deux millions d'occurrences. En ce qui concerne la génération des listes à partir de ces enregistrements, nous avons choisi le logiciel plurilinguistique Micro-OCP (Oxford University). Cet outil a comme principale caractéristique d'être un logiciel programmable par l'utilisateur pour la réalisation d'index et de concordances sur support magnétique. Il nous permet de classer les formes soit alphabétiquement, soit hiérarchiquement (par ordre de fréquence), et de sélectionner la totalité du texte ou une partie de celui-ci.

Cependant, il s'est avéré indispensable, pour dynamiser l'exploration et l'analyse de ces premiers résultats statistiques, de construire un outil de conversion de listes, que nous avons nommé CONVARIT, dans les formats de base de données les plus répandues, et en particulier celle que nous utilisons (works), car elle nous semble la plus adaptée et la plus conviviale pour nos travaux. En effet, le logiciel intégré Works comporte, outre la base de données, deux outils supplémentaires que nous employons toujours dans plusieurs phases de la recherche : le traitement de texte et le tableur graphique. Works facilite, de façon tout à fait considérable les échanges d'informations entre ses trois outils, permettant ainsi, non seulement de dynamiser les procédures d'exploration des textes, mais également de simplifier de beaucoup la présentation des résultats. Ce logiciel est utilisé pour la

correction et la codification éditoriale, structurelle et situationnelle de nos corpus exploratoires (traitement de texte), pour la constitution des tableaux et des graphiques automatiques (tableur), et comme nous l'avons dit précédemment, pour les analyses linguistiques opérées sous base de données, grâce à notre convertisseur CONVARIT.

Ces procédures d'exploration dynamique des données textuelles nous ont donc permis d'opérer davantage d'analyses linguistiques que dans la période durant laquelle cette identification des formes se réalisait encore manuellement, à partir des concordances éditées sur papier. Comme il est impossible d'interpréter directement les résultats obtenus, nous procédons alors à une analyse statistique à partir de l'Analyse Factorielle des Correspondances proposée par SPAD-N et rendant possible l'interprétation.

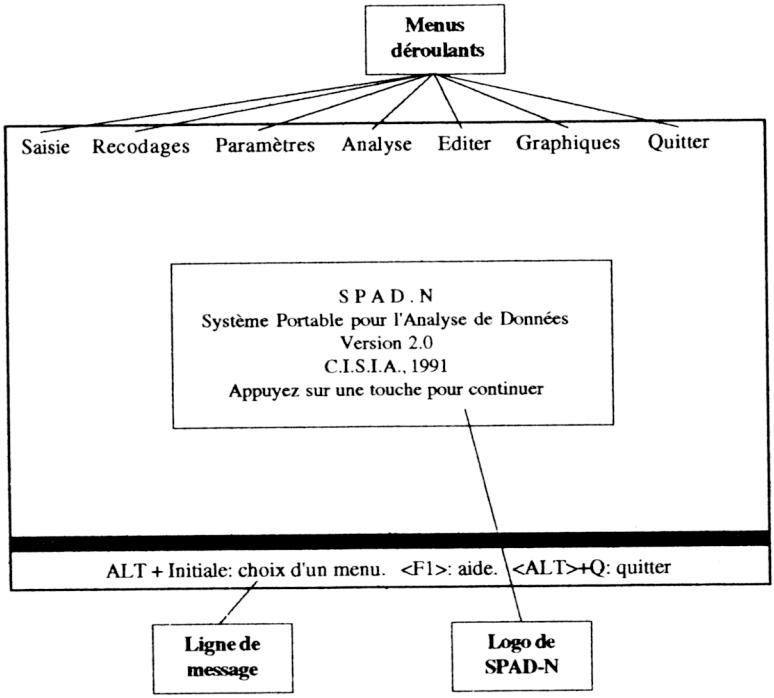
2. PRESENTATION DU LOGICIEL SPAD-N

SPAD-N est un logiciel disponible dans la gamme des ordinateurs compatibles IBM que nous utilisons pour notre recherche. Il propose des procédures d'analyse statistique de tableaux de grande dimension et il est donc orienté principalement vers l'analyse multidimensionnelle des données qui s'inscrit dans la tradition de nos travaux.

Nous proposerons ici une présentation afin d'illustrer les principes de fonctionnement de ce logiciel. Nous rappelons que la version 2 de cet outil comporte des menus déroulants intégrés qui rendent le travail plus convivial.

