Low lexical frequencies, problems, descriptions and predictions

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.

1. Introduction

The description of tables crosstabulating vocabulary and texts is commonly performed through correspondence analysis (CA), well suited to frequency profiles and lexical tables, thanks to the distributional equivalence property of the chi-2 distance. The AC is then complemented by clustering (including additive trees).

The distances of Evrard (1966) (cf. also Brunet, 2011) derived from the Phi coefficient of Yule-Pearson (1912), are simply based on the presence or absence of words (or lemmas). They provide more meaningful representations for discriminating between texts or making attributions of authors. Brunet et al. (2020) have thus shown from a corpus of 50 novels written by 25 authors of the twentieth century (two novels per author) that a flawless pairing of novels by author could be obtained from the Evrard distance matrix. This matrix is easily deduced from the correlation matrix of binary variables (presence-absence). It is involved in the classic principal component analysis (PCA) of Hotelling (1933) applied to binary data. Section 2 is a brief review of the problems entailed by low frequencies (and large frequency disparities) in exploratory analyzes of text, whereas section 3 deals with some solutions proposed in practice. Section 4 shows, with an example, how PCA can provide a complementary point of view to that of AC, emphasizing the role played by the presence or absence of words, while making it also possible to decline the Evrard distances according to the dimension of the principal space and thus enriches the approaches more widespread in the JADT community.

2. Evrard distance, Pearson-Yule ϕ and Pearson χ^2 :

In statistics, the *phi* coefficient (or ϕ) is a measure of association for two binary variables. Based on the correlation coefficient r of Karl Pearson (1900), this measure was proposed by Yule (1912) who had previously published a similar measure of association (Yule, 1900). This measure is closely related to the chi-square (χ^2) calculated on the same contingency table to

test the independence between rows and columns. It coincides with the Pearson correlation coefficient *r* between two binary variables.

Table 1. (2 x 2) contingency table confronting two texts

		Text 2		
		Present words	Absent words	Total
Text 1	Present words	n_{11}	n_{10}	$n_{1.}$
	Absent words	n_{01}	n_{00}	$n_{0.}$
	Total	$n_{.1}$	$n_{..0}$	n

Two binary variables *x* and *y* are considered positively associated if the data concentrates in the diagonal cells and considered negatively associated if they concentrate outside the diagonal. If we have a 2 × 2 table for two texts, the coefficient ϕ which describes the association of *x* and *y* is given by the formula (Yule, 1912), with the notations of table 1:

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

Note that as early as 1900, Yule proposed the similar formula: $Q_{Yule} = \frac{n_{11}n_{00} - n_{10}n_{01}}{n_{11}n_{00} + n_{10}n_{01}}$.

Cohen (1960) proposed to replace the geometric mean of the denominator of formula [1] by an arithmetic mean:

$$s(1,2) = \frac{2(n_{11}n_{00} - n_{10}n_{01})}{n_{1.}n_{0.} + n_{.0}n_{.1}}$$

We can consult Warren (2008) and Baulieu (1989) for an overview of the flora of the many coefficients of association proposed over the years and disciplines.

2.1 Link of ϕ *with* χ^2 *:*

The square of the coefficient ϕ is linked to Karl Pearson's χ^2 statistic for the same *2 × 2* contingency table by the classical relationship (where *n* is the total number of observations: here number of distinct words).

$$\phi^2 = \frac{\chi^2}{n}, \qquad \text{since we have:} \quad n\phi^2 = \frac{n(n_{11}n_{00} - n_{10}n_{01})^2}{n_{1.}n_{0.}n_{.0}n_{.1}}$$

(classical formula of χ^2 for a 2 × 2 table, with 1 degree of freedom [which therefore has 5 chances out of 100 of exceeding 3.84 under the independence assumption]).

2.2 Equivalence of ϕ with Pearson's r.

The classic Karl Pearson correlation coefficient r calculated on the binary data of the incidence table **X** (Table 2) (licit calculation in the case of two variables with two categories) coincides with the coefficient ϕ.

