Arbres ou Axes pour les Analyses de Textes ?

Ludovic Lebart[1]

[1]Mines-Télécom-ParisTech, Paris – France

Abstract

Both principal axes and trees are common visualization tools in the field of textual data analysis. But these methods (trees and axes) are on the one hand complementary with respect to the obtained results and on the other often synergistic in the technical phases of their implementation. To prove, illustrate and discuss these properties, we limit ourselves to dealing with Correspondence Analysis, and as regards tree descriptions, with Minimum Spanning Trees, Hierarchical Clustering (Ward criterion) and Additive Trees. Three examples corresponding to three distinct leading cases are briefly discussed.

Résumé

Les représentations en axes principaux et par arbres constituent des outils de visualisation fréquemment utilisés dans le champ de l'analyse des données textuelles. Or ces méthodes (arbres et axes) sont d'une part complémentaires quant à la nature des résultats obtenus, et d'autre part souvent synergiques lors des phases techniques de leur mise en oeuvre. Pour le prouver et l'illustrer, on utilise, en ce qui concerne les axes principaux, l'analyse des correspondances, et pour les descriptions par arbre, l'arbre de longueur minimale (*Minimum spanning tree*), la classification hiérarchique (critère de Ward) et l'analyse arborée (*Additive trees*). Trois exemples de situations typiques différentes sont présentés à l'appui de cette complémentarité.

Key words : Additive trees, Hierarchical Clustering, Minimum Spanning Tree, Correspondence Analysis.

1. Introduction

On qualifie souvent d'instruments d'observation les outils de visualisation statistiques, une dénomination qui induit une certaine neutralité vis-à-vis de la réalité observée. Mais chaque instrument a ses avantages, ses faiblesses, et surtout ses vocations, c'est-à-dire ses aptitudes à révéler ou décrire certaines structures privilégiées. L'opposition entre les instruments que sont les axes et arbres évoque les oppositions entre continu et discret, entre géométrique et algorithmique. Il existe entre les analyses en axes principaux et les arbres de classification une incontestable complémentarité (dans l'interprétation). Il existe même des relations théoriques, dans certains cas (Cazes, 1984 ; Lebart et Mirkin, 1993), mais celles-ci ne seront pas abordées ici. Après des rappels sommaires sur ces deux familles de méthodes (section 2 et 3), on s'attachera à montrer en section 4 comment se modulent leurs contributions respectives à propos de trois applications volontairement choisies dans des contextes très différents.

2. Les axes principaux

Les méthodes d'analyses en axes principaux (ou encore méthodes factorielles) ont un noyau théorique commun qui est la Décomposition aux Valeurs Singulières (*Singular Value Decomposition* ou SVD). La SVD se diversifie en s'adaptant au contexte : elle devient l'Analyse en composantes principales pour les variables numériques (les mesures), l'analyse des correspondances (CA) pour les tables de contingence (et donc aussi les tableaux lexicaux), l'analyse des correspondances multiples pour les variables nominales. La SVD n'est pas un modèle, mais un théorème d'algèbre linéaire, et aussi un outil de compression de données. En analyse des données textuelles, on utilise surtout la variante CA, principalement

comme outil de visualisation (plans factoriels (1, 2), (3, 4) etc.). On ne présentera pas ces méthodes supposées connues du lecteur. On rappelle seulement ici qu'il existe des méthodes de validation des visualisations par ré-échantillonnage : ce sont les méthodes de validation par *bootstrap*, devant être déclinées en fonction des unités statistiques (formes, lemmes, phrases, réponses, paragraphes, chapitres…). Enfin il existe des relations algébriques simples entre les lignes et les colonnes des tableaux analysés : les relations de transitions qui permettent des représentations simultanées de ces lignes et ces colonnes (souvent de mots et de textes) et qui autorisent la projection d'éléments supplémentaires sur les graphiques, et, ce faisant, permettent de tester certaines hypothèses.

3. Les méthodes de classification

On évoquera la classification hiérarchique (critère de Ward et critère du saut minimal), l'arbre de longueur minimale, et les arbres additifs (analyse arborée). Les partitions, elles, jouent un grand rôle en Analyse des Données et en data mining. Elles sont pourtant moins utilisées dans le domaine textuel, sauf dans les recueils comportant de nombreuses unités (analyses de questions ouvertes, webmining).

Les techniques de classification font appel à une démarche algorithmique et non à des calculs formalisés. Alors que les axes principaux (ACP, AC, ACM) sont la solution d'une équation pouvant s'écrire sous une forme très condensée (même si sa résolution est complexe), la définition des classes ne se fera qu'à partir d'une formulation algorithmique: une série d'opérations définie de façon récursive et répétitive. Il en découle que la mise en œuvre de la plupart des techniques de classification ne nécessite que des notions mathématiques assez élémentaires. Mais on perd la transparence analytique qui est en faveur de l'analyse géométrique des données : axes, plans, projections, espace, barycentres, éléments projetés *a posteriori* (éléments supplémentaires), etc.

