Chapitre 7

Partitions longitudinales, contiguïté

Les textes rassemblés en corpus constituent souvent un ensemble sur lequel nous possédons de nombreuses informations externes (auteur, genre, date de rédaction, dans le cas des corpus rassemblant des textes dus à différents auteurs, catégories socioprofessionnelles, classes d'âge, degré d'instruction dans le cas des agrégats de réponses libres). Ces informations nous permettent d'établir, avant toute expérience de caractère quantitatif visant à rapprocher certaines parties sur la base du stock lexical qu'elles utilisent, une série de relations (même auteur, dates voisines, même genre, etc.).

Le présent chapitre introduit des méthodes qui permettront de confronter analyses lexicales et informations a priori pour un corpus dont les parties sont munies de telles relations : il ne s'agit plus de découvrir ou de redécouvrir des structures, mais d'éprouver la réalité de structures connues, et d'évaluer le niveau de dépendance des typologies obtenues à partir des textes vis-à-vis de ces informations a priori.

7.1 Les trois structures de base

On peut voir sur la figure 7.1 des représentations qui correspondent à trois structures distinctes mises en relation avec un même corpus de textes muni d'une partition en neuf parties. Ces représentations seront formalisées en *graphes de contiguïté* ultérieurement dans ce chapitre.

Les schémas 1 et 2 peuvent correspondre à une situation dans laquelle trois éditorialistes (A, B et C) d'un même quotidien ont écrit à tour de rôle un éditorial pendant neuf jours de parution du journal.

Ces textes ont été rassemblés ensuite en un même corpus lexicométrique qui compte neuf parties. On peut désigner la suite de ces articles par :

A1, B1, C1, A2, B2, C2, A3, B3, C3

Le corpus ainsi constitué peut être examiné par rapport à deux structures a priori, notées S1, et S2.

La première de ces structures, S1, tient compte exclusivement de la variable *auteur de l'article*. C'est-à-dire que l'on considère comme contigus pour cette structure, les articles dûs à un même auteur.

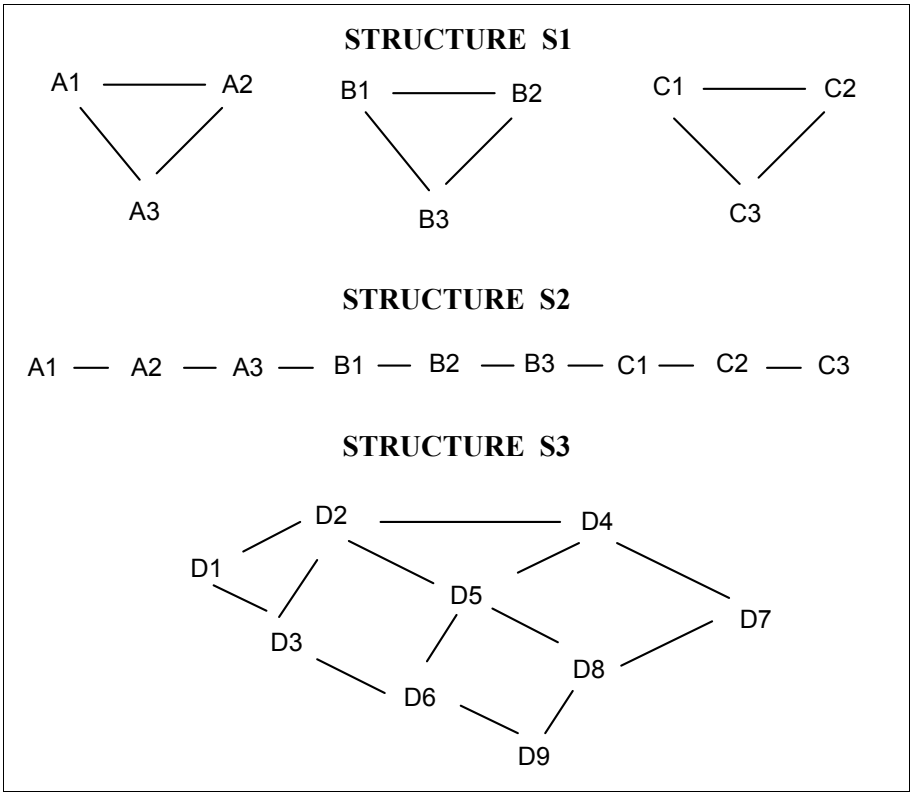

Figure 7.1

Graphes correspondant à trois types de structures courantes : Partition, chronologie, graphe non orienté.

La structure S2 rend compte de la variable *consécutivité dans le temps*. Sont considérés comme contigus, dans ce second cas, les articles rédigés consécutivement sans considération d'auteur.

On rencontre des structures plus générales, représentées par le schéma S3 qui correspond au cas d'un corpus entre les parties duquel existeraient des relations de contiguïté (thématiques, appartenance politique, documents relatifs à des régions limitrophes, etc.).

Les structures de *partition* (type S1) sont d'un usage courant en statistique. L'outil privilégié dans le cas où les variables étudiées sont numériques est

l'*analyse de la variance*, technique classique qui permet de tester l'hétérogénéité de certains regroupements de variables.

Les structures de *séries chronologiques* (type S2) sont étudiées par les méthodes d'une autre branche importante de la statistique : les *séries temporelles et processus stochastiques.*

Les structures de *graphe* (type S3), plus générales, relèvent de techniques moins largement répandues et moins développées que les précédentes, les *analyses statistiques de contiguïté*. Cependant, ces techniques permettent de procéder facilement aux inférences correspondant aux situations les plus courantes. Dans la mesure où elles embrassent comme cas particuliers les deux types S1 et S2, ces techniques fournissent un outil unique pour la prise en compte de la plupart des structures a priori susceptibles d'être observées par le praticien de la statistique textuelle.

Dans la phase délicate de l'interprétation des résultats, les méthodes de l'analyse statistique de la contiguïté interviennent comme complément des analyses purement exploratoires. Elles permettront en fait de prolonger un temps la démarche formelle et de réduire encore la part de subjectivité inhérente à tout commentaire.

Ce chapitre se poursuit par un rappel qui concerne l'analyse statistique de la contiguïté dans le cas, le plus simple, de l'analyse des valeurs d'une seule variable (7.2). La démarche est ensuite étendue à l'étude des facteurs issus d'une analyse des correspondances réalisée à partir d'un tableau lexical (7.3). Le paragraphe (7.4) est consacré à l'application des méthodes présentées à un corpus expérimental composé d'agrégats de réponses libres à une question ouverte. On introduit ensuite (7.5) la problématique particulière des structures longitudinales et l'on présente une application particulièrement importante des analyses de données longitudinales dans le domaine textuel : les séries textuelles chronologiques (7.6). Enfin, une application de l'analyse de la contiguïté au problème de la *recherche en homogénéité d'auteurs* terminera ce chapitre (7.7).

7.2 Homogénéité des valeurs d'une variable par rapport à une partition.

Commençons par le cas le plus simple d'une variable prenant des valeurs numériques sur les éléments d'un ensemble. Par la suite, les notions introduites seront utilisées pour étudier l'ensemble des facteurs sur l'ensemble des parties issus de l'analyse des correspondances d'un tableau lexical. Dans ce qui suit, J symbolise un ensemble d'éléments, p le nombre

de ces éléments. Il est possible de définir sur l'ensemble (J ∞ J) une relation dite : *relation de contiguïté*. Nous noterons cette relation R(j,j'). Cette relation est symétrique (si R(j,j'), alors R(j',j)). Nous supposerons ici de plus que cette relation n'est pas réflexive (R(j,j) n'est jamais vrai) ce qui revient à poser, de façon conventionnelle, que chaque partie j n'est pas contiguë à elle même.[1]

7.2.1 Graphe associé à une structure de contiguïté

Cette relation de contiguïté peut être représentée de différentes façons. Une des représentations traditionnelles s'effectue à l'aide d'un graphe. Dans ce qui suit nous désignerons ces *graphes de contiguïté* par les symboles G, éventuellement indicés (G_1, G_2, G_3...).

Pour cette représentation, chaque élément j de l'ensemble J correspond à un *sommet* du graphe. Chacun des couples de sommets du graphe liés par la relation de contiguïté R(j,j') est relié par une *arête*. Le nombre des arêtes adjacentes à un même sommet j est appelé le *degré* de ce sommet. Nous noterons ce nombre m_j. Si tous les sommets sont reliés par une arête, le graphe est dit *complet*. Un tel graphe possède *p(p-1)* arêtes.

Il est usuel de distinguer les arêtes selon leur sens. Ainsi, l'arête qui va du sommet j au sommet j' est, en principe, distincte de celle qui joint j' à j. On note *m* le nombre des arêtes du graphe.

En fait on se limitera ici, pour les relations a priori qui nous intéressent, aux graphes symétriques (tels que R(j,j') □ R(j',j)), que l'on peut aussi considérer comme des graphes non-orientés avec *m/2* arêtes.

7.2.2 Matrices de contiguïté

On peut également représenter la relation de contiguïté R(j,j') par le biais d'une matrice carrée **M** dite *matrice de contiguïté*. Cette matrice dans laquelle chaque élément vaut 0 ou 1 possède autant de lignes et de colonnes que l'ensemble J compte d'éléments.

Son terme général $m_{jj'}$ vaut :

[1] Dans la littérature sur le sujet on définit parfois des structures de contiguïté qui incluent des proximités à distance 1, 2, ...n , cf. par exemple Lebart (1969). Nous nous limiterons ici aux structures de contiguïté pour lesquelles deux parties sont immédiatement contiguës (distance 1) ou disjointes bien que les résultats soient également généralisables à des structures de contiguïté plus complexes.

$m_{jj'} = 1$ si j et j' sont contigus ;
$m_{jj'} = 0$ dans le cas contraire.

On voit que cette matrice est symétrique du fait de la symétrie de la relation de contiguïté. Selon la convention posée plus haut, les termes situés sur la diagonale principale de la matrice **M** sont tous nuls.

Tableau 7.1 **Matrice de contiguïté relative à la structure S1**

	A1	B1	C1	A2	B2	C2	A3	B3	C3
A1	0	0	0	1	0	0	1	0	0
B1	0	0	0	0	1	0	0	1	0
C1	0	0	0	0	0	1	0	0	1
A2	1	0	0	0	0	0	1	0	0
B2	0	1	0	0	0	0	0	1	0
C2	0	0	1	0	0	0	0	0	1
A3	1	0	0	1	0	0	0	0	0
B3	0	1	0	0	1	0	0	0	0
C3	0	0	1	0	0	1	0	0	0

Tableau 7.2 **Matrice de contiguïté relative à la structure S2**

	A1	B1	C1	A2	B2	C2	A3	B3	C3
A1	0	1	0	0	0	0	0	0	0
B1	1	0	1	0	0	0	0	0	0
C1	0	1	0	1	0	0	0	0	0
A2	0	0	1	0	1	0	0	0	0
B2	0	0	0	1	0	1	0	0	0
C2	0	0	0	0	1	0	1	0	0
A3	0	0	0	0	0	1	0	1	0
B3	0	0	0	0	0	0	1	0	1
C3	0	0	0	0	0	0	0	1	0

Les relations suivantes lient le nombre d'arêtes m, les degrés m_j , et les éléments $m_{jj'}$ de la matrice **M** :

$$m = \sum_{j=1}^{p} m_j = \sum_{j=1}^{p} \sum_{j'=1}^{p} m_{jj'},$$

Les tableaux 7.1 et 7.2 donnent les matrices de contiguïté correspondant aux structures S1 et S2.

Si la matrice du tableau 7.1 voit ses lignes et ses colonnes réordonnées simultanément, de façon à regrouper les symboles par leurs premières lettres A, B ou C, on voit mieux apparaître les trois groupes qui constituent cette structure (tableau 7.3).

Tableau 7.3

Matrice de contiguïté relative à la structure S1, après reclassement des lignes et des colonnes

	A1	A2	A3	B1	B2	B3	C1	C2	C3
A1	**0**	**1**	**1**	0	0	0	0	0	0
A2	**1**	**0**	**1**	0	0	0	0	0	0
A3	**1**	**1**	**0**	0	0	0	0	0	0
B1	0	0	0	**0**	**1**	**1**	0	0	0
B2	0	0	0	**1**	**0**	**1**	0	0	0
B3	0	0	0	**1**	**1**	**0**	0	0	0
C1	0	0	0	0	0	0	**0**	**1**	**1**
C2	0	0	0	0	0	0	**1**	**0**	**1**
C3	0	0	0	0	0	0	**1**	**1**	**0**

7.2.3 Le coefficient de contiguïté

Imaginons maintenant une variable *Z* prenant des valeurs z_j sur chacun des sommets du graphe G défini plus haut. Ces valeurs peuvent correspondre à des mesures faites sur chacune des parties du corpus (la longueur moyenne des phrases dans chaque partie par exemple ou toute autre grandeur ne dépendant pas directement de la longueur de chacune des parties[1]).

Nous commencerons par calculer la moyenne $\bar{z}$ des valeurs de cette variable sur les sommets du graphe G.

$$\bar{z} = \frac{1}{p} \sum_{j=1}^{p} z_j$$

La *variance empirique* de la variable Z vaut :

$$v(Z) = \frac{1}{2p(p-1)} \sum_{j'=1}^{p} \sum_{j=1}^{p} (z_j - z_{j'})^2 \qquad (1)$$

que l'on notera aussi :

$$v(Z) = \frac{1}{2p(p-1)} \Sigma^{jj'} (z_j - z_{j'})^2$$

On retrouve en développant la partie droite, la formule plus classique :

[1] En effet, si la variable z dépend directement de la longueur des parties du corpus l'étude de sa variation au fil des parties risque fort de se confondre avec l'étude de la différence de longueur entre les parties.

$$v(Z) = \frac{1}{(p-1)} \sum_{j=1}^{p} (z_j - \bar{z})^2 ,$$

La variance *v(Z)*, que l'on appelle ici variance totale, mesure la dispersion des valeurs z_j autour de leur moyenne $\bar{z}$ sur l'ensemble des sommets du graphe G.