Table 2. Incidence table **X** *of general term* x_{ij}
(x_{ij} = 1 if the word i [row i] is present in text j [column j])

Words	***Text 1***	***Text 2***
word 1	1	0
word 2	1	1
word 3	0	0
word 4	1	0
…….	….	….
word n	0	1
	$n_{1.}$	$n_{.1}$

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

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

And, for instance for text 1:
$$s_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2 \quad (=\frac{n_{0.}n_{1.}}{n^2})$$

From formula [2] and table 2, we find: $r_{12} = \frac{n_{11}n - n_{1.}n_{.1}}{\sqrt{n_{1.}n_{0.}n_{.0}n_{.1}}}$, and we get again formula [1] in replacing n, $n_{1.}$ et $n_{.1}$ with their values as functions of n_{11}, n_{01}, n_{10} et n_{00}.

This equivalence with the classic test of independence of χ^2 and the identity of $\phi(1,2)$ with the linear correlation coefficient r_{12} give the coefficient ϕ, and therefore the Evrard distance which directly derives from it, a special position among association measures.

2.3 The chi-square distance (χ^2)

The chi-square distance (χ^2 distance) used in correspondence analysis (CA) is an approximation of a measure of mutual information (derived from the theory of Shannon, 1948) evaluating the information provided by an empirical contingency table with respect to the hypothesis independence of rows and columns (see for instance Benzécri, 1973, Tome 1 B n ° 5). This distance shares with a few others (see Escofier, 1978) the property of *distributional equivalence* which ensures stability of results by aggregating rows or columns with the same profiles. CA has become one of the basic tools for describing lexical tables. The χ^2 distance involves however inverses of frequencies which can be problematic in the case of very low frequencies (the adjustment criterion which gives each point a mass equal to its frequency partially compensates for this weakness).

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

3. Low frequencies or frequency discrepancies

In this section, we briefly review three approaches which aim to remedy strong frequency disparities or to involve binary coding of words.

3.1 Logarithmic analysis

Logarithmic analysis (LA) also complies with the distributional equivalence property of CA on arrays of positive numbers. Kazmierczak (1985) based the LA on the principle of Yule (1912) according to which one does not change the distance between two rows or between two columns of a table by replacing the rows and columns of this table by other proportional rows and columns (generalization of distributional equivalence).

LA consists in taking the logarithms of the data (after possible addition of a constant in the case of negative or zero data), then, after having centered them both in rows and in columns, to submit them to an unstandardized principal component analysis (PCA), which coincides in such a case with a singular value decomposition (SVD). If **X** is a (*n, m*) data matrix, and if **A** and **B** are two diagonal matrices respectively of dimensions (*n, n*) and (*m, p*) with positive diagonal elements, the logarithmic analysis of the new array **AXB** coincides with that of **X**. This property of strong invariance, together with the shrinking effect of the logarithm function, makes this technique robust, well suited to applications to massive data, for which the frequency disparities (from 1 to 105 for example) constitute a technical obstacle. In fact, this method dates back to Aitchison (1983) in a different setting. A similar, but not identical, variant had been proposed initially under the name of Spectral Analysis by Lewi (1976), then by Greenacre and Lewi (2009).

3.2 TF-IDF coefficients and LSA

The general term of the (words x texts) lexical table can be replaced by the coefficient TF-IDF (Salton and McGill, 1983). Recall that the coefficient TF-IDF (Term frequency x Inverse of Document frequency) is the product of the frequency of a term (TF) by the logarithm of the quotient: "total number of documents / number of documents in which the term is present ". This quotient (IDF) therefore involves the inverse of the proportion of documents in which the term appears. The logarithm, as with the LA mentioned above, helps to cushion extreme situations, such as when the term is only present in one document out of thousands. In other words, the TF-IDF coefficient combines an indicator of the dominance of the term (TF component) with an indicator of its specialization in the corpus (IDF component), the latter indicator varying from 0 (the term is in all the documents) to a maximum (when the term is in a single document) which depends on the size of the set of documents.

