Lexical Recount between Factor Analysis and Kohonen Map: Mathematical Vocabulary of Arithmetic in the Vernacular Language of the Late Middle Ages

Nicolas Bourgeois, Marie Cottrell, Benjamin Deruelle, Stéphane Lamassé, Patrick Letrémy

Lexical recount between Factor Analysis and Kohonen Map: mathematical vocabulary of arithmetic in the vernacular language of the late Middle Ages

Nicolas Bourgeois[1], Marie Cottrell[1], Benjamin Déruelle[2], Stéphane Lamassé[2], and Patrick Letrémy[1]

[1] SAMM - Université Paris 1 Panthéon-Sorbonne
90, rue de Tolbiac, 75013 Paris, France
nbourgeo@phare.normalesup.org
marie.cottrell,patrick.letremy@univ-paris1.fr
[2] PIREH-LAMOP - Université Paris 1 Panthéon-Sorbonne
1, rue Victor Cousin, Paris, France
benjamin.deruelle,stephane.lamasse@univ-paris1.fr

Abstract. In this paper we present a combination of factorial projections and of SOM algorithm applied to a text mining problem. The corpus consists of 8 medieval texts which were used to teach arithmetic techniques to merchants. Classical Factorial Component Analysis (FCA) gives nice representations of the selected words in association with the texts, but the quality of the representation is poor in the center of the graphs and it is not easy to look for the successive projections to conclude. So using the nice properties of Kohonen maps, we can highlight the words which seems to play a special role in the vocabulary since their are associated with very different words from a map to another. Finally we show that combination of both representations is a powerful help to text analysis.

Keywords: Text Analysis, Factorial Component Analysis, Kohonen Map

1 Introduction

1.1 Context

One approach to the understanding of the evolution of science is the study of the evolution of the language used in a given field. That is why we would like to pay attention to the vernacular texts dealing with practical arithmetic and written for the instruction of merchants: such texts are known since the XIII$^{\text{th}}$ century, and from that century onwards and especially after the diffusion of the Latin Leonard of Pisa's *Liber Abaci*, the vernacular language appears more and more as the medium of practical mathematics.

Treaties on arithmetical education were therefore mostly thought and written in local languages. In this process, the XVth century appears as a time of exceptional importance because we can see then how the inheritance of two hundred years of practice transfers into words[1]. For the authors of these texts, the purpose was not only to teach merchants but also to develop knowledge in vernacular language, and their books were circulated far beyond the shopkeepers' world, as far as the humanists' circles for example.

1.2 An objective of historical research: the study of specialized languages

The work previously done (Lamassé [2012]) consisted in the elaboration of a dictionary of the lexical forms found in all the treaties in order to identify the different features of the mathematical vernacular language of the time. This done, we have worked on the contexts of some especially important words in order to understand the lexicon in all its complexity, and on the specificities of each text to study the proximities and the differences between them. In other words, we should like to determine the common language that forms the specialized language beyond the specificities of each text.

2 The data, the objectives, the protocol

In order to delimit a coherent corpus among the whole European production of practical calculation education books, we have chosen to pay attention to those treaties which are sometime qualified as commercial (*marchand* in French) which have been written in French between 1415 and 1520. This last date is the date of the publication of *L'arismetique novellement compose* of Estienne de La Roche which is closely related to the works of Nicolas Chuquet and which provides in the same time an opening towards the Italian authors, such as Fra Luca Pacioli. In this way, our corpus is in conformity with the rules of the discourse analysis: homogeneity, contrastiveness and diachronism[2]. It contains eight treaties on the same topic, written in the same language and by different XVth century authors. The following table 1 describes some elements of the lexicometric characteristics of the corpus and shows its main quantitative imbalance.

2.1 Humanities and Social Sciences traditional protocol

Traditionally on this kind of textual data, HSS researchers use to work on the statistical specificities and on the contextual concordances, since they allow an easy discovery of the major lexical splits within the texts of the corpus while remaining close to the meanings of the different forms. Then, the factorial and

[1] These treaties were written not only in French but also in Italian, Spanish, English and in German.

[2] For further explanations about texts, methodology and purpose of the analysis see Lamassé [2012].