2.1. L'écran initial

Dès le lancement de SPAD-N, apparaît un écran initial que nous reproduisons ci-dessous :



Comme nous pouvons le voir dans cette illustration, le logiciel offre trois possibilités : soit choisir un menu avec la touche [ALT] accompagnée de l'initiale correspondante, soit afficher l'aide à partir de la fonction [F1], soit quitter.

2.2. L'importation des tableaux

Avant tout traitement statistique, il est nécessaire que le logiciel SPAD-N dispose en mémoire du tableau source des données que l'on souhaite interpréter. Pour ce faire, deux cas de figures sont alors envisageables :

- soit il n'existe pas de fichier contenant les données, et par conséquent l'utilisateur doit effectuer une saisie sous SPAD-N,
- soit les données se trouvent stockées dans un fichier externe à SPAD-N et n'ont plus qu'à être récupérées.

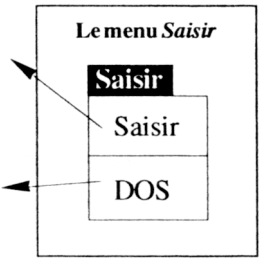
Etant donné que, dans nos recherches, les tableaux récapitulatifs sont constitués par le tableur du logiciel intégré WORKS, dont nous

avons parlé précédemment, nous n'exposerons ici que les procédures d'importation des données vers SPAD-N.

La première opération à effectuer pour cette importation, est l'ouverture du menu Saisie de l'écran initial (à travers la commande ALT+S). Comme nous le voyons ci-contre, ces menus proposent deux commandes :

■ la première commande permet d'obtenir le sous-écran spécialisé de saisie et l'importation des tableaux des données,

■ la deuxième commande offre la possibilité de retourner momentanément au système d'exploitation MS-DOS.



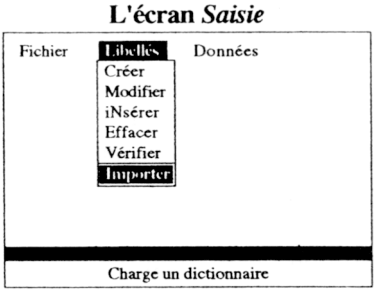
Nous choisirons cette première commande. Elle fait apparaître un écran que nous représentons ci-après :

Comme le montre l'illustration ci-contre, ce sous-écran spécialisé présente trois menus [2] :

■ Le menu Fichier permet de charger et de sauver les fichiers de travail. Il assure donc la gestion des fichiers.

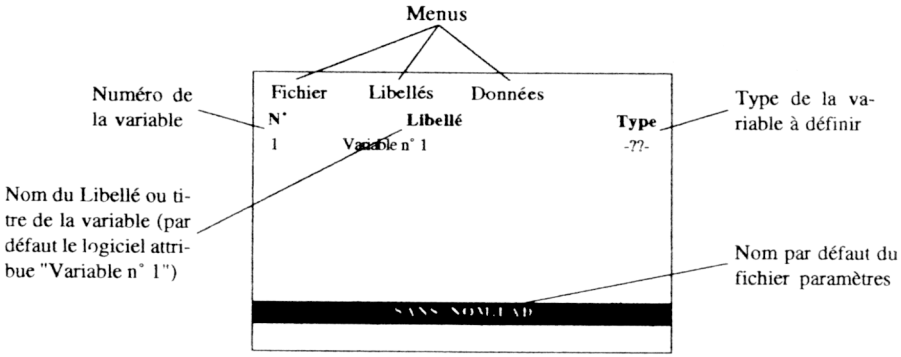
■ Le menu Libellés sert à définir les libellés, c'est-à-dire à donner les différents paramètres des variables ou modalités colonnes.

■ Le menu Données propose, après la définition des libellés, la création et l'importation du tableau des données ou modalités lignes (individus).



