Les basses fréquences lexicales, problèmes, descriptions et prédictions.

Ludovic Lebart

Télécom-Paris – ludovic@lebart.org

Abstract

The description of lexical tables (cross-tabulating vocabulary and texts) is commonly performed through correspondence analysis (CA) [generally supplemented by clustering and / or additive trees]. CA involves in particular the chi-square distance with its property of distributional equivalence. In many cases, however, Evrard's (1966) distance matrix, based more simply on the presence or absence of words in texts (and closely related to the phi coefficient of Pearson-Yule, 1912) provides more meaningful visualizations. The Evrard distance matrix, easily derived from the correlation matrix of binary variables (presence-absence) matrix is involved in the popular principal component analysis (PCA). After a review of the problems entailed in text analysis when dealing with low frequencies (and high discrepancies of frequencies), we show how the use of binary coding of lexical tables enriches and supplements other descriptive approaches.

Keywords: Low lexical frequencies, presence-absence data, PCA, Pearson Phi coefficient.

Résumé

La description des tableaux lexicaux se fait couramment par analyse des correspondances (AC) qui est adaptée aux tableaux de fréquences en raison notamment de la propriété d'équivalence distributionnelle de la distance du chi-2. Souvent cependant, la matrice des distances d'Evrard (1966), fondées plus simplement sur la présence ou l'absence de mots et dérivée directement du coefficient phi de Pearson-Yule (1912), donne des représentations plus intéressantes s'il s'agit de discriminer entre textes ou de procéder à des attributions d'auteurs. La matrice des distances d'Evrard est dérivée de la matrice des corrélations des variables binaires, matrice qui intervient dans la classique analyse en composantes principales (ACP). Après avoir passé en revue les problèmes posés par les basses fréquences (et les fortes disparités de fréquences) dans les applications aux textes, nous montrons comment les distances sur données binaires donnent un point de vue différent et complémentaire.

Mots-clés: Basses fréquences lexicales, codage présence-absence, ACP, Pearson Phi coefficient.

1. Introduction

La description des tableaux croisant vocabulaire et textes se fait couramment par analyse des correspondances (AC), bien adaptée aux profils de fréquences et aux tableaux lexicaux, en raison notamment de la propriété d'équivalence distributionnelle de la distance du chi-2, l'AC étant ensuite complétée par des classifications et/ou des analyses arborées.

Dans de nombreux cas cependant, les distances d'Evrard (1966) (cf . aussi Brunet, 2011) dérivées du coefficient Phi de Yule-Pearson (1912), fondées plus simplement sur la présence ou l'absence de mots (ou de lemmes), donnent des représentations plus intéressantes s'il s'agit de discriminer entre textes ou de procéder à des attributions d'auteurs. Brunet (Brunet et al., 2020) a ainsi montré à partir d'un corpus de 50 romans de 25 auteurs de la seconde moitié du vingtième siècle (deux romans par auteur) qu'un appariement sans défaut des romans par auteur pouvait s'obtenir à partir de la matrice des distances d'Evrard. Cette matrice se déduit de la matrice des corrélations des variables binaires (présence-absence). Elle intervient dans la très classique analyse en composantes principales (ACP) de Hotelling (1933) appliquée aux données binaires. Après avoir passé en revue les problèmes posés par les basses fréquences

(et les fortes disparités de fréquence) dans les applications aux textes, et examiné les solutions proposées en pratique, nous montrons et illustrons à propos d'un exemple comment l'ACP apporte un point de vue complémentaire à celui de l'AC, permet de décliner les distances d'Evrard en fonction de la dimension de l'espace retenue et de ce fait enrichit les approches plus répandues dans la communauté JADT.

2. Distance d'Evrard, ϕ de Pearson-Yule et χ^2 de Pearson :

En statistique, le coefficient phi (ou ϕ) est une mesure d'association pour deux variables binaires. Fondée sur le coefficient de corrélation *r* de Karl Pearson (1900), cette mesure a été proposée par Yule (1912) qui avait publié antérieurement une mesure d'association voisine (Yule, 1900). Cette mesure est étroitement liée au chi-2 (χ^2) calculé sur la même table de contingence pour tester l'hypothèse d'indépendance, et coïncide avec le coefficient de corrélation *r* de Pearson estimé pour deux variables binaires.

Tableau 1. Bilan de la confrontation de deux textes

		Texte 2		
		Mots présents	Mots absents	Total
Texte 1	Mots présents	n_{11}	n_{10}	$n_{1.}$
	Mots absents	n_{01}	n_{00}	$n_{0.}$
	Total	$n_{.1}$	$n_{.0}$	n

Deux variables binaires sont considérées comme associées positivement si les données se concentrent dans les cellules diagonales et considérées comme négativement associées si elles se concentrent hors de la diagonale. Si nous avons un tableau 2 × 2 pour deux textes, le coefficient ϕ qui décrit l'association de x et y est donné par la formule (Yule, 1912), avec les notations du tableau 1:

$$\phi(1,2) = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1.}n_{0.}n_{.0}n_{.1}}} \qquad [1]$$

Notons que dès 1900, Yule proposait la formule assez voisine : $Q_{Yule} = \frac{n_{11}n_{00} - n_{10}n_{01}}{n_{11}n_{00} + n_{10}n_{01}}$.

Cohen (1960) a proposé de remplacer la moyenne géométrique du dénominateur de la formule [1] par une moyenne arithmétique : $s(1,2) = \frac{2(n_{11}n_{00} - n_{10}n_{01})}{n_{1.}n_{0.} + n_{.0}n_{.1}}$. On pourra consulter Warren (2008) et Baulieu (1989) pour un panorama de la flore des très nombreux coefficients d'association proposés au fil des années et des disciplines.

