*Chapitre 8*

# Analyse discriminante textuelle

Les méthodes statistiques évoquées dans les chapitres précédents s'appliquent pour la plupart dans le cadre d'une démarche exploratoire, ou, si l'on préfère une terminologie plus traditionnelle, d'une démarche descriptive. Plus dynamique et plus interactive que la simple description, *l'exploration* recourt à la statistique multidimensionnelle pour obtenir des visualisations ou des regroupements d'éléments qui peuvent être soit des textes, soit des unités décomptées à l'intérieur de ces textes : c'est une recherche d'organisations, de traits structuraux, de résumés suggestifs.

Ces méthodes complètent une panoplie (beaucoup plus étendue) de techniques dévolues à des démarches que l'on qualifie souvent de décisionnelles, encore qu'il ne s'agisse en réalité que d'aide à la décision. Dans le domaine de la statistique textuelle, l'aide à la décision pourra concerner l'attribution d'un texte à un auteur ou à une période, la réponse à une requête sous forme de choix d'un document dans une base de donnée, la codification d'une information exprimée en langage naturel.

Les procédures statistiques qui permettent de réaliser ces attributions, ces choix, ces codifications, relèvent de l'analyse discriminante. Elles visent classiquement à prédire l'appartenance d'un "individu" à une classe ou une catégorie (parmi plusieurs possibles) à partir de variables mesurées sur cet individu.[1] Cette prédiction est rendue possible par une phase d'apprentissage réalisée sur un ensemble d'individus pour lesquels les variables et les catégories sont simultanément connues (*learning or training sample*). Nous

[1] Les premières analyses discriminantes furent réalisées à propos de mesures biométriques et anthropométriques par les statisticiens Fisher et Mahanalobis, qui tentaient de prévoir une appartenance ethnique à partir de mensurations effectuées sur le squelette (Fisher, 1936; Mahalanobis, 1936). Ils furent les premiers à utiliser la technique que l'on désigne parfois sous le nom d'analyse factorielle discriminante, ou encore de discrimination linéaire : c'est la méthode la plus ancienne, et c'est encore une des méthodes les plus utilisées actuellement.

avons signalé (Chapitre 2) que les notions de variable et d'individu sont moins évidentes pour les chercheurs qui étudient les matériaux textuels qu'elles ne le sont pour un biométricien ou un démographe. Le découpage et parfois l'élaboration des unités statistiques constituent, on l'a vu, une phase importante de la recherche. L'analyse discriminante textuelle fera grand cas de ces étapes préparatoires.

On distinguera au paragraphe 8.1 deux grandes familles de préoccupations qui s'apparentent, dans le domaine textuel, aux procédures statistiques de discrimination et de reconnaissance des formes. La première famille, fait abstraction du contenu des textes et s'intéresse essentiellement à la forme. Elle relève du domaine de la stylométrie et sera étudiée aux paragraphes 8.2 (brève revue des unités et indices de la stylométrie) et 8.3 (un exemple de modèle statistique en stylométrie, avec suggestion de méthodes alternatives). La seconde famille que l'on désigne ici par *discrimination globale*, parce qu'elle prend en compte le contenu, et parfois le contenu et la forme des textes, est présentée au paragraphe 8.4, puis développée avec le support d'un exemple au paragraphe 8.5.

## 8.1 Deux grandes familles de problèmes dans le domaine textuel

Parmi les applications les plus courantes des techniques qui s'apparentent à la discrimination, on peut distinguer deux grandes séries de préoccupations :

- Les applications à des corpus littéraires (attributions d'auteurs, datation, par exemple) cherchent à s'affranchir du contenu pour saisir des caractéristiques de *forme* (souvent : de style) à partir des distributions statistiques de vocabulaire, d'indices ou de ratios. Il s'agit de saisir des "invariants" d'un auteur ou d'une époque, dissimulés ou peu apparents, comme des habitudes ou des tics insignifiants de prime abord permettent aux détectives de Simenon ou de Conan Doyle d'identifier un individu à son insu.

- A l'opposé, les applications réalisées en recherche documentaire, en codification automatique, dans le traitement des réponses à des questions ouvertes, visent essentiellement le *contenu,* le sens, la substance des textes. La façon dont une réponse est formulée importe moins que son classement dans un groupe de documents présentant une certaine homogénéité au plan du contenu. Nous appellerons *discrimination globale* cette seconde famille d'application.

Dans la pratique, comme nous l'avons signalé plus haut, il n'est pas toujours utile de chercher à maintenir cette distinction. Lors du traitement statistique

de réponses à des questions ouvertes, la forme des réponses, les connotations véhiculées par des termes apparemment équivalent constituent répétons-le une information très riche qui permet souvent de nuancer et d'infléchir les contenus plus manifestes que l'on peut repérer.

### 8.1.1 Discrimination à partir de la forme : la stylométrie

Dans de nombreuses applications à des corpus littéraires, politiques, historiques, tel le problème d'attribution que nous avons vu au chapitre précédent, les analyses discriminantes ont été utilisées pour attribuer à des auteurs connus des textes (poésies, pièces de théâtre, romans, textes religieux ou sacrés, etc.) d'origine mal établie, ou pour déterminer la date ou l'époque de leur écriture. Ces procédures d'attribution n'ont pas toujours fait appel à l'analyse discriminante des biométriciens, comme en témoigne le travail de pionnier de Yule (1944).

C. R. Rao (1989) dans son ouvrage de réflexion générale *Statistics and Truth* cite les travaux plus récents de Thisted et Efron (1987) à propos de l'attribution à Shakespeare d'un poème découvert en 1985, travaux sur lesquels nous reviendrons plus loin. Il cite également l'imposant travail de Mosteller et Wallace (1964) sur l'attribution des "Federalist papers".

Parmi ces 77 textes politiques, publiés anonymement à New York à la fin du dix-huitième siècle, 12 textes n'ont pu être attribués à un auteur parmi deux possibles. Le traitement statistique permet de désigner l'auteur le plus probable de ces 12 textes litigieux. On pourra aussi consulter sur un thème voisin un travail plus récent de Holmes (1992) sur l'homogénéité des "Mormon Scriptures", et une revue assez complète de ce même auteur sur les méthodes de la *stylométrie* (1985).[1]

La plupart de ces méthodes utilisent des indices synthétiques construits à partir des longueurs des mots, des longueurs des phrases, des fréquences de mots outil, de la richesse du vocabulaire, des distributions de fréquence des mots.

L'utilisation systématique de l'analyse des correspondances et des techniques de classification automatique (cf. Benzécri et al., 1981; 1992) a considérablement enrichi ces approches.

[1] La seconde version du travail de Mosteller et Wallace (1984) contient également un panorama général des tentatives d'attribution d'auteur.

### 8.1.2 Discrimination globale : recherche documentaire, codification, validation

De nos jours, la documentation automatique s'est pratiquement érigée en discipline autonome, avec ses revues, congrès, logiciels, sa terminologie et ses concepts (cf. Salton and Mc Gill, 1983 ; Salton, 1988).

Ce sont les dimensions et le contexte des problèmes qui ont induit cette spécificité. En général, on a affaire à de très grands tableaux clairsemés issus de comptages de vocabulaire et de mots-clés spécialisés, selon les domaines, au sein d'ensembles qui comportent plusieurs centaines de milliers de documents souvent courts et parfois stéréotypés. La finalité de cette opération a souvent un caractère très pragmatique. Il s'agit de faire fonctionner un outil, avec des taux d'échecs, des systèmes de coûts et de contraintes préalablement définis.

Dans ce cadre particulier, il est possible de faire appel à plusieurs sources extérieures d'information pour résoudre les problèmes de classement : des analyseurs syntaxiques, premiers pas vers une *compréhension* de la requête, des dictionnaires ou des réseaux sémantiques pour lemmatiser et désambiguïser les requêtes[1], éventuellement à des corpus artificiels faisant appel à des experts.[2]

Mais beaucoup de techniques utilisées, et parmi les plus efficaces si l'on en croit leurs auteurs, ont recours à des outils multidimensionnels très proches de ceux préconisés par Benzécri (1973, 1977, 1981), ou dans le cadre des grands tableaux clairsemés par Lebart (1982a). Deerwester et al. (1990) proposent ainsi sous le nom de *Latent Semantic Analysis* une méthode très semblable à la discrimination d'après les premiers facteurs d'une analyse des correspondances (ces auteurs utilisent en fait une décomposition aux valeurs singulières, technique qui est à la base de l'analyse des correspondances et de l'analyse en composantes principales).

D'autres auteurs insistent sur la complémentarité entre modèles et méthodes descriptives dans la recherche documentaire (cf. Fuhr et al., 1991), et sur l'intérêt de visualisations les plus synthétiques possibles (Fowler et al., 1991). On peut dire que la tendance actuelle est à l'effacement progressif des barrières interdisciplinaires, et que, de plus en plus, les "décisionistes purs et durs" reconnaissent l'importance des phases de description et d'exploration.

Nous allons présenter ces deux volets (stylométrie et discrimination globale) en nous appuyant sur deux exemples de méthodes et d'applications qui

---

[1] Cf. par exemple, en matière de classification de documents, les travaux de Blosseville et al. (1992), Hebrail et al. (1990, 1993).

[2] Cf. les travaux de pionniers de Palermo et Jenkins (1964). Cf. également Bouroche et Curvalle (1974).

mettent en évidence, pensons-nous, la divergence des démarches, et leurs intérêts respectifs.

Après un bref exposé des spécificités des unités statistiques et indices de la stylométrie, le premier exemple (paragraphe 8.3) reprend et reconsidère le problème stylométrique, évoqué plus haut, d'attribution éventuelle à Shakespeare d'un poème récemment découvert. Le second exemple (paragraphe 8.5) présente pour illustrer l'analyse discriminante globale une application multilingue dans le cadre d'une enquête internationale.

## 8.2 Les unités et indices de la stylométrie

Ici encore, les unités de bases seront dérivées de la forme graphique, mais il convient de noter que de nombreux travaux stylométriques ont pu utiliser, pour comparer des textes, des auteurs ou des genres, des comptages qui fractionnent encore cette unité[1].

Smith (1983) a montré les limites de la distribution du nombre de lettres par mot, alors que la distribution du nombre de syllabes par mots, proposée de façon systématique par Fuchs (1952), a été jugée utilisable, au moins pour les attributions d'auteurs anglophones, par Brainerd (1974).

Même les distributions de fréquences globales des graphèmes (lettres et ponctuation) peuvent avoir un pouvoir discriminant important entre textes (Brunet, 1981 ; Abi Farah, 1988 [textes en langue arabe] ; Salem, 1993). Il s'agit surtout d'un exercice méthodologique, car on lit à travers les graphèmes la spécificité des vocabulaires. Brunet a cependant montré que l'effet discriminant des graphèmes subsistait même après élimination des formes les plus fréquentes, ce qui rend encore plus délicate l'interprétation de tels effets.

### 8.2.1 "Mots-outil", parties du discours

Comme nous l'avons expliqué plus haut, la notion de mot-outil n'est pas une notion de la statistique textuelle dans la mesure ou elle ne se prête à aucune formalisation satisfaisante. De nombreux auteurs utilisent cependant cette notion en s'appuyant sur l'intuition commune que l'on peut dresser, dans chaque langue, une liste de formes que l'on appelle parfois encore "mots-

[1] Hors du cadre de la stylométrie, on trouvera une analyse intéressante des constituants du mot français dans Gruaz (1987).

grammaticaux" et qui ont en commun la propriété d'être moins marqués au plan sémantique[1].

Sur la base d'une telle liste Demonet et al. (1975) ont proposé de mesurer un "taux de fonctionnalité" propre à chaque discours ou à chaque type de discours en recensant le nombre des occurrences qui correspondent à des occurrences d'une liste de "mots-outil" préalablement dressée par eux. Cette propriété a par ailleurs, été largement utilisée, dans les études informatisées, pour écarter de la liste des formes considérées des unités correspondant en grande partie aux formes les plus fréquentes, réputées peu dignes d'intérêt.

A l'inverse, le courant qui s'occupe des problèmes d'attribution d'auteur a longtemps privilégié l'étude de ce type d'unité, posant que leur emploi, moins maîtrisé lors de la rédaction du texte pouvait constituer une marque d'auteur privilégiée[2].

C'est le sens des travaux de pionnier de Ellegard (1962), qui compare les proportions de "mots-outil" dans le corpus des "Junius Letters" (pamphlets publiés à la fin du dix-huitième siècle comportant environ 150 000 occurrences), avec celles calculées dans un autre corpus de la même époque. Cette démarche constitue également une phase importante des travaux de Mosteller et Wallace (1964, op. cit.) ainsi que de ceux de Holmes (1992).

Benzécri a réalisé des typologies de textes en grec ancien (1991a), latins (1991b), et espagnols (1992b) à partir d'ensembles de mots outil, mettant en évidence à la fois les problèmes que pose la sélection de ces unités statistiques, et le pouvoir discriminant des profils de mots outil lorsque ceux-ci interviennent comme éléments actifs d'une analyse des correspondances.

Une catégorisation des unités graphique du texte (analyse syntaxique qui permet d'affecter à chaque forme une catégorie grammaticale[3]), permet de calculer la proportion de noms, de verbes, d'adjectifs, etc. Ces proportions ont également été utilisées en stylométrie. Ainsi, Somers (1966) affirme que la proportion de substantifs dénote instruction et aisance dans l'expression, Brainerd (1974) étudie le pouvoir discriminant de la proportion d'articles.

---

[1] Ce dernier trait n'exclut pas de s'intéresser à de telles unités dans certains domaines d'étude. Que l'on songe, par exemple au rôle important que peuvent jouer dans l'analyse de textes politiques les mots-outil *pour* et *contre*.

[2] Cf. par exemple les travaux de Radday (1974) et Morton (1963) évoqués au chapitre 7, paragraphe 7.7, à propos de l'homogénéité du livre d'Isaïe.

[3] Aujourd'hui, la réalisation d'une telle opération ne peut être entièrement confiée à un ordinateur. Des progrès importants ont été réalisés dans le domaine de l'analyse syntaxique automatisée des textes, comme en témoigne, par exemple, l'amélioration constante des correcteurs orthographiques que l'on trouve désormais sur la plupart des machines de traitement de texte.

Précisons qu'une telle catégorisation est évidemment nécessaire pour identifier d'éventuels mots-outil.[1]

### 8.2.2 La richesse du vocabulaire

Pour conclure sur ce bref survol des unités statistiques de la stylométrie, il faut mentionner les indices de richesse de vocabulaire : Si $V$ désigne le nombre de formes différentes, et $T$ le nombre total d'occurrences, pour chaque texte, les premiers indices proposés furent les quotients du nombre de mots distincts $V$ par le nombre total $T$ d'occurrences : $R = V / T$ (*type -token ratio*) ou encore $R' = LogV / LogT$ .

