Classification et catégorisation automatiques : application à l'analyse thématique des données textuelles

Dominic Forest, Jean-Guy Meunier

UQÀM – LANCI – C.P. 8888, Succ. Centre-Ville – Montréal – Québec – Canada, H3C 3P8
forest.dominic@courrier.uqam.ca, meunier.jean-guy@uqam.ca

Abstract

Since a few years, several literature and humanities research projects have tried to integrate automatic data-processing dimensions into their objectives. In spite of each project's specificities, most of the projects' objectives concern the comprehension and the automatic processing of thematic analysis of textual data. In this paper, we present a data processing sequence adapted to thematic analysis of textual data. The specificity of this data processing sequence lies in its use of data classification and automatic categorization techniques. We present results of an experiment on a philosophical corpus.

Résumé

Depuis quelques années, plusieurs projets de recherche dans les domaines des sciences humaines et des lettres ont tenté d'intégrer des dimensions informatiques à leurs objectifs. Malgré les spécificités propres à chacun de ces projets recherche, on constate qu'un axe de recherche partagé par l'ensemble des disciplines sensibles à l'analyse de textes par ordinateur relève de la compréhension et de l'informatisation du processus d'analyse thématique des données textuelles. Dans cet article, nous présentons une chaîne de traitement adaptée à l'analyse thématique des données textuelles. La spécificité de cette chaîne de traitement réside dans son utilisation de techniques de classification et de catégorisation automatiques des données textuelles. Nous présentons les résultats d'une expérimentation sur un corpus de textes philosophiques.

Mots-clés : catégorisation, classification, analyse thématique, lecture et analyse de textes assistées par ordinateur.

1. Introduction

Depuis environ trente ans, mais surtout durant les dix dernières années, plusieurs recherches menées dans les domaines des sciences humaines et des lettres ont tenté d'intégrer des dimensions informatiques à leurs objectifs. Grâce à ces efforts d'intégration technologique, les sciences humaines ont su développer plusieurs méthodologies et applications d'analyse de textes assistées par ordinateurs. Parmi les types d'applications les plus fréquemment cités, on trouve entre autres ceux portant sur l'analyse qualitative et quantitative des données (Alexa et Zuell, 1999a et 1999b), sur l'analyse de contenu assistée par ordinateur et, de manière plus générale, sur l'analyse des données textuelles et sur la Lecture et l'Analyse de Textes Assistées par Ordinateur (LATAO) (Meunier, 1997 ; Popping, 2000).

D'autre part, le domaine des lettres et de la littérature a, lui aussi, été le lieu de réflexions théoriques et d'efforts pratiques visant à permettre l'émergence d'applications informatiques adaptées aux études littéraires (Rastier *et al.*, 1995 ; Bernard, 1999 ; Kastberg Sjöblom et Brunet, 2000). Ainsi, la rencontre entre les littéraires et les informaticiens a été le lieu d'émergence d'un nouvel axe de recherche dont les principales manifestations, inspirées de travaux en statistique et en mathématique, ont pris la forme de méthodologies d'analyse et de

logiciels destinés spécifiquement à l'analyse de textes littéraires (Bernard, 1999 ; Hockey, 2000).

Malgré les spécificités propres à chacun de ces projets recherche, on constate qu'un axe de recherche partagé par l'ensemble des disciplines sensibles à l'analyse de textes par ordinateur (tant dans le domaine des sciences humaines que des lettres) relève de la compréhension et de l'informatisation du processus d'analyse thématique des données textuelles (Louwerse et van Peer, 2002). En effet, comme le souligne Popping (2000), « in case one is interested in concept occurrences, one might use thematic analysis. This is the kind of analysis that is still applied most today » (p. ix). Ainsi, nombreux sont les chercheurs qui, à l'instar de Popping, n'hésitent pas à exprimer l'importance présentement attribuée à l'analyse thématique des documents. Pettersson (2002 : 237) est, à cet égard, des plus explicites: « [Thematics is] one of the most rapidly expanding fields in the arts, and literary studies in particular. »

