`¸Université de Montréal

Extraction automatique et visualisation des thèmes abordés dans des résumés de mémoires et de thèses en anthropologie au Québec, de 1985 à 2009

par
Anne-Renée Samson

École de bibliothéconomie et des sciences de l'information
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études
en vue de l'obtention du grade de Maître
en science de l'information

Juin 2013

Université de Montréal
Faculté des études supérieures et postdoctorales

Ce mémoire intitulé :

Extraction automatique et visualisation des thèmes abordés dans des résumés de mémoires et de thèses en anthropologie au Québec, de 1985 à 2009

Présenté par :
Anne-Renée Samson

évalué par un jury composé des personnes suivantes :

Yvon Lemay, président-rapporteur
Dominic Forest, membre du jury (Directeur de recherche)
Robert Crépeau, membre du jury (Codirecteur de recherche)
Éric Leroux, membre du jury

Résumé

S'insérant dans les domaines de la Lecture et de l'Analyse de Textes Assistées par Ordinateur (LATAO), de la Gestion Électronique des Documents (GÉD), de la visualisation de l'information et, en partie, de l'anthropologie, cette recherche exploratoire propose l'expérimentation d'une méthodologie descriptive en fouille de textes afin de cartographier thématiquement un corpus de textes anthropologiques. Plus précisément, nous souhaitons éprouver la méthode de classification hiérarchique ascendante (CHA) pour extraire et analyser les thèmes issus de résumés de mémoires et de thèses octroyés de 1985 à 2009 (1240 résumés), par les départements d'anthropologie de l'Université de Montréal et de l'Université Laval, ainsi que le département d'histoire de l'Université Laval (pour les résumés archéologiques et ethnologiques). En première partie de mémoire, nous présentons notre cadre théorique, c'est-à-dire que nous expliquons ce qu'est la fouille de textes, ses origines, ses applications, les étapes méthodologiques puis, nous complétons avec une revue des principales publications. La deuxième partie est consacrée au cadre méthodologique et ainsi, nous abordons les différentes étapes par lesquelles ce projet fut conduit; la collecte des données, le filtrage linguistique, la classification automatique, pour en nommer que quelques-unes. Finalement, en dernière partie, nous présentons les résultats de notre recherche, en nous attardant plus particulièrement sur deux expérimentations. Nous abordons également la navigation thématique et les approches conceptuelles en thématisation, par exemple, en anthropologie, la dichotomie culture/biologie. Nous terminons avec les limites de ce projet et les pistes d'intérêts pour de futures recherches.

Mots-clés : Fouille de textes, forage de textes, analyse thématique assistée par ordinateur, classification automatique, visualisation graphique, analyse réseaux, anthropologie.

Abstract

Taking advantage of the recent development of automated analysis of textual data, digital records of documents, data graphics and anthropology, this study was set forth using data mining techniques to create a thematic map of anthropological documents. In this exploratory research, we propose to evaluate the usefulness of thematic analysis by using automated classification of textual data, as well as information visualizations (based on network analysis). More precisely, we want to examine the method of hierarchical clustering (HCA, agglomerative) for thematic analysis and information extraction. We built our study from a database consisting of 1 240 thesis abstracts, granted from 1985 to 2009, by anthropological departments at the University of Montreal and University Laval, as well as historical department at University Laval (for archaeological and ethnological abstracts). In the first section, we present our theoretical framework; we expose definitions of text mining, its origins, the practical applications and the methodology, and in the end, we present a literature review. The second part is devoted to the methodological framework and we discuss the various stages through which the project was conducted; construction of database, linguistic and statistical filtering, automated classification, etc. Finally, in the last section, we display results of two specific experiments and we present our interpretations. We also discuss about thematic navigation and conceptual approaches. We conclude with the limitations we faced through this project and paths of interest for future research.

Keywords : Text mining, automated thematic analysis of textual data, hierarchical clustering, concept extraction, information visualization, anthropology.

Table des matières

Liste des tableaux

Liste des figures

À ma famille,
et particulièrement à mon père,
cet homme dévoué, persévérant,
intéressé et aidant.

Remerciements

> *Sans opposer un relativisme frileux à un universalisme dogmatique, cela conduit à affirmer l'autonomie et la spécificité de la sphère culturelle, et à poursuivre, dans la direction tracée notamment par Cassirer, l'entreprise d'une philosophie des formes symboliques. Elle dessine les contours d'une sémiotique des cultures, et laisse discerner un projet refondateur pour les sciences sociales, encore victimes de diverses idéologies.* (Rastier : 2004)

D'entrée de jeu, je n'aurais pu mener à bien ce projet sans l'appui, l'intérêt et la collaboration de plusieurs professeurs, de collègues d'étude, d'amis et de proches. Je souhaite remercier chaleureusement tous ceux et celles qui m'ont épaulé, de près ou de loin, à la réalisation de ce projet de maîtrise.

J'aimerais remercier plus particulièrement mes directeurs de recherche; d'abord, Dominic Forest, professeur agrégé de l'École de bibliothéconomie et des sciences de l'information, chercheur en fouille de textes et dans les domaines de l'analyse et de la gestion de l'information textuelle et de la visualisation de l'information. Je le remercie pour l'intérêt qu'il m'a témoigné tout au long de ce projet de recherche et pour sa patience. J'ajouterais avoir grandement apprécié sa passion pour les techniques de fouille de textes et pour l'analyse thématique, ainsi que l'environnement intellectuel riche qu'il m'a offert.

En deuxième lieu, je remercie mon codirecteur, Robert Crépeau, professeur titulaire au Département d'anthropologie de l'Université de Montréal. Je le remercie spécialement pour son intérêt et sa vision de l'anthropologie, notamment en ce qui a trait à la compréhension, à l'interprétation et à la diffusion de cette discipline telle que véhiculée au Québec. Je le remercie également pour son écoute, sa façon de relativiser, pour son humour, pour ses bons conseils et pour son infaillible soutien lorsque la motivation me manquait.

J’offre également mes remerciements à Marie-Andrée Couillard, professeure titulaire et directrice du Département d’anthropologie (Ulaval), et à Michel Fortin, professeur en archéologie et directeur du Département d’histoire (Ulaval). Puis, aussi, à François Beaudet, professionnel et responsable du laboratoire d'ethnologie et des ressources en anthropologie visuelle, du Département d’anthropologie de l’Université de Montréal. Je les remercie pour avoir chaleureusement accepté de me supporter lors de la collecte des données, me donnant également accès aux collections départementales de mémoires et de thèses, tout en m’offrant des espaces de travail.

Finalement, j’aimerais aussi remercie Marcela Baiocchi, candidate au doctorat à l’École de bibliothéconomie et des sciences de l’information, de l’Université de Montréal. Sans sa précieuse collaboration, je n’aurais pu réaliser une aussi bonne visualisation graphique de mes résultats. Elle m’a en effet initiée au logiciel Gephi et m’a grandement aidée, pour ne pas dire secourue, face aux difficultés techniques et informatiques du processus de visualisation.

Introduction

Le dernier quart de siècle a connu des développements prolifiques sur les plans technologiques, favorisant la numérisation et l'exploitabilité de l'information. Pour plusieurs, les innovations du numérique et de l'informatique ont été gages de préservation et de pérennité de notre patrimoine informationnel. Aussi, la qualité des documents s'est considérablement accrue au fil des ans et l'évolution du numérique et des technologies du web ont favorisé la création et la mise en place d'infrastructures documentaires. Mais l'avènement du numérique a aussi mené à une problématique de surcharge documentaire, voire d'*infobésité* (Forest : 2009, 77). En quelque sorte, nous nous retrouvons aujourd'hui avec des quantités de documents informatiques et de bases de données considérables, pour ne pas dire démesurées, et il devient souvent ardu de s'y retrouver et d'effectuer de la recherche d'information qui soit à la fois pertinente et efficace.

Néanmoins, durant les dernières années, plusieurs stratégies ont été mises de l'avant pour contrer l'hypertrophie documentaire et nous remarquons, entre autres, la volonté de normalisation de l'encodage et de l'indexation des documents, la création d'outils de gestion électronique des documents (GED), le développement de protocoles d'échanges et de diffusion des données (entrepôts de données, moissonneurs de données, etc.). Nonobstant ces innovations, peu permettent l'exploitation des contenus sémantiques, restant plutôt sur une perspective de surface et aux métadonnées. Ainsi, voulant précisément porter notre regard sur les aspects sémantiques et l'analyse thématique de corpus de documents, nous proposons d'utiliser des techniques de fouille de textes. Nous croyons en effet que les méthodes de classification hiérarchique constituent une avenue intéressante pour la création d'outils d'analyse prenant en compte le contenu informationnel des documents numériques.

Le domaine de la fouille de textes représente actuellement une source importante de recherche et ses fondements théoriques et pratiques proviennent de plusieurs disciplines et est, en ce sens, multidisciplinaire. La fouille de données (*Data Mining*) et

l'extraction de connaissances dans des bases de données (*ECBD, Knowledge discovery in databases*) sont certainement les domaines les plus influents quant au développement de la fouille de textes. Ces champs de recherche prennent pour objectifs l'exploration et l'analyse automatique ou semi-automatique de bases de données informatiques pour détecter des structures et des tendances (Tufféry : 2010, 4). Nous devons également remarquer les apports d'autres disciplines, à commencer par le repérage de l'information (*information retrieval*), la linguistique computationnelle et le traitement automatique des langues (TAL), l'intelligence artificielle et l'apprentissage machine (*machine learning*), les sciences de l'information et les sciences cognitives (Forest : 2006, 2). Pour illustrer quelques exemples de développements marquants de ces disciplines pour la fouille de textes, mentionnons les techniques de création de résumés automatiques, les méthodes de filtrage de l'information textuelle et la création de taxinomies et d'ontologies à partir de procédés informatiques. Finalement, la fouille de textes découle aussi, en partie, de la bibliométrie et de la scientométrie, où l'on applique les statistiques et les mathématiques aux documents scientifiques pour quantifier et faire ressortir des informations pertinentes quant aux auteurs, aux citations, aux découvertes et aux théories scientifiques (Archambeault : 2002, 2).

Les méthodes de fouille de textes qui, somme toute, sont relativement récentes (une quinzaine d'années) sont destinées au traitement automatique de données textuelles, normalement sur de grands corpus (pouvant traiter des millions de documents), afin d'en dégager des structures et des relations thématiques. Ces techniques permettent, entre autres, la découverte d'informations inconnues, quasiment imperceptibles sans support informatique ou, encore, la reconnaissance stylistique littéraire (par exemple, la reconnaissance des auteurs, de leur style). Donc, en d'autres termes, la fouille de textes tente d'identifier automatiquement des contenus thématiques pour permettre l'analyse sémantique de textes et ainsi, elle représente en quelque sorte l'union de la lexicométrie et de la fouille de données (Tufféry : 2010, 629).

Dans le cadre de notre recherche, nous prenons comme objectif d'expérimenter et d'évaluer la pertinence des méthodes de fouille de textes pour analyser les thèmes provenant des résumés de mémoires et de thèses anthropologiques (1 240 résumés), des départements d'anthropologie et d'histoire de l'Université Laval, et du département d'anthropologie de l'Université de Montréal, de 1985 à 2009. Plus précisément, notre but est de valider l'hypothèse selon laquelle la méthode de classification hiérarchique ascendante (CHA) peut permettre le repérage et l'extraction de mots significatifs issus du lexique, menant ultimement à une visualisation thématique du domaine de l'anthropologie québécoise.

L'organisation de ce mémoire se présente ainsi : en première partie, nous abordons la problématique de recherche et le cadre théorique que nous avons adopté. De la sorte, nous exposons comment et pourquoi nous avons choisi d'utiliser des méthodes de fouille de textes dites descriptives pour l'analyse thématique de notre corpus. Il sera également question des concepts importants en fouille de textes, dont la notion de thème. Ensuite, nous traiterons de l'origine, de l'évolution et des définitions des techniques de fouilles de textes et nous ferons la distinction entre deux grandes familles de méthodes; soit les techniques supervisées et les techniques non supervisés. Puis, nous parlerons des étapes et des processus propres à la fouille de texte, pour compléter finalement cette section avec une brève revue de la littérature.

En deuxième partie de ce mémoire, nous présentons le cadre méthodologique de notre projet. Nous expliquons en détail les différentes étapes du processus de recherche : la collecte des données, le prétraitement des documents, le filtrage du lexique, la vectorisation, la classification automatique, l'extraction automatique des termes thématiques, puis la visualisation des résultats. Nous traitons en dernier lieu de ce qu'est à nos yeux l'anthropologie québécoise et nous faisons une brève mise en contexte de nos données.

Dans la troisième partie, nous abordons d'abord l'approche de navigation thématique et les conséquences qui en découlent. Nous exposons ensuite les résultats de

recherche, en nous concentrant plus précisément sur deux expérimentations. Nous discutons ensuite de l'interprétation globale des résultats et des approches interprétatives en thématisation. Nous concluons avec les limites rencontrées lors de notre recherche et nous proposons quelques pistes possibles pour de futures recherches.

Finalement, il est à remarquer que les résultats de recherche des différentes expérimentations de classification réalisées lors de ce projet sont présentés en pièces jointes de ce mémoire.

Première partie

1. Problématique et cadre théorique

1.1. Définitions de la fouille de textes

La fouille de textes, nommée également *forage de texte* ou *text mining*, est de nature interdisciplinaire et les définitions peuvent ainsi varier selon plusieurs perspectives. Cette discipline est également pluraliste, car elle s'applique à la fois aux milieux universitaires et aux milieux industriels et commerciaux. Indépendamment des points de vue, l'objectif poursuivi en fouille de texte est l'expérimentation et le développement de méthodes, d'algorithmes, de protocoles et d'applications informatiques dans le but d'extraire automatiquement des informations et des structures inhérentes à de grands corpus de textes (Forest : 2009, 79).

En prenant le point de vue informatique, les chercheurs placent beaucoup d'importance à la création d'algorithmes de fouille, comme les algorithmes *k-means* et *k-plus proches voisins* qui, sommairement, servent à partitionner des ensembles de textes grâce à des mesures de distance entre les objets. Nous pouvons également remarquer le développement d'algorithmes pour les méthodes descriptives, tels que les approches de classifications hiérarchiques descendantes (*divisive algorithm*) et ascendantes (*agglomerative algorithm*), tout comme les méthodes de partitionnement (Ibekwe-SanJuan : 2007, 60). Et selon la perspective du traitement automatique des langues (TAL), les objectifs touchent davantage aux questions linguistiques et terminologiques, dont, par exemple, le développement de filtres linguistiques ou la création automatisée d'ontologies. Du côté du repérage de l'information (*information retrieval*), il y a une importance accordée au « *topic spotting »* (détermination d'un seul thème pour décrire un texte) et à l'indexation automatique à partir des techniques de fouille de textes. La définition qui englobe le mieux les différentes perspectives, selon nous, est celle d'Hearst :

> « Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information

> together to form new facts or new hypotheses to be explored further by more conventional means of experimentation » (Hearst: 2003).

Dans un même ordre d'idées, nous aimerions remarquer la définition de Kodratoff, chercheur en extraction de connaissances à partir de bases de données (ECBD) : « *la science qui découvre les connaissances dans les textes […] les connaissances découvertes doivent être ancrées dans le monde réel et doivent modifier le comportement d'un agent humain ou mécanique »* (Ibekwe-SanJuan : 2007, 33). Cet auteur utilise le mot « connaissance » au lieu d'information, car l'ECBD et la fouille de textes repèrent des motifs et des structures informationnelles et lorsque des motifs dépassent un certain seuil d'intérêt (*interestingness*), ils deviennent des connaissances. Mais puisque cette mesure dénote de la subjectivité, la détermination des connaissances par rapport aux informations moins pertinentes relève assurément de l'utilisateur final (Ibid.).

1.2. Origines et évolution

Il n'est pas nécessairement aisé d'identifier les premières traces et utilisations des méthodes de fouille de textes. En fait, cette discipline est récente et résulte, en grande partie, de la problématique de surcharge des documents textuels en format numérique. Il est maintenant admis que 80 % des informations disponibles dans les mémoires électroniques ou sur le web sont textuelles, d'où la nécessité de développer des stratégies de gestion documentaire (Ibekwe-SanJuan : 2007, 31). Selon Kadratoff, l'exploitation des textes en tant que « réservoirs de connaissances », à l'aide de méthodes automatisées, remonte à une quinzaine d'années. Auparavant, l'analyse thématique et qualitative de textes faisait inévitablement intervenir l'esprit humain et il était donc ardu, voire subjectif, de travailler sur de grands corpus. Graduellement, les chercheurs se sont intéressés aux textes comme sources de connaissances modélisables et dans les années 1995-2000, plusieurs techniques ont vu le jour dont, entre autres, la

création automatique d'ontologies (Kodratoff : 1999). Finalement, les premiers travaux en *Knowledge Discovery in Textual data* et en fouille de textes sont attribués à Feldman et Dagan, participants de la première conférence annuelle en recherche de connaissances et en fouille de données, qui eut lieu à Montréal en 1995 (Ibekwe-SanJuan : 2007, 31).

Étant donné le caractère multidisciplinaire de la fouille de textes, nous aborderons son évolution selon ces disciplines pionnières. D'abord, la fouille de données, qui est une spécialisation de l'extraction de connaissances à partir de bases de données (ECBD, Knowledge discovery in databases), est sans aucun doute la discipline qui a le plus influencé le développement de la fouille de textes. Aux débuts et compte tenu des difficultés en analyses sémantiques et syntaxiques profondes, les dimensions linguistiques des documents n'étaient pas considérées et l'ECBD s'appuyait uniquement sur des techniques statistiques pour identifier des distributions textuelles. « *En somme, il s'agissait d'apposer des étiquettes aux textes afin d'en dériver des données structurées* » (Ibekwe-SanJuan : 2007, 31).

Plusieurs disciplines en informatique et en statistiques ont également eu des répercussions importantes sur le développement de la fouille de textes. Sans aborder en profondeur l'évolution des apports disciplinaires parce que nous ne pourrions être exhaustifs, nous aimerions néanmoins souligner les contributions suivantes : l'automatisation des traitements des données textuelles, les méthodes et concepts statistiques comme les règles d'associations, la pondération, les modèles probabilistes, les mesures de distance (par exemple, la similarité cosinus), les modèles vectoriels et le développement d'algorithmes.

Par ailleurs, la linguistique a aussi eu des impacts sur le développement des techniques de fouilles de textes, principalement au regard des questions du traitement automatique des langues. Nous pouvons ainsi penser aux apports importants des techniques de filtrage statistique (ordonnancements et tris) et de filtrage des relations de variations linguistiques dont, notamment, les différentes flexions des verbes et les relations de genre en français (petit / petite ou développer / développa). Nous pouvons

aussi remarquer les techniques de lemmatisation et de racinisation, tout comme l'utilisation des dictionnaires de mots fonctionnels et vides de sens, servant à épurer les lexiques. Ces techniques sont aujourd'hui des étapes méthodologiques incontournables en fouille de textes (Ibekwe-SanJuan : 2007, 128).

Finalement, l'intelligence artificielle a aussi contribué à l'amélioration des techniques de fouille de textes, principalement quant aux notions et méthodes d'apprentissage. Ainsi, les techniques prédictives en fouille de textes (catégorisation automatique) emploient des algorithmes d'apprentissage pour prédire l'appartenance d'un document à une ou des classes préexistantes. La notion d'*apprentissage machine* nécessite un corpus préalablement créé et catégorisé pour entraîner les systèmes à effectuer des prédictions de catégorisation et de classification. Les méthodes basées sur les réseaux neuronaux et les machines à vecteurs supports sont de bons exemples de techniques découlant de l'intelligence artificielle (Ibekwe-SanJuan : 2007, 38).

1.3. Description des principales méthodes de fouille de textes

Essentiellement, les méthodes de fouille de textes peuvent se présenter selon deux grandes approches : les méthodes descriptives et les méthodes prédictives. Dans le premier cas, c'est-à-dire l'approche descriptive, le but général est de fournir une vue synthétique des données pour faire ressortir les agrégations et les relations de dépendance entre les objets, grâce à des règles d'association et de dépendance fonctionnelle (Ibekwe-SanJuan : 2007, 59). Autrement dit, l'objectif consiste à découvrir des structures inconnues d'ensembles de données, en utilisant des algorithmes qui mettent en évidence les tendances d'affinités ou d'éloignements entre les documents. De plus, ces tendances sont calculées selon les mots significatifs et discriminants, utilisés dans chacun des textes, ainsi que par rapport à l'ensemble du corpus. Ces méthodes sont dites non supervisées, car aucun modèle de regroupement n'est utilisé (dictionnaire de catégories, préalablement créer par les chercheurs). D'une certaine façon, nous pourrions imaginer un ensemble de points représentant chacun des

documents textuels puis, les algorithmes serviraient à regrouper automatique en agrégats les documents qui possèdent des indices de similitudes supérieurs à un certain seuil.

Toujours selon l'approche descriptive, quoique ce soit également valable pour les méthodes prédictives, il est aussi possible d'émettre une deuxième distinction en fonction de deux axes, l'un paramétrique et l'autre non paramétrique. L'approche non paramétrique répond à l'hypothèse voulant que plus deux objets sont proches, plus ils ont de chances d'être regroupés dans une même classe. Il n'y a alors aucune hypothèse sur une loi de distribution que révéleraient les données. Les méthodes de classification hiérarchique et de partitionnement sont basées sur l'approche non paramétrique. (Ibekwe-SanJuan : 2007, 60).

À l'inverse, l'approche paramétrique fait intervenir les lois de distribution entre les objets :

> « Les approches paramétriques, basées sur des probabilités, font l'hypothèse que les objets à classer suivent une loi de distribution. La difficulté est alors de trouver la forme (loi normale, loi hypergéométrique, etc.) et les paramètres (moyenne, variance) de cette loi et de déterminer ensuite à quelle classe appartient un objet » (Ibekwe-SanJuan : 2007, 59).

Finalement, une dernière distinction des méthodes descriptives doit être faite quant au mode de fonctionnement méthodologique ; nous retrouvons d'un côté la classification hiérarchique et de l'autre côté, la classification non hiérarchique ou de partitionnement. En bref, la classification hiérarchique permet de créer une hiérarchie de classes emboîtées et peut être ascendante ou descendante, alors que les méthodes de partitionnement utilisent les algorithmes de la famille des *k-means* pour regrouper les données en des partitions non-emboîtantes, grâce à des calculs de distance entre les objets (Ibekwe-SanJuan : 2007, 66).

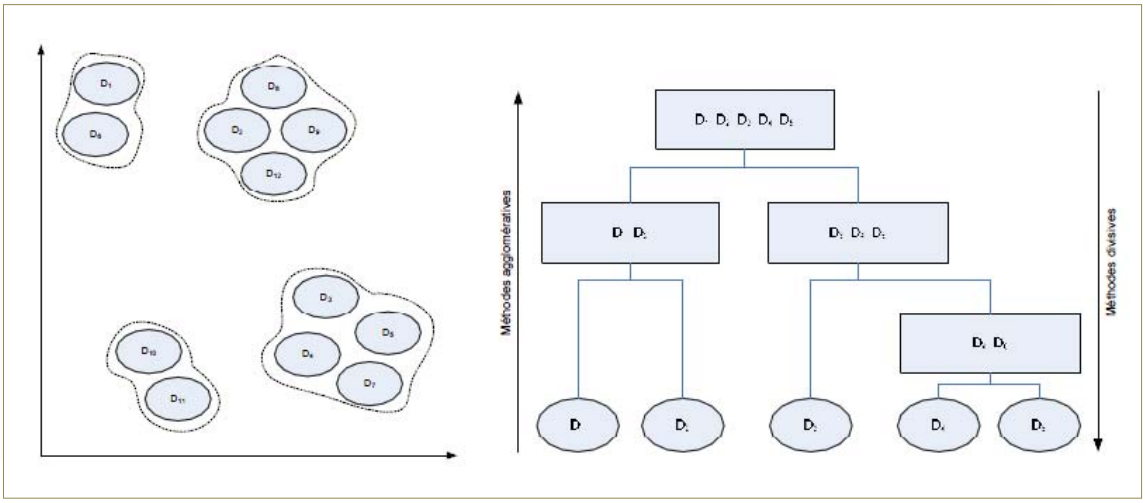


Figure 1. Représentation de la classification par partitionnement et de la classification hiérarchique. Forest : 2006, 56.

Pour les méthodes de fouille de textes prédictives, l'objectif général est de prédire la catégorie des objets à classer, à la suite de l'étape d'apprentissage machine. Ainsi, il y a l'utilisation d'un ensemble de données ayant reçu des valeurs manuellement (catégorisé par un chercheur) permettant d'entraîner les algorithmes à prédire les valeurs de nouvelles données. Tous comme l'approche descriptive, les méthodes prédictives peuvent être paramétriques ou non paramétriques. Du point de vue paramétrique, les données suivent un modèle probabiliste en lien avec une hypothèse à vérifier et l'objectif est alors d'y découvrir une fonction ou un modèle inconnu, en constituant le paramétrage à partir des données d'apprentissage. « *Dans l'approche non paramétrique, aucune hypothèse n'est formulée a priori sur un modèle que suivraient les données. On cherche simplement à rapprocher des objets en fonction d'un critère de similarité ou de distance.* » (Ibekwe-SanJuan : 2007, 89). Finalement, remarquons que les *machines à vecteurs supports*, les *classifieurs bayésiens* et les *réseaux neuronaux* sont des méthodes souvent employées selon l'approche prédictive (Ibekwe-SanJuan : 2007, 92).

Dans le cadre de notre projet, nous avons privilégié l'approche descriptive, car nous n'avions pas de corpus d'apprentissage. En fait, nous sommes les premiers

chercheurs à effectuer de la fouille de textes sur des résumés de mémoires et de thèses anthropologiques du Québec. Ainsi, nous souhaitons explorer nos données de la façon la plus objective possible, sans faire intervenir aucun ensemble de textes préalablement catégorisés.

