Rotated Canonical Correlation Analysis for Multilingual *Corpora*

Simona Balbi, Michelangelo Misuraca

Dipartimento di Matematica e Statistica, Università "Federico II"

Via Cinthia – Complesso Monte Sant'Angelo – 80126 Napoli – Italy

Abstract

This paper aims at proposing the joint use of Canonical Correlation Analysis and Procrustes Rotations (RCA), when we deal with a text and its translation into another language. The basic idea is representing words in the two different natural languages on a common reference space. The main characteristic of this space is to be language independent, although Procrustes Rotation is performed transforming the lexical table derived from translation by minimizing its distance from the lexical table belonging to the original *corpus*, while the subsequent Canonical Correlation Analysis treats symmetrically the two word sets. The most interesting RCA feature is building a unique reference space for representing the correlation structure in the data, inducing the two systems of canonical factors to lie on the same space. These graphical representations enables us to read distances between corresponding points in terms of different way of translating the same word in relation with the general context defined by the canonical variates. Trying to understand the distances between matched points could represent an useful tool for enriching lexical resources in a translation procedure. In this paper we propose the comparison of the most frequent content bearing words in the two languages, analyzing one year (2003) of *Le Monde Diplomatique* and its Italian edition.

Keywords : multilingual corpora, Procrustes rotations, Canonical Correlation Analysis.

1. Introduction and motivations

The globalization of economy, together with the globalization of our everyday life, makes the problem of dealing with multilingual textual information a common situation.

As a consequence, there is a strong interest in developing computational and analytical tools, in order to automate the joint use of documents in different languages or to translate a document into another language. An interesting research field consists in making possible language-independent queries in search engines, or in international companies developing text mining procedures able to discover knowledge in multilingual textual bases.

Researchers working in the field of Textual Data Analysis have already proposed some solutions in order to develop some language-independent tools, often based on statistical methods developed for solving numerical problems.

Aim of this paper is to face a peculiar problem, when we deal with a text and its translation into another language. Here we come back to a previous proposal (Balbi and Esposito, 1998) in order to revise it, taking into account about ten years of developments in literature and in technical solutions, as in the society. Briefly, we propose the joint use of two different techniques, Procrustes rotations and Canonical Correlation Analysis, as basically in Lafosse (1989), in order to perform a canonical analysis in a language-independent space, trying to show how this can be useful in evaluating a translation, by identifying "critical" words.

The proposed instrument seems to be useful also in monolingual problems, when it is not possible (or useful) to pre-treat the documents for avoiding the cases of lexical ambiguity.

An application to an international newspaper collection will show the effectiveness of our proposal.

2. The data structure

In analyzing a set of documents, and its translation into another language, we can principally deal with two different kinds of situation : we can find an official translation (e.g. by United Nations, EU, or official bilingual countries such as Canada), or an "unofficial" translation, specially dealing with e-documents.

Technically, the main difference is the data structure. In the first case we have "parallel" *corpora*, and it can be interesting to consider documents as "aligned" ones, in the second case we have "comparable" *corpora*, focusing the analysis on a global evaluation of the translation process.

We direct our attention to comparable *corpora*, but imposing a word-by-word structure, by representing words belonging to different vocabularies in a common language-independent reference space. The aim is to identify the presence of different systems of synonyms, and enhance the richness of one language with respect to the other one in some peculiar topics.

The data structure consists of two matrices, **X** (*n*,*p*) and **Y** (*n*,*p*). **X** is the lexical table having in row the *n* documents in which the original *corpus* is organized, and in columns some *content bearing words* (as in Widdows et al., 2003), selected among the most frequent terms in the original *corpus*. **Y** is the lexical table having in row the same corresponding *n* documents, and in columns the corresponding *content bearing words* selected as the more natural translation in the translated *corpus*.

3. Theoretical framework

The analysis of multilingual *corpora* has been discussed both in Textual Statistics and in Text Mining with different goals. Particularly in Textual Data Analysis has been faced up to deal with the open-ended questions in multinational surveys, by visualizing in a same referential space the different association structures (Lebart, 1998) ; in Information Retrieval has been faced up to develop language-independent representations of terms, in the frame of natural language and machine translation applications (Grefenstette, 1998).

