Explorer l'espace des mots : du linéaire au non-linéaire

Ludovic Lebart

CNRS – ENST, 46 rue Barrault, 75013, Paris.

Abstract

To visualize the associations between words within a series of texts, it is frequent to use principal axes techniques (of which latent semantic analysis is a particular case) sometimes complemented with clustering techniques. The Kohonen self organizing maps (SOM) can be viewed as a non-linear visualisation tool that performs a sort of compromise between a high-dimensional set of clusters and the 2-dimensional plane generated by principal axes techniques. We show in this paper how both linear and non-linear approaches can be used simultaneously, allowing for a reciprocal assessment of both visualizations. The example of application concerns the statistical processing of responses to open-ended questions in the context of an international survey.

Résumé

Pour représenter les associations entre mots à l'intérieur de textes d'un même corpus, il est fréquent d'utiliser des méthodes en axes principaux (analyses en composantes principales ou des correspondances, toutes deux dérivées de la décomposition aux valeurs singulières, dont le *latent semantic indexing* est un cas particulier), éventuellement complétées par des classifications automatiques. Les cartes de Kohonen constituent une sorte de compromis non linéaire entre les analyses en axes principaux et la classification. Nous essayons de montrer dans ce papier comment les analyses linéaires et non-linéaires peuvent être utilisées simultanément avec profit, et se valider mutuellement. L'application concerne les réponses à une question ouverte dans une enquête internationale.

Mots-clés : Lexicométrie, Visualisation, Classification, Contiguïté, *Self Organizing maps (SOM)*.

1. Introduction

Il existe de nombreux outils de visualisation de données ayant chacun leurs avantages et leurs inconvénients. Ils ont en commun de se matérialiser par une représentation bidimensionnelle (écran ou feuille de papier). Avec les possibilités accrues d'interactivité, de couleur, d'animation, les implémentations logicielles peuvent améliorer à la fois la puissance et l'ergonomie de ces outils. Une représentation tridimensionnelle en perspective animée est bien sûr fondamentale dans des domaines comme l'industrie, la médecine, l'architecture, mais elle ne résout pas tous les problèmes lorsque l'on travaille, comme cela est le cas dans le domaine textuel, dans des espaces ayant plusieurs dizaines, voire plusieurs centaines de dimensions.

Actuellement, les méthodes les plus utilisées pour visualiser les tables lexicales du type : mots x textes (mot désigne ici aussi bien une forme de surface qu'un lemme, un segment ou une locution) sont les méthodes en axes principaux toutes dérivées de la *décomposition aux valeurs singulières* : analyse des correspondances (AC), Latent semantic indexing (LSI), analyse en composantes principales (ACP). Ces méthodes peuvent être qualifiées de linéaire dans ce contexte, parce qu'elles projettent les points sur des droites ou des plans.

Dans plusieurs logiciels, ces méthodes sont complétées par des procédures de classification automatique (classification autour de centres mobiles ou k-means, classification hiérarchique,

classification mixte combinant les deux approches). L'usage conjoint de ces deux techniques préconisé dès les années soixante du siècle dernier (travaux relatés dans Benzécri, 1981) est enrichissant, mais n'est pas toujours facile en pratique, car les représentations obtenues sont complexes.

On dispose dans ce cas de deux points de vue sur les données : des représentations sur une série de plans factoriels, des regroupements en classes ou en hiérarchies de classes. Les plans factoriels décrivent les traits les plus saillants en termes de forme globale du nuage de points. Les classes, calculées à partir de distances dans tout l'espace de départ, décrivent aussi la texture du nuage ; elles sont plus sensibles aux variations locales de densité. On choisit souvent de projeter les centres de classes dans les plans factoriels, pour faire un pont entre la description des classes (à partir de leurs éléments les plus caractéristiques) et les visualisations. Ceci a l'avantage de décrire les positions relatives des classes dans l'espace.