1.4. Étapes méthodologiques typiques en fouille de textes

D'entrée de jeu, il est important de considérer les étapes méthodologiques de la fouille de textes, peu importe les approches utilisées, comme étant un processus itératif, laissant ainsi une marge de manœuvre pour des ajustements et des corrections.

La première étape consiste en la collecte des données. Aussi est-il souhaitable, lors de cette phase, de constituer notre corpus selon nos objectifs de départ et de prendre en considération certaines caractéristiques : des caractéristiques générales ou contextuelles (provenance, date de création, taille du texte, etc.), technologiques (support et format), informationnelles (thématiques) et linguistiques (langue, genre) (Forest : 2002, 79). Finalement, la constitution du corpus est une étape importante, car la qualité des expérimentations et des résultats en sont directement liées.

La deuxième opération à effectuer est la normalisation et le filtrage du lexique. Pour ce faire, il y a construction d'un antidictionnaire ou, autrement dit, d'une liste de mots fonctionnels et vides de sens qui seront enlevés du processus d'analyse. Les mots fonctionnels et vides sont des mots très communs, tels que les prépositions, certains verbes comme « avoir », et certains adverbes et adjectifs moins pertinents, comme « ainsi », « troisièmement » ou « donc ». Ensuite, il peut-être souhaitable de normaliser le lexique en travaillant et en modifiant les variations de certaines expressions, dont les mentions d'époques (19^{e} siècle versus dix-neuvième ou XIX^{e} siècle) ou, encore, des expressions comme « socio-économique » pour « socioéconomique ». Finalement, la dernière sous-opération du filtrage du lexique est la lemmatisation ou la racinisation (*stemming*). La lemmatisation est la réduction des mots en *lemmes* ; un lemme est un

assemblage de morphèmes, possédant un signifiant et un signifié. Par exemple, les flexions des verbes en français peuvent être éliminées par la lemmatisation, et alors les verbes « adoptèrent » et « adoptera » deviennent « adopter ». La racinisation est plus drastique et consiste en un processus d'amputation des terminaisons pour obtenir les racines des mots. Donc, lors du processus de racinisation, le mot « maisonnée » devient « maison ».

La troisième étape méthodologique est la vectorisation du corpus initial, c'est-à-dire la conversion du corpus en matrices de vecteurs, faisant en sorte que les algorithmes puissent être appliqués. Autrement dit, nous donnons des valeurs statistiques aux mots du lexique grâce à des mesures de distance ou de similarité entre les mots qui composent les documents.

> « Cette opération est réalisée en structurant les documents du corpus en une matrice de vecteurs dans laquelle chaque document (ou segment de document) est représenté par l'absence ou la présence, binaire ou pondérée, de chaque unité lexicale retenue à l'étape précédente » (Forest : 2009, 80).

Donc, la vectorisation amène à la création d'une matrice vectorielle (table statistique) pour représenter la distribution des termes discriminants à travers le corpus. En associant aux mots des vecteurs, il est ainsi possible de percevoir leur valeur statistique et selon l'hypothèse de Salton (système SMART [*System for the Mechanical Analysis and Retrieval of Text*]), les documents et les mots peuvent être positionnés dans un espace vectoriel.

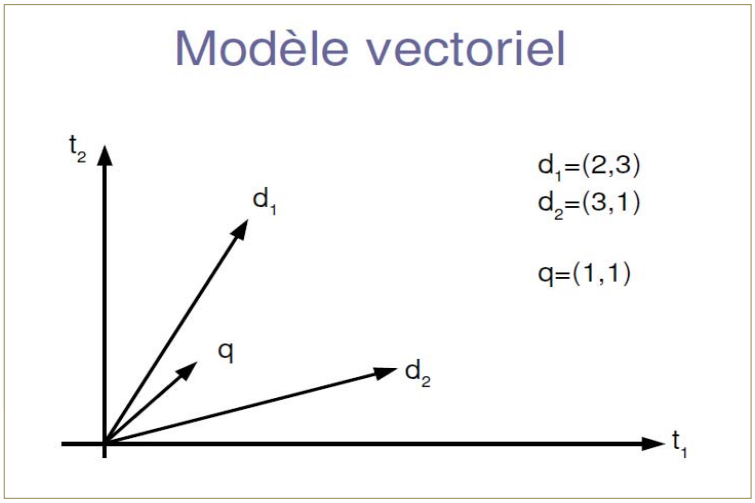


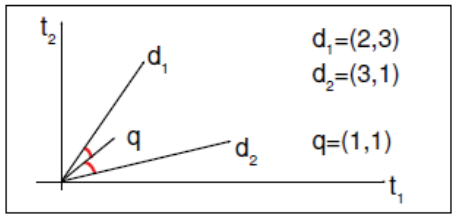
Figure 2. Modèle vectoriel, où « d » représente les documents, « t » les termes, et « q » les requêtes. (Rosenknop : 2001, 14).

À partir du modèle, la pertinence des termes est évaluée par la mesure de similarité dans l'espace vectoriel, à l'aide des valeurs pondérées (ou binaires) des vecteurs. De plus, la valeur discriminante d'un terme dans un document se fonde sur trois règles :

- L'importance du terme dans le document (pondération locale)
- L'importance du terme dans l'ensemble du corpus (pondération globale)
- L'importance du document (taille du document)

Finalement, il existe plusieurs méthodes pour calculer les pondérations, dont les plus fréquentes sont le facteur *tf* (term frequency), le facteur *idf* (inverse document frequency), le facteur fréquentiel (nombre d'occurrences) et le facteur logarithmique. Remarquons également que plus les documents sont similaires quant à la distribution de leurs mots, et plus les coefficients de similarités sont élevés entre ces documents. Dans le cadre de notre recherche, les pondérations furent calculées selon le cosinus (Rosenknop : 2001, 26).

Cosinus :



Formule du calcul de similarité :

$$s(d,q)=\frac{\sum_{i=1}^{n} d_i \times q_i}{\sqrt{\sum_{i=1}^{n} d_i^2}\sqrt{\sum_{i=1}^{n} q_i^2}}$$

Figure 3. Représentation du cosinus, calculé par l'angle entre les vecteurs des documents et la requête, et formule de similarité (Rosenknop : 2001, 26).

À la quatrième étape du processus de fouille de textes, il est question de la classification afin de structurer et de regrouper le corpus, permettant ensuite l'extraction des termes discriminants, c'est-à-dire les termes qui distinguent et caractérisent bien un ou plusieurs documents. Ainsi, par la classification automatique, nous faisons ressortir des structures sémantiques et thématiques, visibles à travers les regroupements et les classes créés par les algorithmes. Autrement dit, nous recherchons un schème, un modèle ou un patron thématique, qui illustre le regroupement et le positionnement des documents en fonction de la similarité de distribution fréquentielle de leurs mots significatifs. Les algorithmes permettent donc à cette étape de constituer des classes de documents et de ces groupes, nous extrayons les mots présentant des indices de similarité élevée.

Finalement, la dernière étape du processus de fouille de textes est l'interprétation et l'évaluation des résultats. Cette étape est complexe et il est indispensable de considérer la nature des documents traités, le contexte de réalisation et les objectifs de recherche (Forest : 2002, 81). Brièvement, il peut y avoir plusieurs perspectives quant à la façon d'interpréter les résultats, mais généralement, l'idée est d'étudier et de chercher à comprendre comment les classes ont été formées, qu'elles sont les mots significatifs, comment les documents et les mots sont liés entre eux et quels modèles sémantiques se dégagent des résultats (cartographie thématique). Remarquons toutefois que

l'interprétation demeure jusqu'à un certain point subjective et qu'au final, ce sont les utilisateurs qui, en naviguant et en étudiant les résultats d'analyse, jugent de la pertinence des termes extraits et des interprétations qu'il est possible de faire.

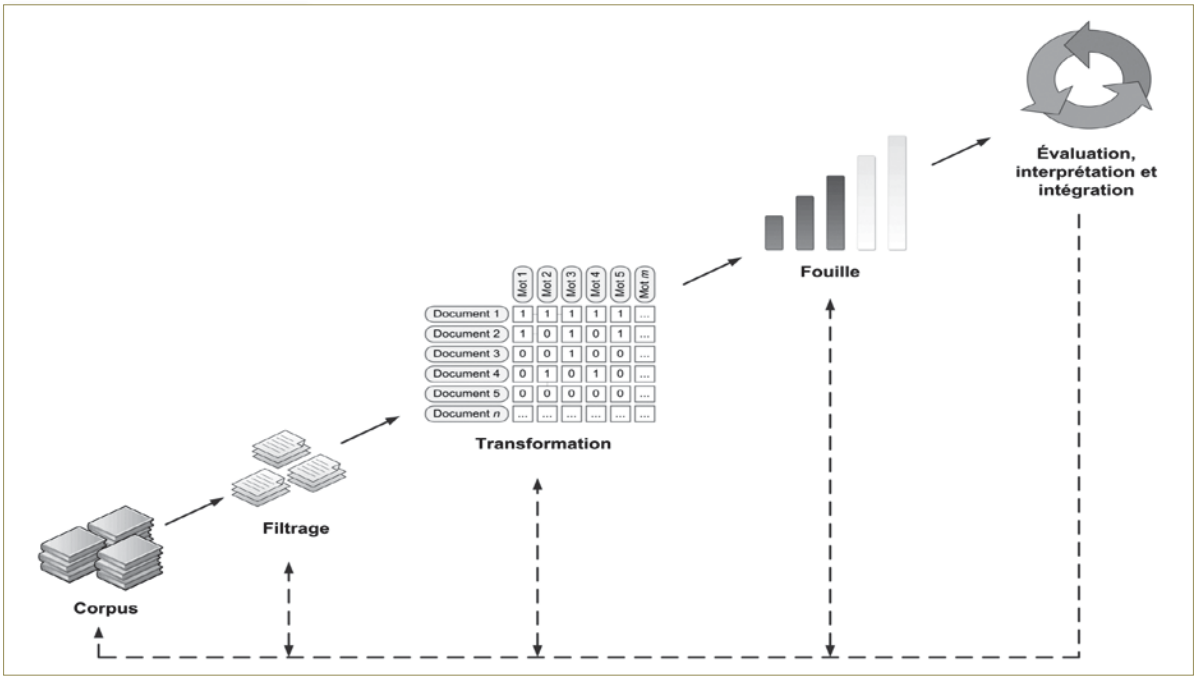


Figure 4. Méthodologie générique de la fouille de textes. Figure inspirée de Fayard et al., adaptée par Forest, 2009 : 81.

1.5. Notion de thème

Puisque nos techniques visent l'extraction des termes significatifs de notre corpus, nous croyons qu'il est important de clarifier ce que nous entendons par « thème » et les conséquences qui en découlent.

En 1915, Ferdinand de Saussure publia l'ouvrage « *Cours de linguistique générale* », dans lequel il présentait sa théorie linguistique. Selon son approche, les mots sont des objets complexes, possédant un signifiant (forme phonologique) et un signifié (sens) et à travers le système langagier, les mots ont des propriétés paradigmatiques (l'emploi des mots dans des phrases est tributaire des situations dans lesquelles sont

construites ces phrases et l'utilisation de tel ou tel mot varie selon des règles morphosyntaxiques) (Tuite : 1999, 9). La notion de « morphème », quant à elle, renvoie à l'unité minimale d'analyse grammaticale et ainsi, le mot *inacceptable* se décompose selon les trois morphes suivants : *in*, *accept* et *able*. Puis, la notion de *lexème* se définit comme étant une unité plus abstraite, prenant différentes formes flexionnelles selon les règles syntaxiques employées dans les phrases (Lyons, Linguistique générale : 1970, 230). En quelque sorte, cette interprétation du lexème est très proche de la distinction entre « thème » (l'objet du discours) et « rhème » (l'information relative au thème), proposée par l'École linguistique de Prague (Forest et Meunier : 2004, 435). Finalement, en décomposant au maximum, le concept *lemme*, pour les langues indo-européennes, représente l'assemblage de phonèmes en morphèmes. Les lemmes sont donc des unités minimales de sens et peuvent constituer des entrées dans un dictionnaire. En bref, les mots *maison*, *femme*, *enfant*, *ossement* sont tous des lemmes.

Par ailleurs, dans les années 1970-1980, Van Dijk et Kintsch ont proposé une approche d'analyse linguistique permettant de distinguer les niveaux micro-structurels (l'analyse de la phrase) et macro-structurels (l'analyse se situe dans l'ensemble du texte, considéré comme un tout cohérent et structuré), accordant ainsi de l'importance à la contextualisation sémantique.

> « Selon cette perspective, l'analyse thématique des textes relève d'un effort d'abstraction se situant au niveau macrostructurel 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 » (Forest : 2006, 11).

Dans le cadre de notre recherche, nous concevons qu'un thème est un lemme ou, dans certains cas, la combinaison de deux ou trois lemmes (par exemple, les expressions suivantes : marché du travail, Colombie-Britannique, États-Unis, Première Guerre mondiale, etc.). Les mots extraits du corpus sont donc des lemmes discriminants et représentatifs de notre corpus ; plus précisément, ce sont les mots ayant les plus grandes fréquences pour chacune des classes générées par le processus classificatoire.

1.6. Domaines d'application

Les domaines d'application faisant intervenir des méthodes de fouille de textes sont aujourd'hui en croissance et touchent aussi bien les milieux universitaires, gouvernementaux, industriels que commerciaux. Évidemment, cela s'explique en grande partie par l'expansion fulgurante du numérique, dont découlent des problématiques de gestion des documents numériques et de recherche d'information. Ne pouvant être exhaustifs quant à la description des différentes utilisations de la fouille de textes, nous citerons uniquement quelques exemples fréquemment invoqués.

Le milieu économique (commerce, marketing, domaine bancaire) est un domaine prolifique pour le développement d'applications de la fouille de textes. Alors que le "*data mining"* permet, entre autres, de détecter les comportements irréguliers (frauduleux) ou de calculer les « notes » de leurs clients en fonction des risques financiers, la fouille de textes est employée principalement en relations clients. En étudiant des sondages, des plaintes ou, encore, des dossiers clients, les entreprises bancaires peuvent effectuer de la classification et de la catégorisation automatique de textes, programmer le routage automatique des courriels, améliorer les services aux clients, etc. Par exemple, les banques peuvent construire leurs campagnes en marketing, par l'identification des clients plus réceptifs à un produit ou à une promotion (Delorme : 2002, 3).

Les domaines se rapportant à la médecine, à la santé et à la biologie ont également introduit rapidement les méthodes de fouille de textes dans leurs pratiques. Généralement, les termes et les définitions sont assez normalisés et se prêtent bien aux analyses textuelles. Divers objectifs sont visés par la fouille de textes, dont, entre autres, la découverte de modèles ou de patrons thématiques. En guise d'exemple, nous citons le projet NeuroSynth, qui présente aux utilisateurs une plateforme spécialisée en neuroscience, permettant l'extraction de termes à partir d'articles, couplée à un outil de visualisation. Ainsi, à l'aide de cette plateforme, il est possible de projeter les résultats de recherche (recherche par mots-clés) sur une carte visuelle du cerveau, situant les

thématiques abordées et la documentation disponible. « *NeuroSynth is a platform for large-scale, automated synthesis of functional magnetic resonance imaging (fMRI) data extracted from published articles.* » (Yarkoni et al. : 2011, Neurosynth Beta). À ce jour, ce système permet de faire de la fouille de textes et de l'extraction de termes à partir de 4 393 études, tout en favorisant le développement d'imageries neuronales, en fonction des termes ayant les plus fortes fréquences.

Finalement, remarquons que les méthodes de fouille de textes sont de plus en plus utilisées dans le domaine de la recherche d'information, particulièrement sur Internet. Ainsi, la classification automatique peut servir à extraire des termes issus des pages web trouvées par les moteurs de recherche ou les métamoteurs, mettant en évidence les thématiques traitées. À la suite d'une classification, il est possible d'affiner les requêtes de recherche, de les réorienter ou d'identifier les ambiguïtés sémantiques. En guise d'exemple, nous citons les moteurs de recherche Grokker (interrogation des index de Yahoo!, Wikipedia et Amazon), Clusty (partie visible du Web en entier) (Forest : 2009, 84) et Cluuz (partie visible du Web et bases de données corporatives) (Sprylogics International : 2008). Ces systèmes, à leur façon, permettent de structurer l'information, tout en améliorant les processus de recherche et de repérage de l'information.

1.7. Revue de la littérature

Comme nous l'avons mentionné précédemment, le domaine de la fouille de textes puise ses origines de plusieurs disciplines et, par conséquent, il y a une quantité importante de documentation abordant ce domaine, sous diverses perspectives. Ainsi, que ce soit à travers la recherche et le repérage de l'information, le traitement automatique des langues (TAL), l'ingénierie linguistique, la traduction et la création automatique de résumés, la lecture et l'analyse de textes assistées par ordinateur (LATAO) ; la documentation s'enrichit toujours.

De ce fait, nous ne pouvons être exhaustifs quant à la présentation des travaux abordant la fouille de textes et la visualisation de l'information. Donc, pour notre revue de la littérature, nous nous concentrerons sur les principales publications qui nous ont permis d'acquérir les connaissances nécessaires et utiles à notre projet, selon deux grands axes : l'analyse et la gestion de l'information textuelle (AGIT) et la visualisation de l'information.

1.7.1. Littérature du domaine de l'AGIT

Dans le domaine de la lecture et de l'analyse de textes assistées par ordinateur, Dominic Forest a analysé et développé des méthodologies et des approches en classification et en catégorisation automatiques, afin de faciliter l'analyse thématique de corpus de textes. D'une part, il propose une chaîne de traitement méthodologique ; Numexco, pour analyser les textes philosophiques *Discours de la méthode* et *Méditations métaphysiques*, de Descartes. De plus, à travers ce projet, il démontre qu'une approche ascendante (classification automatique) est plus performante que l'approche descendante (classification de type traditionnelle telle que l'analyse de contenu), afin d'effectuer de la classification de textes et de l'analyse thématique. De la sorte, il confirme l'hypothèse voulant que « *[…] certains outils informatiques de pointe, lorsque appliqués à des tâches d'analyse de données textuelles, permettent d'assister le chercheur dans la découverte de thèmes au sein de textes philosophiques*. » (Forest : 2002, 3).

D'autre part, prenant pour objectif de développer, d'évaluer et de comparer deux méthodologies en classification et en catégorisation automatiques, Forest énonce l'hypothèse suivante :

> « L'identification automatique des thèmes et l'analyse thématique des données textuelles peuvent être assistées efficacement en employant certaines techniques issues de l'intelligence artificielle et de l'apprentissage machine à des fins de classification et de catégorisation automatiques des données textuelles » (Forest : 2006, 16).

En bref, en appliquant des théories et des méthodologies multidisciplinaires en analyse thématique (psycholinguistique, sémantique, linguistique textuelle) et en classification automatique, il développa un protocole opérationnel en fouille de textes, mettant en perspective des opérations de classification neuronale non supervisée et de catégorisation hybride neurofloue, sur des documents non structurés (articles de journaux). De plus, pour pouvoir appliquer plusieurs thématiques à un document, il propose de segmenter les paragraphes en plusieurs documents distincts. Finalement, il démontre qu'il est possible d'employer certains termes issus directement du lexique comme étiquettes thématiques afin de ne pas avoir à utiliser une taxinomie ou un corpus d'apprentissage.

Par ailleurs, nous remarquons les travaux en analyse thématique de données textuelles réalisés par Meunier (Meunier et Forest : 2004). À travers ces recherches, l'auteur aborde la problématique de l'analyse thématique en expérimentant des méthodes de classification et de catégorisation automatiques dans le but d'assister les chercheurs en sciences humaines et en littérature pour l'analyse thématique de textes (Meunier et Forest : 2004, 434). Ainsi propose-t-il des méthodologies dont la chaîne de traitement *Thématico*, réalisable en sept étapes : 1) constitution du corpus et identification des unités d'information à analyser (segmentation des textes), 2) vectorisation, 3) classification automatique des segments de textes, 4) extraction du lexique des classes obtenues à l'étape précédente, 5) catégorisation automatique (réalisée à partir de catégories thématiques puisées à même les textes, en fonction d'outils statistiques comme l'indice TF-IDF). 6) Projection des catégories thématiques sur le corpus, 7) Navigation, découverte et visualisation des thèmes identifiés. Brièvement, en conclusion de ses travaux, l'auteur confirme que les techniques de classification et de catégorisation automatiques peuvent aider les chercheurs pour l'analyse thématique de grands corpus de textes, mais remarque la composante subjective de la navigation thématique, variant selon les chercheurs qui thématisent (Meunier et Forest : 2004, 437).

Toujours sous la perspective du domaine LATAO, Jean Archambeault propose une application en fouille de texte fondée sur les réseaux neuronaux afin de représenter visuellement l'univers terminologique d'un domaine scientifique. Son projet de recherche s'apparente grandement au nôtre dans le sens où l'auteur tente de tracer l'évolution scientifique d'un domaine d'étude, celui de la spectroscopie par laser, selon une approche en fouille de textes (catégorisation automatique). De plus, en guise de corpus, il emploie également des résumés de recherche scientifique. En somme, l'objectif est de construire un réseau sémantique conceptuel pour permettre l'exploration et la visualisation thématique du domaine étudié. Les méthodes employées sont prédictives et non supervisées. En utilisant les outils d'analyse *Text Analyst 2.0, PAJEK* et *UCINET 5.0* (applications de fouille de texte, méthodes *k-Neighbours* et dénombrement *CORE+Degree*), il présente ses résultats graphiquement, en trois dimensions (logiciel *Mage*). Finalement, malgré des difficultés vis-à-vis l'interprétation des cartes conceptuelles produites, l'auteur en arrive à la conclusion que les outils et méthodes ont permis de démontrer la dynamique du vocabulaire du domaine scientifique de la spectroscopie par laser (Archambeault : 2003, 97).

Se situant précisément en sémantique interprétative, nous remarquons les travaux de Mathias Rossignol, où il est question de l'extraction automatique d'informations lexicales sémantiques. L'auteur prend pour objectif d'identifier les thèmes dans des segments de textes, en fonction de la cooccurrence de mots-clés. De plus, il suggère un découpage de l'espace sémantique à trois niveaux d'analyse : le domaine (mots parlant de la même chose), le taxème (« *rassemble des mots qu'il est possible d'employer les uns à la place des autres dans un texte au prix d'une variation de sens mineure* » (Rosignol : 2005, 10)) et les sèmes spécifiques (à l'intérieur d'un taxème, les sèmes sont des nuances qui permettent de distinguer les mots les uns des autres). Le corpus utilisé se constitue de 5 704 articles de la revue *Le Monde Diplomatique*, de 1985 à 1998 et regroupe 11 380 197 lexies (lemmes). Aussi, un élément que nous trouvons intéressant ; les thématiques du corpus sont variées, traitant aussi bien de politique, de géopolitique, de macroéconomie, de culture et de sociologie.

Pour réaliser ses travaux, l'auteur utilise une méthode de classification hiérarchique par analyse de la vraisemblance de liens (CHAVL). Sa démarche consiste à effectuer une classification hiérarchique sur les segments de textes, pour ensuite calculer la cooccurrence entre les mots de ces segments. Les mots retenus sont ceux dont l'indice de cooccurrence est la plus élevé (Rossignol : 2005, 7).

Abordons maintenant les nombreuses campagnes d'évaluation « DEFT » (DÉfi Fouille de Textes), créées en 2005 par le Laboratoire de Recherche en Informatique de l'Université Paris-Sud, à Orsay. Ces campagnes prennent pour principal objectif l'évaluation francophone de protocoles et de systèmes de fouille de textes (méthodes d'apprentissage machine), développés afin de résoudre des problématiques proposées par les organisateurs (inspirées des conférences TREC (TextREtrieval Conference : 2000). Ainsi, plusieurs équipes provenant tant des milieux privés que publics s'affrontent en soumettant leurs systèmes face à des problématiques scientifiques, cela finalement pour l'avancement de méthodes de fouille de textes. Se présentant en fonction des grandes thématiques abordées, la première édition en 2005 proposait d'identifier les locuteurs de plusieurs allocations politiques intégrées et mélangées dans un seul texte (Jacques Chirac ou François Mitterrand). Les éditions 2006 et 2008 portaient pour leur part sur la classification en genre et en nombre : en 2006, il était question d'identifier les ruptures thématiques de corpus issus de différents domaines (politique, juridique et scientifique), alors qu'en 2008, les participants devaient distinguer les genres journalistique (*Le Monde*) et encyclopédique (Wikipédia). Les éditions de 2007 et 2009 ont porté sur la fouille d'opinions ; dans le premier cas, la fouille était réalisée sur des avis argumentés (critiques de livres et de films), alors que pour l'autre édition, il était question de distinguer les caractères objectifs et subjectifs exprimés dans des textes multilingues (français, anglais et italien, articles de journaux et débats parlementaires). Puis, les éditions 2010 et 2011 ont traité de la variation diachronique de corpus d'archives de presse publiés sur une période de 144 ans (1801 à 1944). En 2010, l'objectif était d'identifier les décennies de parution d'extraits d'articles, ainsi que le pays d'origine de l'article, tandis qu'en 2011, il fallait identifier

l'année précise de publication. Finalement, en 2012, l'édition visait l'extraction de mots-clés pour indexer les contenus des articles utilisés en 2011. En bref, peu importe les éditions, les campagnes DEFT permettent la comparaison, l'évaluation et la diffusion de méthodes et de systèmes de traitement automatique des documents textuels, se situant principalement en classification automatique (Grouin et Forest : 2012, 28).

