L'articulation entre exploration et inférence en analyse statistique de textes

Ludovic Lebart[1]

[1] Télécom-ParisTech, ludovic@lebart.org

Abstract

After a brief reminder about the geometrical aspects of data analysis, we contrast the supervised approach (leading to straightforward external validation) and the unsupervised approaches (leading to several methods of internal validation based on re-sampling techniques). We present then in the unsupervised case some validation procedures allowing for a critical use of the methods and thus providing an assessment of the results. These procedures could be described as variants of *Bootstrap* techniques adapted to the complex nature of textual data.

Résumé

Après un rappel sur le paradigme de l'analyse des données et la fécondité de son application aux corpus de textes, on opposera les approches supervisées et non supervisées en insistant sur le rôle et la taille des unités de contextes dans le cas non supervisé. On présente ensuite les outils de validation qui nous paraissent les plus adaptés. Ceux-ci font appel aux techniques de rééchantillonnage (*Bootstrap*). Un exemple d'application à des textes politiques américains illustre les approches et les méthodes précédentes.

Mots-clés : Inférence statistique, validation, bootstrap, analyse des données.

1. Introduction

Il y a sûrement autant de variétés de statisticiens que de linguistes, c'est dire toutes les combinaisons potentielles de compétences impliquées dans notre domaine de recherche. Ces combinaisons, croisées à leur tour avec des champs d'application qui croissent et se multiplient, donnent lieu à de très nombreux travaux dont les actes des JADT témoignent pour une part importante. Le paradigme initial, dès les premières réunions des JADT à Barcelone (1990) puis à Montpellier (1993) et Rome (1995) fut celui de l'analyse des données, locution qui s'est un peu figée dès 1965 sous l'impulsion de Jean-Paul Benzécri pour désigner, au moins en France, l'analyse exploratoire multidimensionnelle des données. L'enrichissement conceptuel pour les textes a été immédiat. La fécondité du paradigme a été source d'enthousiasme et de passions, pas toutes éteintes aujourd'hui, même si l'on doit rebaptiser la démarche *Text Mining* pour pouvoir communiquer plus facilement, ou même simplement exister dans certains cadres académiques.

Le passage de la notion de scalaire à celle de vecteur, c'est-à-dire, en bref, des mots aux profils lexicaux, a été déterminant : on peut calculer des profils lexicaux à partir de phrases,

de paragraphes, de chapitres, de livres, de réponses, d'articles, de discours, et on peut calculer des distances entre ces profils lexicaux. Des outils permettent de représenter et de classer ces distances. Les comptages et les fréquences des pionniers de l'analyse statistique de texte qui pouvaient paraître ternes ou arides se complètent maintenant avec des formes, des structures, des typologies, des arborescences. Les données sont certes plus vastes et complexes, mais les résultats aussi. Surtout, ces résultats ne sont plus binaires, comme à l'issue d'un test d'hypothèse. Une grande ambition : ils sont censés être plus près de la pensée que les données brutes. Mais il y aura un prix à payer pour ces progrès techniques : investissement et formation pour le chercheur, et/ou division du travail, jamais bienvenue ni d'ailleurs souhaitable dans un processus de connaissance.

1.1. Histoires de réticences

La démarche de l'analyste de données, que celles-ci soient numériques ou textuelles, est souvent mal comprise, pour des raisons diverses, parfois opposées. Prenons un exemple historique, qui est celui des premiers balbutiements de l'analyse des correspondances multiples (ACM). En 1941, le génial Louis Guttman (physicien de formation) propose dans un article (Guttman, 1941) une technique pour construire une échelle (scaling) qui n'est autre que l'ACM, dans tous ses détails de formulation analytique, de calcul, de notations (modernes). Dans cet article séminal qui passe assez inaperçu en pleine guerre, il préconise de n'utiliser que la première dimension extraite (on dirait actuellement, le premier axe factoriel) pour construire une échelle. Presque dix ans plus tard, le psychologue Cyril Burt redécouvre la méthode (Burt, 1950), à qui il prête des vertus exploratoires, en préconisant de retenir plusieurs dimensions. Une polémique s'ensuit, concernant l'antériorité de la méthode (Guttman, 1953 ; Burt, 1953) et ses potentialités. Burt concède effectivement la paternité à son prédécesseur, mais a beaucoup de mal à le convaincre de regarder et d'interpréter plusieurs dimensions. Beaucoup de critiques ont été adressées pour d'autres raisons à Burt, beaucoup plus tard, mais force est de reconnaître que le psychologue, habitué à une réalité complexe et au concept de dimension, voyait des choses que le mathématicien, soucieux de rigueur et de quantification, concevait visiblement assez mal. Cet exemple est emblématique d'un premier type d'incompréhension : l'exploration paraît soit inutile, soit indigne d'un statut scientifique, soit simplement incongrue, parce qu'on ne sent pas qu'il y a un espace complexe à explorer.

1.2. Richesses et difficultés

L'analyse des données textuelles se heurte aux mêmes difficultés et ostracismes. C'est vrai que d'analyser des textes, mêmes avec des outils scientifiques, n'est pas forcément une activité scientifique. Mais se promener dans des textes avec en bandoulière quelques puissants outils de visualisation et de synthèse, c'est une activité fort agréable pour ceux qui aiment les textes (sentiment qui caractérise d'ailleurs la communauté JADT) mais c'est aussi une phase indispensable à toute démarche scientifique sur les textes. C'est la phase de systématique qui précède la formation de toute science, comme ce fut le cas, par exemple, pour la botanique ou la géologie. En termes simples, il faut regarder les données et les textes avant de modéliser. Mais ce n'est pas simple à mettre en pratique, car si on utilise ce qu'on a appris des données dans un modèle, on ne peut plus légitimement tester le modèle sur les mêmes données, situation embarrassante analysée depuis longtemps par Cox (1977). On va montrer dans la section 2 qui suit (*Aspects non-statistiques de l'analyse des données*) que les outils que l'on utilise ne sont

pas uniquement statistiques, et que le statut des résultats que l'on obtient reste encore à élucider. Puis on évoquera certains concepts élémentaires de la théorie de l'apprentissage qui peuvent nous aider à travailler sur les textes (section 3 : *Modèles supervisés et non-supervisés*). On rappellera dans la section 4 les outils de validation qui peuvent s'appliquer au multidimensionnel, et donc aux textes (*Les épreuves de validité adaptées aux textes*). Enfin la dernière section sera consacrée à un exemple d'application qui s'efforcera d'illustrer les phases précédentes, et de nous aider à répondre à la question suivante : Comment passer d'une contemplation à une exploration, puis à des conclusions ?

2. Aspects non-statistiques de l'analyse des données

Peut-il exister une statistique sans fréquence ni répétition ? Dans quelle mesure les schémas statistiques appliqués aux textes sont valides ? réalistes ? utiles ?