Latent Semantic Indexing (LSI, Deerwester et al., 1990) and other Principal Components related approaches are commonly used to obtained lower dimensional representation for documents. Being based on the decomposition of an association structure in the word set, they are able to infer semantic relations, such as synonymy (if words co-occur often in documents, they are highly correlated, and their proximity is represented onto a dimension in the factorial space). Factorial spaces (i.e. lower dimensional spaces spanned by Principal Components, linear document/word combinations) can be interpreted as topic spaces.

When we deal with a set of document and its translation into another language, we can be interested in finding a reference space where comparing if the previously defined semantic relations are preserved in the translation process, or better, to understand the behavior of some words and their corresponding translations. Furthermore, it could be interesting to have some measures, in order to evaluate the quality of the overall translation process (Balbi and Misuraca, 2005). Canonical Correlation Analysis (CCA, Hotelling, 1936) have been often

proposed for dealing with documents and their translation, mainly in its kernel versions (Hardoon et al., 2003), or in other cross-language features in text mining procedures.

In this direction, a promising research path concerns the Cross-Language LSI (Littman et al., 1998), developed for retrieving documents written in different languages. Procrustes CL-LSI has been envisaged as a further development.

3.1. The classical Canonical Correlation Analysis

CCA can be seen as the problem of finding basis vectors for two sets of variables such that the correlation between the projections of the variables onto these basis vectors are mutually maximized. Canonical Correlation Analysis and Principal Component Analysis (PCA) share similar properties and computational tools, but they differ because PCA considers interrelationships *within* a set of variables (here, the analyzed *corpus*), while CCA focuses the relationships *between* the two groups of variables (here, the two multilingual comparable *corpora*) (Mardia, et al., 1979).

Given two matrices **X** (*n*,*p*) and **Y** (*n*,*q*) containing two different sets **x** and **y** of *p* and *q* variables observed on the same *n* units, the aim is to find the linear combinations :

$$\eta=\mathbf{a}^{\mathrm{T}}\mathbf{x} \quad \text{and} \quad \xi=\mathbf{b}^{\mathrm{T}}\mathbf{y}$$

such that η and ξ are maximally correlated. The vectors $\mathbf{a}_i$ and $\mathbf{b}_i$ are the *i*-th *canonical factors*, while η_i and ξ_i are the *i*-th *canonical correlation variables*, where *i* depends on the dimension chosen for the canonical subspace

From a computational viewpoint, CCA consists in finding the eigenstructure of the matrix :

$$\mathbf{V}_{XX}^{-1}\mathbf{V}_{XY}\mathbf{V}_{YY}^{-1}\mathbf{V}_{YX}$$

where $\mathbf{V}_{XX}$ is the covariance matrix of the **x** variables, $\mathbf{V}_{YY}$ is the covariance matrix of the **y** variables, $\mathbf{V}_{XY}$ is the cross-covariance matrix between **x** and **y**, and $\mathbf{V}_{YX}=(\mathbf{V}_{XY})^{\mathrm{T}}$.

In CCA, the choice of scaling is arbitrary. Therefore, in order to deal with canonical factors of unit variance (and uncorrelated), the following orthonormalization constraints are posed :

$$\mathbf{a}_i^{\mathrm{T}}\,\mathbf{V}_{XX}\,\mathbf{a}_i = \mathbf{b}_i^{\mathrm{T}}\,\mathbf{V}_{YY}\,\mathbf{b}_i = 1$$

Note that there is no assumption of causal asymmetry in the mathematics of CCA. An interesting application of (kernel) CCA for automatically extending lexical resources in a bilingual *corpus* is in Widdows et al. (2003).

3.2. Procrustes Analysis

In statistical literature, when the aim is to investigate the agreement between totally or partially paired tables, different versions of Procrustes Analysis are carried out (Gower, 1975). Procrustes Analysis has been often utilized in the geometric frame of multidimensional data analysis, for comparing factorial configuration obtained analyzing different data sets.

Specifically, aim of this technique is to compare two sets of coordinates by optimizing a goodness-of-fit criterion, i.e. by translating, rotating (and in case reflecting), and dilating one configuration, in order to minimize its distance from the other one.