Ce quotient a été proposé en particulier par McKinnon et Webster (1971) pour discriminer des textes écrits sous différents pseudonymes par le philosophe danois Sören Kierkegaard. Appliqué à un ensemble de seize échantillons qui varient environ du simple au double - de 6 548 occurrences à 15 432 - un tel ratio permet à ces auteurs de conclure à la possibilité de caractériser les différentes catégories de textes à partir de ce seul critère.

Dans les corpus de textes, les variations de longueurs entre les parties peuvent de plus être considérables, et les deux quotients précédents ont manifestement l'inconvénient de trop dépendre de la longueur des textes (pour une langue donnée, $V$ est borné supérieurement, alors que $T$ n'a pas de limite a priori).

L'indice $D$ de Simpson (1949) est le quotient :

$$D = \Sigma_r \, r(r-1)V_r \, / \, T(T-1)$$

où $V_r$ est le nombre de formes distinctes apparaissant exactement $r$ fois dans le texte. C'est donc le quotient du nombre de paires d'occurrences d'une même forme par le nombre total de paires d'occurrences. Le numérateur n'est plus borné, et $D$ dépend beaucoup moins de $T$ que $R$ ou $R'$. $D$ est simplement la probabilité que deux occurrences prises au hasard dans le corpus correspondent à une même forme.
Cet indice bien connu des statisticiens est très proche dans son principe de la caractéristique $K$ introduite, très antérieurement, par Yule dans le domaine des études stylistiques et qui peut être définie comme :

$$K = 10^4 \; D \, (T-1)/T$$

[1] Notons que si l'isolement de mots-outil demande une catégorisation et une désambiguïsation du texte (cas de la forme *pas*, par exemple) il existe aussi des locutions contenant des mots pleins qui sont des substituts de mots-outil, et qu'une lemmatisation préliminaire pourrait masquer (cf. par exemple Benzécri, 1992a).

## 8.3 Modèles statistiques en stylométrie : un exemple

On comprend, à la lecture de ce qui précède, que domaines, préoccupations et méthodologie ne sont pas indépendants. Ainsi, l'essentiel des travaux stylométriques met en oeuvre des méthodes statistiques unidimensionnelles. Il s'agit le plus souvent de calculs de paramètres discriminants, à l'exception des travaux, en général plus récents, mettant en oeuvre des vecteurs de mots-outil, ou des combinaisons d'indices stylométriques divers.

A côté des études empiriques utilisant la flore d'indices descriptifs évoquée précédemment, il existe une démarche plus "modélisatrice", moins facile d'accès, mais qui ne manque pas d'intérêt[1].

Nous allons nous appuyer sur le cas déjà cité de l'étude d'attribution d'auteur réalisée à propos du poème (peut-être écrit par Shakespeare) découvert en 1985 (Thisted et Efron, op.cit.) en retraçant brièvement l'histoire de la démarche, les hypothèses, le modèle, les résultats, et en suggérant d'autres approches.

### 8.3.1 Modélisations de la gamme des fréquences

Nous mentionnerons sous ce chapitre certaines applications au domaine textuel des modèles de capture de l'écologie des années cinquante. Les écologistes disposent des pièges ou des trappes pour capturer des spécimens de diverses espèces.

Si l'on fait l'hypothèse, raisonnable et acceptable empiriquement, que le nombre de spécimens effectivement capturés d'une espèce donnée suit une *loi de Poisson* dont l'unique paramètre dépend de l'espèce, on peut estimer, après un certain temps de capture, à la fois le nombre de spécimens par espèce, et également, ce qui peut paraître surprenant, le nombre probable d'espèces "piègeables" dans la zone de capture...

Avant de préciser ces résultats à l'intention des lecteurs plus mathématiciens (pour lesquels le modèle - non-paramétrique - sera présenté en section 8.3.3) traduisons-les en termes de statistique textuelle.

A l'espèce correspond ici la forme graphique, au spécimen une occurrence de cette forme. Il va de soi qu'il ne s'agit que de modéliser le bilan des

[1] On insistera ici sur les modèles de génération de la gamme des fréquences, et non sur les modèles d'ajustement. A cette dernière catégorie appartient par exemple le modèle de distribution de Waring-Herdan, dont on trouvera par exemple une présentation détaillée dans le manuel classique de Muller (1977).

distributions lexicales d'une oeuvre à un instant donné, et non le processus de création. Nous n'avons pas jusqu'à présent fait d'hypothèse d'indépendance entre les nombres d'occurrences par forme, ce qui eût été difficilement acceptable, même si les notions d'associations et de cooccurrences disparaissent au niveau d'un bilan.

Ce modèle permet donc d'estimer, pour un *texte* donné l'évolution de la gamme des fréquences (nouvelles fréquences des formes déjà utilisées, et nouvelles formes) à mesure que le volume du texte augmente. Etant donné un corpus de textes sur lesquels des comptages ont déjà été effectués et une oeuvre nouvelle, il est donc possible de voir si elle satisfait au canon défini par toutes les oeuvres antérieures, y compris dans l'accroissement du vocabulaire qu'elle implique.

En fait, le modèle n'est valide que si le processus est homogène dans le temps : il ne s'applique pas a priori (sauf vérification empirique) au cas de changement de genre (théâtre en vers versus théâtre en prose, par exemple).

### 8.3.2 Les matériaux disponibles pour le problème d'attribution

A partir de la somme de Spevack (1968), il a été possible de connaître la distribution du nombre de mots distincts en fonction de leur fréquence d'apparition dans l'oeuvre de Shakespeare (tableau 8.1).

Selon les décomptes effectués par cet auteur, Shakespeare aurait employé 31 534 formes distinctes, pour un total de 884 647 occurrences. On notera que prés de la moitié des mots distincts sont des hapax.[1]

A partir de cette distribution, et du modèle esquissé plus haut et que l'on considérera plus loin en détail, il est possible d'estimer non seulement les fréquences d'apparition de chacune des formes, mais également les fréquences de nouvelles formes. C'est ce qui a conduit Efron et Thisted (1976) à tenter de répondre à la question *"How many words did Shakespeare know?"*, et à estimer à 35 000 le nombre de mots distincts supplémentaires si le processus de "création" s'était poursuivi indéfiniment.

Cette extrême sollicitation du modèle le conduit très certainement en dehors de son domaine de validité, car comme nous l'avons précisé plus haut, l'hypothèse d'homogénéité du processus est fondamentale, et celle-ci est peu vraisemblable sur longue période. De plus, considérer que les mots qui

[1] Bien entendu, cette précision est tout-à-fait illusoire, puisque pour une pièce aussi étudiée que *Hamlet*, il existe des paragraphes entiers absents dans certaines sources, alors que des désaccords sur l'identification des mots subsistent pour les parties de textes communes à toutes les sources.

pourraient être utilisés plus tard sont des mots tous connus actuellement constitue une hypothèse supplémentaire. Mais le plus surprenant, pour un statisticien, est l'ordre de grandeur somme toute raisonnable du résultat obtenu.

**Tableau 8.1**

**Gamme des fréquences dans l'oeuvre de Shakespeare (tableau partiel)**

| *Fréq. $f_i$* | *Effectif $V_i$* | *Fréq. $f_i$* | *Effectif $V_i$* |
|---|---|---|---|
| *1* | 14376 | *8* | 519 |
| *2* | 4343 | *9* | 430 |
| *3* | 2202 | *10* | 364 |
| *4* | 1463 | *11* | 305 |
| *5* | 1043 | *12* | 259 |
| *6* | 837 | ..... | |
| *7* | 638 | *>12* | 846 |

*Nombre des formes* $= \Sigma_i V_i$ 31 534

*Nombre des occurrences* $= \Sigma_i f_i V_i$ 884 647

C'est en 1985, neuf années après la parution de l'article précité qu'a été découvert par Gary Taylor un poème formé de neuf strophes et de 258 mots (429 occurrences) pouvant être attribué à Shakespeare. Ce fut l'occasion pour Thisted et Efron (1987) d'appliquer leur modèle sur un accroissement très modeste (t = 0,05 %) de l'oeuvre de cet auteur.

Par prudence, ces auteurs traitent simultanément le cas de sept autres poèmes élisabéthains définitivement attribués, dont quatre sont dus à Shakespeare.

***Les huit poèmes élisabéthains :***

| | |
|---|---|
| **Ben Jonson** | An Elegy |
| **C. Marlowe** | Four poems |
| **J.Donne** | The Ecstasy |
| -------- | |
| **Shakespeare** | Cymbeline (extraits) |
| **Shakespeare** | A Midsummer Night's Dream (extraits) |
| **Shakespeare** | The Phoenix and Turtle |
| **Shakespeare** | Sonnets (extraits) |
| **Shakespeare** | **(?)** Taylor's Poem |

Les textes ont été choisis de façon à être sensiblement équivalents en ce qui concerne la longueur, et comparables du point de vue du genre et des grands thèmes abordés.

Le tableau 8.2 ci-après contient les distributions des formes de ces huit poèmes selon le canon Shakespearien : ainsi, le poème de Taylor contient 9 formes jamais utilisées par Shakespeare, et 140 formes utilisées par lui plus de 100 fois.

## Tableau 8.2

**Distribution des formes dans les huit poèmes, selon leur fréquence d'apparition dans l'oeuvre de Shakespeare.**

| ***Freq.*** | **BJon** | **Marl** | **Donn** | **Cymb** | **Mids** | **Phoe** | **Sonn** | **Tayl** | ***Total*** |
|---|---|---|---|---|---|---|---|---|---|
| *0* | 8 | 10 | 17 | 7 | 1 | 14 | 7 | 9 | 73 |
| *1* | 2 | 8 | 5 | 4 | 4 | 5 | 8 | 7 | 43 |
| *2* | 1 | 8 | 6 | 3 | 0 | 5 | 1 | 5 | 29 |
| *3-4* | 6 | 16 | 5 | 5 | 3 | 9 | 5 | 8 | 57 |
| *5-9* | 9 | 22 | 12 | 13 | 9 | 8 | 16 | 11 | 100 |
| *10-19* | 9 | 20 | 17 | 17 | 6 | 18 | 14 | 10 | 111 |
| *20-29* | 12 | 13 | 14 | 9 | 9 | 13 | 12 | 21 | 103 |
| *30-39* | 12 | 9 | 6 | 12 | 4 | 7 | 13 | 16 | 79 |
| *40-59* | 13 | 14 | 12 | 17 | 5 | 13 | 12 | 18 | 104 |
| *60-79* | 10 | 9 | 3 | 4 | 9 | 8 | 13 | 8 | 64 |
| *80-99* | 13 | 5 | 10 | 4 | 3 | 5 | 8 | 5 | 53 |
| *+100* | 148 | 138 | 145 | 120 | 103 | 111 | 155 | 140 | 1060 |
| ***Total*** | 243 | 272 | 252 | 215 | 156 | 216 | 264 | 258 | 1876 |

Les auteurs ont calculé, pour chaque case de ce tableau, la valeur théorique fournie par le modèle dans l'hypothèse selon laquelle chaque texte serait un texte supplémentaire de Shakespeare. Ainsi, par exemple, pour le poème *Taylor*, le tableau 8.3 donne les valeurs théoriques calculées dans cette hypothèse, selon une formule qui est explicitée dans la section technique 8.3.3 ci-après.

L'adéquation des valeurs observées aux valeurs théoriques est calculée pour les huit poèmes en utilisant une grille plus fine (les 100 valeurs de fréquence de 0 à 99) en utilisant un modèle linéaire généralisé. Les auteurs sont ainsi conduits à rejeter comme non-shakespeariens tous les poèmes connus comme tels, mais aussi *The Phoenix and Turtle*, de Shakespeare.

Le poème *Taylor* n'est pas rejeté. Le test apparaît donc comme plutôt trop sévère, puisqu'il rejette même un poème de Shakespeare, ce qui renforce pour ces auteurs la présomption d'appartenance du nouveau poème au corpus shakespearien.

**Tableau 8.3**

**Valeurs théoriques et observées pour le poème *Taylor***

| *Fréquences* | *Taylor observé* | *Taylor théorique* |
|---|---|---|
| 0 | 9 | 6.97 |
| 1 | 7 | 4.21 |
| 2 | 5 | 3.33 |
| 3-4 | 8 | 5.36 |
| 5-9 | 11 | 10.24 |
| 10-19 | 10 | 13.96 |
| 20-29 | 21 | 10.77 |
| 30-39 | 16 | 8.87 |
| 40-59 | 18 | 13.77 |
| 60-79 | 8 | 9.99 |
| 80-99 | 5 | 7.48 |

### 8.3.3 Précision sur le modèle non-paramétrique d'estimation

A l'usage du lecteur statisticien, on précise dans ce paragraphe le modèle esquissé plus haut[1]. L'hypothèse de base concerne la distribution de la forme s qui est une loi de Poisson de paramètre $\lambda_S$. Il y a potentiellement $S$ formes, mais toutes ne sont pas observées.

On désignera par $G(\lambda)$ la fonction de répartition empirique (inconnue) des nombres $\lambda_1$ , $\lambda_2$ , ... $\lambda_S$. Désignons également par $n_x$ le nombre de formes distinctes observées exactement $x$ fois.

L'espérance mathématique $\eta_x$ du nombre exact de formes distinctes observées exactement $x$ fois dans le passé, représenté ici conventionnellement par l'intervalle [-1, 0] s'écrit :

$$\eta_x = E(n_x) = s\int_0^{\infty} e^{-\lambda} \frac{\lambda^x}{x!} dG(\lambda) \qquad (8.1)$$

Notons que l'espérance $E_t$ du nombre de formes distinctes observées à l'instant $t$ mais non observées auparavant (formes nouvelles) s'écrit de façon simple.

[1] Le lecteur non-matématicien pourra se dispenser de la lecture de ce paragraphe et se reporter directement au paragraphe suivant.

$$E_t = s\int_0^{\infty} e^{-\lambda}(1-e^{-\lambda t})dG(\lambda) \qquad (8.2)$$

En substituant le développement :

$$1-e^{-\lambda t} = \lambda t - \frac{\lambda^2 t^2}{2!} + \frac{\lambda^3 t^3}{3!} - \ldots \qquad (8.3)$$

dans l'équation 8.2, et en tenant compte de 8.1, on trouve que $E_t$ peut s'écrire comme la somme de la série :

$$E_t = \eta_1 t - \eta_2 t^2 + \eta_3 t^3 + . \qquad (8.4)$$

Cette étonnante relation, découverte par Good et Toulmin (1956) donne donc l'espérance du nombre de nouvelles formes pour une proportion supplémentaire *t* de texte, en fonction des espérances des nombres de formes $\eta_x$ observées *x* fois au cours du passé, sans qu'il n'ait été fait aucune hypothèse sur la forme de *G(λ),* qui n'intervient pas dans le calcul.