La première opération que l'on doit réaliser, avant l'importation des données, à partir de cet écran, consiste à définir les paramètres des libellés dont nous avons parlé. Pour cela nous emploierons la commande Créer du menu Libellés qui affichera l'écran suivant :

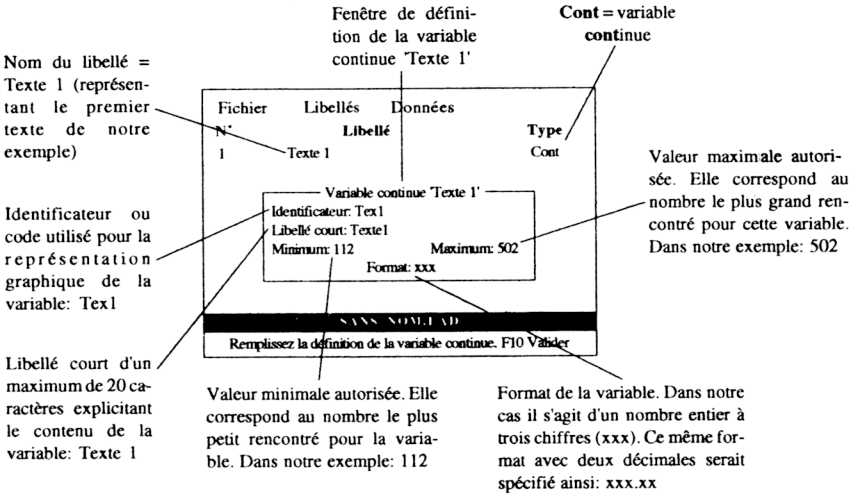
2 Nous ne donnerons lors de cet exposé que les commandes nécessaires à l'importation des fichiers-tableaux réalisés sous le tableur du logiciel intégré works qui est utilisé à plusieurs phases de notre protocole de recherche.



Nous pouvons à partir de là définir deux types de variables :

- les variables continues ou quantitatives sont celles qui ne prennent que des valeurs numériques. Elles figurent dans nos tableaux récapitulatifs des fréquences et de fonctions des formes comme nous pourrons le constater dans les exemples que nous donnons.
- les variables nominales ou qualitatives expriment des modalités non-numériques, c'est-à-dire, celles qui correspondent aux différentes variables que l'on peut définir. Exemple : la typologie des textes ou des discours, les catégories grammaticales et syntaxiques, etc.

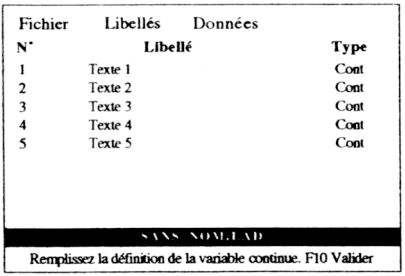
Nous définirons, après avoir donné le nom du libellé, la variable continue en spécifiant dans la colonne Type le code Cont. C'est ainsi qu'à partir de notre exemple sur la distribution des douze formes dans cinq textes, nous définirons la première variable ainsi :



SPAD-N au service d'une méthodologie pour l'analyse des données textuelles

Une fois que l'on a paramétré les cinq variables, on obtient l'écran ci-contre, comportant

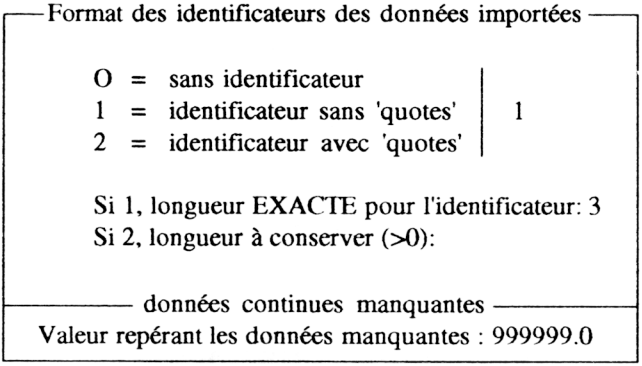
- les numéros d'ordre des variables
- les noms des libellés de variables,
- le type de variables (cont).



Il ne reste plus qu'à importer les données qui se trouvent dans un fichier que nous appellerons ici Tab.txt. Pour cela, il faut utiliser deux commandes du menu Données [3] :

- Premièrement, nous validerons la commande intitulée Format d'importation permettant de définir le format du fichier que l'on veut intégrer dans SPAD-N. Cette opération permet l'affichage de la fenêtre suivante :

3 Le menu Données se situe en troisième position dans la barre de menus (voir l'illustration en haut de cette page).



A partir de cette zone de dialogue, nous pouvons définir si les lignes du tableau des données à importer comportent ou non des identificateurs (noms des lignes) et si oui, s'ils sont repérés ou non par des quotes (').