Par ailleurs, nous prenons également en considération les travaux d'Amal Zouaq, qui utilise une approche en ingénierie ontologique pour extraire et exploiter des connaissances à partir de textes. L'objectif principal poursuivi par l'auteur est de créer une architecture ontologique pour générer dynamiquement des objets de connaissance et d'apprentissage, en couplant des éléments des systèmes tutoriels intelligents et des systèmes e-Learning traditionnels (*Learning Knowledge Objects, LKO*). Ainsi, selon l'auteur, les LKO sont basés sur une structure sémantique et possèdent une capacité adaptative à un apprenant, faisant en sorte qu'ils puissent être exploités parallèlement aux systèmes tutoriels (Zouaq : 2007, iii). Finalement, ce qui permet de dire que cette recherche s'apparente à la nôtre réside dans les étapes méthodologiques effectuées ; l'auteur emploi des techniques de classification semi-supervisées (utilisation d'un corpus d'apprentissage), basées sur des méthodes statistiques TF-IDF (*Term Frequency-Inverse Document Frequency)*, modèle vectoriel, analyse sémantique latente, techniques de catégorisation et de classification), tout en intégrant des techniques linguistiques (bases de patrons, listes, thésaurus, glossaires, ontologies) (Zouaq : 2007, 34).

Puis, nous ne pourrions faire fi des récents travaux intitulés « Culturomics », effectués sous la direction de Jean-Baptiste Michel et d'Erez Lieberman Aiden. « *Culturomics is the application of high-thoughput data collection and analysis the study of culture* » (Michel: 2011, 181). L'équipe de recherche travaille sur la création d'un logiciel de fouille de textes à partir d'un sous-ensemble des documents numérisés par Google Books (5 millions de livres issus de 40 universités à travers le monde), sur une période allant de 1500 à 2000 apr. J.-C. Le but premier de ce projet est de permettre l'analyse des phénomènes linguistiques et culturels : « *Computational analysis of this*

corpus enables us to observe cultural trends and subject them to quantitative investigation » (Michel : 2011, 176). Un exemple cité par les auteurs est la comparaison fréquentielle des expressions « the Great War », «World War I» et «World War II», où les références à la Grande Guerre chutent dans les années 1940, car à partir de ce moment, celle-ci est davantage associée à la Première Guerre mondiale.

> « These examples highlight two central factors that contribute to culturomic trends. Cultural change guides the concepts we discuss (such as "slavery"). Linguistic change, which, of course, has cultural roots, affects the words we use for those concepts ("the Great War" versus "World War I"). In this paper, we examine both linguistic changes, such as changes in the lexicon and grammar, and cultural phenomena, such as how we remember people and events » (Michel et al.: 2011, 176).

Ce sont précisément les techniques statistiques employées qui attirent notre attention ; filtrage du lexique, fréquences, fréquences médianes, n-grams, etc. Aussi, selon nous, l'un des aspects fort intéressants de cette recherche est l'analyse chronologique des phénomènes, qui fait apparaître des tendances évolutives. Par exemple, les auteurs analysent la popularité de personnalités comme Charles Darwin et en viennent à la conclusion que jusqu'au milieu du 20^{e} siècle, le temps moyen avant qu'une personne devienne populaire (citée dans les livres) était de 50 ans après son décès, alors qu'aujourd'hui, les individus deviennent plus rapidement populaires, mais tombent également dans l'oubli plus rapidement (Michel et al. : 2011, 180). En bref, comme les publications utilisées couvrent une large période de temps, il est possible de percevoir des tendances et des phénomènes culturels et littéraires qui seraient moins visibles sans les techniques de fouilles de textes.

1.7.1. Littérature en visualisation de l'information

Pour commencer, nous proposons d'aborder l'article de Borgatti et al, à travers lequel les auteurs établissent une synthèse évolutive des méthodes d'analyses des réseaux en sciences sociales. Aussi, puisque les méthodes de visualisation sous forme de graphes sont fortement basées sur l'analyse réseau, nous trouvons judicieux de débuter par cette revue méthodologique. Donc, en premier lieu, les auteurs présentent comment la théorie des graphes et la modélisation schématique et graphique furent introduites en sciences sociales afin d'expliquer des phénomènes sociaux. Ainsi est-il remarqué, par exemple, les apports d'Émile Durkheim pour expliquer l'influence de l'environnement social sur l'individu :

> « Fifty years after Comte [aux environs de 1890], the French sociologist Durkheim had argued that human societies were like biological systems in that they were made up of interrelated components. As such, the reasons for social regularities were to be found not in the intentions of individuals but in the structure of social environments in which they were embedded» (P. Borgatti et al.: 2009, 892).

Dans les années 1950, l'intrusion des matrices algébriques ainsi que le développement de laboratoires expérimentaux solidifièrent les fondements de l'approche réseau et rendit possible la découverte de groupes émergents à l'intérieur de données réseaux. Dans les décennies suivantes, de plus en plus d'anthropologues s'intéressèrent à l'analyse réseau dont, entre autres, Claude Lévi-Strauss, qui représenta la parenté tel un système relationnel pouvant être soumis aux mathématiques et à l'algèbre. « *It was soon discovered that the kinship systems of such peoples as the Arunda of Australia formed elegant mathematical structures that gave hope to the idea that deep lawlike regularities might underlie the apparent chaos of human social systems* » (P. Borgatti et al., 2009: 893). Au tournant des années 1980, l'analyse réseau fut définitivement reconnue et s'implanta dans différents champs disciplinaires et milieux professionnels dont, notamment, en santé publique ou en criminologie. Remarquons en terminant qu'à travers leur exposé, les auteurs abordent plusieurs concepts de l'analyse réseau, tels que les notions d'ouverture et de fermeture des

réseaux, la notion de similarité ou, encore, le concept de connectivité (P. Borgatti et al. : 2009, 893).

D'autres travaux en visualisation de l'information que nous souhaitons souligner sont ceux de Santo Fortunato, concernant les méthodes de détection des communautés dans des graphes. Selon lui, sous forme de graphes, les communautés peuvent être considérées comme des classes ou « *clusters* » et permettent de visualiser des systèmes structuraux, représentant une issue analytique considérable pour les sciences sociales, la biologie et l'informatique. De plus, s'intéressant aux récents développements algorithmiques, l'auteur prend comme objectif de présenter les principaux travaux et avancements pour le repérage des communautés, prêtant une attention particulière aux contributions de la physique statistique. Ainsi, à travers les différentes méthodes exposées, nous retrouvons, entre autres, les méthodes traditionnelles de classification hiérarchique, celles basées sur l'inférence statistique, les techniques de partitionnements graphiques, ou, encore, les méthodes fondées sur le concept de modularité (« *modularity-based methods*»). Finalement, en conclusion de ses travaux, l'auteur arrive au constat que malgré les avancements, les méthodes d'analyses réseaux ne possèdent pas de cadre théorique défini et unanime aux différents champs disciplinaires. Il deviendra alors important, dans les années à venir, d'établir des mécanismes de contrôle pour l'évaluation de la qualité et permettre la comparaison des méthodes algorithmiques de classification (Fortunato : 2010, 90).

Par ailleurs, nous souhaitons mentionner les travaux en visualisation de Rafols et al (2010), par lesquels les auteurs présentent une approche de cartographie (« *overay maps* »), favorisant la visibilité des structures relationnelles. L'objectif premier est de présenter leur méthode tel un outil pour assister les chercheurs dans l'exploration graphique de données, que ce soit tant dans les domaines scientifiques que commerciaux. Pour illustrer leur approche, les auteurs proposent quelques exemples dont l'exploration des transformations sociocognitives de 18 disciplines universitaires (provenant de 3 universités), en catégorisant des articles et journaux scientifiques afin

de générer une matrice de similarité. Bien qu'il ne soit pas question du même logiciel de visualisation (Pajek) que dans nos travaux, plusieurs aspects demeurent pertinents pour notre recherche ; d'abord, les auteurs reconnaissent que les disciplines ne se compartimentent pas toujours en branches disciplinaires (nous parlons alors d'inter-, de multi; et de transdisciplinarité), ensuite, ils démontrent que les cartes globales et locales mettent en évidence les relations de similarité entre les éléments, et, finalement, qu'elles facilitent les comparaisons. Remarquons en terminant que les cartes produites dans leurs recherches tendent à démontrer que les sciences sociales sont liées à la psychologie, aux sciences de la santé, aux sciences cognitives, aux sciences économiques, politiques et géographiques, ainsi qu'à l'informatique (Rafols et al. : 2010).

Le projet *Réseaux, Traces et Controverses*, dirigé par Franck Ghitalla (2009), propose une approche en analyse et en visualisation de l'information, basée sur les réseaux numériques d'information. Le but général de ce projet est de produire des hypothèses et des données expérimentales sur les controverses « science-société », telles que véhiculées sur les réseaux sociaux du Web. Plus précisément, l'objectif de l'auteur est d'analyser qualitativement et quantitativement des masses d'information afin de produire des cartographies d'agrégations des controverses. Aussi, ce projet se situe, d'une part, entre la sociologie et l'anthropologie pour l'interprétation théorique (l'analyse qualitative), et d'autre part, entre le *Web Mining* et l'exploration de réseaux pour l'analyse statistique. Les données extraites, 7 670 mots, proviennent de grands moteurs de recherche et le processus de classification a permis la formation de 113 classes distinctes (clusters). Finalement, les visualisations produites ont été réalisées à l'aide du logiciel Gephi, sous forme de graphes dynamiques, permettant l'exploration des résultats de recherche, d'où l'intérêt que nous portons à ce projet (Ghitalla : 2009).

Par ailleurs, nous aimerions souligner les travaux de Dmitry Paranyus Kin, du Nodus Labs de Berlin, portant sur l'analyse réseau de textes. L'auteur présente une méthodologie d'analyse réseau et des algorithmes pour analyser les relations sémantiques de textes. La méthodologie proposée se décompose ainsi : il y a

suppression des mots vides, racinisation, normalisation, exportation des textes en XML pour les visualiser en graphes, application d'algorithmes de visualisation (dont l'algorithme *Modularity class*) et l'analyse des structures graphiques pour interpréter les résultats. Même si ces travaux s'appliquent précisément sur l'analyse réseau, ils demeurent forts pertinents pour notre recherche du fait que nous utilisons le même système de visualisation des résultats (Gephi). Qui plus est, l'auteur démontre comment interpréter les résultats graphiques découlant du processus algorithmique, en expliquant par exemple comment l'algorithme *Force Atlas* fonctionne et comment celui-ci influence les graphes (spatialisation des communautés) (Paranyus Kin : 2011, 17).

Se rapprochant de notre projet par les thématiques abordées et la méthodologie de classification (à l'aide de Gephi), nous remarquons les travaux de Zack Batist sur l'analyse réseau afin d'explorer la thématique de festivité de l'Âge de Bronze, dans la région de la Grèce (Aegean), à travers des artefacts archéologiques. Plus précisément, l'auteur cherche à explorer le mouvement de hiérarchisation sociétaire, visible par la présence d'objets luxueux sur des sites funéraires (2995 poteries trouvées sur 10 sites). Aussi, afin de créer le réseau d'analyse sous forme de graphe, l'auteur utilisa les sites archéologiques et les différents types de poteries comme nœuds et les liens furent calculés par des mesures de degré, de centralité et de classification (algorithme *Modularity Class*). Brièvement, les résultats obtenus démontrent le groupement de deux principaux groupes de poteries ; l'un se situant en Crète et l'autre dans le sud de la Grèce. De plus, l'auteur en arrive à la conclusion que certains types de poterie sont communs à tous les sites, suggérant l'homogénéité régionale. Par contre, les sites de la Crète présentent dans l'ensemble des assemblages plus variés et luxueux que les sites du sud de la Grèce (Batist, Zack : 2012, 31).

Nous remarquons en dernier lieu les travaux de Vincent Labatut et de Jean-Michel Balasque (2012), portant sur la détection de communautés en analyse réseau. Devant le fait de la profusion des algorithmes de détection des communautés, ils démontrent la difficulté inhérente aux choix d'utilisation de tel ou tel algorithme, selon

les données et les structures à analyser. Ainsi prennent-ils comme objectif d'aborder cette problématique sous l'angle de l'utilisateur et présentent des approches et des outils méthodologiques pouvant être utilisés pour analyser structurellement les communautés. Notamment, ils visent à démystifier les différentes approches algorithmiques telles que l'approche Louvain, et ils expliquent comment interpréter et évaluer les classes formées, selon, d'une part, des notions topologiques (densité des graphes, la transitivité [*clustering coefficient*], le degré de distribution, etc.), et d'autre part, selon les structures des communautés (l'homogénéité / l'hétérogénéité, les attributs des nœuds, des liens inter et intra-classes). Finalement, pour appuyer leur description méthodologique, les auteurs utilisent 552 fiches d'un sondage effectué auprès d'étudiants de l'Université d'Istanbul, en Turquie, et démontrent comment ils extraient et interprètent les représentations graphiques obtenues avec l'algorithme *«Fast Greedy*» (famille des algorithmes hiérarchiques) (Labatut, Vincent et Jean-Michel Balasque : 2012, 16).

Deuxième partie

2. Cadre méthodologique

Dans cette deuxième section, nous présentons d'abord une description des données utilisées pour notre recherche, nous amenant ensuite à définir l'anthropologie québécoise. Ainsi souhaitons-nous mettre en contexte nos données et notre approche d'analyse. Puis, dans un deuxième temps, nous exposons les étapes méthodologiques avec lesquelles nous avons conduit ce projet de recherche.

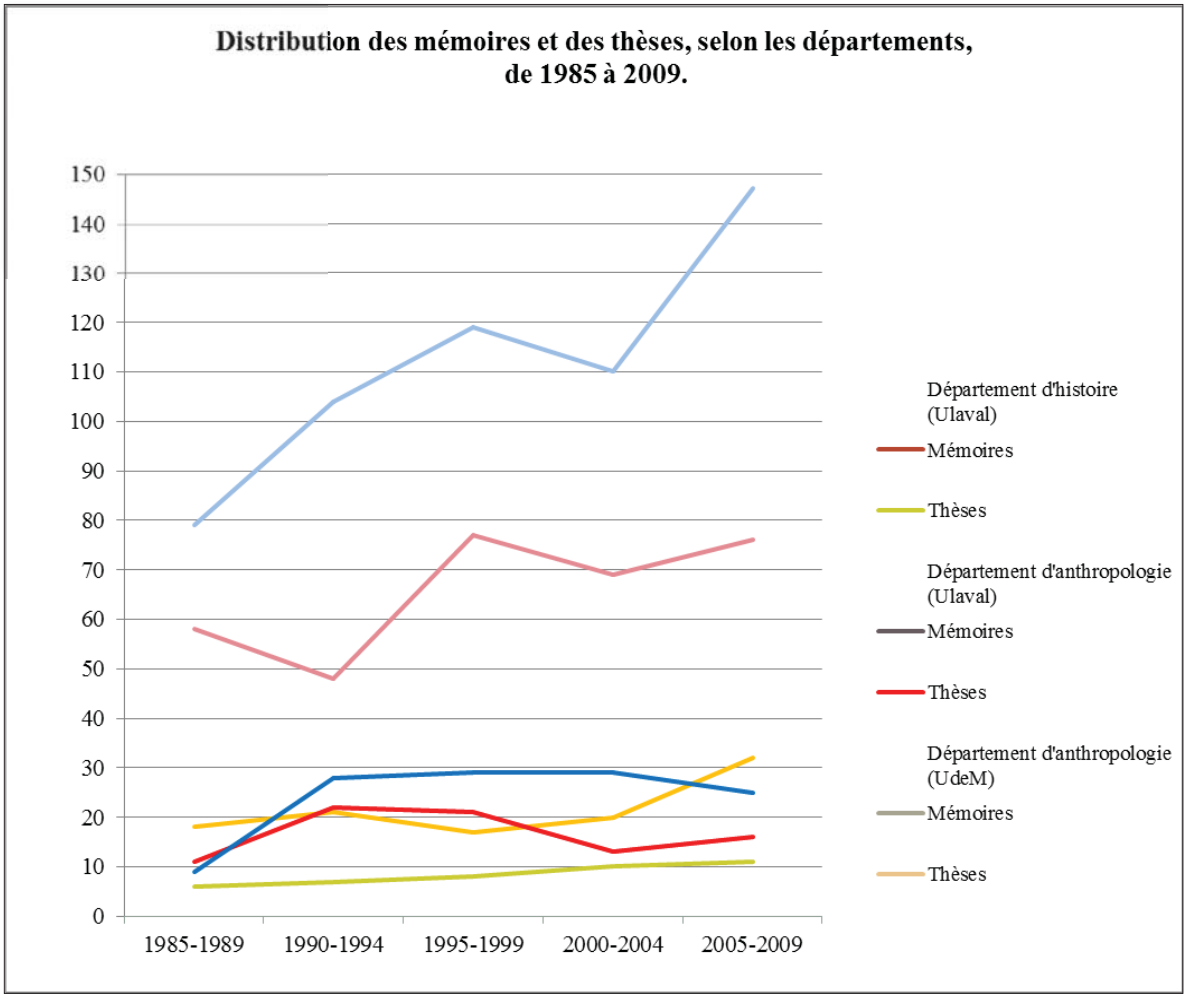
2.1. Description des données

Tel que mentionné précédemment, les textes que nous analysons consistent en des résumés de mémoires et de thèses en anthropologie, octroyés par les départements d'anthropologie et d'histoire de l'Université Laval, ainsi que par le département d'anthropologie de l'Université de Montréal, de 1985 à 2009.

La longueur des textes varie généralement entre une demi-page et une page et demie (150 à 350 mots), mais certains résumés présentent plusieurs pages, pour un maximum de 14 pages (4 528 mots, résumé de Cruciatti : 1999). Selon nos observations, les résumés de plus de 350 mots représentent environ 12 % de notre corpus; ils proviennent souvent du département d'anthropologie de l'Université de Montréal et ils ont été écrits, pour la majorité, dans les années 1985 à 1990. Nous expliquons cette disparité de longueur de textes par l'intégration de normes de présentation des mémoires et des thèses, instaurées au sein des facultés.

Remarquons qu'à travers les 1240 résumés, les quatre branches disciplinaires de l'anthropologie sont abordées : l'archéologie, la bioanthropologie ou anthropologie biologique, l'ethnologie, aussi dite anthropologie socioculturelle, et l'ethnolinguistique.

Voici quelques statistiques descriptives de nos données :



	1985-1989	1990-1994	1995-1999	2000-2004	2005-2009	Totaux
Département d'histoire (Ulaval)						
Mémoires	18	21	17	20	32	108
Thèses	6	7	8	10	11	42
Département d'anthropologie (Ulaval)						
Mémoires	58	48	77	69	76	328
Thèses	11	22	21	13	16	83
Département d'anthropologie (UdeM)						
Mémoires	79	104	119	110	147	559
Thèses	9	28	29	29	25	120

Figure 5. Répartition chronologique des mémoires et des thèses, selon les départements universitaires.

Le graphique précédent montre une tendance générale à l'accroissement dans le temps du nombre de dépôts de mémoires et de thèses. Cela s'explique certainement par l'augmentation constante du nombre d'étudiants universitaires, ainsi que par la croissance des départements qui ont, au fil du temps, accueilli davantage de chercheurs et d'enseignants (Beaudoin : 2010).

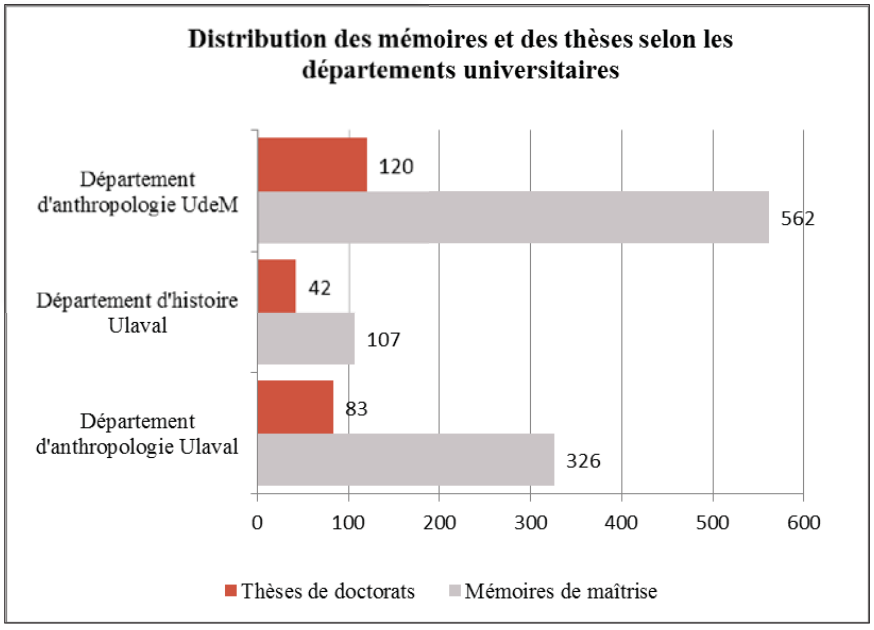


Figure 6. Distribution des mémoires et des thèses octroyés par les départements, de 1985 à 2009.

Ce graphique-ci montre bien la disparité entre le nombre de thèses de doctorat et le nombre de mémoires de maîtrise. Ainsi, au département d'anthropologie de l'UdeM, il y a 82,3 % de dépôts de mémoires par rapport à 17,7 % pour les thèses ; au département d'anthropologie de l'ULaval, il y a 79,8 % de mémoires contre 20,2 % de thèses, et finalement, au département d'histoire de l'ULaval, les proportions sont de 72,0 % de mémoires contre 28,0 % de thèses. Bref, il y a en moyenne 22,0 % des étudiants qui poursuivent au doctorat et qui déposent des thèses.

Distribution des mémoires et des thèses (grappes d'années), selon les départements universitaires et en fonction des sous-disciplines.

	1985-1989	1990-1994	1995-1999	2000-2004	2005-2009	Total général
UdeM	88	132	148	139	172	679
Archeologie	18	20	30	27	35	130
Bioanthropologie	13	21	7	16	25	82
Ethnolinguistique	5	3	11	10	8	37
Ethnologie	52	88	100	86	104	430
UlavalA	69	70	98	82	92	411
Ethnolinguistique	1	1	3	0	1	6
Ethnologie	68	69	95	82	91	405
UlavalH	24	28	25	30	43	150
Archeologie	14	18	13	20	18	83
Ethnologie (folklore)	10	10	12	10	25	67
Total général	181	230	271	251	307	1240

Tableau 1. Distribution des mémoires et des thèses sur des périodes de cinq ans, selon les départements universitaires et en fonction des sous-disciplines.

Ce tableau présente la répartition dans le temps des sous-disciplines abordées dans les mémoires et les thèses. Pour les trois départements, l'ethnologie demeure la branche disciplinaire la plus étudiée. Ensuite viennent l'archéologie, l'anthropologie biologique et l'ethnolinguistique. Il nous est difficile d'émettre une explication complète quant à la popularité disciplinaire, mais nous pouvons néanmoins avancer que l'ethnologie est apparue dans les universités québécoises dès les débuts du 20^{e} siècle (souvent associée à la sociologie), alors que les autres sous-disciplines se sont développées et imposées plus tardivement au sein des facultés. « L'anthropologie naissante [décennie 1960] sera rapidement vue par les départements de sciences sociales comme une discipline dont la mission est d'ouvrir le Québec sur le monde, à travers l'étude des autres sociétés et une réflexion fondamentale sur l'humain » (Bibeau et al. : 2011, 3).

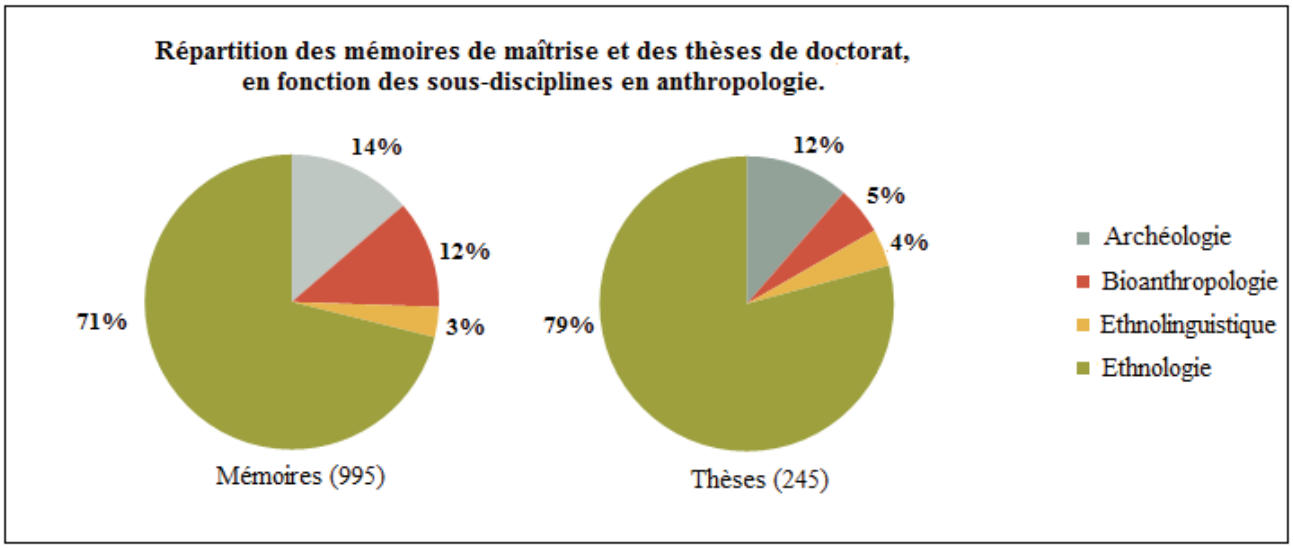


Figure 7. Répartition des mémoires de maîtrise et des thèses de doctorat selon les sous-disciplines.