Cette formule a permis de calculer le premier terme du tableau 8.3, en substituant aux valeurs théoriques $\eta_x$ les valeurs calculées $n_x$. La valeur de *t*, quotient des nombres d'occurrences, est de :

$$t = 4.85\ 10^{-4}\ \ (= 429 / 884\,647).$$

Avec une valeur si petite, seul le premier terme *14 376 x t (= 6.97)* est non-négligeable.

Ainsi, avec ce modèle, le nombre de formes nouvelles, dans le cas d'un très petit accroissement relatif du volume des écrits, s'obtient par une simple règle de trois à partir du nombre d'hapax antérieurement observé.

Intéressons-nous plus généralement à l'espérance mathématique $\nu_x$ du nombre exact de formes distinctes observées exactement *x* fois dans l'intervalle [-1 , 0], et présentes dans l'intervalle [0 , t]. Elle vaut :

$$\nu_x = E(n_x) = s\int_0^{\infty} e^{-\lambda}\frac{\lambda^x}{x!}(1-e^{-\lambda t})dG(\lambda)$$

La même procédure de substitution utilisant le développement 8.3 nous donne cette fois-ci une formule un peu plus complexe :

$$\nu_x = \sum_{k=1}^{\infty} (-1)^{k+1} \binom{x+k}{k} t^k \eta_{x+k}$$

Pour obtenir une estimation de cette dernière quantité, on peut substituer aux valeurs théoriques $\eta_X$ les valeurs calculées $n_X$ :

$$\nu_x^* = \sum_{k=1}^{\infty} (-1)^{k+1} \binom{x+k}{k} t^k n_{x+k}$$

Il s'agit d'un *estimateur empirique bayésien* de $\nu_X$ .

Pour $x = 0$, (formes nouvelles) on retrouve la formule 8.4.

C'est à partir de cette formule que sont calculées les valeurs théoriques telles que celles qui figurent dans le tableau 8.3.

### 8.3.4 Autres approches du problème

Considérons de nouveau le tableau 8.2 donnant la distribution des formes des huit poèmes selon les classes de fréquences d'apparition des mêmes formes dans l'oeuvre de Shakespeare.

Si la distribution des formes par fréquence d'emploi dans l'oeuvre de Shakespeare peut être considérée comme caractéristique de cet auteur, on doit s'attendre à une certaine hétérogénéité dans les profils des colonnes.

Effectivement, le chi-2 (test d'indépendance classique sur les tables de contingence), calculé (malgré la faiblesse de certains effectifs), vaut, pour 77 degrés de liberté (77 = [8 - 1] x [12 - 1] ) , et avec les notations usuelles (voir chapitre 3) :

$$\chi^2 = 1876 \times \Sigma_{ij} (f_{ij} - f_{i.} f_{.j})^2 / f_{i.} f_{.j} = 104.7$$

Cette valeur a sensiblement une chance sur 100 d'être dépassée dans l'hypothèse d'homogénéité, qui correspond ici à l'indépendance des lignes et des colonnes de la table. Cette hypothèse est donc douteuse.

L'un des mérites de l'analyse des correspondances est de nous décrire, dans le cas d'un rejet de l'hypothèse d'indépendance, comment cette hypothèse est rejetée.
La figure 8.1 décrit donc les positions des poèmes et des classes de fréquences dans le plan factoriel. Le poème litigieux *Taylor* est proche de l'origine, donc du profil moyen.

Les poèmes de Shakespeare sont alignés le long d'une droite joignant *The Phoenix* à *Midsummer*.

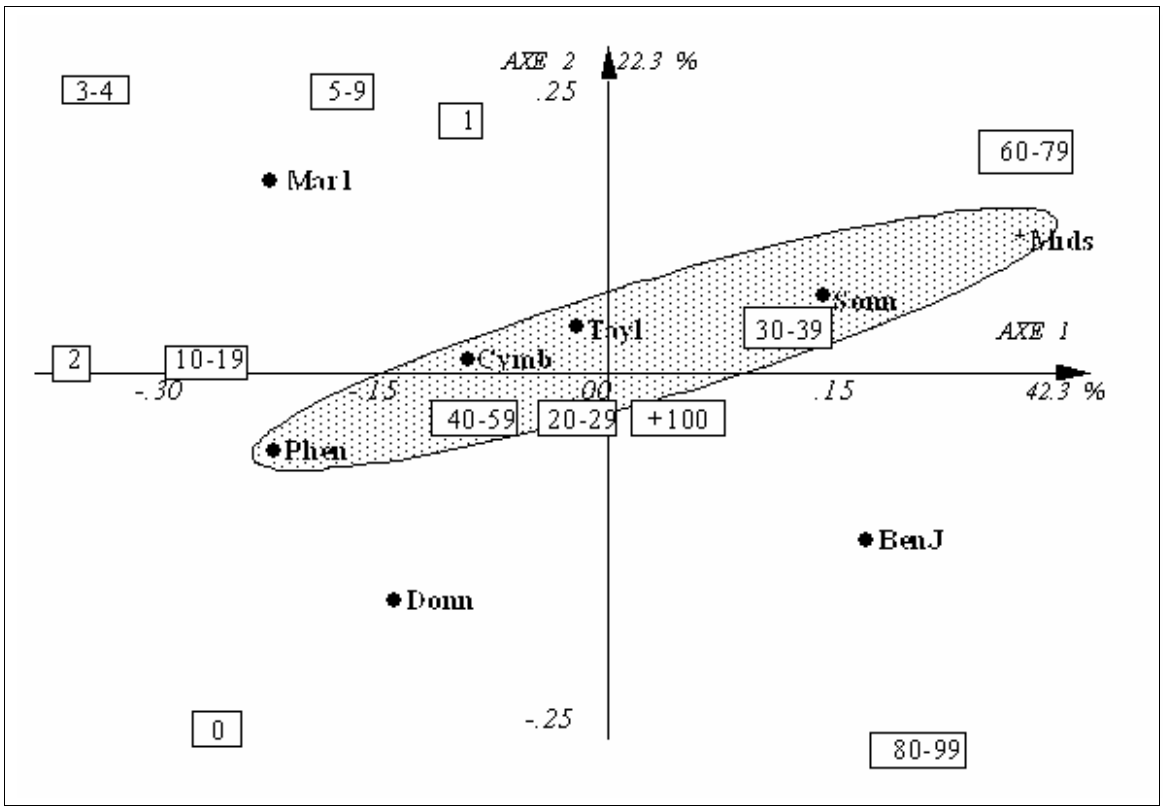


**Figure 8.1**

**Premier plan factoriel de l'A.C. de la table 12 x 8**
**Classes de fréquence x Poèmes (tableau 8.2)**
**La zone ombrée contient les poèmes attribués à Shakespeare et le poème *Taylor***

*The Phoenix*, qui a été rejeté comme non-shakespearien par les tests issus du modèle poissonien, est effectivement assez périphérique, entre *Marlowe* et *Donne*, dans une zone riche en mots nouveaux (fréquence "0" chez Shakespeare) et en fréquences "2" et "10-19". A l'opposé, l'extrait de *Midsummer* est anormalement pauvre en formes nouvelles (exclusives).

Cette visualisation nuancée, fondée sur des comptages n'ayant subi aucune transformation, donne l'idée d'une approche empirique élargie, qui consisterait à multiplier les poèmes analysés (aussi bien de Shakespeare que d'autres auteurs élisabéthains) de façon à pouvoir déterminer des zones de rejets dans un plan factoriel analogue à celui de la figure 8.1. Ces zones seraient fondées sur des variations de densité dans le plan, et non sur un tout petit nombre de jalons.
Il s'agit là d'une approche voisines de l'analyse discriminante telle qu'elle sera évoquée dans la section suivante, à ceci près que les variables de bases utilisées jusqu'à présent (répartition des formes selon des classes de fréquence pour un auteur donné) sont réputées indépendantes du contenu.

## 8.4 Analyses discriminantes globales

Dans les analyses qui vont suivre, à l'usage de la recherche documentaire ou des analyses ou codifications de réponses libres dans les enquêtes, la discrimination se fera avec l'ambition de prendre en compte forme et contenu, ou parfois contenu seul. Précisons, s'il en est besoin, que prendre en compte n'est pas comprendre.

### 8.4.1 Principe général

Le principe de base commun aux méthodes d'analyse discriminante les plus répandues est le suivant :

a) on dispose de $n$ points dans un espace de dimension élevée $p$, et ces $n$ points sont répartis en $K$ catégories, étiquetées de 1 à $K$.

- *exemple 1* : 65 textes, chacun caractérisé par son profil lexical calculé à partir des 200 formes graphiques les plus fréquentes du corpus global (espace à $p = 200$ dimensions : $\mathbf{R}^{200}$). Les $n = 65$ textes correspondent à $K = 2$ auteurs.

- *exemple 2* : les 1500 points sont des réponses à une question ouverte caractérisées une sélection de 100 formes ou segments (espace à $p = 100$ dimensions : $\mathbf{R}^{100}$). Les répondants ont (ou n'ont pas) acheté un certain produit ($K = 2$).

- *exemple 3* : chaque point, caractérisé par quelques mots-clés choisis parmi les 1 500 d'un thesaurus, correspond à un document dans un ensemble qui en comprend 10 000 (espace à $p = 1\,500$ dimensions : $\mathbf{R}^{1500}$). Ces documents sont répartis en $K = 200$ sous-matières.

b) On dispose également de $m$ points, dans le même espace à $p$ dimensions que les $n$ points précédents (c'est-à-dire décrits par les mêmes variables ou les mêmes profils), mais ces points sont sans étiquette. Et l'on désire affecter à chacun de ces $m$ points l'étiquette la plus probable parmi un ensemble d'étiquettes définies a priori.

L'exemple 1, pour $m = 12$, correspondant au problème cité des *Federalist papers* : il s'agit de savoir lequel de deux auteurs a écrit chacun des 12 textes.

L'exemple 2, pour $m = 1$ est, une tentative de prévision de comportement d'achat à partir d'une réponse libre.

L'exemple 3 est un problème de classement de document.

Une méthode simple consiste à calculer dans chaque cas les $K$ profils moyens dans $\mathbf{R}^p$ et à affecter chaque nouveau point au profil qui lui est le plus proche.

On peut aussi tenir compte de la forme de chacun des $K$ *sous-nuages* de points, car, comme l'illustre la figure 8.2 la proximité au point moyen n'est pas toujours le meilleur indice d'appartenance à la classe.

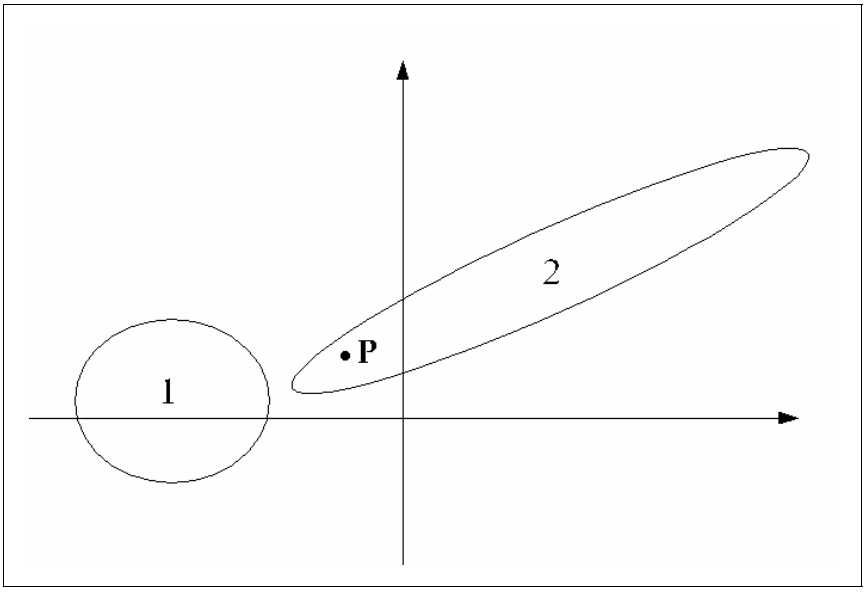


**Figure 8.2**

**Le point P est plus proche du centre 1 que du centre 2, mais appartient plus vraisemblablement à la classe 2, compte tenu des densités observables.**

Comme on le pressent, cette notion d'indice d'appartenance n'est pas une chose simple, et de nombreuses variantes techniques seront possibles pour que les distances calculées puissent s'interpréter en termes de probabilités. Ainsi, dans le cas de la figure 8.2, la *distance de Mahalanobis locale*, qui tient compte de la forme des ellipsoïdes[1] correspondant à chaque sous-nuage, rendra le point P plus proche du point 2 que du point 1.
Les principales méthodes de discrimination adaptées aux vastes tableaux de données qualitatives sont :

- la *discrimination sous indépendance conditionnelle* (cf. Goldstein et Dillon, 1978 ; Bochi et al., 1993),

[1] La distance de Mahalanobis locale du point $\mathbf{X}$ au groupe k, qui est utilisée en *analyse discriminante quadratique*, s'écrit $d_k(\mathbf{X}) = (\mathbf{X}-\mathbf{m}_k)' \mathbf{S}_k^{-1}(\mathbf{X}-\mathbf{m}_k)$ où $\mathbf{S}_k$ est la matrice de covariance interne au groupe $\mathbf{k}$, de point moyen (centre de gravité) $\mathbf{m}_k$.

- la *discrimination par estimation directe de densité* (Habbema et al., 1974 ; Aitchison et Aitken, 1976),
- la *discrimination par la méthode des plus proches voisins* (cf. par exemple, Hand, 1986),
- la *discrimination sur coordonnées factorielles* (qui permet d'appliquer aux variables qualitatives toutes les méthodes qui s'appliquent aux variables numériques : analyse factorielle discriminante, discrimination quadratique, etc.) sur laquelle nous reviendrons (Benzécri, 1977 ; Wold, 1976).

On trouvera une revue récente et complète de la plupart de ces méthodes dans l'ouvrage de McLachlan (1992). Mais ces méthodes ne s'appliquent qu'après la détermination des unités statistiques de base, détermination dont on connaît l'importance dans le cas des données textuelles.