Il existe plusieurs familles d'algorithmes de classification : les *méthodes de partitionnement* comme les méthodes d'agrégation autour de centres mobiles; les méthodes hiérarchiques qui comprennent les *algorithmes ascendants* (ou encore agglomératifs) qui procèdent à la construction des classes par agglomérations successives des objets deux à deux, et qui fournissent une hiérarchie de partitions des objets; enfin les *algorithmes descendants* (ou encore divisifs) qui procèdent par dichotomies successives de l'ensemble des objets, et qui peuvent encore fournir une hiérarchie de partitions. Parmi ces derniers, la méthode divisive proposée par Reinert (1983) occupe une position hybride intéressante, puisque la division à chaque étape de la classification se fait en utilisant un axe principal (premier axe d'une AC).

La carte de Kohonen (1984) (ou carte auto-organisée) fournit une représentation non linéaire : la classification selon une grille peut dans certains cas s'adapter à des formes de nuage de points dans un espace de dimension élevée. Disons en bref que les cartes auto-organisées réalisent un autre compromis intéressant entre classification et visualisation.

3.1. L'arbre de longueur minimale (ALM)

Il s'agit de l'arbre (graphe connexe sans cycle) dont les sommets sont les objets à classer eux-mêmes, et dont la longueur (somme des longueurs des arêtes, c'est-à-dire des distances entre paires d'objets) est minimale. Une référence classique est Kruskal (1956), mais l'algorithme développé par Florek *et al.* (1953), est plus ancien et beaucoup plus performant. Cet algorithme est à l'origine de la *Wroclaw taxonomy*, école polonaise pionnière d'analyse des données, antérieure à l'apparition des ordinateurs. Cf. aussi : Graham et Hell (1985) pour une

histoire (passionnante) de cet algorithme, par ailleurs très utilisé en recherche opérationnelle (comment optimiser le câblage d'une ville ?).

A la première étape, on joint chaque sommet à son voisin le plus proche. Cela revient à prendre la plus petite distance dans chaque ligne du tableau des distances. Cette opération rapide produit directement une forêt F_1 (famille d'arbres, c'est-à-dire simplement : graphe sans cycle). A l'étape *k*, chaque arbre de la forêt F_{k-1} (chaque composante connexe du graphe) est joint à son plus proche voisin en prenant comme distance entre arbres la plus petite distance entre un sommet quelconque de l'un et un sommet quelconque de l'autre. Le processus s'arrête dès que le graphe F_k est connexe. Cet algorithme peut même être mis en œuvre manuellement sur des tableaux de distances assez grands (d'où la *Wroclaw taxonomy* précitée).

3.1.1. Equivalence entre ALM et agrégation suivant le saut minimal

Cette équivalence a été démontrée par Gower et Ross (1969). On peut représenter simultanément la hiérarchie et l'arbre de longueur minimale en perspective lorsque les points sont peu nombreux. Une représentation qualifiée de « squelette arborescent » a été proposée par Benzécri et Jambu, (1976) pour organiser les éléments terminaux de tout arbre hiérarchique. Pour le praticien de l'analyse factorielle, il sera souvent intéressant de porter l'ALM sur les plans factoriels de façon à remédier, dans une certaine mesure, aux possibles déformations imputables à l'opération de projection (*cf.* figures 2 et 4 ci-dessous).

3.2 ***Les arbres additifs (AA) : l'explosion phylogénétique***

Ces arbres ont été proposés à l'origine par Buneman (1971), puis étudiés par Sattath et Tverski (1977). Le concept de hiérarchie à la base de la classification ascendante revenait à approximer les distances initiales par une distance *ultramétrique*, qui vérifie, en plus des axiomes classiques de toute distance, pour tout triplet *(x, y, z)*, l'inégalité :

$$d(x, y) <= Max\ (d(x, z), d(y, z)\ ).$$

Les arbres additifs sont moins exigeants, bien que cela ne soit pas évident *a priori*, en demandant seulement, pour tout quadruplet *(x, y, z, t)*, que soit vérifiée l'inégalité :

$$d(x, y) + d(z, t) <= Max\ (\ \{d(x, z) + d(y,t)\},\ \{d(x, t) + d(y, z)\}\ )$$

Avec une telle distance, un arbre peut être dessiné avec les objets comme éléments terminaux (ou feuilles), tel que la distance entre deux objets soit la longueur du chemin joignant ces deux objets sur l'arbre. Plus souple que l'ALM qui dépend de *n-1* paramètre, l'AA implique *2n – 3* paramètres. Il reste à trouver une approximation des distances initiales qui satisfasse ces conditions.

Stimulées par les travaux de Barthélémy et Guénoche (1988) et de Luong (1988), les méthodes d'analyse arborées ont été très utilisées en France dans le champ des analyses de texte. Toutefois, les premiers algorithmes proposés demandaient un volume de calcul prohibitif pour des nombres d'objets à classer importants.

Saitou et Nei (1987) ont proposé un algorithme intitulé *Neighbor Joining* qui permet de ramener approximativement la recherche de l'arbre additif à une procédure de classification ascendante classique. Cette heuristique, améliorée par la suite par Gascuel (2000) puis par Bryant (2005), puis implémentée par Huson et Bryant (2006), a eu d'immenses répercussions dans l'univers en pleine expansion de la recherche phylogénétique, domaine de prédilection des arbres additifs, comme en témoigne le fait que l'article de Saitou et Nei a été cité plus de

42 000 fois depuis sa publication. Des justifications théoriques de l'efficacité de l'algorithme ont été présentées par Mihaescu et al. (2009).

