La classification mathématique des textes : un outil d'assistance à la lecture et à l'analyse de textes philosophiques

Dominic Forest, Jean-Guy Meunier
LANCI – UQÀM – C.P. 8888, Succ. Centre-Ville, Montréal, Québec, Canada
H3C 3P8

ABSTRACT

Mathematical text classifiers can serve many purposes in the field of computer-assisted reading and analysis of text (CARAT). More than ever, with the large amount of text available in digital format such as CD-Roms, encyclopædia, etc., the use of text classifiers has become an essential tool in the field of natural language information processing technologies to help the reader discover information found in the text. The recent researches are revealing many different uses stemming from the mathematical classification research field. Besides lexical analysis or automatic generation of hypertext links, this recent technology can be use in thematic analysis knowledge extraction. In this paper, we shall present an illustration of thematic analysis from a mathematical text classifier applied on Descartes' s philosophical tex*Le discours de la méthode*.

RÉSUMÉ

La classification mathématique des textes s'avère très utile dans le domaine de la lecture et de l'analyse de textes assistées par ordinateur (LATAO). Avec l'accroissement constant du nombre de textes disponibles sur support numérique (CD-Roms, les encyclopédies, etc.), l'utilisation de classifieurs textuels est devenu un outil essentiel dans le domaine des technologies du traitement du langage naturel aidant le lecteur à découvrir l'information présente dans le texte. De récentes recherches présentent de nombreuses applications issues de la classification mathématique. En plus de l'analyse lexicale ou de la génération automatique de liens hypertextes, cette nouvelle technologie peut être utilisée dans le domaine de l'analyse thématique. Dans cet article, nous présenterons la classification mathématique des textes dans son application à l'analyse thématique du texte philosophique de Descartes *Le discours de la méthode*.

Mots-clés : classification mathématique, analyse thématique, texte philosophique, application.

1. Introduction

La *classification* mathématique des textes est devenue un important lieu de recherche dans le domaine du traitement informatique du langage naturel. Le développement de l'Internet, des CD-Roms contenant de larges corpus textuels ou des encyclopédies numérisées plongent le lecteur dans une mer d'informations dont l'exploration et l'analyse deviennent extrêmement ardues, voire impossibles. Pour accomplir ces tâches, il devient donc nécessaire d'explorer de nouvelles approches d'aide à la lecture et d'analyse de texte assistées par ordinateur (LATAO). Les outils à la disposition du lecteur doivent non seulement l'accompagner tout au long du texte, mais doivent aussi précisément l'assister dans son processus cognitif de découverte du contenu sémantique. Dans la présente recherche, nous explorerons un cas particulier de la classification des textes dans son application à l'analyse thématique.

L'analyse thématique

Un des problèmes majeurs du lecteur utilisant l'ordinateur à des fins d'analyse est son ignorance du contenu potentiel du texte. Ainsi, en raison de cette situation, il faut lui offrir des

outils d'exploration heuristiques de recherche de contenu. Dans cette perspective, certaines approches (Lebart et Salem : 1988, Salton : 1989, Church et Hanks : 1990, Reinhart : 1994, etc.) ont exploré la pertinence d'utiliser des classifieurs numériques pour structurer certaines informations du texte sous forme de réseaux lexicaux, de réseaux sémantiques, etc. Mais ces classifications privilégient surtout la présentation de ces classes lexicales sous la forme de graphes. Ce type de présentation est pertinent, mais il devient cognitivement difficile à maîtriser s'il représente d'un seul coup l'ensemble des interactions de tous les termes retenus pour un corpus donné. Aussi faut-il tenter d'explorer de manière plus ciblée les résultats de ces classifieurs. De récentes recherches en sémantique latente (Deerwester : 1990, Lewis et Gale : 1995) les ont explorés dans la catégorisation automatique des textes en associant ces classes lexicales à des plans généraux de classification. D'autres les utilisent pour de l'indexation. Certains l'appliquent à la génération automatique de liens hypertextes (Balpe, Lelu, et Papy : 1996 ; Meunier, Remaki et Forest : 1999 ; Nault, Rialle, Meunier : 1999). Dans la présente recherche, nous explorons ces méthodes dans une optique d'analyse thématique dynamique. Par analyse thématique de texte, nous entendons le parcours heuristique de découverte du contenu conceptuel qu'un lecteur réalise dans son parcours de lecture. C'est dans cette visée que nous appliquons les méthodes de classification numérique au texte.