### 8.4.2 Unités pour la discrimination globale

Il est possible de faire varier la nature et la qualité de la discrimination en agissant sur le vocabulaire de base de différentes façons :

a) Les formes peuvent être sélectionnées à l'aide d'un seuil de fréquence préalablement fixé.

b) Les décomptes de formes peuvent être enrichis de comptages portant sur les segments et les quasi-segments,

c) Le texte peut être lemmatisé, (avec ou sans élimination des mots-outil),

d) Seules les formes (segments, lemmes, etc.) caractéristiques des classes à discriminer peuvent être sélectionnées a priori,

A partir du vocabulaire de travail ainsi constitué, il est possible :

e) De procéder à une analyse des correspondances préalable de la table (unités statistiques, individus), et de ne retenir que les premiers axes (filtrage et régularisation par analyse des correspondances),

f) De procéder à une classification des unités statistiques (cf. Hand, 1981 ; Celeux et al., 1991), et de travailler sur des agrégats d'unités.

Toutes ces possibilités impliquent donc que l'utilisateur définisse une stratégie. Même dans le cas d'une discrimination globale, les domaines et les préoccupations vont avoir une influence sur ces transformations préliminaire du vocabulaire de base.

Ainsi, dans le cas de réponses à des questions ouvertes (ou de recherche documentaire ou simplement de textes très courts), le fait de lemmatiser (procédure c) ou de regrouper (procédure f) ou encore de combiner les deux (procédure que l'on note c x f) modifie le résultat des calculs de distances entre textes. Deux textes avec des profils disjoints au départ peuvent acquérir des éléments communs après ces transformations : deux flexions d'une même unité lexicale se rassemblant en un élément unique. Ainsi, lemmatisation et/ou regroupements peuvent rapprocher deux éléments (réponses, documents) n'ayant au départ aucune forme en commun lors de la phase de segmentation automatique du texte.

### 8.4.3 Discrimination et réponses modales

Le résultat de la numérisation et des étapes de transformations préliminaires conduit à une matrice **T**, d'ordre (*n,V*) analogue à celle qui a été définie au chapitre 5, dont les lignes correspondent aux *n* individus, documents ou textes, et les colonnes aux *V* différentes unités textuelles (formes, segments, lemmes, etc..). Les *K* groupes à discriminer définissent une partition des individus, documents ou textes. Ils peuvent être décrit par une matrice binaire **Z** d'ordre (*n, K*) comme celle qui a été définie pour caractériser une question fermée.

On a vu dans ce chapitre que, pour toute question fermée dont les réponses sont codées dans un tableau **Z**, il est possible de calculer le tableau lexical agrégé **C** par la formule :

$$\mathbf{C = T'Z}$$

et donc de comparer les profils lexicaux de différentes catégories de population.

Plusieurs outils permettent d'aider la lecture des tableaux lexicaux agrégés : l'analyse des correspondances, les techniques de classification qui lui sont complémentaires, les listes de formes caractéristiques, les listes de réponses modales.

Chacun de ces outils va permettre de construire une ou plusieurs techniques de discrimination. En particulier, la procédure de calcul des réponses modales est en elle-même une analyse discriminante.

#### *Réponses modales utilisant les formes caractéristiques*

Les spécificités ou formes caractéristiques (cf. chapitre 6) sont, rappelons-le, les formes "anormalement" fréquentes dans une partie du corpus ou dans les réponses d'un groupe d'individus (les formes anti-caractéristiques que l'on a

également appelées spécificités négatives sont au contraire anormalement rares dans ce groupe).

Une *réponse modale* pour une catégorie donnée a été définie comme une réponse d'un individu appartenant à cette catégorie, et contenant des formes qui caractérisent l'ensemble de la catégorie. Ainsi, une réponse contenant beaucoup de formes caractéristiques, et très peu de formes anti-caractéristiques, si elle existe, sera très caractéristique pour la catégorie.

On peut donc associer à toute réponse i du groupe $g$ la somme $s_i(g)$ des valeurs-test relatives aux formes composant la réponse (les formes anti-caractéristiques étant affectées de valeurs-test négatives). Les $m$ réponses modales correspondront aux $m$ individus i associés aux plus grandes valeurs de $s_i(g)$. On peut considérer que $s_i(g)$ définit une *règle de discrimination*, puisque pour un individu donné i, la valeur de $g$ pour laquelle $s_i(g)$ est maximale peut définir le groupe le plus probable d'appartenance de i.

***Les réponses modales d'après un critère de distance***

Rappelons le principe de ces sélections présenté au chapitre 6. Une réponse (ou un document, un texte) est un point-ligne i de **T**, donc un vecteur à $V$ composantes.

On a appris à calculer des distances entre des réponses et les regroupements de ces réponses, puisque les réponses (lignes de **T**) et les regroupements de réponses (colonnes de **C** , ou lignes de **C**', transposée de **C**) sont tous représentés par des vecteurs d'un même espace. Toujours avec les notations définies au chapitre 6, la distance du chi-2 entre un point-ligne i de **T** et un point-colonne g de **C** est donnée par la formule:

$$d^2(i,g) = \sum_{j=1}^{p} \left(\frac{t_{..}}{t_{.j}}\right)\left(\frac{t_{ij}}{t_{i.}} - \frac{c_{jg}}{c_{.g}}\right)^2$$

Alors que la sélection des réponses caractéristiques se traduit par la recherche, pour un groupe $g$ donné, l'individu $i$ qui rend minimale la distance $d^2(i,g)$, la discrimination (une méthode possible de discrimination) consiste à chercher, de façon inverse, pour un individu $i$ donné, la valeur de $g$ qui rend minimale la quantité $d^2(i,g)$.

### 8.4.4 Discrimination régularisée par analyse des correspondances préalable

Les analyses des correspondances peuvent décrire les tableaux **C** qui sont des tables de contingence (dont les "individus" sont des occurrences de formes) ou encore les tableaux clairsemés **T**. Elles permettent de visualiser les associations entre formes et groupes ou modalités.

Elles donnent aussi la possibilité de remplacer des variables qualitatives (simple présence ou fréquence d'une forme) par des variables numériques (les valeurs des coordonnées factorielles), et donc d'appliquer les méthodes classiques d'analyse discriminante (linéaire ou quadratique) après changement de coordonnées.

L'analyse des correspondances peut donc servir de "pont" entre les données textuelles (qualitatives, et parfois clairsemées) et les méthodes usuelles d'analyse discriminante.

Mais surtout, comme cela sera illustré lors de l'exemple du paragraphe 8.5, le fait d'abandonner les derniers axes factoriels opère une sorte de filtrage de l'information (cf. : Wold, 1976 ; Benzécri, 1977) qui renforce le pouvoir de prédiction de l'ensemble de la procédure[1]. Ces propriétés sont utilisées en recherche documentaire (cf. Furnas et al., 1988 ; Deerwester et al., 1990).

C'est en ce sens que l'on parle de régularisation de l'analyse discriminante. Les techniques de régularisations (cf. par exemple Friedman,1989) visent en effet à rendre possible des discriminations dans des cas que les statisticiens décrivent comme "pauvrement posés" (à peine plus d'individus que de variables) ou mal posés (moins d'individus que de variables). De tels cas se présentent lorsque les échantillons d'individus sont petits, ou lorsque le nombre de variables est important, ce qui est souvent la situation de la statistique textuelle. Le fait de travailler sur un petit nombre de facteurs est alors un avantage décisif[2].

Dans les exemples qui vont suivre, on procédera à une analyse des correspondances préalable du tableau **T**, et l'on travaillera sur les $p$ premiers axes factoriels. Le calcul du nombre d'axes principaux à retenir peut être déterminé de façon à optimiser les résultats de la discrimination (cf. Lebart,

---

[1] L'expérience relatée au paragraphe 7.4 du chapitre 7 illustre cette propriété de filtrage de la méthode : la plupart des traits structuraux sont restitués par les premiers axes principaux.

[2] Notons que les distances du chi-2 entre profils utilisées pour calculer les réponses modales deviennent des distances euclidiennes usuelles lorsqu'elles sont calculées à partir des coordonnées factorielles.

1992a), avec les précautions méthodologiques concernant la validité des résultats qui seront évoquées au paragraphe suivant.

### 8.4.5 Validation d'une discrimination

Les nombreuses possibilités de choix laissées aux utilisateurs (choix et élaboration des unités statistiques, choix de la technique de discrimination) ne doivent ni les rendre perplexes ni les décourager. Car la qualité d'une discrimination peut être évaluée avec des critères précis, plus simples que ceux qui interviennent dans le cas des méthodes exploratoires.

Pour établir la validité d'une discrimination, il est nécessaire de procéder à des contrôles sur une partie de l'échantillon (échantillon-test) qui n'a pas servi à établir la règle de discrimination. Cette règle est établie sur la partie de l'échantillon initial que l'on nomme *échantillon d'apprentissage*. Le critère le plus simple est le "pourcentage de bien classés", qui sera calculé à la fois pour l'échantillon d'apprentissage (ce qui donnera toujours une idée trop optimiste de la qualité de la discrimination) et pour l'échantillon-test.

Cette distinction entre sous-échantillons est fondamentale, et finalement assez intuitive si l'on décrit la phase d'apprentissage avec les termes des procédures d'apprentissage en intelligence artificielle.

La procédure de discrimination, en réalisant un maximum du pourcentage de bien classés sur l'échantillon d'apprentissage, va en fait utiliser toutes les particularités de cet échantillon, et donc des informations accidentelles (du bruit) pour arriver à ses fins... Dans certains cas, si le nombre de paramètres est grand, on peut atteindre un pourcentage de 100% d'individus bien classés sur l'échantillon d'apprentissage, avec un pourcentage très faible sur tout autre échantillon : c'est le phénomène de l'*apprentissage par coeur*. La procédure, pourrait-on dire, a appris sans comprendre, c'est-à-dire ici sans avoir saisi les traits structuraux généraux, sans avoir fait la part du structurel et du contingent dans l'échantillon d'apprentissage. On comprend que la vraie valeur d'une analyse discriminante ne se mesure que sur un, ou mieux sur une série d'échantillons test.

McLachlan (1992) fait une revue intéressante des méthodes de calcul d'erreur en analyse discriminante. Des techniques comme le Jackknife et le Bootstrap (cf. Efron (1982), pour une revue de ces deux familles de méthodes) permettent effectivement d'apprécier la confiance que l'on peut accorder à une procédure de discrimination, comme leurs variantes que sont les techniques de *Leave-one-out* (ou de *cross-validation*) (cf. Lachenbruch and Mickey, 1968 ; Stone, 1974 ; Geisser, 1975), qui consistent à construire n échantillons en supprimant à chaque fois un individu, et à appliquer la règle de

discrimination à cet individu (ces dernières techniques très coûteuses ne sont applicables qu'à de très petits échantillons).
Nous appliquerons lors de l'exemple du paragraphe 8.5 qui va suivre une technique de *leave-x%-out*, beaucoup plus économique dans le cas de grands échantillons, qui revient à tirer plusieurs échantillons test représentant chacun un pourcentage donné de l'échantillon total. Ceci nous permettra de représenter la qualité de la prédiction et les erreurs d'affectation par des *matrices de confusion* croisant les catégories initiales (réelles) avec les catégories construites à l'aide des fonctions discriminantes.

## 8.5 Discrimination globale et validation d'après un exemple

Dans quelle mesure peut-on prévoir l'appartenance d'un individu à une catégorie (une catégorie démographique, par exemple), à partir de ses réponses à une question ouverte ? Dans ce qui suit, on montrera comment éprouver le pouvoir de prédiction des profils lexicaux construits à partir des réponses aux questions ouvertes d'une enquête.

Dans l'exemple ci-dessous, les catégories démographiques, au nombre de six, sont obtenues par croisement des modalités de la variable genre (masculin, féminin) avec trois classes d'âge. Ces catégories sont communes à trois enquêtes, réalisées dans trois pays, et dans trois langues différentes pour ce qui concerne les questions ouvertes.

Les matrices de confusion calculées sur échantillons-test caractériseront, pour chacun des trois pays, non seulement la qualité de la prédiction, mais aussi la structure des erreurs de prédiction (quelles sont les catégories bien séparées, lesquelles sont confondues, etc.). Elles permettront donc de procéder à des comparaisons entre les trois pays, malgré le caractère très hétérogène (et par conséquent peu comparable a priori) de l'information de base.

### 8.5.1 L'exemple et le problème

Une enquête internationale[1] a eu lieu en 1989-1990 sur les comportements, habitudes et préférences alimentaires de trois grandes métropoles : Paris, New York et Tokyo (Cf. Akuto, 1992).

Dans chacune de ces trois villes, cette enquête a donné lieu à un millier d'entrevues environ (entretiens face à face).

[1] Cette enquête s'est déroulée à l'initiative de l'Institute of Research on Urban Life (Institut de Recherche Japonais dépendant de la Tokyo Gas Company Ltd) sous la direction du Pr.H. Akuto.

Le questionnaire était composé de nombreuses questions fermées (communes aux trois pays) décrivant de façon détaillée les caractéristiques socio-démographiques des répondants, le rythme, les horaires, les lieux et circonstances des repas ou collations. Le questionnaire comportait en outre deux questions ouvertes, également communes aux trois pays.

Le libellé de ces questions, dans la version française, était le suivant :

- *Quels sont les plats que vous aimez et mangez fréquemment?*
  avec une relance :
  *Y-a-t-il d'autres plats que vous aimez et mangez fréquemment?*

- *Pouvez-vous donner des exemples de ce que vous considérez comme un "repas idéal" ?* avec la relance :
  *Y-a-t-il d'autres choses qui feraient partie de votre "repas idéal" ?*

Dans chacune des enquêtes, on a pu constater que la structure syntaxique de la réponse fournie à cette même question variait fortement d'un répondant à l'autre. Si certaines des réponses sont constituées par simple énumération d'une liste d'items (noms de produits alimentaires ou de spécialités culinaires, en l'occurrence), d'autres se présentent sous forme de phrases relativement construites, motivant les choix retenus par des considérations qui peuvent toucher à la diététique ou même à l'art de vivre.

Pour chacun des "corpus de texte" recueillis dans les trois enquêtes, nous avons pu vérifier que les caractéristiques d'ensemble de la gamme des fréquences du vocabulaire ressemblaient fortement à celles que l'on observe lors du dépouillement des questions ouvertes que nous avons traitées antérieurement (présence de quelques formes très fréquentes, profusion des hapax, etc.).

On trouvera ci-dessous quelques exemples de réponses qui relèvent de chacun des types que nous venons de décrire.

***Réponses de Paris***

— poulet, légumes, viande en sauce / plats surtout équilibrés avec des ingrédients le plus possible naturels.

— boeuf, canard / beaucoup de vitamines moins de matière grasses et d'additifs artificiels.