On calcule la *variance locale v*(Z)* de la variable Z sur les seuls sommets contigus du graphe G au moyen d'une formule analogue à la formule (1) mais pour laquelle la sommation s'effectue cette fois pour les seuls couples j et j' qui sont reliés par une arête du graphe (au lieu de la sommation sur tous les j et j'). La sommation selon ce dernier critère est représentée par le symbole Σ^*.

$$v^*(Z) = \frac{1}{2m} \Sigma^* (z_j - z_{j'})^2$$

ou de façon équivalente :

$$v^*(Z) = \frac{1}{2m} \sum_{j'=1}^{p} \sum_{j=1}^{p} m_{jj'} (z_j - z_{j'})^2 ,$$

que l'on note aussi :

$$v^*(Z) = \frac{1}{2m} \Sigma^{jj'} m_{jj'} (z_j - z_{j'})^2,$$

Rappelons que dans ces formules[1]:

— $m = \sum_{j=1}^{p} m_j = \Sigma^{jj'} m_{jj'}$, est le nombre total des connexions.

— Σ^* indique une sommation faite pour l'ensemble des sommets j et j' tels que j et j' soient contigus.

— $\Sigma^{jj'}$ indique une sommation sur l'ensemble des sommets j et j' pris deux à deux.

— $\bar{z} = (1/p) \sum_{j=1}^{p} z_j$, est la moyenne des z_j.

[1] Il faut remarquer que, compte tenu des notations que nous avons adoptées, la distinction (ou la non-distinction) des arêtes selon leur orientation amène un calcul légèrement différent pour un résultat identique.

Le coefficient de contiguïté $c(Z)$ (Geary, 1954) se calcule alors comme :

$$c(Z) = \frac{v^*(Z)}{v(Z)} = \frac{p(p-1)\sum^{*}(z_j - z_{j'})^2}{m\sum^{jj'}(z_j - z_{j'})^2}$$

Ce rapport est proche de l'unité si la variance locale (la variance mesurée sur les sommets contigus) est proche de la variance mesurée sur l'ensemble des sommets du graphe, ce qui indique une présomption de répartition aléatoire sur le graphe. Plus il est proche de zéro, plus on peut dire que la variable z prend sur les sommets contigus des valeurs voisines par rapport aux sommets pris deux à deux de façon quelconque. Si le coefficient est nettement supérieur à l'unité on en conclura que les valeurs de la variable *z* sont, au contraire, "anormalement" différentes sur les sommets voisins.

7.2.4 Calcul des moments du coefficient de contiguïté

Sous l'hypothèse selon laquelle les valeurs z_j peuvent être considérées comme des réalisations de variables aléatoires normales indépendantes, on peut calculer les quatre premiers moments du coefficient $c(Z)$[1].

A partir de ces moments, on détermine μ_1, μ_2, μ_3 et μ_4, moments centrés d'ordre 2, 3 et 4 pour cette même distribution.

Le coefficient d'asymétrie de Fisher :

$$\gamma_1 = \mu_3 / \sigma^3 = \mu_3 / \mu_2^{3/2}$$

et le coefficient d'aplatissement :

$$\gamma_2 = \mu_4 / \sigma^4 - 3 = \mu_4 / \mu_2^2 - 3$$

permettent de comparer la distribution empirique à une loi normale. Si γ_1 et γ_2 sont proches de zéro on appliquera, en première approximation, le test de l'écart réduit pour juger de l'écart du coefficient $c(Z)$ par rapport à l'unité.

7.2.5 Un cas particulier : les séries temporelles

Le coefficient $c(Z)$ constitue une généralisation du coefficient de Von Neumann (1941). En effet, dans le cas où la relation de contiguïté se réduit à une relation de consécutivité comme c'est le cas, par exemple, pour la structure S2 présentée plus haut, le numérateur du coefficient de contiguïté peut s'écrire :

[1] Pour un exposé plus complet cf. Lebart (1969). Pour d'autres applications de la notion de contiguïté, cf. Aluja Banet et al. (1984), Burtschy et al. (1991).

$$c = \frac{1}{2(p-1)} \frac{\sum_{j=2}^{p} (z_j - z_{j-1})^2}{v(Z)}$$

puisque le nombre total des connexions dans une structure de consécutivité avec p éléments est égal à *p-1*.

Dans ce dernier cas, le coefficient de Geary est une des formes du coefficient (plus classique) de Von Neumann qui mesure l'autocorrélation d'une suite de valeurs z_j.

7.2.6 Utilisation du coefficient *c*

Revenons à l'exemple évoqué plus haut et imaginons une variable *Z* prenant des valeurs sur les neuf parties d'un corpus (par exemple, la fréquence relative des occurrences de la forme *de*). Nous commencerons par calculer le coefficient *c(Z)* pour les structures de contiguïté S_1 et S_2. Notons ces nombres respectivement $c(Z \mid S_1)$ et $c(Z \mid S_2)$.

Si le coefficient $c(Z \mid S_1)$ est très proche de l'unité, nous dirons que la variable z_j ne subit pas de variation notable liée à la différence d'auteur. Dans le cas où le coefficient $c(Z \mid S_2)$ serait très proche de l'unité nous dirions que la variable z_j ne subit pas d'effet chronologique.

La proximité à l'unité s'analyse dans la pratique en fonction des moments calculés au paragraphe 7.2.4. Dans tous les cas cependant, si les valeurs z_j varient moins pour les textes produits par un même auteur qu'elles ne varient par rapport au temps, le calcul du coefficient $c(Z \mid S_1)$ amènera une valeur nettement plus petite que la valeur $c(Z \mid S_2)$. Dans le cas contraire, d'un coefficient $c(Z \mid S_2)$ plus petit nous dirons que c'est l'effet chronologique qui prime au contraire sur l'effet "auteur".

7.3 Homogénéité des facteurs d'une analyse des correspondances en fonction d'une structure a priori

Comme plus haut, on supposera que l'on est en présence d'un corpus de textes divisé en *p* parties par la partition P. De plus, on dispose d'une structure S1, structure a priori de contiguïté sur l'ensemble des parties.

A cette structure correspond la matrice de contiguïté **M**. A partir de ce corpus, nous avons construit la table de contingence (formes x parties) correspondant à cette partition. Ce tableau qui compte n lignes et p colonnes

a ensuite été soumis à une analyse des correspondances afin d'en extraire *(p-1)* facteurs.

Les facteurs sur l'ensemble des parties sont notés $F_\alpha(j)$ où l'indice α indique le numéro du facteur et varie de 1 à *(p-1)*. L'indice j indique ici le numéro de partie et varie entre 1 et *p*.

7.3.1 Homogénéité d'un facteur par rapport une structure.

Le calcul du coefficient de contiguïté pour la suite des valeurs $F_\alpha(j)$ permet de mesurer le lien entre ce facteur et la structure de contiguïté S1. Pour chaque facteur ce coefficient se calcule selon la formule employée pour les variables numériques :

$$c(F_\alpha) = \frac{(p-1)\ \sum^{*}_{j} (F_\alpha(j) - F_\alpha(j'))^2}{2m \sum_{j} (F_\alpha(j) - F_\alpha(.))^2} \qquad (2)$$

Dans ce qui suit nous noterons ce coefficient c_α.

On a noté $F_\alpha(.) = \frac{1}{p} \sum_{j=1}^{p} F_\alpha(j)$ la moyenne des valeurs du facteur. Cette quantité est nulle si toutes les parties ont le même poids.

7.3.2 Homogénéité dans l'espace des k premiers facteurs.

L'analyse réalisée pour chacun des facteurs pris isolément par rapport à une structure de contiguïté peut être heureusement complétée par une analyse portant sur la distance entre les éléments de l'ensemble J dans l'espace engendré par les *k* premiers facteurs.

Le carré de la distance du chi-2 entre deux points j et j' s'écrit en fonction du tableau des profils :

$$d^2(j,j') = (1/f_i) \sum_{i=1}^{N} (f_{ij}/f_{.j} - f_{ij'}/f_{.j'})^2$$

Cette même quantité s'écrit :

$$d^2(j,j') = \sum_{\alpha=1}^{n_f} (F_\alpha(j) - F_\alpha(j'))^2$$

en fonction des n_f facteurs issus de l'analyse des correspondances.

Le carré de la distance entre deux points j et j' dans l'espace engendré par les k premiers facteurs s'écrit :

$$d_k^2(j,j') = \sum_{\alpha=1}^{k} \left(F_\alpha(j) - F_\alpha(j')\right)^2$$

La quantité :

$$G_k = \frac{p(p-1) \sum^{*} d_k^2(j,j')}{m \sum^{jj'} d_k^2(j,j')} \qquad (3)$$

constitue donc un coefficient *c* calculé à partir des distances $d_k^2(j,j')$ prises cette fois dans le sous-espace des *k* premiers facteurs.

Comme on le voit d'après les formules (2) et (3), dans le cas ou *k* est égal à 1 on a l'égalité $G_1 = c\ (F_1)$. Lorsque *k* est égal à *(p-1)* nombre total des facteurs issus de l'analyse des correspondances, le coefficient G_k rend compte du rapport entre la variance "sur la structure" et la variance totale pour la distance du chi-2 calculée sur le tableau des fréquences. Les valeurs intermédiaires de *k* correspondent à des distances filtrées par les premiers facteurs.

Dans la mesure où les parties n'ont pas toutes la même importance numérique, on calculera dans le cas de corpus dont les parties sont de taille très dissemblable la quantité :

$$G_k = \frac{p(p-1) \sum^{*} p_{.j}\, p_{.j'}\, d_k^2(j,j')}{m \sum^{jj'} p_{.j}\, p_{.j'}\, d_k^2(j,j')} \qquad (4)$$

pour laquelle les distances $d_k^2(j,j')$ sont pondérées par le poids relatif de chacune des parties.

7.4 Les agrégats de réponse et l'analyse de la contiguïté

Les coefficients que nous venons d'introduire vont nous permettre de formaliser le dépouillement des résultats de l'analyse des correspondances sur le tableau des agrégats présenté au chapitre 5 (table 5.3, puis figure 5.1). Dans le cas des agrégats de réponses constitués à partir de la variable

nominale composite (âge x diplôme), la matrice de contiguïté $\mathbf{S_{ad}}$ (ad : comme âge x diplôme) peut être représentée par le graphe de la figure 7.2. Pour chacun des axes factoriels issus de l'analyse du tableau (formes x agrégats), on calcule alors les coefficients c_α et G_α

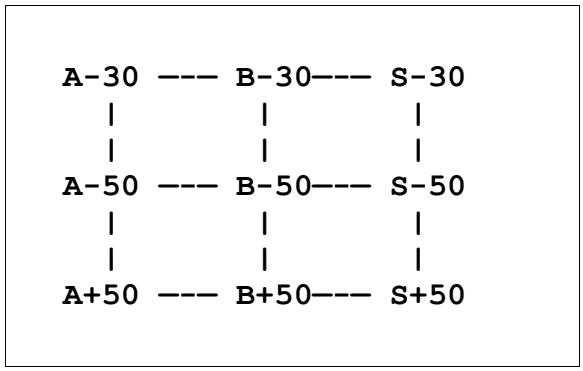

Figure 7.2

Graphe de contiguïté décrivant la structure des colonnes de la table 5.3 (chapitre 5)

Structure et filtrage

L'exemple qui suit concerne l'analyse d'un tableau dont les lignes correspondent aux 321 formes les plus fréquentes pour la question ***Enfant*** (au lieu de 117 pour l'exemple du chapitre 5) et les colonnes correspondent à la même partition en 9 classes croisant *âge* et *niveau de diplôme*.

Tableau 7.4

Valeurs propres issues de l'analyse d'un tableau (321 formes x 9 agrégats) et coefficients c_α et G_α calculés par rapport à la structure de contiguïté S_{ad} pour chacun des facteurs

Axe	*Valeur propre*	c_α	G_α
1	0.057	0.397	0.397
2	0.037	0.400	0.398
3	0.028	1.489	0.698
4	0.027	1.071	0.761
5	0.025	1.190	0.849
6	0.022	1.120	0.873
7	0.021	1.206	0.909
8	0.017	1.093	0.926

Comme on pouvait le prévoir au vu des résultats de l'analyse de ce tableau (figure 5.1) les coefficients c_α et G_α qui correspondent aux deux premiers facteurs ont ici des valeurs faibles, signe d'une bonne représentation de la structure $\mathbf{S_{ad}}$ par le plan des deux premiers axes. Le calcul de ces mêmes coefficients pour les axes suivants permet de vérifier que ces axes ne possèdent pas cette propriété constatée à propos des deux premiers.

Dans le cas particulier du corpus ***Enfant*** muni de la partition (âge x diplôme) que nous venons d'étudier, les calculs de contiguïté (tableau 7.4) confirment un résultat aisément perceptible lors de l'examen du plan des deux premiers facteurs (figure 5.1, du chapitre 5). Mais ils montrent surtout que ces *deux premiers facteurs* ont d'une certaine façon *l'exclusivité* de cette structure. La croissance du coefficient G_α montre également que cette structure est de plus en plus floue lorsque la dimension augmente : le coefficient de contiguïté vaut 0.926 pour les distances calculées dans l'espace à 8 dimensions.

Autrement dit, le lien entre la structure a priori et les distances (du chi-2) globales est faible. La propriété de *filtrage* des premiers axes de l'analyse des correspondances, si souvent remarquée par les praticiens, et cependant si difficile à définir et étayer théoriquement, est encore vérifiée sur cet exemple. Le squelette schématisé par la figure 7.2 est à peine décelable dans l'espace, et pourtant représenté sans interversion par le premier plan factoriel (figure 5.1). On emploie ici à dessein le mot squelette, car l'analyse des correspondances a été souvent comparée à un appareil radiographique.