Etienne Brunet (1984), dans un article profond et savoureux intitulé « Le viol de l'urne » répond, avec patience et pédagogie à un mathématicien qui questionne avec véhémence le schéma d'urne utilisé par la statistique lexicale. Etienne Brunet rappelle avec force arguments et exemples que « le schéma d'urne est une figure idéale, sans cesse démentie par la réalité du discours ». On peut reformuler son argumentation en disant qu'un modèle permet de concevoir des outils plus utiles que le modèle lui-même. Ce fut le cas de l'analyse factorielle classique (*factor analysis*) découverte à partir d'un modèle psychologique considéré comme simpliste (Spearman, 1904) : le modèle a été critiqué et invalidé au fil des années, mais la méthode a survécu et s'est diversifiée.

La référence au schéma d'urne se complique et s'enrichit avec l'utilisation de l'analyse des données car il existe une composante géométrique (et non statistique) dans les outils exploratoires qui nous éloigne cependant beaucoup du schéma d'urne et des modèles statistiques, sans pour autant nous dispenser de pièges et d'écueils. On va esquisser deux exemples : l'analyse d'un graphe, et la compression d'une image.

2.1. Description de graphes

Supposons que l'on demande aux 95 préfets des départements français de répondre à la question ouverte : «Quelles sont vos départements voisins ?». Les trois premières réponses (laconiques !) figurent dans le tableau ci-dessous.

**** Ain
Ain Isere Jura Rhone Hte_Saone Savoie Hte_Savoie
**** Aisne
Aisne Ardennes Marne Nord Oise Seine_Marne Somme
**** Allier
Allier Cher Creuse Loire Nievre Puy_de_Dome Hte_Saone

Tableau 1 : Codage textuel de la contiguïté pour trois départements français

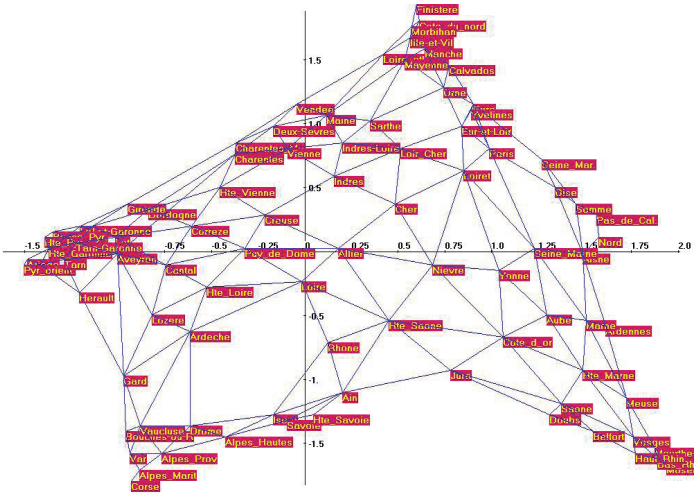

Figure 1. Carte de France schématique (sud à gauche et nord à droite) reconstituée à partir des textes ci-dessus

Il s'agit, certes, de textes bien spéciaux, qui, soumis à une analyse des correspondances lexicale (analyse de la table « préfets x mots »), nous donne la carte de la figure 1, où les départements sont positionnés les uns par rapport aux autres sans aucune interversion.

Il y a effectivement des données sous forme de texte, mais pas de schéma d'urne ni de fréquence sous-jacents. La structure initiale est reconstituée à partir d'une relation d'incidence. Donc certaines structures peuvent être détectées (celle de graphe planaire ou approximativement planaire constitue un cas assez favorable). Notons qu'il n'existe pas d'outils statistiques permettant de valider une telle représentation[1]. La carte est reconnue, reconnaissance qui résulte d'une validation externe. L'intervention de « variables supplémentaires » pourra aussi jouer le rôle de validation externe, comme on le verra dans les exemples d'application de la section 5.

2.2. Compression d'images

Ce second exemple[2] porte sur un recueil de données qui n'est pas plus aléatoire que des textes comme la *Déclaration des droits de l'homme* ou que *Booz endormi* : une photographie, c'est-à-dire ici une table à 145 lignes et 723 colonnes (241 pixels x 3 couleurs) représentant un oiseau. Chaque cellule contient un nombre compris entre 0 et 255 (niveau de : Rouge, Vert, Bleu) L'exemple illustre les propriétés de compression de l'analyse des correspondances.

1 Comme toujours dans le cas de codage binaire clairsemé (*sparse matrix*), les deux premières valeurs propres sont peu détachées des autres valeurs propres et n'expliquent qu'une part très faible de la variance totale.

2 Les données correspondant à ces deux exemples et à tous ceux qui suivront en section 5, ainsi que le logiciel de traitement (Dtm-Vic) peuvent être librement téléchargés à partir du site : www.dtm-vic.com.

Figure 2. Cardinal de l'île Maurice. Cas de l'analyse des correspondances : Images reconstituées successivement avec deux axes principaux, 10 axes et 100 axes.

Avec deux axes (figure de gauche), on obtient une simple tache de couleur, avec 10 axes (centre), ce qui correspond à seulement 7% du volume de la table initiale, l'oiseau est déjà reconnaissable. Donc l'outil que nous utilisons a le pouvoir de produire des résumés (plus numériques que statistique) à partir de tableaux de données qui décrivent des réalités non-statistiques. Le plus remarquable dans cet exemple, est que la compression (comme toute diagonalisation) ne dépend pas de l'ordre des lignes et des colonnes (ce n'est pas le cas des algorithmes de compression utilisés couramment en photographie). C'est rassurant pour ceux qui considèrent les textes comme des sacs de mots et qui s'en inquiètent trop : la redondance est partout, dans les images comme dans les textes.

2.3. Puissance et limite de l'outil

Nous disposons donc d'outils mathématiques permettant de détecter d'éventuelles formes ou structures et d'effectuer certaines synthèses. Leur intérêt purement heuristique est donc incontestable. Nous avons affaire à des instruments d'observations et non à une modélisation. Dans cette optique, comme en microscopie, il y a indépendance entre les lois ou règles qui régissent l'instrument et celles qui régissent la réalité observée. Dans une optique de validation des structures observées, on pourra en amont préparer les données (prétraitements, perturbations ad hoc) et aussi, comme on le verra en section 4, en aval traiter les résultats.

3. Modèles supervisés et non-supervisés

Rappelons que dans la théorie de l'apprentissage statistique (voir par exemple Vapnik, 1998 ; Hastie *et al.*,2001), il est d'usage de distinguer entre l'«approche non supervisée » (Ce qui signifie approximativement: «approche exploratoire ou descriptive») et «Approche supervisée (étroitement liée à l'«approche confirmatoire ou explicative «). Généralement, les techniques d'analyse factorielle et de classification sont non supervisées, alors que l'analyse discriminante (attribution d'éléments à des classes existantes) ou la régression sont des méthodes supervisées. La validation externe est la procédure normale dans le cas de modèles d'apprentissage supervisés. Une fois que les paramètres du modèle ont été estimés (phase d'apprentissage), la validation externe sert à évaluer le modèle (phase de généralisation), généralement avec des méthodes de validation croisée.