Manuscripts and **Title**	**Date**	**Author**	**Number of occurrences**	**Words**	**Hapax**
Bibl. nat. Fr. 1339	ca. 1460	anonyme	32077	2335	1229
Bibl. nat. Fr. 2050	ca. 1460	anonyme	39204	1391	544
Cesena Bibl. Malest. S - XXVI - 6, *Traicté de la praticque*	1471?	Mathieu Préhoude?	70023	1540	635
Bibl. nat. Fr. 1346, Commercial appendix of *Triparty en la science des nombres*	1484	Nicolas Chuquet	60814	2256	948
Méd. Nantes 456	ca. 1480-90	anonyme	50649	2252	998
Bibl. nat. Arsenal 2904, *Kadran aux marchans*	1485	Jean Certain	33238	1680	714
Bib. St. Genv. 3143	1471	Jean Adam	16986	1686	895
Bibl. nat. Fr . Nv. Acq. 10259	ca. 1500	anonyme	25407	1597	730

Table 1. Corpus of texts and main lexicometric features (Hapax are words appearing once in a text).

clustering methods, combined with co-occurrences analysis - see Martinez and Salem [2003] help us to cluster the texts without breaking the links with semantic analysis. However, such a method of data processing requires a preliminary treatment of the corpus, the lemmatization. It consists in gathering the different inflected forms of a given word as a single item. It offers the possibility to work at many different levels of meaning, depending upon the granularity adopted: forms, lemma, syntax. We can justify this methodological choice here by its effect on the dispersion of the various forms which can be linked to the same lemma, a high degree of dispersion making the comparison between texts more difficult. It must also be remembered that in the case of medieval texts, this dispersion is increased by the lack of orthographic norms. In our case, this process has an important quantitative consequence on the number of forms in the corpus, which declines from 13516 forms to 9463, a reduction of some 30%.

The factorial analysis allowed us to establish a typology of the complete parts of the corpus, based upon all the forms. However, it can be useful to improve this global analysis with a probabilistic calculation for each component of the corpus, by using the table of the under-frequencies (Lebart and Salem [1994]). It makes it possible to compare the parts of the corpus with each other, taking into account the occurrences of the words and their statistical specificities.

This process has been made with a particular attention to meaning of the word in order to suppress ambiguities : a good example is the French word *pouvoir* which can be a verb translated by "can" or "may", and which is also a substantive meaning "power".

Finally, to realize a clustering of the manuscripts, we have only kept the 219 words with the highest frequencies. The set of words thus selected for text classification relate both to mathematical aspects, such as operations, numbers and theirs manipulations, as well as to didactic aspects. Their higher frequencies reflect the fact that they are the language of the mathematics as they appear to be practiced in these particular texts. Thus, in what follows the data are displayed in a contingency table T with $N = 219$ rows (the words) and $p = 8$ columns (the texts) and the entry $t_{i,j}$ is the number of occurrences of word i in text j.

2.2 Factorial Correspondence Analysis (FCA)

Factorial Correspondence Analysis is a factorial method which provides the simultaneous representation of both the individuals and their characteristics, that is to say the columns and the rows of a table, in our case the texts (columns) and the words (rows). Figure 1 and 2 show the projection of the data on the first two factorial axes.