In the context of documentary research, the aim here is to find one or more documents in a database (short and numerous) using a few terms. One must penalize the documents which do not contain these terms (element TF in the formula). If we denote by *d* the number of documents, *d (i)* the number of documents which contain the word *i*, by f_{ij} the frequency of the word *i* in the document *j*, $f_{i.}$ the total frequency of word *i*, and $f_{.j}$ the total frequency of document *j* , we have :

- frequency of term *i* in document j: TF (*i,j*) = ($f_{ij} / f_{.j}$).

- logarithm of the inverse of the frequency of documents containing the term i: IDF (*i,j*) = log(*d / d(i)*).

Like CA and LA, *Latent Semantic Analysis* (LSA) [or *Latent Semantic Indexing* (LSI)] (Deerwester et al., 1990) is a decomposition into singular values (SVD) of a transformed lexical table.

Here, it will be the general term TF-IDF coefficient matrix:

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

We also show that the AC can be deduced from the SVD of the general term matrix:

$$w(i, j) = \frac{f_{ij}}{\sqrt{f_{i.} f_{.j}}} \text{ which can be written: } w(i, j) = \frac{f_{ij}}{f_{.j}}\left(\sqrt{\frac{f_{.j}}{f_{i.}}}\right) \qquad [4]$$

The formulas [3] and [4] differ by the factors represented by their right parentheses which both penalize the words *i* very frequent in the corpus: by the number of documents *d(i)* which contain them for *t(i, j)*, by their overall frequency $f_{i.}$ for *w (i, j)*. The concepts of number *d* of documents, and of number *d(i)* of documents containing a word *i* are especially operative for numerous and short documents.

3.3 Alceste methodology

Reinert (1983) proposed to create new statistical units in a text corpus. Such corpus is divided into Elementary Context Units (ECU) having similar lengths (for example 20 consecutive words, one or more lines of 120 characters, a sentence). The analysis of these new units is the basis of a procedure known as ALCESTE.

This methodology is implemented in the ALCESTE and IRaMuTeQ software (Ratinaud, 2014). As long as we are working on short fragments, all word frequencies are low within a fragment, and the presence or absence of a term can be taken into account. In this case, the binary coding occurs after transformation of the corpus.

General remark:

We have seen that low frequencies occur naturally in short texts, whether they are documents or summaries in a database, fragments or units of context, pages of novels, or even answers to open questions. Presence-absence coding is an acceptable and empirically proven option. It can also be modulated by thresholding ("present" if more than *s* occurrences, for example). On the other hand, for applications to large texts in volume, coding the presence or absence of a word is a deliberate option which provides a specific point of view on the texts of a corpus, complementary to the global processing of original frequencies.

4. Illustrative example

To show the relevance of presence-absence coding, and of the use of PCA in this case, we will use the classic STATE OF THE UNION corpus which brings together the speeches on the State of the Union delivered by the American presidents. Americans in office before Congress, from George Washington (1790) to Barack Obama (2009) [42 speeches]. The (up-to-date) corpus is available at http://stateoftheunion.onetwothree.net/index.shtml, also accessible from the *nltk* site (Natural Language Tool-kit, cf.: http://www.nltk.org/nltk_data/, topic: C-Span Inaugural Address Corpus).

The corpus used here comprises 1,746,702 occurrences and 25,246 distinct words. (It can also be downloaded from the "Complementary material" button on the site: https://www.puq.ca/catalogue/livres/analyse-des-donnees-textuelles-3651.html). For this methodological example, we will work on the tokens of the plain text (without lemmatization).

We are talking here about illustration rather than application because this corpus is meant as a benchmark allowing comparisons and is not an object of study in itself. Its strong chronological structure means that other methodologies can be applied with profit, and the problematic authorship of certain speeches would require interpretative precautions which go beyond the present example.

The process of global description of the corpus after coding in the form of presence-absence of words will be schematized by a few graphs.

4.1 Principal component analysis of the presence-absence table (figure 1)

Table 1 presented in section 1 now has 42 columns (Presidents) and 10,030 rows (tokens). The lines must have at least two 1s (presence) (this eliminates the hapaxes) and at least two 0s (absence) (we eliminate the terms used in all the texts or absent in a single text). Such trimming has the effect of reducing the size of the table by removing most of tool-words (or function words) and auxiliaries, as well as a lot of common terms. The loss of raw information can seem considerable. But the only information that interests us at this point is what description tools can use.