En termes logiques, la classification des textes est définie comme une opération qui, appliquée à des segments de texte, identifie des classes d' équivalence entre ces segments eut égard à leur contenu informationnel (mots, n-gram, etc.). En cela, elle est différente de la catégorisation des textes qui associe aux classes trouvées une étiquette quelconque. Bien que les mots soient les descripteurs des segments textuels, les classifieurs sont appliqués aux segments et non aux mots comme tels. La classification ne représente pas une fin en soi. Ce traitement classificatoire doit constituer une étape précise dans un processus cognitif beaucoup plus complexe lié à la lecture et l' analyse de textes assistées par ordinateur (LATAO) (Meunier 1996 : 289-305).

2. Méthodologie

La première étape consiste à préparer le texte numérisé afin qu'il soit analysable. Dans un premier temps, le texte est segmenté. Cette étape est fondamentale (Lebart, Salem : 1988), car elle identifie les unités lexicales qui seront comparées à des fins de classification. Dans la présente expérimentation, nous avons privilégié une segmentation de 100 mots par fragment. Celle-ci s'est avérée très efficace. Par la suite, une étape de filtration où le lexique est réduit pour ne retenir que les mots pertinents. Cette étape permet aussi de réduire au maximum la présence de bruit à l'intérieur du texte. Les critères de filtration sont : la fréquence et la longueur des termes, la sélection élective de termes à partir d'une base de données de termes identifiés préalablement, etc. Enfin, le texte subit une lemmatisation classique. Dans certains cas, une pondération est ajoutée aux unités choisies.

Suite à ces étapes de traitement du texte, vient la classification proprement dite. Celle-ci s'effectue à partir de la matrice constituée des segments de texte et des unités d' information textuelle (voir figure 1). Autrement dit, le texte est traduit dans un modèle de Salton (1989), lequel est de type vectoriel.