— sole au vin / repas avec entrées, plats de résistance, dessert, avec des amis dans la bonne humeur.

***Réponses de New York***

— baked potatoes with veal chops, fried shrimp with yellow rice, steaks, lobster with curry rice and potatoes, french onion.
— collard greens, lamb chops, rice and chicken.

— pot roast,mashed potatoes, lamb chops, salads/cucumber and pepper salad, pot roast with mashed potatoes and gravy, apple pie.

***Réponses (romanisées) de Tokyo***

— NIMONO / EIYO NO BARANSU GA TORE, MITAME NI UTSUKUSHII KOTO.
— YASAISUPU / ESA DE NAKU SHOKUJI DE ARU KOTO, KAZOKU SOROTTE SHOKUJI O TORU, ANKA DE ARU KOTO.
— NIZAKANA, SARADA, NABERYORI, CHUKARYORI / ANZEN NA ZAIRYO O TSUKAI BARANSU NO YOI OISHIIMONO O KAZOKU YUJIN TO TANOSHIKU.

[— Pot-au-feu / équilibré au plan de la nutrition, beau à regarder.
— Soupe de légume / pas de nourriture préparée n'importe comment, des choses bon-marché.
— Poisson cuit, salade, bouillons, cuisine chinoise / C'est agréable de manger en famille et avec des amis, de bonnes choses équilibrées à base de produits naturels.]

Dans les premières analyses qui sont présentées ici, les réponses à ces deux questions ont été regroupées, et considérées comme une seule réponse. Ce regroupement permet d'avoir des réponses individuelles relativement riches, au prix d'une certaine perte d'homogénéité. Bien entendu, les trois langues seront traitées séparément. Les réponses en japonais ont été romanisées, pour assurer une compatibilité avec les logiciels disponibles.

Certaines des catégories de répondants, communes aux trois enquêtes, comme le genre (sexe) et l'âge, peuvent être caractérisées à l'intérieur de chaque ensemble par des profils lexicaux que l'on comparera ensuite.

Bien que repérés au départ dans des espaces différents, les points qui représentent ces catégories pourront être caractérisés de façon plus intrinsèque par leurs distances respectives, et par les configurations géométriques qui résument ces distances.

L'analyse discriminante permettra d'évaluer, pour une ville donnée, et donc dans notre cas pour une langue donnée, le pouvoir de prédiction de la réponse textuelle sur ces catégories socio-démographiques. Elle permet ainsi de répondre à la question : dans quelle mesure la réponse ouverte d'un individu nous apporte de l'information sur son genre et son âge ? Cette notion de pouvoir de prédiction peut être mesurée de façon intrinsèque, et donner lieu à des comparaisons entre pays.

Nous étudierons ici une variable composite comprenant six catégories (trois catégories d'âge, croisées avec le genre) qui sera notre "grille de passage" entre les trois enquêtes. Les six classes de cette variable sont les suivantes :

| | | | | | |
|---|---|---|---|---|---|
| 1 | moins de 30 ans | hommes | 4 | de 30 à 50 ans | femmes |
| 2 | moins de 30 ans | femmes | 5 | plus de 50 ans | hommes |
| 3 | de 30 à 50 ans | hommes | 6 | plus de 50 ans | femmes |

En somme, à partir d'individus inconnus et de mots inconnus, (au moins considérés comme tels dans une première phase), on peut espérer observer et comparer des configurations de catégories connues. On peut donc finalement conduire des comparaisons internationales, comme cet exemple va tenter de le montrer, sans qu'il soit toujours nécessaire de procéder à une traduction pendant certaines phases de traitement statistique.[1]

***Sous-échantillon de Tokyo***

Le texte total des 1 008 réponses contient 6 219 occurrences de 832 formes graphiques distinctes.

L'analyse utilisera les 139 formes apparaissant au moins 7 fois dans l'échantillon des 1 008 réponses aux deux questions ouvertes. Ces 139 formes donnent lieu à 4 975 occurrences, soit 80% du texte total.

***Sous-échantillon de Paris***

Le texte total des 1 000 réponses contient maintenant 11 108 occurrences de 1 229 formes distinctes.

Les 112 formes apparaissant au moins 18 fois donnent lieu à 7 806 occurrences, soit 70% du texte total.

***Sous-échantillon de New York***

Le sous-échantillon des personnes interrogées à New York est sensiblement plus réduit que les précédents (634 personnes). Le corpus des réponses libres contient 6 511 occurrences de 638 formes distinctes.

Sur ces 6 511 occurrences, 5 034 (soit 77%) concernent les 83 formes les plus fréquentes (apparaissant plus de 12 fois), qui serviront de base à notre étude.

## 8.5.2 Vocabulaire et analyse pour Tokyo

Seule, l'analyse relative au volet japonais de l'enquête sera exposée ici de façon détaillée. Le lecteur pourra se reporter aux travaux cités plus haut pour l'exposé des résultats correspondant aux deux autres enquêtes.

Le tableau 8.4 nous montre les 139 formes apparaissant au moins 7 fois dans l'échantillon des 1 008 réponses aux deux questions ouvertes.

On note la présence de mots-outil (conjonctions, prépositions comme TO, NO, NI, DE ), de mots étrangers adoptés par la langue japonaise.

[1] Le détail des analyses relatives aux trois pays est donné dans Akuto (1992) et Akuto et Lebart (1992) ; un point de vue critique sur ces analyses est donné dans Benzécri (1992a).

# Tableau 8.4

## Réponses libres du sous-échantillon de Tokyo (texte romanisé) Les 139 formes graphiques les plus fréquentes

| NUM. | MOTS EMPLOYES | FREQ. | NUM. | MOTS EMPLOYES | FREQ. | NUM. | MOTS EMPLOYES | FREQ. |
|---|---|---|---|---|---|---|---|---|
| 1 | AGEMONO | 14 | 48 | KOTO | 128 | 95 | SOBA | 10 |
| 2 | AJI | 21 | 49 | MENRUI | 9 | 96 | SOROTTE | 24 |
| 3 | ANKA | 7 | 50 | MISOSHIRU | 26 | 97 | SOZAI | 20 |
| 4 | ARI | 15 | 51 | MITAME | 14 | 98 | SUKI | 9 |
| 5 | ARU | 40 | 52 | MO | 9 | 99 | SUKINA | 47 |
| 6 | ATTA | 11 | 53 | MONO | 209 | 100 | SUKIYAKI | 50 |
| 7 | BARANSU | 152 | 54 | NABE | 14 | 101 | SUNOMONO | 9 |
| 8 | CHAHAN | 7 | 55 | NABEMONO | 54 | 102 | SUPAGETTEI | 42 |
| 9 | CHIRASHIZUSHI | 7 | 56 | NADO | 7 | 103 | SURU | 24 |
| 10 | CHUUKA | 41 | 57 | NAI | 11 | 104 | SUSHI | 65 |
| 11 | CHUUSHIN | 9 | 58 | NAKU | 7 | 105 | SUTEKI | 13 |
| 12 | DE | 127 | 59 | NANDEMO | 15 | 106 | TABERARERE BA | 8 |
| 13 | DEKIRU | 8 | 60 | NI | 97 | 107 | TABERARERU | 26 |
| 14 | DK | 14 | 61 | NIHON | 44 | 108 | TABERU | 127 |
| 15 | EIYOU | 90 | 62 | NIHONSHOKU | 21 | 109 | TABETAI | 25 |
| 16 | EIYOUKA | 8 | 63 | NIHONSOBA | 7 | 110 | TANOSHII | 18 |
| 17 | FUNIKI | 10 | 64 | NIKU | 36 | 111 | TANOSHIKU | 37 |
| 18 | FURAI | 14 | 65 | NIKURUI | 10 | 112 | TEMPURA | 38 |
| 19 | FURANSU | 10 | 66 | NIMONO | 96 | 113 | TENKABUTSU | 8 |
| 20 | GA | 106 | 67 | NITSUKE | 9 | 114 | TO | 92 |
| 21 | GOHAN | 11 | 68 | NIZAKANA | 20 | 115 | TOKI | 12 |
| 22 | GURATAN | 20 | 69 | NO | 300 | 116 | TOMONI | 12 |
| 23 | GYOUZA | 15 | 70 | O | 223 | 117 | TONKATSU | 22 |
| 24 | HAMBAGA | 7 | 71 | ODEN | 15 | 118 | TORE | 15 |
| 25 | HAMBAGU | 42 | 72 | OHITASHI | 9 | 119 | TORETA | 73 |
| 26 | IROIRONO | 7 | 73 | OISHII | 95 | 120 | TORETEIRU | 8 |
| 27 | ISSHONI | 14 | 74 | OISHIKU | 75 | 121 | TORI | 8 |
| 28 | ITAMEMONO | 14 | 75 | OOI | 7 | 122 | TORIIRETA | 7 |
| 29 | JIBUN | 27 | 76 | OOKU | 16 | 123 | TORU | 23 |
| 30 | JIKAN | 12 | 77 | PIZA | 11 | 124 | TSUKATTA | 8 |
| 31 | KAISEKI | 8 | 78 | RAISU | 15 | 125 | TSUKEMONO | 14 |
| 32 | KAKETE | 8 | 79 | RAMEN | 27 | 126 | UDON | 15 |
| 33 | KAKIFURAI | 7 | 80 | RYOURI | 258 | 127 | WA | 17 |
| 34 | KARAAGE | 11 | 81 | SAKANA | 63 | 128 | WASHOKU | 87 |
| 35 | KARADA | 15 | 82 | SAKANARUI | 14 | 129 | YAKINIKU | 62 |
| 36 | KARE | 79 | 83 | SAKE | 13 | 130 | YAKITORI | 8 |
| 37 | KARORI | 11 | 84 | SANDO | 12 | 131 | YAKIZAKANA | 57 |
| 38 | KATEI | 14 | 85 | SAPPARISHITA | 8 | 132 | YASAI | 95 |
| 39 | KATSU | 7 | 86 | SARADA | 20 | 133 | YASAIITAME | 29 |
| 40 | KATSUDON | 7 | 87 | SASHIMI | 94 | 134 | YASUKU | 10 |
| 41 | KAZOKU | 64 | 88 | SHABUSHABU | 7 | 135 | YOI | 62 |
| 42 | KENKOU | 28 | 89 | SHICHUU | 21 | 136 | YOKU | 42 |
| 43 | KENKOUTEKI | 8 | 90 | SHINSENNA | 11 | 137 | YUKKURI | 17 |
| 44 | KENKOUTEKINA | 7 | 91 | SHOKUHIN | 8 | 138 | YUUJIN | 12 |
| 45 | KI | 11 | 92 | SHOKUJI | 184 | 139 | ZAIRYOU | 16 |
| 46 | KISETSU | 14 | 93 | SHOKUSURU | 7 | | | |
| 47 | KONOMI | 7 | 94 | SHURUI | 10 | | | |

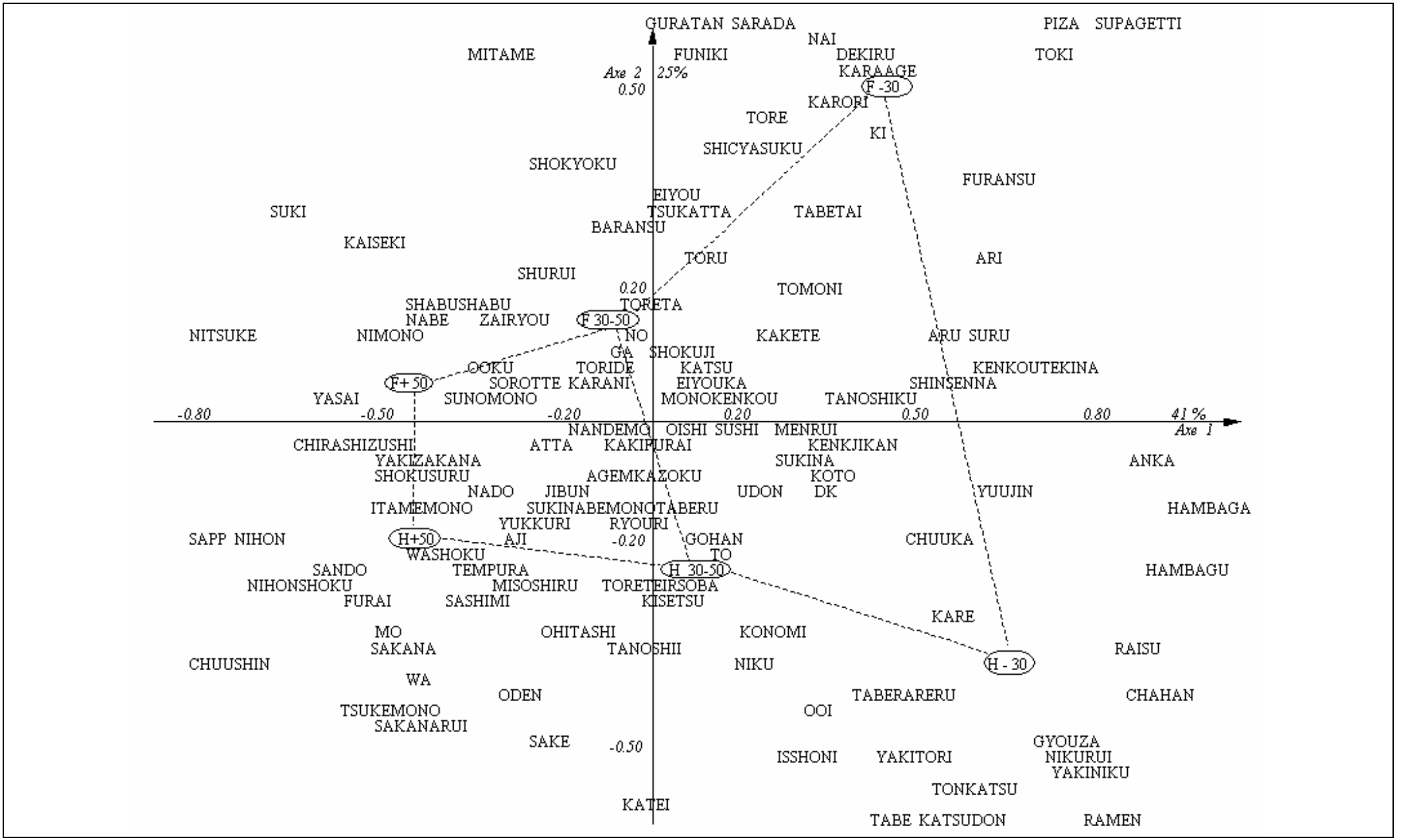


**Figure 8.3 Analyse des correspondances de la table Formes- Catégories (Sexe-Age). Enquête alimentaire, Tokyo.**

Citons, comme exemples de transcriptions : BARANSU (balanced : équilibré), GURATAN (gratin), HAMBAGU, HAMBAGA (variétés de hamburgers), KARE (curry), SUPAGETTEI, SUTEKI, PIZA (spaghetti, steak, pizza), RAISU (riz), SARADA (salade).