The parabolic shape of the sequence of presidents in Figure 1 is not just a pure *Guttman effect* (or horseshoe effect). The 10,030 words cloud does not follow this shape, and the area inside the curve contains many words common to extreme periods.

We present the plane of axes 2 and 3 to ensure a comparison with the plane (1, 2) of the correspondence analysis which follows. We will study separately the first axis of the PCA (so-called "size" factor) below.

4.2 Comparison with the CA of the entire lexical table (figure 2)

Figure 2 shows the principal plane (1, 2) of a correspondence analysis of the original lexical table, larger than the table with "presence-absence" coding.

The sequence of the first twenty presidents (right part of the figure) is less clearly represented in this space.

4.3 The "size factor" of the PCA.

The origin of the principal axes in PCA is the mean-point of the individuals (here: the individuals are the words) in one space, but it is not the mean-point of the variables in the other space. When there is a positive correlation between all the pairs of variables (here: the presidents) we obtain a "size factor". This is the famous "general aptitude factor" (supposed to measure intelligence) described by Spearman (1904): some students have good marks in all subjects, and the first dimension pits them against those who have bad marks in all subjects (schematic situation largely discussed since). Here, some words are common among all presidents (left part of Figure 3), others are rare.

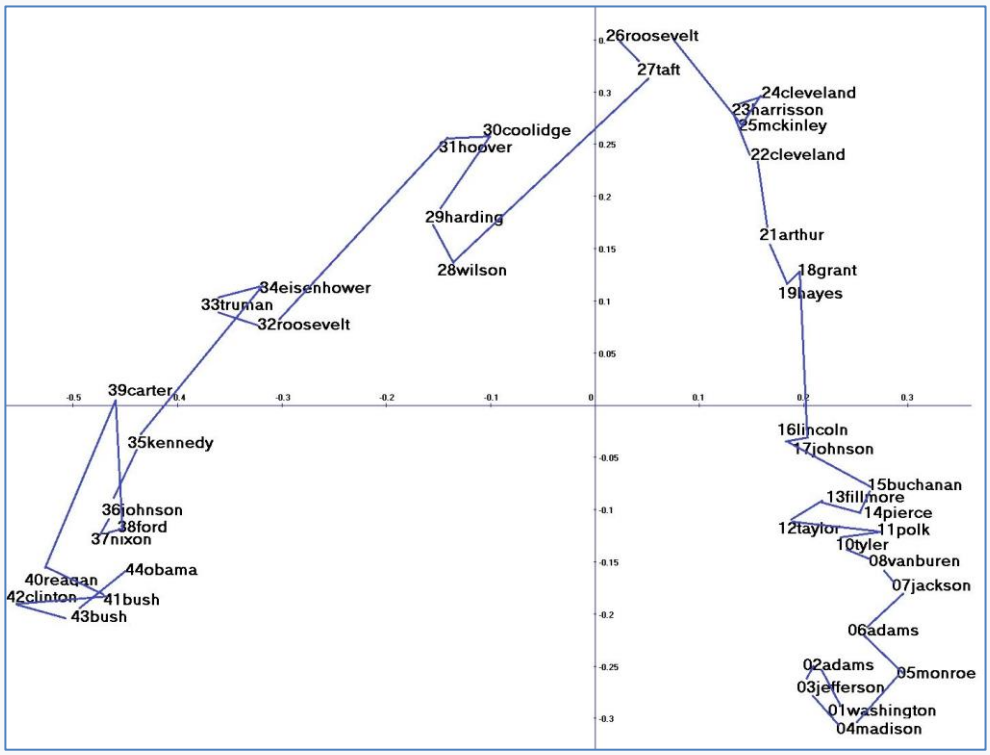

Figure 1. Plane (2, 3) of the PCA of the table (10030 x 42) Words x Presidents.