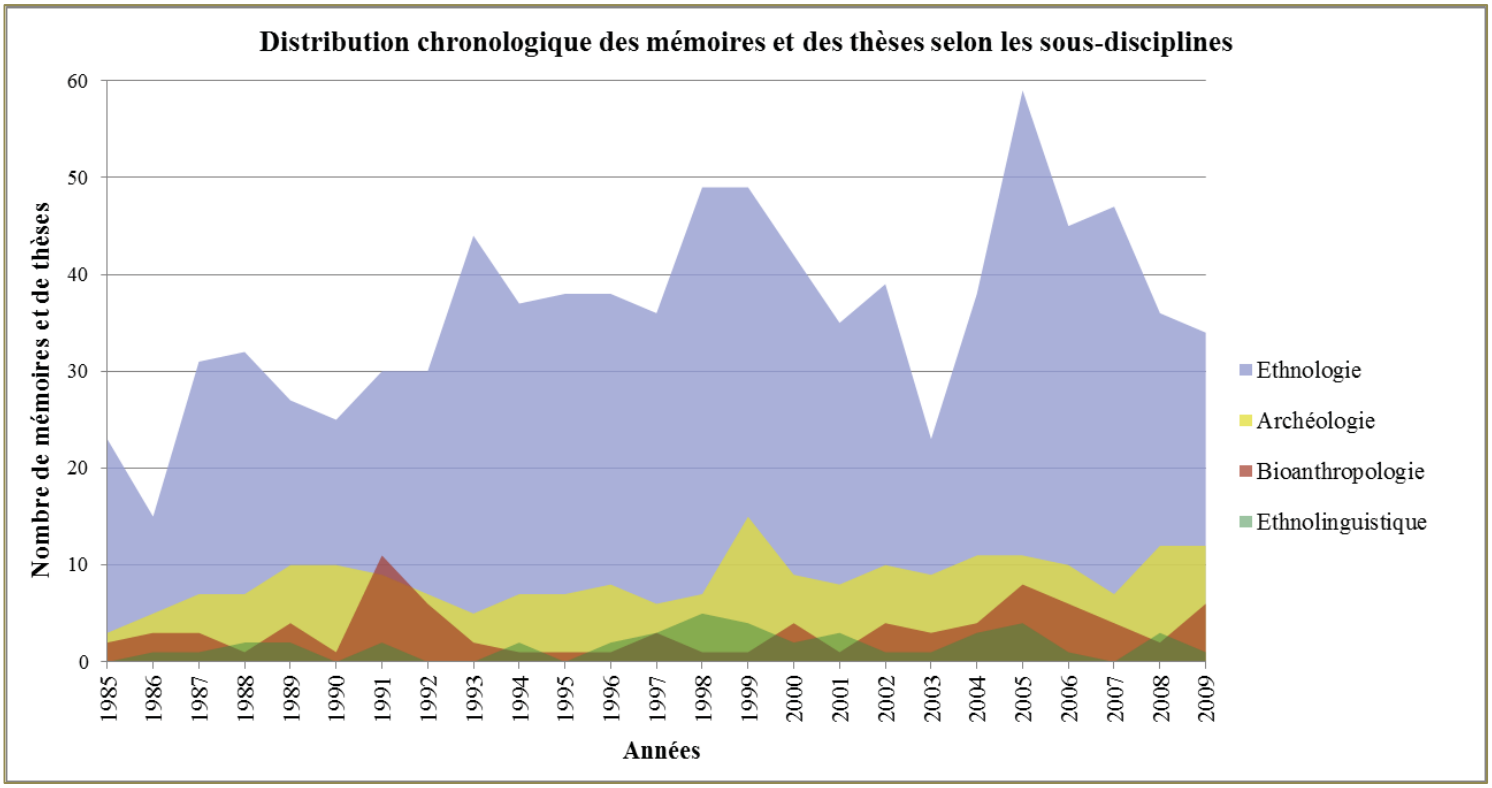


Figure 8. Distribution chronologique des mémoires et des thèses en fonction des sous-disciplines.

2.2. Définition de l'anthropologie québécoise

Globalement, nous pouvons définir l'anthropologie comme étant la discipline qui étudie la culture et l'humanité sous toutes ses dimensions : biologique, historique, comportementale et communicative;

> « Réflexion scientifique et humaniste portant sur la variabilité et la similitude du fait humain, l'anthropologie offre un regard original sur la culture, se caractérisant surtout par l'étendue de sa vision, qui se pose un peu partout dans le monde, et par son attention aux particularités et aux détails du quotidien. Pour l'anthropologue, derrière chaque geste individuel ou derrière chaque objet, c'est toute la société et la culture qui se profilent et qu'il faut décoder » (Département d'anthropologie de l'Université de Montréal, Site Internet : Présentation du département).

Typologiquement, le terme « anthropologie » vient des mots grecs *anthropos* et *logos*, qui signifient respectivement « homme » et « science ». Donc, d'un côté, nous avons l'être humain vu sous la perspective biologique et de l'autre, l'être humain est étudié dans sa diversité culturelle (St-Denis : 2006, 2).

En Amérique du Nord, il est couramment reconnu que l'anthropologie se divise en quatre grands champs disciplinaires : l'ethnologie, aussi dite l'anthropologie sociale et culturelle, l'anthropologie physique ou, autrement dit la bioanthropologie (aussi nommée l'anthropologie biologique), l'archéologie et, finalement, l'ethnolinguistique (approche nord-américaine). Dans le reste du monde et particulièrement en Europe, l'anthropologie n'est pas perçue comme une discipline unifiée; elle est davantage axée sur le culturel et l'ethnographie, laissant l'archéologie à l'histoire et l'anthropologie biologique à la biologie et aux sciences de la santé (approche européenne) (Dumonchel : 2009, 5). Brièvement, l'ethnologie s'intéresse aux faits culturels, à la variabilité socioculturelle, ainsi qu'aux sociétés humaines, la bioanthropologie traite des origines, de l'évolution et de la variabilité de l'espèce humaine, l'archéologie aborde l'évolution et l'histoire humaine à partir des restes archéologiques (périodes historiques et préhistoriques), et l'ethnolinguistique étudie la variabilité des pratiques linguistiques, passées comme présentes (Bonte, Pierre et al. : 1991). Finalement, nous devons remarquer une distinction qui existe entre l'anthropologie socioculturelle (aussi dite l'ethnologie, enseignée aux

départements d'anthropologie de UdeM et Ulaval) et l'ethnologie qui est enseignée au département d'histoire de Ulaval, qui puise ses origines du folklore. La nuance n'est pas évidente et peut-être discutée, mais essentiellement, l'ethnologie prise dans le sens du folklore se concentre sur l'étude de « sa culture » (la culture québécoise en l'occurrence), alors que l'anthropologie socioculturelle, telle qu'enseignée dans les départements d'anthropologie, élargit son objet d'étude à toutes les cultures existantes. Il faut remarquer que les approches, les méthodes et les techniques de recherche sont très souvent les mêmes dans ces deux disciplines : travail de terrain, observations, collectes de données, analyses qualitatives et quantitatives, analyses de contenu, analyses de récits de vies, analyses comparatives…

> « L'ethnologie mène à la découverte de Soi et de l'Autre proche, c'est-à-dire qu'elle privilégie l'étude de sa propre culture dans une perspective d'ouverture au métissage des cultures. Elle prend pour objet d'observation et d'étude les productions et expressions culturelles et symboliques et les savoirs et savoir-faire, d'où l'intérêt porté au patrimoine matériel et immatériel. » (Département d'histoire de l'Université Laval, présentation de l'ethnologie : http://www.hst.ulaval.ca/le-departement/disciplines/ethnologie/ (consulté le 20 novembre 2012)).

Par ailleurs, en prenant l'approche nord-américaine de l'anthropologie, le folklore est généralement perçu comme faisant partie de l'anthropologie socioculturelle. Ainsi n'y a-t-il pas de grandes différences entre un folkloriste québécois et un ethnologue se spécialisant sur l'étude de la culture québécoise. En somme, nous avons favorisé cette approche (nord-américaine) de l'anthropologie dans notre décision de collecter et d'intégrer à notre corpus les résumés en ethnologie, du département d'histoire de l'Université Laval.

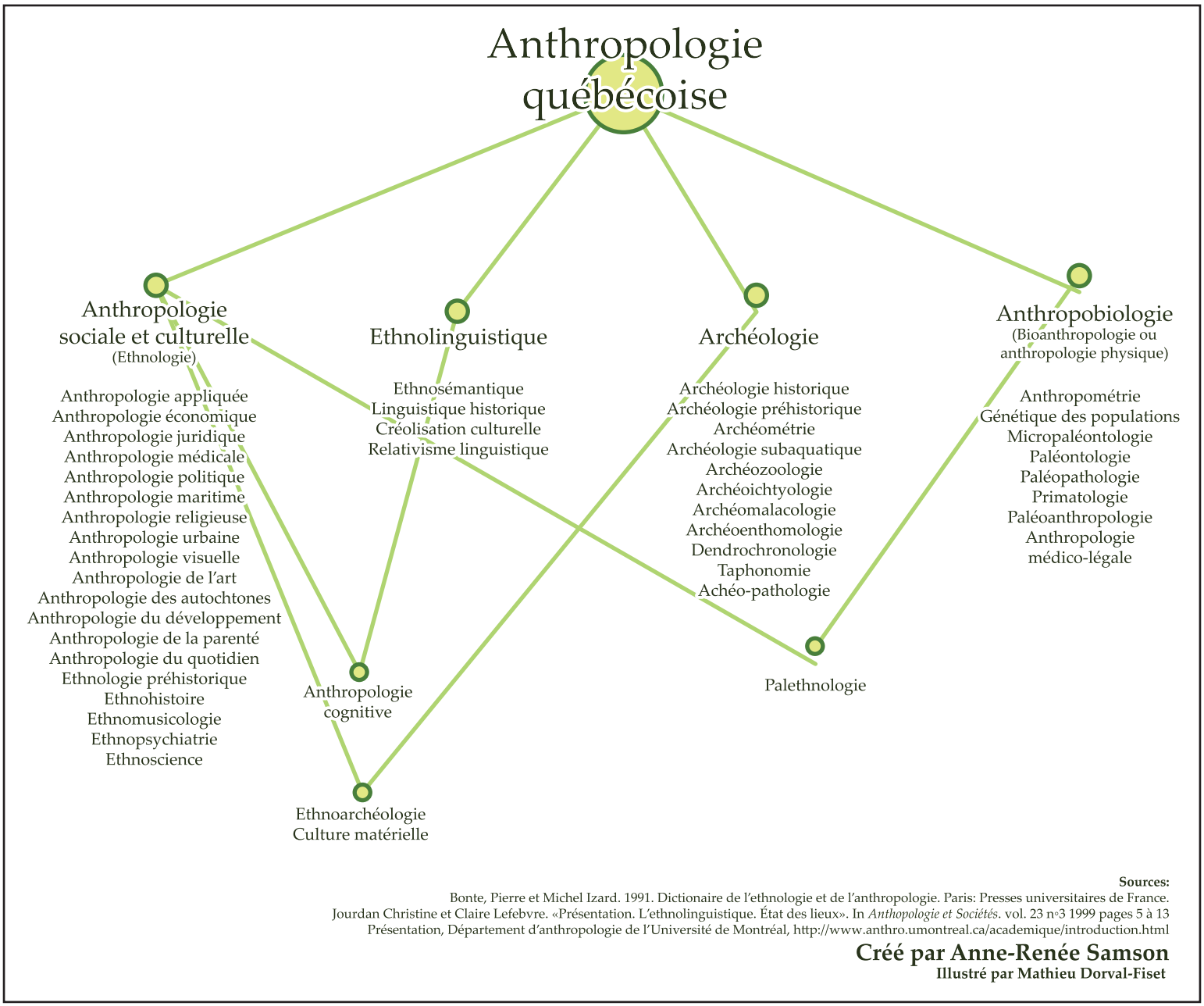


Figure 9. Schéma représentant, selon-nous, les sous-disciplines de l'anthropologie québécoise en fonction de l'approche nord-américaine.

Il est à remarquer que dépendamment des perspectives et des points de vue, certaines sous-disciplines sont partagées par plusieurs branches disciplinaires. Par exemple, l'étude de la culture matérielle peut-être contemporaine, employée en ethnologie, ou historique et préhistorique, prise dans une perspective archéologique. En fait, les sous-disciplines s'empruntent souvent entre elles des théories et des approches, provoquant un métissage scientifique. Conséquemment, il peut être ambigu de catégoriser manuellement les sous-disciplines anthropologiques. Néanmoins, cette représentation schématique a le mérite de démontrer la diversité des approches et des objets d'études de l'anthropologie québécoise.

2.3. Processus méthodologique

2.3.1. Collecte des données

La collecte des données s'est déroulée sur une période d'un peu plus de cinq mois et nous avons utilisé les collections départementales de mémoires et de thèses pour numériser les résumés se trouvant en format papier. En outre, environ le tiers des mémoires et des thèses étaient déjà en format numérique, disponibles à partir de sites Internet départementaux, de catalogues de bibliothèques ou à partir de la collection *Thèses Canada*, de Bibliothèque et Archives Canada (BAC). Ainsi, pour tous les résumés déjà numérisés, nous les avons importés et convertis en format texte pour construire notre corpus. Quant aux résumés que nous avons numérisés (numérisation en images « *JPEG* »), nous les avons convertis en format texte à l'aide d'un logiciel de reconnaissance optique des caractères (*TopOCR*, version 3.1). Par ailleurs, bien que ce type de logiciels se soit grandement amélioré depuis quelques années, les conversions présentaient des erreurs de lecture, surtout pour les textes dactylographiés (par exemple, les « I » étaient souvent pris pour des « 1 », ou les « e » pour des « c ») et nous devions vérifier chacun des textes pour réduire au maximum ces erreurs.

Finalement, dans le but d'avoir un recensement le plus juste possible des mémoires et des thèses rédigés sur la période étudiée, nous nous sommes appuyés sur des listes de dépôts, produites à partir de l'*HypoThèse*, base de données créée en 2008 par l'Association des anthropologues du Québec, qui regroupe les références de mémoires et de thèses anthropologiques de McGill, Concordia, l'Université de Montréal et l'Université Laval (plus de 2 200 références, de 1944 à 2009). La construction de cette base de données s'est faite, quant à elle, à partir de listes fournies par les départements (Association des anthropologues du Québec : 2008).

Remarquons en terminant que nous avons nommé nos fichiers de façon à conserver une trace de nos variables ; départements, années, grappes d'années, niveaux

de scolarité et sous-disciplines. Cette dernière variable, les champs disciplinaires, fut catégorisée manuelle par les départements universitaires.

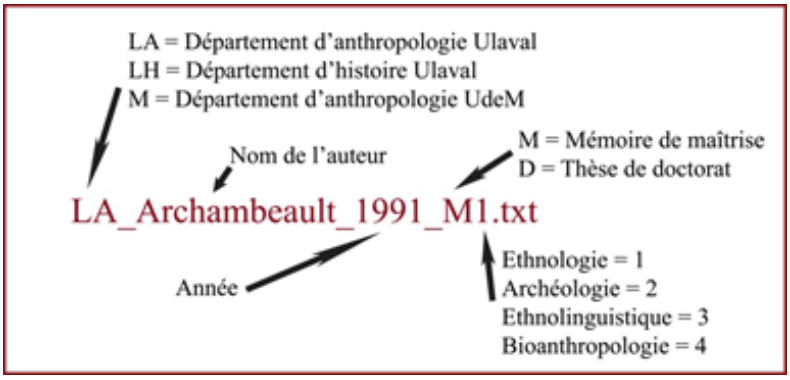


Figure 10. Exemple d'un nom de fichier, avec ses variables codées.

2.3.2. Prétraitement des données

Après avoir rassemblé tous les résumés en fichiers textes, nous avons importé ceux-ci dans le logiciel *SimStat* pour construire notre base de données, à partir de laquelle nous pouvions réaliser nos analyses. Remarquons brièvement que *SimStat* est un module d'analyse quantitative faisant partie de la suite *Provalis*, qui intègre également *QDA Miner*, un module d'analyse qualitative et *WordStat*, un logiciel de fouille de textes. Ces trois modules fonctionnent conjointement. Pour en revenir à notre base de données, nous avons importé les 1 240 textes et pour chacun, nous avons codifié nos variables : Départements, années, grappes d'années, niveau de scolarité et sous-disciplines. La codification était de type nominal (exemple, 1 = ethnologie, 2 = archéologie, etc.) et numérique.

Ensuite, nous avons commencé le prétraitement des données, afin de normaliser les résumés et de les préparer au processus d'analyse. « *Pour étudier un texte, le « bon corpus » est d'abord constitué des textes qui partagent le même genre* » (Rastier : 2011, 34). Concrètement, nous avons supprimé les titres « Résumé », « Sommaire » et les

signatures d'auteurs et de directeurs lorsque présentes, car ces informations ne sont finalement pas pertinentes pour nos analyses. Nous avons également éliminé les références et les mots-clés se trouvant en bas de page, et ce, pour des raisons de rigueurs. Selon notre corpus, c'est à partir des années 1995 que les mots-clés deviennent systématiques, probablement en conséquence à l'introduction, au sein des facultés, de normes de présentation des mémoires et des thèses. En bref, environ le tiers des résumés de notre corpus n'avaient pas de mots-clés, d'où la nécessité de les supprimer dans notre processus de normalisation. Aussi croyons-nous que les auteurs utilisent des mots significatifs à même les textes des résumés et ainsi, la suppression des mots-clés en bas de page n'a pas, selon nous, de répercussions notables, car les mots extraits du processus sont véritablement tirés des textes et non de listes artificielles de mots-clés.

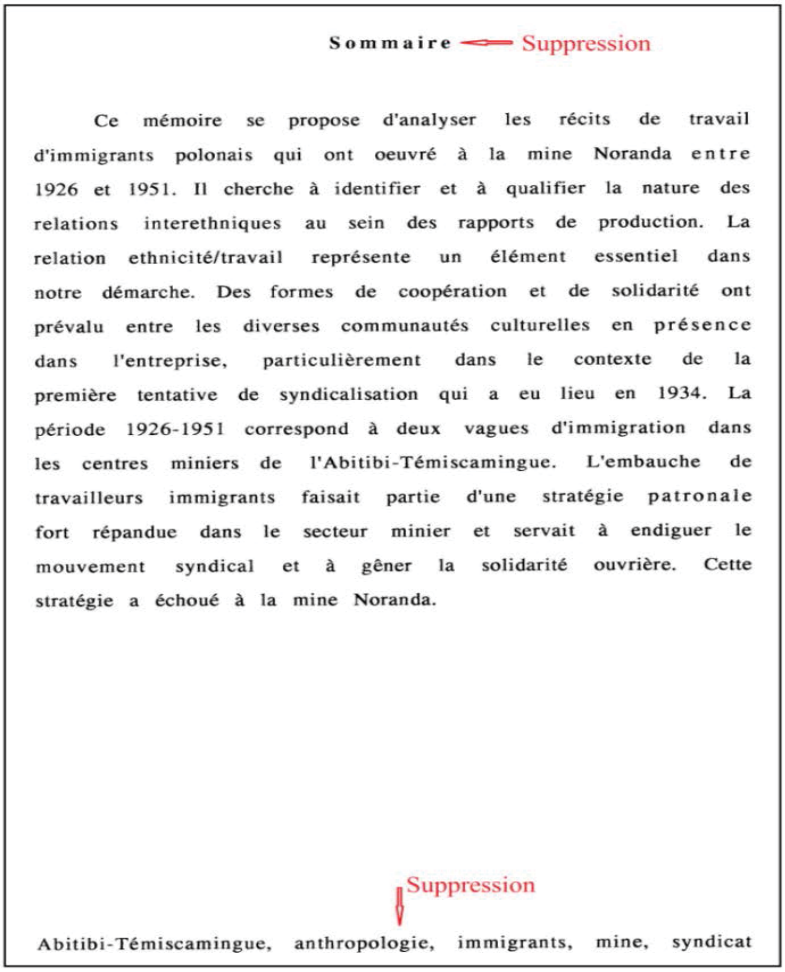


Figure 11. Suppressions effectuées lors du prétraitement. Beaupré : 1998.

Par ailleurs, dans le but de réduire les effets de variances linguistiques de la langue, nous avons également entrepris, lors du prétraitement, la normalisation manuelle de plusieurs expressions. En fait, comme toutes les langues vivantes, le français évolue dans le temps et présente beaucoup de variances morphosyntaxiques et synonymiques ; ethno-culturel et ethnoculturel, 18^{e} siècle / XVIIIe siècle / dix-huitième siècle ou, encore, inter-culturel et interculturel.

De plus, puisque WordStat ne reconnaît pas les traits d'union et les lignes de soubassement (par exemple, Abitibi-Témiscamingue est considérée comme deux mots

distincts), nous avons lié les expressions les plus utilisées de notre corpus. Ainsi, « pays en développement devenait « PaysEnDéveloppement » et « Tiers monde » devenait « TiersMonde ». Ce processus de normalisation manuelle nous permettait d'augmenter la fiabilité et la précision de nos résultats et nous l'avons effectué avec l'application simple mais efficace du *Rechercher / remplacer*, dans le module QDA Miner.

2.3.3. Filtrage des données

Notre troisième étape méthodologique réside dans le filtrage du lexique, qui se décline en plusieurs sous-opérations et selon deux approches : l'une linguistique, l'autre statistique. Peu importe l'approche, le but est de retirer du lexique les termes indésirables, lesquels ne sont pas pertinents pour l'analyse. Le filtrage de type statistique est assez simple; il conduit à ordonnancer les termes candidats et de cette liste, les chercheurs choisissent les termes à rejeter, selon des indices tels que la fréquence absolue ou le TF-IDF. Quant à l'approche linguistique, elle permet de filtrer les relations de variation, dont les déclinaisons de genre et de nombre en français (Ibekwe-SanJuan : 2007, 128).

Dans un premier temps, nous avons supprimé les mots fonctionnels, c'est-à-dire les mots outils, aussi dits les mots-vides, les mots grammaticaux ou "*trivial-words"* (Lerot : 1997, 317). Ce sont donc des prépositions, des pronoms et certains adjectifs, verbes et adverbes. Pour supprimer ces mots-vides, nous avons utilisé un dictionnaire de mots outils, aussi appelés *antidictionnaire* et "*stoplist"*, qui était intégré au logiciel WordStat. Puis, afin d'adapter ce dictionnaire à notre corpus, nous avons ajouté un peu plus de 650 mots aux 240 mots fonctionnels qui se trouvaient initialement dans le dictionnaire.

Tel que mentionné par Forest (2006), la pertinence des termes en fouille de textes résulte de leur valeur "discriminante", c'est-à-dire la valeur qu'un terme a de représenter un ou plusieurs documents. Selon la figure ci-dessous, le processus de filtrage est représenté en fonction de la distribution des mots du lexique. Les mots qui sont rejetés

sont soit des mots rares, ayant une faible fréquence, soit des mots fonctionnels qui sont très fréquents.

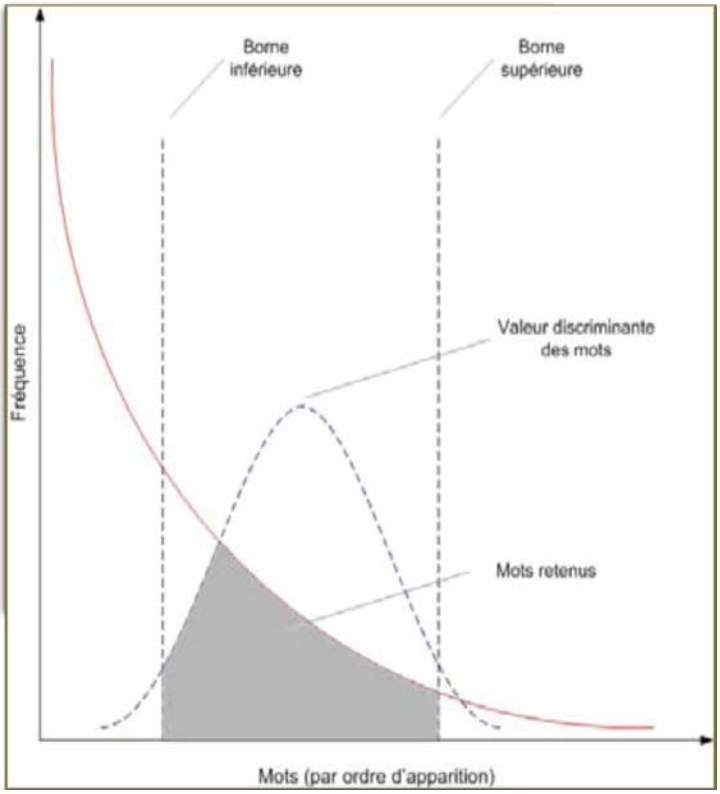


Figure 12. Les termes à retenir suite aux opérations statistiques de filtrage. Schultz : 1968, 120 ; Van Rijsbergen : 1979.

Dans un deuxième temps, nous nous sommes attardés à la suppression des relations de variances et de flexions linguistiques du français. Tel que précédemment mentionné, ce processus se nomme la *lemmatisation*, et nous avons ainsi réduit les principales flexions, qui se caractérisent selon : A) la conjugaison B) la déclinaison ou variation des noms, adjectifs, déterminants et pronoms en genre et en nombre » C) les degrés de comparaison des adjectifs et des adverbes (Lerot : 1997, 318). Tout comme la technique de *stemming* (réduction des termes selon leurs racines), la lemmatisation

permet de réduire le nombre de lexèmes du corpus, optimisant les résultats d'analyses (Forest : 2006, 39).