Given two configurations **X** and **Y** of *n* points in a *p*-dimensional space, conventionally shifted to the origin, the best rotation (in a *least square* sense) of **Y** to **X** is obtained by considering the Polar Decomposition of $\mathbf{Z} = \mathbf{Y}^T\mathbf{X}$:

$$\mathbf{Z} = \mathbf{U}\Lambda\mathbf{V}^T$$
with the constraints
$$\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$$

in which **U** and **V** are square matrices of order *p*, having in columns the left and right singular vectors of **Z**, respectively, normalized to 1. From the previous decomposition it is possible to derive the so called rotation factor $\mathbf{R} = \mathbf{U}\mathbf{V}^T$. It is possible to prove that :

$$\mathbf{R} = \mathbf{U}\mathbf{V}^T = \mathbf{Y}^T\mathbf{X}\,(\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X})^{-1/2}$$

3.3. Orthogonal Procrustes Rotations and Rotated Canonical Analysis

Lafosse (1989) proposes a joint use of Procrustes rotations and CCA in order to analyse similarities, dealing with repeated measurements of some variables on the same cases. The basic idea is to combine the Procrustes rotation objective of finding the "maximal agreement" by adapting one matrix to the other one, chosen as reference, with CCA's aim of finding the main structure of similarity, enabling to represent how the two configuration differ.

In the following we refer to this method as Rotated Canonical Analysis (RCA). Note that RCA works on totally paired matrices.

Given two matrices **X** (*n*,*p*) and **Y** (*n*,*q*), shifted to the origin and standardized, the best rotation of **Y** to **X** is given by **YR**, where after elementary algebraic operations :

$$\mathbf{R} = \mathbf{V}_{YX}\left(\mathbf{V}_{XY}\mathbf{V}_{YX}\right)^{-1/2}$$

The most interesting consequence of RCA (Balbi and Esposito, 2000) consists in building a unique reference space for representing the correlation structure in the data, because if we perform a CCA after rotating one configuration to the other, we induce the two systems of canonical factors to lie on the same space.

Computationally, as $\mathbf{V}_{XY}\mathbf{R} = \mathbf{V}_{YX}\left(\mathbf{V}_{XY}\mathbf{V}_{YX}\right)^{1/2}$, we decompose a symmetric matrix and its spectral decomposition is given by :

$$\mathbf{V}_{XY}\mathbf{R} = \sum_{j=1}^{p} \zeta_j \mathbf{u}_j \mathbf{u}_j^T$$

where ζ_j is the *j*-th singular value of both $\mathbf{V}_{XY}$ and $\mathbf{V}_{YX}$.

As we consider two standardized sets of variables, an Euclidean metric is considered in the decomposition, and a $(\mathbf{X}^T\mathbf{X})^{1/2}$ Mahalanobis metric is introduced in projecting the variables on the common subspace, for normalizing the axes.

4. Rotated Canonical Analysis for a bilingual *corpus*

In this work we propose the joint use of Procrustes Rotation and Canonical Correlation Analysis, in order to represent the main consequence of a translation process. As previously stated both methods are independently used in Cross-Language Information Retrieval.

The motivation of this strategy relies on the basic idea that it is important to develop specific tools related to the peculiar objective of the analysis. As our interest lies on lexical tables

related to the same documents written in two different languages, the correlation of words belonging to different sets is investigated, by representing them in a language invariant space.

By setting the analysis in a proper geometrical framework, the definition of a common plane in which displaying the similarities between the plots of words becomes relevant. These graphical representations enables us to read the above mentioned similarities in terms of different way of translating the same terms in relation with the general context defined by the canonical variates. Moreover, trying to understand the distances between matched points could represent an useful tool for enriching lexical resources in a translation procedure.

5. The analyzed bilingual *corpus* : *Le Monde Diplomatique*

Le Monde Diplomatique (LMD) is a monthly review edited in France since 1954, characterized by a critical viewpoint dealing with worldwide economical, political, social and cultural issues. At present, LMD is published in 20 different languages and distributed in about 30 countries, on paper or electronic support.

Regarding the Italian edition, the editorial staff, together with the translation from the French edition, draws up some book reviews. The language is quite homogeneous because the same few persons translate the revue from French. Since 1998, an electronic edition is available on the website *http://www.ilmanifesto.it/MondeDiplo/*, together with a paper edition sold as monthly supplement of the newspaper *Il Manifesto*.

The two *corpora* we deal with are a subset of 240 articles published in 2003 on the French edition and the corresponding articles in the Italian edition. The whole collection has been downloaded from the newspaper website and it has been automatically converted in text format with a script written in Java language. For each article, only the body has been considered (i.e. titles and subheadings have been ignored). The articles have been normalised in order to reduce the possibility of data splitting, for example by converting all the capital letters to the lower case or conforming the transliteration of words coming from other alphabets, mainly proper nouns, or using the same notations for acronyms or dates.