Donnons quelques exemples de ces trois formats d'importation :

1) Si nous choisissons l'option sans identificateur, les données du tableau seront présentées ainsi [4] :

502	629	885	379	904
486	666	807	327	863
427	582	1034	404	1052
274	323	520	226	565
251	356	607	291	641
183	246	448	160	430

2) Si nous choisissons l'option identificateur avec 'quotes', les données du tableau seront les suivantes :

4 Comme nous pouvons le voir dans ces exemples, l'utilisateur n'inclut pas le calcul des marges. Le logiciel SPAD-N réalise ces opérations automatiquement. Nous pouvons remarquer également que le blanc est le séparateur utilisé pour délimiter les colonnes. Il s'agit donc, pour ces échantillons de données, de définir l'identificateur de chaque ligne.

'de'	502	629	885	379	904
'y'	486	666	807	327	863
'que'	427	582	1034	404	1052
'a'	274	323	520	226	565
'la'	251	356	607	291	641
'en'	183	246	448	160	430

3) Si nous choisissons l'option identificateur sans 'quotes', il faudra définir dans la fenêtre la longueur (en nombre de caractères) de ces identificateurs. Les données du tableau auront le format suivant :

de	502	629	885	379	904
y	486	666	807	327	863
que	427	582	1034	404	1052
a	274	323	520	226	565
la	251	356	607	291	641
en	183	246	448	160	430

■ La deuxième opération que l'on doit réaliser, après la définition de ces paramètres sur les identificateurs, est la validation de la commande Importer du menu Données. Une fenêtre apparaît dans laquelle il faut indiquer le nom du fichier à importer :

Nom des données : Tab.txt

Immédiatement, les données sont inscrites à l'écran dans le tableur intégré à SPAD-N. Voici une illustration à partir de notre exemple :

Fichier Libellés Données

N°	Identif.	1	2	3	4	5
1	de	502	629	885	379	904
2	y	486	666	807	327	863
3	que	427	582	1034	404	1052
4	a	274	323	520	226	565
5	la	251	356	607	291	641
6	en	183	246	448	160	430
7	los	179	158	283	111	246
8	el	149	231	366	168	418
9	por	122	118	212	87	210
10	le	121	107	158	77	220
11	las	112	111	112	61	151
12	se	112	177	221	92	274

TAB.TAD TAB.DAD

Identificateur de la donnée puis Enter. Esc pour annuler

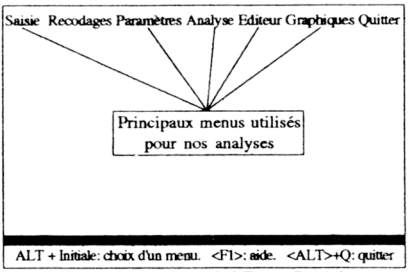
Il ne reste plus qu'à sauvegarder ces données sous SPAD-N. Pour cela nous utiliserons les commandes du menu fichier suivantes :

- Sauver sous... pour une première sauvegarde du fichier importé,
- Sauver pour enregistrer toute modification des données à partir de la deuxième sauvegarde.

2.3. Lancement du traitement

Pour pouvoir lancer le traitement, nous devons encore réaliser un certain nombre d'opérations :

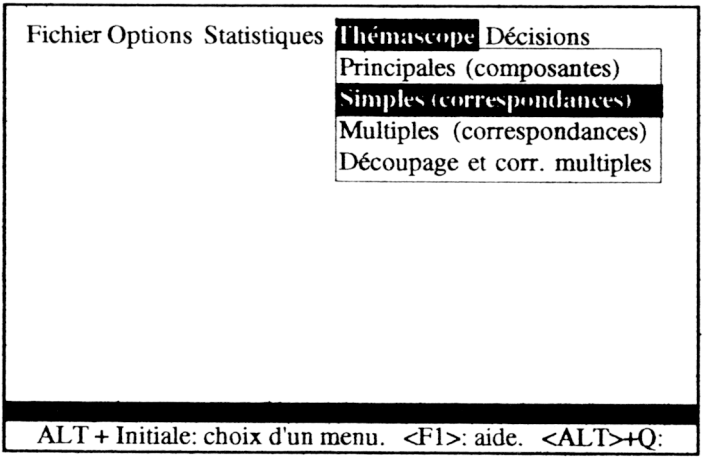
■ Premièrement, nous quitterons l'Ecran Saisie que nous venons de présenter. Pour cela nous valide-rons la commande Retour du menu Fichier. Nous nous retrouverons alors dans l'Ecran Initial qui comporte les principaux menus.