4. Confrontations illustrées par trois situations-type.

On présentera un exemple de textes longs en nombre limité (section 4.1), un exemple de réponses à une question ouverte (section 4.2) et un exemple plus méthodologique d'analyse de thésaurus (recueils courts) (section 4.3).

4.1 Les discours sur l'Etat de l'Union des 19 derniers présidents des USA (1900, 2009).

Cet extrait du corpus classique disponible sur http://www.usa-presidents.info/union/ et accessible à partir de ***nltk*** a été lemmatisé à partir du logiciel TreeTagger (Schmid, 1994), avec élimination des mots outils, prépositions, déterminants, formes élidées, pronoms. Après ces transformations (discutables, mais ce n'est pas le thème de cet article), le corpus retenu a une longueur de 546 321 mots et 14 486 mots distincts. On se restreindra ici au texte de 364 180 mots généré par les 487 mots qui apparaissent plus de 199 fois.

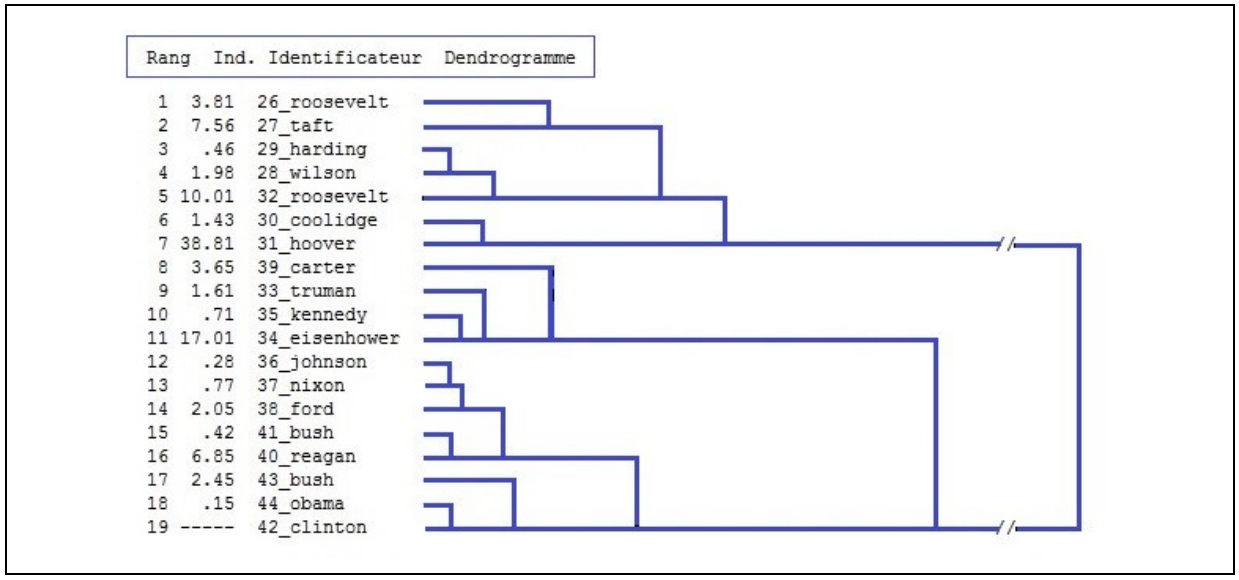

Figure 1. Classification hiérarchique des 19 présidents. CAH, critère de Ward.

La figure 1 contient l'arbre d'une classification ascendante hiérarchique classique utilisant le critère de Ward pour les 19 colonnes de la table de contingence (487 x 19) (lemmes, présidents). Une section verticale des trois grandes branches horizontales de l'arbre met en évidence trois groupes : les présidents 26 à 32 (rangs 1 à 7) ; un groupe 33-35 + 39 (Carter), enfin tous les présidents post-Kennedy (sauf Carter).

La figure 2 contient le premier plan principal de l'AC de la même table, avec aussi le tracé de l'Arbre de Longueur Minimale. Les deux valeurs propres correspondantes, très dominantes représentent 56 % de la somme des valeurs propres (il ne s'agit pas d'un pourcentage d'information - dénomination répandue - car il faudrait alors préciser de quelle *mesure* de l'information il s'agit. L'information statistique au sens de Shannon-Wiener ou de Kullback dépend en fait de la petitesse des dernières valeurs propres). Le tracé de l'ALM ne comporte aucun croisement d'arêtes, situation peu fréquente due à la dominance des 2 axes. Il permet de noter, par exemple, que Bush (43) est plus proche dans l'espace total de Bush (41) que de Reagan (40), contrairement à la suggestion du plan principal. Les trois groupes précédents

sont facilement décelables, avec toujours la dissonance chronologique du président Carter (qualifié de président *OVNI* par certains historiens).