After the pre-treatment step two vocabularies of about 2400 terms for each language have been obtained. We decide to deal with *graphical forms*, in the sense of Lebart and Salem (1994), because a lexicalization and a lemmatization in this case can lead to incomparable vocabularies. Two separate Correspondence Analysis (CA), on the French and Italian lexical tables, have been performed in order to evaluate the presence of a common structure between the two separate monolingual corpora. Results are very similar in terms of explained inertia (Balbi and Misuraca, 2005).

We have selected the 60 most frequent content bearing words in the French vocabulary and the corresponding main translations in the Italian vocabulary, building two totally paired lexical tables.

After rotating the Italian table to the French, as explained in § 3.3, we perform a spectral decomposition on the covariance matrix $\mathbf{V}_{XY}\,\mathbf{R}$, cross-tabulating the French words and their Italian translation.

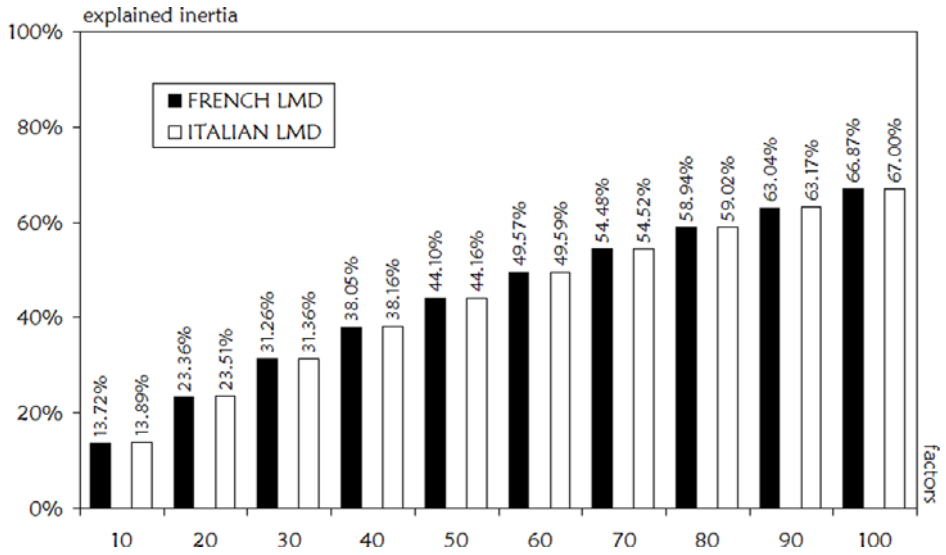

Figure 1. Explained inertia distribution in the CA on French and Italian corpora.

The first factorial axis can be seen in the frame of a domain opposition between the words belonging to the "people" and the words belonging to the "establishment" (Figure 2). For this reason it seems to be more interesting to observe further factorial axes for discovering possible mistranslation.