Rappel des menus de l'Ecran Initial

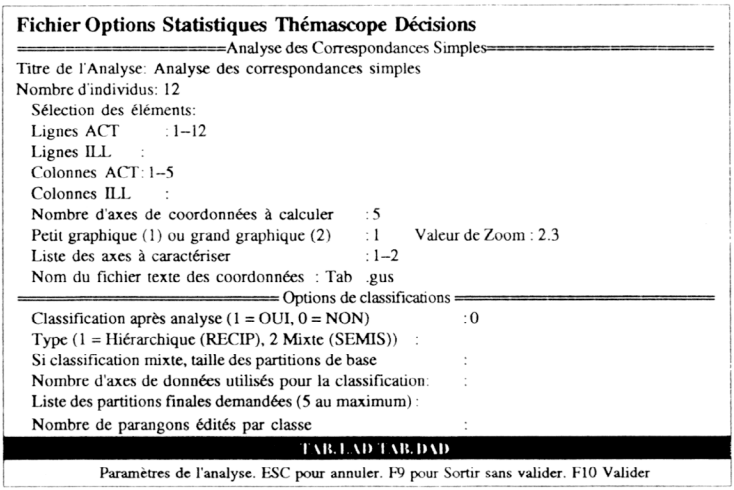
■ Deuxièmement, nous appellerons l'écran Paramètres en exécutant la commande Paramètres de l'Ecran Initial. Nous obtenons alors un nouvel écran qui comporte cinq menus :

Ecran *Paramètres*



Dans ce menu, comme nous pouvons le voir dans l'illustration ci-dessus, nous n'utiliserons que le menu Thémascope ou sont insérées les commandes qui permettent de choisir l'analyse factorielle. Deux types sont proposés : l'analyse de correspondances simple et multiple. Dans nos

recherches nous n'appliquons que le premier. Ainsi, après avoir validé la commande Simples (correspondances) la fenêtre suivante apparaît :



Nous ne nous intéresserons qu'à la première partie de cette fenêtre. Nous pouvons remarquer qu'elle propose le titre de l'analyse ainsi que le nombre d'individus qui entrent en jeu, dans notre cas 12. De plus, elle nous permet de définir les lignes (ensemble I) et les colonnes (ensemble J) que l'on veut sélectionner pour le traitement. Dans cet exemple toutes les lignes [5] et les colonnes [6] seront prises en compte. Si par exemple nous souhaitions sélectionner seulement les lignes 3, 7 et de 10 à 12, nous rentrerions la séquence suivante dans l'option Lignes ACT :3,7,10--12 ; et pareillement pour les colonnes.

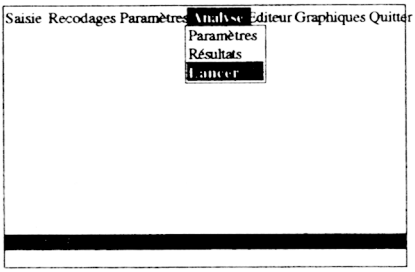
Les options Lignes ILL et Colonnes ILL correspondent aux éléments supplémentaires que nous avons présentés antérieurement. Nous pouvons également définir le nombre d'axes factoriels que l'on désire calculer. Par défaut, le logiciel calcule 5 axes. D'autres options sont disponibles, tels que le choix entre petit et grand graphique ainsi que les axes factoriels à caractériser. Une dernière option permet de créer un fichier de coordonnées auquel il ne faut jamais oublier de donner un nom et qui comporte toujours l'extension .GUS ; pour notre exemple nous l'avons nommé Tab.gus. Nous validerons cette fenêtre avec la touche de

5 La séquence 1--12 signifie la sélection de la ligne 1 à 12. Par conséquent la totalité des lignes de notre exemple .

6 La séquence 1--5 signifie la sélection de la colonne 1 à 5. Par conséquent la totalité des colonnes de notre exemple.

fonction F10 qui proposera une nouvelle zone de dialogue pour donner un nom au fichier contenant les résultats du traitement.

