NUMEXCO: A Text Mining Approach to Thematic Analysis of a Philosophical Corpus

Dominic Forest and Jean-Guy Meunier
Laboratoire d'Analyse Cognitive de l'Information (LANCI) ;
Université du Québec à Montréal (UQAM)
dominic.forest@internet.uqam.ca and
meunier.jean-guy@uqam.ca

Abstract

In this paper, we present the results of a research in which our main objective was to explore text mining techniques in their application to thematical analysis of philosophical corpus. In the first part, we present the software (Numexco) used in our experiment, focusing on the modular approach of the software. The second part presents the results obtained when we applied Numexco to the thematical analysis of Descartes' Discours de la méthode. This paper also demonstrates how this computer assisted technology can allow the user to rapidly identify, explore and navigate through the different themes found in humanities texts.

KEYWORDS: Text classification, thematic analysis, categorization, text mining, philosophy.

Introduction

Since the 1970's, many researches in cognitive sciences and information technologies have had important impacts on the reading and analysis of humanities texts. In numerous of these researches, various artificial intelligence classification and categorization strategies have been explored to assist the description and the interpretation of texts. Some projects have tried to apply predefined rules and concepts to texts. This procedure is known as a top-down analysis process. This type of strategy still highly influences what is called "qualitative analysis" (Glaser and Strauss, 1967), "grounded theory", "computer assisted qualitative data analysis" (Barry, 1998, Alexa and Zuell, 1999a) or, in more general terms, "Computer Assisted Reading and Analysis of Text" (CARAT). From this point of view, "qualitative research usually means the collection and analysis of unstructured textual material in order to develop concepts, categories, hypotheses, theories. Thus, most of the time during "qualitative data analysis" is spent on reading, rereading, interpreting, comparing and thinking on texts" (Kelle, 1997a).

On the other hand, more and more bottom-up strategies are appearing for inductive text classification and categorization. This second type of strategies is traditionally applied to information retrieval problems, but their power can also be used in computer assisted reading and analysis of text for lexical analysis, automatic generation of hypertext links, knowledge extraction and thematic analysis. In this paper, we present an application of these classification and categorization technologies to one specific field: the thematic analysis of philosophical texts. The computer application presented in this paper is called Numexco. It is an example of analysis chain that was conceived within the computer platform called SATIM, devoted to the conception of computer assisted reading and analysis of text tasks.

1. The nature of classification

In general terms, a classifier is an abstract machine or function that realizes some type of grouping or sorting on a set of objects. In a set theoretical definition, classification is the projection of a partition on objects so that they are as homogenous as possible. In categorical terms, a classifier can be defined as a quadruplet (O, X, I, G) where:

O is the set of objects ($o_1 \ldots. o_n$);

X is the set ($x_1 \ldots. x_n$) of features describing each object. O;

I is the set of types ($i_1 \ldots. i_n$);

G is a discriminant function.

In this perspective, a classifier is conceived as a categorical operation defined in the following manner:

For each object $O((G(x_1 \ldots. x_n) i)x_j)$.

In other terms, a classifier is an operation G that takes as input a set of objects of type i described by their feature X and delivers objects of another type i. The real challenge for building a good classifier is twofold: first, the classifier must be given the right descriptors for its objects and, secondly it must possess a good discriminant function. This discriminant function is what allows the classifier to concretely build a class; in other words, it is the criterion of sameness, similarity, homogeneity, equivalence etc.

Thus, text classification can be defined as an operation, that, when applied to texts described by some of their characteristics, builds equivalent classes on them. “Text classification is the automated grouping of textual or partially textual entities” (Lewis and Gale, 1994).

It is important to note that a classification process can be applied to different kinds of objects found in a text. For example, one can classify words, paragraphs, pages, etc. However, one of the main applications in text analysis is to apply it not to entire texts but to parts or fragments of texts described by their “representors” (words or n-grams). Secondly, classification is not an end in itself. It is a first step in a complex cognitive process of computer assisted text interpretation (Meunier, 1996). Concretely, this means that text classification must be linked to the dynamic process of text interpretation and hence must be appreciated according to this end.