Une autre option consiste à élaborer une méthode intermédiaire qui construit des classes en imposant à celles-ci des contraintes de proximités pouvant conduire à une représentation plane unique. C'est le principe des cartes auto-organisées de Kohonen, dont nous allons étudier les relations avec les méthodes linéaires plus classiques.

2. Les cartes auto-organisées

Les cartes cartes auto-organisées de Kohonen (1989) cherchent à représenter dans un espace à deux (rarement trois) dimensions les lignes ou les colonnes d'un tableau en respectant la notion de voisinage dans l'espace des éléments à classer.

Le principe est de considérer une carte comme une grille rectangulaire (parfois hexagonale) aux mailles déformables, laquelle, une fois dépliée épouse au mieux les formes du nuage de points. Les nœuds de la grille sont les *neurones* de la carte. Chaque point du nuage est projeté sur le nœud dont il est le plus proche. De fait, chaque point, décrit initialement dans un espace multidimensionnel est représenté à la fin par deux coordonnées donnant la position du *neurone* sur la carte : l'espace est réduit. Tous les points affectés à un même *neurone* sont proches dans l'espace initial. Ils décrivent et regroupent des individus semblables.

On définit *a priori* une notion de voisinage entre classes et les observations voisines dans l'espace des variables de dimension q appartiennent après classement à la même classe ou à des classes voisines. Ces voisinages peuvent être choisis de diverses manières mais en général on les suppose directement contigus sur la grille rectangulaire (ce qui représente alors 4 ou 8 voisins pour un *neurone*).

L'algorithme d'apprentissage pour classer m points est itératif. L'initialisation consiste à associer à chaque classe k un centre provisoire $\mathbf{C}_k$ à p composantes choisi de manière aléatoire dans l'espace à p dimensions contenant les m mots à classer. À chaque étape on choisit un mot i au hasard que l'on compare à tous les centres provisoires et l'on affecte le mot au centre $\mathbf{C}_k$ le plus proche au sens d'une distance donnée *a priori*. On rapproche alors du mot i le centre $\mathbf{C}_k$ et les centres voisins sur la carte ce qui s'exprime à l'étape t par :

$$\mathbf{C}_k(t+1)=\mathbf{C}_k(t)+\varepsilon(i(t+1)-\mathbf{C}_k(t))$$

où i(t+1) est le mot présenté à l'étape t+1, ε un paramètre d'adaptation positif et inférieur à 1. Cette expression n'intervient que pour le centre $\mathbf{C}_k$ et ses voisins.

Cet algorithme est analogue à celui des centres mobiles (ou k-means, ou nuées dynamiques), mais dans ce dernier cas, il n'existe pas de notion de voisinage entre classes et on ne modifie à chaque étape que la position du centre $\mathbf{C}_k$.

3. Visualisation factorielle versus carte auto-organisée

On va comparer dans ce paragraphe les deux types de visualisations (analyse des correspondances d'une part, carte auto-organisée de l'autre) pour une table lexicale croisant 113 mots et 9 catégories.

3.1. Contexte de l'application

Une question ouverte a été posée lors d'une enquête multinationale (cf. Hayashi *et al*., 1992) avec le libellé suivant "*Que signifie pour vous la culture de votre pays?*". C'est le sous-échantillon relatif à la France qui est traité ici (effectif : 1009 personnes représentatives des personnes de 15 ans ou plus). L'ensemble des réponses à cette question représente 14742 occurrences de 2248 mots (formes graphiques) distincts. Sont traités ici les 153 mots apparaissant au moins 12 fois, qui représentent 10648 occurrences. On a retranché à cette liste 40 mots-outils (opération jugée parfois contestable, mais qui ne modifiera en rien nos conclusions, et permettra d'alléger les graphiques) ce qui porte à 113 le nombre définitif de mots gardés. Les réponses sont ici regroupées en neuf catégories obtenues par croisement de l'âge en trois classes (moins de 30 ans, 30-55 ans, plus de 55 ans) et du niveau d'éducation en trois classes (bas, moyen, haut).