7.5 Partitions longitudinales d'un corpus

La relation de consécutivité constitue le cas le plus simple de relation de contiguïté sur les parties d'un corpus. On obtient des partitions de ce type, lorsque l'on construit, par exemple, des agrégats de réponses à une question ouverte à partir des valeurs d'une même variable numérique (âge, revenu mensuel, ou nombre d'enfants) ou encore à partir d'une variable dont les différentes modalités peuvent être rangées dans un ordre qui correspond à un degré d'intensité pour cette variable (niveau d'instruction par exemple). Tel est le cas encore lorsqu'on range les parties d'un corpus de textes selon un ordre chronologique (rédaction, publication, etc.)

Dans ce qui suit on appellera ce type de partition : des *partitions longitudinales* d'un corpus en fonction d'une variable. On examinera successivement le cas de réponses à une question ouverte regroupées selon les modalités d'une variable ordonnée (classes d'âges), puis, dans le

paragraphe suivant, dans le cadre plus général des séries textuelles chronologiques, le cas d'une série de discours politiques.

7.5.1 Exemple de partition longitudinale

On désigne par *corpus* ***Life*** un ensemble des réponses fournies par 1 000 personnes de langue anglaise à une question ouverte portant les choses qu'elles jugent personnellement importantes pour elles dans la vie[1]. Le libellé de la question était : *"What is the single most important thing in the life for you ? "* suivie d'une relance *"What other things are very important to you ?"*.

En constituant des agrégats de réponses fondés sur l'appartenance des répondants à une même classe d'âge, nous pouvons espérer mettre en évidence des variations liées à la variable "âge du répondant".

Tableau 7.5

Découpage en 6 classes du corpus des réponses à la question *Life* d'après l'âge des répondants.

code	val.	occ.	formes	hapax	fmax	
Ag1	18-24	2003	429	247	112	my
Ag2	25-34	2453	512	295	148	my
Ag3	35-44	2529	510	290	151	family
Ag4	45-54	2519	493	258	159	my
Ag5	55-64	1864	459	264	110	my
Ag6	65-++	3326	623	333	144	health

Le tableau 7.5 montre les bornes retenues pour effectuer le découpage des agrégats en 6 classes notées (Ag1, Ag2, …, Ag6) en fonction de l'âge. Ce découpage amène, on le voit, des parties dont la taille est comparable. Remarquons que le simple calcul de la fréquence maximale suffit à faire apparaître une préoccupation *(health)* propre aux personnes les personnes les plus âgées.

[1] Cette question est extraite d'une enquête par sondage réalisée à la fin des années quatre-vingt au Royaume Uni. Il s'agit du volet britannique d'une enquête comparative internationale réalisée auprès de cinq pays (Japon, Allemagne, France, Grande Bretagne et USA,) sous la direction des Pr. C. Hayashi, M. Sasaki et T. Suzuki. Pour la série de travaux dans lesquels s'inscrit cette opération, cf. Hayashi et al.(1992). Cf. aussi Sasaki et Suzuki (1989), Suzuki (1989). Cette enquête se situe elle-même dans la lignée d'une enquête permanente sur le *Japanese National Character* réalisée régulièrement depuis 35 ans au Japon (Hayashi, 1987). Les réponses des individus à la question et à la relance ont été regroupées ici. Les différences dans l'ordre des réponses fournies par les sujets interrogés font par ailleurs l'objet d'une étude particulière.

Classification sur les agrégats

La figure 7.3 montre le résultat d'une classification ascendante hiérarchique (selon le critère de Ward généralisé, comme les méthodes présentées au chapitre 4) obtenue à partir de cette partition. Le dendrogramme obtenu à partir de l'analyse du tableau (6 classes d'âge, 483 formes de fréquence supérieure ou égale à 3) possède une propriété particulière : ses noeuds rassemblent toujours des classes d'âge consécutives.

Les classes extrêmes (Ag1 et Ag6) s'agrègent très haut, dans l'arbre de classification, à l'ensemble constitué par les autres classes, signe d'une particularité plus marquée par rapport aux classes intermédiaires. Comme plus haut, nous interpréterons cette circonstance en concluant que les répondants qui appartiennent à des classes d'âge voisines produisent des réponses plus semblables entre elles que les répondants qui appartiennent à des classes d'âge plus distantes.

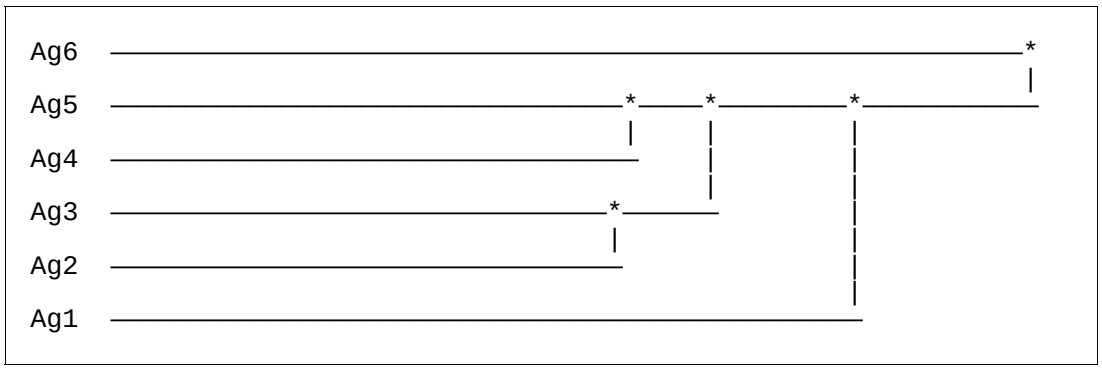

Figure 7.3

Classification ascendante hiérarchique réalisée à partir d'une partition en 6 classes d'âge de l'ensemble des réponses à la question *Life*.

7.5.2 Analyse de la gradation "classe d'âge"

L'analyse des correspondances réalisée à partir d'un tableau, pour lequel l'ensemble des distances entre agrégats est dominé par l'existence d'une gradation, produit des résultats d'un type particulier.

Dans le cas ou la gradation est absolument régulière, la représentation des éléments sur les plans engendrés par les premiers facteurs acquiert une forme très caractéristique. Les différents facteurs à partir du second se révèlent des fonctions du premier (quadratique pour le facteur numéro 2, cubique pour le facteur numéro 3, etc.). L'analyse des correspondances représente de manière

plutôt complexe (en introduisant des facteurs qui sont des fonctions polynomiales du premier) un phénomène qui est fondamentalement de nature unidimensionnelle. Dans la littérature statistique ce phénomène est connu sous le nom d'*effet Guttman*, du nom du statisticien qui étudia de manière systématique[1] les tableaux de données correspondants.

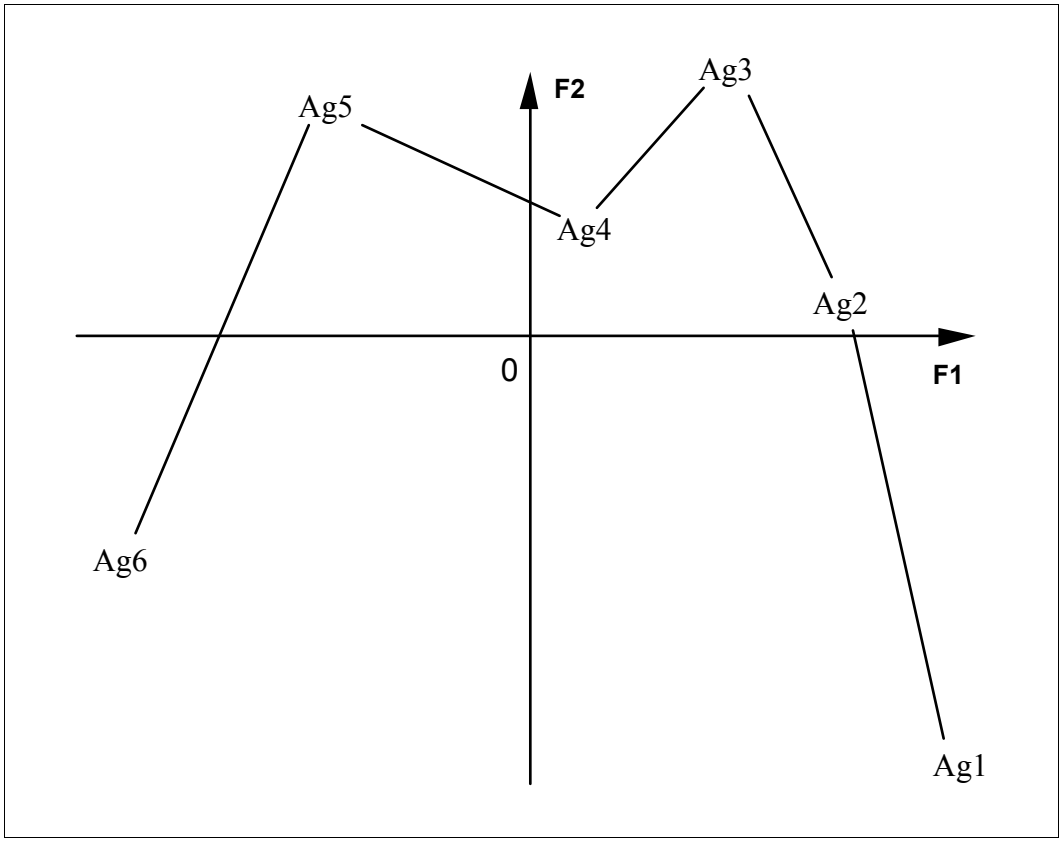

Figure 7.4

L'effet Guttman : représentation schématique du plan factoriel 1 x 2 tableau (6 classes d'âges x 483 formes).
Question ouverte *Life*

Devant des résultats de ce type, on évitera donc, en général, de commenter séparément les oppositions constatées sur chacun des axes factoriels. Ici, le schéma classique de l'interprétation d'une typologie que l'on affine au fur et à mesure par la prise en compte de nouveaux axes factoriels doit faire place à la reconnaissance d'une situation caractéristique globale liée à l'existence et à la dominance d'une gradation.
Lors de l'analyse des correspondances du tableau des distances calculées sur les stocks lexicaux correspondant à chacune des parties du corpus des

[1] Cf. Benzécri (1973) Tome IIb, chapitre 7, Cf. aussi Guttman (1941) et Van Rijckevorsel (1987).

réponses à la question ***Life*** - figure 7.4 - les points correspondants aux classes d'âge (Ag1, ... , Ag6) dessinent sur le plan des deux premiers facteurs une courbe incurvée en son centre qui rappelle le schéma théorique mentionné plus haut. L'examen des facteurs suivant confirme l'impression que l'analyse est bien dominée par l'existence d'une gradation.

Les formes et les segments pourvus d'une coordonnée factorielle extrême correspondent dans ce cas à des termes utilisés par les plus jeunes ou les plus âgés des répondants. Cependant, pour mettre en évidence les termes qui sont responsables, au plan quantitatif, de la gradation d'ensemble constatée sur les premiers facteurs, il faudra mettre en oeuvre des méthodes plus spécifiques.

7.5.3 Spécificités connexes

Pour répondre à cette dernière préoccupation, on va construire des indicateurs permettant de repérer les termes, formes ou segments, dont la ventilation présente des caractéristiques qui peuvent être mises en relation avec la gradation d'ensemble. Considérons un terme de fréquence totale F dont la ventilation dans les n parties d'un corpus est donnée par la liste des sous-fréquences :

$$f_1,\ f_2,\ f_3,\ \ldots,\ f_n.$$

Comme toujours, T désigne la longueur totale du corpus et la liste : $t_1,\ t_2,\ t_3,\ \ldots,\ t_n$, donne la longueur de chacune des parties.

On a vu au chapitre 6 que ces éléments permettent de calculer, pour chacune des sous-fréquences f_i, et pour un seuil de probabilité s fixé, un diagnostic de spécificité S_i qui sera symbolisé par "S+", "S-" ou "b" suivant que nous aurons affaire à une spécificité positive, négative ou à une forme banale pour la partie i.

Convenons encore d'affecter à ces diagnostics un indice supérieur, égal à 1, qui indiquera que nous avons affaire à des spécificités de premier niveau, c'est-à-dire des spécificités portant sur des périodes élémentaires du corpus prises isolément.

Dans ces notations, la liste des spécificités de premier niveau s'écrit donc :

$$S^1_1,\ S^1_2,\ S^1_3,\ \ldots\ S^1_n,$$

On appellera spécificités de niveau 2 les spécificités calculées de la même manière pour chaque couple de parties consécutives.