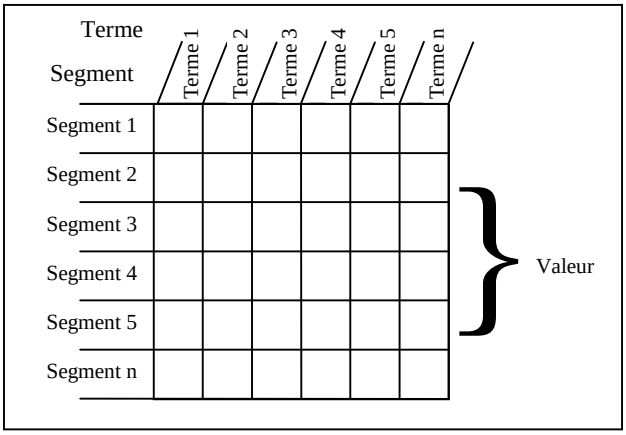


Figure 1. Matrice terme-segment

Les entrées de la matrice varient en fonction du critère sélectionné (absence, présence, poids, etc.). Les paramètres utilisés sont essentiellement de nature statistique, c' est-à-dire que l'on évalue l' occurrence (pondérée ou non) de chaque mot (du texte) en fonction de sa présence dans chacun des segments. Plusieurs types de classifieurs ont été explorés avec succès sur ce type de matrice. Ils vont de l'analyse en composante principale (Reinheirt : 1994, Lebart et Salem : 1988) aux K-means (Balpe, Lelu et Papy : 1996), aux réseaux de neurones (Veronis : 1990 ; Williams : 1990 ; Salton et Buckley : 1994 ; Kohonen : 1982 ; Nault et Meunier : 1999 ; Memmi et Meunier : 1997) et même aux algorithmes génétiques (Rialle, Ousedik, Nault et Meunier 1997). Dans la présente expérimentation, nous avons utilisé le classifieur ART I (Grossberg : 1988 ; Grossberg et Carpenter : 1991). Ce classifieur est de type réseau de neurones auto associatifs sans supervision. Par rapport aux autres approches, il possède l'avantage d'être dynamique, c' est-à-dire d'accepter de nouveaux intrants sans obliger la reprise du calcul ; ce qui le rend très pertinent pour une analyse thématique qui se déploie essentiellement de manière incrémentielle et dynamique. Le modèle ART impose cependant un mécanisme de contrôle sévère exprimé par un seuil de vigilance qui est une fonction discriminante spécifique. Le principe fondamental de ce modèle repose principalement sur une interaction entre deux niveaux de neurones entrant en phase de résonance (voir figure 2).

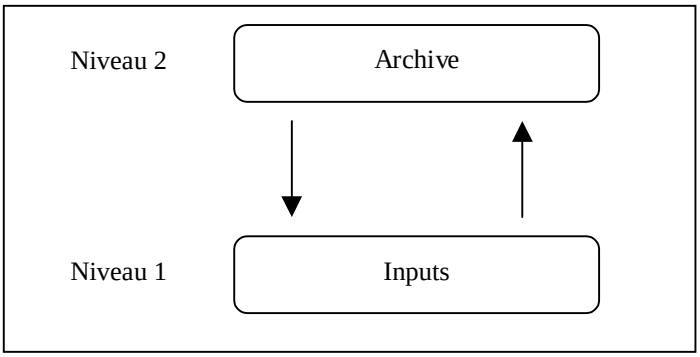


Figure 2

Le système reçoit, au niveau 1, les fragments du texte qu'il renvoie, par la suite, après attribution d'une pondération, au second niveau. Cette transmission du niveau 1 vers le niveau 2 implique une série complexe d'opérations. Cette opération de renvoie aura comme conséquence de créer une série de *patterns* au sein du niveau 2. Ces *patterns* serviront comme prototypes pour les entrées subséquentes au niveau 1. Le second niveau aura donc un impact significatif sur les nouvelles entrées qui se produiront au premier niveau. Ainsi, les nouvelles entrées au niveau 1 seront comparées aux prototypes en fonction d'un critère de résonance

RHO. Si la correspondance est positive, la nouvelle entrée sera incluse dans la classe déjà existante (prototype). Dans le cas inverse, cette nouvelle entrée constituera à son tour un nouveau prototype. La fonction d'adaptabilité émergera donc de la modification constante des connections entre les deux niveaux. La consolidation de cette résonance évoluera progressivement avec l'apprentissage.

Ainsi, le cœur du modèle ART consiste à mesurer la différence entre les segments intrants (input) en fonction du poids accordé entre les différents mots les constituant, et ceci en regard d'un seuil souvent sévère. Par conséquent, les segments appartenant à une même classe seront sélectionnés en raison d' un comparaison systématique entre eux. Lorsqu' une nouvelle entrée survient, le classifieur ART I l' insère parmi les classes préalablement formées. Il y a donc ajout d' un élément au sein de la matrice. Ce principe repose sur l' interaction entre deux niveaux de neurones entrant en phase de résonance. Les résultats, en terme informatiques, prennent la forme de classes de texte.

À la fin du processus, ART produit des classes de segments. Chaque classe C contient alors une liste de segments (S_1….. S_n), ce qui permet alors de construire un réseau lexical, lequel est constitué de l'intersection des termes entre les différents segments de (S_1….. S_n). Chaque segment n'appartient qu'à une seule classe. Sur le plan de l'interprétation, ceci signifie que cette classe contient une liste de lexèmes qui ont la caractéristique de se retrouver ensemble dans les fragments ainsi réunis par le classifieur. Ceci correspond en fait à une cooccurrence de lexèmes dont la co-présence est déterminée par un réseau de classification neuronale (voir figure 3).