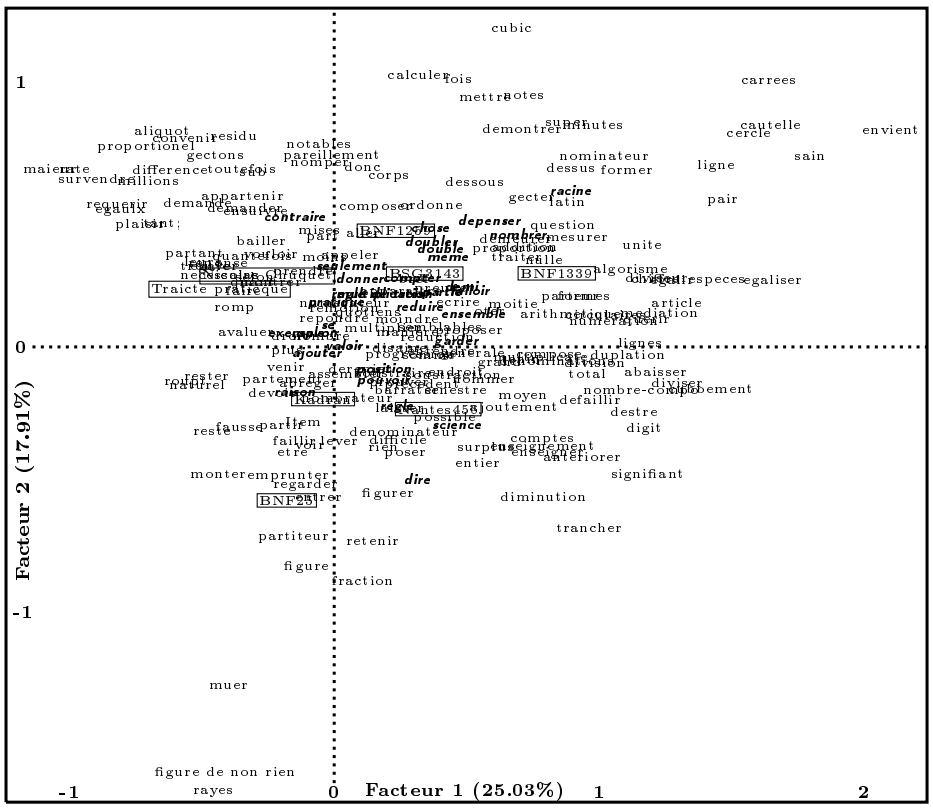

Fig. 1. Projection on first two factors of the FCA. The eight texts appear in frames, the slanted words stand for fickle words (this notion will be defined in section 4)

The first two factors (43.94% of the total variance) show the diversity of the cultural heritages which have informed the language of these treaties. The first factor (25.03%) discriminates between the vocabulary according to its relation to the university legacy on the left, and to the tradition of mathematical problems on the right.

On the left, we can observe a group whose strong homogeneity comes from its orientation towards mathematical problems (*trouver* that is to say "to find",

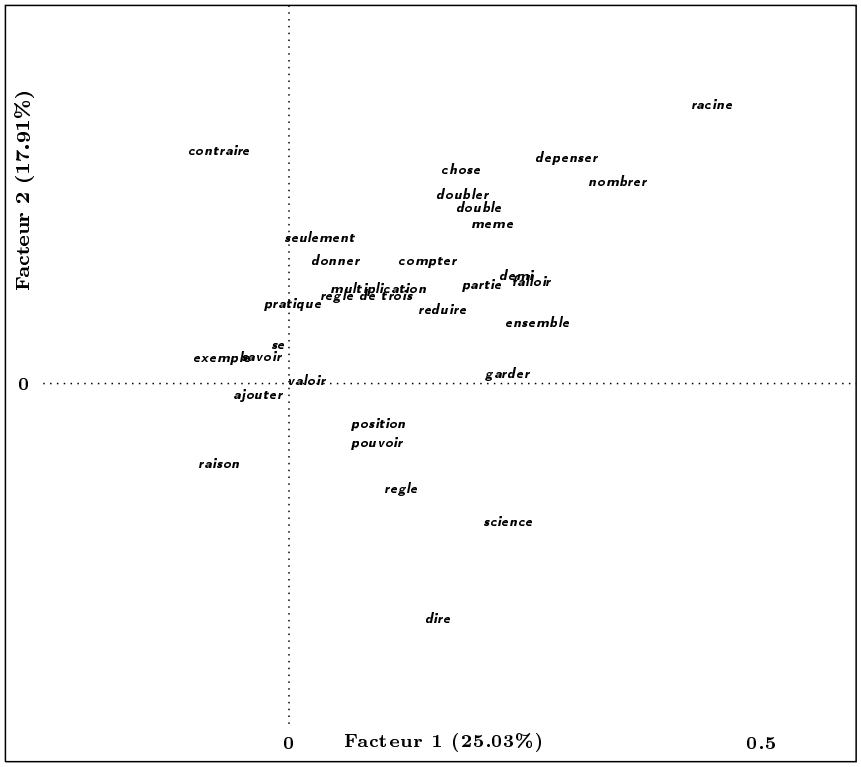

Fig. 2. Projection on first two factors of the FCA (zoom on the central part). Only the fickle words are represented.

demander which we can translate as "to ask") and their iteration (*item, idem*). That vocabulary can be found most often in both the appendix of *Triparty en la science des nombres* and *Le Traicte de la praticque*. Furthermore, there are more verbal forms on this side of the axis than on the other. And we can find verbs like *requerir* which means "to require", *convenir* "to agree", *faire* "to do", *vouloir* "to want". Some of them are prescriptive, as *devoir* "to have to" or *vouloir* "to want" for example, while others introduce examples, as *montrer* "to show". All these texts contain a lot of mathematical problems and in a way that texts are more practical.