Finalement, une troisième sous-opération fut nécessaire pour adapter notre filtre à notre corpus, afin de diminuer les dernières ambiguïtés sémantiques, souvent causées par la polysémie (un mot possède plusieurs sens) et l'homonymie (un signifiant possède plusieurs sémèmes) (Rossignol : 2005, 32). Un exemple flagrant de polysémie de notre lexique est le suivant; le mot « mémoire » pourrait être intéressant pour nos analyses, principalement lorsqu'il est pris dans le sens « mnésique » (mémoire vivante, mémoire patrimoniale, etc.). Toutefois, dans plus de la moitié des résumés, les auteurs utilisaient l'expression « Dans ce mémoire, il sera question de… », ce qui n'était pas pertinent pour nos analyse. Donc, une dernière vérification de la pertinence des termes du lexique fut effectuée manuellement afin d'augmenter la fiabilité de nos résultats d'analyse.

À la fin de notre processus de filtrage, nous éliminions 63,7 % de mots de notre lexique initial de 416 522 mots. Cela semble avoir du sens à nos yeux, car tout compte fait, les mots qui sont réellement significatifs et discriminants d'un corpus ne forment qu'un sous-ensemble, proportionnel à l'ampleur du corpus initial.

Statistiques du corpus	
Nombre total de documents	1240
Nombre total d'occurrences	416 522
Nombre total de formes	13 880
Total d'occurrences exclues	265 346
Moyenne d'occurrences par phrases	26,6
Moyenne d'occurrences par paragraphes	78,4
Moyenne d'occurrences par document	336

Tableau 2. Statistiques descriptives du corpus d'étude.

2.3.4. Vectorisation

Comme nous l'avons vu en première partie, l'objectif de la vectorisation est la création d'une matrice vectorielle pour représenter la distribution des termes discriminants à travers le corpus. Ainsi, en associant aux mots des vecteurs, il est possible de percevoir leur valeur statistique, et cela permet également d'accélérer considérablement le temps de traitement des algorithmes de fouille de textes, car ceux-ci vont comparer les vecteurs entre eux au lieu de comparer chacun des mots de chaque document.

Lors de notre recherche, la vectorisation se faisait automatiquement par le logiciel WordStat et les matrices vectorielles étaient produites en même temps que le processus de classification. Conséquemment, pour chaque classe produite, nous avions une matrice vectorielle, basée sur une pondération du facteur *TF* (*term frequency*), autrement dit, sur la fréquence des mots. Aussi, la mesure de similarité que nous avons utilisée s'appuie sur le cosinus, c'est-à-dire la similarité des angles entre les vecteurs (WordStat : 1998-2010, 100). Donc, en vulgarisant, les matrices vectorielles démontraient la présence des mots discriminants, en termes de fréquence absolue, au sein de chacun des documents.

2.3.5. Classification automatique et extraction des termes discriminants

Pour ce qu'il en est du processus de classification, nous avons utilisé la méthode de classification hiérarchique ascendante (CHA), aussi dite agglomérative.

> « Au départ, une matrice de similarité est calculée en utilisant des mesures de similarité existantes dans la littérature (indice d'équivalence, indice d'inclusion, information mutuelle). La matrice indique la force d'association entre deux unités textuelles à classer (en général des mots). Il en découle une liste de paires de mots, ordonnées par indice de similarité décroissante » (Ibekwe-SanJuan : 2007, 65).

Donc, l'approche ascendante crée d'abord une classe pour chaque document et à la première itération, l'algorithme regroupe en paire de deux les documents ayant les indices de similarité les plus élevés. Ensuite, lors des autres itérations de l'algorithme,

d'autres documents peuvent être joints aux classes existantes ou, encore, des classes peuvent être fusionnées entre elles. Finalement, il est possible d'effectuer les itérations de l'algorithme indéfiniment, jusqu'à l'obtention d'un nombre optimal de classes (Ibid).

Voici l'algorithme de classification hiérarchique ascendante :

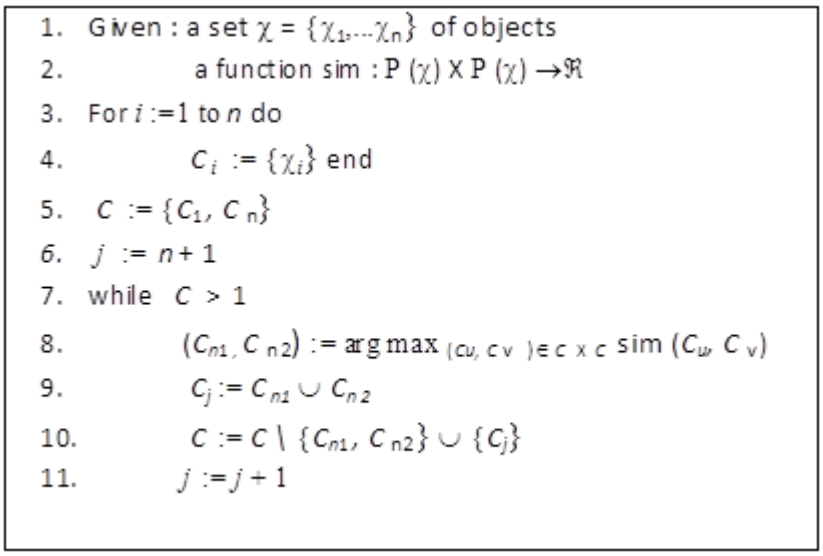


Figure 13. Algorithme de classification hiérarchique ascendante. Mannin and Schütze : 1999, 502.

Pouvant également s'écrire ainsi :

1. Commencer par la paire des classes κ et χ ayant la plus petite dissimilarité $d\ (i, j)$

2. Modifier $\mathcal{D}$ en supprimant les lignes et colonnes κ, χ et en ajoutant une nouvelle ligne et colonne $\kappa \cup \chi$;

3. Recalculer la dissimilarité entre $\kappa \cup \chi$ et chaque classe de la matrice. Ainsi, la dissimilarité entre la nouvelle classe $\mathcal{K} \cup x$ et une classe existante z est calculée de la manière suivante : $\big(d(z)\big), (k, x)) = \min\ (d(z), (x))$;

4. Retourner à (1) tant qu'il reste plus d'une classe, sinon s'arrêter.

Figure 14. Étapes de l'algorithme CHA. Ibekwe-SanJuan : 2007, 65.

Pour en revenir à notre processus, nous avons commencé en expérimentant une classification hiérarchique sur trois niveaux, en formant d'abord 5 classes ayant ces proportions : 1=405 documents - 32,6 %, 2=78 documents - 6,3 %, 3=372 documents - 30,0 %, 4=281 documents - 22,7 %, 5=104 documents - 8,4 %. Précisons que c'est en nous inspirant de l'approche nord-américaine de l'anthropologie (division en quatre grandes branches disciplinaires) que nous avons décidé de nous limiter qu'à un petit nombre de classes pour le premier niveau hiérarchique. Aussi, en ayant un premier niveau plus général avec ces grandes classes, nous cherchions à préciser, aux niveaux subséquents, les termes discriminants sur lesquels le processus classificatoire s'est opéré.

D'autre part, il est à noter que nous devions paramétrer notre système afin de lancer l'algorithme de classification (exemple de paramètres : l'utilisation des mots apparaissant dans plus de 40% des documents). Toutefois, la détermination des paramètres suivait une méthode dite d'essais-erreurs et nous devions donc tester et adapter les paramètres pour chacune des classifications. De façon générale, nous utilisions des paramètres moins contraignants au premier niveau pour obtenir des résultats plus généraux. Ensuite, les paramètres étaient davantage précisés au fur et à mesure de l'avancement de l'expérimentation. Donc, plus nous descendions dans les niveaux hiérarchiques, plus nous avions des ensembles de classes restreintes et plus nous pouvions établir des paramètres efficaces pour obtenir des résultats plus précis et extraire des mots davantage discriminants et représentatifs des classes formées.

Afin de pouvoir passer du premier niveau hiérarchique aux niveaux subséquents, nous avons procédé de la façon suivante : une fois que nous avions les classes de documents formées, nous exportions les fichiers de classification en format texte (tous les documents regroupés dans une classe sont ordonnés dans un même fichier texte). Ensuite, nous divisions automatiquement tous les documents du fichier source afin de reconstituer les classes de résumés, nous permettant de réitérer le processus de classification. Par exemple, à partir de la première classe de documents du premier niveau, nous avons divisé le fichier source pour obtenir les 405 résumés de cet ensemble, que nous avons

importés dans WordStat pour réinitialiser la classification au deuxième niveau, dans lequel nous obtenions 6 classes.

En ce qui concerne les autres expérimentations de classification, nous nous sommes concentrés sur nos différentes variables : départements, niveaux de scolarité, grappes de cinq ans et sous-disciplines. Cela nous a permis d'explorer plus en détail notre corpus. Aussi, en utilisant des filtres dans notre base de données SimStat, nous obtenions les ensembles désirés, par exemple, seulement les résumés de mémoires et de thèses en archéologie, que nous traitions ensuite dans WordStat pour effectuer la classification et l'extraction des termes discriminants. Remarquons que toutes les expérimentations se faisaient selon une classification hiérarchique agglomérative, mais nous restions qu'au premier niveau hiérarchique.

2.4. Visualisation des résultats

2.4.1. Analyse et préparation des fichiers à des fins de visualisation

La dernière étape de notre processus méthodologique est la visualisation des résultats. Puisque les techniques de fouille de textes permettent l'analyse de grands corpus, il peut être difficile d'obtenir une image d'ensemble des résultats. La suite Provalis et le module de WordStat, en particulier, offrent des outils intéressants pour visualiser les analyses, par exemple avec des cartes thermiques sur la fréquence des mots.

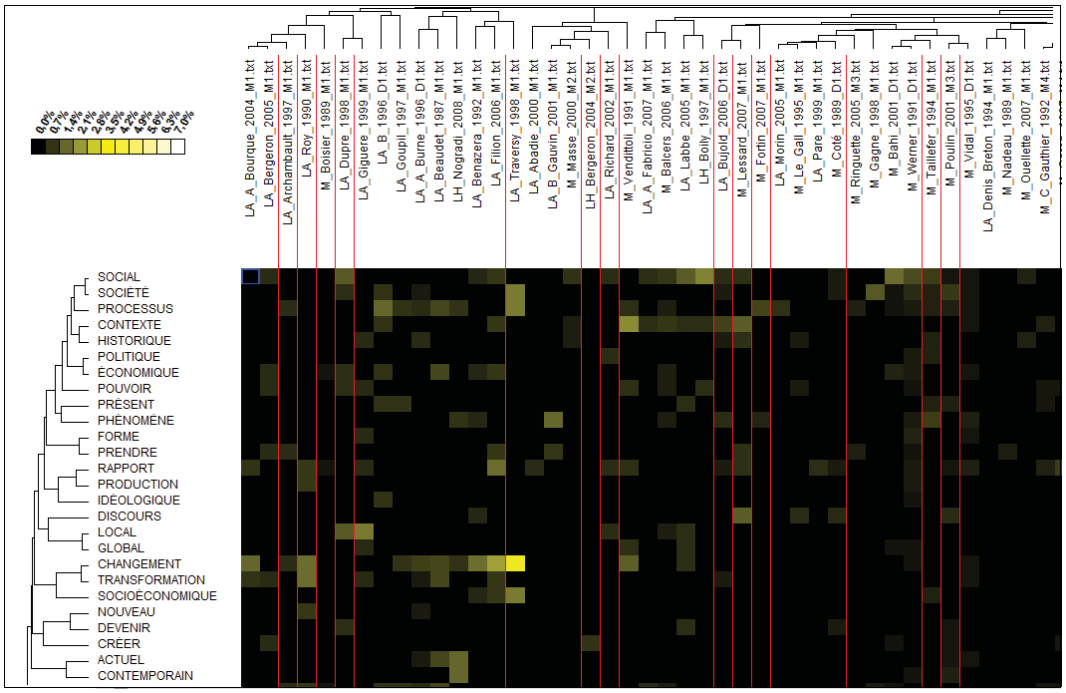


Figure 15. Carte thermique « *heat map* », sur la fréquence des mots.

Toutefois, comme nous travaillions à partir de matrices vectorielles, il devenait possible pour nous d'explorer la visualisation sous forme de graphes en trois dimensions, ouvrant la voie sur l'analyse réseau. Nous croyons en effet que les graphes illustrent bien les caractéristiques de liaisons thématiques entre les documents. En positionnant les résultats de classification dans un espace vectoriel, chaque document a un espace précis : ils sont représentés par les points, dits sommets ou nœuds « *nodes* », et ils possèdent des liens les reliant à d'autres documents. Ces liens, lorsque dirigés, sont dits arc « *vertices* », alors que lorsqu'ils sont non dirigés, ils se nomment arêtes « *edges* » (Paranyus Kin : 2011). En bref, l'avantage premier que nous trouvons à utiliser des graphes est que ceux-ci permettent de percevoir les résultats en tant que réseaux relationnels et sémantiques, menant à une visualisation structurelle et thématique des données.

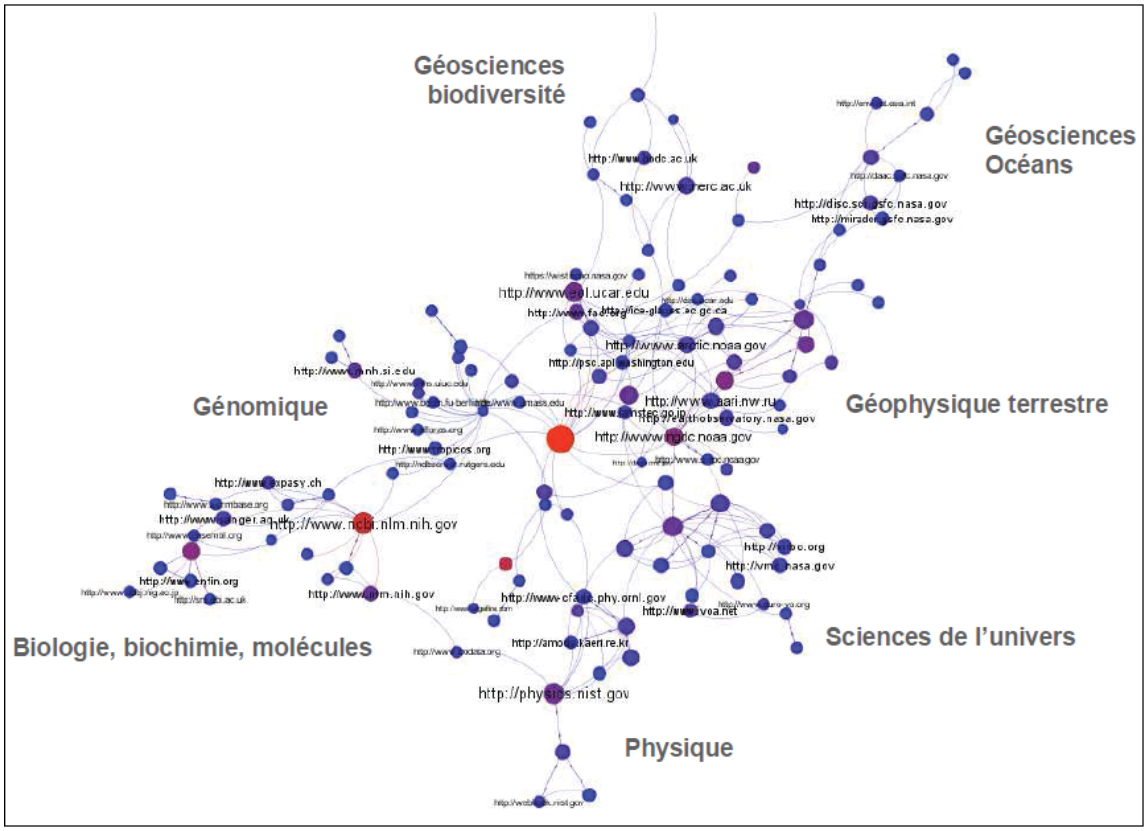


Figure 16. Le projet Web-Datarium, figure tirée de Ghitalla : 2010.

Pour réaliser la visualisation sous forme de graphe, nous avons utilisé le logiciel libre Gephi. Ce logiciel est une plateforme interactive de visualisation et d'exploration de réseaux et d'analyse textuelle, sous forme de graphes dynamiques. Ce logiciel est en perpétuel développement et il fut mis en œuvre, initialement, par le Médialab, de l'Université SciencesPo.

> « The goal is to help data analysts to make hypothesis, intuitively discover patterns, isolate structure singularities or faults during data sourcing. It is a complementary tool to traditional statistics, as visual thinking with interactive interfaces is now recognized to facilitate reasoning. This is a software for Exploratory Data Analysis, a paradigm appeared in the Visual Analytics field of research » (Gephi. <http://gephi.org/features/> (consulté en 2012)).

Différentes méthodes existent pour importer des fichiers dans Gephi et celles-ci exigent toutes quelques manipulations ou sous-opérations afin d'être exploitables. Voici les étapes méthodologiques que nous avons effectuées pour intégrer nos résultats de classification dans Gephi :

- À partir de WordStat, exportation du fichier d'analyse (matrice de données) en format Pajek (.net).
- Conversion du fichier Pajek en format GML (Graph Modeling Language) afin d'ajouter et de modifier les balises de lecture (Les balises *graph, node, edge, id, source, label, target* et *value* ont été ajoutées).
- Importation du fichier GML dans le logiciel Gephi

graph[							
node [	id	1	label	""	]		
node [	id	2	label	""	]		
node [	id	3	label	""	]		
…	…	…	…	…	…		
node [	id	679	label	""	]		
edge [	source	1	target	2	value	0.294	]
edge [	source	1	target	3	value	0.43	]
edge [	source	1	target	4	value	0.351	]
…	…	…	…	…	…	…	…
edge [	source	677	target	679	value	0.092	]
]							

Tableau 3. Balisage du fichier de classification en format GML.

2.4.2. Limite de cette méthode

Lors de l'exportation du fichier de classification de WordStat, le format Pajek n'a malheureusement pas conservé les classes résultant du processus de classification; en lisant le fichier, nous constatons que nous avons les nœuds (documents), les liens, la direction des liens (target) et leurs valeurs. Ainsi, il manque une balise pour les classes

(exemples du codage des balises de classes : <attibute id="modularity_classe" title="Modularity class" type="integer"> et <attvalue for="modularity_class" value="3">).

```
<?xml version="1.0" encoding="UTF-8"?>
<gexf xmlns="http://www.gexf.net/1.2draft" version="1.2" xmlns:viz="http://www.gexf.net/1.2draft/viz"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.gexf.net/1.2draft
http://www.gexf.net/1.2draft/gexf.xsd">
  <meta lastmodifieddate="2013-06-05">
    <creator>Gephi 0.8</creator>
    <description></description>
  </meta>
  <graph defaultedgetype="directed" timeformat="double" mode="dynamic">
    <nodes>
      <node id="1.0" label="M_A_Coloma_1998_D1.txt ">
        <attvalues></attvalues>
        <viz:size value="10.0"></viz:size>
        <viz:position x="201.41875" y="189.26877"></viz:position>
        <viz:color r="153" g="153" b="153"></viz:color>
      </node>
      <node id="2.0" label="M_A_Fournier_2000_M1.txt ">
        <attvalues></attvalues>
        <viz:size value="10.0"></viz:size>
        <viz:position x="-213.67982" y="-458.93875"></viz:position>
        <viz:color r="153" g="153" b="153"></viz:color>
      </node>

...     [Et ainsi pour tous les noeuds]

      <node id="679.0" label="M_Robert_2009_M43.txt ">
        <attvalues></attvalues>
        <viz:size value="10.0"></viz:size>
        <viz:position x="-322.8144" y="215.72433"></viz:position>
        <viz:color r="153" g="153" b="153"></viz:color>
      </node>
    </nodes>     [Dernier noeud]

    <edges>
      <edge source="1.0" target="2.0" weight="0.294">
        <attvalues>
          <attvalue for="weight" value="0.294"></attvalue>
        </attvalues>
      </edge>
      <edge source="1.0" target="3.0" weight="0.43">
        <attvalues>
          <attvalue for="weight" value="0.43"></attvalue>
        </attvalues>
      </edge>

...   [Et ainsi pour tous les liens]

 </edge>
      <edge source="45.0" target="268.0" weight="0.202">
        <attvalues>
          <attvalue for="weight" value="0.202"></attvalue>
        </attvalues>
      </edge>
                    [Fin du fichier]
```

Figure 17. Extrait du fichier source produit par WordStat, sans les balises des classes (« modularity »).

Pour contrer cette difficulté, nous avons utilisé l'algorithme de classification *Modularity class*, qui est intégré à Gephi. Ensuite, pour valider les résultats de cette deuxième classification, nous avons reconstitué les classes produites par Gephi dans notre base de données SimStat et nous avons ensuite extrait les termes discriminants pour chacune des classes. Les résultats de la classification et de l'extraction de termes obtenus dans Gephi nous ont semblé assez similaires à ceux obtenus auparavant dans WordStat,

mis à part la proportion des classes générées, qui était plus stable dans Gephi. Ainsi avons-nous décidé de poursuivre nos travaux de visualisation sous forme de graphes, en nous limitant toutefois qu'à une portion des résultats d'analyse ; les 679 résumés de mémoires et de thèses du département d'anthropologie de l'Université de Montréal. En troisième partie de ce mémoire, nous traitons plus particulièrement des limites rencontrées et des voies de solutions possibles.

2.4.3. Classification dans Gephi

L'algorithme de classification *Modularity class* se base sur la méthode « Louvain » (communauty detection mechanism), développée par Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte et Étienne Lefebvre (2008). Cette méthode regroupe en classes les nœuds qui présentent une densité de connexion (liens) plus élevée. Aussi, cet algorithme fonctionne en deux temps :

> « First, it looks for "small" communities by optimizing modularity in a local way. Second, it aggregates nodes of the same community and builds a new network whose nodes are the communities. These steps are repeated iteratively until a maximum of modularity is attained. The partition found after the first step typically consists of many communities of small sizes. At subsequent steps, larger and larger communities are found due to the aggregation mechanism. This process naturally leads to hierarchical decomposition of the network » (Blondel et al.: 2008, 2).

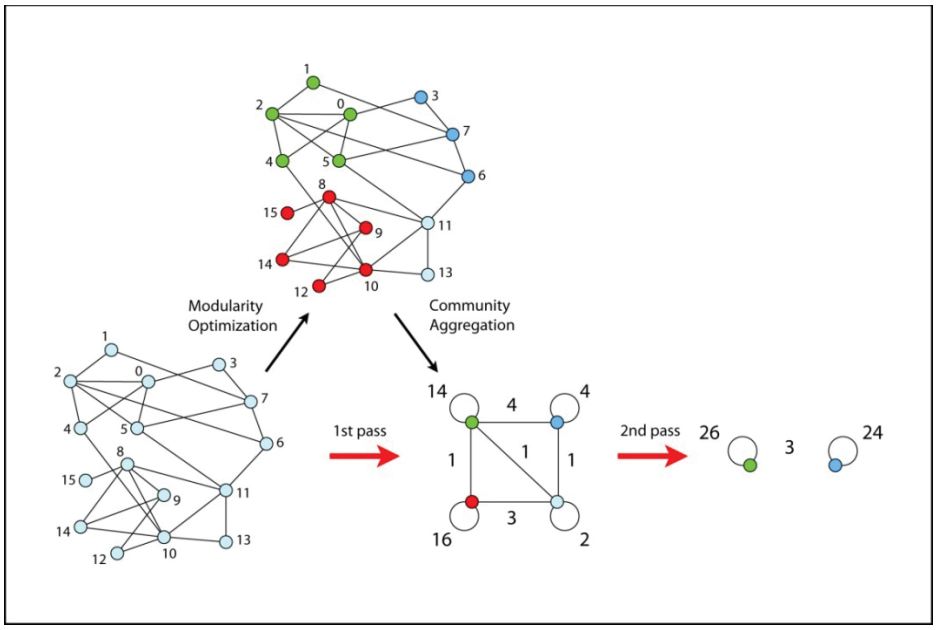


Figure 18. Visualisation des deux phases algorithmiques du *Modularity class*. Blondel et al. : 2008, 3.

Suite au processus de classification de Gephi, nous avions cinq classes de formées (tout comme initialement dans WordStat), selon les proportions suivantes : classe 1 = 156 documents - 23.00 %, classe 2 = 89 documents - 13.10 %, classe 3 = 112 documents - 16.48 %, classe 4 = 184 documents - 27.10 %, classe 5 = 138 documents – 20.32 %.

2.4.4. Processus de visualisation graphique

Lorsque les communautés ou classes étaient déterminées, nous pouvions explorer notre graphe et expérimenter les outils de visualisation de Gephi. Plusieurs algorithmes et protocoles de mise en forme (*layout*) sont proposés par le logiciel pour faire ressortir les structures et les particularités des réseaux et pour améliorer le rendu et la lisibilité des graphes (par exemple, ForceAtlas, Fruchterman-Reingold, Yifan Hu multilevel layout, OpenOrd layout, Circular layout, etc.). Nous avons choisi d'utiliser l'algorithme « Open Ord », car celui-ci fonctionne sur les graphes non dirigés et met l'emphase sur la distinction des classes (notre graphe ou réseau est non dirigé).

The algorithm is originally based on Frutcherman-Reingold and works with a fixed number of iterations controlled via a simulated annealing type schedule (liquid, expansion, cool-down, crunch, and simmer). Long edges are cut to allow clusters to separate. (Gephi. Tutorial Layouts).