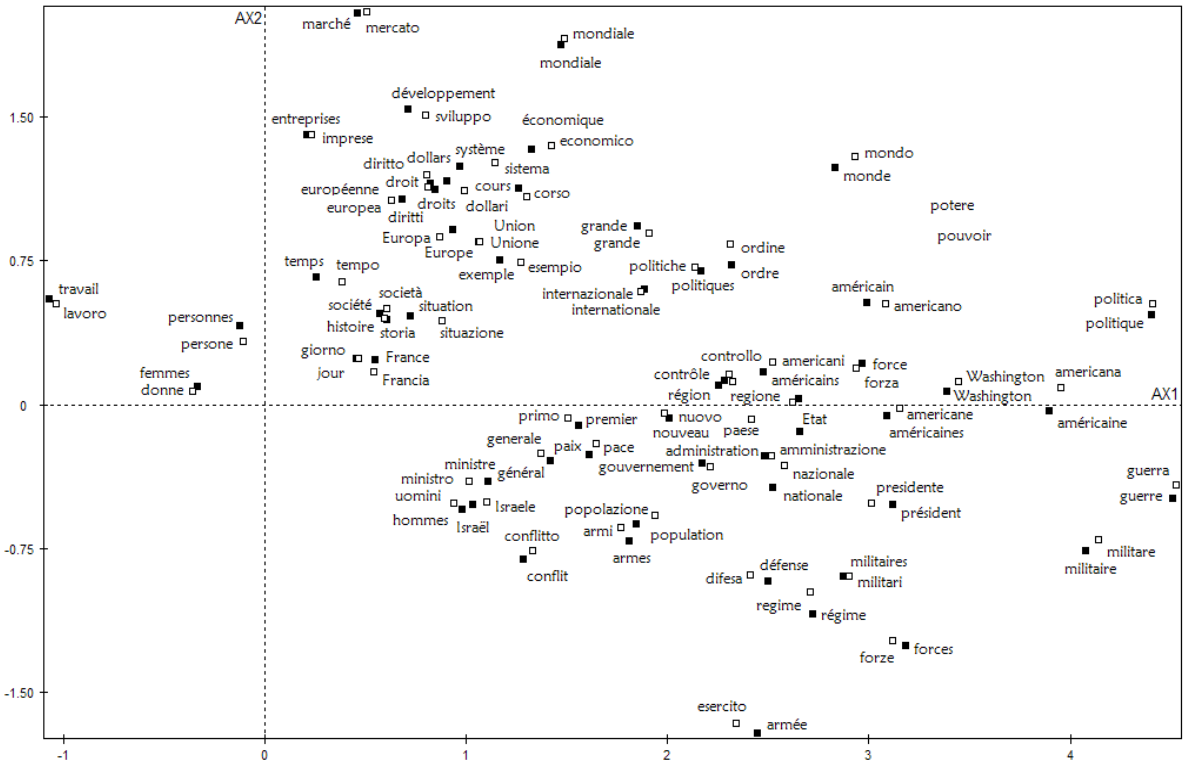

Figure 2. Joint words space : first and second factorial axes.

In Figure 3 we show the joint word factorial map obtained by crossing the second and the third axes. It is possible to see that on the second axis we have an opposition between the words related with the "market" and "war" issues, while on the third axis we have an opposition between the words related with the internal policies and the foreign affairs issues.

When a word is correctly translated the two points are very close. This can be proved also by considering a "bad translation" as *paese* (country) with respect to *Etat* (state). The two concepts are sometime considered as synonym but in different contexts they assume a different meaning. As a consequence the two words are no so close.