3.1. Validation externe dans le contexte de l'analyse des correspondances (AC).

La validation externe peut être utilisée dans le contexte non supervisée de l'AC dans les deux circonstances pratiques suivantes:

a) lorsque l'ensemble de données peut être divisé en deux ou plusieurs parties, une partie étant utilisée pour estimer le modèle, l'autre partie (s) servant à vérifier l'adéquation de ce modèle,

b) lorsque certaines méta-données ou des informations externes sont disponibles pour compléter la description des éléments à analyser. Nous supposerons que l'information externe a la forme d'éléments supplémentaires (des lignes ou des colonnes supplémentaires du tableau de données). Dans la pratique de l'analyse des données, les éléments supplémentaires sont projetés par la suite sur les plans de visualisation principaux. Leurs positions peuvent être évalués au moyen d'outils statistiques classiques (valeurs-test ou écart-réduit) ou à partir de validation *Bootstrap* (voir section 4). La technique des variables supplémentaires est une régression visualisée[3]. En ce sens, c'est une technique supervisée. Elle permet de conclure, à partir de la question : ces variables supplémentaires sont elles indépendantes du texte ?

3.2. À propos d'une option de fragmentation du corpus

On peut créer de nouvelles «observations artificielles » dans un corpus de textes. Nous utilisons volontairement l'oxymore: «observations artificielles» pour souligner l'originalité de l'approche proposée par Reinert (1983, 1986) à la base d'une procédure connue sous le nom de méthodologie ALCESTE, et maintenant IRAMUTEQ (Ratinaud, 2011 ; voir aussi Marchand, 2011, pour des applications) .

L'idée principale est de considérer le texte comme un «fournisseur potentiel d'observations ». Le texte est un peu arbitrairement découpé en unités, nommées *unités de contexte élémentaire* (UCE) ayant des longueurs égales ou similaires (par exemple 20 mots consécutifs). L'hypothèse sous-jacente est que de telles unités méritent d'être prises en considération : elles contiennent des informations précieuses sur les cooccurrences locales. Cette hypothèse implique une certaine homogénéité du texte. Notons que la création de ces observations artificielles n'est possible que parce que les corpus de textes ont une structure séquentielle ou chronologique. Si nous traitons un ensemble de, par exemple, 50 discours politiques par l'intermédiaire de l'analyse des correspondances (AC) du tableau de contingence lexicale (50 x 1000) croisant les 50 discours avec les 1000 mots les plus fréquents, nous sommes en fait dans le cas d'une approche supervisée à l'égard de l'ensemble des discours. Nous profitons de notre connaissance de la partition du corpus en discours afin d'agréger les mots, et, ce faisant, nous limitons le calcul des distances entre les mots à leurs fréquences globales au sein de chaque discours. [En fait, l'analyse des correspondances d'une table de contingence est également un cas particulier d'analyse discriminante linéaire (cf. Lebart *et al*, 1984). du point de vue des données textuelles, cette analyse produit la meilleure fonction pour discriminer entre les discours à partir des mots utilisés].

3 En régression, on projette une variable à expliquer dans le sous-espace engendré par les variables explicatives. Les coefficients de régression sont les coordonnées de cette projection dans la base des variables explicatives. Une variable supplémentaire est projetée sur une approximation du sous-espace des variables explicatives (sous-espace factoriel) ce qui permet une visualisation de sa projection.

Sinon, si nous analysons, par exemple, une partition du corpus en 2000 UCEs, en ignorant la partition en discours, nous sommes dans le cas d'une phase d'analyse non supervisée vis-à-vis des discours. Si nous projetons ensuite les 50 centres de gravité (moyennes) des UCEs appartenant à chacun des discours (comme catégories supplémentaires), nous effectuons une validation externe de l'analyse non supervisée. Notons que les courtes réponses aux questions ouvertes dans une enquête par sondage peuvent être considérées comme des UCEs naturelles tandis que les catégories de répondants permettent de définir des discours (artificiels). En fait, les deux approches, distinctes, nécessaires, se complètent mutuellement : d'une part, l'analyse supervisée du tableau de contingence (50 x 1000) [discours x mots], d'autre part l'approche non supervisée de la table (2000 x 1000) [UCEs x mots], avec sa confrontation ultérieure à la partition en discours.

3.3. Les avantages de la fragmentation du corpus

Résumons les avantages de la fragmentation du corpus en unités de contexte élémentaires:

- La structure du texte à l'intérieur de chaque discours est prise en compte, un élément d'information négligé dans l'approche classique sur le seul tableau agrégé.

- Une compréhension plus profonde de la structure interne de chaque texte, une granularité plus fine. Prise en compte de cooccurrences locales, en complément d'éventuels segments répétés (Salem, 1987).

- Une validation externe probante peut être réalisée en utilisant la partition du corpus initial en textes (unités de contexte initiales). Cependant, lors de cette validation externe, la qualité des visualisations reste à être évaluée. Les outils de rééchantillonnage présentés dans la section suivante seront un complément indispensable à la méthode traitée dans cette section.

4. Les épreuves de validité adaptées aux textes

4.1. Principe des zones de confiance "bootstrap"

Les visualisations permises par les analyses en axes principaux [composantes principales (ACP), correspondances (AC)] n'ont de sens que si elles sont accompagnées de la confiance que l'on peut accorder à la position de chaque point. Or les calculs à la base de ces visualisations sont d'une grande complexité analytique, et il est exclu de procéder à des évaluations selon les méthodes de la statistique classique.

La technique de *bootstrap* (cf. Efron, 1979 ; Efron et Tibshirani, 1993) va permettre de tracer des zones de confiance (ellipses ou enveloppes convexes de réplications) autour des points représentés sur les plans factoriels, que ces points représentent des mots ou des textes. La méthode consiste à construire *n* "réplications" de l'échantillon par tirage aléatoire *avec remise* des unités statistiques. Dans une réplication, certaines unités apparaîtront ainsi deux fois ou plus, d'autres n'apparaîtront pas. On perturbe ainsi le tableau de données de départ. Sur chacune des *n* réplications, on peut refaire les calculs complexes que nécessitent les techniques de visualisation (i.e. : actionner notre instrument d'observation). Sous des hypothèses faibles, on montre que la variabilité observée sur les *n* réplications est de l'ordre de grandeur de celle que l'on aurait observée dans la population. Autrement dit, on peut disposer de *n* réplications de paramètres complexes, et donc obtenir des intervalles de confiance pour ces paramètres.