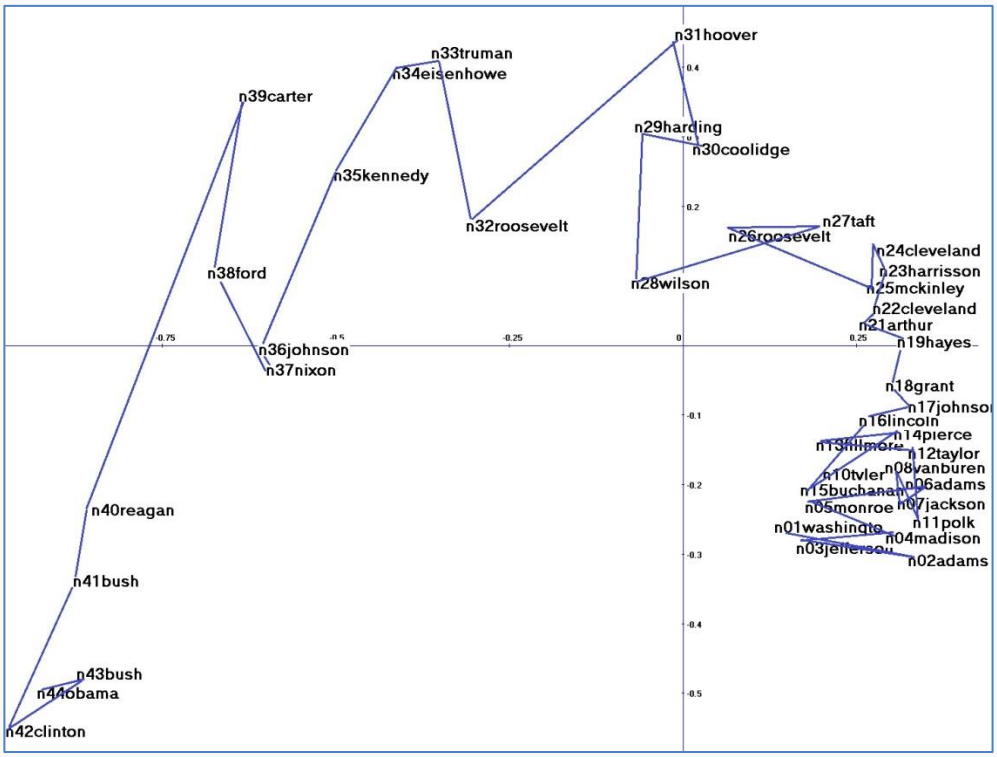

Figure 2. Plane (1, 2) from the CA of the lexical table (10 682 x 42) Words x Presidents.