2. La multidisciplinarité de la problématique de l'analyse thématique

Bien que la problématique de l'analyse thématique puise ses racines dans les travaux de Platon et d'Aristote (Sollors, 1993), plusieurs des principales recherches concernant cette problématique furent réalisées au cours du 20^{e} siècle. Parmi les principales réalisations contemporaines qui ont fortement modélisé et contribué à la compréhension des thèmes (identification et organisation) figurent celles de Tomashevsky (1925) sur la consistance et la décomposition (causale et temporaire) des thèmes en motifs (ces derniers constituant les plus petits composants d'un thème), de Aarne et Thompson (1928) sur les index de motifs, de Thompson (1946) sur la classification des motifs, de Propp (1928) sur la nécessité de développer une méthode formelle et structurée de segmentation basée sur une approche multidisciplinaire, afin d'identifier les différents motifs présents dans un texte, etc. Ces travaux fondateurs constituent les assises théoriques sur lesquelles reposent les récentes contributions à l'analyse thématique, lesquelles ont aussi largement été influencées par la distinction élaborée par l'école linguistique de Prague (principalement grâce aux travaux de N. S. Troubetskoï, de R. Jakobson et de S. O. Kartsevski) entre les concepts de « thème » (l'objet du discours) et de « rhème » (l'information relative au thème). Ce qui caractérise toutefois les principaux travaux actuels portant sur la question de l'analyse thématique concerne le niveau d'analyse, la manière dont les auteurs ont fréquemment abordé la question. En effet, tant en littérature qu'en linguistique, la majorité des travaux ont traditionnellement abordé la problématique du thème à partir d'analyses phrastiques en tentant d'identifier les éléments, principalement linguistiques, permettant de comprendre l'organisation thématique des textes (Rastier *et al.*, 1995 ; Prince, 1985 ; Louwerse et van Peer, 2002).

Malgré la richesse de la perspective linguistique, les travaux de Van Dijk (1972), Kintsch et Van Dijk (1978) et Van Dijk et Kintsch (1983) ont aussi mis en relief une approche fondée sur la distinction entre les niveaux micro-structurels (reposant, de manière générale, sur l'analyse de la phrase) et macro-structurels (dont l'objet d'analyse réside dans l'ensemble du texte, pris comme un tout cohérent et structuré) du texte. Cette approche, fondée sur des travaux provenant de plusieurs disciplines des sciences humaines semble des plus fécondes. Selon cette perspective, l'analyse thématique des textes relève d'un effort d'abstraction se situant au niveau macro-structurel et est régie par quatre principes : 1) la suppression de l'information non pertinente, 2) la sélection de l'information pertinente, 3) la généralisation des propositions retenues et 4) l'intégration des propositions dans un tout structuré et cohérent (Louwerse et van Peer, 2002). Cette théorie qui a l'avantage de « contextualiser » l'identification et l'interprétation des thèmes d'un texte a, en outre, motivé l'application de la théorie

de la sémantique latente, des calculs de cooccurrences et de corrélation au domaine de l'analyse thématique afin d'assister le chercheur dans l'interprétation des textes auxquels il est confronté. La présente recherche s'inspire partiellement de ces quatre principes en tentant toutefois d'en dépasser les limites. L'objectif de cette recherche consiste à développer une hypothèse méthodologique permettant de mettre en valeur la complémentarité des approches reposant, d'une part, sur une analyse micro-sémantique de la phrase et, d'autre part, sur une analyse macro-sémantique de la textualité (Rastier *et al.*, 1994 ; Rastier, 1996 et 2001).

On constate également qu'à travers leurs diverses perspectives d'analyse, les récents travaux dans le domaine de l'analyse thématique ont clairement démontré l'importance de la variété de points de vue que peuvent apporter les différentes disciplines concernées par cette problématique. Comme le soulignent Louwerse et van Peer (2002 : 9), « it seems more likely that in many disciplines thematics has always occupied a place but we may not have recognize it as such. » Suivant Rimmon-Kenan (1985), il semble donc nécessaire d'opter pour une approche pluridisciplinaire fondée non pas exclusivement sur des considérations linguistiques, mais plutôt sur une perspective plus large permettant de tenir compte tant des phénomènes linguistiques en jeu dans le texte que des phénomènes relevant de la textualité (processus discursif, pragmatique, etc.) sur lesquels reposent l'organisation et la structure des divers thèmes d'un corpus. Comme le soutient Giora (1985) : « la notion de thème (*topic*) discursif est indépendante de la notion de thème (*topic*) phrastique ». « Therefore, instead of describing thematics as an "undisciplined discipline" we prefer describing it as an interdisciplinary discipline, working at different levels in different areas but with a unified goal. » (Louwerse et van Peer, 2002 : 9)

3. Objectif de recherche

L'objectif général de la recherche que nous présentons ici consiste à explorer certaines méthodes de classification et de catégorisation automatiques à des fins d'analyse thématique de textes théoriques. Il vise ainsi à effectuer un transfert de concepts et de méthodologies provenant, d'une part, des recherches théoriques sur l'analyse thématique et, d'autre part, des recherches appliquées sur la catégorisation des données (intelligence artificielle, apprentissage machine, forage de textes, classification, etc.) afin de développer diverses méthodologies et applications visant à assister le chercheur en sciences humaines et en littérature dans son travail d'analyse thématique des textes. Bien que cet objectif de recherche se déploie en plusieurs volets (modélisation formelle de concepts reliés au domaine de l'analyse thématique, etc.), nous nous attarderons essentiellement dans le cadre de cet article à la présentation détaillée et à l'application exploratoire d'une chaîne de traitement spécifique adaptée à l'analyse thématique des données textuelles. Cette recherche – laquelle approfondit certaines des réflexions déjà présentées antérieurement (Forest, 2002 ; Forest et Meunier, 2000) – ne vise donc pas à procéder à une validation en profondeur de cette méthodologie d'analyse, mais plutôt d'explorer le potentiel et la fécondité d'une telle démarche.