Tableau 7.6

Spécificités longitudinales majeures pour le corpus des réponses *Life* découpé en 6 classes d'après l'âge des répondants.

terme	F	fp	sp	code
to do	25	13	+E06	Ag1-
being	116	32	+E05	Ag1-
friends	116	32	+E05	Ag1-
good job	13	8	+E05	Ag1-
having	70	20	+E04	Ag1-
a good job	6	5	+E04	Ag1-
a good	54	18	+E04	Ag1-
a	300	62	+E04	Ag1-
money	170	80	+E06	Ag1-Ag2
career	11	9	+E04	Ag1-Ag2
be happy	41	23	+E04	Ag1-Ag2
to get a	6	6	+E04	Ag1-Ag2
being happy	30	19	+E04	Ag1-Ag2
my job	37	29	+E04	Ag1-Ag3
education	25	20	+E04	Ag1-Ag3
job	143	128	+E12	Ag1-Ag4
work	118	100	+E07	Ag1-Ag4
family	686	582	+E07	Ag1-Ag5
my	809	670	+E05	Ag1-Ag5
my family	225	194	+E04	Ag1-Ag5
happiness	227	198	+E04	Ag1-Ag5
the children	25	21	+E07	Ag2-Ag3
children	125	66	+E05	Ag2-Ag3
wealth	11	10	+E04	Ag2-Ag3
of	312	193	+E05	Ag2-Ag4
can	35	8	-E04	Ag2-Ag4
do	47	17	-E05	Ag2-Ag5
security	40	37	+E05	Ag2-Ag5
want	31	10	-E04	Ag2-Ag5
to keep	27	8	-E04	Ag2-Ag5
keep	50	19	-E04	Ag2-Ag5
health	611	564	+E06	Ag2-Ag6
welfare	22	11	+E04	Ag3-
welfare of	13	8	+E04	Ag3-
the	331	155	+E06	Ag3-Ag4
the family	78	47	+E06	Ag3-Ag4
need	10	9	+E04	Ag3-Ag4
to	522	203	-E05	Ag3-Ag5
are	65	57	+E04	Ag3-Ag6
me	33	31	+E04	Ag3-Ag6
be	137	10	-E04	Ag4-
grandchildren	30	30	+E09	Ag4-Ag6
I	248	158	+E04	Ag4-Ag6
they	24	21	+E04	Ag4-Ag6
religion	13	13	+E04	Ag4-Ag6
good health	176	117	+E04	Ag4-Ag6
as	83	44	+E04	Ag5-Ag6
worry	8	7	+E04	Ag6-

Guide de lecture du tableau 7.6

Les spécificités calculées concernent les sous-fréquences des termes (formes ou segments) dont la fréquence dépasse 5 occurrences dans le corpus. Seuls les diagnostics les plus importants ont été retenus dans ce tableau (indice de probabilité inférieur à 10^{-4}).

Les diagnostics retenus ont été reclassés en fonction des catégories caractéristiques sélectionnées par la méthode (des plus jeunes aux plus âgés).

La colonne F indique la fréquence du terme dans l'ensemble du corpus.

La colonne fp indique la fréquence partielle du terme dans la partie ou le groupe de parties consécutives pour lesquelles le diagnostic de spécificité a été établi.

Le diagnostic de spécificité est constitué par un signe et un exposant (+Exx renvoie à une spécificité positive à laquelle est attachée une probabilité d'ordre 1/10 à la puissance xx).

La partie, ou le groupe de parties consécutives, du corpus pour laquelle la spécificité positive à été établie apparaît entre deux barres verticales.

Dans le cas des spécificités de niveau 1, la comparaison porte, pour un terme i et une partie j donnés, sur les quatre nombres $T, t_i, F, f_i,$ la spécificité de niveau 2 se calcule donc à partir des nombres : $T, t_i+t_{i+1}, F, f_i+f_{i+1}$. Les spécificités de niveau 2 sont au nombre de $n-1$. Elles sont notées :

$$S^2_1, S^2_2, S^2_3, \ldots S^2_{n-1},$$

Comme plus haut, l'indice supérieur indique le niveau des spécificités auxquelles nous avons affaire. L'indice inférieur indique, de son côté, le numéro de la première des deux parties consécutives sur lesquelles porte le calcul.

On définit de la même manière, les spécificités de niveau k, c'est-à-dire les n-k+1 spécificités calculées pour des séquences de k parties consécutives qui s'écrivent :

$$S^k_1, S^k_2, S^k_3, \ldots S^k_{n-k+1},$$

Nous conviendrons que le niveau supérieur est celui qui regroupe les deux spécificités de niveau *n-1*. Remarquons que les deux nombres obtenus sont respectivement égaux aux spécificités de niveau 1 calculées pour la dernière et la première partie du corpus.

En effet, les calculs, effectués séparément sur les sous-fréquences d'une forme dans deux sous-ensembles créés par la division du corpus en deux parties connexes dont la réunion est égale à l'ensemble, conduisent nécessairement au même résultat puisque la connaissance de la sous-fréquence f dans l'une des parties permet de calculer la sous-fréquence *(F-f)* dans la partie complémentaire. Ce qui entraîne :

$$S_1^k = S_{k+1}^{n-k}$$

Les spécificités connexes, ainsi calculées, permettent de repérer certaines des irrégularités liées au temps dans la ventilation d'un terme dans les parties d'un corpus.

Application à la question Life

Les diagnostics de spécificités rassemblés au tableau 7.6 concernent des termes dont la fréquence est supérieure à 5 occurrences dans l'ensemble du corpus ***Life***. Pour chaque terme on a commencé par calculer les spécificités simples (cf. chapitre 6) qui concernent la ventilation du terme dans chacune des parties du corpus. Dans un deuxième temps, on a calculé les spécificités qui concernent les parties consécutives (niveau 2, 3, etc.). A la fin de ce processus, on a retenu, pour chaque terme, le diagnostic correspondant à la probabilité la plus faible.

Les diagnostics présentés au tableau 7.6 correspondent donc aux faits de répartition les plus remarquables du point de vue de leur distorsion par rapport à une ventilation qui résulterait d'un tirage au sort.

Les spécificités longitudinales ainsi calculées mélangent des constats que l'on pouvait aisément prévoir et des découvertes plus inattendues. Les formes *job* et *work* sont rapprochées par leur répartition dans les catégories d'âge (plutôt fréquentes dans les catégories 1 à 4).

Good job, *career* et *money* mais aussi *friends* et *being happy* sont plutôt mentionnées par des catégories de répondants plus jeunes (Ag1 et Ag2). Comme dans l'enquête ***Enfants***, et pour des raisons que l'on peut comprendre, le terme *health* est majoritairement mentionné par les catégories les plus âgées

La présence importante dans les réponses des classes Ag3 et Ag4 de l'article *the* est liée au fait que ceux-ci, dont le niveau d'instruction est supérieur en moyenne, font le plus souvent des phrases plus longues, à la syntaxe plus

précise, avec aussi un degré de nominalisation supérieur (cf. au chapitre 5, les remarques qui suivent le paragraphe 5.5).

7.6 Séries textuelles chronologiques

Dans le domaine des analyses textuelles, et dans le domaine socio-politique en particulier, les analyses de presse, les études à caractère historique, conduisent souvent à la constitution d'un type de corpus réalisé par l'échantillonnage au cours du temps d'une même source textuelle sur une période plus ou moins longue.

Les textes ainsi réunis en corpus sont souvent produits dans des conditions d'énonciation très proches, parfois par le même locuteur. Leur étalement dans le temps doit permettre de les comparer avec profit, de mettre en évidence ce qui varie au cours du temps.

Nous appelons *séries textuelles chronologiques* ces corpus homogènes constitués par des textes produits dans des situations d'énonciation similaires, si possible par un même locuteur, individuel ou collectif, et présentant des caractéristiques lexicométriques comparables.

Selon les cas, l'étude de cette dimension peut se révéler plus ou moins intéressante. Ainsi, on peut décider d'entreprendre l'étude de l'évolution du vocabulaire d'un même auteur durant toute la période qui l'a vu produire son oeuvre, en s'appuyant dans un premier temps, quitte à la remettre en cause par la suite, sur la date de rédaction présumée de chacune de ses productions.[1]

Les études réalisées sur des séries textuelles chronologiques ont mis en évidence l'importance d'un même phénomène lié à l'évolution d'ensemble du vocabulaire au fil du temps : *le Temps lexical*. Ce renouvellement constitue, la plupart du temps, la caractéristique lexicométrique fondamentale d'une série chronologique[2]. Tout émetteur produisant des textes sur une période de temps assez longue utilise sans cesse de nouvelles formes de vocabulaire qui

[1] C'est ce qu'a fait Ch. Muller (1967) dans son étude sur le vocabulaire du théâtre de Pierre Corneille.

[2] Après les recherches, désormais classiques, de Ch. Muller sur l'évolution du lexique de Pierre Corneille et celles d'Etienne Brunet sur les données du Trésor de la Langue Française, cf. Brunet (1981). Citons parmi des études plus récentes : Habert, Tournier (1987) et Salem (1987), sur un corpus de textes syndicaux français (1971-1976), Peschanski (1981) sur un corpus d'éditoriaux de *l'Humanité* (1934-1936), Romeu (1992) sur des éditoriaux parus dans la presse espagnole entre 1939 et 1945, Gobin, Deroubaix (1987) sur des déclarations gouvernementales en Belgique (1961-1985), Bonnafous (1991) sur 11 ans d'éditoriaux de presse française (1974-1984) sur le thème de l'immigration, Labbé (1990) sur le premier septennat de F. Mitterrand.

viennent supplanter, du point de vue fréquentiel, d'autres formes dont l'usage se raréfie.

Il s'ensuit que les vocabulaires des parties correspondant à des périodes consécutives dans le temps présentent en général plus de similitudes entre eux que ceux qui correspondent à des périodes séparées par un intervalle de temps plus long. Bien entendu, il peut arriver que des clivages lexicaux plus violents relèguent au second plan cette évolution chronologique. Cependant, sa mise en évidence lors de plusieurs études portant sur des corpus chronologiques nous incite à lui prêter un certain caractère de généralité et à penser que sa connaissance est indispensable pour chaque étude de ce type.

7.6.1 La série chronologique *Discours*

Le corpus ***Discours***, présenté au chapitre 2, rassemble, comme nous l'avons signalé plus haut, les textes de 68 interventions radiotélévisées de F. Mitterrand survenues entre juillet 1981 et mars 1988. Ce corpus constitue un bon exemple de *série textuelle chronologique*.

Tableau 7.7

Les 7 périodes du corpus *Discours*

	code	*années*	*occurrences*	*formes*	*hapax*	*fmax*
1	Mit1	(81-82)	30867	4296	2298	1388
2	Mit2	(82-83)	41847	5077	2611	1807
3	Mit3	(83-84)	53050	5622	2818	2040
4	Mit4	(84-85)	36453	4601	2362	1285
5	Mit5	(85-86)	53106	5509	2726	1975
6	Mit6	(86-87)	41205	4920	2499	1523
7	Mit7	(87-88)	40730	4748	2417	1526

Partition en 7 années

On obtient une partition chronologique du corpus en regroupant l'ensemble des interventions en 7 parties qui correspondent chacune à une période d'un an (du 21 mai au 20 mai de l'année suivante).

On trouve au tableau 7.7 les principales caractéristiques lexicométriques des parties ainsi découpées.

Comme le montre le tableau 7.7, ce découpage produit une partition relativement équilibrée du corpus tant du point de vue du nombre des occurrences affectées à chacune des parties que de celui du nombre des formes.

La sélection des 1 397 formes dont la fréquence est supérieure à 20 occurrences représente 256 855 occurrences soit 86% des occurrences du corpus.

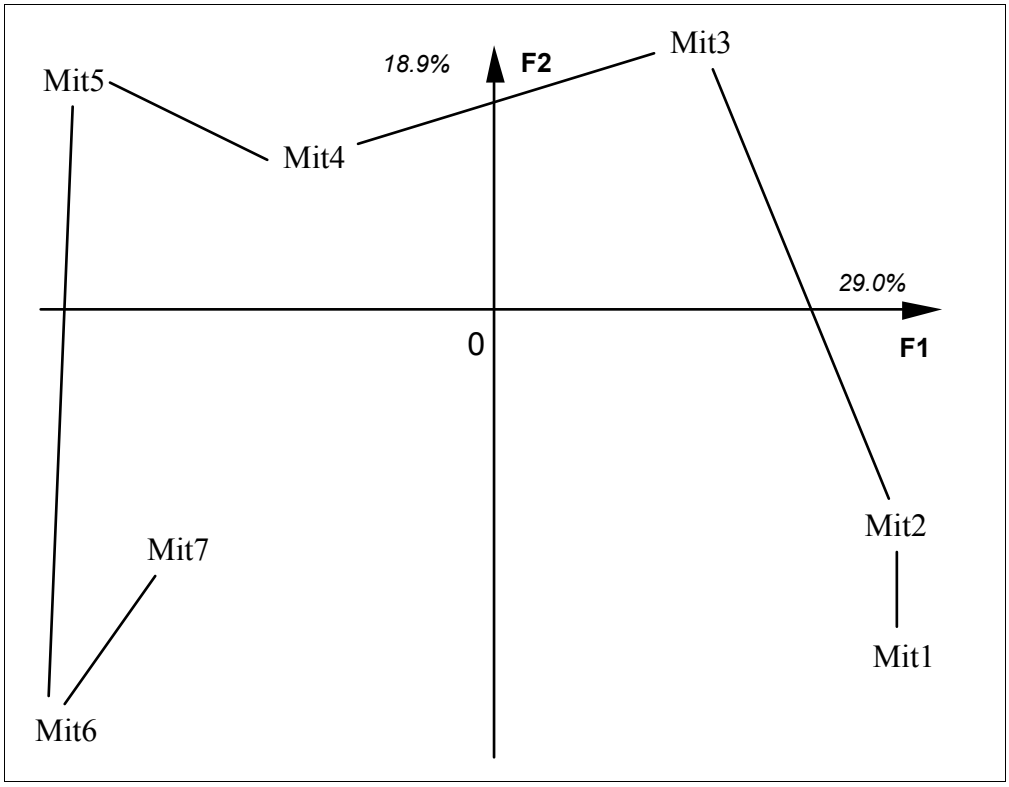

Figure 7.5

Corpus *Discours* : Esquisse des facteurs 1 et 2 issus de l'analyse des correspondances du tableau [1 397 formes (F≥ 20) ∞ 7 périodes]

Les résultats de l'analyse des correspondances du tableau (1 397 formes x 7 périodes) en ce qui concerne les différentes périodes sont résumés par la figure 7.5

L'examen des valeurs propres relatives à chacun des axes montre que trois axes se détachent de l'ensemble (λ_1=0.026, τ_1=29.0%, λ_2=0.017, τ_2=18.9%, λ_3=0.014, τ_3=16.2%)

On vérifie immédiatement que la disposition des périodes sur les premiers axes factoriels obéit dans les grandes lignes au schéma du temps lexical

exposé plus haut. Le décalage le plus marquant par rapport à ce schéma vient de la position de la dernière période sur le second axe.