2.1 Lien de ϕ *avec le chi-2 :*

Le carré du coefficient ϕ est lié à la statistique du χ^2 de Karl Pearson pour la même table de contingence 2 × 2 par la relation (où *n* est le nombre total d'observations : ici nombre de mots distincts).

$$\phi^2 = \frac{\chi^2}{n}, \qquad \text{on a en effet :} \quad n\phi^2 = \frac{n\,(n_{11}n_{00} - n_{10}n_{01})^2}{n_{1.}n_{0.}n_{.0}n_{.1}}$$

(formule classique du χ^2 pour une table 2 × 2, avec 1 degré de liberté [qui a donc 5 chances sur 100 de dépasser 3.84 sous l'hypothèse d'indépendance]).

2.2 Equivalence de ϕ avec le r de Pearson.

Le classique coefficient de corrélation *r* de Karl Pearson calculé sur les données binaires du tableau d'incidence **X** (tableau 2) (calcul légitime pour deux variables à deux modalités) coïncide avec le coefficient ϕ.

Tableau 2 Table d'incidence X de terme général x_{ij}
(x_{ij} = 1 si le mot j [ligne j] est présent dans le texte i [colonne i])

Mots	***Texte 1***	***Texte 2***
Mot 1	1	0
Mot 2	1	1
Mot 3	0	0
Mot 4	1	0
…….	….	….
Mot n	0	1
	$n_{1.}$	$n_{.1}$

$$r_{12} = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{s_1 s_2} \qquad [2]$$

Avec, $\bar{x}_1 = \frac{1}{n}\sum_{i=1}^{i=n} x_{i1} \quad (= \frac{n_{1.}}{n}), \qquad \bar{x}_2 = \frac{1}{n}\sum_{i=1}^{i=n} x_{i2} \quad (= \frac{n_{.1}}{n})$

Et, par exemple, pour s_1 : : $s_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2 \quad (= \frac{n_{0.}n_{1.}}{n^2})$

A partir de la formule [2] et du tableau 2, on trouve de façon élémentaire (mais un peu laborieuse) : $r_{12} = \frac{n_{11}n - n_{1.}n_{.1}}{\sqrt{n_{1.}n_{0.}n_{.0}n_{.1}}}$, et on retrouve exactement la formule [1] en remplaçant n, $n_{1.}$ et $n_{.1}$ par leurs valeurs en fonction de n_{11}, n_{01}, n_{10} et n_{00}.

Cette équivalences avec le classique test d'indépendance du χ^2 et l'identité de ϕ *(1,2)* avec le coefficient de corrélation linéaire r_{12} donnent au coefficient ϕ, et donc à la distance d'Evrard qui en dérive directement, une position privilégiée dans les mesures d'association.

2.3 La distance du chi-2 (χ^2)

La distance du χ^2 utilisée en analyse des correspondances (AC) est une approximation d'une mesure d'information mutuelle dérivée de la théorie de Shannon (1948) évaluant l'information apportée par une table de contingence empirique par rapport à l'hypothèse

d'indépendance des lignes et des colonnes (Benzécri, 1973, Tome 1 B n°5). Cette distance partage avec quelques autres (cf. Escofier, 1978) la propriété d'équivalence distributionnelle qui assure une stabilité des résultats par agrégation de lignes ou de colonnes ayant mêmes profils. L'AC est devenue un des outils de base de la description de tableaux lexicaux. Mais la distance du χ^2 fait intervenir des inverses de fréquences qui peuvent poser problème dans le cas de fréquences très faibles (le critère d'ajustement qui donne à chaque point une masse égale à sa fréquence compense partiellement cette faiblesse).

$$d^2(j,j') = \sum_{i=1}^{i=n} \frac{1}{f_{i.}} \left[\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right]^2$$

3. Fréquences basses ou disparités de fréquences

On passe brièvement en revue dans cette section trois approches qui visent à remédier aux fortes disparités de fréquences ou à faire intervenir des codages en présence-absence de mots.

3.1 L'analyse logarithmique

L'analyse logarithmique (AL), vérifie aussi la propriété d'équivalence distributionnelle de l'AC sur des tableaux de nombres positifs. Kazmierczak (1985) s'appuie pour l'AL sur le principe de Yule (1912) selon lequel on ne change pas la distance entre deux lignes ou entre deux colonnes d'un tableau en remplaçant les lignes et les colonnes de ce tableau par d'autres lignes et colonnes proportionnelles (généralisation de l'équivalence distributionnelle).

L'AL consiste à prendre les logarithmes des données (après addition éventuelle d'une constante en cas de données négatives ou nulles), puis, après les avoir centrées *à la fois en ligne et en colonne*, à les soumettre à une analyse en composantes principales (ACP) non normée, qui coïncide ici avec une simple décomposition aux valeurs singulières (SVD). Si **X** est un tableau de données (*n, m*) et si **A** et **B** sont deux matrices diagonales respectivement de dimensions (*n, n*) et (*m, p*) à éléments diagonaux positifs, l'analyse logarithmique du nouveau tableau **AXB** coïncide avec celle de **X**. Cette propriété de forte invariance, jointe à l'effet contractant de la fonction logarithme, rend robuste cette technique, bien adaptée aux applications à des données massives, pour lesquelles les disparités de fréquences (de 1 à 10^5 par exemple) constituent un obstacle technique. Cette méthode remonte aussi à Aitchison (1983) dans un cadre différent. Une variante voisine avait été proposée initialement sous le nom de *Spectral Analysis* par Lewi (1976), puis par Greenacre et Lewi (2009).