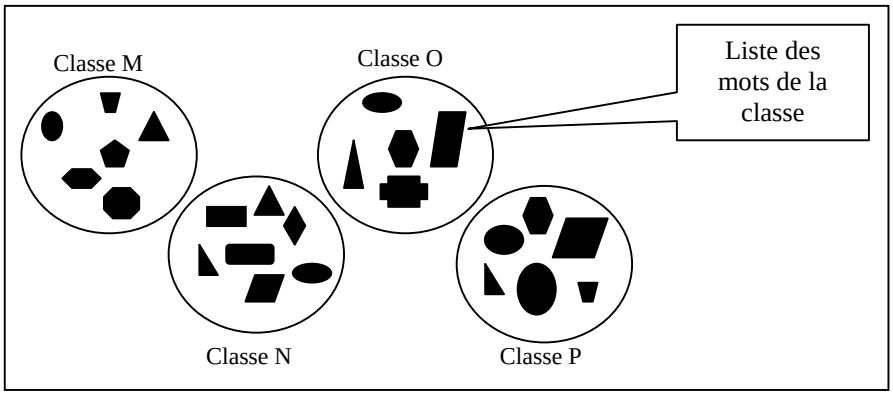


Figure 3. Représentation graphique des différentes classes

Dans cette figure, chaque classe est constituée d'une liste de lexèmes (représentés par une forme). Plusieurs cheminements sont alors possibles pour une analyse thématique. Nous en explicitons ici qu'un seul.

Certaines unités lexicales présentes dans une classe particulière peuvent se retrouver aussi dans un autre classe, indiquant par là qu'il opère dans un autre contexte. Ainsi, dans une classe de départ, le lecteur peut partir d'un terme choisi pour son intérêt thématique et « sauter » dans un autre classe ou le même terme se retrouve, mais cette fois dans un nouveau contexte. Ce contexte, à son tour, est constitué de lexèmes nouveaux qui peuvent servir de départ pour aller vers d'autres classes. Et ce processus recommence indéfiniment jusqu' à la clôture ou la saturation du parcours.

Ainsi, au bout de son parcours, le lecteur aura exploré son texte, de segments en segments, mais sans nécessairement connaître au préalable vers quel but. Le parcours est heuristique et

s'adapte aux résultats obtenus. Comme ART peut accepter de nouveaux documents, le calcul n'est pas à refaire à chaque fois. De nouveaux chemins sont alors possibles, ouvrant l'analyse thématique vers de nouveaux horizons (voir figure 4).

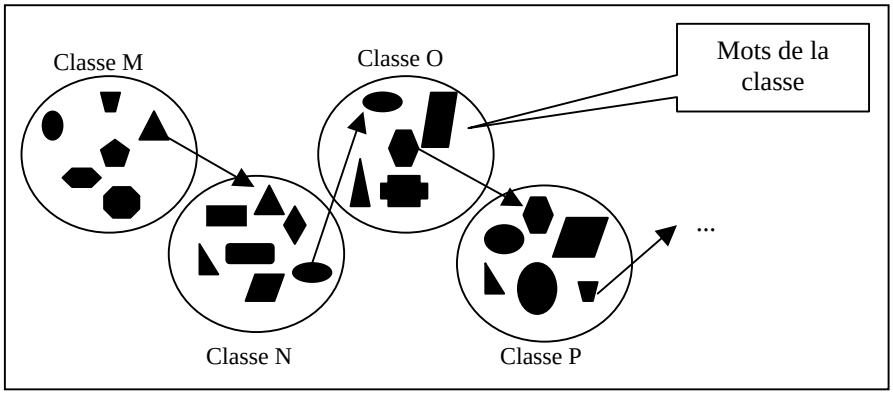


Figure 4. Représentation graphique de liens unissant les classes

3. L'expérimentation

L'expérimentation consistait à appliquer cette méthodologie à un texte spécifique pour en explorer la pertinence. Ce texte était le *Discours de la Méthode*, composé de 36 pages et de 21453 mots. Après une étape de segmentation du texte à raison de 100 mots par segments, nous avons procédé au filtrage et à la lemmatisation du texte. Suite à ces étapes, le texte de Descartes comportait 139 segments.