Nous obtenions le graphe qui suit :

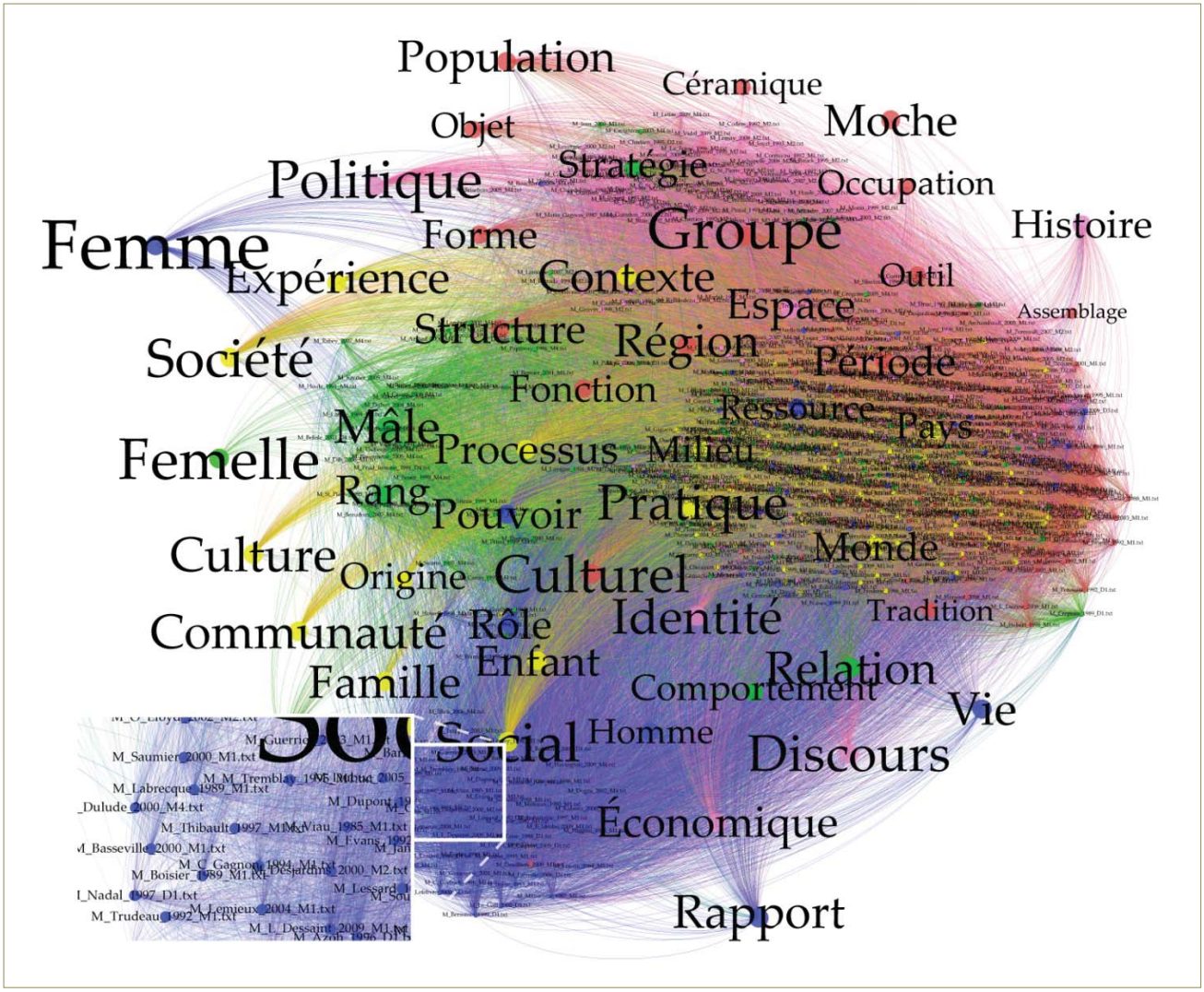


Figure 19. Graphe initial après l'application de l'algorithme *OpenOrd*. Classe 1 =rose, classe 2 = jaune, classe 3 = bleu, classe 4 = rouge, classe 5 = vert.

Statistiques graphiques	
Noeuds (Nodes)	729.000
Liens	33 538.000
Degré	46.005
Degré pondéré	21.062
Diamètre	3.00
Coefficient de classification	0.325

Tableau 4. Statistiques descriptives du graphe initial.

Selon les statistiques de ce graphe, nous remarquons le nombre élevé de liens. Le degré qui est élevé montre également qu'il y a beaucoup de liens entrants et sortants par nœud, ce qui signifie que les documents sont fortement interreliés. Le degré pondéré indique une distance maximale entre les sommets qui est modérée et quant au diamètre, celui-ci indique une distance maximale de 3.00 entre toutes les paires de nœuds. Le coefficient de classification, bien que modéré, montre que les résultats de classification sont significatifs (Paranyuskin : 2011).

Par ailleurs, afin d'améliorer le rendu de notre graphe et d'améliorer sa lisibilité, nous avons appliqué quelques autres algorithmes de mise en forme (*layouts*). Par exemple, nous avons réduit la superposition des nœuds (algorithme « Node overlapping »), ajusté la taille des étiquettes des noms de fichiers et modifié la taille des mots discriminants selon l'indice TF.IDF.

À cette étape, notre graphe possédait encore trop de liens et d'information, faisant en sorte qu'il était très difficile de le manipuler ; les systèmes comme « Adobe Reader », « Illustrator » et même « Gephi » devaient recalculer chaque nœud et chaque lien lors des manipulations (par exemple, le « zoom »). Ainsi, le téléchargement des données contenues dans le fichier de graphe était très lent, rendant difficile l'exploration thématique.

La solution que nous avons trouvée fut de diviser notre graphe en partitions ou « *subgraphs* », selon nos cinq classes. En obtenant cinq partitions, nous réduisions

considérablement le poids des fichiers, mais la répercussion inhérente à ce choix fut d'éliminer les liens interclasses (un mot ou un document pouvait avoir des liens avec d'autres mots ou d'autres documents provenant des autres classes).

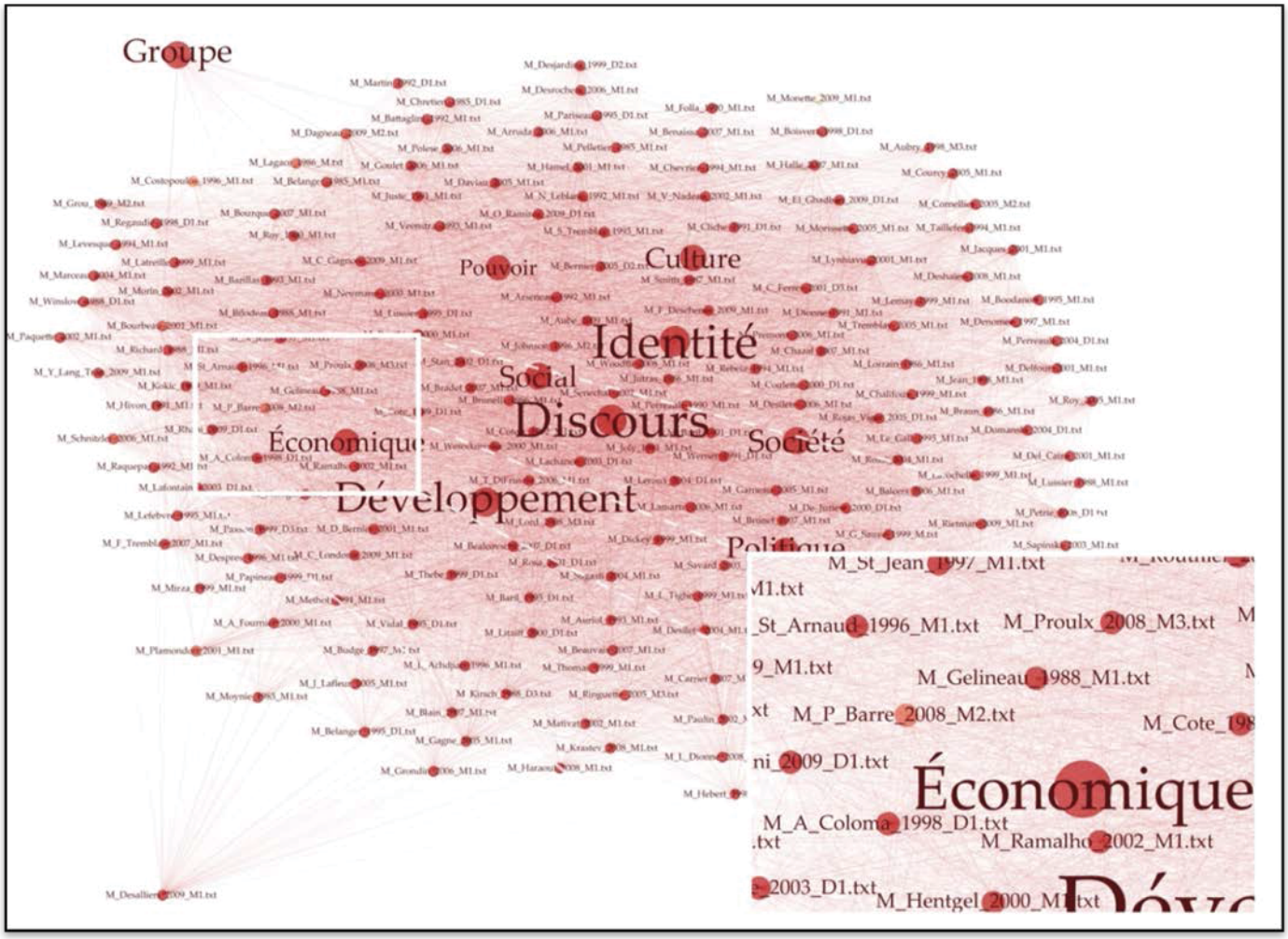


Figure 20. Exemple d'une partition graphique, classe 4, représente 27,1 % des 679 résumés, département d'anthropologie de l'UdeM, 1985 à 2009.

Troisième partie

3. Présentation des résultats

Au fil des sections précédentes, nous avons présenté notre cadre méthodologique, fondé sur la technique de classification hiérarchique ascendante. Aussi, l'objectif de notre recherche était de démontrer que la méthodologie adoptée peut permettre la classification automatique, le repérage et l'extraction de mots significatifs, menant ultimement à l'analyse et à la visualisation thématiques d'un domaine scientifique; celui de l'anthropologie québécoise.

L'organisation de ce chapitre se résume ainsi : dans un premier temps, nous présentons les stratégies et les outils employés afin de permettre la visualisation conceptuelle et graphique des résultats. Dans un deuxième temps, nous traitons de la question de la « navigation thématique » et des conséquences qui découlent de cette approche. En fait, la visualisation sous forme de graphe permet l'exploration des résultats tels qu'un réseau, avec les résumés et les mots-clés en guise de nœuds avec des liens qui les unissent. Cependant, la navigation est conduite par les intérêts des chercheurs et demeure donc, jusqu'à une certaine limite, subjective à ceux-ci. Ainsi proposerons-nous, dans un troisième temps, de présenter nos résultats de recherche et notre propre interprétation, en nous limitant toutefois sur deux exemples précis : la classification hiérarchique sur trois niveaux de l'ensemble du corpus (1 240 résumés), ainsi que la classification hiérarchique des 679 résumés du département d'anthropologie de l'Université de Montréal. Le choix de présenter en précision que deux des expérimentations réalisées réside dans une volonté de simplifier et de faciliter l'interprétation des résultats.

Nous complétons ce troisième chapitre par une discussion sur l'interprétation globale des résultats de recherche, ainsi que sur les approches théoriques de l'anthropologie. Puis, pour clore ce mémoire, nous traiterons des limites rencontrées dans ce projet et des pistes d'intérêts pour de futurs travaux.

3.1 Outils et stratégies de visualisation

Afin d'exposer les résultats d'analyse, nous avons décidé, en premier lieu, d'utiliser des cartes conceptuelles (logiciels *Microsoft Illustrator* et *Cmap* (Florida Institute for human & machine cognition)). Ainsi, nous présentons visuellement à plat (en deux dimensions) les structures inhérentes aux classifications, les mots obtenus de l'extraction, leur fréquence d'occurrence, ainsi que le nombre de documents constituant chacune des classes.

En deuxième lieu, nous proposons d'aborder les visualisations graphiques des résultats de la classification du département d'anthropologie de l'Université de Montréal. Aussi, tel que mentionné précédemment, nous présentons les cinq partitions produites ("*subgraphs*") à l'aide du logiciel *Gephi*, et qui reflètent les cinq classes obtenues de la classification hiérarchique ascendante, que nous avons effectuée avec l'algorithme « Modularity Class ».

3.2. Navigation thématique

Pour amorcer notre réflexion sur la navigation thématique, nous proposons de commencer par une citation, car les mots de l'auteur cité reflètent bien notre conception de la navigation thématique et soulève la problématique de la subjectivité :

> « Certains auteurs, comme René Thom, estiment que la modélisation graphique résout certains problèmes épistémologiques des sciences sociales : « By using the " distanciation " effet of geometric representation, to break the hermeneutic circle which has kept imprisoned so manys social science thinkers » (Thom : 1980). Cette mise à distance permettrait, confirme Petitot, de « briser le cercle herméneutique » (1985, p. 83) : une représentation serait alors conforme aux choses mêmes, parce qu'elle en exprime la forme, structurant selon les mêmes principes les deux « couches de l'Être », physique et sémiotique. Mais on peut voir dans ce vœu une illusion diagrammatique, car les représentations géométriques elles-mêmes doivent être interprétées, et, loin de le briser, elles renforcent et complexifient le cercle herméneutique » (Rastier : 2005).

Donc, à travers notre processus de classification et d'extraction des mots significatifs, nous ouvrons la voie à la thématisation afin de structurer, d'organiser et de situer les différents thèmes entre eux. Dans cette perspective, il devient possible d'identifier les classes qui partageaient des lexiques particuliers et d'explorer les associations. « [Nous concevons] une quelconque manifestation du thème comme la manifestation d'un autre, et de parcourir ainsi, de proche en proche, toute la série des variations du thème. » (Bremond : 1985, 419). Autrement dit, la navigation thématique peut se concevoir comme un processus de découverte et de parcours thématique à travers un réseau de documents. Cela implique donc une certaine subjectivité, car les chercheurs qui thématisent le font en fonction de leurs intérêts et des objectifs visés (Forest : 2006). Ainsi, d'autres personnes qui étudieraient notre lexique pourraient arriver à des interprétations différentes des nôtres, et ce, selon les parcours thématiques abordés.

> « 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 » (Prince : 1985).

Pour illustrer l'approche de navigation thématique, nous pouvons prendre le mot « Vie ». Selon nos résultats de classification, nous avons remarqué que cette thématique est employée principalement en ethnologie et que quelques résumés en bioanthropologie utilisent également ce terme (voir la partition graphique 3, à la page 93). De plus, selon les résultats obtenus de la classification et de l'extraction des mots significatifs pour cette classe, nous découvrons que plusieurs termes cooccurrent; du point de vue biologique, nous relevons les termes « humain », « groupe », « comportement », « épidémie », etc., alors que du point de vue ethnologique, ce sont les termes suivants que nous remarquons; « femme », « travail », « social », « société », « pouvoir », « relation », « communauté », …

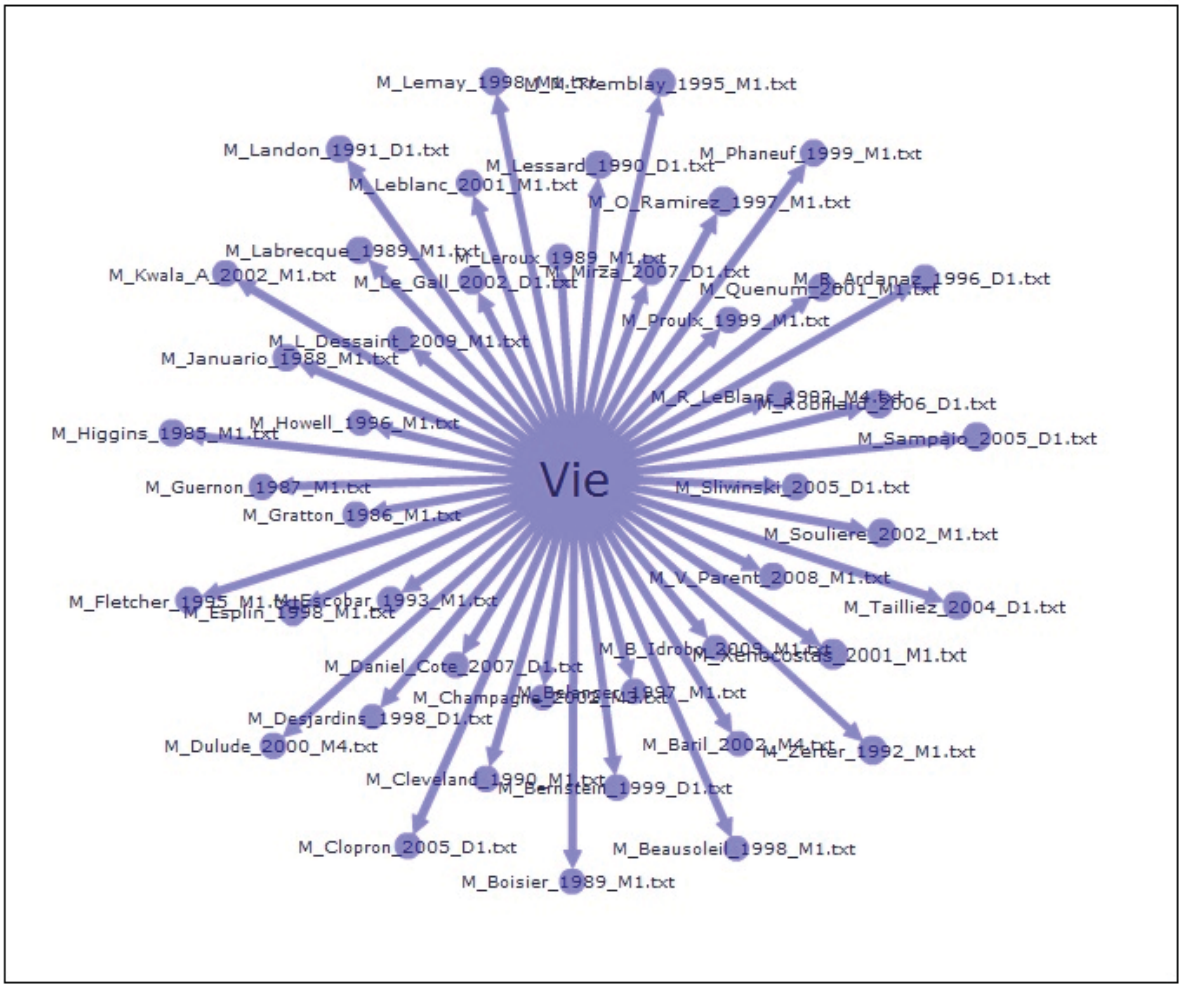


Figure 21. Représentation des documents ayant l'occurrence « vie ». Réalisée à l'aide de l'application web *Gexf Walker* (Jacomy : 2011).

Donc, par cet exemple, nous souhaitions montrer que plusieurs interprétations sont possibles et la navigation thématique appartient finalement aux utilisateurs, qui parcourent et interprètent les résultats selon leurs attentes et leurs besoins. « *La nature de ce que l'on pourrait appeler plus généralement l'étude thématique des textes est d'abord fonction de l'objectif visé* » (Martin : 1995, 18).

3.3. Première expérimentation : classification hiérarchique ascendante (CHA) à trois niveaux,sur l'ensemble des données

3.3.1. Vue d'ensemble de la classification

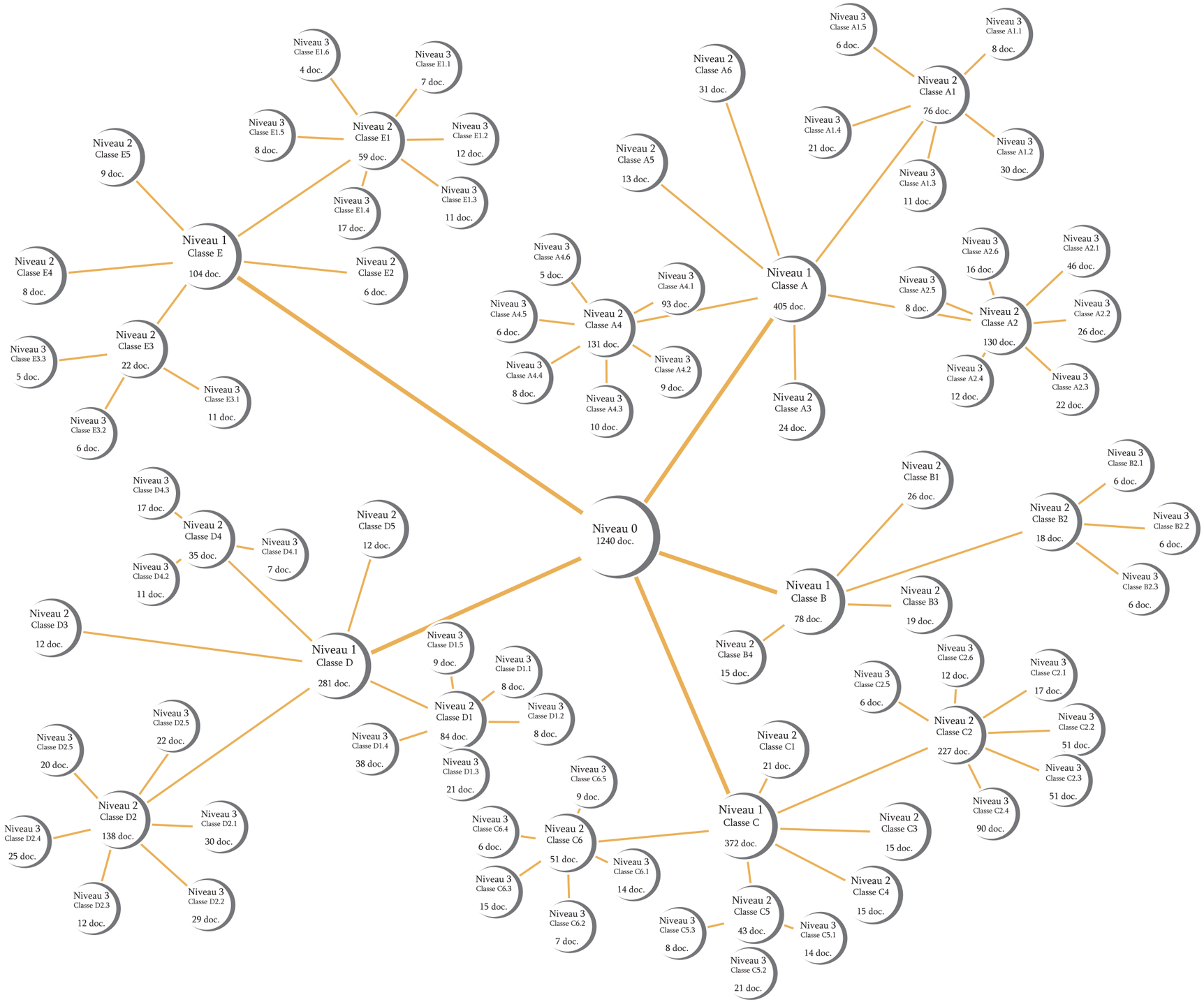


Figure 22. Représentation de la classification hiérarchique sur trois niveaux de hiérarchie, 679 résumés, département d'anthropologie de l'UdeM

3.3.2. CHA : 1er niveau hiérarchique

Figure 23. Représentation de la classification au premier niveau hiérarchique.

Observations :

- **Mots dont la fréquence d'occurrences est ≥ à 15% du corpus** : Femme, social, travail, développement, politique, économique, société, production, rapport, famille, discours, groupe, modèle, identité, culture, pratique, contexte, communauté.
- **Mots se répétant dans plusieurs classes** : Femme, social, société, maladie, culturel, pratique, contexte, groupe, communauté.

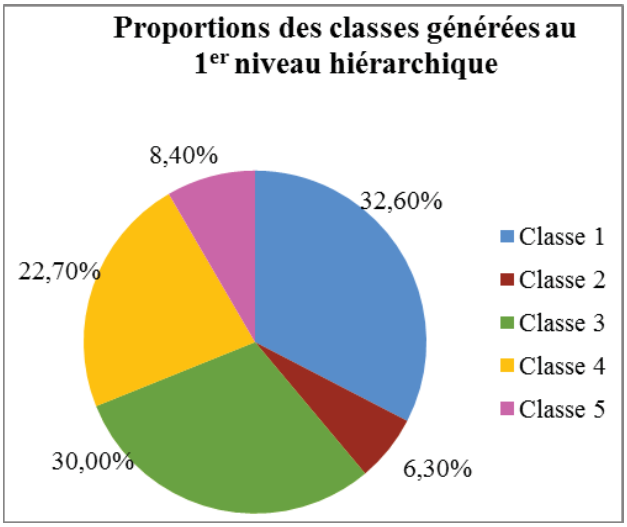


Figure 24. Proportions des classes obtenues au 1er niveau hiérarchique

Dans l'ensemble, nous remarquons que les termes extraits au premier niveau sont assez généraux et il est difficile de percevoir des patrons particuliers que suivraient les données. Bien sûr, nous observons qu'il y a une prépondérance de termes utilisés généralement en ethnologie ; « culturel », « social », « politique », « discours », « société », « économique », etc. Cela a du sens à nos yeux puisque l'ethnologie représente 72,7 % des données et que cette discipline reste pionnière pour le

développement de l'anthropologie au Québec. De plus, peu importe les sous-disciplines, nous croyons qu'il y a utilisation de termes communs et que ce sont alors les perspectives d'emploi des thématiques qui diffèrent. Par exemple, les termes « rapport », « groupe » et « modèle » peuvent très bien être employés dans les différents champs disciplinaires, selon des perspectives biologiques, historiques, culturelles, actuelles ou passées.