3.2 Coefficients TF-IDF et LSA

Le terme général du tableau lexical mots x textes peut être remplacé par le coefficient TF-IDF (Salton et McGill, 1983). Rappelons que le coefficient TF-IDF (*Term frequency x Inverse of Document frequency*) est le produit de la fréquence d'un terme (TF) par le logarithme du quotient : « nombre total de documents / nombre de documents dans lesquels le terme est présent ». Ce quotient (IDF) fait donc intervenir l'inverse de la proportion des documents où le terme apparaît. Le logarithme, comme pour l'AL évoquée plus haut, permet d'amortir les situations extrêmes, comme le cas où le terme n'est présent que dans un document sur des milliers. Autrement dit, le coefficient TF-IDF combine un indicateur de dominance du terme (composante TF) avec un indicateur de sa spécialisation dans le corpus (composante IDF), ce dernier indicateur variant de 0 (le terme est dans tous les documents) à un maximum (lorsque le terme est dans un seul document) qui dépend de la taille de l'ensemble des documents.

Dans le cadre de la recherche documentaire, il s'agit ici de trouver un ou des documents dans une base de documents (courts et nombreux) à partir de quelques termes. On doit pénaliser les documents qui ne contiennent pas ces termes (élément TF dans la formule). Si on désigne par *d* le nombre de documents, *d(i)* le nombre des documents qui contiennent le mot *i*, par f_{ij} la fréquence du mot *i* dans le document *j*, $f_{i.}$ la fréquence totale du mot *i*, et $f_{.j}$ la fréquence totale du document *j*. , on a :

- fréquence du terme *i* dans le document *j* : TF (*i*,*j*) = ($f_{ij} / f_{.j}$).

- logarithme de l'inverse de la fréquence des documents contenant le terme *i* : IDF (*i*,*j*) = log(*d* / *d(i)*).

Comme l'AC et comme l'AL, l'analyse sémantique latente (*Latent Semantic Analysis*, LSA) [ou indexation sémantique latente (*Latent Semantic Indexing*, LSI)] (Deerwester *et al.*, 1990) est une décomposition en valeurs singulières (SVD) d'un tableau lexical transformé.

Ici, il s'agira de la matrice de coefficients TF-IDF de terme général :

$$t(i,j) = \frac{f_{ij}}{f_{.j}}\left(\log\left(\frac{d}{d(i)}\right)\right) \qquad [3]$$

On montre par ailleurs que l'AC peut se déduire de la SVD de la matrice de terme général :

$$w(i,j) = \frac{f_{ij}}{\sqrt{f_{i.}\,f_{.j}}} \quad \text{que l'on peut écrire:} \quad w(i,j) = \frac{f_{ij}}{f_{.j}}\left(\sqrt{\frac{f_{.j}}{f_{i.}}}\right) \qquad [4]$$

Les formules [3] et [4] diffèrent par les facteurs représentés par leurs parenthèses de droite qui pénalisent toutes les deux les mots *i* très présents dans le corpus : par le nombre de documents *d(i)* qui les contiennent pour *t*(*i*,*j*), par leur fréquence globale $f_{i.}$ pour *w*(*i*,*j*). Les concepts de nombre *d* de documents, et de nombre *d(i)* de documents contenant un mot *i* sont surtout opératoires pour des documents nombreux et courts.

3.3 Méthodologie Alceste

Reinert (1983) a proposé de créer de nouvelles unités statistiques dans un corpus de textes.. Celui-ci est découpé en *unités de contexte élémentaire* (UCE) ayant des longueurs similaires (par exemple 20 mots consécutifs, une ou plusieurs lignes de 120 caractères, une phrase). L'analyse de ces nouvelles unités est à la base d'une procédure connue sous le nom de ALCESTE.

Cette méthodologie est implémentée dans les logiciels ALCESTE et IRaMuTeQ (Ratinaud, 2014). A partir du moment où on travaille sur des fragments courts, toutes les fréquences sont basses à l'intérieur d'un fragment, et la présence ou l'absence d'un terme peut être prise en compte. Dans ce cas, le codage binaire intervient après transformation du corpus.

Remarque générale :

On a vu que les fréquences basses interviennent de façon naturelle dans les textes courts, qu'il s'agisse de documents ou de résumés dans une base, de fragments ou unités de contexte, de pages de romans, voire de réponses à des questions ouvertes. Le codage par présence-absence de mots est une option acceptable et qui a fait ses preuves empiriquement. Il peut de plus être modulé par seuillage (« présent » si plus de *s* occurrences, par exemple). En revanche, pour des applications à des textes importants en volume, coder la présence ou l'absence d'un mot est une option délibérée qui donne un autre point de vue sur les textes d'un corpus, point de vue distinct et complémentaire de celui de la prise en compte globale des fréquences.

4. Exemple illustratif

Pour montrer la pertinence des codages de type présence-absence, et de l'utilisation de l'ACP dans ce cas, on utilisera le corpus classique **STATE OF THE UNION** qui rassemble les discours sur l'État de l'Union prononcés par les présidents américains en cours de mandat devant le congrès, de George Washington (1790) à Barack Obama (2009) [42 discours]. Le corpus (à jour) est disponible sur http://stateoftheunion.onetwothree.net/index.shtml, également accessible à partir du site de nltk (Natural Language Tool-kit, cf. : http://www.nltk.org/nltk_data/, rubrique : *C-Span Inaugural Address Corpus*).
Le corpus utilisé ici totalise 1 746 702 occurrences et 25 246 mots distincts. (Il est téléchargeable à partir du bouton « Matériel complémentaire » du site : https://www.puq.ca/catalogue/livres/analyse-des-donnees-textuelles-3651.html). Pour cet exemple méthodologique, on travaillera sur les formes graphiques du texte brut (sans lemmatisation).