3.2. Visualisation simultanée mots-catégories par analyse des correspondances

Il s'agit d'une représentation classique, (axes 1 et 2 expliquant respectivement 25.6 % et 16.8 % de la variance totale), encore assez lisible parce que le nombre de mots est relativement limité, et parce que de petits déplacements ont été autorisés pour éviter les points doubles.

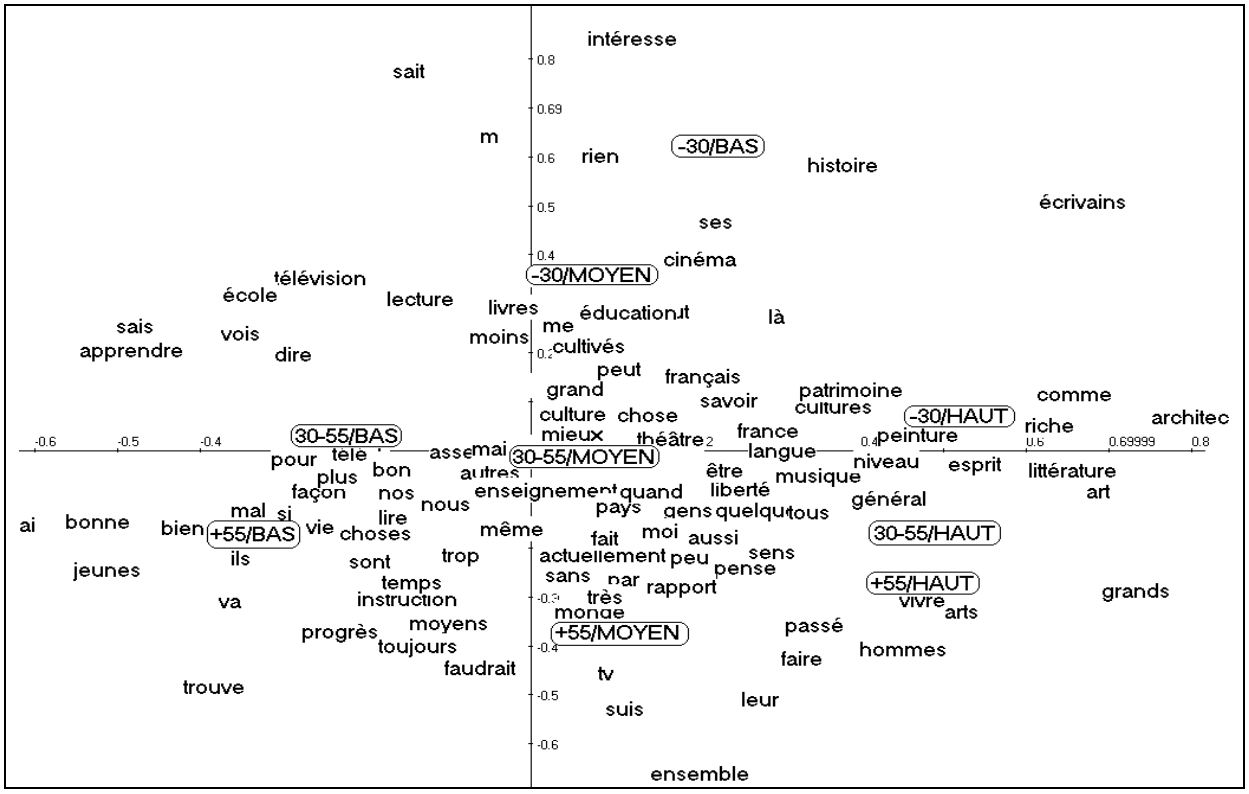

Figure 1. *Plan (1, 2) de l'analyse des correspondances de la table lexicale croisant les 9 catégories de répondants avec les 113 mots apparaissant plus de 12 fois.*

3.3. Visualisation simultanée mots-catégories par cartes de Kohonen

La carte de Kohonen utilise ici le même système de distance que celui de l'analyse des correspondances de la table lexicale qui croise les occurrences des 113 mots les plus fréquents et les neuf catégories précédentes (les identificateurs des catégories sont abrégés). On obtient ainsi (figure 2) une synthèse assez comparable à celle de l'analyse des correspondances. Cette synthèse nous permet de repérer assez rapidement des cooccurrences de mots dans une même catégorie, et des proximités lexicales entre catégories.