4. Méthodologie

4.1. *La plate-forme SATIM*

Au niveau informatique, la réalisation du projet présenté consiste à concevoir et à développer une chaîne de traitement (nommée Thématico) adaptée spécifiquement à l'analyse thématique des documents textuels. Ce projet repose sur la technologie de la plate-forme informatique SATIM (Système d'Analyse et de Traitement de l'Information Multidimensionnelle). Cette plate-forme consiste en une interface permettant de structurer des modules informatiques

indépendants, de les organiser, de les faire communiquer entre eux et de les appliquer à divers domaines de traitement de l'information. La plate-forme SATIM offre des outils pour explorer des modules de traitement déjà existants afin de les agencer de façon à atteindre un but particulier (en l'occurrence la découverte et l'analyse des thèmes présents dans un corpus textuel).

4.2. *La chaîne de traitement Thématico*

La chaîne de traitement Thématico constitue une opérationnalisation informatique de l'analyse thématique (l'assistance à la découverte des principaux thèmes d'un corpus, ainsi que l'exploration et la navigation entre ces derniers) permettant d'assister le chercheur dans sa tâche d'analyse thématique des textes. La chaîne de traitement respecte une architecture classique en sept étapes.

1) L'identification des unités d'information pertinentes. Cette première étape vise à extraire du corpus les unités d'information, c'est-à-dire les unités linguistiques servant d'ancrage à l'analyse. Ces unités peuvent être des mots ou des chaînes de caractères (n-grammes) (Cavnar et Trenkle, 1994) et les domaines d'information (paragraphes, chapitres, etc.) qui seront analysés ultérieurement dans la chaîne de traitement. Certaines opérations linguistiques (suppression des mots fonctionnels, lemmatisation, étiquetage morphologique et sémantique) ou statistiques sont aussi appliquées au lexique extrait.

2) La vectorisation. Suite au découpage du corpus en unités d'information et en domaines d'information, le texte est traduit en une matrice de vecteurs (modèle vectoriel) (Salton, 1989 ; Manning et Schütze, 1999) qui représente alors chaque domaine d'information par la présence ou l'absence (binaire ou floue) des unités d'information (figure 1).

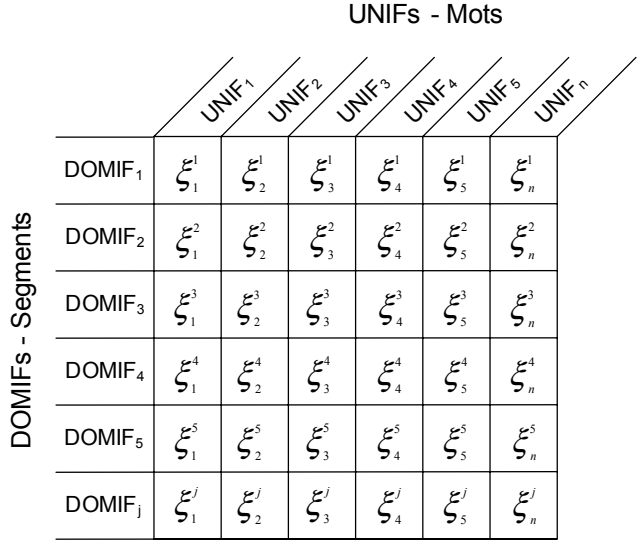

Figure 1. La matrice domaines d'information / unités d'information

3) La classification des segments. Par la suite, des classifieurs numériques sont appliqués sur la matrice. Par classifieurs numériques, nous entendons les stratégies mathématiques qui permettent la production de classes d'équivalence sur des segments. Plusieurs techniques numériques ont déjà été explorées sur le texte. Malgré certaines limites, ces approches ont présenté des résultats très positifs et se comparent avantageusement aux approches exclusivement linguistiques (Salton, 1989). Elles permettent de plus une immense économie de temps dans le parcours exploratoire d'un corpus. Technologiquement, ces approches sont incontournables

lorsqu'elles sont confrontées à de vastes corpus textuels. Dans la présente recherche, en raison de la technologie explorée et développée antérieurement par l'équipe du LANCI, nous avons privilégié comme classifieur le réseau de neurones Art (Grossberg et Carpenter, 1988).

4) La production de sous-classes lexicales différenciées. Les classifieurs produisent des regroupements (classes) de segments en raison de propriétés similaires. De ces classes, on extrait alors le lexique. Par des opérations ensemblistes simples (intersection, union, etc.), on différencie divers types de sous-classes spécifiques.