En effet, la période Mit7 prend sur cet axe une position plus centrale que celle qui correspondrait à une évolution lexicale homogène par rapport à l'évolution d'ensemble. A partir de ce constat, on étudiera l'hypothèse d'une inflexion particulière du discours présidentiel dans la dernière année du septennat.

On remarque par ailleurs que l'évolution est loin d'être homogène dans le temps entre les différentes périodes. Ainsi les périodes Mit1 et Mit2 sont-elles rapprochées sur le plan des deux premiers axes factoriels alors que les périodes Mit5 et Mit6 sont au contraire très distantes. Les périodes Mit3 et Mit4 se signalent par le fait qu'elles rompent l'alignement de l'ensemble Mit1-Mit6.

7.6.2 Spécificités chronologiques

L'analyse des spécificités chronologiques (spécificités longitudinales dans le cas d'un corpus chronologique), permet de mettre en évidence les termes particulièrement employés (et les termes sous-employés) au cours d'une période ou d'un groupe de périodes consécutives dans le temps.

Le tableau 7.8, donne la liste des spécificités chronologiques majeures du corpus ***Discours***. Seules quelques formes extrêmement significatives d'un point de vue statistique[1] ont été sélectionnées pour ce tableau. La liste s'ouvre sur un diagnostic de spécificités relatif à la forme *nous* dont la méthode nous signale l'abondance relative dans les premières parties. On note, qu'a l'inverse, la fréquence de la forme *je*, moins forte dans les premières périodes à tendance à augmenter avec le temps.

La période numéro 1 est caractérisée par la présence des termes *nationalisations, reformes, relance*, lesquels disparaissent par la suite comme on peut le constater.

Un examen plus détaillé de cette liste permettra de dégager des thèmes relatifs à chacune des périodes et de suivre leurs fluctuations dans le temps. Le classement en probabilité permet comme toujours de hiérarchiser les écarts.

Comme plus haut, ces listes, dont on pourra régler le volume global en faisant varier le seuil de sélection des termes, peuvent être triées en fonction

[1] Il s'agit de formes dont l'indice de probabilité est inférieur à $1/10^{-14}$.

des périodes du corpus si l'on désire illustrer les fluctuations du vocabulaire au fil des périodes du corpus.

Tableau 7.8

Spécificités chronologiques majeures du corpus *Discours*
(cf. guide de lecture du tableau 7.6)

terme	F	fp	sp	code
nous	2059	1148	+E35	01-03
l' iran	50	41	+E27	07-
nationalisations	42	31	+E22	01-
étudiants	28	27	+E22	06-
chaîne	39	34	+E21	05-
israël	71	54	+E20	01-02
majorité	212	130	+E19	05-06
réformes	39	27	+E18	01-
a	2863	1875	+E18	04-07
relance	32	24	+E17	01-
avons	523	458	+E17	01-05
de	11544	3195	+E17	01-02
c	3240	2643	+E17	03-07
je ne	794	567	+E16	04-07
nous avons	413	366	+E16	01-05
pour 100	204	193	+E16	01-05
cela	1636	1096	+E15	04-07
pas	4067	2092	+E15	05-07
pays	748	358	-E15	03-06
cinquième	35	27	+E14	05-
je	5024	3156	+E14	04-07

7.6.3 Les accroissements spécifiques

Pour compléter l'étude de l'évolution d'un vocabulaire au fil du temps, il reste encore à construire des outils qui permettront de signaler des changements brusques dans l'utilisation d'un terme lors d'une période donnée par rapport aux périodes précédentes - et non à l'ensemble du corpus comme c'est le cas dans les utilisations que nous venons de décrire.

Calcul des accroissements spécifiques

Dans ce qui suit, nous sommes toujours confrontés à un tableau lexical du type de ceux que nous avons étudiés plus haut. Cependant, la fréquence f_{ij} de la forme i dans la partie j va être appréciée cette fois en fonction de quantités qui ne correspondent pas à celles que nous avons employées pour le calcul des spécificités simples.
En effet, si l'on se fixe une période j (j varie entre 2 et p - nombre total des périodes du corpus, puisqu'il s'agit d'accroissement), il est possible pour un

terme donné de comparer la fréquence f_{ij} à la fréquence de ce même terme dans l'ensemble des parties précédentes.

Pour effectuer cette comparaison, on commencera par calculer les quatre paramètres du modèle hypergéométrique. Deux de ces paramètres sont identiques par rapport au calcul des spécificités simples, ce sont les paramètres:

f_{ij} — fréquence du terme i dans la partie j;
t_j — nombre des occurrences de la partie j;

Les deux autres sont des analogues des paramètres employés lors du calcul des spécificités simples pour lesquels la sommation s'effectue cette fois non plus pour l'ensemble des parties mais pour les j premières parties.

T^j — le nombre des occurrences du corpus réduit aux seules j premières parties du corpus;
F^j_i — la fréquence du terme i dans ce même sous-corpus.

Guide de lecture du tableau 7.9

Les spécificités calculées concernent les sous-fréquences des termes (formes ou segments) dont la fréquence dépasse 20 occurrences dans le corpus. Seuls les diagnostics dont l'indice de probabilité est inférieur à 10^{-6} ont été retenus.

La colonne Fx indique la fréquence du terme dans les x premières parties du corpus. Cas particulier, pour la dernière période du corpus Fx=F ; fréquence totale de la forme dans le corpus.

La colonne f indique la fréquence partielle du terme dans la partie pour lesquelles le diagnostic d'accroissement spécifique a été établi (ici la dernière partie).

Le diagnostic d'accroissement spécifique affecté d'un exposant

– / Exx renvoie à un accroissement spécifique positif, sur-emploi spécifique par rapport aux parties précédentes, à laquelle est attachée une probabilité d'ordre 1/10 à la puissance xx.
– \ Exx renvoie à un accroissement spécifique négatif, sous-emploi spécifique par rapport aux parties précédentes, à laquelle est attachée une probabilité d'ordre 1/10 à la puissance xx.

Les diagnostics retenus ont été classés en fonction des spécificités positives ou négatives mises en évidences puis en fonction de l'indice probabilité.

Tableau 7.9

Corpus *Discours* accroissements spécifiques majeurs de la partie Mit7 (période : 12/04/87 - 4/03/88)

terme	F7	f	spec.
l iran	50	41	/E27
iran	53	41	/E25
arabe	34	23	/E13
monde arabe	21	17	/E12
d instruction	20	16	/E11
instruction	23	17	/E11
l irak	29	18	/E09
irak	32	18	/E08
élection	35	18	/E07
président	303	73	/E07
d armes	27	15	/E07
un président	28	15	/E07
politiques	105	34	/E07
armes	93	32	/E07
juge	35	17	/E07
pays	748	151	/E07
manière	25	13	/E06
oui	256	63	/E06
candidat	26	14	/E06
y	1749	305	/E06
vraiment	143	40	/E06
il y	1014	191	/E06
désarmement	38	17	/E06
nous avons	413	27	\E06
inflation	83	0	\E06
avons	523	35	\E07
jeunes	134	2	\E07
nous	2059	182	\E12

Ces deux derniers paramètres ont été pourvus d'un indice supérieur qui rappelle que la sommation est une sommation partielle, limitée aux premières périodes du corpus.

Remarquons tout de suite que pour la dernière des parties du corpus, et pour elle seulement :

$$T^p = T; \quad F^p{}_i = F_i \, .$$

En d'autres termes, pour la dernière des parties le calcul des accroissements spécifiques amène des résultats identiques au calcul des spécificités simples pour cette partie. Pour chaque période donnée, on peut alors obtenir une liste des termes dont l'emploi a notablement augmenté ainsi que celle des termes brutalement disparus.

L'étude des accroissements spécifiques majeurs du corpus permet de localiser les changements les plus brusques sur l'ensemble du corpus. On a rassemblé au tableau 7.9 les diagnostics les plus importants qui concernent la seule partie Mit7 laquelle, comme nous l'avons vu sur le premier plan factoriel issu de l'analyse du tableau (1 397 formes x 7 parties) manifeste une position particulière par rapport à l'évolution d'ensemble.

Ce tableau illustre très nettement l'émergence au cours de l'année 1987-1988 d'un vocabulaire lié à des conflits internationaux du moment *(iran, monde arabe, irak, armes)* et aux affaires juridiques elles-mêmes liées aux attentats terroristes *(juge, instruction)* alors que refluent de manière spectaculaire les préoccupations liées à l'*inflation* et aux *jeunes*. On note également la spécificité négative de formes liées à la première personne du pluriel (*nous, avons*). En ne retenant que quelques formes liées aux diagnostics les plus marqués d'accroissement spécifique, nous nous sommes bornés à une illustration de la méthode. Il va sans dire qu'une sélection plus large de termes (que l'on obtient sans peine en choisissant un seuil de probabilité plus élevé) permet de remarquer des accroissements dont l'évidence est moins nette pour le politologue.

7.6.4 Étude parallèle sur un corpus lemmatisé

Nous avons vu au chapitre 2, à propos de la série chronologique ***Discours***, comment les décomptes parallèles réalisés d'une part en formes graphiques et d'autre part sur des unités lemmatisées amenaient des résultats très éloignés les uns des autres en ce qui concerne les principales caractéristiques lexicométriques du corpus : nombre des formes, des hapax, fréquence maximale, etc.[1]

Le présent paragraphe nous permettra d'évaluer l'incidence de l'opération de lemmatisation sur les résultats auxquels aboutissent les analyses multidimensionnelles présentées dans les chapitres qui précèdent.

Pour cette comparaison effectuée à partir du décompte des occurrences de chacun des lemmes dans les 7 périodes du corpus, nous avons conservé les seuils de sélection établis lors de l'analyse effectuée sur les formes graphiques. Nous avons ensuite soumis aux diverses méthodes d'analyses multidimensionnelles un tableau croisant les 1 312 lemmes dont la fréquence est au moins égale à 20 occurrences avec les sept périodes considérées dans l'expérience relatée plus haut.

---

[1] Rappelons que la série chronologique ***Discours*** que nous considèrons actuellement à fait l'objet d'une lemmatisation, partiellement assistée par ordinateur, réalisée par D. Labbé. Cf. Chapitre 2 du présent ouvrage, et Labbé (1990). On trouvera un commentaire plus développé de l'ensemble de ces résultats dans Salem (1993).

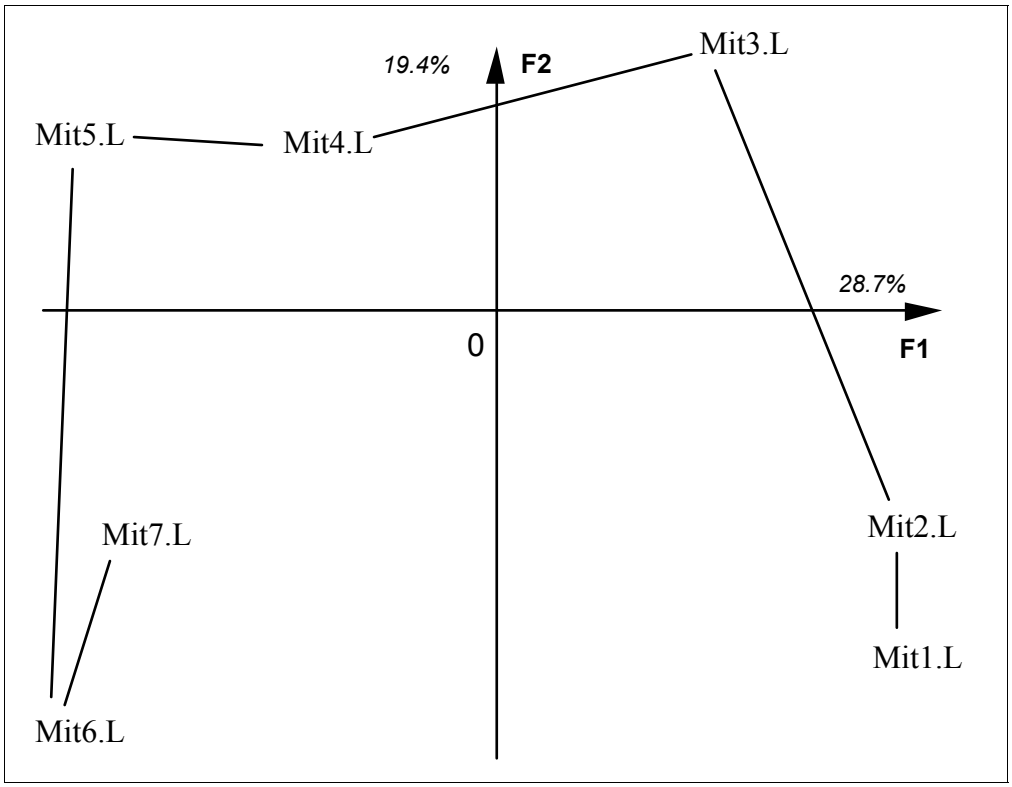

Figure 7.6

Corpus *Discours lemmatisé*
Facteurs 1 et 2 issus de l'analyse des correspondances
du tableau (1 312 lemmes (F≥ 20) x 7 périodes)

Les résultats de l'analyse des correspondances de ce dernier tableau sont résumés, pour ce qui concerne les différentes périodes, par la figure 7.6. Les symboles qui désignent les 7 périodes du corpus (Mit1.L, Mit2.L, ..., Mit7.L) ont été affectés de la lettre "L" qui rappelle que les décomptes sont effectués cette fois ci à partir de lemmes.
Ici encore trois axes se détachent de l'ensemble auxquels correspondent des valeurs propres et des pourcentages d'inertie très voisins de ceux obtenus lors de l'analyse sur les formes graphiques (λ_1=0.024, τ_1=28.7%, λ_2=0.016, τ_2=19.4%, λ_3=0.014, τ_3=16.6%)

L'examen de la disposition relative des périodes sur le plan des premiers axes factoriels confirme l'impression que nous nous trouvons bien devant des résultats extrêmement proches de ceux obtenus à partir des formes graphiques (cf. fig 7.5 du présent chapitre).