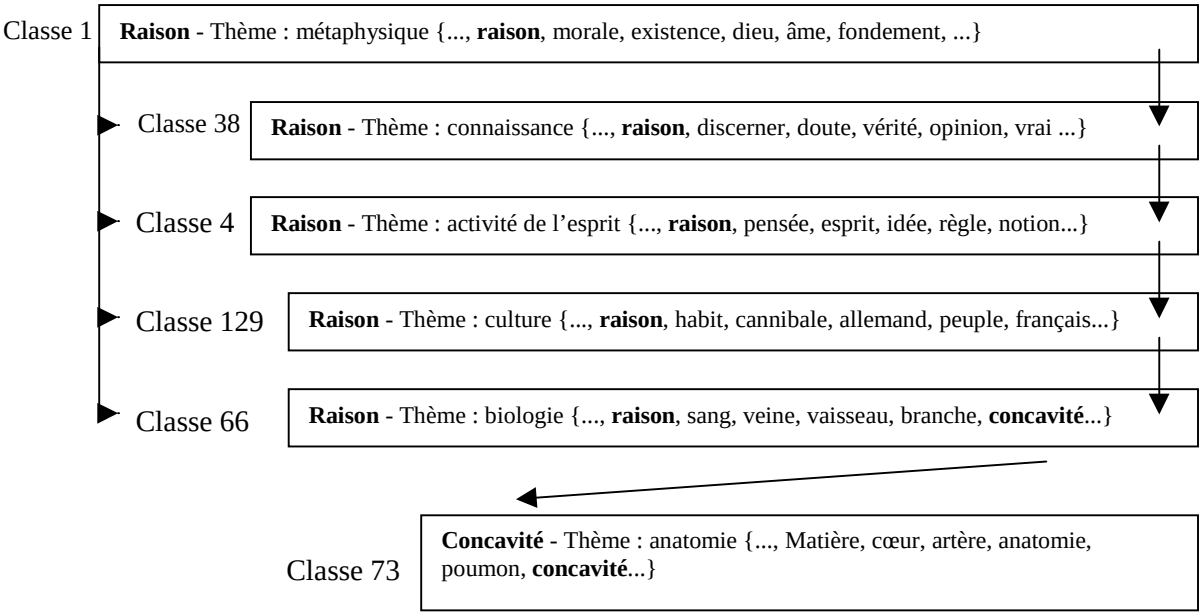
Toute la chaîne de traitement proposée pour ce parcours thématique est réalisée avec un logiciel développé spécifiquement à des fins de classification de texte: CONTERM. Ce logiciel offre à l'utilisateur des paramètres différents pour effectuer les diverses tâches d'analyse de texte. Mais surtout, il met a sa disposition plusieurs classifieurs adaptés à cette tâche.

4. Résultats

Tel qu'énoncé précédemment, les résultats de cette expérimentation nous ont permis de découvrir de nombreux liens thématiques possiblement pertinents pour le lecteur au sein du texte de Descartes. Notons toutefois que la pertinence de ces liens thématiques proposés, suite à la classification, relève, en dernière instance, au lecteur, en fonction de ces intérêts particuliers de recherche ou de lecture. Voici quelques échantillons de résultats liés à un choix particulier de thèmes.

Classe 1	***Classe 4***	***Classe 38***	***Classe 66***	***Classe 73***	***Classe 129***
Morale	Pensée	Vrai	Sang	Matière	Habit
Existence	Esprit	Vérité	Veine	Cœur	Coutumes
Dieu	Notion	Opinion	Goutte	Artère	Cannibale
Âme	Idée	Probable	Vaisseau	Anatomie	Allemand
Fondement	Règle	Français	Enfler	Poumon	Français
Vérité	**Raison**	**Raison**	Branches	Concavités	**Raison**
Raison	...	...	**Raison**	Animaux	Peuple
...			...	...	...