Nous parlons ici d'illustration plutôt que d'application car ce corpus est surtout un repère permettant des comparaisons et non un objet d'étude en soi. Sa forte structure chronologique fait que d'autres méthodologies peuvent s'appliquer avec profit, et l'auctorialité parfois problématique de certains discours demanderait des précautions d'interprétation qui dépassent le présent exercice.

On schématisera par quelques graphiques la démarche de description globale du corpus après codage sous forme de présence-absence de mots.

4.1 Analyse en composantes principales du tableau présence absence (figure 1)

Le tableau 1 présenté en section 1 a maintenant 42 colonnes (présidents) et 10 030 lignes (formes graphiques). Les lignes doivent comporter au moins deux 1 (présence) [on élimine ainsi les hapax] et au moins deux 0 (absence) [on élimine les termes utilisés dans tous les textes ou absents dans un seul texte] ce qui a pour effet de délester le tableau en supprimant bon nombre de mots-outils et d'auxiliaires, et de termes très courants. La perte d'information brute peut paraître considérable. Mais la seule information qui nous intéresse à ce stade est celle que les outils de description sont capables d'utiliser.
La forme parabolique de la séquence des présidents sur la figure 1 n'est pas qu'un simple « effet Guttman » car le nuage des 10030 mots n'épouse pas cette forme, et la zone située à l'intérieur de la courbe contient les mots communs aux périodes extrêmes.

On présente le plan des axes 2 et 3 pour assurer une comparaison avec le plan (1, 2) de l'analyse des correspondances qui va suivre. Nous étudierons l'axe 1 de l'ACP (facteur dit « de taille ») plus bas.

4.2 Comparaison avec l'AC du tableau lexical entier (figure 2)

La figure 2 donne le plan factoriel (1, 2) d'une analyse des correspondances de la table lexicale originale, plus volumineuse qu'une table avec codage « présence-absence ».

La séquence des vingt premiers présidents (partie droite de la figure) est moins clairement représentée dans cet espace.

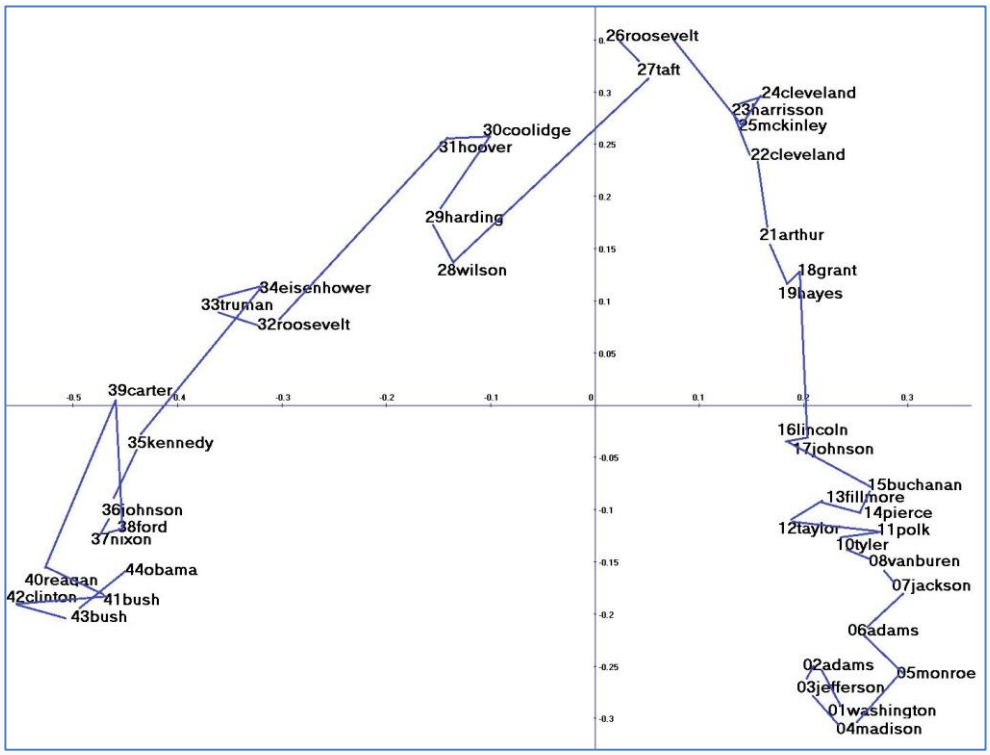

Figure 1. Plan (2, 3) de l'ACP de la table binaire (10030 x 42) Mots x Présidents.