Cette propriété est valable pour les vecteurs propres, les réplications des valeurs propres étant biaisées (cf., dans le domaine textuel, Alvarez *et al.*, 2002 – 2004).

Il faut souligner que le *bootstrap* ne se fonde pas sur l'hypothèse d'indépendance, il part de la loi empirique observée (avec toutes les dépendances qui ont pu être observées) et la considère comme loi théorique. Pour reprendre l'image de l'urne, les boules que l'on tire dans l'urne ne sont pas des mots, mais des paires (mot, texte), beaucoup plus nombreuses. Nous prenons en compte l'intérieur de la table de contingence (mots x textes), et non plus seulement ses marges. C'est le ré-échantillonnage par ordinateur qui permet de « coller à la réalité » de cette façon, car il était impossible d'utiliser analytiquement des lois aussi complexes. Le débat précité sur le « Viol de l'urne » mériterait ainsi d'être revisité avec de nouveaux arguments en faveur de la position d'Etienne Brunet.

4.2. Le "bootstrap" partiel

La technique de bootstrap que l'on appellera *bootstrap partiel* (sans recalcul des valeurs propres) proposée notamment par Greenacre (1984) dans le cadre de l'analyse des correspondances, répond à plusieurs des préoccupations des utilisateurs[4]. Une réplication consiste en un tirage avec remise des n individus (vecteurs-observations), suivi du positionnement des p nouvelles variables ainsi obtenues, ces variables ayant le statut de «variables supplémentaires», sur les q premiers axes de l'analyse de base. Les procédures décrites ci-dessus peuvent être mises en oeuvre avec un programme classique de projection d'éléments supplémentaires. On calcule donc les réplications de ce coefficient, ce qui revient à repondérer les individus avec les «poids Bootstrap» 0, 1, 2, ... qui caractérisent un tirage sans remise. On obtient, comme sous-produit, des réplications de la variance sur l'axe, qui sont évidemment distinctes de ce que seraient des réplications des valeurs propres. Les s réplications étant projetées sur un repère commun (celui de l'analyse initiale), on caractérisera graphiquement la dispersion des réplications d'une variable donnée soit par l'enveloppe convexe de l'ensemble de ses réplications, soit par un ellipsoïde d'ajustement du nuage des réplications (qui résultera en fait d'une analyse en composantes principales de ce dernier nuage). L'enveloppe convexe a l'avantage de l'exhaustivité (toutes les réplications sans exception sont enveloppées), l'ellipsoïde a l'avantage de prendre en compte la densité du nuage des réplications, et d'être moins sensible à d'éventuelles rares réplications aberrantes. L'exemple d'application ci-après (section 5) comporte des exemples de tracés d'ellipses.

4.3. Le Bootstrap total

Le *bootstrap total* consiste à réaliser autant d'analyses factorielles qu'il y a de réplications. Mais le système d'axes n'est plus le même d'une analyse à une autre. Il peut y avoir des changements de signes (les axes factoriels sont définis aux signes près), des interversions d'axes, des rotations d'axes[5]. Il faut donc procéder à une série de transformations afin de retrouver des axes homologues au cours des diagonalisations successives des s matrices issus des échantillons répliqués $\mathbf{C}_k$ ($\mathbf{C}_k$ correspond à la *k-ème* réplication). Les trois types de transformations possibles, conduisant à trois types de tests de stabilité, sont :

4 Voir aussi Gifi (1981). Pour une discussion du bootstrap partiel en analyse en composantes principales, cf. Chateau et Lebart (1996).

5 Cf. Milan et Whittaker (1995).

4.3.1. Bootstrap total de type 1

Bootstrap total de type 1 (épreuve sévère, très pessimiste) : simple changement (éventuel) de signes des axes homologues pour les réplications. Il s'agit seulement de remédier au fait que les axes sont définis au signe près. Un simple produit scalaire entre axes originaux et axes répliqués de mêmes rangs permet de rectifier le signe de ces derniers.

4.3.2. Bootstrap total de type 2

Bootstrap total de type 2 (épreuve assez sévère, plutôt pessimiste) : changement de signe et correction des interversions d'axes. Les axes répliqués sont affectés (séquentiellement, sans remise en cause d'affectations antérieures) du rang des axes originaux avec lesquels ils sont les plus corrélés en valeur absolue. Puis on procède à un éventuel changement de signe des axes, comme en bootstrap de type 1.

4.3.3. Bootstrap total de type 3

Bootstrap total de type 3 (épreuve plutôt laxiste si on s'intéresse à la stabilité des axes, mais apte à décrire la stabilité des sous-espaces de dimension supérieure à 1) : une rotation dite procrustéenne (cf. Gower et Dijksterhuis, 2004) permet de rapprocher de façon optimale les systèmes d'axes répliqués et les systèmes d'axes initiaux.

4.3.4. Récapitulation des utilisations

Le bootstrap total de type 1 ignore les possibles interversions d'axes et rotations d'axes. Il permet de valider des structures stables et robustes. Chaque réplication doit produire les axes initiaux avec les mêmes rangs (ordre des valeurs propres).

Le bootstrap total de type 2 est idéal si on veut valider des axes, c'est-à-dire des dimensions cachées, sans attacher une importance particulière aux rangs de celles-ci.

Enfin le bootstrap de type 3 permet de valider globalement un sous-espace engendré par les axes principaux correspondant aux premières valeurs propres. Si par exemple le sous-espace des quatre premiers axes répliqués coïncide avec celui des quatre premiers axes initiaux, on pourra trouver une rotation dans cet espace à quatre dimensions qui fera coïncider les axes (ce qui nous ramène au cas du bootstrap partiel). Comme le bootstrap partiel, le bootstrap total de type 3 peut être qualifié de laxiste par les utilisateurs qui s'intéressent à l'individualité des axes, et pas seulement aux sous-espaces engendrés par plusieurs axes consécutifs. On peut effectivement discuter le principe de compensation des rotations accidentelles, qui sont la cause de l'instabilité des axes.

4.4. Le Bootstrap spécifique (ou hiérarchique)

Le bootstrap spécifique intervient quand il existe plusieurs niveaux d'unités statistiques, ou une hiérarchie de niveaux. Dans le cas des réponses aux questions ouvertes, il existe une population de répondants, et une « population » de mots. On peut travailler sur les tables lexicales mots x catégories de répondants (profession, région, sexe-âge, etc.). Les méthodes bootstrap décrites précédemment stipuleront des tirages avec remise de mots dans une table de contingence. Mais si l'on désire faire une inférence des résultats à la population générale d'où est extrait l'échantillon de répondants, il faut procéder à un tirage avec remise des répondants eux-mêmes. Chaque

répondant est pour nous un « sac de mots », et l'on conçoit que la perturbation des données de la table de contingence est alors plus forte, surtout si ces « sacs » sont de tailles différentes, si certains mots sont fréquents à l'intérieur d'une même réponse, etc. Naturellement, ce type de bootstrap peut lui aussi être partiel ou total, ce qui ne facilite pas la tâche de l'utilisateur.