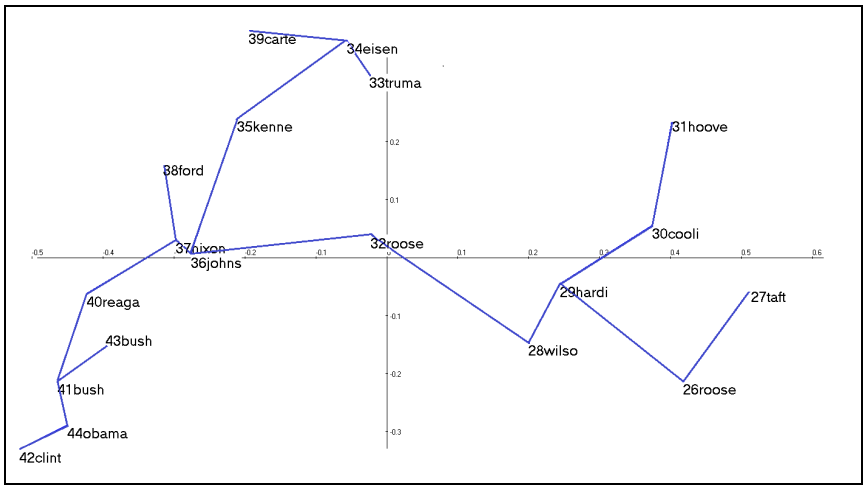

Figure 2. Arbre de longueur minimale tracé dans le premier plan factoriel (valeurs propres 0.12 et 0.05, plan correspondant à 56 % de la trace)

La figure 3 contient le tracé (élégant) de l'arbre additif qui confirme sans innovation importante dans ce cas particulier, les résultats précédents. Notons que l'ALM a $n = 19$ sommets, et donc $n-1 = 18$ arêtes, alors que l'AA a $n = 19$ éléments terminaux et $2n-3 = 35$ arêtes. Son tracé s'avère vite plus difficile dans les plans principaux.

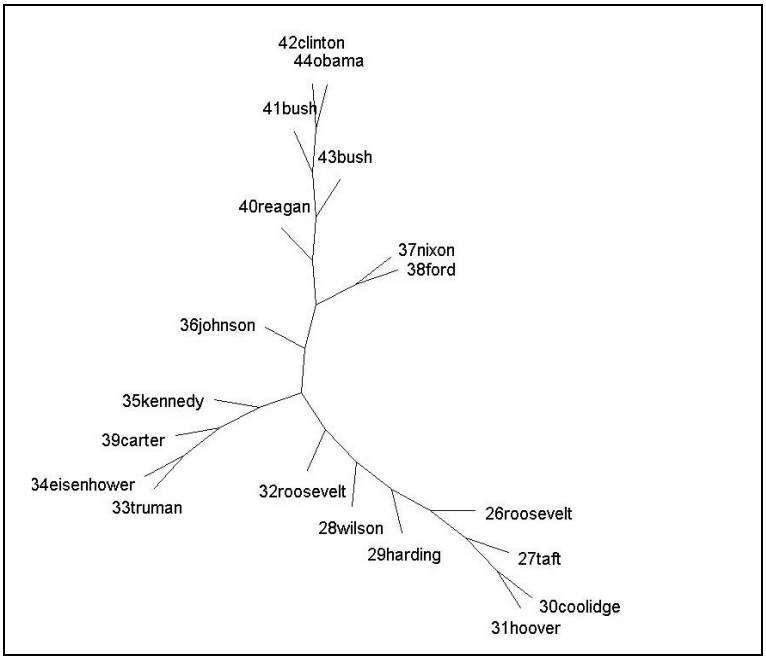

Figure 3 : Arbre additif (Analyse arborée). Méthode Neighbor Joining. *(SplitsTree: Huson and Bryant, 2006)*

Les trois grands groupes décelables sur les figures 1 et 2 se retrouvent, comme les couples de plus proches voisins (Obama – Clinton), (Eisenhower – Truman), (Nixon – Ford), etc.

Enfin, la figure 4 fait intervenir la dualité entre les deux espaces (Lemmes-Présidents). Cette dualité ne concerne que les axes principaux et s'exprime par les relations de transitions entre les deux espaces. Les distances entre présidents étant calculées à partir de leurs profils

lexicaux, il est naturel de positionner quelques lemmes choisis ici parmi les plus responsables. Ici, présidents et lemmes sont assortis de leurs enveloppes convexes de confiance établies par re-échantillonnage *Bootstrap*. On note la précision de la position des présidents dans ce plan. Dans le cas présent d'un plan principal expliquant une part notable de la variance, la représentation simultanée validée par les zones *bootstrap* est un argument non pas en faveur du choix unique des axes principaux, mais en faveur de leur présence indispensable en vue d'une interprétation complète.