On the right, the texts of BnF. fr. 1339 and Med. Nantes 456 are clearly more representative of the university culture, containing latin words sequences. They describe basic operations and numbers through words developed around the XII$^{\text{th}}$ and the XII$^{\text{th}}$ century in the universities, such as *digit*, *article* and *nombre composé*[3].

[3] *Digit* is used for 0 to 9, *article* for every multiple of ten, and *nombre composé* is a mixed between *article* and *digit*.

The second axis (17.91% of the variance) is mostly characterized by the text of BNF. fr. 2050 and also by *Kadran aux marchans*. It displays words of Italo-Provencal origin, like *nombrateur* which refers to the division's numerator. Designation of the fraction and operation of division take a significant part of the information while the most contributory words (for ex. *figurer* "to draw") allow us to examine another dimension of these works: the graphical representation as a continuation of writing.

Correspondence Analysis displays the particularities of each text, but leaves untouched some more complex elements of the data. For instance, we cannot conclude from this opposition that the authors of the appendix of *Triparty en la science des nombres* and the *Traicte de la praticque* are ignorant of the university texts that inspire the other books. Correspondence analysis does not make fully possible the analysis of the attractions. Moreover, we cannot assert that the words which appear in the center of the graph represent a "common vocabulary": as a matter of fact, we ought to analyze all the successive factors in order to build the list of the words constituting the "normal" vocabulary.

3 SOM algorithm for contingency table

One way to overcome the limitations of the Factorial Correspondence Analysis (FCA) consists of using a variant of the SOM algorithm which deals with the same kind of data, that is a contingency table (see Oja and Kaski [1999] for other applications of SOM to text mining). See Cottrell et al. [1998] for a definition of this variant of SOM, we called KORRESP.

The KORRESP algorithm consists in a *normalization* of the rows and of the columns in order to sum to 1, the *definition* of an extended data table by associating to each row the most probable column and to each column the most probable row, a *simultaneous classification* of the rows and of the columns onto a Kohonen map, by using the rows of the extended data table as input for the SOM algorithm.

After convergence of the training step, the modalities of the rows and of the columns are simultaneously classified. In our example, one can see proximities between words, between texts, between words and texts. It is the same goal as in Factorial Correspondence Analysis. The advantage is that it is not necessary to examine several projection planes : the whole information can be read on the Kohonen Map. The drawback is that the algorithm is a stochastic one, and that apparent contradictions between several runs can be troublesome.

In fact, we can use this drawback to improve the interpretation and the analysis of relations between the studied words. Our hypothesis is that the repetitive use of this method can help us to identify words that are strongly attracted/repulsed and fickle pairs.

In its classical presentation Kohonen [1995], Cottrell et al. [1998], the SOM algorithm is an iterative algorithm, which takes as input a dataset $\mathbf{x}_i, i \in \{1, \ldots, N\}$ and computes prototypes $\mathbf{m}_u, u \in \{1, \ldots, U\}$ which define the map.

We know that self-organization is reached at the end of the algorithm, which implies that close data in the input space have to belong to the same class or to neighboring classes, that is to say that they are projected on the same prototypes or on neighboring prototypes on the map. In what follows we call neighbors data that belong either to the same unit or to two adjacent units. But the reciprocal is not exact : for a given run of the algorithm, two given data can be neighbor on the map, while they are not in the input space. That drawback comes from the fact that there is no perfect fit between a two-dimensional map and the data space (except when the intrinsic dimension is exactly 2). Moreover, since the SOM algorithm is a stochastic algorithm, the resulting maps can be different from one run to another.

We address the issue of computing a reliability level for the neighboring (or no-neighboring) relations in a SOM map. More precisely, if we consider several runs of the SOM algorithm, for a given size of the map and for a given data set, we observe that most of pairs are almost always neighbor or always not neighbor. But there are also pairs whose associations look random. These pairs are called *fickle* pairs. This question was addressed by de Bodt et al. [2002] in a bootstrap frame.

According to their paper, we can define: $NEIGH^l_{i,j} = 0$ if x_i and x_j are not neighbor in the l-th run of the algorithm, and $NEIGH^l_{i,j} = 1$ if x_i and x_j are neighbor in the l-th run of the algorithm, where (x_i, x_j) is a given pair of data, l is the number of the observed run of the SOM algorithm.