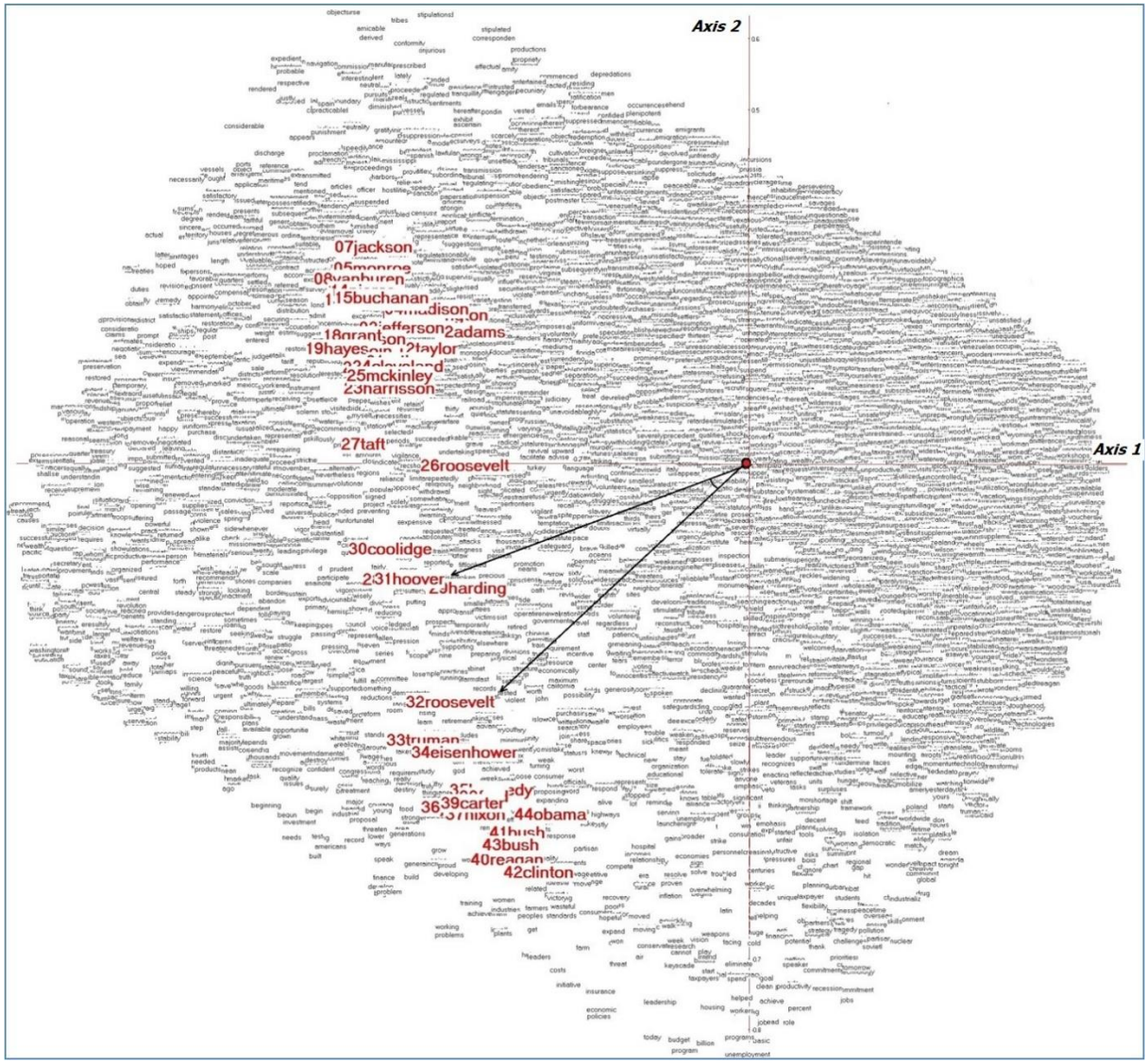

Figure 3: Simultaneous positioning of the 10,030 active words and the 42 presidents in the plane (1, 2) of the PCA. (This figure can only be a sketch. Obviously, it must be strongly magnified to be readable).

In our case (Figure 3), horizontal axis 1 is a consensus axis (axis absent from a CA which is based on profiles that are conditional frequencies). This axis tells us as a first approximation that the presidents all speak the same language (share most of the words, quite simply because these are frequent in the language), while the second vertical axis tells us that they do not all say the same thing.

In Figure 3, consider as an example the two vectors joining the origin of the axes to the two presidents Hoover (23) and Roosevelt (32). The cosine of their angle is an estimate of the correlation coefficient *r* of the corresponding binary vectors, which is proportional (see section 1.1 above) to the χ^2 [calculated in a table such as Table 1, where the two texts are the two discourses]. Paradoxically, we read more easily χ^2 tests on a PCA on binary data than on an CA essentially based on the distance of χ^2.

Table 3. Identification of the first axis of the PCA by the ranking of the 42 presidents and the 50 tokens (words) occupying the most extreme positions on this axis (out of 10,030 active words…)

Ranking of presidents according to axis 1
("Size factor" of the PCA)
Arrows indicate the left-right direction on axis 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

Extreme tokens on axis 1
("Size factor" of the PCA)
Arrows indicate the left-right direction on axis 1

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

Although the presidents all occupy the negative half of axis 1 in Figure 3, we can see however a slight shift from the more recent presidents (from Theodore Roosevelt, # 26), to the right. The left part of Table 3 which describes their ranking according to horizontal axis 1 confirms this slight but significant opposition (exacerbated opposition on vertical axis 2) (with, however, one exception, the second president Adams, and, to a lesser extent, Presidents Washington and Madison).