Par ailleurs, les termes qui se démarquent le plus, au niveau fréquentiel, sont « femme », « social », « travail » et « groupe ». Malheureusement, le terme « travail » doit être rejeté, car il est trop sujet à la polysémie ; les auteurs employaient trop souvent ce terme ainsi ; « Dans ce travail, il sera question de… ». Quant au mot « femme », celui-ci revêt une importance primordiale et demeure, dans l'ensemble, le mot le plus fréquent et significatif. Comme nous le verrons subséquemment, nous croyons que la prédominance de cette thématique peut résulter, en partie, des mouvements féministes au Québec, tant lors de la première moitié du 20^{e} siècle qu'aux alentours de 1960-1970 (Dumont : 1997, 2). Aussi pourrait-on ajouter, pour expliquer la force de la thématique, que les anthropologues auront étudié la femme sous tous ses angles, au présent comme au passé, culturellement comme biologiquement. Les termes qui suivent de par leur force fréquentielle et leur dispersion au sein du corpus sont « social », « société » et « groupe ». Cela va de soi à nos yeux et même si un chercheur étudiait qu'un seul individu, il le comparerait probablement à un groupe ou à l'environnement socioculturel, afin de contextualiser sa recherche.

Finalement, même si le schème thématique n'est pas très précis, nous remarquons les termes « période », « objet », « historique » et « contexte » (classe D), souvent employés en archéologie.

3.3.3. CHA, 2e et 3e niveaux, branche A

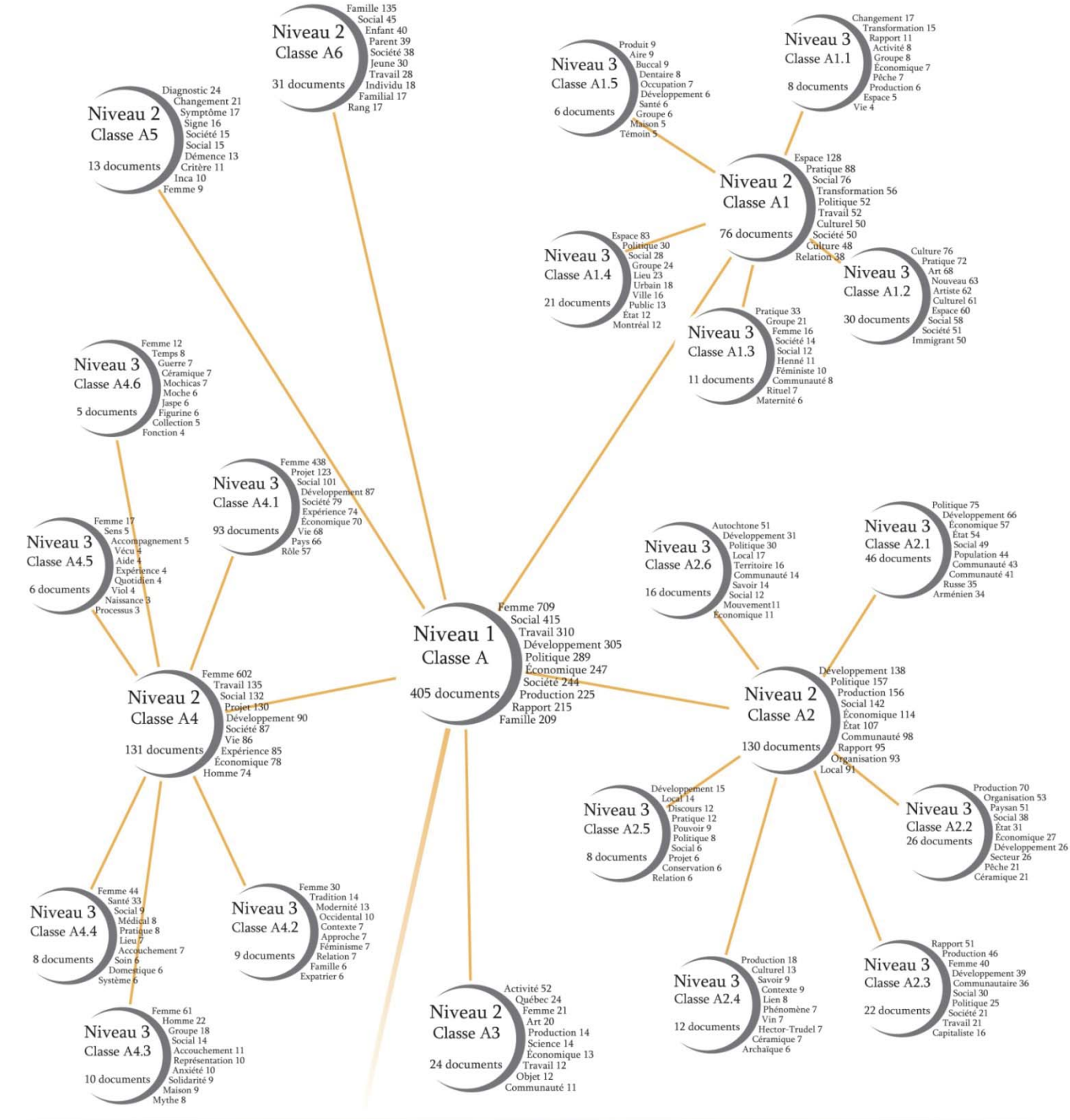


Figure 25. Première décomposition de la classification hiérarchique à trois niveaux. Branche A.

Observations :

- **Mots dont la fréquence d'occurrences est ≥ à 15% du corpus** : Femme, espace, pratique, social, culture, art, nouveau, artiste, culturel, politique, développement, production, économique, état, communauté, rapport, organisation, local, projet, société, expérience, vie, pays, famille.
- **Mots se répétant dans plusieurs classes** : Social, politique, travail, société, développement, production, économique, communauté.

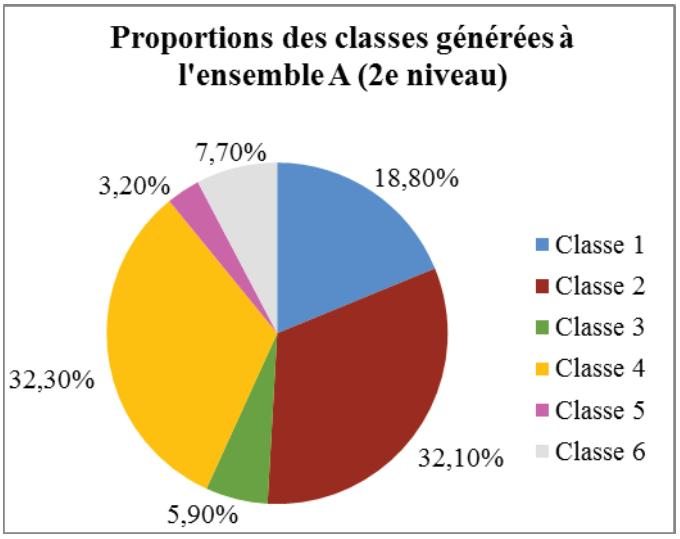


Figure 26. Proportions des classes obtenues au 2e niveau de la branche A.

Pour cette première branche de classification hiérarchique, nous remarquons que les thématiques semblent fortement orientées vers l'ethnologie et l'étude de phénomènes socioculturels, généralement contemporains. La thématique de la femme se retrouve précisément dans cette branche et elle semble souvent associée à des termes utilisés dans le domaine ethnologique, tels que « communauté », « politique » et « art ».

Nous remarquons néanmoins l'occurrence de ce terme dans une perspective archéologique, en association avec la culture « mochica », à la classe A4.2 (période précolombienne, au Pérou).

Nonobstant la thématique féminine, d'autres termes extraits nous semblent essentiels quant à la formation des différentes classes ; « famille », « pratique », « social », « espace », « développement », « politique », « production », « économique » et « projet ». De par leurs fortes occurrences fréquentielles, nous croyons qu'ils sont discriminants et représentatifs des grands schèmes thématiques. Puis, pour les mots extraits qui sont moins fréquents, tels qu' « autochtone », « expatrier », « capitalisme » et « maternité », ils amènent une précision thématique intéressante et sont facilement associables aux grands patrons thématiques qui ressortent.

Finalement, nous pouvons remarquer la présence de thématiques qui tendent vers des perspectives plus historiques, ancestrales et archéologiques. Ainsi retrouvons-nous aux classes A1.5 et A2.4 des occurrences de mots comme « céramique », « archaïque », « Hector-Trudel » (site archéologique), « mochica », « moche », « jaspe » (matériel d'orfèvrerie), « figurine », « collection », etc.

3.3.4. CHA, 2e et 3e niveaux, branche B

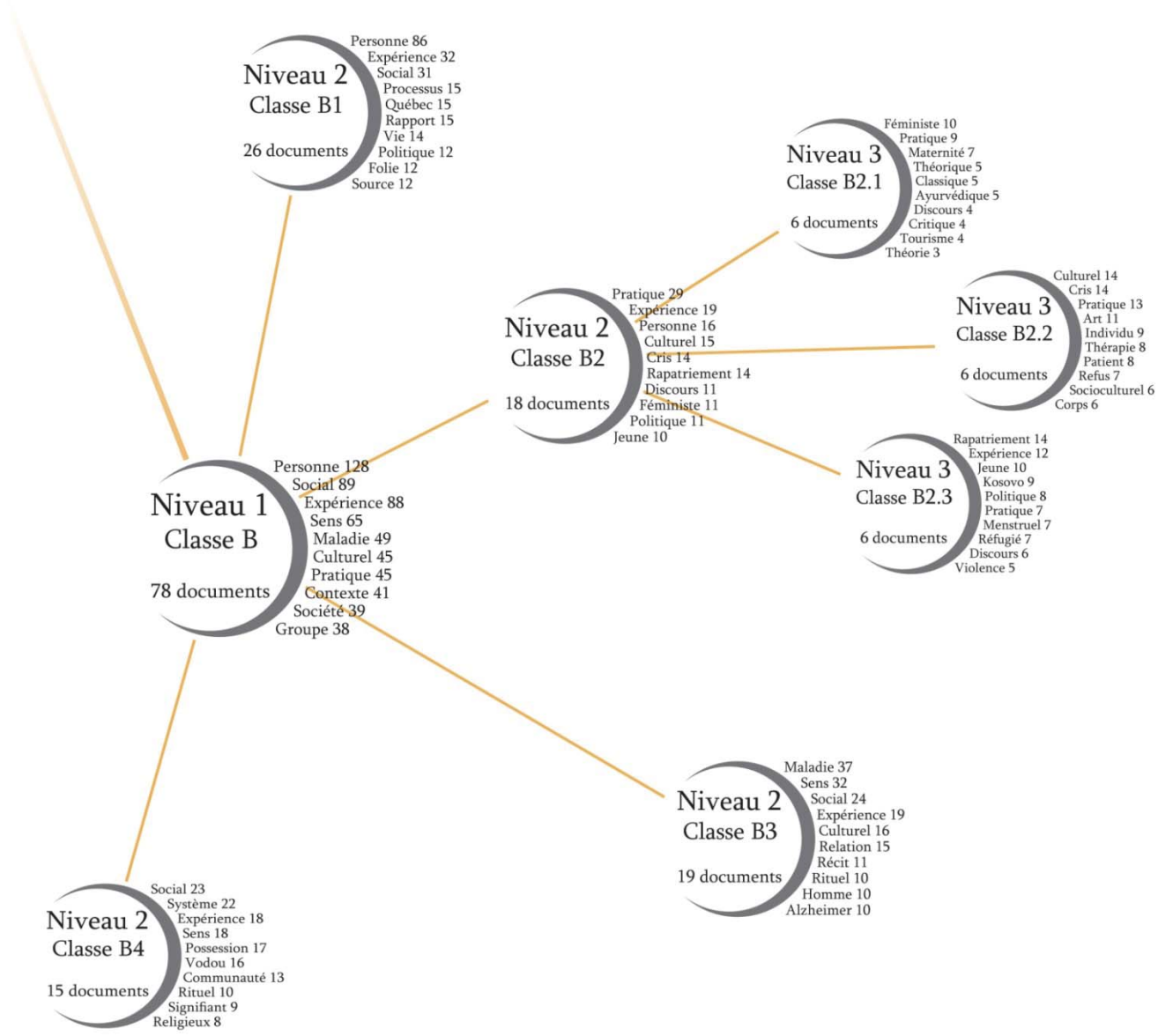


Figure 27. Deuxième décomposition de la classification hiérarchique à trois niveaux. Branche B.

Observations :

- **Mots dont la fréquence d'occurrences est ≥ à 15% du corpus** : Personne, expérience, social, processus, Québec, rapport, vie, politique, folie, source, pratique, Cris, culturel, rapatriement, maladie, sens, relation, système, possession, Vodou, communauté.
- **Mots se répétant dans plusieurs classes** : Social, discours, identité, culture, communauté, politique, phénomène, Montréal, système, culturel, représentation, modèle, ressource, femelle, individu, sens.

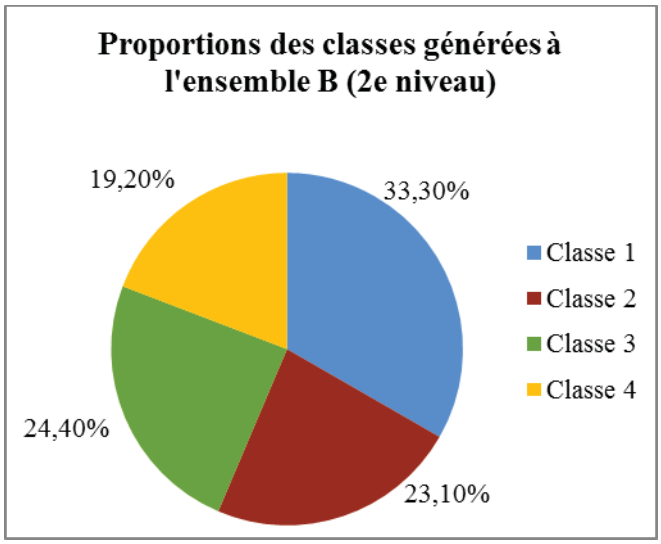


Figure 28. Proportions des classes obtenues au 2e niveau de la branche B.

Pour ce deuxième ensemble classificatoire, nous remarquons une fois de plus que les termes extraits semblent abordés des thématiques essentiellement ethnoculturelles. En effet, à partir de la classe centrale (classe B), nous avons l'impression que les thématiques s'orientent assez fortement vers le vécu, les sensations et l'expérientielle. Cela porte à croire que les résumés présentent des approches plus

herméneutiques des phénomènes étudiés. Ensuite, viennent d'autres thématiques pouvant facilement s'y associer ; « maladie », « rapatriement », « folie », « maternité », « possession », « rituel » et encore.

Par ailleurs, nous remarquons que les thématiques de la maladie et de la médecine prennent une certaine importance, principalement eu égard à ces termes extraits : « maladie », « Ayurvédique » (médecine traditionnelle indienne), « thérapie », « patient » et « Alzheimer » (classes B2.1, B2.2, B3). De plus, comme l'anthropologie médicale puise ses origines de l'anthropologie sociale et culturelle, étudiant largement la santé chez l'homme à travers des phénomènes tels la maladie et la guérison, nous croyons qu'il est normal que cette thématique apparaisse en cooccurrence avec des thèmes ethnographiques.

Finalement, nous remarquons la présence de certains termes faisant allusion à l'identité ou à l'individu : « personne », « individu », « Cris » (amérindien), « patient », « homme », « réfugié » et « jeune » (Classes B1, B2, B2.2, B2.3, B3). Nous croyons ainsi que les résumés de cet ensemble classificatoire abordent plus fortement l'individu en tant qu'objet d'étude, en comparaison avec les groupes, les communautés et les sociétés.

3.3.5. CHA, 2e et 3e niveaux, branche C

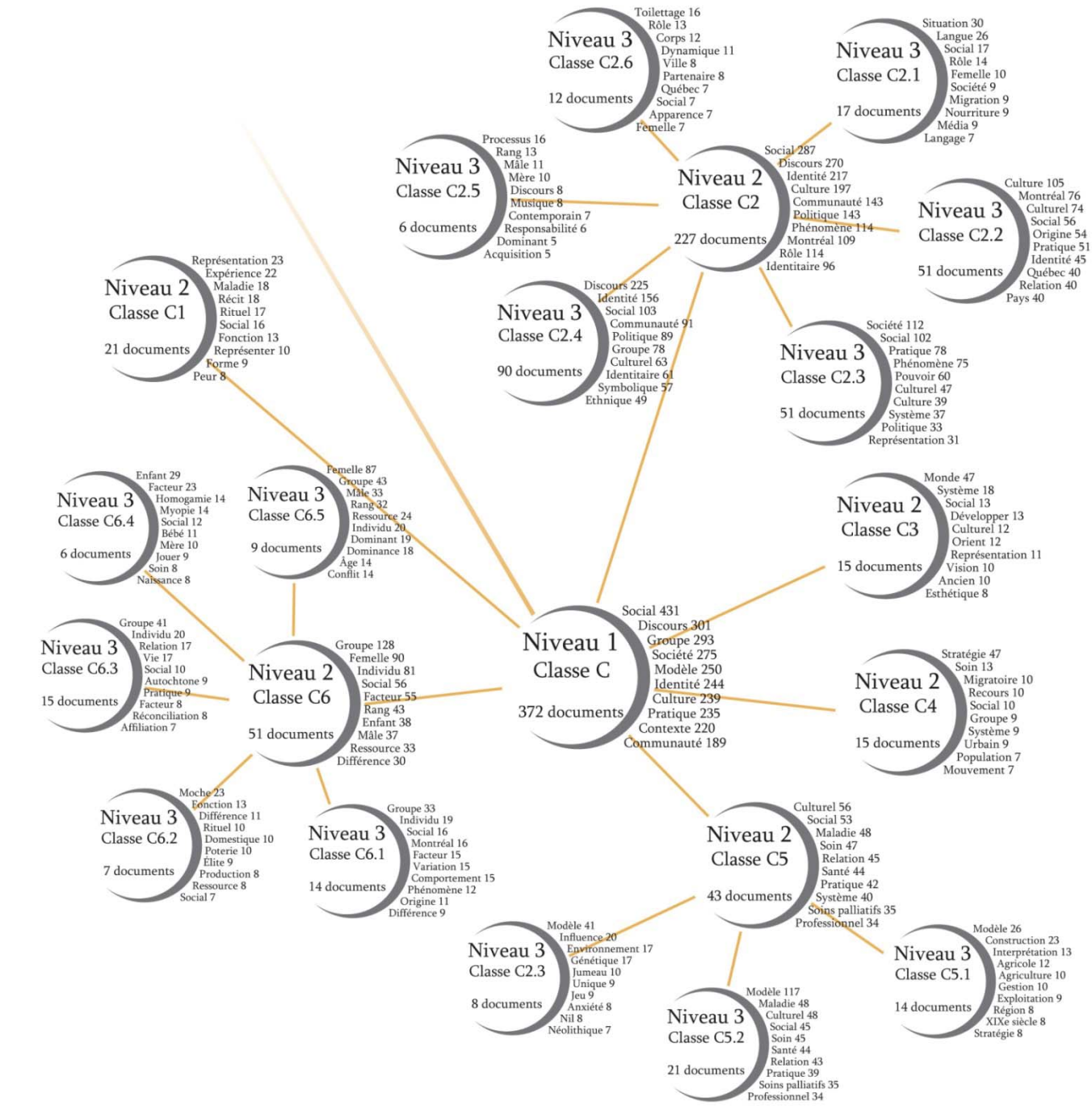


Figure 29. Troisième décomposition de la classification hiérarchique à trois niveaux. Branche C.

Observations :

- **Mots dont la fréquence d'occurrences est ≥ à 15% du corpus** : Social, discours, identité, culture, communauté, politique, phénomène, Montréal, rôle, identitaire, identité, groupe, culturel, symbolique, société, pratique, pouvoir, modèle, femelle, individu.
- **Mots se répétant dans plusieurs classes** : Social, discours, identité, culture, communauté, politique, phénomène, Montréal, système, culturel, représentation, modèle, ressource, femelle, maladie.

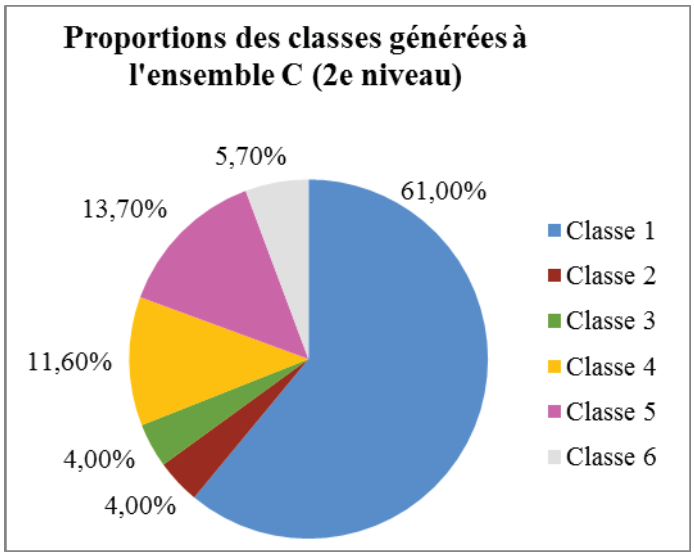


Figure 30. Proportions des classes obtenues au 2e niveau de la branche C.

Dans cet ensemble, il nous apparaît que la perspective ethnoculturelle est encore très forte. Toutefois, certains termes extraits se dissocient de cette tendance et s'orientent plutôt vers l'archéologie et l'étude d'époques passées. Nous remarquons en effet des termes comme « Moche », « néolithique », « poterie », « domestique » (poterie domestique), « fonction », « élite » (en archéologie, l'élite laisse plus de vestiges que les individus moins nantis), « construction » et « XIXe siècle ». Par les mots extraits, nous

discernons deux principales périodes d'étude : la préhistoire (culture mochica, au Néolithique (environ 9 000 ans av. J.-C. à 3000 ans av. J.-C.)) et la période historique.

Dans un autre ordre d'idées, nous remarquons l'intrusion de termes généralement employés en primatologie : « toilettage », « mâle » et « femelle », « dominant », « rang », « acquisition » (acquisition du rang de dominance), « ressource » (classes C2, C2.1, C2.4, C2.5, C2.6, C6 et C6.5), etc. Cela est intéressant compte tenu que la primatologie n'est enseignée qu'au département d'anthropologie de l'Université de Montréal. Nous pouvons aussi remarquer la proximité des classes de résumés en primatologie par rapport aux classes ayant des termes plus ethnologiques (par exemple, « discours », « identité » et « communauté »). Nous expliquons ces associations par le fait que les primatologues étudient des comportements chez les primates, tout comme le font finalement les ethnologues avec les individus et les groupes sociaux.

Par ailleurs, nous observons que la thématique de la maladie est toujours présente, principalement à la classe C5. Par contre, au lieu d'être associée à des mots exprimant l'individu ou la personne en tant qu'objet d'étude, ce sont plutôt des termes exprimant des structures, des systèmes, des pratiques médicales et, plus largement, des aspects socioculturels en anthropologie médicale.

Finalement, nous relevons certains termes extraits qui semblent ne pas concordés avec les autres occurrences, amenant une certaine confusion thématique. Par exemple, à la classe C2.5, plusieurs mots tendent vers la primatologie alors que l'apparition du terme « musique », au sein de cette même classe, laisse un peu perplexe. Autre exemple, nous relevons l'occurrence du mot « ville » à la classe C2.6, apparaissant en cooccurrence avec plusieurs mots issus de la primatologie.

3.3.6. CHA, 2e et 3e niveaux, branche D

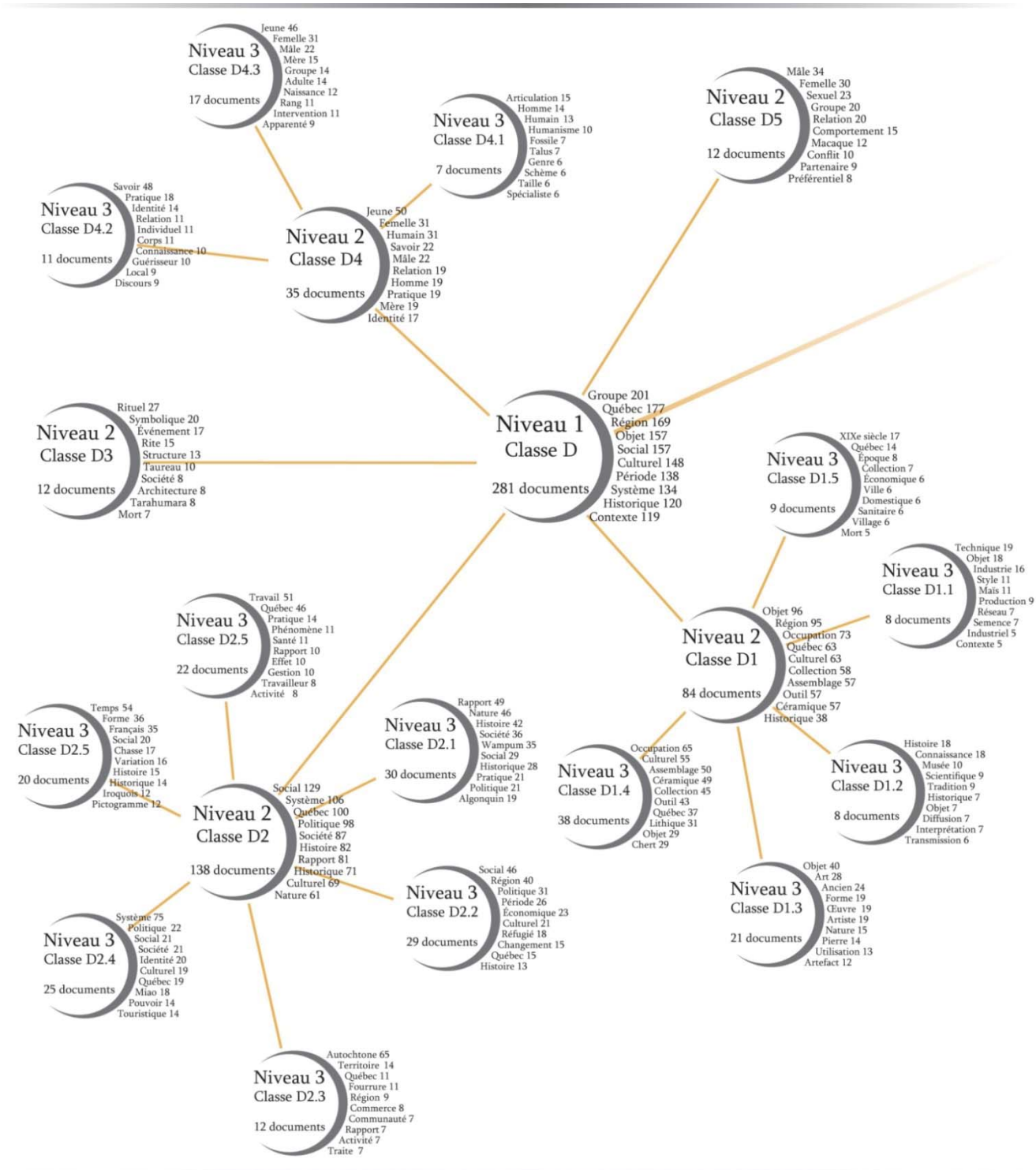


Figure 31. Quatrième décomposition de la classification hiérarchique à trois niveaux. Branche D.