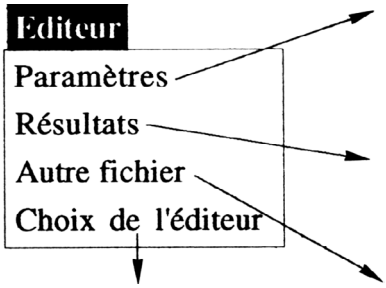
Pour lancer le traitement, l'utilisateur devra revenir à l'Ecran Initial en exécutant la commande Retour du menu Fichiers, puis la commande Lancer du menu Analyse (voir ci-contre). Deux autres commandes sont disponibles, Paramètres et Résultats, qui permettent respectivement de sélectionner des fichiers paramètres et résultats.



Le menu *Analyse* de l'Ecran Initial

2.4. La consultation des résultats

Comme nous l'avons indiqué précédemment, lors de tout traitement, l'ordinateur construit une série de résultats intermédiaires. SPAD-N offre la possibilité de les consulter et éventuellement de les imprimer. Rappelons que ces résultats sont indispensables pour l'interprétation des représentations graphiques. Pour les consulter, nous utiliserons le menu Editeur de l'écran initial. Il propose l'édition à l'écran des fichiers paramètres et résultats, ainsi que le choix d'un autre éditeur que celui de SPAD-N pour les parcourir. Nous présentons ce menu :



Permet d'afficher à l'écran le contenu du fichier paramètres. Rappelons que ce fichier correspond à la programmation des procédures de traitement.

Permet d'afficher à l'écran le contenu du fichier résultats. Ce fichier comporte tous les tableaux intermédiaires permettant d'interpréter les graphiques (Histogramme de valeurs propres, coordonnées ; contributions et cosinus carrés)

Nous pouvons changer d'éditeur. Ainsi, par exemple, l'utilisateur pourra travailler avec le traitement sont en cours.

Avec cette commande, on peut choisir, pour consultation, un fichier différent de ceux qui de texte qu'il utilise couramment.

Ainsi, si nous demandons à consulter le fichier résultats nous obtenons l'écran suivant à partir duquel on peut parcourir l'ensemble de ces informations :

```
------ SPAD.N - MODIFICATION DE PARAMETRES ------------------------------------------

LISTAGE DES PARAMETRES DE COMMANDE
----------------------------------

   ---+---1---+---2---+---3---+---4---+---5---+---6---+---7---+---8
 1 LISTP = OUI, LISTF = NON, LERFA = OUI : Paramètres généraux
 2 NDICZ = "SPADEX.LAD"      : Fichier des libellés
 3 NDONZ = "SPADEX.DAD"      : Fichier des données
 4
 5 NDICA = "SPADEX.LAR"
 6 PROC ARDIC       : Lecture des libellés
 7 Archivage des libellés
 8 LDICZ = EXT, LTYPE = LARGE, LEDIT = COURT, NQEXA = 5, NXMOD = 0
 9
10 NDONA = "SPADEX.DAR"
11 PROC ARDON       : Lecture des données
12 Archivage des données
13 NIDI = 1, NQEXA = 5, NIEXA = 12 >
14 LDONZ = EXT, NEDIT = NON, LEXTR = OUI >
15 TEST = 999999.00, NLFOR = -1, NCOLZ = 80
16
TAB.LST Lecture (^E:Ecriture.) 100%      <F1>: Aide  <ALT X>: Fin
```

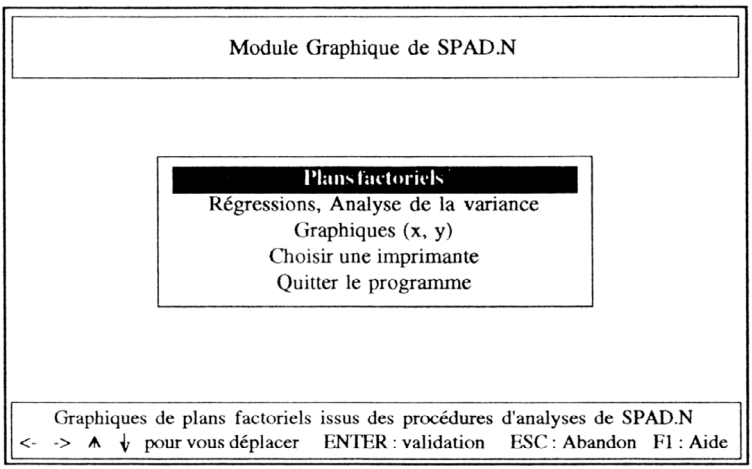
Comme nous pouvons le voir dans la ligne de message, la commande <ALT X> permet de quitter l'éditeur avec un retour à l'écran Initial.