Apparaissent également quelques classiques de la cuisine japonaise, bien connus des occidentaux comme les SASHIMI, SUSHI, SUKIYAKI, YAKITORI, TEMPURA. D'autres formes seront commentées plus loin.[1]

***Lecture de la figure 8.3***

La figure 8.3 représente le premier plan factoriel issu de l'analyse des correspondances de la table de contingence à 139 lignes et 6 colonnes obtenues en regroupant les 1 008 profils en 6 classes correspondant à la variable composite à six modalités *âge-genre* évoquée plus haut.

On doit tout d'abord remarquer que les profils lexicaux permettent de reconstituer sans interversion les positions relatives des classes : l'âge augmente de la droite vers la gauche, les hommes occupent le bas du graphique, les femmes la partie supérieure. Le premier axe (41% de l'inertie) est donc un axe qui rend compte de l'âge, le second (25% de l'inertie), un axe qui sépare les hommes et les femmes.

Du côté des jeunes, et surtout des garçons, les réponses mentionnent les hamburgers, de la viande grillée, (YAKI = grillé, NIKU = viande, YAKITORI = poulet grillé), du riz, des nouilles (RAMEN). Pour les jeunes filles ou femmes, on peut noter PIZA, SUPPAGETTEI, FURANSU (français). D'une manière générale, les jeunes sont surtout ceux qui s'intéressent à la cuisine étrangère, les filles étant peut-être encore plus sensibles à la cuisine européenne que les garçons.

Du côté des classes les plus âgées, on trouve NIHON (japonais), et les diverses sortes ou préparations de poissons (SAKANA ou ZAKANA), avec mention de la KAISEKI (cuisine Kyotoïte raffinée).

Si les distances représentées dans ce plan factoriel sont une bonne représentation des distances dans l'espace initial à cinq dimensions (il y a six catégories), on constate de plus une convergence progressive des points relatifs aux deux genres pour une même classe d'âge lorsque l'âge augmente, ce qui donne une forme trapézoïdale à la configuration. Nous reviendrons sur cette constatation plus loin.

---

[1] La romanisation de l'écriture japonaise a introduit une confusion que permettaient de lever les Kanjis chargés de sens, ou même les accents. La forme graphique romanisée SAKE, par exemple, désignera aussi bien dans notre codage le vin de riz que le saumon. Ce type de mutilation de l'information de base n'enlève que peu de chose à la richesse des profils lexicaux agrégés, comme on va pouvoir en juger.

Les explications ne manquent pas à cette observation : traditionalisme, vie commune, et donc homogénéité des modes de vie des personnes plus âgées, ouverture à la fois culturelle (goûts éclectiques) et matérielle (non cohabitation ou cohabitation partielle, fréquentation des restaurants, invitations) des plus jeunes. Une telle ouverture favorise probablement une différentiation des goûts et des habitudes alimentaires.

## Tableau 8.5

**Sélection des formes caractéristiques (Tokyo, classes d'âge extrêmes).**

| | *LIBELLE DE LA FORME GRAPHIQUE* | *POURCENTAGE INTERNE* | *GLOBAL* | | *FREQUENCE INTERNE* | *GLOBALE* | *V.TEST* | *PROBA* |
|---|---|---|---|---|---|---|---|---|
| ***moins de 30 ans - hommes*** | | | | | | | | |
| 1 | YAKINIKU | 3.95 | 1.25 | | 26. | 62. | 5.509 | 0.000 |
| 2 | KARE | 3.95 | 1.59 | | 26. | 79. | 4.427 | 0.000 |
| 3 | HAMBAGU | | 2.58 | 0.84 | | 17. | 42. | 4.268 | 0.000 |
| 4 | CHAHAN | 0.91 | 0.14 | | 6. | 7. | 3.991 | 0.000 |
| 5 | RAMEN | 1.82 | 0.54 | | 12. | 27. | 3.812 | 0.000 |
| 6 | RAISU | | 1.21 | 0.30 | | 8. | 15. | 3.483 | 0.000 |
| ***plus de 50 ans - hommes*** | | | | | | | | |
| 1 | SASHIMI | | 4.00 | 1.89 | | 23. | 94. | 3.421 | 0.000 |
| 2 | SAKANA | | 2.78 | 1.27 | | 16. | 63. | 2.938 | 0.002 |
| 3 | CHUUSHIN | | 0.87 | 0.18 | | 5. | 9. | 2.926 | 0.002 |
| 4 | NIHON | 2.09 | 0.88 | | 12. | 44. | 2.720 | 0.003 | |
| 5 | NABEMONO | | 1.91 | 1.09 | | 11. | 54. | 1.731 | 0.042 |
| 6 | YASAI | 2.96 | 1.91 | | 17. | 95. | 1.720 | 0.043 | |
| | ---------- | | | | | | | | |
| ***moins de 30 ans - femmes*** | | | | | | | | |
| 1 | SUPAGETTEI | | 3.57 | 0.84 | | 24. | 42. | 6.536 | 0.000 |
| 2 | PIZA | 1.04 | 0.22 | | 7. | 11. | 3.595 | 0.000 |
| 3 | SARADA | | 1.34 | 0.40 | | 9. | 20. | 3.239 | 0.001 |
| 4 | GURATAN | | 1.04 | 0.40 | | 7. | 20. | 2.237 | 0.013 |
| 5 | EIYOU | 2.97 | 1.81 | | 20. | 90. | 2.160 | 0.015 | |
| 6 | TOKI | 0.74 | 0.24 | | 5. | 12. | 2.155 | 0.016 | |
| ***plus de 50 ans - femmes*** | | | | | | | | |
| 1 | YASAI | 4.19 | 1.91 | | 38. | 95. | 4.910 | 0.000 |
| 2 | NIMONO | | 3.64 | 1.93 | | 33. | 96. | 3.712 | 0.000 |
| 3 | WASHOKU | | 3.09 | 1.75 | | 28. | 87. | 3.056 | 0.001 |
| 4 | NITSUKE | | 0.66 | 0.18 | | 6. | 9. | 2.905 | 0.002 |
| 5 | NIHON | 1.76 | 0.88 | | 16. | 44. | 2.719 | 0.003 | |
| 6 | JIBUN | 1.21 | 0.54 | | 11. | 27. | 2.559 | 0.005 | |

Reste cependant à vérifier la réalité du fait statistique, manifestée par cette forme trapézoïdale dont on observe peut-être une projection déformée. On va tout d'abord consulter les compléments usuels des analyses des correspondances présentés au chapitre 6 : les formes caractéristiques, qui permettent de chiffrer en termes probabilistes les liaisons formes-catégories,

et les réponses modales, qui permettent de situer ces formes dans leur contexte.

***Lecture du tableau 8.5 (formes caractéristiques)***

Le tableau 8.5 complète la figure 8.3 en faisant apparaître les formes caractéristiques de chaque catégorie. Pour lire ce tableau, prenons par exemple la forme YAKINIKU, qui apparaît caractéristique des hommes de moins de 30 ans.

On lit sur cette table que le pourcentage des occurrences de cette forme chez les hommes de moins de 30 ans (3,95%) est plus de trois fois supérieur à son pourcentage global (1,25%). L'avant dernière colonne donne à cet écart une valeur-test de 5,51. La dernière colonne nous fournit d'ailleurs les premiers chiffres de ce seuil.[1]

On observe que certaines formes, comme GURATAN, caractérisent les femmes des deux premières classes d'âge. En revanche, une forme comme YASAI (légumes) caractérise les personnes âgées des deux sexes, tout en étant beaucoup plus caractéristique des femmes.

***Lecture du tableau 8.6 (réponses modales)***

Le tableau 8.6 nous donne quelques réponses caractéristiques pour chacune des catégories étudiées. Rappelons que les réponses caractéristiques d'une classe sont des réponses originales sélectionnées comme étant les plus proches du centre de cette classe, (au sens de la distance du chi-2 entre le profil lexical de la réponse et le profil lexical moyen de la classe).[2]

Les réponses caractéristiques sont utiles car elles plongent les formes dans leur contexte réel, et remettent en mémoire la complexité de l'information de base. Ainsi, la forme SUPAGETTEI, très caractéristique des jeunes femmes, se rencontre parfois chez leurs homologues masculins. Si la réponse 3 est quand même typique des garçons, c'est qu'elle est lestée par ailleurs de formes très caractéristiques des hommes jeunes.

---

[1] Rappelons que la valeur-test convertit, pour plus de lisibilité, une probabilité critique en variable normale standardisée : la valeur 1.96 correspond ainsi au seuil 0.025, alors que la valeur 5.51 correspond à une probabilité de l'ordre de $10^{-6}$.

[2] Un autre mode de calcul consiste à caractériser une réponse par la valeur-test moyenne des formes qu'elle contient, mais ce critère qui favorise trop les réponses lapidaires, n'a pas été utilisé ici (cf. chapitre 6, paragraphe 6.2).

## Tableau 8.6

**Sélection des réponses caractéristiques**
**(Echantillon de Tokyo, classes d'âges extrêmes)**

***moins de 30 ans - hommes***

-- 1 YAKINIKU, KARE RAISU, RAMEN
-- 1 OISHII TO OMOTTE TABERARERU KOTO

-- 2 RAMEN,YAKINIKU,KARE
-- 2 OISHIKUTE RYOU NO ARU KOTO

-- 3 KARE RAISU, HAMBAGU, SUPAGETTEI
-- 3 NIKURUI, ABURAPPOI RYOURI, RAMEN NO YOUNA MONO

***plus de 50 ans - hommes***

-- 1 NABEMONO, MAGURO NO SASHIMI
-- 1 YASAI TO SAKANA GA ISSHONI NATTA RYOURI

-- 2 WASHOKU
-- 2 SAKANA, YASAI,TOUFU, NIMONO, JUNNIHONTEKINA MONO

-- 3 NABEMONO, TONKATSU, YAKIZAKANA
-- 3 YASAI SAKANA CHUUSHIN NO WASHOKU

***moins de 30 ans - femmes***

-- 1 KARE, GURATAN, PIZA,UNAGI, SUPAGETTEI
-- 1 EIYOU NO BARANSU TORETA SHOKUJI

-- 2 GURATAN, PIZA, SUPAGETTEI, CHIRASHI, KARE, HAMBAGA,
-- 2 SARADA, OISHIKU EIYOU GA ARI MAINICHI HENKA NI
-- 2 TONDA KONDATE NO SHOKUJI

-- 3 SUPAGETTEI
-- 3 BARANSU TORETA SHOKUJI

***plus de 50 ans - femmes***

-- 1 YASAI NO NIMONO
-- 1 WASHOKU

-- 2 YASAI NO NIMONO, SASHIMI, NIZAKANA, YAKIZAKANA
-- 2 JIBUN NI ATTA SUKINA MONO O TABERU KOTO

-- 3 YASAI NO NIMONO, KORUDOBIFU, SUTEKI
-- 3 JIBUN DE RYOURISHITA MONO O OISHIKU AJIWAU KOTO GA
-- 3 DEKIREBA RISOU DE ARU

### 8.5.3 Réalité des configurations

Les analyses opérées sur Paris et New York donnent des *patterns* analogues à ceux observés pour la Figure 8.3. Pour Tokyo et New York, on observe un éloignement marqué des deux catégories les plus jeunes : les garçons et les filles de moins de 30 ans ont, semble-t-il des réponses très différentes. On n'observe pas cette divergence pour Paris (Figure 8.4).