La comparaison des listes d'unités les plus contributives pour chacun des ensembles d'unités de décompte (lemmes et formes graphiques) qui servent de base à la typologie dans l'une et l'autre des analyses montre que les différences sont peu importantes dans l'ensemble. De la même manière la

liste des diagnostics de spécificité chronologique et d'accroissements spécifiques les plus forts, calculés à partir des unités lemmatisées recoupent dans une grande mesure les diagnostics obtenus à partir des formes graphiques.[1]

Ainsi, à partir de tableaux qui présentent des différences importantes dans la mesure ou ils résultent de comptages effectués sur des unités elles mêmes très différentes, les méthodes de la statistique textuelle (analyse des correspondances, classification hiérarchique, spécificités chronologiques) produisent des résultats très proches.

L'expérience présentée ci-dessus ne constitue évidemment pas à elle seule une preuve définitive de ce que les analyses multidimensionnelles mettent en évidence des contrastes entre les textes peu dépendants du choix des unités de décompte. Ici encore c'est la finalité de l'étude entreprise qui soufflera les choix à opérer.

La décision est aussi, si l'on peut dire, d'ordre économique. Il est dans l'absolu toujours préférable de disposer d'un double réseau de décomptes (en formes graphiques et en lemmes).[2] Une *lemmatisation complète*, sur un corpus important, reste une opération coûteuse. Indispensable dans un travail de recherche, elle est beaucoup moins justifiée s'il s'agit d'obtenir rapidement des visualisations ou des typologies de parties de corpus d'une certaine richesse lexicale. On peut alors travailler sur la base des formes graphiques sans crainte de passer à côté de l'essentiel. En revanche, si l'on a affaire à des textes courts non regroupés, les analyses sur texte lemmatisé donneront un point de vue parfois différent, souvent complémentaire.

7.7 Recherches en homogénéité d'auteur

Application au problème du livre d'Isaïe.[3]

Dans ce paragraphe, on appliquera les méthodes de l'analyse de la contiguïté, à un problème philologique ancien : celui de l'homogénéité du livre de la Bible que l'on nomme : *Le livre d'Isaïe* (cf. encadré ci-dessous). Sans insister sur les particularités lexicométriques de l'hébreu biblique, langue dans laquelle le texte est parvenu jusqu'à nous, on se consacrera essentiellement à l'éclairage qu'apportent à ce problème les typologies empiriques réalisées sur la base des comptages de formes.[4]

---

[1] Cf. Salem (1993) p 740-746.

[2] Mais il ne faut pas perdre de vue que la lemmatisation d'un corpus de plusieurs centaines de milliers d'occurrences est une opération de longue haleine.

[3] Pour plus de détails, cf. Salem (1979). Les résultats des analyses quantitatives ont été repris sous leur forme originale ; l'exposé méthodologique a été réactualisé.

[4] Rappelons cependant que cette langue, comme la plupart des langues de la famille sémitique, s'est donné une écriture consonantique (i.e. on ne transcrit que les consonnes

Un problème vieux de deux mille ans . . .

Combien d'auteurs ont-ils participé à la rédaction du livre de la Bible attribué au prophète Isaïe? Depuis longtemps les différentes écoles de la critique biblique divergent. Pour les uns, le livre entier doit être attribué à un même auteur. D'autres avancent des découpages en deux, trois, six, voire dix-sept parties qui selon eux doivent chacune être attribuées à un auteur différent.

Si l'oeuvre dont tous les commentateurs s'accordent à reconnaître la grande qualité littéraire, se présente comme rédigée avant l'exil de Babylone (environ 600 av J.C.), la mention du roi Cyrus (Is.Ch44.28) qui vécut quelques cent cinquante ans plus tard, laisse planer un doute sur l'éventualité d'un auteur unique pour l'ensemble de l'oeuvre.

A partir de remarques de ce type, l'hypothèse d'une oeuvre constituée par compilation de plusieurs textes d'origines différentes investit rapidement le milieu des études bibliques et donne lieu à une inflation sur le nombre possible des auteurs. Avec le développement des études bibliques, le livre se trouve fragmenté en parcelles de plus en plus fines dont les critiques affirment que chacune d'entre elles doit être attribuée à un auteur différent.

Signalons enfin que parmi les *Manuscrits de la Mer Morte* exhumés en 1947, les archéologues ont découvert une copie du livre d'Isaïe qui se trouve, du point de vue de l'établissement du texte, dans un état très proche de celui qui est parvenu jusqu'à nous à travers les manuscrits de l'époque médiévale. Avant cette découverte le manuscrit le plus ancien de la Bible était un manuscrit daté de l'an 1009 (après JC). Cette découverte signifiant pour nous que s'il y a eu compilation, elle a été réalisée à une date très ancienne qui ne peut être postérieure à 150 avant J.C.

Recherches textuelles en homogénéité d'auteur.

De nombreuses études "textuelles", c'est à dire se fondant exclusivement sur le matériau textuel du livre et non sur les recoupements d'ordre historique que l'on peut faire à partir du texte, ont été publiées à propos de cette oeuvre. La plupart de ces études s'appuie sur des comptages réalisés manuellement ou automatiquement. Ces comptages portent sur des unités

de la chaine vocalique laissant le loisir au lecteur de rétablir les voyelles ; opération que le lecteur expérimenté effectue sans grande difficulté). En hébreu, on a coutume de diviser les mots du discours en trois catégories : noms, verbes et particules libres. La base des verbes est constituée par des racines trilitères. Aux mots, peuvent s'adjoindre différentes particules proclitiques ou enclitiques. Dans les années 70, le Centre Analyse et de Traitement Automatique de la Bible (CATAB), dirigé par G.E.Weil, a établi une transcription sur cartes perforées de l'ensemble de la Bible hébraïque.

très différentes dont les auteurs affirment, dans chacun des cas, qu'elles permettent l'identification d'un auteur de manière privilégiée[1].

Nous évoquerons également au chapitre 8, dévolu à l'analyse discriminante textuelle, les problèmes d'attribution d'auteurs, en insistant sur certains modèles statistiques sous-jacents. On réservera la terminologie *homogénéité d'auteur* aux cas pour lesquels il n'y a pas plusieurs auteurs potentiels identifiés, mais simplement incertitude sur l'unicité de l'auteur d'une série de textes.

Cette tradition de recherche en homogénéité d'auteur dans le domaine biblique prend largement ses sources dans les recherches analogues réalisées dans d'autres domaines parfois très éloignés[2].

Au plan méthodologique, le schéma qui régit la plupart de ces études peut-être ramené à trois phases principales :

a) promotion au rang de variable discriminante, pour l'attribution d'un texte à un auteur donné, d'un type de comptage particulier portant sur les occurrences d'un phénomène textuel de préférence assez fréquent, (occurrences d'une particule, d'une tournure syntaxique, proportion d'emploi des voix du verbe, etc.).
b) discussion d'exemples montrant que, sur un corpus composé de groupes de textes dont on sait avec certitude que chaque groupe est composé de textes écrits par un auteur différent, la variable privilégiée prend effectivement des valeurs qui discriminent les groupes.
c) application du critère ainsi validé à un problème n'ayant pas de solution connue.

La difficulté majeure créée par ce genre d'étude vient précisément de la diversité des critères d'homogénéité proposés. Le point le plus discutable est que ces critères n'ont été, en général, élaborés et utilisés qu'en vue de résoudre un seul problème. On est en droit de se poser la question : pour justifier une partition quelconque d'une oeuvre n'est-il pas possible de

[1] Cf. par exemple une étude portant sur une codification de la morphologie du verbe dans le livre d'Isaïe, Kasher (1972), ou encore une étude portant sur la répartition dans ce même livre d'une particule que l'on nomme "waw conversif", Radday (1974).

[2] Radday (1974) signale d'ailleurs que le critère portant sur les occurrences du waw conversif lui a été inspiré par une étude du Rev. A.Q. Morton : *The authorship of the Pauline corpus* (Morton, 1963), sur des textes évangéliques en grec ancien. Morton estime en effet que la variation du nombre des occurrences de la particule και dans différents textes des évangiles permet de conclure que certains des textes attribués à Paul n'ont pu être écrits par ce dernier, d'où la transposition proposée dans une langue complètement différente.

trouver un critère discriminant, parmi la masse des critères imaginables, et ce, quelle que soit cette partition ?

Pour sortir de ces difficultés, il est donc important de ramener l'ensemble de ces problèmes à l'intérieur du cadre plus général du décompte exhaustif des formes graphiques d'un texte soumis à l'analyse multidimensionnelle des tableaux ainsi obtenus. La statistique textuelle ne peut prétendre trancher de manière indiscutable ce débat entre érudits. Cependant les méthodes exposées plus haut peuvent apporter sur ces questions un éclairage quantitatif dont on a plusieurs fois prouvé la pertinence.

7.7.1 Le corpus informatisé.

Le texte sur lequel nous avons travaillé compte 16 931 occurrences. Plusieurs possibilités de découpage du corpus en unités de base se présentent alors. Plutôt que de choisir une partition en tranches de longueur fixe (1 000 mots, 100 versets, etc.) nous avons opté pour la partition du livre en 66 chapitres[1] que l'on désignera par la suite (Is1, Is2, ..., Is66) partition presque universellement reçue de nos jours. C'est dans ce langage des chapitres que s'expriment les partitions en fragments homogènes proposés par la plupart des critiques.

Dans la mesure où nous envisageons de faire apparaître, à l'aide de méthodes quantitatives, des fragments homogènes à l'intérieur du livre, nous éviterons par tous les moyens d'introduire en qualité d'hypothèses les découpages de la critique biblique que nous nous proposons de mettre à l'épreuve[2].

Au cours de l'étude, l'unité de chacun des 66 chapitres ne sera plus remise en cause. Cette partition servira ensuite à construire des groupes de chapitres homogènes au plan statistique.

---

[1] Signalons que cette partition en chapitres n'a été établie que vers le début du 13ème siècle par Stephen Langton, futur archevêque de Canterbury. Elle est utilisée aujourd'hui par la plupart des spécialistes de la Bible.

[2] On trouvera une description des différentes partitions de ce livre proposées au cours des sciècles par la critique dans l'article (Weil et al. 1976). Il est facile de constater que les chapitres Is36 à Is39 constituent un groupe très particulier puisque le texte de chacun de ces chapitres correspond, souvent mot pour mot, à un passage du Deuxième livre des Rois (autre livre de la Bible). Certains des critiques considèrent que ce groupe de chapitres ne fait pas réellement partie du livre d'Isaïe et l'écartent purement et simplement de leur partition de l'oeuvre en fragments homogènes. Nous n'écarterons pas d'emblée ces chapitres de nos analyses ; leur présence nous permettra au contraire de juger, à travers les classements qui leur seront attribués, de la pertinence des méthodes utilisées.

En effet, pour être utile, la *partition de travail* qui sert de base aux regroupements futurs doit répondre à plusieurs critères :

a) les éléments de base doivent être choisis suffisamment ténus pour que la majorité des parties ne chevauchent pas les frontières, inconnues a priori, entre ce qui constitue des fragments homogènes au plan statistique.

b) cette partition ne doit pas introduire subrepticement une partition a priori, résultant de considérations extra-textuelles, puisque le but final de l'analyse est précisément de dégager une telle partition avec un minimum de considérations a priori.

c) enfin, les éléments de base doivent être choisis suffisamment longs pour que leurs profils lexicaux représentent bien des moyennes statistiques susceptibles d'être rapprochées par des algorithmes de classement.

Pour cette division en chapitres, qui respecte l'unité thématique de petits fragments, le chapitre le plus long, Is37, compte 566 occurrences, le plus court Is12 n'en compte que 62. Cependant, la grande majorité des chapitres oscillent entre 150 et 400 occurrences soit environ du simple au triple.

A partir du décompte des occurrences de chacune des formes les plus fréquentes dans chacun des 66 chapitres du livre, on construit le tableau à double entrée de dimensions (89 formes x 66 chapitres).[1]

L'objectif principal de cette étude étant d'aboutir à un découpage en parties homogènes, on se gardera d'introduire trop tôt, même à des fins de validation, des structures de contiguïtés par trop inspirées des découpages traditionnels de la critique biblique, puisqu'il s'agit précisément de mettre ces découpages à l'épreuve. Il est néanmoins possible d'introduire une hypothèse de contiguïté particulière qui servira de point d'appui pour mesurer la pertinence des typologies obtenues.

L'hypothèse d'homogénéité globale des chapitres consécutifs

Le découpage en unités relativement fines que sont les 66 chapitres permet en effet de supposer que la grande majorité des coupures entre les chapitres ne constituent pas également des coupures entre fragments dus à des auteurs différents.

[1] La liste des 100 formes les plus fréquentes nous avait été fournie par l'équipe du CATAB. A l'époque nous en avions écarté d'emblée une liste de 11 noms propres, ne souhaitant pas y trouver un appui pour le découpage recherché, précaution que nous jugerions superflue aujourdhui.

Sous cette hypothèse, les mesures de contiguïté, nous permettent de dire, pour toute variable numérique prenant ses valeurs sur l'ensemble des chapitres, si cette variable prend des valeurs voisines sur les chapitres consécutifs.