13 école va nos même moins là lire enseignement	14 temps sont si monde jeunes instruction ils choses bonne bien ai +55/BAS	15 vie trouve faut faudrait	16 tv très trop toujours suis sans ont leur ensemble +55/MOYEN
notre moyens grand dire bon beaucoup autres 30-55/MOYEN	9 10 plus nous mieux enfants	11 rapport quand par moi française	12 être sens savoir quelque pense gens faire
vois télévision télé surtout sais pour pas mal apprendre	5 6 éducation me français culture chose assez	théâtre peu pays passé musique fait connaissance aussi	7 8 tous peinture niveau arts 30-55/HAUT
sait rien progrès peut mais livres lecture intéresse façon cinéma 30-55/BAS -30/MOYEN	1 2 ses patrimoine m liberté général cultivés actuellement -30/BAS +55/HAUT	3 vivre langue histoire france	4 écrivains riche littérature hommes grands esprit cultures comme art architecture -30/HAUT

***Figure 2**. Carte auto-organisée faite à partir des 8 axes (mots + catégories) de l'analyse des correspondances de la section 3.2 – dont les deux premiers axes sont représentés sur la figure 1.*

Le format d'édition de la figure 2, composée de fenêtres rectangulaires, est particulièrement apte à recevoir des listes de mots, sans superposition ni empiètement. S'agissant d'une classification (avec les contraintes, pour les classes, de figurer dans une grille), les regroupements prennent en compte des informations qui vont au-delà de celles données par les deux dimensions d'un premier plan factoriel.

Parmi les inconvénients de cette technique de visualisation, citons le caractère « non déterministe » (comme les méthodes de classification autour de centres mobiles ou *k-means* usuelles) des algorithmes qui produisent, dans ce cas particulier, des cartes différentes en fonction de l'initialisation aléatoire de départ. Enfin, il n'existe pas vraiment, dans les implémentations disponibles, de méthodes de validation permettant de décrire la confiance à accorder à la fois à la composition et aux positions relatives des différentes cases de la grille.

La seule validation possible, commune avec les méthodes de classification du type *k-means* ou nuées dynamiques, consiste à opérer divers démarrages aléatoires de l'algorithme, et à observer visuellement la stabilité des résultats. De simples clics de souris génèrent instantanément de nouvelles cartes, qui peuvent différer (légèrement, peut-on espérer) à la fois par les regroupements à l'intérieur des cellules et par les positions relatives des cellules.

3.4. Dissonances entre les deux représentations (figures 1 et 2)

Commençons par noter les signes de cohérence les plus évidents : la classe 4 de la figure 2 caractérisée par les jeunes instruits (-30 / HAUT, 90 répondants) et la classe 14 de cette même figure caractérisée par les personnes plus âgées peu instruites. (+55 / BAS, 150 répondants). Ces deux classes sont caractérisées par des mots voisins dans les deux représentations (*architecture, écrivains, riche, littérature, grands, hommes,* pour la classe 4 - *bonne, jeunes, ils, trouve, instruction*… pour la classe 14). Les deux catégories concernées sont éloignées sur les deux figures.