5) La catégorisation thématique. Ici, l'hypothèse est que certaines techniques de classification et de catégorisation automatiques peuvent être utilisées afin d'identifier l'organisation thématique des documents. Sur les sous-classes lexicales différenciées (étape 4) peuvent être appliquées diverses techniques de catégorisation automatique. Traditionnellement (Sebastiani, 1999 et 2003 ; Jackson et Moulinier, 2002), la catégorisation consiste à attribuer une catégorie aux domaines d'information à partir d'un ensemble prédéfini de catégories. Cette catégorisation associe une étiquette catégorielle à chaque sous-classe lexicale différenciée. Ce prédicat peut soit résumer le contenu signifiant (catégorie descriptive), soit en définir la fonction (catégories fonctionnelles). Cette technique de catégorisation requiert de prendre le vecteur représentant chaque classe et de le comparer à un gabarit (l'ensemble des catégories prédéfinies). Ceci peut se faire de trois manières. Une première méthode est manuelle : c'est l'analyste qui, en fonction de son répertoire propre, attribue la catégorie à assigner. Une seconde est automatique : les vecteurs des sous-groupes sont comparés (par un calcul de similarité) à une définition (en extension) d'une liste catégorielle ou d'un plan de classification, c'est-à-dire un ensemble de catégories prédéfinies. Une troisième approche procède par apprentissage : l'analyste assigne des catégories sur des échantillons de textes et le système les redistribue sur les items lexicaux ayant des contextes similaires. Dans cette recherche, nous avons voulu expérimenter une technique de catégorisation thématique fondée sur l'extraction automatique des catégories à partir des documents traités. Comme l'ont souligné entre autres Louwerse et van Peer (2002 : 4), les index thématiques (catégories thématiques prédéfinies) posent plusieurs problèmes : « The index was conceived to be a practical reference, but trying to classify tales in the […] index proved problematic. » Afin de dépasser les limites de la catégorisation automatique effectuée à partir d'ensembles de catégories thématiques prédéfinies, nous exploitons certains outils statistiques permettant de faire émerger les catégories thématiques à partir des documents analysés. Cette méthode consiste à appliquer certains critères statistiques utilisés dans les domaines du repérage de l'information (pondération distribuée, tf · idf, taux d'information, entropie, etc.) (Salton, 1989 ; Yates et Ribiero, 1999) à chacune des sous-classes lexicales différenciées afin d'identifier au sein de chacune de ces sous-classes les termes les plus significatifs pouvant (suite à une évaluation de l'utilisateur) servir d'étiquette thématique pour la découverte des principaux thèmes d'un corpus.

6) Projection des catégories thématiques sur le texte. Une fois la catégorisation thématique effectuée, chaque domaine d'information peut se voir étiquetée automatiquement avec les étiquettes thématiques. Le texte est alors soumis à des analyses classiques soit qualitatives (regroupements, listes, arbres, graphes, etc.) soit quantitatives (statistiques, etc.), mais, cette fois, ce sont les catégories thématiques qui en sont l'objet.

7) La découverte, la navigation et la visualisation des thèmes identifiés. Pour assister l'analyse et l'interprétation des résultats, il est de plus en plus utile d'offrir aux analystes des moyens de visualiser de manière ergonomique ces classes et les relations entre les thèmes. Ce sont, comme le dit Barry (1998), des « *mind mapping tools* » ou des cartographies cognitives du contenu thématique des textes. Diverses technologies commencent à apparaître pour

assister ce type d'analyse (Spence, 2000 ; Fayyad, Grinstein et Wierse, 2001). Cette dimension ergonomique de représentation est un atout précieux dans le soutien de l'activité interprétative du chercheur.

5. Expérimentation

Dans le cadre de l'expérimentation présentée, nous avons appliqué la méthodologie décrite précédemment à un texte spécifique afin d'en explorer la pertinence. Nous présentons ici les résultats obtenus à partir du *Discours de la méthode* de Descartes.

Le lexique initial du *Discours de la méthode* est constitué de 2914 termes dont la fréquence varie entre 1 et 925 occurrences. Nous avons effectué un filtrage du lexique où furent supprimés les termes fonctionnels et ceux dont la fréquence a été jugée non pertinente pour la classification (c'est-à-dire les termes trop fréquents et pas assez fréquents). De plus, le lexique fut lemmatisé. Suite à ces opérations, le lexique épuré fut composé de 306 termes. La fréquence des termes retenus varie entre 5 et 56 apparitions. De plus, nous avons privilégié une segmentation par mots, à raison de 150 mots par segment. Pour la classification, nous avons utilisé le classifieur neuronal ART1.

6. Résultats

Le processus de segmentation a permis de découper le corpus initial en 154 segments de 150 mots. En utilisant le classifieur ART1, ces 154 segments furent regroupés en 83 classes. D'un point de vue strictement technique, la qualité des résultats de cette classification (moyenne de 1.86 segment par classe) est manifestement discutable. Cependant, l'objectif de notre recherche ne consiste pas à valider le processus de classification, mais bien la fécondité de notre démarche à l'égard d'un processus interprétatif beaucoup plus complexe lié à l'analyse thématique des données textuelles. Dans cette optique, les résultats obtenus s'avèrent acceptables afin d'atteindre, comme nous le verrons, notre objectif.