Considérons en effet, une variable z_j prenant ses valeurs sur l'ensemble J des chapitres (cette variable peut être le résultat du décompte au fil des chapitres d'une unité textuelle, un facteur, ou encore toute autre fonction numérique). Si l'on stipule l'hypothèse d'homogénéité globale des chapitres consécutifs, c'est à dire si l'on accepte, dans notre cas, que deux chapitres consécutifs ont de fortes chances d'appartenir au même fragment-auteur, il s'ensuit que pour toute variable z_j prenant ses valeurs sur l'ensemble J des chapitres, les différences du type $z_j - z_{j-1}$ sont, pour la plupart d'entre elles, des différences entre chapitres appartenant à un même fragment.

Dans le cas où la variable z_j discrimine réellement les fragments-auteur, c'est à dire dans le cas ou elle prend des valeurs similaires pour les chapitres j et j - 1 situés à l'intérieur d'un même fragment-auteur, et des valeurs différentes pour les fragments-auteur distincts, le coefficient de contiguïté calculé sur les chapitres consécutifs doit être nettement inférieur à l'unité.

Commençons par appliquer le calcul de ces coefficients à l'étude de la variation au fil des chapitres de la variable *D*, mesurée pour chacun des chapitres et définie comme suit :

$$D_j = (V_j - P_j) * 100 / t_j$$

où :

— V_j est le nombre des occurrences de formes verbales du chapitre j.

— P_j est le nombre des occurrences de particules dans ce même chapitre.

— t_j est le nombre des occurrences dans le chapitre.

En d'autres termes D_j représente la différence entre le nombre des verbes et celui des particules rapportée au nombre des occurrences dans chacun des chapitres.[1]

[1] Une étude parallèle (Salem 1979) a montré que les livres réputés anciens comportaient, proportionnellement, plus de particules par rapport aux verbes que les livres plus récents. Ces proportions se modifiant au fur et à mesure de l'évolution de la langue en raison vraisemblablement de la formation de verbes nouveaux par l'intégration en qualité d'affixes de particules précédemment libres aux racines anciennes. Cette circonstance permet, sous certaines conditions, de considérer la variable *VP* comme un critère de datation dont l'efficacité varie avec la longueur des fragments considérés. Ce critère ne tient pas compte de la possibilité de réécriture à une date plus rapprochée de nous de textes beaucoup plus anciens; procédé dont on peut supposer qu'il a été fréquemment utilisé avant la fixation définitive des textes que nous considérons.

La figure 7.7 montre comment la quantité D_j varie pour les 66 chapitres du livre d'Isaïe et pour les 40 chapitres que contient un autre livre de la Bible, le livre de l'Exode, dont on a de nombreuses raisons de croire qu'il est beaucoup plus ancien.

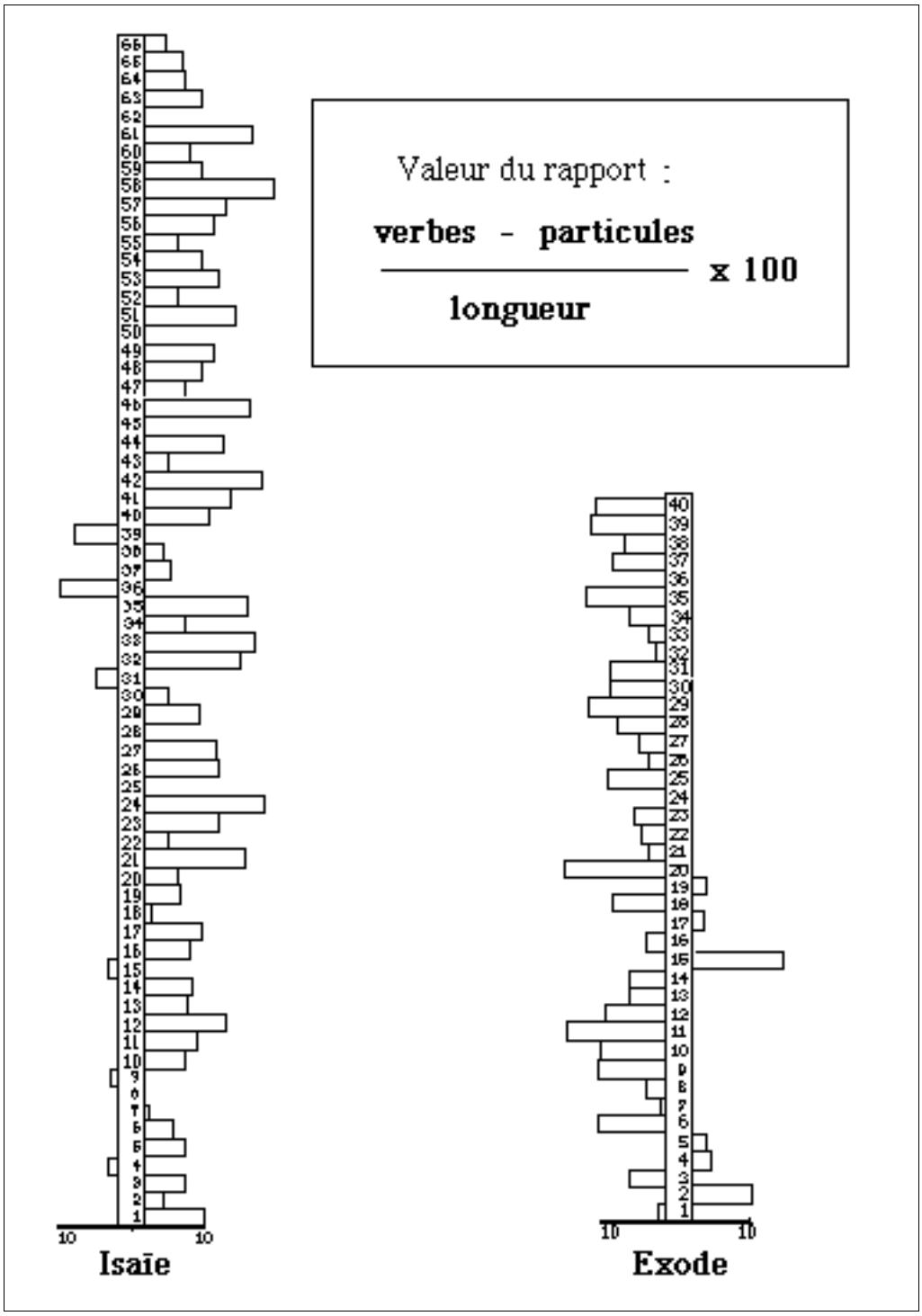

Figure 7.7

Etude d'homogénéité pour la différence *Verbes-Particules* pour les chapitres du livre d'Isaïe et du livre de l'Exode.

Guide de lecture de la figure 7.7

Les chapitres d'un même livre sont représentés par ordre croissant de bas en haut. Pour chacun des chapitres, les écarts de la variable D_j sont représentés par un rectangle qui s'écarte sur la droite du numéro correspondant au chapitre si le nombre des verbes excède celui des particules, et sur la gauche de ce numéro dans le cas inverse.

On note que les chapitres du livre d'Isaïe comptent dans l'ensemble plutôt plus de verbes, à l'inverse des chapitres du livre de l'Exode dans lesquels les particules sont plutôt plus abondantes. Ce qui confirmerait, si besoin était, les présomptions d'antériorité du livre de l'Exode.

Pour le livre d'Isaïe, les exceptions les plus marquantes sont constituées par les chapitres Is36 et Is39 dont on a déjà signalé la particularité à l'intérieur d'un groupe Is36-Is39. Pour ces chapitres, l'excédent de particules par rapport aux verbes pourrait conforter l'hypothèse d'une écriture antérieure aux autres chapitres du livre. Les chapitres Is37 et Is38 qui appartiennent visiblement au même groupe thématique que Is36 et Is39 ne présentent pas la même répartition que ces derniers.

Pour chacun des deux livres, on a calculé le coefficient de contiguïté pour les valeurs de la variable D_j :

pour les 66 chapitres du livre d'Isaïe, $c(D_j.) = 0.86$,

pour les 40 chapitres du livre de l'Exode, $c(D_j.) = 0.93$.

Dans les deux cas le coefficient est très faiblement inférieur à l'unité. L'examen des tables de l'écart type pour la loi de ce coefficient, en fonction du nombre des parties, conduit à la conclusion que le coefficient calculé pour les deux livres s'écarte de l'unité de manière non significative.

La conclusion qui s'impose est donc que la variable D_j distingue les deux séries de chapitres en raison, pouvons-nous supposer, de leur date d'écriture différente. Par contre, cette variable ne varie pas de manière très homogène au fil des chapitres consécutifs. Il semble donc difficile d'utiliser la variable D_j. pour délimiter des fragments homogènes à l'intérieur de chacun des livres.

Analyse sur les chapitres

L'examen de l'histogramme des valeurs propres, figure 7.8, montre que les deux premières valeurs propres se détachent très nettement. Les valeurs propres correspondant aux deux axes suivants (3 et 4) se détachent ensuite

des valeurs propres qui correspondent au reste des facteurs lesquelles décroissent ensuite régulièrement.

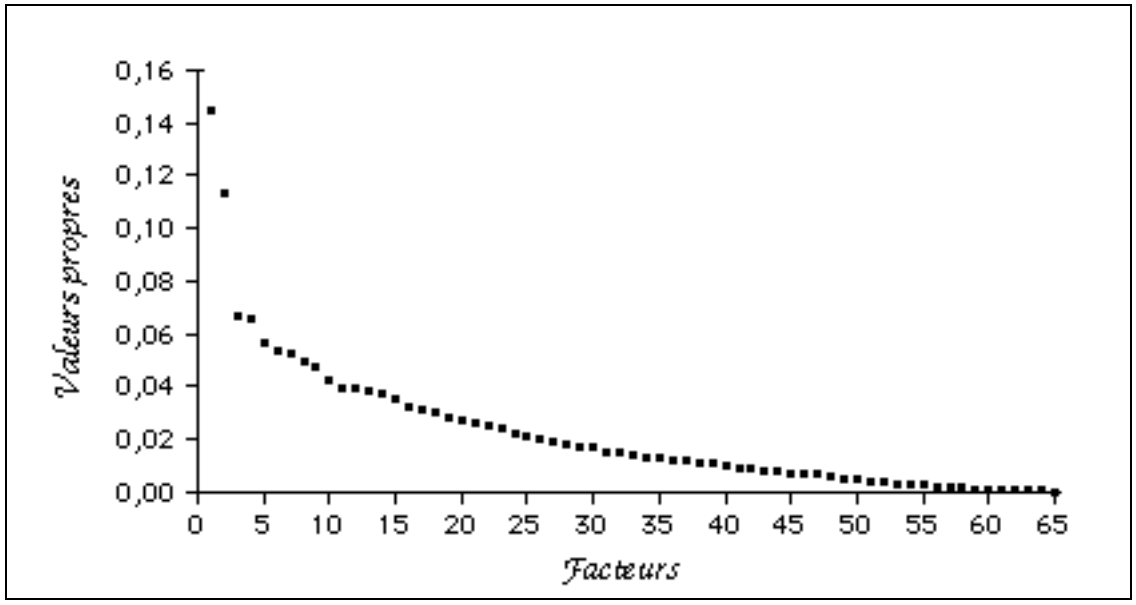

Figure 7.8

Isaïe : Les valeurs propres issues de l'analyse du tableau (89 formes x 66 chapitres)

En considérant les facteurs sur l'ensemble des chapitres on s'aperçoit vite que les chapitres consécutifs occupent sur ces facteurs des positions similaires.

7.7.2 Validation des résultats

Pour chacun des 65 facteurs issus de l'analyse des correspondances du tableau (89 formes x 66 chapitres), le coefficient de contiguïté va nous permettre d'apprécier dans quelle mesure le facteur possède ou non la propriété de "proximité des valeurs sur les chapitres contigus".

Le coefficient à été calculé à partir des 66 coordonnées factorielles relatives à chacun des chapitres. On trouve sur la figure 7.9 le résultat de ces calculs.

La figure 7.10 fournit les valeurs du coefficient G_α (coefficient de contiguïté mesuré sur les α premiers facteurs, cf. paragraphe 7.3.2).

Comme le montrent nettement ces figures, les premiers facteurs issus de l'analyse des correspondances possèdent à un très haut degré la propriété d'homogénéité sur les chapitres consécutifs que nous avons étudiée plus haut (en plus de leur propriété spécifique qui est de maximiser l'inertie du nuage).

Cette propriété est particulièrement frappante pour ce qui concerne les deux premiers facteurs.

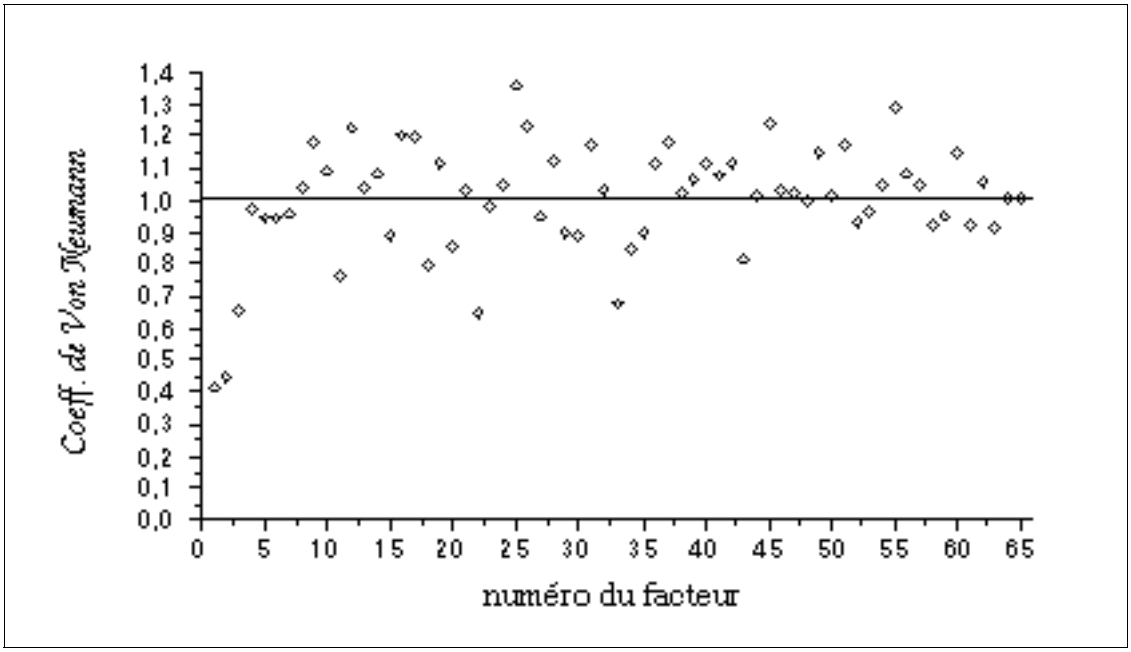

Figure 7.9

***Isaïe* : Le calcul du coefficient de Von Neumann pour les 65 facteurs issus de l'analyse des correspondances du tableau (89 formes x 66 chapitres)**

Les deux méthodes proposées ci-dessous permettent de préciser le diagnostic que l'on peut faire à partir de ces données.