Mais des incohérences notables sont observées. Citons par exemple la catégorie (-30 / BAS, 13 répondants) qui est en haut du graphique sur la figure 1, donc fort éloignée de la catégorie (+55 / HAUT, 31 répondants) qui est située en bas à droite de ce plan factoriel. Ces deux catégories éloignées sur la figure 1 se trouvent cependant dans une même case (la classe 2) sur la figure 2. Ce sont bien sûr des catégories rendues instables par le faible nombre de répondants, surtout pour la première catégorie citée.

La présence des trois catégories ayant un haut niveau d'éducation sur la droite du premier axe de la figure 1 témoigne de la puissance des axes principaux et de leur capacité de filtrage de l'information. La disposition des catégories semble plus cohérente dans le plan factoriel de la figure 1 que sur la grille de la figure 2.

4. Représentation des 16 classes dans le plan factoriel

On peut représenter, sur le plan, factoriel de la figure 1 chacune des 16 classes de la figure 2 par leur centre : chaque centre est le point moyen des éléments (mots ou catégories) qui appartiennent à la classe. On peut ensuite représenter la grille elle-même en joignant par une arête les centres de classes adjacentes sur la figure 2. Comme cela était prévisible par suite des incohérences observées, la grille est partiellement déformée, repliée (figure 3).