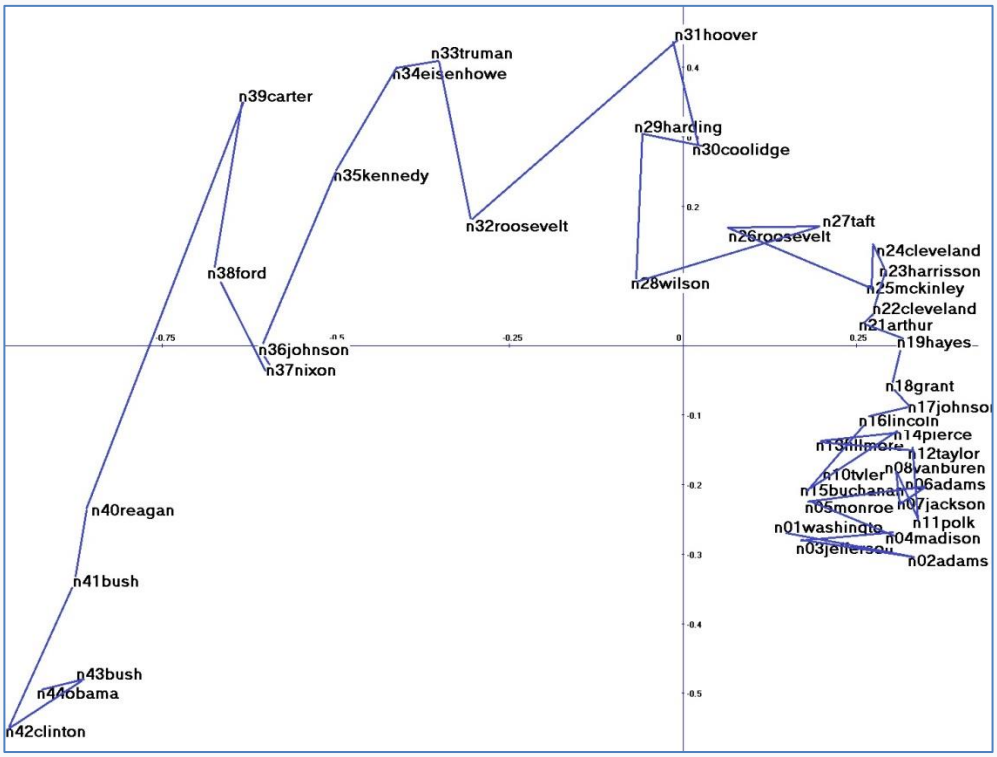

Figure 2. Plan (1, 2) de l'AC de la table lexicale (10 682 x 42) Mots x Présidents.

4.3 Le « facteur taille » de l'ACP.

Le calcul des axes principaux en ACP se fait à partir du point moyen des individus (ici : les mots) dans un espace, mais à partir de l'origine des axes dans l'autre espace. Lorsqu'il existe

une corrélation positive entre tous les couples de variables (ici : les présidents) on obtient un facteur de taille. C'est le fameux « facteur général d'aptitude » (censé mesurer l'intelligence) de Spearman (1904) : certains élèves ont des bonnes notes dans toutes les matières, et la première dimension les opposent à ceux qui ont des mauvaises notes dans toutes les matières (situation schématique largement nuancée depuis). Ici, certains mots sont fréquents chez tous les présidents (partie gauche de la figure 3), d'autres sont rares.