5. Applications : Variations autour de huit présidents

5.1. Le corpus traité

Nous allons tenter d'illustrer les considérations précédentes à partir d'un unique corpus de taille moyenne, qui est celui des discours sur l'Etat de l'Union des 8 derniers présidents américains extrait d'un corpus plus large qui nous a été communiqué par André Salem [voir par exemple le site : http://www.usa-presidents.info/union/ qui contient l'intégralité des textes depuis le discours de George Washington en 1790]. Ce corpus ne pourra évidemment pas représenter toutes les situations typiques que l'on peut rencontrer en analyse des textes (séries chronologiques longues, enquêtes avec questions ouvertes et questions fermées, entretiens, bases documentaires). Dans le cadre de cette application purement illustrative, et sans que cela soit un préalable systématique pour nous, le corpus a été lemmatisé à partir du logiciel Treetagger (Schmid, 1994), avec élimination des mots outils, prépositions, déterminants, formes élidées, pronoms. Après ces transformations, il a une longueur de 160 117 mots et 8311 mots distincts (dans la suite, on parlera indifféremment de mots ou de lemmes). On se restreindra en fait au texte de 117 099 mots générés par les 583 mots qui apparaissent plus de 49 fois. Ce corpus prétraité contient 12 854 lignes d'environ 120 caractères, détail qui aura de l'importance pour nous car nous allons considérer successivement comme unité de contexte chacune de ces lignes, puis des blocs de 2, 5, 20, 100 lignes consécutives, avant d'étudier la classique table de contingence lexicale (8 X 583) (présidents X lemmes).

5.2. Analyse (non supervisée) des 12 854 lignes

On ne représentera pas les lignes ni les lemmes. La figure 3 représente les points moyens des lignes correspondant à chaque président (i.e. : projection de la variable nominale : « président » en tant que variable supplémentaire). Les cartes factorielles complètes représentant les lignes et les lemmes sont riches d'enseignement, mais impubliables sous ce format. Notons qu'il ne peut s'agir ici que de *bootstrap partiel*, puisque la variable supplémentaire ne participe pas à la construction des axes. Il s'agit de plus de *bootstrap spécifique*, car ce sont les lignes qui seront tirées pseudo-aléatoirement avec remise, et non les mots.

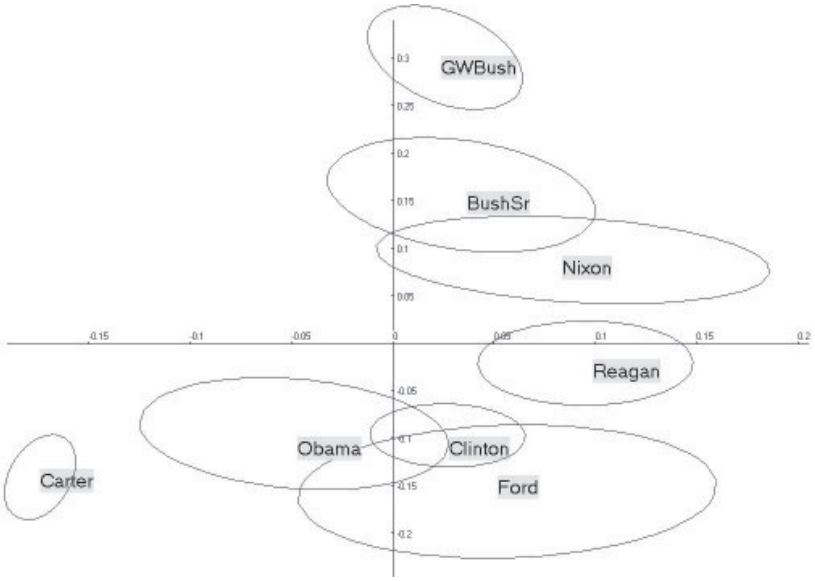

Figure 3. Projection de la variable supplémentaire « Président » sur le premier plan factoriel de l'analyse des 12 854 lignes (considérées comme unités de contexte), avec ellipses bootstrap.

Comme la partition du corpus en 8 présidents n'a pas été utilisée pour construire les axes factoriels, la projection des centres de classes (les présidents) a un caractère de preuve. Les ellipses de confiance *bootstrap* sont construites ici en tirant plusieurs fois avec remise les 12 854 lignes, et en repérant ainsi la variabilité des points-présidents. Les points Carter, G.W.Bush et Reagan sont ainsi assez typés sur les axes, en revanche, les points (Clinton ; Ford, Obama) ne sont pas significativement différents (sur ce plan factoriel), comme les points (Nixon, Bush_Sr),

5.3. Analyse (non supervisée) des 6430 paires de lignes

Le fait de prendre des unités de contexte deux fois plus longues (suites de lemmes ne dépassant pas 240 caractères) a pour effet de modifier de façon notable le premier plan factoriel (figure 4). Nous allons voir que cette modification va dans le sens d'une stabilisation progressive de la structure vis-à-vis de l'agrégation des lignes.

Les points Clinton et Obama restent indiscernables, mais occupent maintenant une position typée. On verra qu'ils resteront indiscernables à tous les niveaux d'agrégation.

Un triangle dont les trois sommets sont Carter, GW.Bush, et le couple (Clinton, Obama) va en fait être conservé en passant à des blocs de 5, 20, 100 lignes. Le point Carter est particulièrement isolé quelque soit la taille des blocs (ancien exploitant agricole avant d'être président, puis plus tard prix Nobel de la paix, Carter a été considéré comme un président hors normes, parfois décrit comme « un OVNI » par les commentateurs politiques. Son vocabulaire est effectivement spécifique : il suremploi : *administration, development, policy, international* et il sous-emploi *America, child*, et les verbes *to do, to say, to let, to know*)

La particularité et l'instabilité des deux analyses qui viennent d'être faites sur des blocs de une et deux lignes sont imputables au phrases très conventionnelles de début et de fin de discours (lemmes typiques : *God, bless, you, America, honor, members, thank, fellow, congress,* etc.). Sous des formes variées, ces vocables sont communs à tous les présidents, à l'exception de

Jimmy Carter (dans le corpus dont nous avons disposé). Ce trait saillant, qui isole cependant Carter, va se dissoudre progressivement au fur et à mesure de l'augmentation de la taille des blocs.