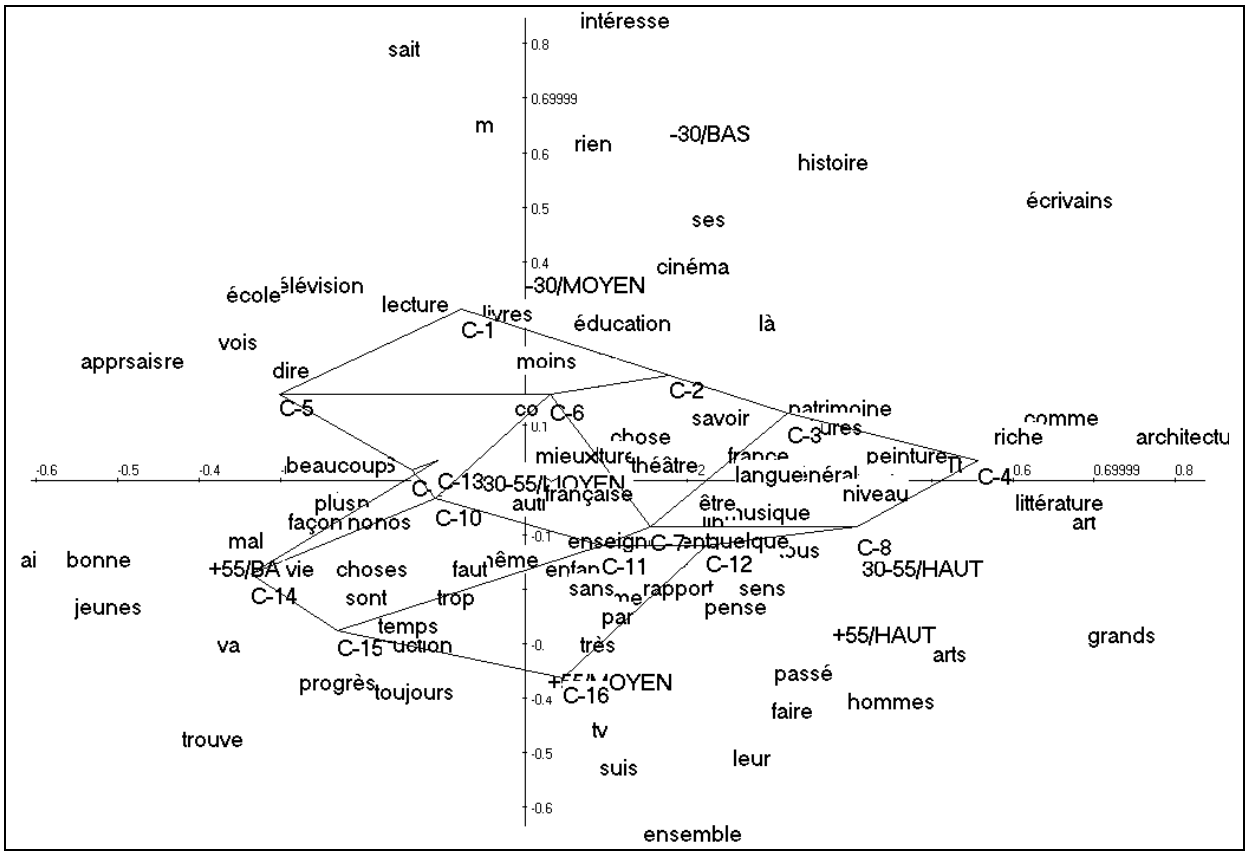

***Figure 3**. Carte auto-organisée de la figure 2 projetée dans le plan de la figure 1.*

La carte est partiellement repliée parce que les distances entre centres de classes ne sont pas compatibles avec les distances dans le premier plan factoriel.

On remarque cependant qu'il existe une bonne compatibilité entre les deux représentations pour les huit premières classes de la carte auto-organisée, notées sur la figure 3 de C-1 à C-8, malgré le changement d'orientation de la carte (le bas de la figure 2 correspond au haut de la figure 3).

5. Sous-espace associé à une grille de Kohonen

La question à laquelle nous allons tenter de répondre maintenant est la suivante : peut-on trouver un plan sur lequel la grille de Kohonen se projette sans repliement ni déformation excessive ?

5.1. Dépliage de la carte par analyse de contiguïté

Cela est parfois possible, et la réponse est alors donnée par l'analyse de contiguïté (cf. Lebart, 2000). Lorsque des observations statistiques multidimensionnelles sont associées à un graphe, ce type d'analyse permet de définir les variances et covariances locales qui prennent en compte la dépendance des observations vis-à-vis de la structure de graphe. L'analyse de contiguïté permet alors de reconstituer autant qu'il est possible la structure de graphe dans un plan (si le graphe est associé à une partition, l'analyse de contiguïté coïncide avec l'analyse linéaire discriminante de Fisher). Appliquée au graphe induit par une carte de Kohonen (deux éléments sont joints par une arête s'ils appartiennent à une même classe ou à deux cases adjacentes de la carte), l'analyse de contiguïté permet ainsi de trouver, dans

l'espace à p dimensions de départ (p = 9 pour notre exemple), un sous-espace à 2 dimensions (un plan) qui respecte au mieux la forme de la grille (Lebart, 2005).