2. Thematic analysis

Many definitions have been proposed for thematic analysis (Louwerse and van Peer, 2002). According to Popping (2000), thematic analysis can be described as the “identification of what, and how frequently, concepts occur in texts”. This definition, like many others, remains very general and fuzzy. Stone (1997) mentioned that the concept of theme “is used in a loose, general way for analysing patterns in text”. Stone’s point of view remains very general, but seems to allow some type of computational perspective on theme analysis. However, instead of describing theme analysis in terms of patterns analysis, the literature on this topic (here, we refer to studies in philosophy, literature, text analysis, etc.) shows that theme analysis lies more in the discovery and identification of the multiples relations between different themes that make a textual corpus consistent and intelligible. We will show how this definition can lead to some type of computer-assisted theme analysis.

3. Methodology and experimentation

Within the Numexco application, the computer-assisted thematic analysis process can be realized through five distinct operations all of which constitute a step in the overall analysis chain: 1) The text is pre-processed, 2) it is then transformed into a vectorial model, 3) a classifier is applied, 4) thematic links are extracted and 5) results are interpreted.

3.1. Text preparation

To be effective, text classifiers need to define two different types of entities: fragments of texts and units of textual information. Although quite simple to understand, this deconstruction of a text into fragments of texts and units of textual information relies on many implicit postulates: some of them are linguistic in nature, others are mathematical.

Units of information can be either words or n-grams (chain of n characters). Choosing the units of information relies on many decisions. For instance, many possibilities are at hand: one can take into account linguistic variants of words (for instance the flexional variants such as "do", "done", "did"). It is also possible to eliminate stop-words and trivial words, to lemmatize the whole corpus, to aggregate or not complex lexical words such as "philosophy of mind", "mind-body problem", etc. More so, one can retain or not all the units of information, eliminate low and high frequency words, etc.

In the following experiment, we applied the Numexco software to a philosophical text: Descartes' Discours de la méthode. This untouched text (except for basic printing noises) contains 21 436 words. The corpus was segmented in 154 segments (or fragments) of 150 words. The main reason for dividing the corpus in fragments of 150 words comes from the choice of the classifier which will be applied to the fragments. The units of textual information were the lemmatized words left after eliminating hapax and functional terms.

3.2. Text transformation into a vectorial representation

From then on, the text is transformed into a vectorial model (Salton, 1989). This procedure requires that each vector represents a segment described by its units of information. From these vectors a matrix is built (figure 1). The values given to each entry of the matrix depends on the model chosen (e.g. presence, absence, fuzziness, weighting, etc.) and the classifier which will be used to classify the set of fragments. On this textual matrix, we now have to define the core of the classification operation; that is, the discriminant function. It is on this matrix that classifiers are applied.