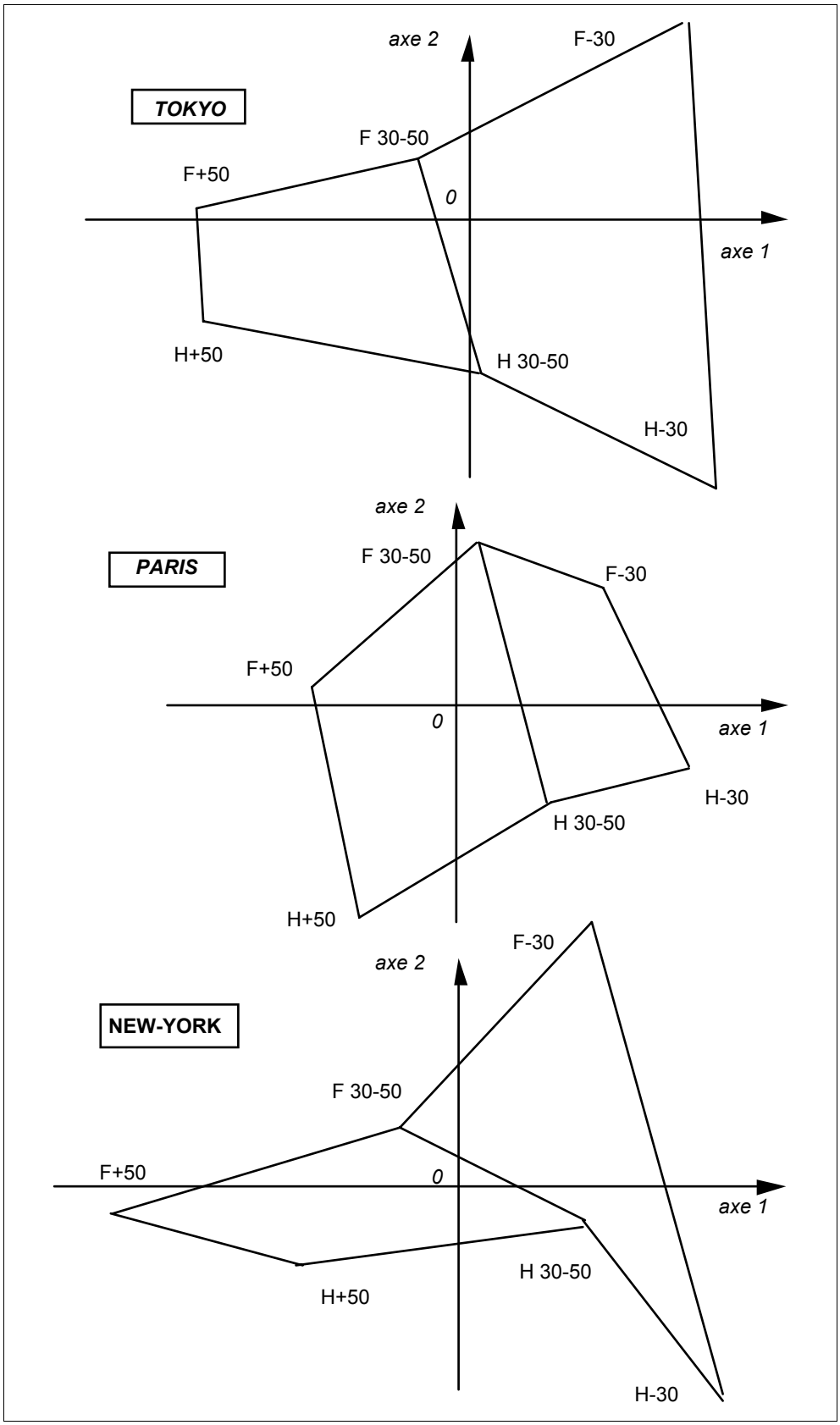


**Figure 8.4.**

**Comparaison des patterns (genre - âge) des plans (1,2) pour les trois villes**

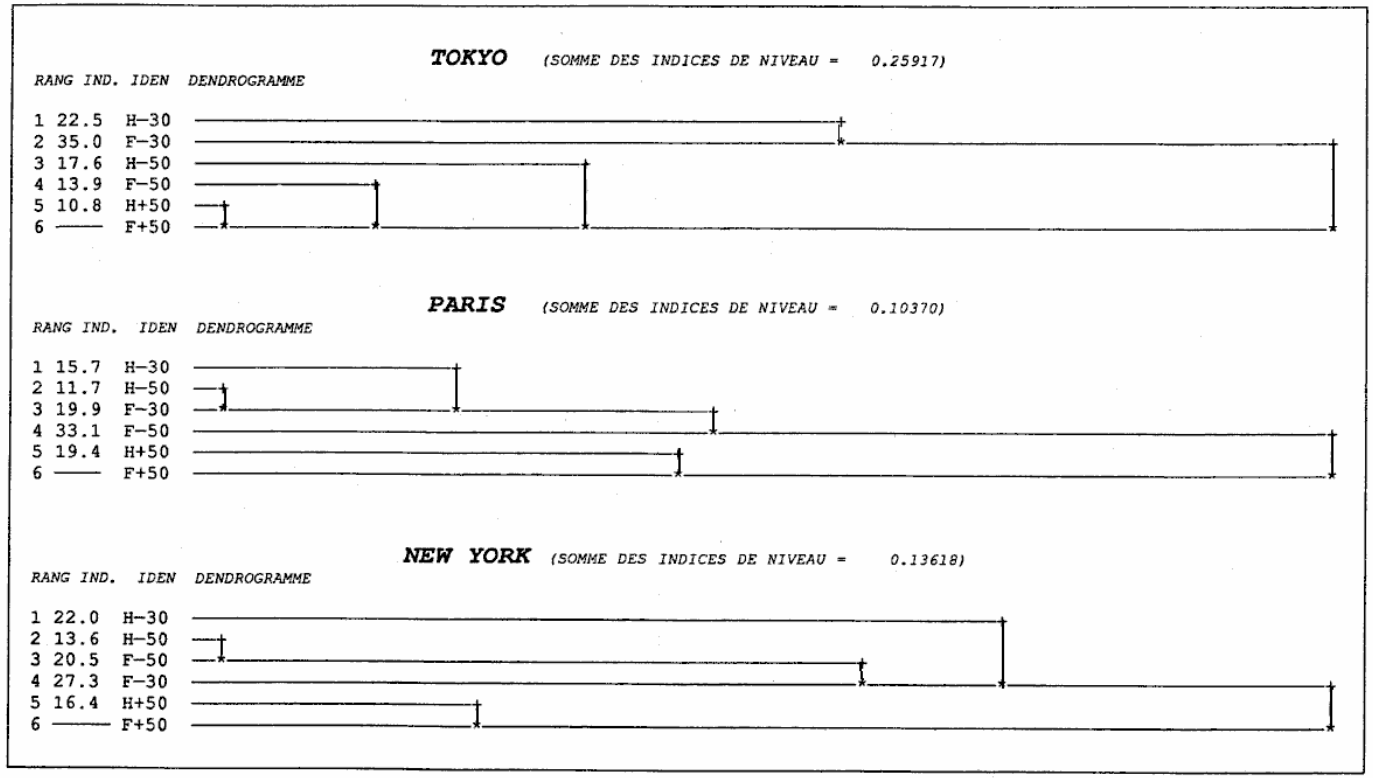


**Figure 8.5**

**Comparaison des trois dendrogrammes (Tokyo, Paris, New York)**
**(Indices exprimés en pourcentages de la somme des indices)**

La Figure 8.4 schématise, en effet, les trois "patterns" observés : notons que dans les trois cas, les âges s'échelonnent sur l'axe 1, et les genres se séparent le long de l'axe 2.

Quelle est donc la réalité des configurations observées ? Quelles sont les limites des interprétations entre différences de configurations ?

Une des façons de répondre à ces questions est de décrire par classification hiérarchique (sur les 5 coordonnées factorielles, en utilisant le critère dit de Ward, cf. chapitre 4) les distances entre points catégories. La Figure 8.5 nous montre ainsi les trois dendrogrammes de Tokyo, Paris et New York. Les indices correspondant à chaque noeud sont exprimés en pourcentages de la somme des indices.

On note par exemple que, pour Tokyo, la *convergence entre les genres à mesure que l'âge augmente* constatée sur l'analyse factorielle (figure 8.3), est bien visible sur la figure 8.5 : les plus âgés s'agrègent pour un niveau bas de l'indice (11% de la somme), et les plus jeunes à un niveau plus élevé (22.5%).

C'est effectivement à Tokyo que les personnes les plus âgées ont les comportements les plus homogènes (niveau 11%, contre 16% à New York et 19% à Paris) Une autre particularité de Tokyo concerne le rattachement tardif des jeunes au reste de la population : la coupure la plus nette est à 30 ans dans cette ville, alors qu'elle est autour de 50 ans dans les deux autres.

A Paris et New York, le rattachement des personnes les plus âgées au reste de la population se fait au plus haut niveau du dendrogramme.

On remarque que les femmes d'une classe d'âge ne se rattachent jamais sur ces dendrogrammes à des hommes d'une classe d'âge inférieure :

- A Tokyo, le point F-50 rejoint les points H+50 et F+50 avant que ne le fasse le point H-50.
- A Paris, le point F-30 s'agrège très bas au point H-50.
- A New York, le point F-30 s'agrège à la classe -50 (H et F) avant de s'agréger au point H-30.

On peut en somme conclure que les préférences alimentaires des femmes, dans les trois pays, les rapprochent plutôt des hommes d'une classe d'âge supérieure (du point de vue du contenu des réponses, à partir des formes caractéristiques et des réponses modales : moindre intérêt pour les viandes, intérêt plus marqué pour les légumes et le poisson). La classification, fondée ici sur des distances dans l'espace à cinq dimensions, a également confirmé que les positions relatives des classes (figure 8.4), notamment les divergences

à l'intérieur des classes de jeunes pour New York et Tokyo, n'étaient pas des artefacts liés à la projection.

### 8.5.4 Analyse discriminante et matrices de confusion

Les unités choisies relèvent des options a) et e) de la section 8.4.2 : La discrimination se fait dans l'espace des formes graphiques seules après sélection suivant un critère de fréquence, et après régularisation par analyse des correspondances préalable.

***Régularisation par les axes principaux***

On illustrera l'effet de filtrage des sous-espaces issus de l'analyse des correspondances par la figure 8.6, qui représente les taux de succès (donc de bien classés) d'une série d'analyses discriminantes en fonction du nombre d'axes factoriels (ou principaux) retenus (nombre porté sur l'axe des abscisses de la figure 8.6) et aussi en fonction de l'importance relative des échantillons test.

Cette discrimination concerne le sous-échantillon de New York, les variables de prédiction sont les 83 formes graphiques, et la variable à prédire est l'âge en trois classes.

- Première constatation : Pour les échantillons d'apprentissage (symboles noirs), les taux de succès augmentent continûment avec le nombre d'axes.

- Seconde constatation, pour les échantillons test (symboles blancs) qui mesurent le vrai pouvoir discriminant en situation réelle, on observe une saturation du taux de succès aux alentours de 40 axes. Ainsi, à partir d'une certaine dimension, les axes supplémentaires n'améliorent plus la vraie discrimination, mais améliorent quand même les taux de succès (illusoire, parce qu'optimiste) de l'échantillon d'apprentissage. C'est le phénomène qualifié en Intelligence Artificielle et en réseaux neuronaux d'*apprentissage par coeur* , qui fonctionne à l'intérieur de l'apprentissage, mais qui ne permet pas d'inférences extérieures.

- Troisième constatation : Le taux de succès sur l'échantillon test est meilleur si celui-ci ne constitue qu'un petit prélèvement (toutefois, la différence est petite entre les taux de 30% (symboles carrés) et 10% (symboles losanges). Ce dernier point est intuitif, car l'apprentissage ne peut qu'être meilleur si l'échantillon est important.

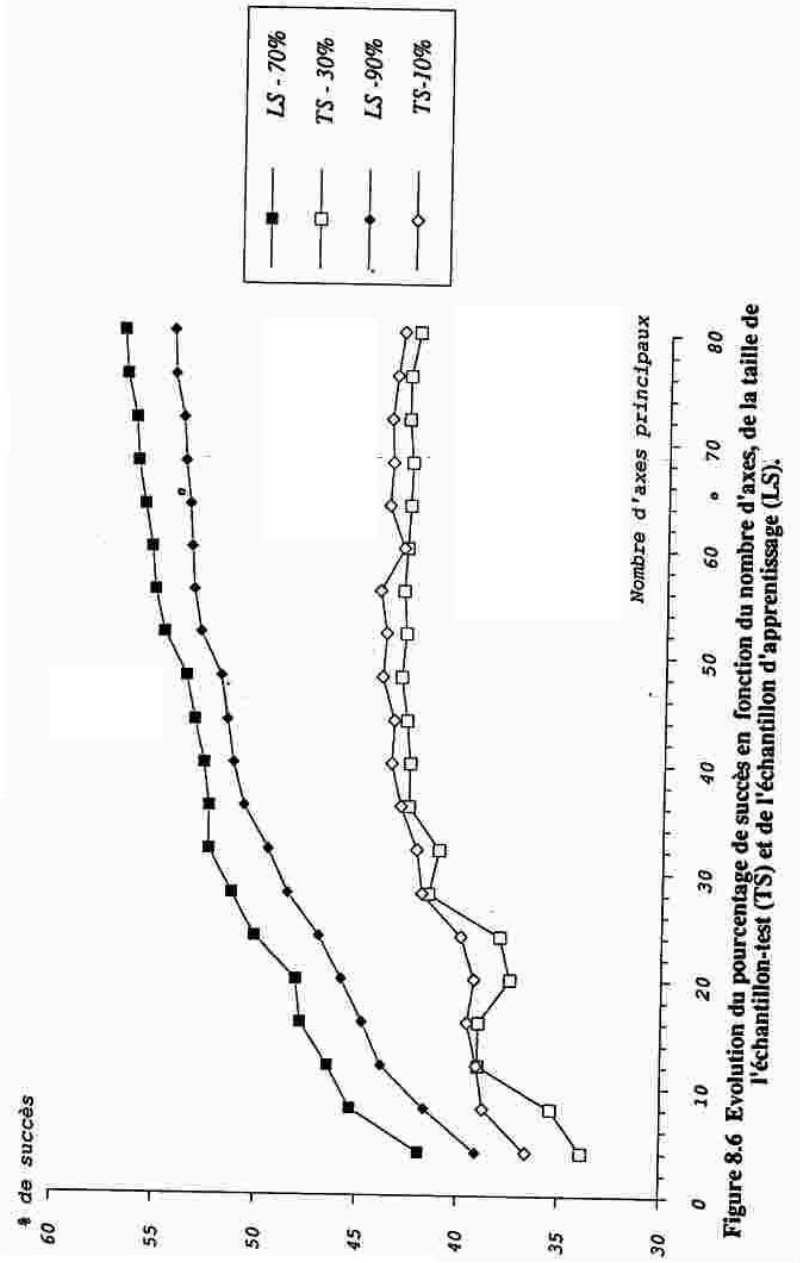


**Figure 8.6 Evolution du pourcentage de succès en fonction du nombre d'axes, de la taille de l'échantillon-test (TS) et de l'échantillon d'apprentissage (LS).**

En revanche, si l'échantillon d'apprentissage est réduit, les taux de succès sur cet échantillon sont alors trop optimistes, et donc trompeurs (cas des carrés noirs).

Ces trois constatations nous ont conduit à travailler sur un nombre d'axes réduits (36 pour l'exemple de Tokyo), avec un prélèvement d'échantillon- test de 10%.

**Tableau 8.7**

**Exemples de Matrices de confusion (Tokyo)**
**Nombre d'axes principaux : 36.**
**(Echantillon-test : 10% de chacun des 30 ech.)**

**a - Echantillon d'APPRENTISSAGE**
*Catégories initiales [lignes] et catégories calculées[colonnes]*

| | H-30 | H-50 | H+50 | F-30 | F-50 | F+50 | Taux de succès |
|---|---|---|---|---|---|---|---|
| **H-30** | 1469 | 541 | 244 | 506 | 694 | 388 | 38.24 % |
| **H-50** | 596 | 1457 | 620 | 676 | 1535 | 808 | 25.60 % |
| **H+50** | 93 | 361 | 999 | 229 | 1097 | 833 | 27.66 % |
| **F-30** | 377 | 222 | 76 | 1342 | 835 | 316 | 42.36 % |
| **F-50** | 361 | 538 | 486 | 1146 | 2610 | 1023 | 42.34 % |
| **F+50** | 268 | 492 | 635 | 391 | 1195 | 1841 | 38.18 % |

**Taux de succès moyen : 35.60 %**

**b - Echantillon TEST**
*Catégories initiales [lignes] et catégories calculées[colonnes]*

| | H-30 | H-50 | H+50 | F-30 | F-50 | F+50 | Taux de succès |
|---|---|---|---|---|---|---|---|
| **H-30** | 125 | 71 | 20 | 59 | 95 | 48 | 29.90 % |
| **H-50** | 61 | 117 | 79 | 71 | 175 | 105 | 19.24 % |
| **H+50** | 16 | 51 | 79 | 37 | 115 | 110 | 19.36 % |
| **F-30** | 51 | 31 | 6 | 118 | 98 | 38 | 34.50 % |
| **F-50** | 54 | 67 | 60 | 143 | 237 | 115 | 35.06 % |
| **F+50** | 26 | 63 | 74 | 49 | 125 | 151 | 30.94 % |

**Taux de succès moyen : 28.13 %**

### *Les matrices de confusion*

Trente échantillons-test d'environ 100 individus ont été tirés au hasard sans remise dans l'ensemble des 1 008 individus.