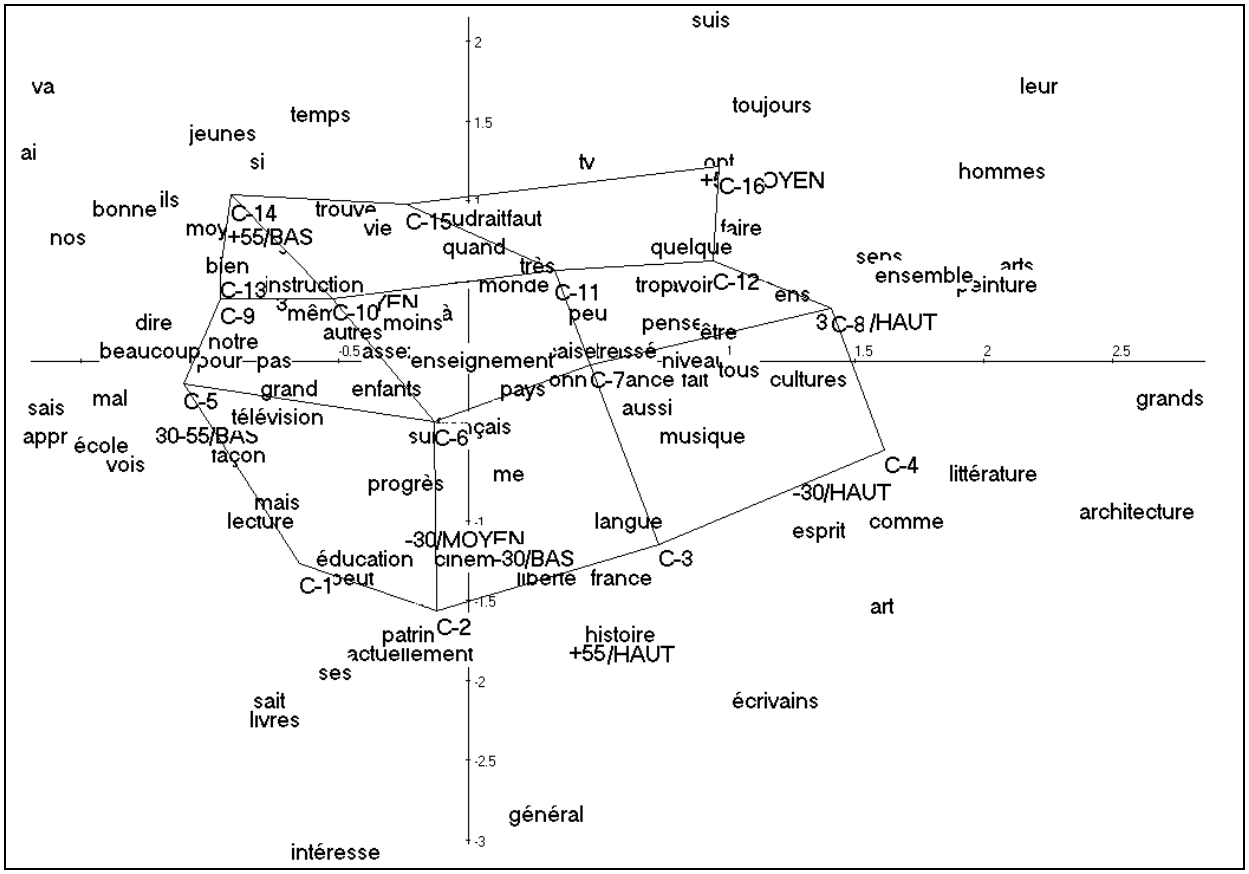

Figure 4. *La même carte auto-organisée projetée sur le sous-espace associé à la grille, qui est le plan lui permettant de se déployer au mieux. (Analyse de contiguïté minimisant la variance locale sur la carte de Kohonen).*

La figure 4 représente un plan qui n'est plus un plan factoriel, mais un plan sur lequel se projette la carte de Kohonen de façon optimale (ce plan est optimal au sens de la variance locale, qui est une variance calculée seulement à l'intérieur des classes et entre classes contiguës). Au vu des mots occupant des positions extrêmes, on voit que l'axe horizontal ressemble beaucoup au premier facteur de l'analyse des correspondances (cf. figure 1).

5.2 Validité de la représentation

Le fait de disposer d'un espace obtenu par projection à partir des données initiales permet de nuancer la position des points à l'intérieur des 16 classes, et aussi d'apprécier la forme des classes en traçant, par exemple, sur la figure 4, les enveloppes convexes des 16 classes.

On peut également valider par la technique de *bootstrap* (Efron et Tibshirani, 1993) la position des points, car l'opérateur projection sur ce sous-espace permet de projeter les réplications du tableau de départ, et d'en déduire des zones de confiance (*bootstrap partiel*, simple projection a posteriori) pour la position des points.

La figure 5 représente de cette façon les zones de confiance des six catégories de niveau d'éducation extrêmes (HAUT ou BAS). On constate ainsi un empiètement des zones de

confiance des deux catégories (-30 / BAS) et (+55 / HAUT) qui étaient séparées sur la figure 1 et très voisines sur la figure 2.