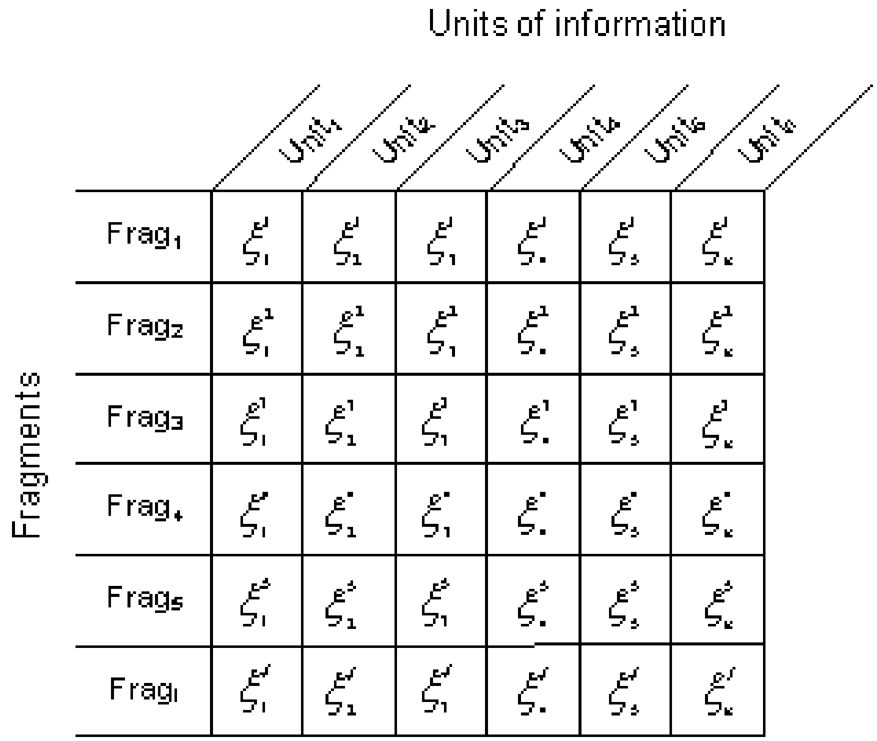

Figure 1. The matrix fragments of text x units of information.

3.3. The classification process

In the literature of mathematical text classifiers, many types of classifiers have been applied to various text analysis tasks. All these different classifiers have their parameters, hence, their fecundity and their limits. The most common ones are statistically oriented (clusterizers, correlators, factorial analysis (Reinert, 1994), principal component analysis (Benzecri, 1973)). One successful implementation of these types of models has been, in the information retrieval community, the SMART system of Salton (Salton, 1989). Also, some more probabilistic techniques have been tested, such as the bayesian classification, nearest neighbours (Hart, 1998), neural networks classifiers (Kohonen, 1982; Anderson, 1976), genetic algorithm (Holland, 1975), markovian fields classifiers (Bouchaffra and Meunier, 1995b). All of these mathematical models have been applied to textual information processing.

In this experiment, we have chosen the neural ART1 classifier (Grossberg and Carpenter, 1987). We here explore this classifier and study its relevance for the computer assisted reading and analysis of text, focusing on its application to thematic analysis of philosophical corpus. The purpose of this experiment is not to demonstrate its validity for texts classification, as it is often done in classical information retrieval experiments, but to show its fecundity in thematic analysis studies.

In this ART1 classifier, the discriminant function measures the difference of the input segments described by their weighted features (words) but through a competition amongst the segments. That is to say, the fragments that belong to a set are the winners of their weight competition. In most models, this computation consists in accepting a segment only if the impact of all the features of an input attains a certain threshold.

The ART1 classifier (Grossberg, 1988) is highly interesting in many ways. Amongst other things, it allows plastic processing of information. Thus, when a new pattern is added, the ART1 system classifies it in the set of classes formed by the means of a competition process undertaken during the training phase. The essential principle of this model is based on the interaction of two levels of neurons that enter into a resonance phase such as shown in figure 2.

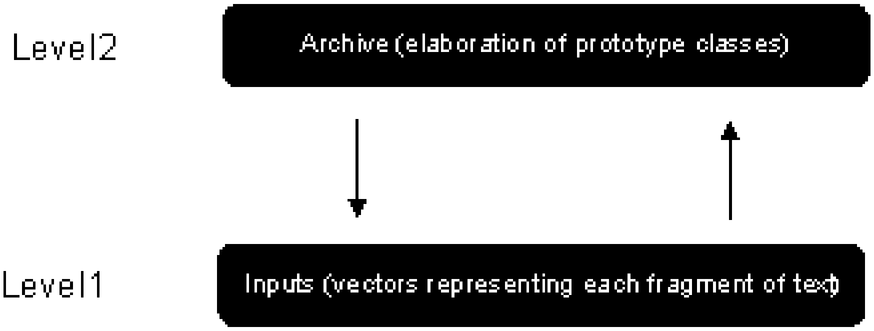
Level2
Archive (elaboration of prototype classes)
Level1
Inputs (vectors representing each fragment of text)

Figure 2. The ART1 resonance phase.

The system receives in its first level (Level1) the text fragments which are sent, after a modification of their distributed weights, to the second level of neurons (Level2). This transmission implies various complex differential operations, variation of the activating forces, degradation, shunting, etc.

This second level will then possess the modified patterns, each of which serves as a prototype for new-comers at the first level. Hence, each new input will be compared to the prototype according to a resonance criterion. If the correspondence is positive then the input is entered in the class of the prototype. If not, it will be considered as an emergent prototype. The adaptability emerges by this constant modification of the levels' interconnections. This new emerging prototype will then serve for starting up a new class. As the learning goes on, there will be a consolidation of this resonance process.