Then they define the stability index $\mathcal{M}_{i,j}$ as the average of $NEIGH_{i,j}$ over all the runs ($l = 1, \ldots, L$), i. e. $\mathcal{M}_{i,j} = (1/L) \sum_{l=1}^{L} NEIGH^l_{i,j}$. The next step is to compare it to the value it would have if the data x_i and x_j were neighbor by chance in a completely random way.

So we can use a classical statistical test to check the significance of the stability index $\mathcal{M}_{i,j}$. Let U be the number of units on the map. If edge effects are not taken into account, the number of units involved in a neighborhood region (as defined here) is 9 in a two-dimensional map. So for a fixed pair of data x_i and x_j, the probability of being neighbor in a random way is equal to $9/U$ (it is the probability for x_j to be a neighbor of x_i by chance once the class x_i belongs to is determined).

Let $Y_{i,j} = \sum_{l=1}^{L} NEIGH^l_{i,j}$ be the number of times when the data x_i and x_j are neighbor for L different, independent runs. It is easy to see that $Y_{i,j}$ is distributed as a Binomial distribution with parameters L and $9/U$. Using the classical approximation of Binomial Distribution by a Gaussian one (L is large and $9/U$ not too small), we can build the critical region of the test of null hypothesis H_0 "x_i and x_j are neighbor by chance" against hypothesis H_1 : " the fact that x_i and x_j are neighbor or not is significant".

We conclude that the critical region for a test level of 5% based on $Y_{i,j}$, is

$$\mathbb{R} - [L\frac{9}{U} - 1.96\sqrt{L\frac{9}{U}(1-\frac{9}{U})}, L\frac{9}{U} + 1.96\sqrt{L\frac{9}{U}(1-\frac{9}{U})}]$$

$$\text{Fix } A = \frac{9}{U} \text{ and } B = 1.96\sqrt{\frac{9}{UL}(1 - \frac{9}{U})}.$$

Practically, in this study, for each pair of words, we can compute (over 40 experiments) the index $\mathcal{M}_{i,j} = Y_{i,j}/L$, and conclude. Henceforth:

- if their index is greater than $A + B$, they are almost always together in a significant way, the words attract each other.
- if their index is comprised between $A-B$ and $A+B$, their proximity depends on the text they belong, they are a fickle pair.
- if their index is less than $A - B$, they are almost never neighbor, the words repulse each other.

4 Analysis of fickle pairs and nodes

4.1 Identification of fickle pairs

We run KORRESP L times and store the result in a matrix $\mathcal{M}$ of size $(N+p) \times (N+p)$. The value stored in a given cell i, j is the proportion of maps where i and j are neighbors.

	abaisser	abreger	addition	ajoutem.	ajouter	algorisme	aliquot	aller	anteriorer
abaisser	1	0	0,025	0,275	0	0,05	0	0	0,525
abreger	0	1	0	0	0,25	0	0,325	0	0,025
addition	0,025	0	1	0	0	0,875	0	0,05	0
ajoutement	0,275	0	0	1	0,025	0	0	0,025	0,7
ajouter	0	0,25	0	0,025	1	0,025	0,15	0,125	0
algorisme	0,05	0	0,875	0	0,025	1	0	0	0
aliquot	0	0,325	0	0	0,15	0	1	0,025	0
aller	0	0	0,05	0,025	0,125	0	0,025	1	0
anteriorer	0,525	0,025	0	0,7	0	0	0	0	1

> 0,179 [0,02 ; 0,179] < 0,02

Fig. 3. Excerpt from matrix $\mathcal{M}$ with $L = 40$ and $r = 1$

Figure 3 displays an example of the nine first rows and columns of such a matrix. We have highlighted with colors three different situations. According to the theoretical study mentioned above:

- Black cells stand for pairs that are neighbors with high probability (proximity happens with frequency greater than $A + B$).
- White cells stand for pairs that are not neighbors with high probability (proximity happens with frequency less than $A - B$).
- Grey cells are not conclusive.