L'identification des principaux thèmes présents dans le corpus est effectuée en appliquant certains calculs statistiques permettant de faire émerger les catégories thématiques à partir des documents analysés. Dans le cadre de la présente expérimentation, nous avons uniquement appliqué la formule classique tf · idf (« *term frequency · inverse document frequency* ») (Salton, 1989) au lexique de chaque classe. Le principe de ce calcul peut être formulé de la manière suivante : un terme sera d'autant meilleur pour représenter le contenu d'une classe s'il est à la fois fréquent dans cette classe et rare dans l'ensemble des classes à analyser. La fréquence inverse du document, idf = log(N/n), où N est le nombre total de documents et n est le nombre de documents contenant le terme, vient donc modérer ou accentuer l'importance de la fréquence de chaque terme. Ainsi, ce calcul est utilisé, dans le cadre de notre analyse, afin d'extraire les termes les plus représentatifs des classes obtenues. Les termes retenus suite à ce calcul sont alors attribués comme « étiquette thématique » à leur classe respective.

Les processus de découverte et de navigation thématique débutent par le choix d'un terme particulier présent dans le corpus à analyser. Ce choix est effectué en fonction des intérêts – de recherche ou de lecture – du chercheur. À titre d'illustration, en sélectionnant le terme « connaissance », on constate que ce dernier opère dans plusieurs regroupements de segments. Ce terme est en effet présent dans des classes de segments (desquels la chaîne de traitement utilisée nous a permis d'extraire le lexique) où se retrouvent les autres termes suivants : « animal, artère, bête, branche, cave, chaleur, cœur, concavité, mouvement, organe, poumon, sang, veine, etc. » (classes 22, 41, 44 et 46, segments 94, 93, 96, 97, 103, 104 et 112), « air,

astre, ciel, lumière, matière, monde, terre, etc. » (classes 1, 24, 38 et 40, segments 1, 49, 80, 81, 84 et 86), « démonstration, entendement, géomètre, mathématique, méthode, philosophie, science, vrai, etc. » (classes 6 et 17, segments 32 et 37), « âme, astre, certitude, dieu, entendement, esprit, existence, idée, grand, parfaire, etc. » (classes 31, 32 et 34, segments 63, 64, 66, 67, 70, 71 et 73). Ces informations nous indiquent que le terme retenu opère dans plusieurs contextes différents. Compte tenu du vocabulaire spécifique de chacun de ces contextes (représentés par des classes de segments), nous pouvons affirmer que chacun de ces regroupements représente en fait un thème présent dans le *Discours de la méthode*. Le passage de la liste des termes du lexique de chaque classe vers l'identification et l'attribution d'un thème à une classe particulière s'effectue en identifiant les termes les plus représentatifs de chaque classe (selon les valeurs obtenues par le calcul tf · idf).

Ainsi, à partir du terme retenu, il nous est possible, dans un premier temps, d'explorer les segments du *Discours de la méthode* étiquetés des termes thématiques suivants : « artère, cœur, peau, sang». En effet, sur la base de leur lexique respectif, on constate que les classes 22, 41, 44 et 46 traitent de ces aspects particuliers de la philosophie de Descartes. Mais, il nous est aussi possible, toujours à partir de ce même terme, de découvrir un tout autre thème présent dans le corpus. En effet, le regroupement constitué des classes 1, 24, 38 et 40 est, quant à lui, caractérisé par les termes suivants : « astre, ciel, matière, physique ».

Si, en contrepartie, notre analyse se dirige vers les classes 6, 17 et 31, 34, il nous est alors possible de découvrir les extraits thématiques dont les termes représentatifs sont les suivants : « entendement, démonstration, géomètre, mathématique » (classes 6 et 17) et « certitude, dieu, entendement, existence » (classes 31 et 34).

Ces résultats semblent en majeure partie concorder avec ceux obtenus lors d'expérimentations antérieures (Forest, 2002 ; Forest et Meunier, 2000) au cours desquelles les termes thématiques de chaque classe ont été inférés subjectivement. En effet, dans le cadre de ces expérimentations, les classes présentées ci-haut s'étaient vu attribuer les catégories thématiques suivantes : « biologie » (classes 22, 41, 44 et 46), « physique » (classes 1, 24, 38 et 40), « mathématique » (classes 6 et 17) et « métaphysique » (classes 31, 32 et 34). Il est a noter cependant que la catégorisation thématique, lorsque qu'effectuée manuellement par le chercheur, semble caractérisée par un plus haut niveau de généralité. Cette observation est entre autres explicable par le fait que l'analyse n'est pas limité dans sa catégorisation uniquement aux termes présents dans le lexique du corpus.