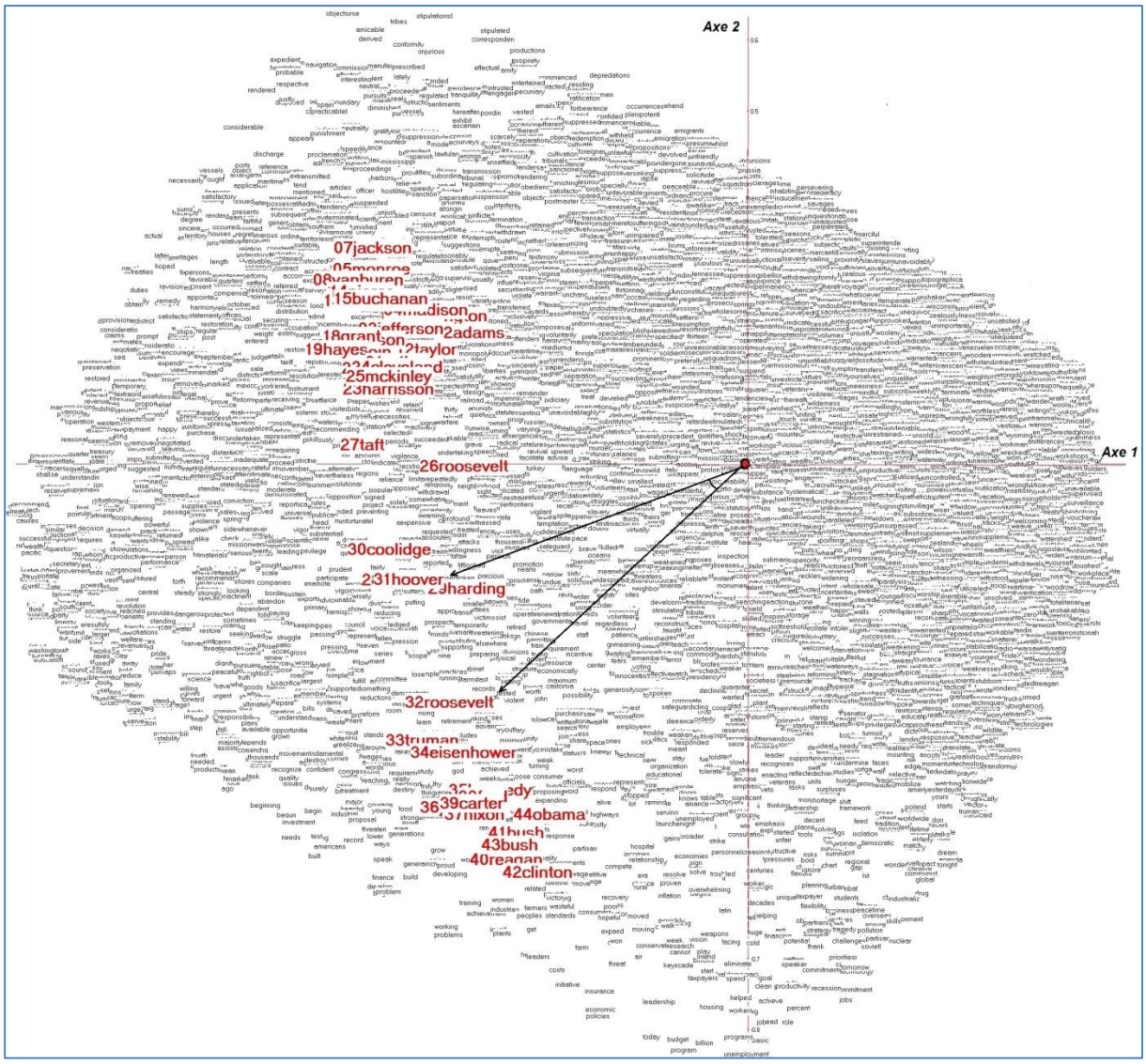

Figure 3: Positionnement simultané des 10 030 mots actifs et des 42 présidents dans le plan (1, 2) de l'ACP. (cette figure ne peut être qu'une esquisse. Elle doit évidemment être grossie pour être lisible).

Dans notre cas (figure 3), l'axe 1 horizontal est un axe de consensus (axe absent d'une AC fondée sur des profils qui sont des fréquences conditionnelles). Cet axe nous dit en première approximation que les présidents parlent tous la même langue (ont en partage la plupart des mots, tout simplement parce que ceux-ci sont fréquents dans la langue), alors que le second axe vertical nous dit qu'ils ne disent pas tous la même chose.
Sur la figure 3, examinons à titre d'exemple les deux vecteurs joignant l'origine des axes aux deux présidents Hoover (23) et Roosevelt (32). Le cosinus de leur angle est une estimation du coefficient de corrélation *r* des vecteurs binaires correspondants, qui est proportionnel (voir

plus haut section 1.1) au χ^2 [calculé sur le tableau 1, où les deux textes sont les deux discours]. Paradoxalement, on lit mieux des tests du χ^2 sur une ACP sur données binaires que sur une AC pourtant fondée sur la distance du χ^2…

Tableau 3. Identification du premier axe de l'ACP par le classement des 42 présidents et les 50 graphies (mots) occupant les positions les plus extrêmes sur cet axe (sur 10 030 mots actifs)

Classement des présidents selon l'axe 1 ("facteur de taille de l'ACP")

Les flèches indiquent la direction gauche -> droite sur l'axe 1

Identifier	axis 1	Identifier	axis 1
19hayes	-626	44obama	-331
08vanburen	-615	42clinton	-347
13fillmore	-601	41bush	-367
16lincoln	-601	43bush	-377
18grant	-601	40reagan	-393
17johnson	-597	38ford	-409
11polk	-596	35kennedy	-423
10tyler	-592	37nixon	-430
14pierce	-592	39carter	-435
05monroe	-587	02adams	-444
07jackson	-586	29harding	-453
15buchanan	-586	36johnson	-463
27taft	-577	26roosevelt	-465
22cleveland	-574	34eisenhower	-477
23harrisson	-573	32roosevelt	-486
25mckinley	-569	12taylor	-497
30coolidge	-566	33truman	-513
24cleveland	-560	04madison	-515
06adams	-552	01washington	-518
03jefferson	-551	31hoover	-529
28wilson	-546	21arthur	-541

Graphies ayant des positions extrêmes sur l'axe 1 ("facteur de taille" de l'ACP)

Coordonnées négatives		Cordonnées positives	
Identifier	axis 1	Identifier	axis 1
thus	-1049	uninsured	469
directed	-1047	prescription	469
treaty	-1044	pell	469
having	-1041	iraqi	469
recommend	-1041	backgrounds	469
effect	-1039	terrorist	469
direct	-1037	ink	469
form	-1036	entitlement	469
subject	-1036	bush	469
either	-1036	basics	469
account	-1036	usa	467
neither	-1034	teen	467
causes	-1034	talented	467
constitution	-1034	stays	467
laid	-1031	saddam	467
successful	-1031	hussein	467
enterprise	-1029	fueled	467
land	-1029	emissions	467
additional	-1029	creativity	467
nothing	-1029	stories	467
influence	-1029	dime	467
sources	-1029	spark	467
labor	-1029	kuwait	467
direction	-1029	targeted	467
pacific	-1027	sanctuary	467

Bien que les présidents occupent tous la moitié négative de l'axe 1, on distingue cependant sur la figure 3 un léger décalage des présidents les plus récents (à partir de Theodore Roosevelt, n°26), vers la droite. Le volet gauche du tableau 3 qui décrit précisément leur classement selon l'axe 1 horizontal confirme cette légère mais significative opposition (opposition exacerbée sur l'axe 2 vertical) (avec cependant une exception, le second président Adams, et, dans une moindre mesure, les présidents Washington et Madison).

Le volet gauche du tableau 3 liste les 25 mots les plus à gauche et les 25 mots les plus à droite sur l'axe 1 de la figure 3 (extrait particulièrement anecdotique compte tenu des 10 030 mots actifs).

Du côté des mots consensuels (colonne de gauche), on trouve comme prévu un vocabulaire peu différencié, alors que la colonne de droite porte l'empreinte des seuls derniers présidents.

La première approximation : axe 1 : « ils parlent la même langue », axes suivants : « ils ne disent pas la même chose » est à réviser : ils ne parlent pas tout à fait la même langue, eu égard au vocabulaire. C'est évident compte tenu de la longueur historique de la période.