The right part of Table 3 lists the 25 leftmost words and the 25 rightmost words on axis 1 of Figure 3 (a particularly small extract from the 10,030 active words).

As for consensual words (left column of right part of the figure), we find, as expected, a poorly differentiated vocabulary, while the right column bears the imprint of only the last presidents. The first approximation: axis 1: “they speak the same language”, following axes: “they do not say the same thing” needs to be revised: they do not speak quite the same language, given the vocabulary. This is evident given the historical length of the period.

4.4 Modulations of additive trees according to dimensions

This type of variation is not specific to PCA, and concerns all the principal axes methods mentioned (logarithmic analysis, LSA, CA). They contribute here to the clarity of interpretation of distances on binarized data.

Figure 4 presents an additive tree (AT) constructed by taking all the main axes of PCA on presence-absence data (SplitsTree procedure by Huson and Bryant, 2006, called from the DtmVic software). These data exactly reconstruct the correlation matrix that corresponds to the Evrard distances. The proximities are therefore interpreted in terms of coefficients ϕ, given by formula [1] of section 1, or in terms of coefficient *r*, given by formula [2]. ϕ and *r* are easier to conceive, conceptualize and interpret than a χ^2 distance.

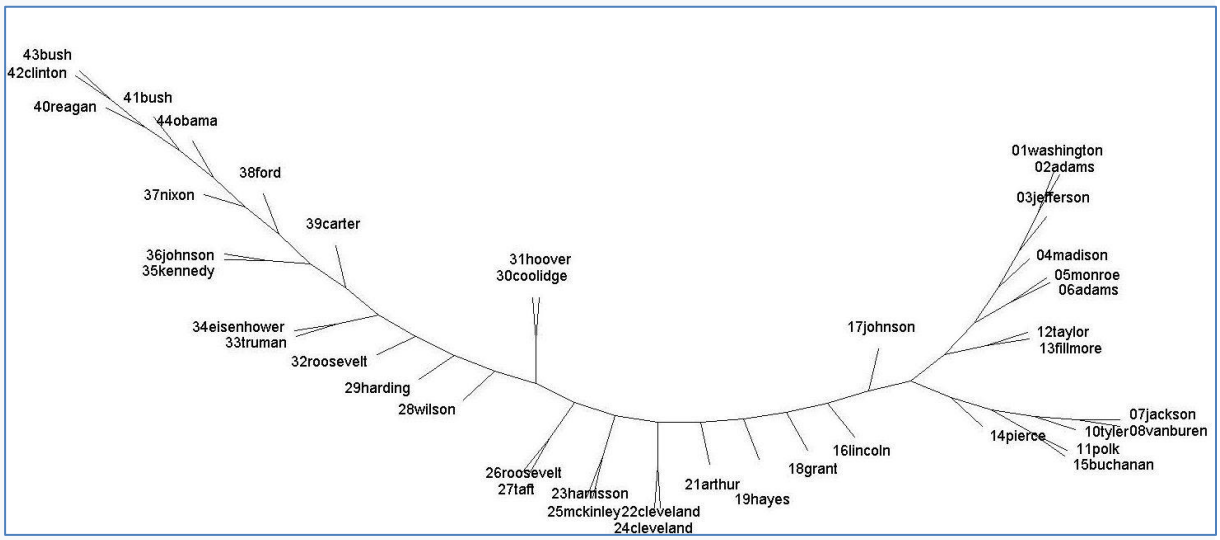

Figure 4. Additive tree (distances calculated on the 42 axes of the PCA on binary data) (distances from Evrard, derived from ϕ and r).

Figure 5 gives us, for example, a similar tree, but the reconstruction of the correlation matrix is limited to the first 4 axes, showing a specific branch of the tree corresponding to a particular period (reconstruction period after the end of the civil war, or *Gilded Age*, industrial development, massive immigration…) This period corresponds to presidents 18 to 26.