On remarque donc que le terme « connaissance » que nous avons utilisé comme éléments clefs dans un processus de découverte thématique du *Discours de la méthode* nous guide vers un premier niveau d'analyse composé de quatre axes principaux (figure 2). Ces quatre regroupements thématiques constituent d'ailleurs des éléments classiques de la pensée de Descartes (Rodis-Lewis, 1966 et 1984).

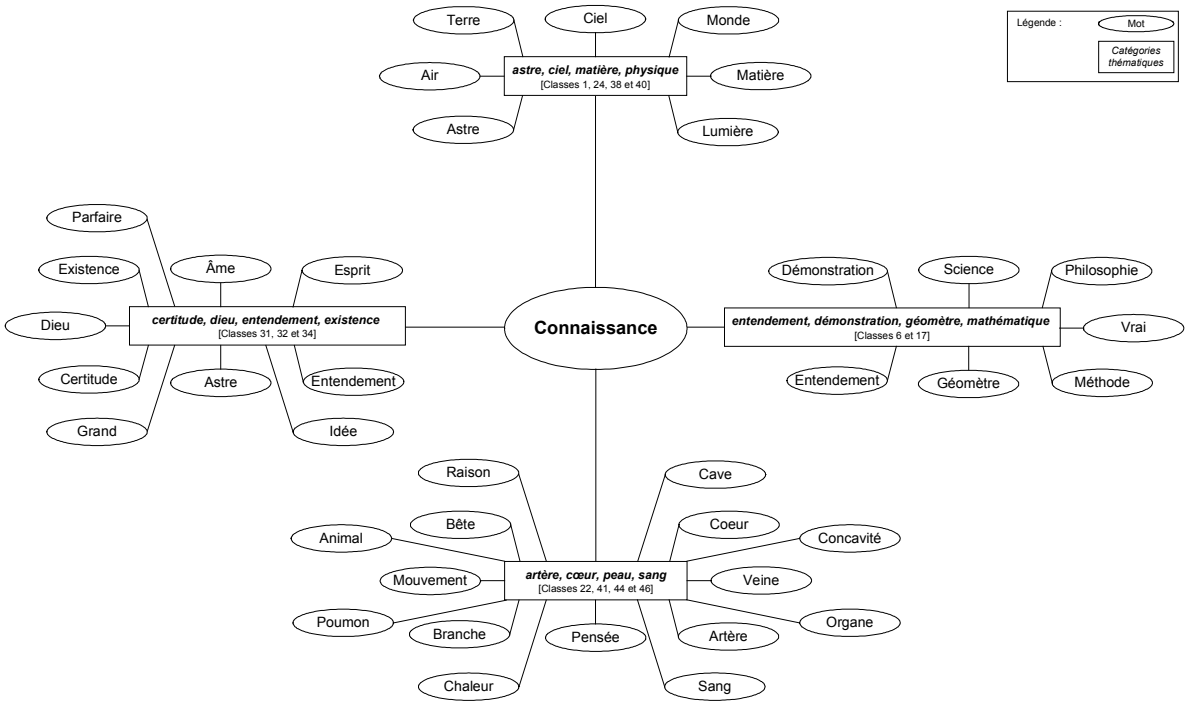

Figure 2. Échantillon des résultats thématiques obtenus à partir du terme « connaissance » du Discours de la méthode

Toutefois, certaines unités lexicales (c'est-à-dire certains mots) présentes dans une classe particulière peuvent se retrouver aussi dans une autre classe, indiquant par là qu'elles opèrent dans un autre thème. Ces mots se retrouvant dans plusieurs classes ou regroupements thématiques jouent un rôle de liaison dans l'exploration des multiples thèmes d'un corpus. C'est, en effet, sur la base de ces mots que le chercheur est amené à découvrir les contenus thématiques du corpus. Ainsi, dans une classe de départ, le lecteur peut partir d'un terme choisi et diriger son analyse vers une autre classe thématique où le même terme se retrouve, mais cette fois dans un nouveau contexte. Ce contexte, à son tour, est constitué de lexèmes qui peuvent servir de départ pour aller vers d'autres classes thématiques. C'est dans la découverte successive de plusieurs classes thématiques que consiste la navigation thématique.

Ce processus de découverte et de navigation thématique n'est pas fermé. Il se poursuit indéfiniment jusqu'à la clôture (c'est-à-dire lorsque l'analyste découvre une classe dont le lexique ne comporte pas de terme(s) servant de lien(s) vers d'autres classes thématiques) ou la saturation du parcours (c'est-à-dire lorsqu'il aura parcouru l'ensemble des thèmes de son corpus).