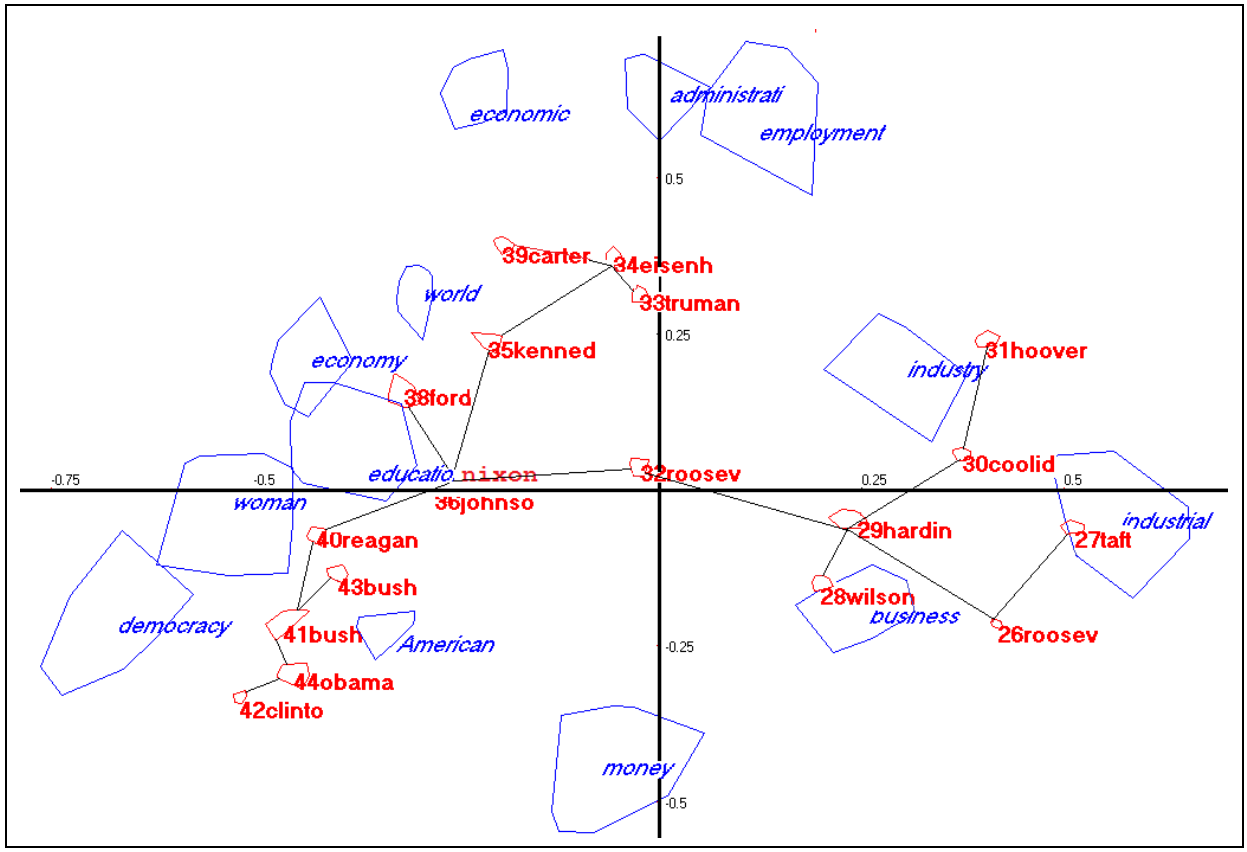

Figure 4 : Après « Comment les distances… », « Pourquoi les distances… » : L'espace des lemmes. Enveloppes convexes de confiance des présidents et de quelques mots avec aussi esquisse de l'ALM.

4.2 Les préférences diététiques des Japonais

Il s'agit maintenant de 1008 réponses à la question ouverte : « *Quels sont les plats que vous aimez et mangez fréquemment?* » posée en 1990 dans le volet japonais de l'enquête internationale sur les comportements, habitudes et préférences alimentaires de trois grandes métropoles : Paris, New York et Tokyo, réalisée sous la direction du Professeur Hiroshi Akuto. Le questionnaire est composé de nombreuses questions fermées donnant notamment les caractéristiques socio-démographiques des répondants. Le « texte » regroupant ces réponses codées en Japonais romanisé comporte 6299 occurrences de 881 mots distincts, l'exemple simple ci-dessous concernant les 4500 occurrences générées par les 83 mots apparaissant plus de 12 fois (cf. Lebart et Salem, 1994, pour une présentation plus détaillée).

La figure 5 représente le premier plan de l'AC de la table lexicale (83 x 6) : elle montre un pattern régulier sexe-âge (âge pour le premier axe, sexe pour le second). Des lignes ont été tracées *a posteriori* pour joindre les catégories d'âge et sexes homologues.