At the end of the process, a set of classes of segments is produced. From each class of segments, the lexicon is extracted. We then have for each class a specific lexical list. It is on these classes that the thematic analysis begins (figure 3).

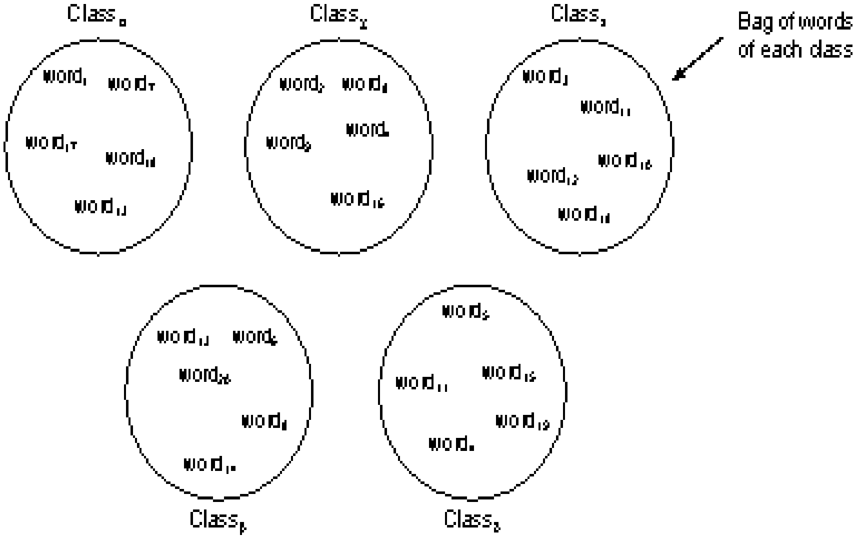

Figure 3. Graphical representation of the lexical lists generated by the classification process.

The various processes presented above (extraction of the units of information, segmentation, etc.) can lead to useful statistical information about the corpus. For example, the following figure illustrating the growing evolution of Descartes' vocabulary in his Discours de la méthode was generated by comparing each word (token) of the Discours de la méthode with the set of fragments. This information can be useful in many tasks such as authorship attribution, corpora comparison, etc.

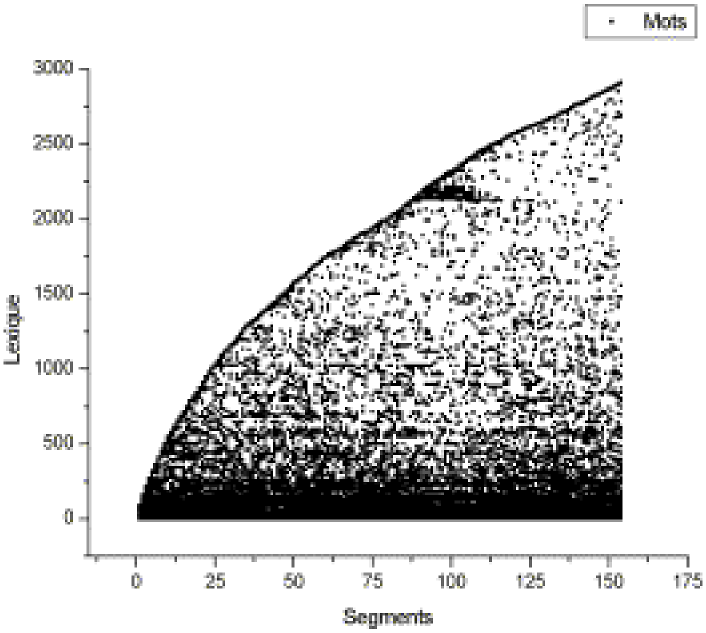

Figure 4. Descartes' vocabulary evolution in the Discours de la méthode.