4.4 Déclinaison des arbres additifs selon les dimensions

Ce type de déclinaisons n'est pas propre à l'ACP, et concerne toutes les méthodes en axes principaux mentionnées (analyse logarithmique, LSA, AC). Elles ajoutent ici à la clarté d'interprétation des distances sur données binarisées.

La figure 4 présente un arbre additif construit en prenant tous les axes principaux de l'ACP sur données de présence-absence (procédure SplitsTree de Huson et Bryant, 2006, appelée à partir du logiciel DtmVic). Ces données reconstituent exactement la matrice des corrélations qui correspond aux distances d'Evrard. Les proximités s'interprètent donc en termes de coefficients ϕ, donnés par la formule [1] de la section 1, ou en terme de coefficient *r*, donné par la formule [2]. ϕ et *r* sont plus faciles à concevoir, conceptualiser et interpréter qu'une distance du χ^2.

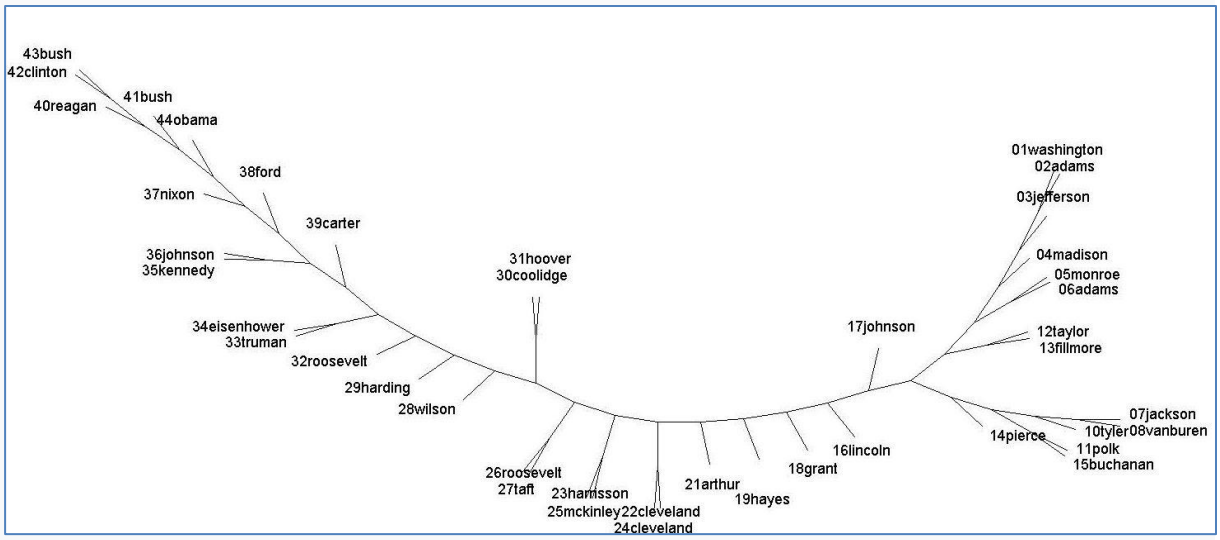

Figure 4. Arbre additif (distances calculées sur les 42 axes de l'ACP sur données binaires) (distances d'Evrard, dérivée de ϕ et r).

La figure 5 nous donne, toujours à titre d'exemple, un arbre similaire, mais la reconstitution de la matrice des corrélations est limitée aux 4 premiers axes, faisant apparaitre une branche spécifique de l'arbre correspondant à une période particulière (période de reconstruction après la fin de la guerre civile, *Gilded Age*, développement industriel, immigrations massives… *)* Cette période correspond aux présidents 18 à 26.

Les déformations de l'arbre additif n'excluent pas la consultation de plans factoriels, mais elles ont sur ceux-ci un avantage considérable : elles résument des espaces ayant plus de deux dimensions, comme de l'exemple de la figure 5 engendré par les 4 premiers axes principaux..