2.5. La création des graphiques

Lors du traitement du tableau que nous avons expliqué, le logiciel crée, dans le fichier résultats, un graphique en format ASCII. Cette représentation n'étant pas en mode graphique, elle ne donne pas la distance exacte des points.

C'est pourquoi SPAD-N comporte un module graphique permettant de créer les représentations au pixel (point) près.

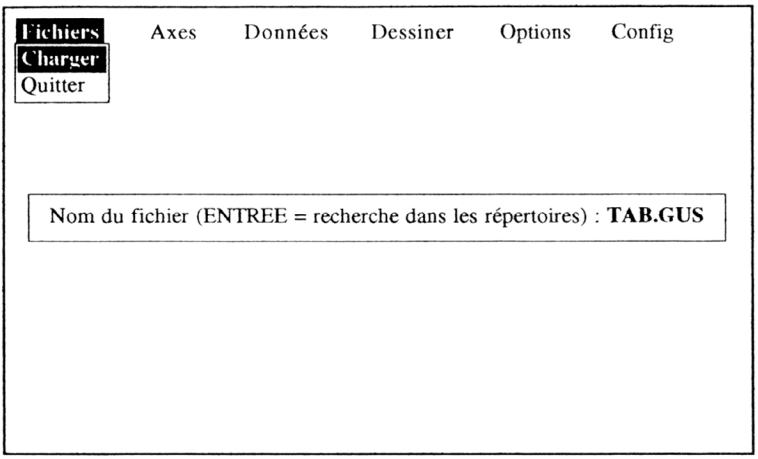
Pour charger ce module nous lancerons la commande Graphiques de l'écran initial. Un premier écran apparaît :



Comme le montre cette illustration, plusieurs choix nous sont offerts :

- Trois types de représentation graphique : Plans factoriels, régression et analyse de la variance ou graphiques x et y.
- Le choix d'une imprimante pour l'impression des graphiques.
- L'abandon de ce module (Quitter) avec un retour au menu initial.

Nous choisirons la première de ces options qui correspond à l'analyse factorielle des correspondances et nous verrons ainsi s'afficher un nouvel écran :

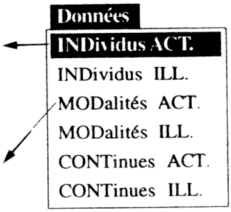


L'écran comporte six menus. Le premier permet de charger un fichier (toujours d'extension .GUS) comme nous le voyons ci-dessus. Il offre également la possibilité de quitter le module graphique (quitter).

Une fois que l'on a choisi le fichier où se trouvent stockées les coordonnées, il est nécessaire de définir les éléments que l'on veut représenter. Pour cela, nous utiliserons deux commandes du menu Données :

■ La commande INDividus ACT. qui permet de charger les individus (lignes) ou ensemble I.

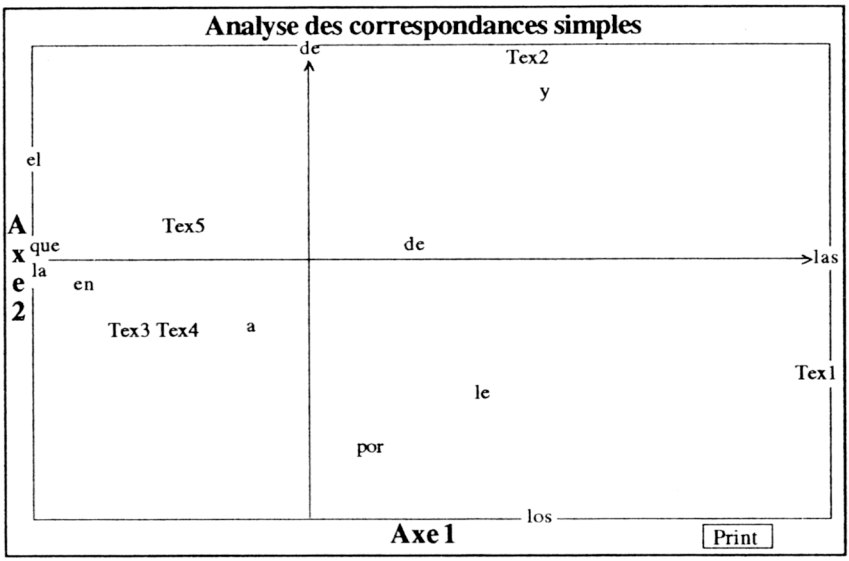
■ La commande MODalités ACT. qui permet également le chargement des modalités (colonnes) ou ensemble J.