Ainsi, si nous partons du terme « raison », nous découvrons qu'il nous mène directement vers différents thèmes très particuliers de la philosophie cartésienne. Par exemple, le terme « raison » se retrouve dans un premier temps dans la classe 1. Le contexte de cette classe, lorsque l'on en interprète les autres termes et ceci malgré le bruit lexical qu' on y trouve, réfère principalement à certains segments du texte où Descartes traite des questions relatives à l'existence, à Dieu, à l'âme et aux fondements premiers de l'existence humaine. Toutefois, ce terme – utilisé aussi dans d'autres contextes – nous permet de découvrir le contenu du texte traitant aussi de l'activité de l'esprit (classe 4), de la connaissance (classe 38) et de la biologie (classes 66 et 73). Ainsi, le point de départ – le terme « raison » – , lorsqu'il est employé dans un contexte métaphysique (classe 1), nous permet d'explorer et de découvrir le thème de la connaissance (classe 38), de l'activité de l'esprit (classe 4), de la « culture humaine » (classe 129) et de la biologie (classe 66). Après avoir parcouru les différents contextes d'utilisation d'un même terme, le lecteur peut alors diriger sa lecture en fonction de ses intérêts. Ainsi, une fois « entré » dans la classe 66 (par l'entremise du terme « raison »), le lecteur peut choisir un second terme qui lui permettra d'explorer certains segments traitant d'une problématique particulière. Ainsi, selon le même principe, le terme « concavité » (présent dans les classes 66 et 73) - lorsqu'il est pris comme élément de départ vers un autre horizon de découverte, rend possible l'exploration du thème de l'anatomie du corps humain. Ce processus s'avère, selon notre hypothèse, applicable à l'ensemble du texte classifié. Bref, le lecteur dont l'activité est guidée par la classification mathématique du texte pourra, en fonction de ses champs d'intérêt propres, parcourir certains thèmes de la philosophie cartésienne. En voici un exemple :



C'est en partant du terme « raison » que le lecteur découvre, par le biais de cette stratégie classifiante, les principaux thèmes du *Discours de la méthode* associés, de près ou de loin, au terme « raison ». Ce processus de découverte du texte assisté par ordinateur s'avère donc un outil pertinent pour l'analyse du texte philosophique. En fonction du terme choisi par le lecteur comme point de départ de son exploration de texte, il sera amené à découvrir tous les thèmes reliés au terme choisi au départ.

5. Analyse des Résultats

Comme nous l' avons mentionné précédemment, l' usage des classifieurs varie en fonction de leur usage dans le cadre de la lecture et de l' analyse de textes assistées par ordinateur (LATAO). Le critère d' association de certains segments dans les différentes classes sera, par exemple, élevé ou faible, en fonction du besoin et de la tâche de l' utilisateur.

Dans le cas présent, nous nous sommes attardés uniquement à une seule application précise de la classification des textes : l' analyse thématique. Ce type d' analyse peut, selon nous, être efficacement assisté par le biais de la classification du texte lorsqu' il est soutenu par une technologie le permettant. Lors de notre expérimentation, les résultats nous ont clairement démontré de nombreuses relations entre les différents segments du texte. Certains liens « intratextes » furent découvert par le classifieur mathématique ART I. Nous avançons l'hypothèse que ce type de relation n'est pas restrictif et peut s'appliquer à l'ensemble du texte. Il y a donc possibilité d'établir un ensemble de liens thématiques regroupant certaines parties du texte. C'est en ce sens que l'application d'un classifieur mathématique s'avère un outil pertinent pour l'analyse thématique d'un texte. Le lecteur choisit un point d'ancrage de départ (un terme présent dans un segment d'une classe) à partir duquel il découvrira, par le biais de liens thématiques, un tout nouveau contenu conceptuel propre au texte.