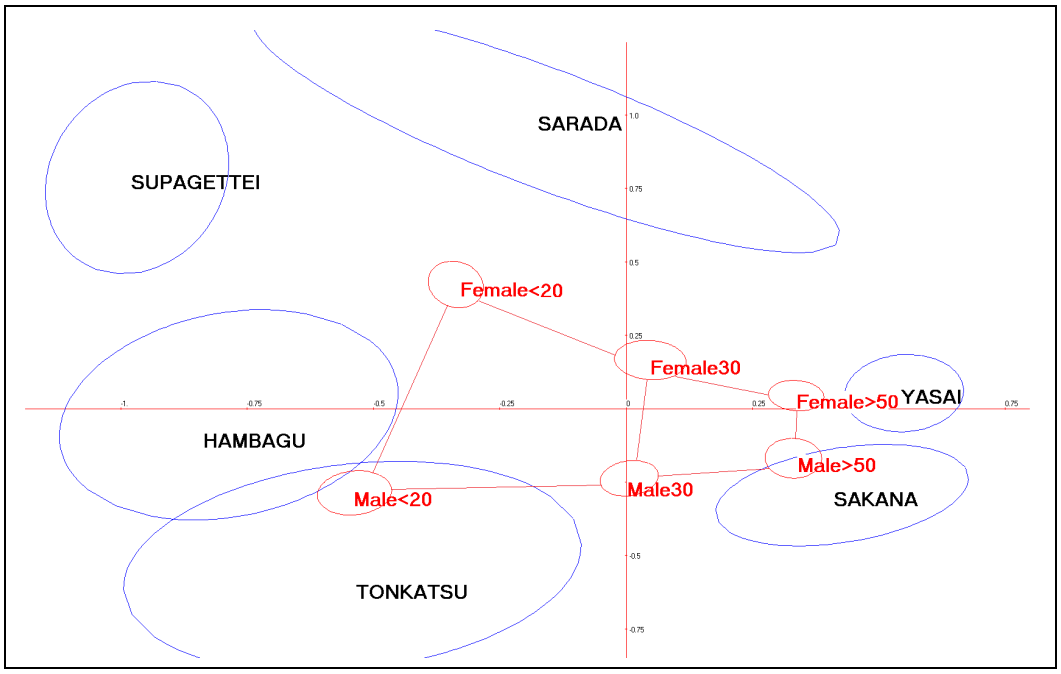

Figure 5: Position de 6 catégories de répondants (sexe – âge) et de quelques mots (romanisés). Ellipse de confiance Bootstrap pour les mots et les catégories. (valeurs propres 0.09 et 0.06, plan correspondant à 71 % de la trace)

La figure 6 donne l'arbre additif des catégories. A propos de la comparaison entre les figures 5 et 6, on peut citer Sattath et Tversky (1977) dans un des articles fondateurs des arbres additifs : *" It is interesting to note that tree and spatial models are opposing in the sense that very simple configurations of one model are incompatible with the other model. For example, a square grid in the plane cannot be adequately described by an additive tree."*

La figure 6 représente correctement les distances entre catégories, mais échoue à séparer les hommes et les femmes. Il est évidemment possible et licite de modifier le tracé de l'arbre en permutant les arêtes *Male>50* et *Female>50* de façon à séparer les sexes, mais rien dans le calcul ni dans le tracé de l'arbre ne permet de faire *a priori* cette opération : une analyse spatiale (par axes) externe est nécessaire.

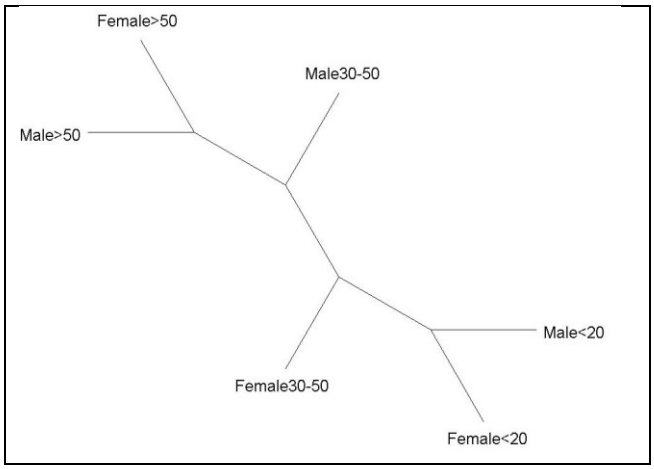

Figure 6. Arbre additif pour les 6 catégories (sexe-âge).

4.3 Les synonymes de 829 verbes français : Succès des arbres additifs

On reprend ici, en la complétant, l'expérience relatée dans l'ouvrage *La Sémiométrie* (Lebart *et al.*, 2003) visant à décrire l'ensemble des verbes français usuels (les 829 verbes les plus fréquents figurant dans le manuel classique de grammaire « Bescherelle ») par l'ensemble de leurs synonymes [pour une approche de ce type de corpus faisant appel à la théorie des graphes, cf. Gaume (2004)]. Le « corpus » formé par les verbes et leurs synonymes comporte 17 446 occurrences de 3 839 verbes distincts. Ce nombre est supérieur aux 829 verbes de départ puisque des verbes moins usités peuvent figurer parmi les synonymes. On traitera ci-dessous les 229 verbes ayant au moins 20 synonymes.