3.4. Thematic extraction

The computed-assisted thematic analysis starts by choosing a preferred "thematic word" that can be found in one of the segments of the classes produced. In the present experiment, we have started with the word (or concept) "connaissance". From this specific word, thematic relations are then identified with other "thematic words"[1] belonging to other classes. In the sample results shown below, a particular analysis of Descartes' Discours de la Méthode starts with the word "connaissance" found in many classes. In classes 38 and 40, the word "connaissance" co-occurs with the following words "astre", "air", "ciel", "lumière", "matière", "monde", "terre". Starting from this observation, the user can thematically categorize (manually or automatically (Forest and Meunier, 2004)) the content of the class with the categorical tag "physique". But, by the same token, we also observe that the same word ("connaissance") operates semantically in other contexts or classes. Accordingly, the reader could then decide to discover (by exploring other classes where the word "connaissance" is also found) that the same concept of “connaissance” operates in many other different fields (or leads to the discovery of different themes) of Descartes’ writing. Thus, in a general way, we can observe that, according to this analysis, the word “connaissance” can also be used to discover the theme “métaphysique” (composed of classes 31 and 34 which are characterized with the words “parfaire”, “existence”, “dieu”, etc.), “mathématique” (composed of classes 6 and 17 which are characterized with the words “démonstration”, “science”, “géomètre”, etc.) and “anatomie” (composed of classes 44 and 46 which are characterized with the words “animal”, “poumon”, “coeur”, etc.).

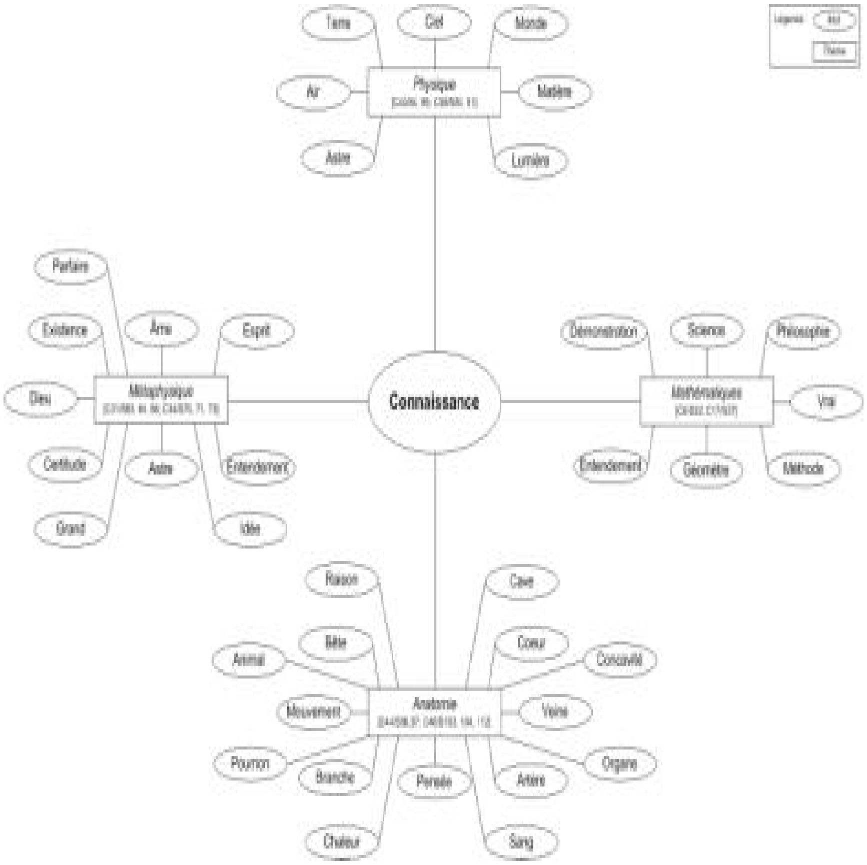

Figure 5. Thematic exploration.

With this preliminary classification function, the reader can then direct his analysis according to a chosen theme. And from then on, he can explore other different themes found in Descartes' philosophy. For instance in the following example, he may switch to another concept inside a particular class (by choosing the word "pensée", for example) and start a new thematic path. This strategy is then applied to the entire text.