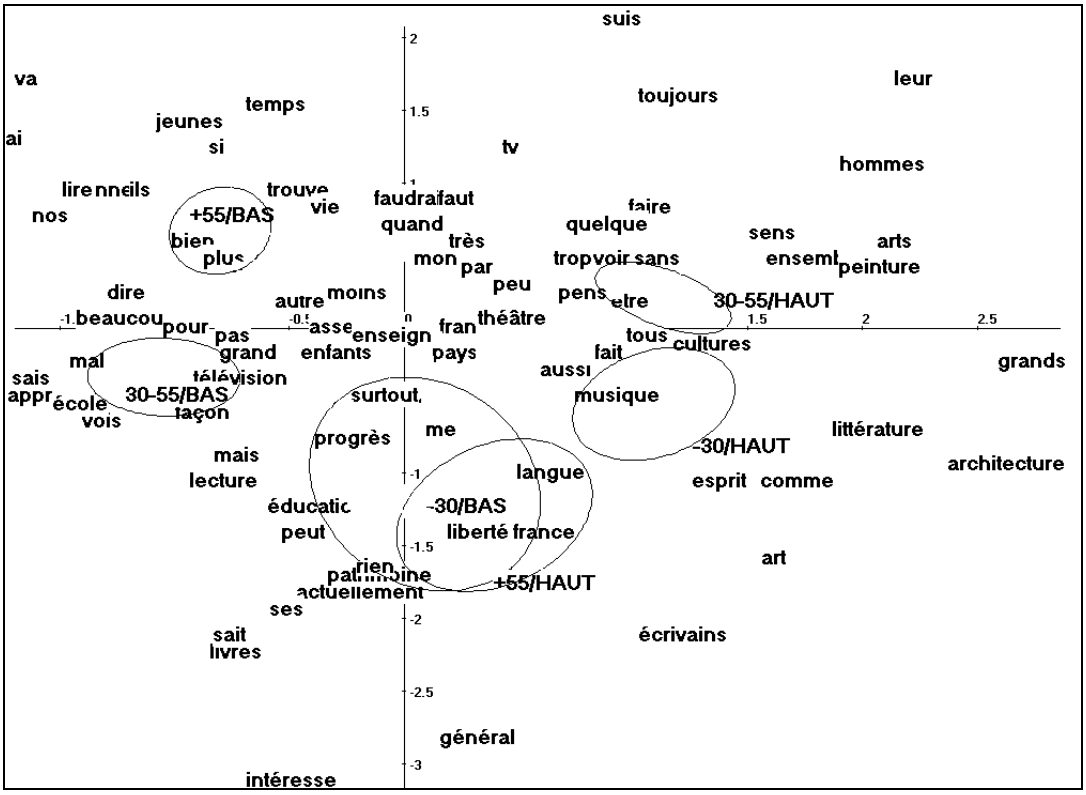

Figure 5. *Zones de confiance bootstrap pour 6 catégories de répondants dans le sous-espace associé à la grille de Kohonen (même plan que la figure 4) : les 3 catégories de niveau d'éducation BAS, les 3 catégories de niveau d'éducation HAUT.*

[Toutes les procédures de calcul et de tracé graphique sont implémentées dans le logiciel académique DTM qui peut être librement téléchargé à partir du site www.lebart.org].

Références

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

Efron B., Tibshirani R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

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

Kohonen T. (1989). *Self-Organization and Associative Memory*. Springer Verlag, Berlin.

Lebart L., Salem A., Berry E. (1998). *Exploring Textual Data*. Kluwer Ac. Publisher, Dordrecht.

Lebart, L. (2000). Contiguity Analysis and Classification. In Gaul W., Opitz O. and Schader M. (Eds), *Data Analysis*, Springer, Berlin : 233-244.

Lebart L. (2005). Visualization of textual data : Unfolding the Kohonen map. In *Applied Stochastic Models and Data analysis*, ENST-Bretagne, Juin 2005, Brest, France. Texte intégral : http://asmda2005.enst-bretagne.fr/IMG/pdf/proceedings/73.pdf