7. Analyse

Dans l'expérimentation présentée, la navigation thématique est perçue comme un processus de découverte et de parcours des différents thèmes présents dans un corpus textuel. Cette démarche est potentiellement multiple et fort complexe. Elle repose en dernière instance sur plusieurs choix, tant théoriques que pratiques. Mais, de manière générale, la navigation thématique consistera en un parcours caractérisé par un compromis entre, d'une part, les attentes du chercheur et, d'autre part, les indices sémiotiques présents dans le texte. Comme le souligne (Bremond, 1985 : 420) : « par quoi suis-je orienté dans la série de mes choix ? On peut

répondre : par le désir d'isoler la ou les bonnes formes du thème. Mais qu'est-ce qu'une bonne forme ? […] la bonne forme, c'est celle qui procure la satisfaction la plus grande à mon attente de lecteur […]. » Pour d'autres (Prince, 1985), cette attente du chercheur prendra le nom de « réalité extra-textuelle ». D'ailleurs plusieurs théoriciens ont noté l'importance, dans l'activité de thématisation et de découverte des contenus thématiques, de cette composante essentiellement subjective.

Dans le cadre de cette recherche, cette composante subjective se manifeste dans l'intérêt du chercheur envers certains thèmes qu'il privilégie et dans les objectifs qu'il désire atteindre lors de son analyse. Ainsi le chercheur peut, par exemple, choisir d'explorer un ou plusieurs thèmes précis du corpus, et ce dans le but d'en démontrer l'organisation, la structure, etc. Peut-être voudra-t-il explorer l'ensemble des thèmes d'un corpus afin d'orienter ou de cibler les passages qu'il voudra analyser plus en détail par la suite.

Mais d'autre part, une seconde contrainte, plus objective cette fois, entre aussi en compte dans le cadre de la tâche d'analyse et de découverte. Cette contrainte repose sur le texte à analyser. Cette composante intra-textuelle limite nécessairement la liberté de l'interprète, car elle guide inévitablement l'ensemble des analyses. En effet, malgré les intérêts et les raisons qui mènent le chercheur vers la découverte d'un thème particulier plutôt que d'un autre, le chercheur ne crée pas les thèmes dans le corpus qu'il analyse. C'est le texte qui expose, à l'aide des différents porteurs sémiotiques qu'il comporte, les thèmes sur lesquels le chercheur posera éventuellement son analyse.

8. Conclusion

L'application de techniques informatiques de catégorisation permet d'assister le chercheur dans son travail d'analyse thématique des documents (Forest, 2002 ; Forest et Meunier, 2000 ; Rossignol et Sébillot, 2002). Cependant, au-delà de la technologie et des méthodes employées, les différentes techniques de catégorisation présentent des problèmes théoriques importants. En effet, il importe de comprendre davantage la nature « catégorisante » de l'activité d'analyse thématique des documents. Il importe, dès lors, de situer le processus de catégorisation à l'égard des théories qui le sous-tendent : la logique de la classification et de la catégorisation, la sémantique cognitive, la sémantique fonctionnelle, la sémantique interprétative des textes, l'analyse de texte, etc. L'hypothèse théorique qui traverse cette recherche liée à l'utilisation des technologies pour l'analyse thématique de texte est avant tout inspirée d'une herméneutique matérielle (Rastier *et al.*, 1994), c'est-à-dire que le texte, bien que présentant une structure linguistique définie, ne révèle son contenu thématique qu'en regard d'un projet (subjectif) de lecture. L'analyse et la lecture des textes assistées par ordinateur sont des opérationnalisations informatiques des actes d'interprétation effectués sur des textes par un analyste, et dont un des moments forts est la découverte de classes de régularités sémantiques, thématiques et discursives. Comme le souligne Martin (1995 : 18), « la nature de ce que l'on pourrait appeler plus généralement l'étude thématique des textes est d'abord fonction de l'objectif visé », et Prince (1985 : 432), « thématiser un texte dépend donc non seulement du ″texte même″ mais aussi (et peut-être davantage) du thématiseur, du cadre adopté, des unités choisies, des opérations accomplies pour les harmoniser, des résumés et paraphrases effectués. » Il faut donc nécessairement lors de l'évaluation des résultats de l'application établir un compromis entre, d'une part, les techniques classiques rigoureuses d'évaluation des résultats et, d'autre part, le caractère subjectif du travail d'analyse thématique (les objectifs d'analyse thématique et les choix théoriques effectués par le chercheur).

Références

Aarnes A. et Thompson S. (1928). The types of folk-tale: a classification and bibliography. *Folklore fellows communications*, vol. (74). Suomalainen.

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

Alexa M. et Zuell C. (1999b). *A review of software for text analysis*. ZUMA arbeitsbericht, ZUMA.

Baeza-Yates R. et Ribeiro B. d. A. N. (1999). *Modern information retrieval*. ACM Press / Addison-Wesley.

Barry C.A. (1998). Choosing qualitative data analysis software: Atlas/ti and Nudist compared. *Sociological Research Online*, vol. (3/3). www.socresonline.org.uk/socresonline/3/3/4.html.

Bernard M. (1999). *Introduction aux études littéraires assistées par ordinateur*. Presses Universitaires de France.

Bremond C. (1985). Concept et thème. *Poétique*, vol. (64) : 415-423.

Bremond C., Landy J. et Pavel T. (dir. publ.) (1995). *Thematics. New approaches*. Suny Press.