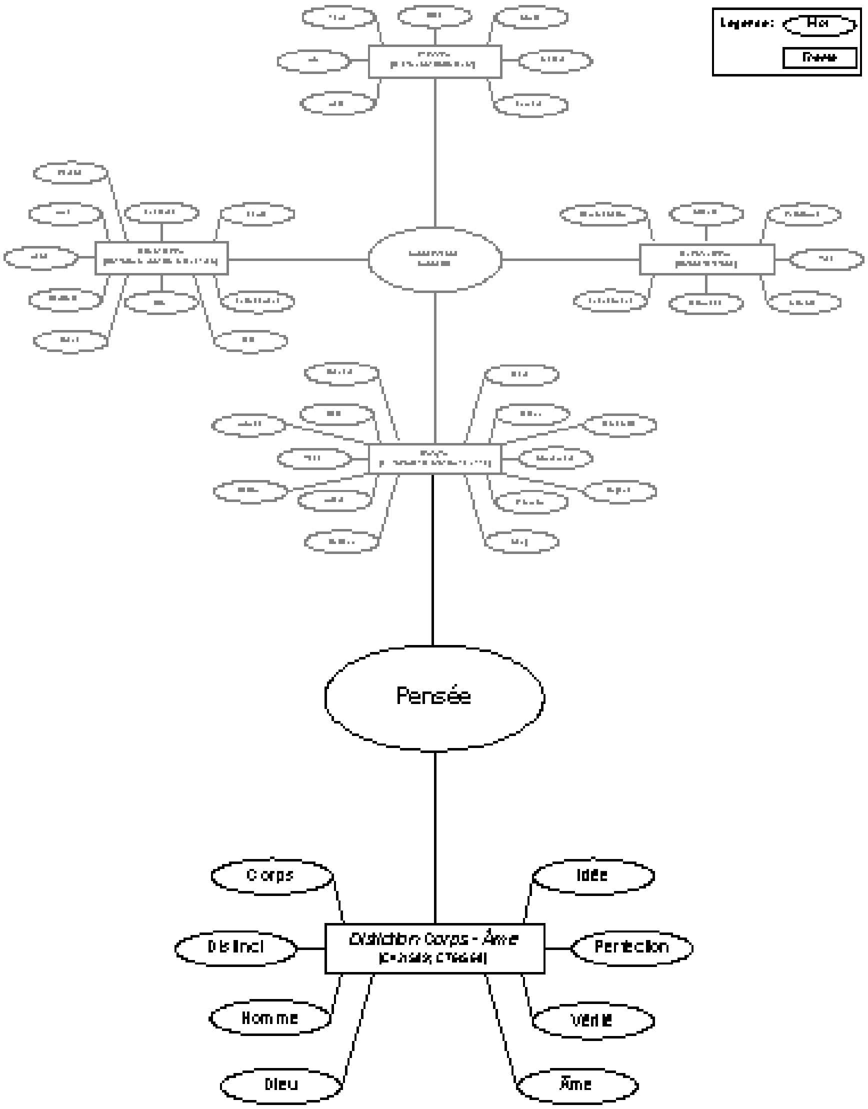

Figure 6. Other path in thematic exploration.

The results have shown that many thematic links could be relevant for the reader. It is worth mentioning that these results, after being compared with the traditional Descartes' commentators (Rodis-Lewis, 1966, 1985; Gueroult, 1953; Laporte, 1950, etc.), were similar with the classical interpretations of Descartes' philosophy for this particular theme.

4. Evaluation of classifiers for thematic analysis

From an epistemological point of view, this classification method usually relies on a benchmark produced by a set of objective experts (judges). Results are then compared to these benchmark results. We however believe that reading and analysis of texts remain a most subjective process. Therefore, this supposed "objective" evaluation methodology seems somehow difficult to apply because the discovery of the semantic content of a text is always a complex cognitive process that is guided by the reader's knowledge and personal reading interests. For each reader, and mainly in philosophical texts, there is a personal interpretative path. Nevertheless, the use of such classification methods gives some "objective" orientation to the analysis. It does not impose any kind of strategy to the reader but offers a multitude of possible paths. One possible way to evaluate the quality of a computer thematic classification would be to compare it to various best practices or to offer some type of relevance feedback measures produced by HMM strategies or genetic algorithms.

Finally, we must keep in mind that these techniques using information technologies must not be seen as replacing tools for reading and analysis of text. Instead, they must be seen as "assisting tools" to help the reader's discovery and interpretation of texts.

Notes

[1] "Thematic words" are the words that are present in more than one class. It is these specific words that allow the user to discover the various themes found in the corpus.