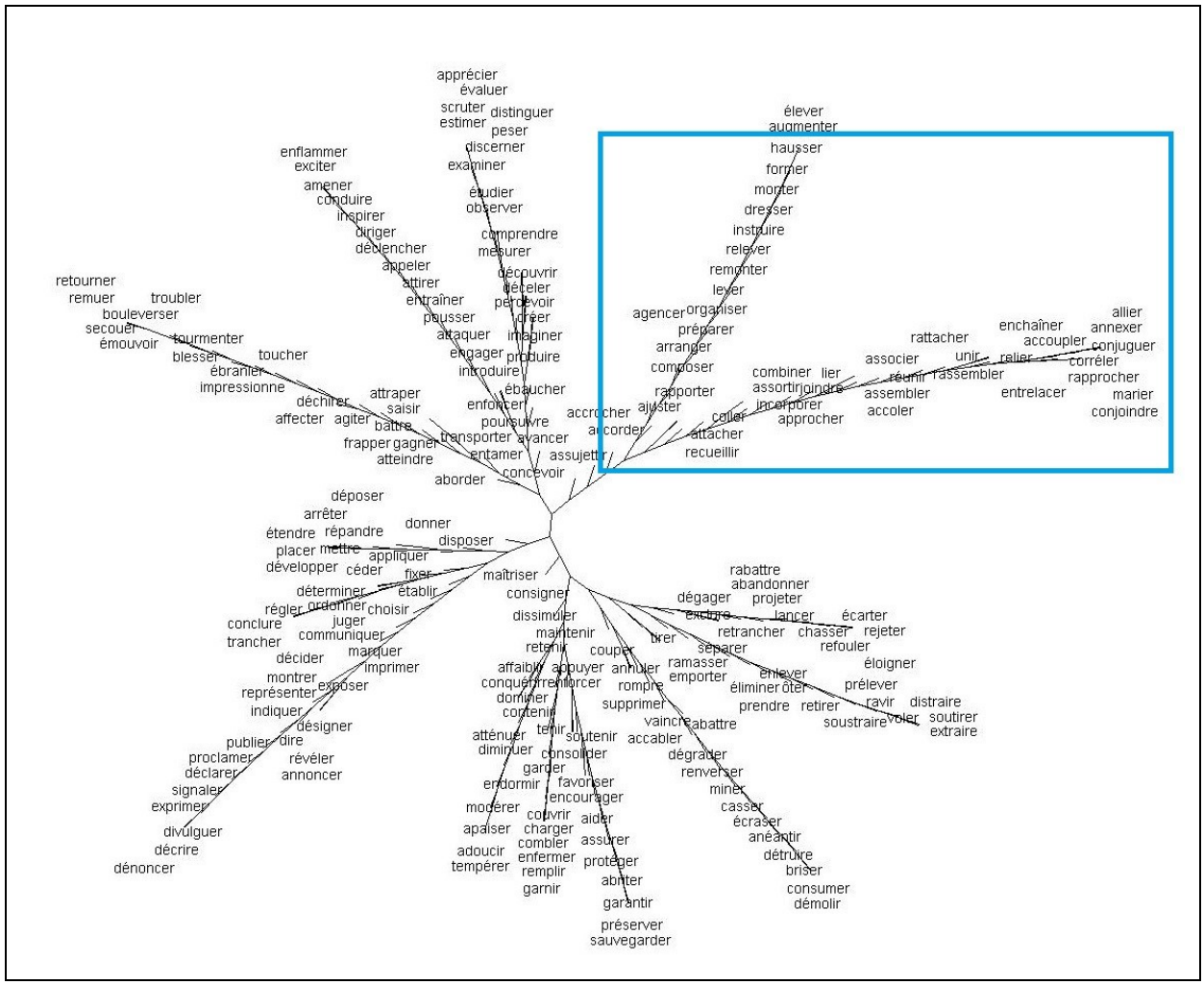

Figure 7. Arbre additif des verbes français (extrait : synonymes apparaissant 19 fois ou plus)

Les conclusions des analyses des correspondances réalisées dans l'ouvrage précité étaient plutôt décevantes : « *En fait, l'analyse géométrique du nuage multidimensionnel de points-verbes montre que ce nuage est presque sphérique (valeur propres voisines). Cette quasi-sphère comporte à la périphérie des « grumeaux » qui sont des amas de verbes sémantiquement voisins. Ces « grumeaux » créent les axes principaux au gré de leur taille, qui dépend d'ailleurs du seuil de fréquence minimale choisi au départ* ». Cette structure péniblement décrite par l'AC est, pourrait-on dire, du pain béni pour les arbres additifs. La figure 7 est un panorama synthétique résumant une dizaine de plans principaux. La plupart des branches de cet arbre correspondent à des axes principaux qui opposent chacun l'une de ces branches à deux ou trois autres.

Autant dire qu'il est artificiel de rechercher les axes principaux d'un oursin (on pense ici aux échinodermes sphériques de nos climats).
Une fenêtre a été grossie (figure 8) pour plus de lisibilité locale. On n'échappe pas aux « effets de chaîne » qui caractérisent souvent les algorithmes ascendants.

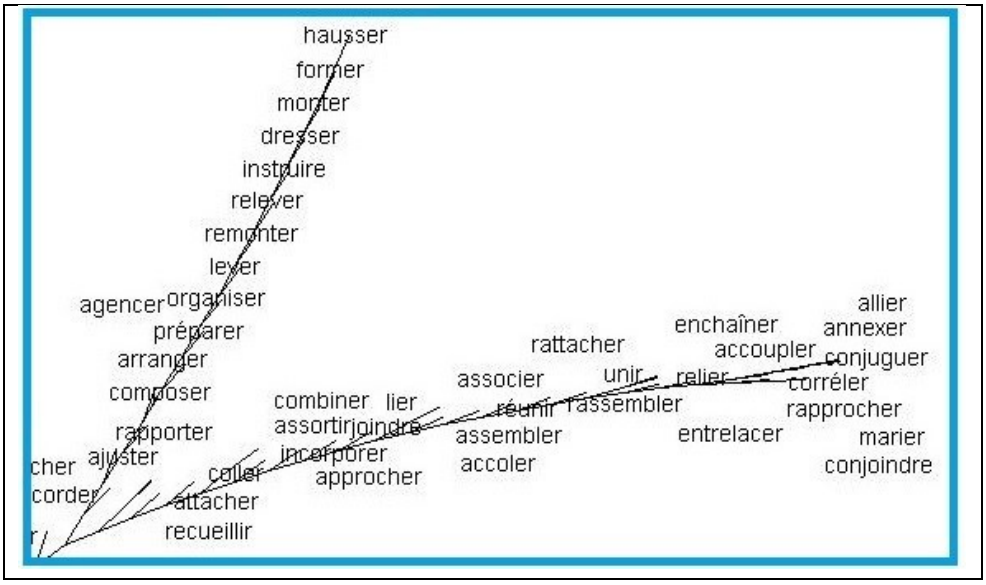

Figure 8. Zoom sur la fenêtre de la figure 7.