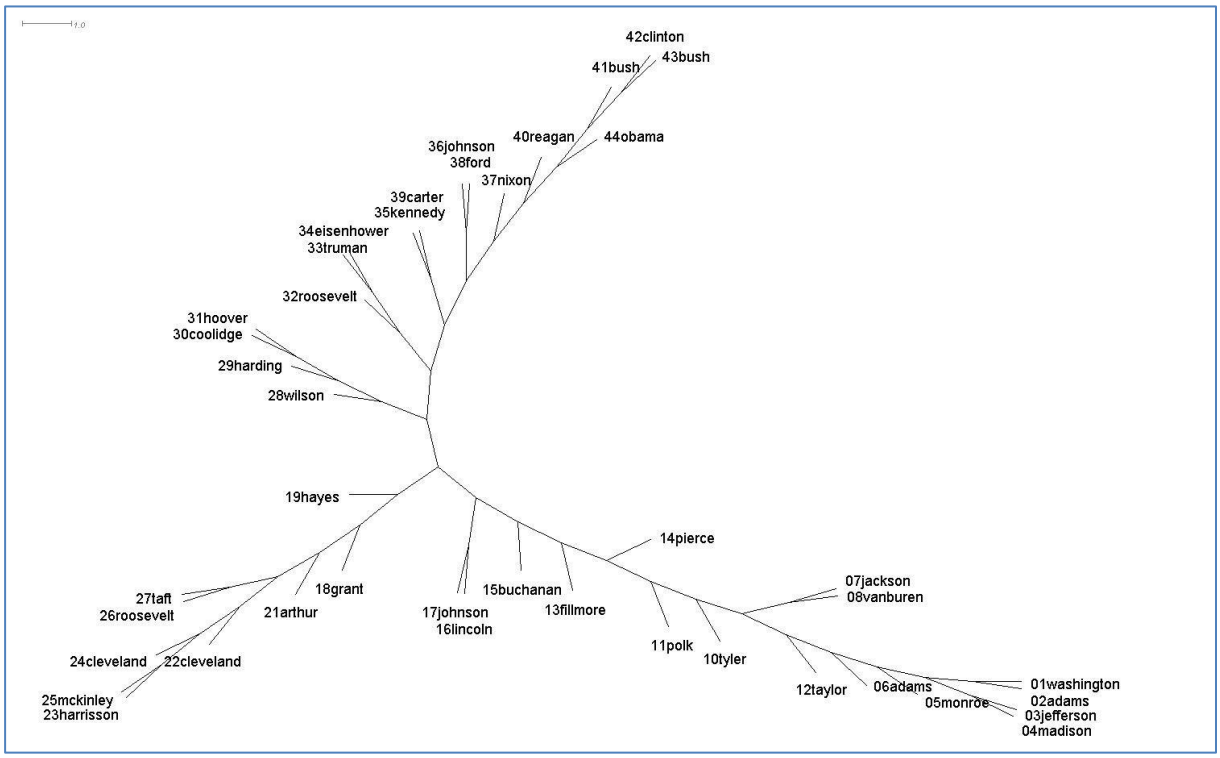

Figure 5. Arbre additif calculé sur les 4 premiers axes de la même ACP, mettant en exergue la période 1870-1910 (bas gauche de la figure)(modulation des distances d'Evrard en fonction des dimensions).

5. Conclusion

L'utilisation des coefficients ϕ et *r*, comme celle du χ^2 [pour tables (2 x 2)] permet de travailler sur les distances désignées comme distances d'Evrard par les linguistes francophones (après les applications pionnières d'Evrard). Au confluent de plusieurs approches statistiques, naturellement liées à l'ACP, ces distances possèdent un pouvoir descriptif et discriminant attesté par de nombreuses applications. La formulation explicite des coefficients assure transparence et qualité de communication des résultats. La mise en œuvre enfin se résume pour l'essentiel à une ACP sur données de présence-absence avec les avantages des implémentations existantes (validation bootstrap, possibilités de variables supplémentaires, synergie avec les méthodes de classification, et en particulier les arbres additifs).

Références

Aitchison J. (1983). Principal Component Analysis of Compositional Data. *Biometrika*, 70, 1,57-65.

Baulieu, F.B. (1989). A classification of presence/absence based dissimilarity coefficients. *Journal of Classification*, 6, 233–246.

Benzécri J.-P. (1973). *L'Analyse des Données*. Tome 1: *La Taxinomie*. Tome 2: *L'Analyse des Correspondances*. Dunod, Paris.

Brunet E. (2011). Les affinités lexicales. Hommage à Etienne Evrard. *Langues anciennes et analyse statistique. Cinquante ans après*, Fialon S., Longrée D., Pietquin P., Rome, Italie. pp.9-31. ⟨hal-01363232⟩.

Brunet E., Lebart L. & Vanni L.(2020). Littérature et intelligence artificielle, in D. Mayaffre et al. (sous la dir.), *L'intelligence artificielle des textes. Points de vue critique, points de vue pratique*, Paris, Champion, Lettres numériques (sous presse).

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K. et Harshman R. (1990). Indexing by latent semantic analysis, *J. of the Amer. Soc. for Information Science*, 41 (6): 391-407.

Escofier B. (1978). Analyse factorielle et distances répondant au principe d'équivalence distributionnelle, *Revue de Statist. Appl.*, vol. 26, n°4: 29-37.

Evrard, E. (1966). Etude statistique sur les affinités de cinquant-huit dialectes bantous. In *Statistique et analyse linguistique: colloque de Strasbour*g, 20-22 Avril, 1964, 85-94. Paris: Presses Universitaires de France.

Greenacre M., Lewi P. (2009). Distributional Equivalence and Subcompositional Coherence in the Analysis of Compositional Data, Contingency Tables and Ratio-Scale Measurements, *Journal of Classification*, Springer, vol. 26(1), 29-54.

Hotelling H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psy.* 24: 417-441 et 498-520.

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

Kazmierczak J.-B. (1985) - Analyse logarithmique : deux exemples d'application. *Revue de Statist. Appl.* , 33, (1), p 13-24.

Lewi P.J. (1976). Spectral mapping, a technique for classifying biological activity profiles of chemical compounds. Arzneim. Forsch. in: *Drug Res.* 26, 1295-1300.

Pearson, K. (1900). Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London*, Series A, 195, 1–47.

Ratinaud, P. (2014). IRaMuTeQ : Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires . http://www.iramuteq.org

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

Salton G., Mc Gill M.J. (1983). *Introduction to Modern Information Retrieval*, International Student Edition., McGraw Hill, New York.

Shannon C.E. (1948). A mathematical theory of communication, *Bell System Technical Journal*, 27 (1948), 379-423, 623-659.

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

Warren M. J. (2008). On association coefficients for 2x2 tables and properties that does not depend on the marginal distributions. *Psychometrika*, 73; 777.

Yule, G.U. (1900). On the association of attributes in statistics. *Philosophical Transactions of the Royal Society A*, 75, 257–319.

Yule, G. U. (1912). On the methods of measuring the association between two attributes, *Journal of the Royal Statistical* Society, *75*, 579-642.