6. Conclusion

La classification mathématique des textes est un outil des plus utiles pour la lecture et l'analyse de textes assistées par ordinateur (LATAO). Toutefois, il importe de bien distinguer l'analyse thématique de la génération automatique des liens hypertextes. La classification repose sur la découverte de liens intratextes. Elle postule implicitement que les segments similaires reconnus par le classifieur présentent un similarité sémantique quelconque et ceci malgré le bruit engendré par la méthode. Ce lien permet alors au lecteur de parcourir étape par étape les autres relations intratextuelles potentielles et ceci en fonction d'un terme d'entrée particulier et du projet de lecture. Une fois un lien reconnu, accepté et pertinent, le lecteur peut alors décider de le figer dans un lien hypertexte au sens technologique de ce terme, c' est-à-dire un lien HTML. Et l'ordinateur pourrait assister cette prise de décision en peaufinant la qualité de la relation via des algorithmes génétiques (Nault, Rialle Meunier, 1999). Autrement dit, tous les liens intratextuels proposés par le classifieur ne deviennent pas des liens hypertextes.

En contre partie, le générateur de liens hypertextes, procèdent à l' inverse. Il projette sur le texte un grand nombre de liens hypertextes HTML. Ceux-ci sont autant de chemins potentiels proposés à un lecteur. La véritable différence entre les deux approches nous semble reposer sur les perspectives cognitives. Les deux n'ouv rent pas sur les mêmes parcours cognitifs.

Références

Balpe, J.P., Lelu, A., Papy, F.(1996). *Techniques avancées pour l'hypertexte*. Paris: Hermes.

Carpenter, G. & Grossberg, G. (1991). *An Adaptive resonnance Algorithm for Rapid Category Learning and Recognition*.

Church K. W., Hanks P. (1990). *Word association norms, mutual information and lexicography*. Comp. Ling., Vol. 16, p. 22-29.

Deerwester, S., Dumais. S. T., Furnas, G. Landauer. T. K. Harshman. (1990). *Indexing by latent semantic analysis*, *Journal of the American Society for Information science*, 391-407.

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

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

Lebart, L., Salem, A. (1988). *Analyse statistique des données textuelles*. Paris, Dunod.

Lewis, D.L., Gale, W.A.A.A (1995). *Sequential Algorithm for training text Classfiers*. SIGIR, p. 2-11.

Memmi, D., Gabi, K., Meunier, J.-G., (1998). *Dynamical Knowledge extraction from texts by Art Networks. Proceedings of Neurap.* Marseille. 1998.

Meunier, J.-G., Remaki, L. Forest, D. (1999). *Use of classifiers in computer-assisted reading and analysis of text (CARAT).* Actes du colloque international CISST 1999, Las Vegas, Nevada, U.S.A.

Meunier, J.G., Biskri, I., Nault, G., Nyongwa M., Septembre (1997), "Exploration de classifieurs connexionnistes pour l' analyse de textes assistée par ordinateur", *Actes du Colloque LTT97*, Tunis, Tunisie. p 289-296.

Nault G., V. Rialle et J.G. Meunier (1999). PROGEN : a Genetic-Based Semi-automatic Hypertext Construction Tool - first steps and experiment. In Smith, R. E. (eds.). GECCO-99: Proceedings of the Genetic and Evolutionary Computation Conference, July 13-17, 1999, Orlando, Florida USA. San Francisco, CA: Morgan Kaufmann.

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

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

Rialle, G, V, Ousedik, Meunier, J. S, Nault. G. (1997). *Semiotics and Modeling Computer Classifications of Text with genetic Algortihm: Analysis and first Results*. In A. M Meystel (ed) ISAS: Int. Conf.. *On intelligent systems and semiotics. A Learning Perspective*. National Institute of standard and technology. (NIST) Washington, DC. p.57-60.

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

Salton, G., Allan, J., Buckley, C. (1994). *Automatic structuring and retrieval of large text file*. Communications of the ACM 37 (2), 97-107.

Veronis, J., Ide, N.M., Harie, S. (1990). *Utilisation de grands réseaux de neurones comme modèles de représentation sémantiques*. Neuronimes.

Williams, M. (1990) *Connectionist models and information retrieval*. 25, 209-259.