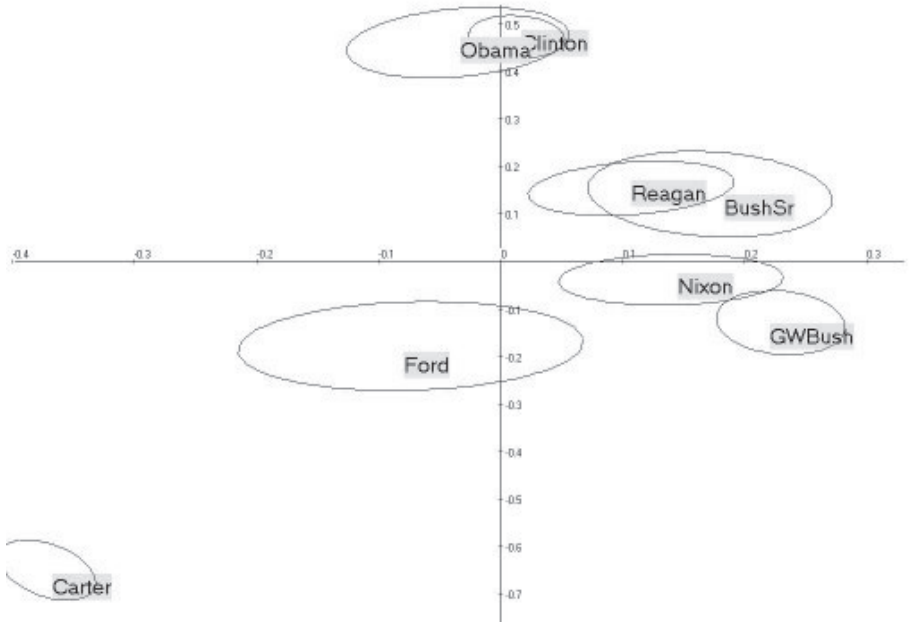

Figure 4. Projection de la variable supplémentaire « Président » sur le premier plan factoriel de l'analyse des 6 430 paires de lignes (considérées comme unités de contexte), avec ellipses bootstrap.

5.4. Analyse (non supervisée) des 2574 blocs de 5 lignes consécutives

Le passage à 5 lignes pour la taille des blocs (maintenant au nombre de 2574) va engendrer un *pattern* des points présidents qui va se stabiliser et se reproduire pour tous les blocs de lignes jusqu'à la taille 60, c'est pourquoi nous ne reproduisons pas la figure ici. On peut avoir une idée du *pattern* au vu de la figure 5 de la section 5.5. Les variations sont minimes : Clinton et Obama, toujours indiscernables, s'opposent à Carter, avec les présidents républicains en position intermédiaire, G.W.Bush étant légèrement excentré. En fait, à une rotation près dans ce plan, cette structure sera stable jusqu'à l'agrégation finale des lignes en 8 blocs présidentiels.

5.5. Analyse (non supervisée) des 646 blocs de 20 lignes consécutives

Comme annoncé, l'agrégation en blocs de 20 lignes reproduit la même configuration de points dans le plan factoriel.

On pourrait s'étonner de trouver des ellipses ayant des tailles similaires, alors que le nombre de blocs diminue dans des proportions importantes. Le *bootstrap* spécifique mis en œuvre ici consiste à tirer les 646 blocs avec remise, alors que les ellipses de la figure 3 étaient obtenues à partir 12 854 tirages de lignes avec remise. D'une part un tirage avec remise induit une perturbation qui ne dépend pas de la taille n de l'échantillon : la probabilité qu'une observation soit absente du tirage avec remise est de 1/e (e = 2.718…) et ne dépend donc pas de n. Certes, le fait d'enlever des blocs paraît plus lourd de conséquence, mais la structure calculée à partir

des blocs est aussi mieux assise, et il y a approximativement compensation entre la sévérité du *bootstrap* et la stabilité de la structure.

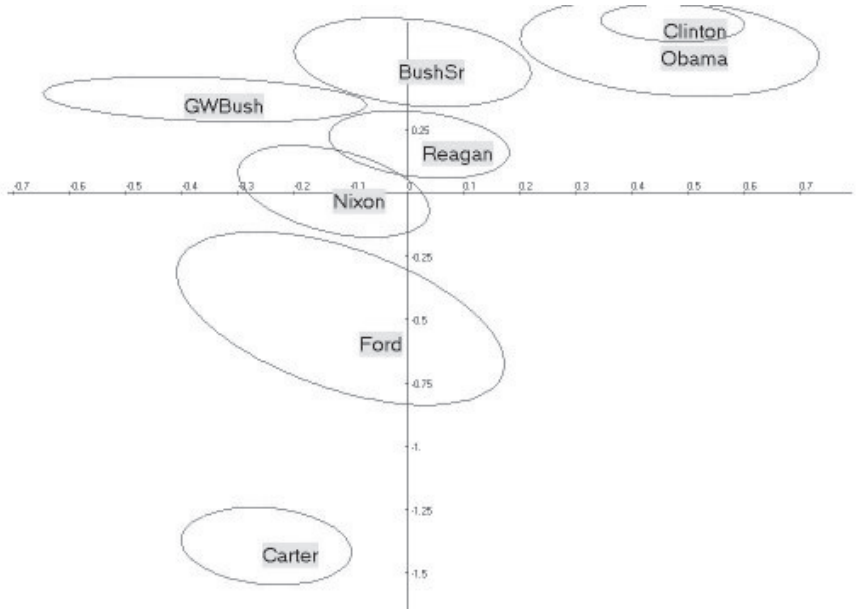

Figure 5. Projection de la variable supplémentaire « Président » sur le premier plan factoriel de l'analyse des 646 blocs de 20 lignes (considérés comme unités de contexte), avec bootstrap.

Mais nous allons tirer avantage du fait que le nombre de blocs devient raisonnable pour voir comment ces blocs se distribuent dans le premier plan factoriel et comment se chevauchent les blocs des différents présidents.

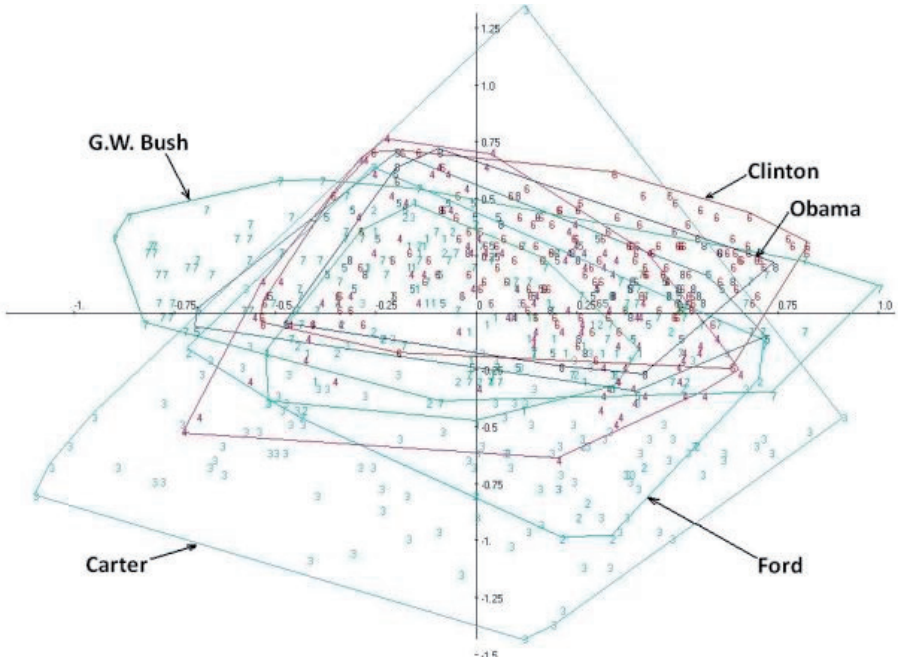

Figure 6. Même plan que figure 6. Enveloppes convexes des blocs de 20 lignes (considérés comme unités de contexte) correspondant à chacun des 8 présidents (au total, 646 blocs).