Cavnar W.B. et Trenkle J.M. (1994). N-gram-based text categorization. In *Proceedings of SDAIR-94*, Las Vegas, Nevada, U.S.A., April 1994. UNLV Publications/Reprographics : 161-175.

Fayyad U., Grinstein G.G. et Wierse A. (sous la direction de) (2001). *Information visualization in data mining and knowledge discovery*. Morgan Kaufmann Publishers.

Forest D. (2002). *Lecture et analyse de textes philosophiques assistées par ordinateur : application d'une approche classificatoire mathématique à l'analyse thématique du* Discours de la méthode *et des* Méditations métaphysiques *de Descartes*. Mémoire de maîtrise, Montréal, Université du Québec à Montréal.

Forest D. et Meunier J.-G. (2000). La classification mathématique des textes : un outil d'assistance à la lecture et à l'analyse de textes philosophiques. In *Actes des JADT 2000* : 325-329.

Giora R. (1985). Notes toward a theory of text coherence. *Poetics Today*, vol (6/4).

Grossberg S. et Carpenter G.A. (1988). The ART of adaptive pattern recognition by a self-organizing neural network. *Computer*, vol. (21/3) : 77-88.

Hockey S. (2000). *Electronic texts in the humanities*. Oxford University Press.

Jackson P. et Moulinier I. (2002). *Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization*. John Benjamins Publishing Company.

Kastberg Sjöblom M. et Brunet Ét. (2000). La thématique. Essai de repérage automatique dans l'œuvre d'un écrivain. In *Actes des JADT 2000* : 457-466.

Kintsch W et Van Dijk T.A. (1978). Toward a model of text comprehension and production. *Psychological Review*, vol. (85/5) : 363-394.

Louwerse M. et van Peer W. (sous la direction de) (2002). *Thematics: Interdisciplinary Studies*. John Benjamins Publishing Company.

Manning C.D. et Schütze H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Martin É. (1995). Thème d'étude, étude de thème. In Rastier Fr. (sous la direction de), *L'analyse thématique des données textuelles : l'exemple des sentiments*. Didier érudition : 13-24.

Meunier J.-G. (1997). La Lecture et l'Analyse de Textes Assistées par Ordinateur (LATAO) comme système de traitement d'information. *Sciences Cognitives*, vol. (22) : 211-223.

Meunier J.-G. et Torres J.M. (2000). Classphères : un réseau incrémental pour l'apprentissage non supervisé appliqué à la classification de textes. In *Actes des JADT 2000* : 365-372.

Pettersson B. (2002). Seven trends in recent thematics and a case study. In Louwerse M. et van Peer W. (sous la direction de), *Thematics: Interdisciplinary Studies*. John Benjamins Publishing Company : 237-252.

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

Prince G. (1985). Thématiser. *Poétique*, vol. (64) : 425-433.

Propp V. (1928/1968). *Morphology of the folktale*. Texas University Press.

Rastier Fr. (1996). *Sémantique interprétative*. Presses Universitaires de France.

Rastier, Fr. (2001). *Arts et sciences du texte*. Presses Universitaires de France.

Rastier Fr. *et al*. (sous la direction de) (1995). *L'analyse thématique des données textuelles : l'exemple des sentiments*. Didier Érudition.

Rastier Fr. *et al.* (1994). *Sémantique pour l'analyse. De la linguistique à l'informatique*. Masson.

Rimmon-Kenan S. (1985). Qu'est-ce qu'un thème ? *Poétique*, vol. (64) : 397-406.

Rodis-Lewis G. (1966). *Descartes et le rationalisme*. Presses Universitaires de France.

Rodis-Lewis G. (1984). *Descartes*. Librairie Générale Française.

Rossignol M. et Sébillot P. (2002). Automatic generation of sets of keywords for theme characterization and detection. In *Actes des JADT 2002* : 185-196.

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

Sebastiani F. (1999). A tutorial on automated text categorisation. In Amandi A. et Zunino A. (sous la direction de), *Proceedings of ASAI-99*, Buenos Aires : 7-35.

Sebastiani F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, vol. (34/1) : 1-47.

Sebastiani F. (2003). Text Categorization. In Zanasi A. (dir. publ.), *Text Mining and its Applications*. WIT Press.

Sollors W. (1993). *The return of thematic criticism*. Harvard University Press.

Spence R. et Press A. (2000). *Information visualization*. Addison-Wesley.

Thomashevsky B. (1925). Thematics. In Lemon L.T. et Reis M.J. (Eds) (1965), *Russian formalist criticism*. University of Nebraska Press : 61-98.

Thompson S. (1946). *The folktale*. Berkeley : University of California Press.

Van Dijk T.A. (1972). *Some aspects of text grammars. A study in theoretical linguistics and poetics.* Mouton.

Van Dijk T.A. et Kintsch W. (1983). *Strategies of discourse comprehension*. Academic Press.