Pour chacun de ces 30 tirages, les 6 centres de gravité (points moyens) des catégories *âge-genre* sont calculés dans l'échantillon d'apprentissage (sur les 1 008 - 100 = 908 individus) dans l'espace à 36 dimensions des 36 premiers facteurs de l'analyse des correspondances du tableau **T** (1 008 lignes, 139 colonnes).

Le tableau 8.7 présente la matrice de confusion moyenne relative à l'échantillon de Tokyo. Les individus des échantillons d'apprentissage (tableau 8.7-a), puis ceux de l'échantillon-test (tableau 8.7-b) sont alloués aux centres les plus proches.

La description des matrices de confusion relatives aux trois pays par analyse des correspondances permet de procéder à une comparaison de structures situées dans des espaces différents (Lebart, 1992b).

Le tableau 8.8 nous montre les taux de succès calculés de façons analogues pour les trois pays. On voit ainsi que la discrimination est bien meilleure à Tokyo que dans les autres villes, ce qui n'était pas visible sur les arbres hiérarchiques. En effet, les échelles des indices d'agrégation ne sont pas comparables car au départ ni les vocabulaires utilisés, ni les échantillons n'ont la même taille.

## Tableau 8.8

**Comparaison des taux de succès pour les trois villes**

| | Ech. d'apprentissage | | | Ech. Test | | |
|---|---|---|---|---|---|---|
| | *TOKYO* | *PARIS* | *NEW YORK* | *TOKYO* | *PARIS* | *NEW YORK* |
| **H-30** | 38.24 | 20.77 | 43.45 | 29.90 | 14.44 | 33.00 |
| **H-50** | 25.60 | 13.16 | 26.23 | 19.24 | 6.89 | 20.05 |
| **H+50** | 27.66 | 31.50 | 17.08 | 19.36 | 23.13 | 10.14 |
| **F-30** | 42.36 | 29.45 | 38.00 | 34.50 | 24.59 | 30.48 |
| **F-50** | 42.34 | 29.55 | 15.78 | 35.06 | 25.28 | 12.17 |
| **F+50** | 38.18 | 30.26 | 42.00 | 30.94 | 29.41 | 37.11 |
| **Total** | 35.60 | 26.23 | 28.60 | 28.13 | 21.39 | 22.64 |

La figure 8.7 donne une visualisation des points lignes des deux matrices de confusion (échantillon d'apprentissage actif en traits pleins, échantillon-test supplémentaire en traits pointillés) relatives à l'échantillon de Tokyo.

La structure correspondant à l'échantillon-test est contractée (donc moins forte) mais reproduit fidèlement celle de l'échantillon d'apprentissage.

C'est une confirmation de la réalité de la structure observée sur la figure 8.3, avec cet élément d'information supplémentaire : la *confusion* semble très importante entre les hommes et les femmes de plus de 50 ans.

Ce que l'on interprète dans les termes suivants : rien, dans leurs réponses à la question ouverte, ne permet de distinguer ces deux catégories, alors que la distinction entre hommes et femmes était aisée pour les deux autres classes d'âge.

Une matrice de confusion sur échantillon-test représente bien le pouvoir prédicteur réel du texte sur les catégories. Etant donné que cette représentation a un caractère intrinsèque, elle peut permettre des comparaisons entre pays différents et entre réponses exprimées dans des langues différentes.

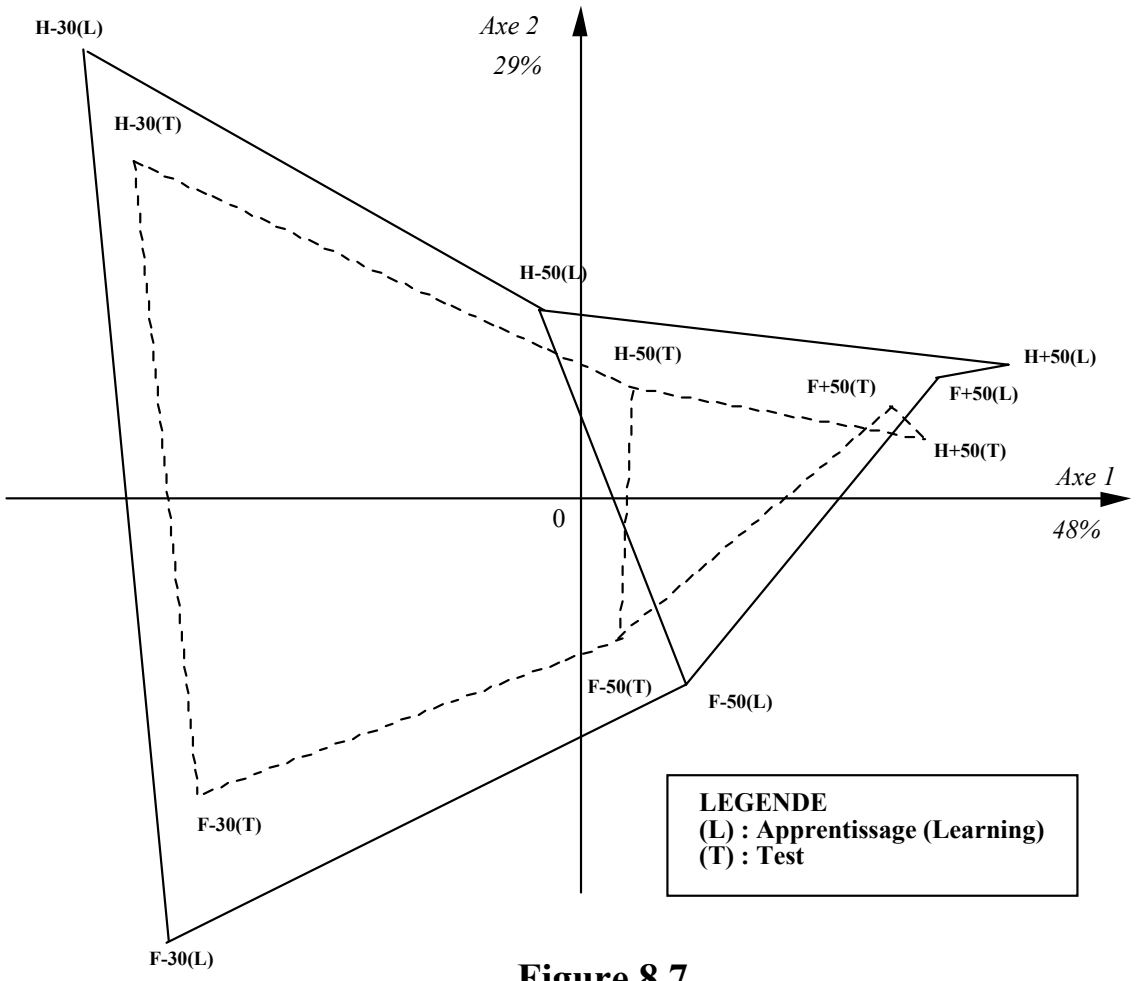


**Figure 8.7**

**Analyse des correspondances des deux matrices de confusion relatives à l'échantillon de Tokyo.**

On peut en effet comparer ces matrices de confusion, et donc les pouvoirs prédicteurs de différents textes sur une même variable. C'est dans cet esprit que la figure 8.8, qui n'est complexe qu'en apparence, représente la synthèse des trois matrices de confusion (sur échantillons-tests) relatives aux trois villes, obtenues en soumettant à l'analyse des correspondances leur juxtaposition en ligne. Cette juxtaposition comprend trois matrices analogues de la matrice de confusion du tableau 8.7b qui ne concernait que Tokyo (il s'agit d'échantillons-test, c'est à dire de pouvoir de prédiction réel). Le tableau décrit a donc 18 lignes (six pour chaque ville) et six colonnes.

***Lecture de la figure 8.8***

La grille en traits gras joignant les points encadrés représente les *catégories affectées* (colonnes de la juxtaposition de matrices) alors que l'on distingue les trois grilles de *catégories réelles* : New York en bas, Paris en pointillé au milieu, et Tokyo en haut (lignes de la juxtaposition de matrices). Remarquons que pour chacune des grilles, les jeunes sont à droite, les plus âgés à gauche, les femmes vers le haut, et les hommes vers le bas.

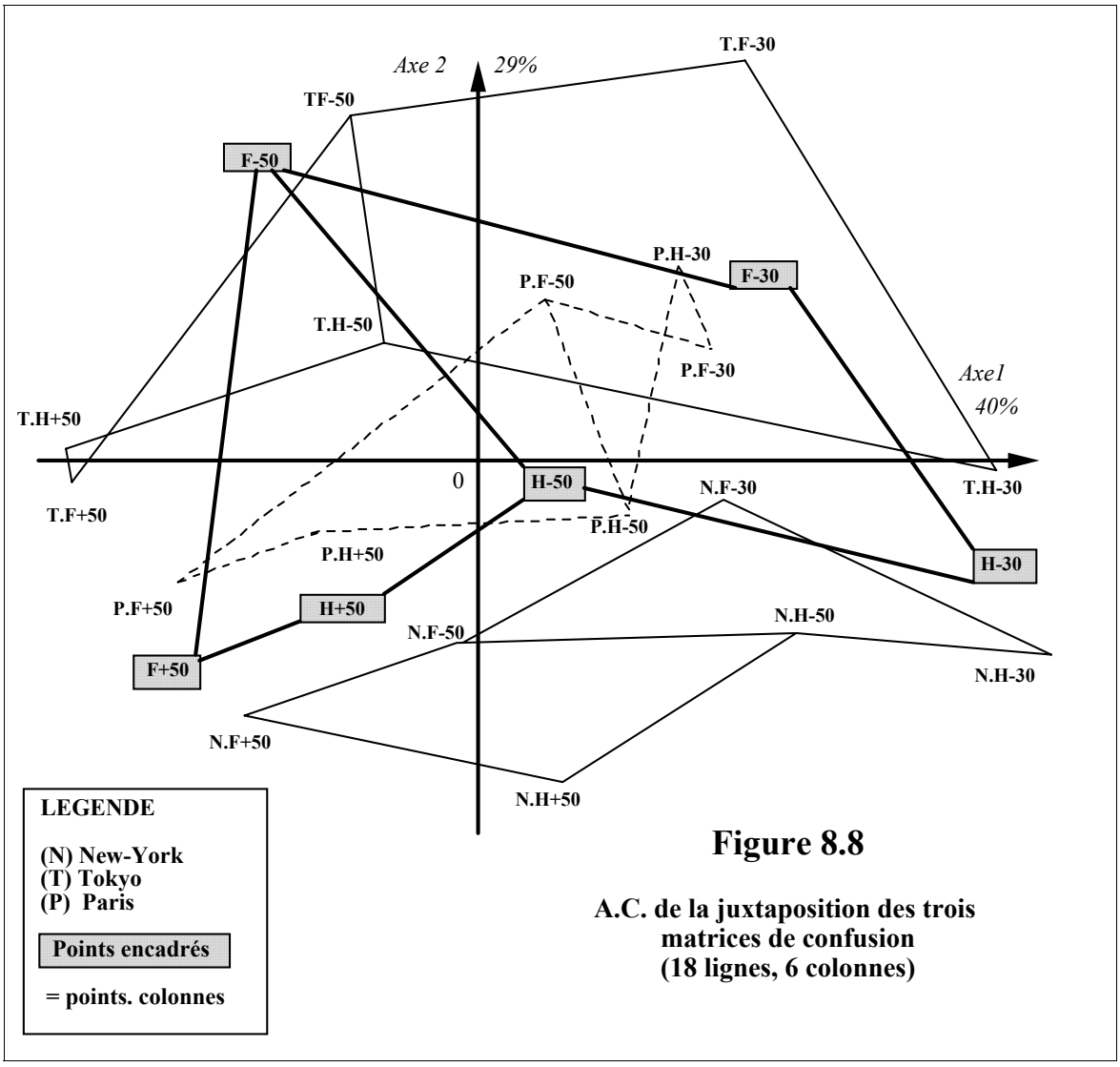


**Figure 8.8**

**A.C. de la juxtaposition des trois matrices de confusion (18 lignes, 6 colonnes)**

En regardant la position d'un point-catégorie réel par rapport à la grille en traits gras, on a une idée de la façon dont l'analyse discriminante répartit les individus correspondants. Ainsi, à gauche, les trois points F+50 correspondant aux trois villes (N.F+50, P.F+50, T.F+50) sont plus proches du point encadré F+50 que des autres points encadrés, ce qui veut dire que la catégorie des femmes de plus de 50 ans est bien reconnue à partir des réponses libres sur l'alimentation dans les trois villes (et les trois langues). De

même, à droite, les jeunes hommes de New York et Tokyo sont bien identifiés par leurs discours (points T.H-30, N.H-30), cela est moins vrai pour Paris (P.H-30), pour lequel les jeunes hommes sont beaucoup moins différenciés des jeunes femmes et de leurs aînés par leurs réponses libres. La grille en pointillé est en effet recroquevillée dans sa partie droite.

Le premier axe, lié à l'âge, ne permet pas de distinguer les trois villes (avec l'exception déjà signalée des jeunes parisiens des deux sexes).

En revanche, le second axe, lié au genre, sépare très nettement Tokyo et New York. Ce sont essentiellement les femmes de 30-50 ans (points F-50) et les femmes de moins de 30 ans (points F-30) qui sont responsables de cet état de choses. Les femmes de la catégorie d'âge intermédiaire de Tokyo sont en effet beaucoup mieux identifiées à partir de leurs réponses libres que celles de Paris, et surtout que celles de New York. Notons que ce résultat était déjà visible sur le tableau 8.8 donnant les pourcentages d'individus bien-classés par ville et par catégories.

### 8.5.5 Conclusions du paragraphe 8.5

La possibilité de comparer (assez facilement, au moins dans une première approche) des comportements à partir de textes dans des langues différentes ouvre évidemment des directions de recherche intéressantes.

Les phases de traduction et de codage de l'ensemble des réponses n'ont pas été nécessaires pour procéder à des descriptions et des inductions confirmées par des procédures de validation puissantes.

Mais une autre caractéristique de cet exemple est que, dans cette première phase exploratoire, l'on n'aurait su que faire des traductions, car beaucoup des variables de bases exprimant des habitudes alimentaires spécifiques n'ont pas d'équivalent d'un pays à l'autre.

Bien entendu, en montrant l'étendue des possibilités d'un traitement aussi formel de l'information de base, on pense mieux préparer les phases ultérieures de travail, et non les remplacer.