Les plans factoriels avec les tracés de l'ALM demanderaient trop de volume pour être publiés dans le cadre matériel de cette contribution. La capacité de synthèse de l'arbre additif pour ce type de structure très multidimensionnelle est remarquable. Ici encore, on peut encore se référer à Sattah et Tverski (1977, p. 338), à propos de leurs propres expériences : *"These observations suggest that the two models may be appropriate for different data and may capture different aspects of the same data".*

4. Conclusion

Brunet (2011), dans un travail de statistique expérimentale beaucoup plus subtil que la présente contribution, a exprimé le point de vue d'un linguiste sur la convergence des analyses arborées et des méthodes factorielles, sans ignorer leur possible complémentarité. Ses propos concernent surtout les « vrais corpus » de textes littéraires, proches de ceux utilisés dans la section 4.1. Nous pensons avoir confirmé ici cette complémentarité, en insistant cependant sur la vocation particulière de chaque outil. La possibilité de représenter simultanément des mots et des textes, puis d'intégrer la méta-information sous formes de variables supplémentaires est un atout indéniable des méthodes en axes principaux. Enrichis de zones de confiance *bootstrap* pour la position des points, ces méthodes sont incontournables, mais gagnent de toute façon à être accompagnées par le tracé d'un arbre de longueur minimale (rectification des distances projetées) et le calcul d'un arbre additif. Ce dernier s'avère en revanche indispensable si la dimensionnalité et la sphéricité du nuage ne permettent pas de visualisation dans un espace de faible dimension. La pire situation, selon Brunet (*op . cit.*) , est celle où l'utilisateur « … que le doute épargne se promène sans vertige sur le parapet de l'interprétation ». Que le petit arsenal présenté ici sème quand même le doute et donne le vertige à ceux qui se contenteraient d'un seul point de vue sur leur corpus.

References

Barthélémy J.-P. and Guénoche A. (1988). *Les arbres et les représentations de proximité*. Masson, Paris.

Benzécri J.-P. and Jambu M. (1976). Agrégation suivant le saut minimum et arbre de longueur minimum. *Les Cahiers de l'Analyse des Données*, vol. (1) : 441-452.

Brunet E. (2011). Le Corpus conçu comme une boule. In : *Ce qui Compte : Ecrits Choisis. Tome II.* Poudat C. (Editor). Honoré Champion, Paris.

Buneman P. (1971). The recovery of trees from measurements of dissimilarity. In: Hodson F. R. D. Kendall G., and Tautu P., (Editors). *Mathematics in the archeological and historical sciences.* Edinburgh University Press, Edinburgh: 387-395.

Bryant D. (2005). On the uniqueness of the selection criterion in Neighbor-Joining. *Journal of Classification,* vol. (22), 1: 3-16.

Cazes P. (1984). Correspondance hiérarchiques et ensembles associés. *Cahiers du B.U.R.O.*, (Université Pierre et Marie Curie) vol. (43-44) : 43-142.

Florek K., Lukaszewicz J., Perkal J., Steinhaus H. and Zubrzycky S. (1951). Sur la liaison et la division des points d'un ensemble fini. *Colloq. Math.*, vol.(2) : 282-285.

Gascuel O. (2000). Data model and classification by trees: The minimum variance (MVR) reduction method. *Journal of Classification,* vol. (17), 1: 67-100.

Gaume, B. (2004). Balades aléatoires dans les Petits Mondes Lexicaux. I3*, Information Interaction Intelligence*, vol. (4), 2 : 1-58.

Gower J. C. and Ross G. (1969). Minimum spanning trees and single linkage cluster analysis. *Appl. Statistics*, vol. (18): 54-64.

Graham R. L. and Hell P. (1985). On the history of the minimum spanning tree problem. *Ann. Hist. Comput*. vol. (7): 43-57.

Huson D.H. and Bryant D. (2006). Application of Phylogenetic Networks in Evolutionary Studies, *Molecular Biology and Evolution*, vol. (23), 2: 254-267.

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

Kruskal J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Amer. Math. Soc.* , vol. (7): 48-50.

Lebart L. and Mirkin B. (1993). Correspondence analysis and classification. In : *Multivariate Analysis: Future Directions 2*, Cuadras C. M. and Rao C. R., Editors, North Holland, Amsterdam, 341-357.

Lebart L., Piron M. and Steiner J.-F. (2003). *La Sémiométrie.* Dunod, Paris.

Lebart L. and Salem A. (1994). *Statistique Textuelle*. Dunod, Paris.

Luong X. (1988). *Méthodes d'analyse arborée. Algorithmes, applications.* Thèse pour le doctorat ès sciences. Université Paris V.

Mihaescu R., Levy D. and Pachter L. (2009). Why Neighbor-Joining works? *Algorithmica*, vol. (54) : 1-24.

Reinert, M. (1983). Une méthode de classification descendante hiérarchique: Application à l'analyse lexicale par contexte. *Cahiers de l'Analyse des Données*, vol. (3) : 187-198.

Sattath S. and Tversky A. (1977). Additive similarity trees. *Psychometrika,* vol. (42), 3: 319-345.

Saitou N. and Nei M. (1987). The neighbor joining method: a new method for reconstructing phylogenetic trees, *Molecular Biology and Evolution*, vol. (4), 4: 406-425.

Schmid H. (1994). Probabilistic part of speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.