Un test d'autocorrélation

L'écart type du coefficient de contiguïté vaut $\sigma = 0.122$. Il nous permet de définir un intervalle de confiance approximatif $1 \pm 2\,\sigma$ qui vaut dans notre cas [0.756, 1.244].

Comme on le voit en se reportant à la figure 7.9 , un très petit nombre de coefficients ainsi calculés à partir des facteurs se trouve à l'extérieur de cet intervalle[1]. Ces coefficients sont, pour ce qui concerne les valeurs inférieures à l'unité : $c_1 = 0.410$, $c_2 = 0.449$, $c_3 = 0.654$, $c_{22} = 0.642$, $c_{33} = 0.677$ et pour ce qui concerne les valeurs supérieures à l'unité : $c_{55} = 1.290$.

Ici encore, la distribution de l'écart n'est pas symétrique dans ce sens que l'on ne rencontre que très peu de facteurs ayant la propriété de prendre sur les chapitres consécutifs des valeurs fortement différentes.

[1] On se trouve ici encore devant un problème de comparaisons multiples (cf. chapitre 6, paragraphe 6.1.3). L'intervalle de confiance approximatif pour les coefficients c est donc trop large, ce qui incite à penser que seuls les deux premiers sont nettement significatifs.

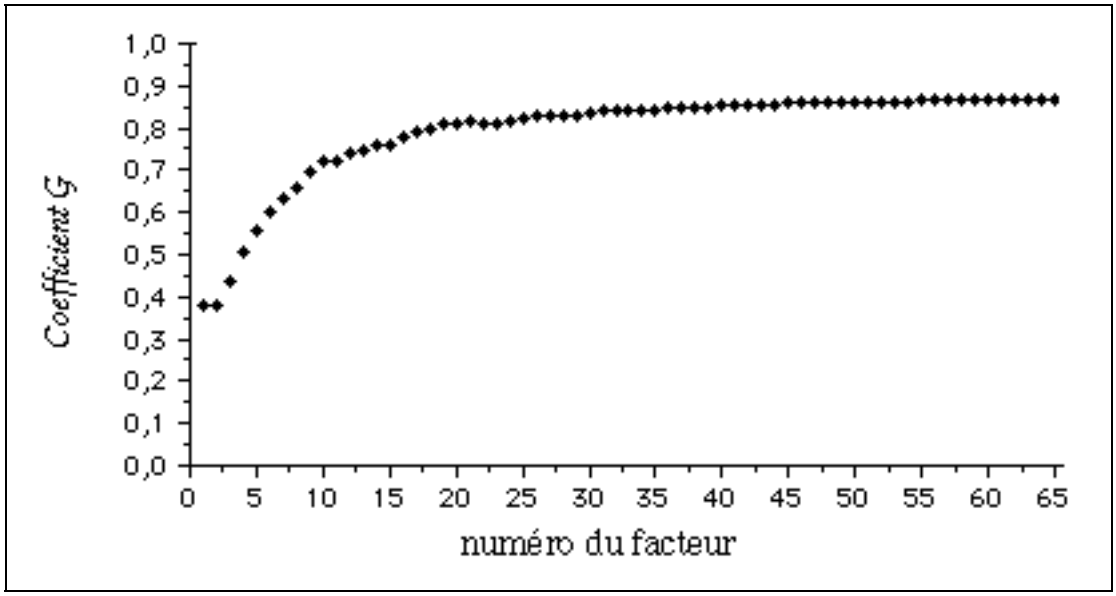

Figure 7.10

***Isaïe* : Le calcul du coefficient G_α pour les 65 facteurs issus de l'analyse des correspondances du tableau (89 formes x 66 chapitres)**

De tout ce qui précède, on retiendra que les deux premiers facteurs, outre le fait qu'ils correspondent aux plus fortes valeurs propres, possèdent plus que les autres la propriété de rapprocher les chapitres consécutifs.

7.7.3 Fragments de *n* chapitres consécutifs.

L'analyse du tableau (89 formes x 66 chapitres) produit plusieurs facteurs (les deux premiers de façon incontestable) qui possèdent globalement, et à un très haut degré, la propriété de rapprocher les chapitres consécutifs.

Cependant, certains des chapitres s'écartent sensiblement des chapitres qui leur sont voisins dans le texte sans que l'on puisse savoir si cet écart doit être mis au compte de ce que l'on pourrait appeler des fluctuations d'échantillonage ou, au contraire, d'une différence profonde dans le stock lexical employé.

Pour tenter d'avancer sur ce point, on procédera à des analyses du même type sur des fragments plus importants. Pour ne pas introduire des découpages qui peuvent nous être suggérés par la tradition de la critique biblique, nous nous sommes orientés vers des découpages en unités sensiblement égales constituées par la réunion d'un même nombre de chapitres consécutifs.

On présentera ici un découpage en 13 fragments de 5 chapitres consécutifs[1]. A l'issue de cette étape qui doit fournir une typologie plus nette, l'unité des fragments constitués de manière aussi brutale sera remise en cause, lors d'une phase consacrée tout particulièrement à l'affinage du découpage.

Tableau 7.10

***Isaïe* : Le regroupement des 66 chapitres en 13 fragments de cinq chapitres consécutifs**

1	*Chap 1 - 5*	*8*	*Chap 36 - 40*
2	*Chap 6 - 10*	*9*	*Chap 41 - 45*
3	*Chap 11 - 15*	*10*	*Chap 46 - 50*
4	*Chap 16 - 20*	*11*	*Chap 51 - 55*
5	*Chap 21 - 25*	*12*	*Chap 56 - 60*
6	*Chap 26 - 30*	*13*	*Chap 51 - 66*
7	*Chap 31 - 35*		

Pour ce dernier découpage, le poids des occurrences décomptées pour l'ensemble des 89 formes varie du simple au triple.

Ici encore, trois axes se détachent nettement avec des pourcentages d'inertie respectivement égaux à : τ_1=29.3 %, τ_2=17.6 % et τ_3=10.0 %.

On retrouve sur le premier axe l'opposition entre les premiers et les derniers fragments. Du côté négatif, les 5 derniers fragments s'opposent fortement au reste. Comme dans la première analyse, les contributions les plus fortes sont dues aux fragments 9 et 10 (Chap 40 à 50) qui portent à eux seuls 60% de l'inertie de l'axe.

Le second axe est constitué par l'opposition du fragment 8 (Chap 36 à 40) à tous les autres. Ici encore, ces fragments sont très bien expliqués par le plan des axes 1 et 2.

Comparaison des deux analyses

En comparant sur l'ensemble des formes les facteurs issus de l'analyse (89 formes x 13 fragments) avec les facteurs correspondants issus de l'analyse des correspondances du tableau (89 formes x 66 chapitres) on s'aperçoit que les deux ensembles présentent de très nombreuses similitudes. Cette circonstance n'est pas surprenante dans la mesure ou ces fragments ont été constitués en opérant, par rapport à la première analyse, des regroupements

[1] Le dernier fragment 13 compte 6 chapitres afin que la partition couvre l'ensemble du livre.

de chapitres consécutifs dont les profils se ressemblent. Dans le cas d'une égalité parfaite des profils à l'intérieur de chaque groupe, le principe d'équivalence distributionnelle (Chapitre 3) nous aurait assuré une totale similitude des deux analyses.

7.7.4 Classification des chapitres

Les résultats des classifications hiérarchiques sont souvent sensibles aux fluctuations dans les données initiales car celles-ci se répercutent à tous les niveaux de la hiérarchie contrairement aux premiers facteurs des analyses des correspondances qui présentent quant à eux, toutes conditions égales conservées, plus de stabilité en général.

La classification ascendante hiérarchique portant sur l'ensemble des 66 chapitres nous a fourni des résultats très médiocres : si nous avons retrouvé les grandes classes dégagées par l'analyse des correspondances, le nombre des exceptions devenait beaucoup trop important pour que nous puissions effectuer une synthèse à partir des classes obtenues.

Dans un second temps on a classé les fragments obtenus plus haut à partir des regroupements de chapitres consécutifs.

On examine le tracé de l'arbre - fig 7.11 - en se souvenant que les seules proximités à interpréter sont celles qui font l'objet de la formation d'une classe.

Ainsi, le fait que les fragments 9 et 10 soient réunis très bas dans l'arbre témoigne d'une proximité relative importante entre eux. Cette classe ne sera réunie au reste des fragments qu'à la dernière étape de l'algorithme d'agrégation.

Par ailleurs, on constate que l'algorithme de classification agrège dans la majorité des cas des fragments consécutifs. La seule exception à cette règle est fournie par la classe 15 qui agrège les fragments 3 et 5 auxquels le fragment 4 vient presque immédiatement se joindre pour former avec eux la classe 17 qui contient les fragments 3, 4 et 5.

Le fait que les classes de la hiérarchie agrègent des fragments consécutifs nous amène à supposer que des parties homogènes, au moins du point de vue de la description à laquelle nous nous livrons dans ce chapitre, existent dans le livre dont la taille dépasse les fragments que nous avons constitués jusqu'à présent.

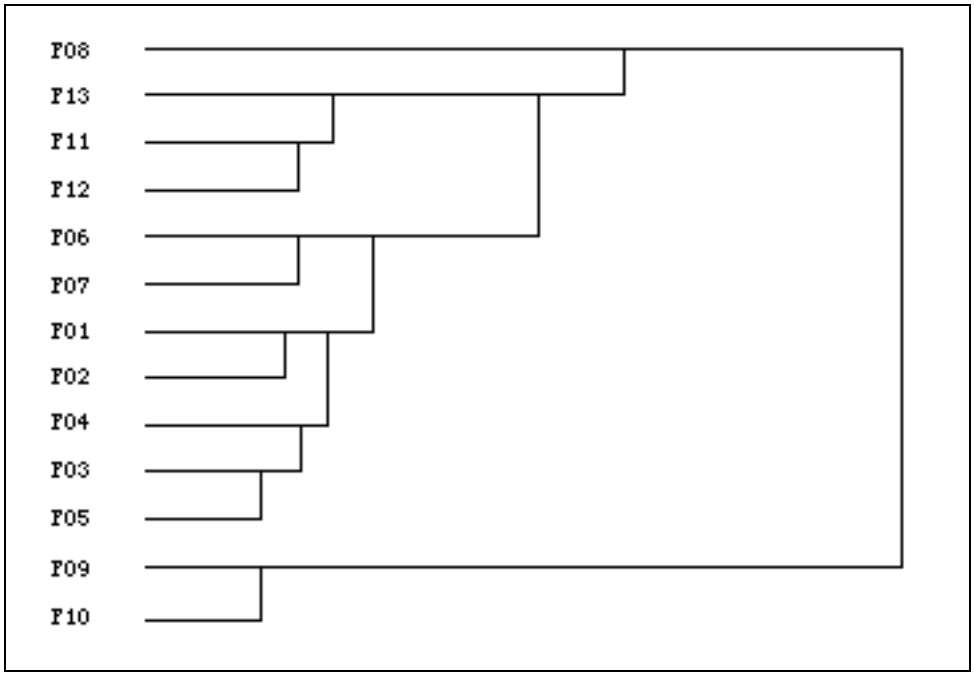

Figure 7.11

***Isaïe* : Classification ascendante hiérarchique à partir du tableau (89 formes x 13 fragments)**

Il est donc possible d'avancer d'un pas et d'étudier l'hypothèse d'une quadripartition provisoire de l'oeuvre selon le schéma présenté au tableau 7.11. Bien entendu les frontières entre les parties ainsi découpées, si elles ne peuvent être précisées, dans cette expérience, au-delà de la frontière de chapitre, peuvent être appréhendées avec plus de précision en analysant les caractéristiques de chacun des chapitres par rapport au fragment auquel il participe.

Tableau 7.11

***Isaïe* : Partition empirique du livre en 4 parties à partir du regroupement de fragments de chapitres consécutifs**

1	Fragments	1 - 7	—	Chap 1 - 35
2	Fragment	8	—	Chap 36 - 40
3	Fragments	9 -10	—	Chap 41 - 50
4	Fragments	11-13	—	Chap 51 - 66

A l'issue de ces expériences, le problème de la détermination du nombre d'auteurs qui ont participé à la rédaction du livre d'Isaïe n'est certes pas résolu. Les facteurs mis en évidence par l'analyse des correspondances à partir du tableau (89 formes x 66 chapitres) possèdent la propriété très

intéressante de rapprocher les chapitres consécutifs et plus généralement les chapitres qui se trouvent être proches dans le texte.

Incontestablement, à l'intérieur de ces grandes classes les chapitres consécutifs présentent des similitudes quant à l'emploi des formes retenues pour l'analyse.