Pour la figure 6, il n'est plus question de *bootstrap*. Bien que les détails ne soit pas discernables après réduction de la taille de la figure, ce sont les 646 blocs de 20 lignes qui sont représentés. Ces blocs sont les lignes de la table (646 x 583) qui vient d'être analysée. La figure 5 représentait les points moyens des blocs correspondant à chaque président, avec les zones de confiance *bootstrap* correspondantes. Sur la figure 6, ce ne sont plus les points moyens qui sont représentés, mais l'ensemble des blocs, avec les enveloppes convexes des blocs correspondant à chaque président. On voit à quel point les décalages entre présidents paraissent subtils, alors que leurs positions dans ce plan sont néanmoins significatives statistiquement, comme nous l'a montré la figure 5.

La figure 6 est purement géométrique (ou mathématique, si l'on préfère), comme l'étaient les figures 1 et 2 relatives au graphe des départements ou à l'image de l'oiseau. La figure 5 était pour une part géométrique (position des points), pour une part statistique (taille des ellipses). Elle nous a appris, ce qui est loin d'être évident à partir de la seule figure 6, que les écarts entre présidents ne sont (probablement) pas imputables au seul hasard (sauf pour le cas Obama et Clinton, dont les ellipses de confiance empiètent largement). En revanche, la figure 6 nous montre la dispersion des blocs à l'intérieur de chaque discours, nous permet de repérer les blocs typiques et nous incite à les consulter. Enfin, il ne faut pas oublier de consulter les axes factoriels suivants, qui peuvent aussi recevoir leurs ellipses de confiance comme leurs enveloppes convexes de blocs. Le troisième axe permet, dans certaines de ces analyses, de séparer les démocrates (Carter, Clinton, Obama) des républicains.

5.6. Analyse (non supervisée) des 132 blocs de 100 lignes consécutives

Voici maintenant l'analyse dont la granularité est la moins fine parmi les exemples pris en compte ici avant l'analyse des regroupements de lignes en huit blocs correspondants aux huit présidents. L'allure du plan factoriel est stabilisée : chaîne opposant Carter, toujours isolé, au couple Obama/Clinton avec, de gauche à droite, Ford, Reagan, Nixon, Bush_Sr, et G.W.Bush légèrement décalé vers le bas (rotation de 45 degrés de la figure 6, puis changement de signe du deuxième axe). Même remarque que précédemment à propose de la taille des ellipses, construites à partir de blocs 100 fois moins nombreux que dans la section 5.2.

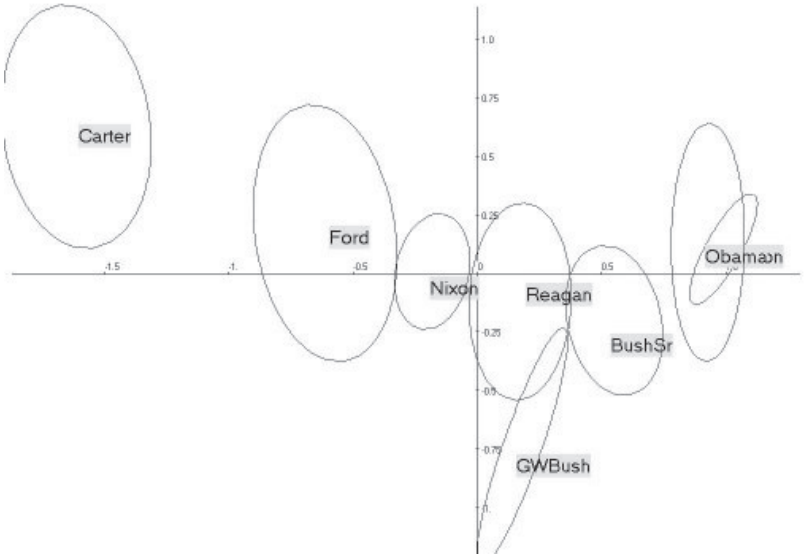

Figure 7. Projection de la variable supplémentaire « Président » sur le premier plan factoriel de l'analyse des 132 groupes de 100 lignes (considérés comme unités de contexte), avec bootstrap.

5.7. Analyse supervisée des 8 discours intégraux des présidents

L'analyse de la table de contingence lexicale (583 x 8) croisant les lemmes et les présidents constitue l'approche la plus courante. Cette phase d'analyse est qualifiée ici de supervisée, car la partition en 8 présidents est utilisée pour construire ce plan factoriel, ce qui n'était pas le cas dans les sections 5.2 à 5.6, pour lesquelles cette partition intervenait *a posteriori* comme une variable nominale supplémentaire dans l'espace des blocs. Ce qui peut frapper à la lecture de la figure 8 est la taille réduite des ellipses de confiance.

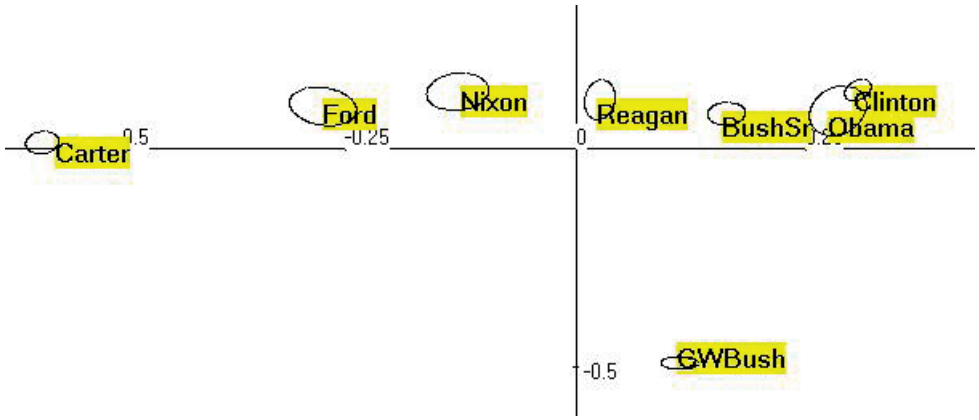

Figure 8. Projection de la ***variable active*** *« Président » sur le premier plan factoriel de l'analyse de la table de contingence (8 présidents, 583 lemmes) avec bootstrap total type 1 (le plus sévère).*