The deformations of the additive tree do not exclude the consultation of factorial planes, but the AT has a considerable advantage over them: it summarizes spaces having more than two dimensions, as in the example of figure 5 generated by the 4 first main axes.

.

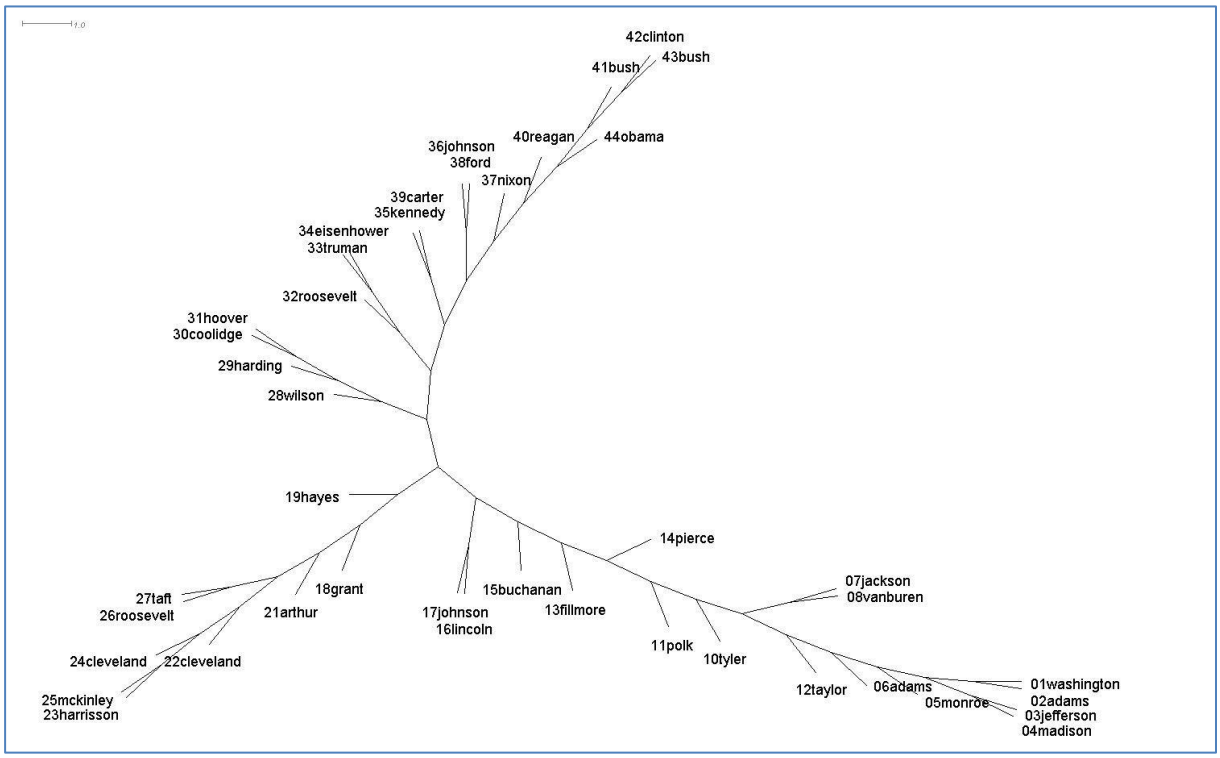

Figure 5. Additive tree calculated on the first 4 axes of the same PCA, highlighting the specificity of the period 1870-1910 (bottom left of the figure)
(modulations of Evrard distances according to the number of kept axes).

5. Conclusion

The use of the coefficients ϕ and *r*, like that of χ^2 [for (2 x 2) tables] makes it possible to work on the distances designated as Evrard distances by French linguists (after the pioneering applications of Evrard). At the confluence of several statistical approaches, naturally linked to PCA, these distances have a descriptive and discriminating power attested by numerous applications. The explicit formulation of the coefficients ensures transparency and quality of communication of the results. Finally, implementation essentially reduces to a PCA after building the presence-absence data from the original corpus, with the advantages of existing implementations (bootstrap validation, possibilities of supplementary variables, synergy with clustering methods, particularly with additive trees).

References

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.