Works Cited

Alexa, M. and C. Zuell. (1999a). "Commonalities, difference and limitations of text analysis software: The results of a review." *ZUMA arbeitsbericht* ZUMA: Mannheim.

Anderson J. R. (1976). *Language, Memory and Thought*. New York: John Wiley & Sons.

Barry, C.A. (1998). "Choosing qualitative data analysis software: Atlas/ti and Nud*ist compared." *Sociological research online* 3(3).

Benzecri, J.-P. et al. (1973). *La Taxinomie.* Vol. I. *l'analyse des correspondances*. Paris : Dunod.

Bouchaffra, D. and J.-G. Meunier. (1995b). *A Thematic Knowledge Extraction Modeling through a Markovian Random Field Approach*. 6th International DEXA 95 Conference and Workshop on Database and Expert Systems Applications, Sept. 19-22, London, UK.

Deerwester, Scott C., Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. (1990). "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science* 41.6 (1990): 391-407.

Forest, D. et J.-G. Meunier. (2004). *Classification et catégorisation automatiques : application à l'analyse thématique des données textuelles.* Actes du colloque JADT 2004 (7èmes Journées internationales d'Analyse statistique des Données Textuelles), 10-12 mars 2004, Louvain-la-Neuve, Belgique.

Glaser, B. G. and A. L. Strauss. (1967). "The Discovery of Grounded Theory." *Strategies for Qualitative Research*. Chicago: Adline.

Grossberg, S. (1988). *Neural Network and Natural Intelligence*. Cambridge: MIT Press.

---. and G. A. Carpenter. (1987). "A massively parallel architecture for a self-organizing neural pattern recognition machine." *Computer Vision, Graphics, and Image Processing* 37: 54-115.

Gueroult, M. (1953). *Descartes selon l'ordre des raison.* Paris: Aubier.

Hart, P. E. (1968). "The condensed nearest neighbor rule." *IEEE, Trans. on Information theory*. IT. 14: 515.

Holland, J. (1975). *Adaptation in Natural, and Artificial Systems*. Ann Arbor, Michigan: University of Michigan Press.

Kelle, U. (1997a). "Theory Building in Qualitative Research and Computer Programs for the Management of Textual Data." *Sociological Research Online* 2 (2).

Kohonen, T. (1982). "Clustering, taxonomy and topological Maps of Patterns." *IEEE Sixth International Conf. Pattern Recognition*: 114-122.

Laporte, J.M.F. (1950). Le rationalisme des Descartes. Paris: PUF.

Lewis, D.D. and W. A. Gale. (1994). "A sequential algorithm for training text classifiers." *Proceedings of the seventeenth annual international ACM-SIGIR conference on research and development in information retrieval.* Eds. W. Bruce Croft and C. J. van Rijsbergen. Dublin: Springer-Verlag, 3-12.

Louwerse, M. and W. van Peer, eds. (2002). *Thematics: Interdisciplinary Studies*. Netherlands: John Benjamins Publishing Company.

Meunier, J. G. (1996). "La théorie cognitive: son impact sur le traitement de l'information textuelle." *Penser l'Esprit, Des sciences de la cognition à une philosophie cognitive*. Eds. V. Rialle et D. Fisette. Grenoble: Presses UG, 289-305.

Popping, R. (2000). *Computer-assisted text analysis*. London: Sage.

Reinert, M. (1994). "Quelques aspects du choix des unités d'analyse et de leur contrôle dans la méthode Alceste." *Analisi Statistica dei Dati Testuali. Eds. L. L. S. Bolasco and A. Salem.* Vol. 1. Rome: CISU, 19-27.

Rodis-Lewis, G. (1966). Descartes et le rationalisme. Paris: PUF.

---. (1985). *Idées et vérités éternelles chez Descartes et ses successeurs*. Paris: J. Vrin.

Salton, G. (1989). *Automatic Text Processing*. Addison Wesley.

Stone, P. (1997). "Thematic text analysis: new agendas for analyzing text content." T*ext analysis for the social sciences: methods for drawing statistical inferences from texts and transcripts*. Ed. C. W. Roberts. Mahwah, NJ: Lawrence Erlbaum Associates, 35-54.