4.2 From fickle pairs to fickle words

We call fickle a word which belongs to a huge number of fickle pairs:

$$|\{i, |\mathcal{M}_{i,j} - A| \leq B\}| \geq T$$

Unfortunately, it is not quite an easy task to find an appropriate threshold T. Here we have decided to fix it according to data interpretation. The 30 ficklest words, whose number of safe neighbors/non-neighbors is between 89 and 119, are displayed in table 2.

contraire "opposite" (89)	*regle de trois* "rule of three" (104)	*depenser* "to expend" (112)
doubler "to double" (89)	*savoir* "to know" (105)	*racine* "root" (113)
falloir "to need" (93)	*partie* "to divide" (105)	*chose* "thing" (113)
meme "same, identical" (93)	*position* "position" (107)	*compter* "to count" (113)
pratique "practical" (94)	*exemple* "for example" (107)	*dire* "to say" (113)
seulement "only" (94)	*demi* "half" (108)	*nombrer* "count" (115)
double "double" (97)	*garder* "to keep"(109)	*raison* "calculation, problem" (116)
multiplication (99)	*science* "science" (109)	*donner* "to give" (117)
reduire "to reduce" (103)	*pouvoir* "can" (111)	*ensemble* "together" (117)
regle "rule" (103)	*se* "if" (111)	*valoir* "to be worth" (119)

Table 2. 30 ficklest words among 219 studied

FCA with fickle pairs The combination of both techniques FCA and SOM whose result is displayed in figure 1 is interesting because it preserves properties from the FCA while giving additional information about the center of the projection - which is usually difficult to interpret. Indeed, the identification on the FCA of the fickle forms allows us to control the general interpretation of the factorial graph, where some words find their place because of the algorithm and not because of their attraction with other forms and with the texts.

Remember that, on the first two factorial axes (see section 2.2), we have observed an opposition between the university legacy, on the right, and a more practical pole with rule, problems and fractions, on the left. It was tempting to support this observation with words such as "practical" or "rule of three". On the other hand, the fickle forms enhancement shows that these words are shared between a lot of different texts and not only linked to the treaty of Nicolas Chuquet and the *Traicté en la praticque*. As a matter of fact, they do belong to all the texts. And we can observe that the first factor opposes two technical languages that are on either side of a set of common words – and these words are obviously not to say necessarily in the center of the FCA (see for instance *racine* "root").

To conclude, we can see that two levels of interpretation are superimposed: the fickle pairs which reveal the shared lexicon and the factorial map which inserts it in a local interaction system. And because the list is not sensitive to the FCA, we can play on this combination for each successive factorial axis. It is the articulation between these two levels which makes this representation

interesting. In the end, the meaning of this new kind of factorial map is quite intuitive and offers easy tools to the argumentation.

5 Perspectives and conclusion

First, we intend to use the proposed method for other corpus to confirm its capacity to extract specialized vocabulary. In particular we want to make a new experimentation on a corpus of medieval and renaissance prologues of epics. This corpus has been constituted in order to discern the political and ideological appropriations of the chivalric culture in the context of the XV$^{\text{th}}$ and XVI$^{\text{th}}$ centuries Renaissance. Secondly, the method has to be appropriated by linguists in order to improve it according to their own paradigms.

Another challenge will be to work on a statistical characterization of a threshold for the definition of fickle pairs. Indeed, while we managed to define (through confidence intervals) a theoretical frame for reliability of a pair, we still need to infer a similar method for each data.

Finally, we think that we have open a new perspective for clustering through Kohonen maps. Indeed, the study of robust attraction/repulsion between data as well as fickle pairs can be translated into a graph. Then, we can apply methods from graph representation and graph mining in order to get visualization containing more information than a single SOM.

References

M. Cottrell, J.-C. Fort, and G. Pagès. Theoretical aspects of the SOM algorithm. *Neurocomputing*, 21:119–138, 1998.

E. de Bodt, M. Cottrell, and M. Verleysen. Statistical tools to assess the reliability of self-organizing maps. *Neural Networks*, 15, 8-9:967–978, 2002.

Teuvo Kohonen. *Self-Organizing Maps*, volume 30. Springer Series in Information Science, Berlin, 1995.

Stéphane Lamassé. Les traités d'arithmétique médiévale et la constitution d'une langue de spécialité. In Joëlle Ducos, editor, *Sciences et langues au Moyen Âge, Actes de l'Atelier franco-allemand, Paris, 27-30 janvier 2009*, pages 66–104. Universitätesverlag, Heidelberg, 2012.

Ludovic Lebart and André Salem. *Statistique textuelle.* Dunod, Paris, 1994. ISBN 2-10-002239-3.

William Martinez and André Salem. *Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels.* Thèse doctorat, 2003.

Erkki Oja and Samuel Kaski. *Kohonen Maps.* Elsevier, 1999.