Le Menu *Données* du Module Graphique

Les commandes intitulées ILL. (ILLustratives) correspondent aux modalités nominales ou illustratives dont nous avons déjà parlé.

Il ne reste donc plus qu'à créer le graphique correspondant en exécutant la commande Dessiner située dans la barre de menu de cet écran. Immédiatement, la représentation s'affiche :

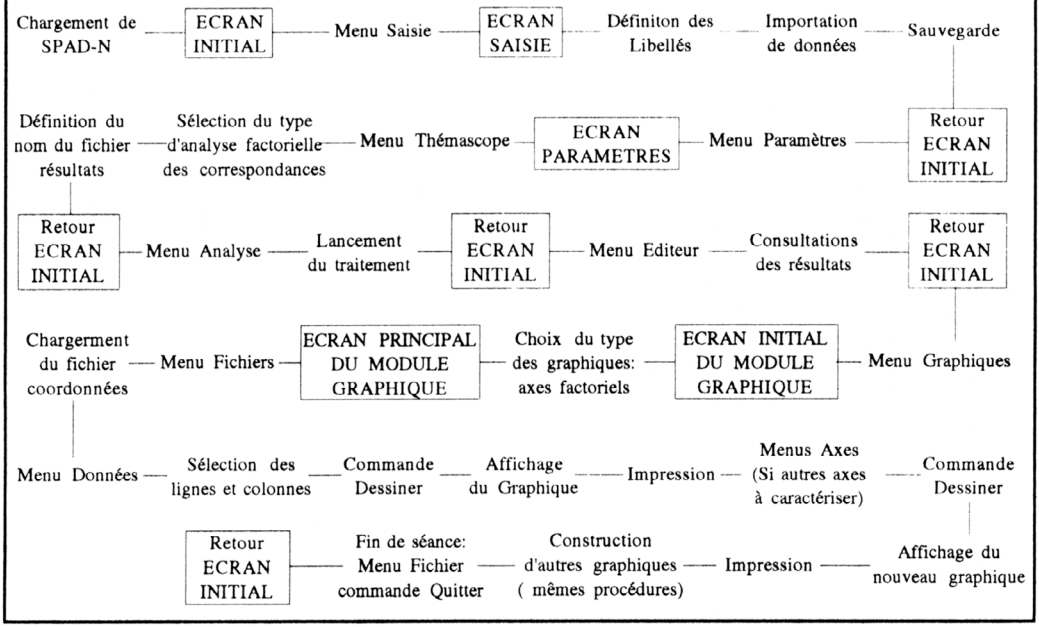


Comme le montre l'illustration ci-dessus, le graphique comporte les informations suivantes :

- le titre du graphique,
- le nom des Axes représentés,
- les points des deux nuages I et J,
- la commande Print qui permet d'imprimer le graphique sur une imprimante matricielle (ou à aiguilles). L'impression sur imprimante laser doit être effectuée avec un logiciel de capture de page.

En conclusion, nous proposons un schéma récapitulatif des étapes que nous réalisons dans le cadre de nos recherches textuelles avec cet excellent Système Portable pour l'Analyse des Données Numériques.

Schéma récapitulatif des étapes pour la construction des graphiques sous SPAD-N



CONCLUSION

Nous espérons que cette présentation, sur les principales procédures de SPAD-N, permettra de montrer l'intérêt de ce programme pour le traitement statistique au lecteur intéressé et ne connaissant pas ce logiciel. Chacun pourra envisager l'utilité selon son domaine d'application, comme par exemple : le traitement des enquêtes, l'analyse des données économiques, techniques, ou encore, en ce qui nous concerne, des données textuelles.

En tout cas, SPAD-N est un outil de statistiques très précieux pour tous ceux qui souhaitent analyser des grandes masses de données. SPAD-N mérite donc notre meilleure appréciation, même s'il est nécessaire d'acquérir une très bonne formation à la pratique de ce logiciel, à laquelle nous espérons avoir déjà contribué à travers cette présentation.

Javier SANCHEZ
Université de Limoges