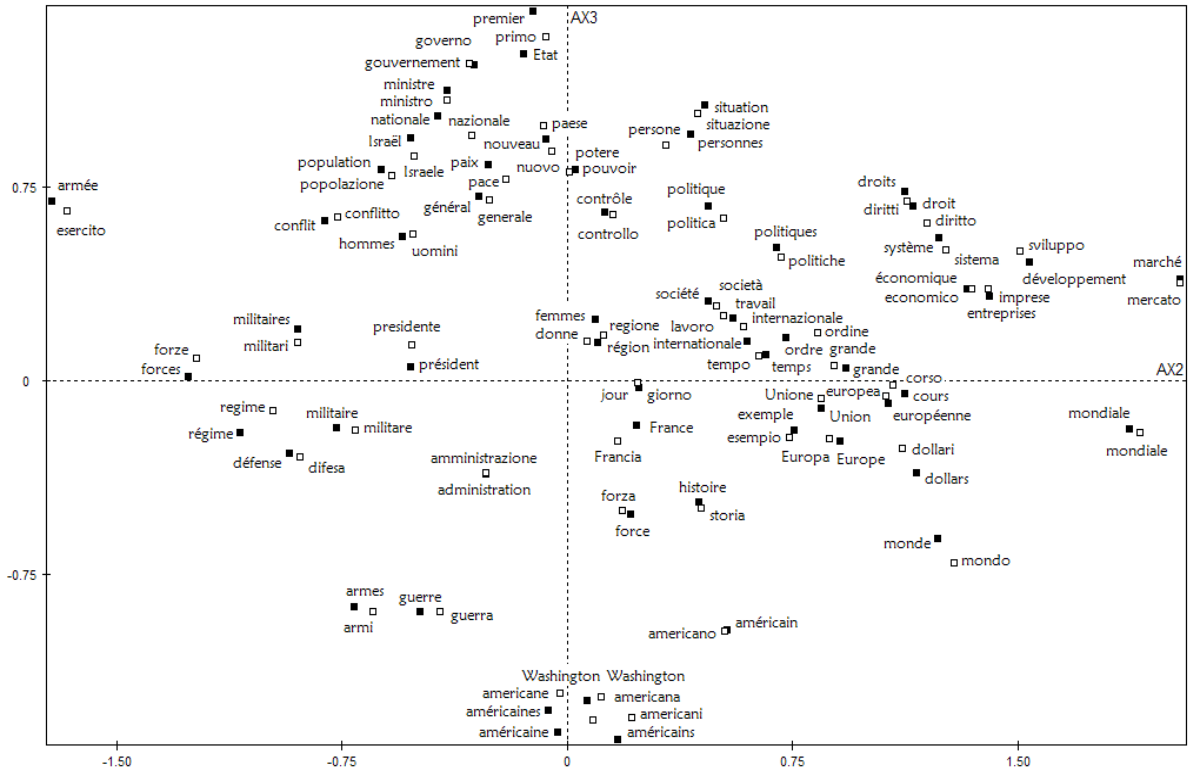

Figure 3. Joint words space : second and third factorial axes.

It is interesting to underline that on this graphical representation (Figure 3) we can evaluate the possibility of lemmatizing some words (the flexion on the word *american* on the lower side of the map) and performing a lexicalization, in terms of minimal sense unit, for visualizing specific contexts and avoiding ambiguous untagged words (e.g. *economical development*).

6. Conclusions and perspective

This paper has been thought in order to show how an intensive use of factorial methods and graphical representation tools can be useful in solving problems connected with the translation in multilingual data set.

Our experiment, related to the edition of a review in two different languages, shows how referring to a language-independent canonical space can be useful in relating a word and its translation in a contextual frame defined by the canonical variables.

We think of this paper as a first attempt to this approach. Further developments can be related to the choice of a proper metrics, or a proper weighting system for words. Here we have considered the lexical tables as intensity tables, with the consequent choices of using Euclidean and Mahalanobis distances in measuring distances. It is well known how a proper choice of the metric, and specifically, a proper way of weighting terms can be important in any textual data analysis.

Moreover, in the original Balbi and Esposito's proposal (2000), external variables are introduced in order to enrich the analysis, by interpreting distances of matched points in terms of external information.

This can be a further research topic to be developed, together with the introduction of more than two languages, following again Balbi and Esposito (1998), by means of Generalised Canonical Correlation and Generalised Procrustes analyses (Gower, 1975).

References

Balbi S. and Esposito V. (1998). Comparing advertising campaigns by means of textual data analysis with external information. In Mellet S., editor, *Actes des 4es Journées internationales d'Analyse statistique des Données Textuelles*, UPRESA, Nice : 39-47.

Balbi S. and Esposito V. (2000). Rotated canonical analysis into a reference subspace. *Computational Statistics and Data Analysis*, vol.(32) : 395-410.

Balbi S. and Misuraca M. (2005). Procrustes Techniques for Text Mining. In Zani S. and Cerioli A., editors, *Book of the Short Papers, Meeting of the Classification and Data Analysis Group of the Italian Statistical Association* (CLADAG05), MUP : 37-40.

Deerwester S., Dumais S., Furnas G., Landauer T. and Harshman R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, vol.(41.6) : 391-407.

Gower J.C. (1975). Generalised Procrustes Analysis. *Psychometrika*, vol.(40) : 33-51.

Grefenstette G., editor (1998). *Cross Language Information Retrieval*. Kluwer Academic Publishers.

Hardoon D.R., Szedmak S. and Shawe-Taylor, J. (2003). Canonical correlation analysis. An overview with application to learning methods. Technical Report CSD-TR-0302, Computer Science Department, Royal Holloway, University of London.

Hotelling H. (1936). Relations between two sets of variables. *Biometrika*, vol.(28) : 321-377.

Lafosse R. (1985). Une nouvelle analyse procrustéenne de deux tableaux. *Data Analysis and Informatics*, vol.(4) : 407-414.

Lafosse R. (1989). Ressemblance et différence entre deux tableaux totalement appariés. *Statistique et analyse des données*, vol.(14) : 1-24.

Lebart L. (1998). Text mining in different languages. *Applied Stochastic Models and Data Analysis*, vol.(14) : 323–334.

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

Littman M.L., Dumais S.T. and Landauer T.K. (1998). Automatic cross-language information retrieval using latent semantic indexing. In Grefenstette G., editor, *Cross Language Information Retrieval*, Kluwer Academic Publishers : 51–62.

Mardia K.V., Kent J.T. and Bibby J.M. (1979). *Multivariate Analyisis*. Academic Press.

Widdows D., Dorow B. and Chan C.K. (2003). Using Parallel Corpora to enrich Multilingual Lexical Resources. In *Proceedings of LREC 3* : 240-245.