Il s'agit pourtant ici d'un *bootstrap* total de type 1 (le plus sévère, voir section 4 ci-dessus), mais c'est un *bootstrap* sur les mots (ou lemmes, ici), et non sur les blocs. Ce sont les 117 099

occurrences de mots qui sont tirés avec remise à l'intérieur de la table de contingence (mots x présidents) pour créer une réplication de cette table. Ce type de *bootstrap* nous montre quand même que les points Clinton et Obama sont indiscernables sur ce plan. Le modèle statistique sous-jacent prend en compte l'interdépendance observée entre les mots et les présidents (le schéma d'urne sous-jacent au *bootstrap* suppose 583 x 8 = 4664 boules de couleurs différentes), mais sur des nombres d'occurrences aussi importants, il nous rappelle que la plupart des individus (ou leurs scribes) sont différents. Enfin, le *bootstrap* spécifique (section 4.4) appliqué en prenant en compte les blocs de 100 lignes précédents redonne des zones de confiance du même ordre de grandeur que celles de la figure 7.

6. Conclusion

L'analyse statistique des textes est sans doute, comme l'Académie fondée par Platon, un domaine où nul ne devrait entrer s'il n'est géomètre, mais aussi un domaine d'où nul ne devrait sortir en criant *Eureka* s'il n'est (pas du tout) statisticien. On a des outils pour découvrir apprendre et décrire, et d'autres pour communiquer, conclure, prouver, vérifier. Les premiers outils sont incontestablement les plus séduisants, les seconds sont plutôt vécus comme un mal nécessaire. On a tenté d'esquisser ce dilemme au cours de l'exemple précédent. Il reste encore beaucoup à faire pour définir une démarche qui ressemble à une stratégie de traitement. Sur les pas d'Etienne Brunet et à la suite de son « Viol de l'urne », on a simplement voulu souligner d'une part l'apport de la fragmentation en blocs de contextes de tailles variées, et d'autre part l'apport du ré-échantillonnage à son plaidoyer pour une analyse des données textuelles pragmatique, où alternent bon sens, finesse et des épreuves techniques intransigeantes.

Références

Alvarez R., Bécue M., Lanero J. J., Valencia O. (2002). Results stability in Textual Analysis: its Application to the Study of the Spanish Investiture Speeches (1979-2000). In*: JADT-2002, 6-th Intern. Conf. on Textual Data Analysis*, Morin A., Sébillot P., (eds), INRIA-IRISA, Rennes , 1-12.

Alvarez, R., Bécue et M., Valencia O. (2004). Etude de la stabilité des valeurs propres de l'AFC d'un tableau lexical au moyen de procédures de rééchantillonnage. In*: « Le poids des mots »,* Purnelle, G., Fairon, C., Dister, A., editors, PUL, Louvain, 42-51.

Brunet E. (1984). Le viol de l'urne. In : *La recherche française par ordinateur en langue et littérature*. Slatkine-Champion, Genève-Paris. (Repris in « *Compte d'Auteur* », Recueil de textes édité par C. Poudat, Honoré-Champion, Paris, 2011).

Burt C. (1950). The factorial analysis of qualitative data. *British J. of Statist. psychol*. 3, 3, p 166-185.

Burt C. (1953). Scale Analysis and factor analysis. Comments on Dr Guttman paper. *British J. of Statist. psychol*. 6, p 5-20.

Chateau, F. and Lebart, L. (1996). Assessing sample variability in visualization techniques related to principal component analysis: *bootstrap* and alternative simulation methods. In : *COMPSTAT96*, A., Prats, editor, Physica Verlag, Heidelberg, 205-210.

Cox D. R. (1977). The role of significance tests. *Scandinavian Journal of Statist*., 4, p 49-70.

Efron, B.(1979). Bootstraps methods: another look at the Jackknife. *Annals of Statistics*, 7, 1-26.

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

Gifi, A. (1990). *Non Linear Multivariate Analysis*, Wiley, Chichester.

Gower, J., C. and Dijksterhuis, G. B. (2004). *Procrustes Problems*, Oxford Univ. Press, Oxford.

Greenacre, M. (1984). *Theory and Applications of Correspondence Analysis*, Academic Press, London.

Hastie, T., Tibshirani R. and Friedman, J. (2001). *The Elements of Statistical Learning*, Springer, New York.

Guttman L. (1941). The quantification of a class of attributes: a theory and method of a scale construction. In : *The prediction of personal adjustment* (Horst P., ed.) p 321 -348, SSCR New York.

Guttman L. (1953). A note on Sir Cyril Burt's Factorial Analysis of Qualitative Data, *British J. of Statist. psychol*. 6, p 1-4.

Lebart, L. (2003). Validation Techniques in Text Mining, in: *Text Mining and its Applications*, Spiros Sirmakessis, editor, Springer. 169-178.

Lebart, L., Morineau, A. and Warwick, K. (1984). *Multivariate Descriptive Statistical Analysis,* Wiley, New York.

Lebart, L. (2007). Which *bootstrap* for principal axes methods? In: *Selected Contributions in Data Analysis and Classification*, P., Brito *et al.*, editors, Springer, 581 – 588.

Lebart, L., Salem, A. and Berry, E. (1998). *Exploring Textual Data*, Kluwer Academic Publishers, Dordrecht.

Marchand P. (2012). [http://pascal.marchand.fr].

Milan, L. and Whittaker, J. (1995). Application of the parametric bootstrap to models that incorporate a singular value decomposition. *Applied Statistics*, 44, 1, 31-49.

Ratinaud P. (2011). [http://www.iramuteq.org/members/pierre.ratinaud].

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

Reinert, M. (1986). Un logiciel d'analyse lexicale: [ALCESTE]. *Cahiers de l'Analyse des Données*, 4, 471–484.

Salem A. (1987). *Pratique des Segments Répétés. Essai de statistique textuelle*. Klincksieck, Paris.

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

Spearman C. (1904). General intelligence, objectively determined and measured. *Amer. Journal of Psychology,* 15, p 201-293.

Tuzzi, A. and Tweedie, F., J. (2000). The best of both worlds: Comparing Mocar and Mcdisp. In: *JADT2000* (*Cinquièmes Journées Internationales sur l'Analyse des Données Textuelles*), Rajman, M., Chappelier, J-C., editors, EPFL, Lausanne, 271-276.

Vapnik W. (1998). *Statistical Learning Theory*. Wiley, New York.