Observations :

- **Mots dont la fréquence d'occurrences est ≥ à 15% du corpus** : Objet, région, occupation, Québec, culturel, collection, assemblage, outil, céramique, rapport, nature, histoire, autochtone, système, temps, travail, savoir, jeune.
- **Mots se répétant dans plusieurs classes** : Groupe, femelle, mâle, Québec, collection, économique, mort, objet, histoire, rapport, nature, société, social, politique, culturel, région.

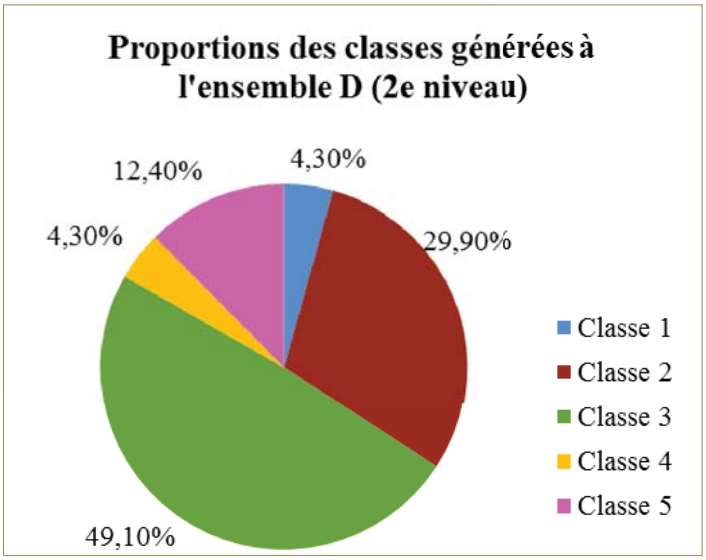


Figure 32. Proportions des classes obtenues au 2e niveau de la branche D.

Globalement, la branche classificatoire D possède des termes extraits convenant à la fois à l'ethnologie, à l'archéologie et à la bioanthropologie. En commençant avec les classes ethnologiques, qui abordent des phénomènes ou objets socioculturels, nous remarquons d'emblée des termes qui font allusion aux nations autochtones. Ainsi retrouvons-nous ces groupes : « Miao » (peuple asiatique vivant en région montagneuse), « algonquin », « iroquois », « tarahumara » (peuple vivant dans la région de la *Barranca del Cobre,* au Mexique) , « autochtone », et plusieurs termes

cooccurrents ou associatifs tels que « wampum » (ceinture traditionnelle algonquienne), « traite » (traite des fourrures), « chasse », « tourisme », « taureau » (rite du taureau chez les Tarahumaras), « pictogramme », etc. (classes D2.1, D2.3, D2.4, D2.5 et D3). Toujours à l'époque contemporaine et selon une perspective ethnologique, nous remarquons quelques autres thématiques, quoique moins fortes que l'étude autochtone ; d'abord l'art : « art », « artiste », « œuvre », « artefact », « musée » (classes D1.2 et D1.3), ensuite les savoirs : « connaissance », « savoir », « guérisseur », « discours » (classe D4.2), et le travail : « travailleurs », « gestion », « activité », « travail » (classe D2.6).

Du point de vue archéologique, les classes issues de D1 présentent des mots extraits plutôt discriminants et significatifs de cette discipline, bien que l'ethnologie s'y rattache aussi. Les termes que nous remarquons plus particulièrement sont : « céramique », « occupation », « collection », « lithique », « artefact », « chert », « assemblage », « outil », « pierre » et « XIXe siècle ». Cela nous porte à croire que l'archéologie, telle que véhiculée dans ces résumés, est plus historique que préhistorique.

Pour ce qu'il en est de la bioanthropologie, cette discipline est surtout perceptible, dans cet ensemble de classes, par la primatologie (classes D4.3 et D5) et l'anthropologie physique (classe D4.1). Pour la primatologie, il est question de termes tels que « mâle », « femelle », « sexuel » (les comportements sexuels des primates sont souvent étudiés en primatologie), « macaque », « partenaire », « conflit », « comportement », « rang », « apparenté » (l'apparentement), … Et du côté plus physique, nous retrouvons des occurrences de mots comme « articulation », « fossile », « taille » et « humain ».

Finalement, nous aimerions remarquer que certaines sous-disciplines se côtoient singulièrement au troisième niveau. C'est le cas de la classe D4 qui se décompose au troisième niveau en trois classes selon trois perspectives distinctes ; ethnologique, biologique et comportemental (les comportements étudiés en primatologie). Cette

situation de métissage scientifique se présente à nouveau à la classe D1, où nous percevons à la fois de l'archéologie et de l'ethnologie. Toutefois, dans ce dernier exemple, les mots extraits semblent plus facilement associables, selon nous parce que ces deux disciplines se rapprochent (étude des objets culturels, présents comme passés) et que l'archéologie, telle que reflétée dans les mots extraits, couvre la période historique, proche de l'époque contemporaine.

3.3.7. CHA, 2e et 3e niveaux, branche E

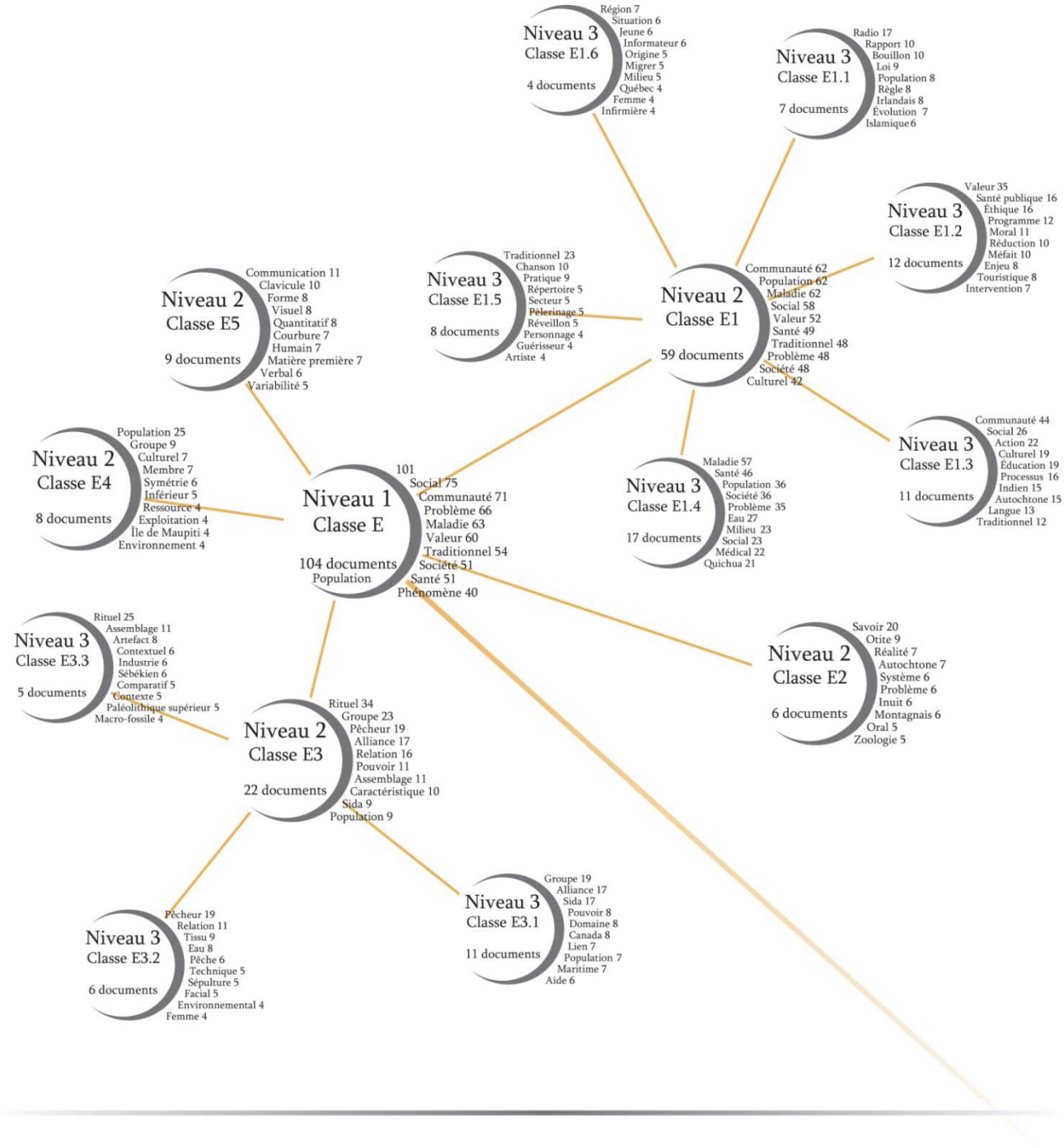


Figure 33. Cinquième décomposition de la classification hiérarchique à trois niveaux. Branche E.

Observations :

- **Mots dont la fréquence d'occurrences est ≥ à 15% du corpus** : Communauté, population, maladie, social, valeur, santé, traditionnel, problème, société, culturel, radio, santé publique, éthique, action, éducation, processus, eau, milieu, médical, Quichua, traditionnel, groupe, alliance, sida, relation, rituel, pêcheur.
- **Mots se répétant dans plusieurs classes** : Communauté, social, culturel, traditionnel, groupe, population, femme, environnement, autochtone.

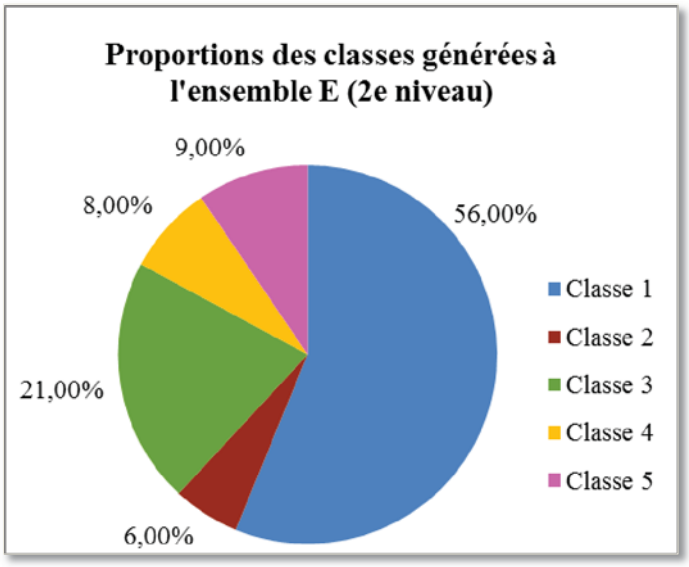


Figure 34. Proportions des classes obtenues au 2e niveau de la branche E.

Pour ce dernier ensemble de classes, les termes extraits tendent à démontrer, une fois de plus, une prédominance thématique de l'ethnologie. En effet, nous pouvons remarquer l'importance en terme fréquentiel de mots comme « communauté », « pouvoir », « social », « culturel », « santé » et « maladie ». Aussi, nous remarquons la

présence de l'étude autochtone (classes E1.3, E1.4 et %2), avec les thématiques suivantes : « Indien », « autochtone », « Quichua » (communautés culturelles d'Amérique du Sud), « Inuit » et « Montagnais ». À ces thématiques faisant référence à des groupes autochtones, nous observons des thèmes associés à la santé, à l'éducation et aux savoirs traditionnels (par exemple, la « zoologie »).

Par ailleurs, nous discernons aussi des thématiques généralement associées à l'ethnologie du Québec et au folklore. Par exemple, à l'intérieur des classes E1.5 et E1.6, nous retrouvons des mots dont « chanson », « traditionnel », « réveillon », « pratique », « Québec », « informateur » (personne interviewée, en ethnologie et folklore).

Eu égard à d'autres perspectives disciplinaires, nous relevons la présence de termes extraits faisant allusion à la bioanthropologie et à l'archéologie. Ainsi, les termes extraits : « clavicule », « courbure », « quantitatif », « forme », « humain », « variabilité », etc., orientent vers l'anthropologie physique et la variation biologique, alors que les mots et expressions suivants : « paléolithique supérieur », « macrofossile », « sébékien » (industrie lithique), « assemblage », « sépulture », font plutôt référence à l'archéologie, et plus particulièrement, à la préhistoire et la paléoanthropologie.

3.4. Deuxième expérimentation : visualisation sous forme graphique

Tel que précédemment mentionné, pour explorer nos résultats de recherche plus profondément, nous avons utilisé le logiciel *Gephi* afin de créer des partitions graphiques, représentant les cinq classes obtenues de la classification hiérarchique du département d'anthropologie de l'Université de Montréal (679 résumés).

De plus, puisque nous avons dû réitérer le processus de classification résultant à un problème d'exportation du fichier d'analyse dans le logiciel Gephi, nous nous retrouvons avec deux expérimentations de classification, l'une réalisée à l'aide de *WordStat* et l'autre avec le second logiciel, à l'aide de l'algorithme *Modularity class* (méthode Louvain).

Le tableau suivant montre les mots extraits à la suite des deux expérimentations. Nous ne cherchons pas à faire une comparaison détaillée des deux algorithmes de classification, mais nous remarquons néanmoins que le *Modularity class* a permis de générer des classes beaucoup plus proportionnelles que celles produites dans WordStat. Les termes extraits restent toutefois relativement semblables; pour plus de la moitié des occurrences de mots extraites, ceux-ci se retrouvent dans les deux expérimentations. Ainsi, il ne semble pas y avoir de grandes différences entre les deux systèmes quant à la pertinence des mots extraits. Les différences se jouent au niveau des proportions des classes générées, où Gephi présente plus de stabilité. En contrepartie, par rapport à WordStat, l'accessibilité aux paramètres de classification est limitée dans Gephi (par exemple, le choix *a priori* du nombre de classes à former dans le processus de classification n'est pas possible dans Gephi).

		Classification WordStat				Classification Gephi		
		Mots	Fréquences	TF-IDF		Mots	Fréquences	TF-IDF
Classe 1	109 documents	Groupe	151,0	46,1	156 documents	Culturel	186,0	57,6
		Région	150,0	49,5		Période	143,0	47,6
		Période	123,0	37,5		Groupe	139,0	53,4
		Culturel	105,0	43,5		Occupation	108,0	55,0
		Occupation	98,0	41,6		Population	90,0	49,2
		Outil	77,0	43,1		Outil	71,0	52,8
		Assemblage	70,0	37,3		Céramique	86,0	66,7
		Collection	62,0	34,7		Structure	85,0	48,2
		Population	60,0	33,6		Objet	84,0	42,0
		Québec	58,0	35,2		Moche	79,0	87,8
Classe 2	16 documents	Tradition	24,0	6,0	89 documents	Pratique	144,0	36,1
		Technique	16,0	4,0		Communauté	117,0	34,7
		Culture	14,0	7,6		Culture	113,0	41,8
		Contexte	12,0	3,6		Social	108,0	42,5
		Pratique	11,0	3,9		Société	71,0	33,5
		Histoire	9,0	4,5		Origine	69,0	32,6
		Société	9,0	4,5		Expérience	59,0	27,0
		Relation	8,0	4,0		Processus	55,0	26,0
		Identification	7,0	5,1		Enfant	51,0	41,0
		Pouvoir	7,0	2,7		Famille	42,0	29,2
Classe 3	458 documents	Social	629,0	185,9	112 documents	Femme	446,0	53,4
		Femme	478,0	316,3		Social	168,0	43,1
		Société	426,0	169,7		Expérience	117,0	56,3
		Politique	411,0	185,0		Vie	113,0	44,7
		Groupe	392,0	180,6		Société	112,0	46,6
		Culturel	390,0	167,6		Rôle	99,0	99,0
		Travail	367,0	155,5		Rapport	88,0	44,5
		Contexte	342,0	136,7		Pouvoir	87,0	87,0
		Pratique	330,0	175,0		Homme	80,0	34,1
		Discours	304,0	188,8		Famille	74,0	44,6
Classe 4	15 documents	Humain	46,0	4,5	184 documents	Politique	334,0	59,6
		Vie	20,0	8,0		Social	283,0	59,9
		Membre	20,0	9,5		Discours	243,0	89,2
		Articulation	14,0	12,3		Identité	187,0	87,0
		Social	10,0	4,0		Société	180,0	62,2
		Comportemer	10,0	8,8		Économique	135,0	57,5
		Santé	10,0	8,8		Développeme	133,0	78,8
		Clavicule	10,0	12,9		Groupe	127,0	63,7
		Épidémie	10,0	11,8		Pouvoir	114,0	46,5
		Population	9,0	3,6		Culture	109,0	57,2
Classe 5	81 documents	Femelle	135,0	82,0	138 documents	Femelle	197,0	134,1
		Groupe	126,0	41,2		Social	176,0	54,6
		Enfant	113,0	62,1		Relation	138,0	56,7
		Relation	110,0	43,8		Mâle	127,0	104,2
		Mâle	107,0	74,8		Rang	116,0	86,4
		Social	100,0	32,7		Modèle	120,0	64,9
		Famille	94,0	55,2		Société	100,0	58,7
		Rang	78,0	52,6		Jeune	79,0	70,1
		Société	70,0	42,5		Mère	73,0	63,1
		Mère	62,0	46,7		Famille	72,0	60,6

Tableau 5. Comparaison des résultats de classification de l'UdeM, produits par les algorithmes CAH de WordStat et Modularity class, de Gephi

3.4.1. Classe 1, département d'anthropologie de l'UdeM

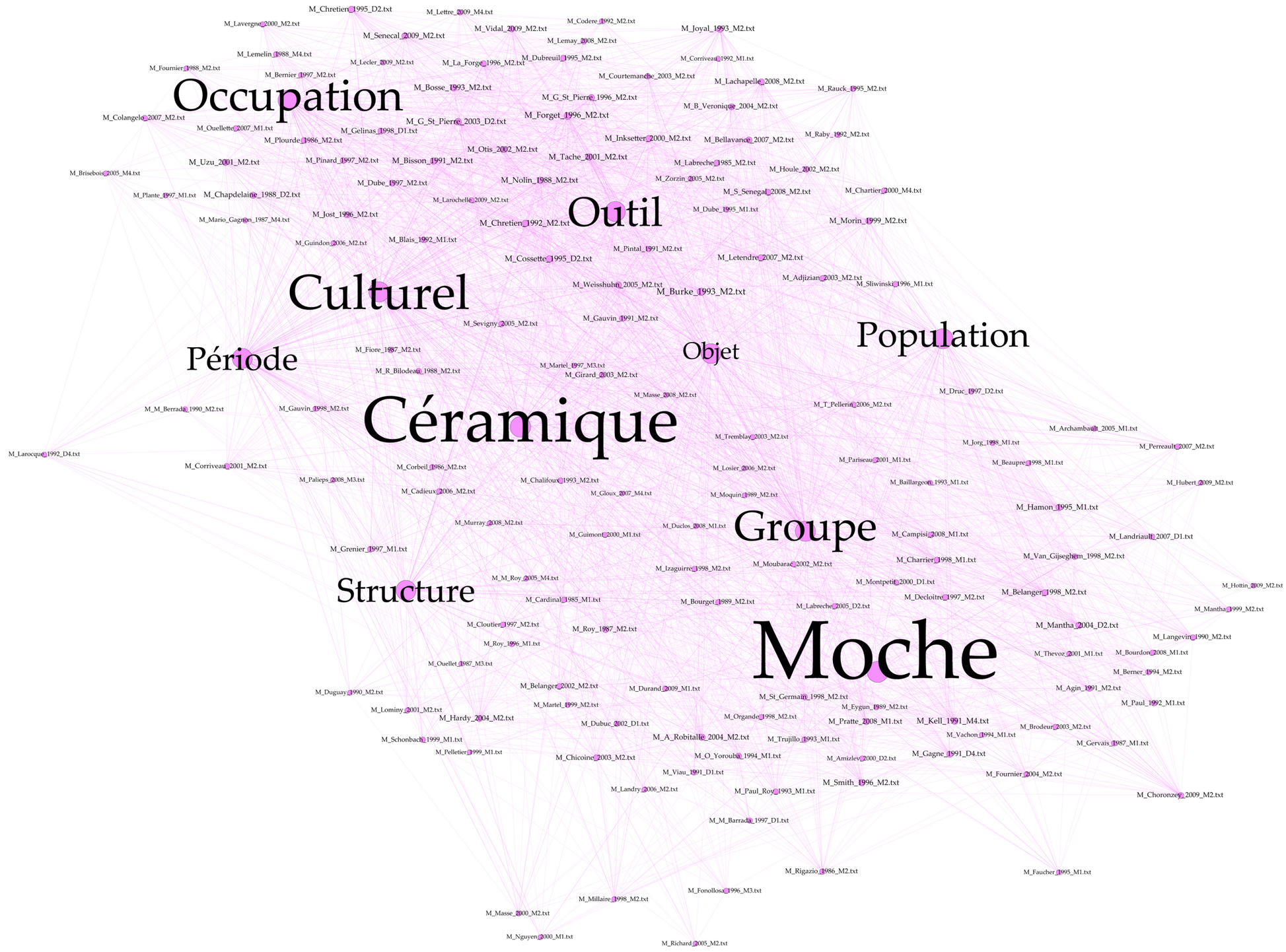


Figure 35. Partition graphique 1 (classe 1), département d'anthropologie de l'Université de Montréal.

Observations :

- Nombre de nœuds (résumés et mots-clés) : 166
- Nombre de liens : 2 723
- Occurrences de mots extraites et indices TF-IDF : « Moche » 87.7, « céramique » 66.7, « culturel » 57.6, « occupation » 55.0, « outil » 52.8, « population » 49.2, « structure » 48.2, « période » 47.6, « objet » 42.0.

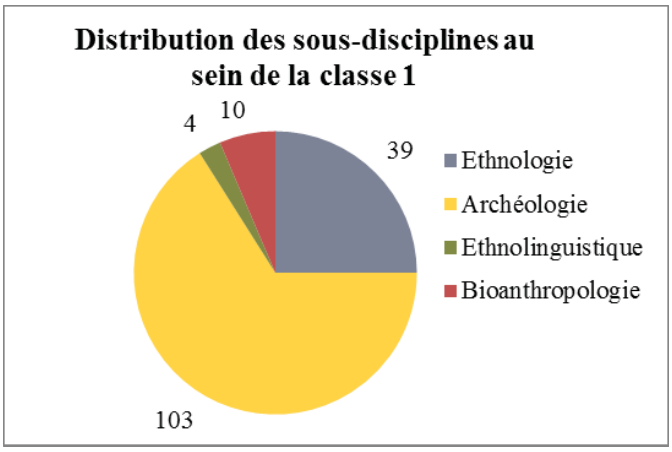


Figure 36. Répartition des résumés (valeurs réelles) en fonction des sous-disciplines, UdeM. (Catégorisation manuelle des sous-disciplines, effectuée par le département d'anthropologie).

Il semble évident que les résumés regroupés dans cette classe portent largement sur des thématiques archéologiques. De plus, l'un des termes extraits nous donne un indice quant à la zone culturelle qui prédomine, tout autant qu'à l'époque couverte; le mot « moche » révèle l'étude de la culture précolombienne (entre l'an 100 et l'an 700 apr. J.-C.), établie le long de la côte nord péruvienne. Le deuxième terme important en termes d'indice TF-IDF est « céramique », et il témoigne en fait de l'importance de ce témoin archéologique, fréquemment trouvé en grande quantité sur des sites de fouille. Aussi, puisque cet objet d'étude permet de dater des sites à moindres coûts par rapport à d'autres techniques comme la datation au carbone 14 ou la spectroscopie du bois, nous croyons qu'il est normal que ce terme soit fort et significatif. Le mot « occupation » est également important; lors de fouilles archéologiques, les chercheurs visent normalement

à comprendre l'occupation spatiale des sites, permettant de contextualiser leur recherche. Quant aux termes « outil » et « objet », faut-il s'étonner de les voir? Comme les archéologues accordent une attention particulière à la culture matérielle pour comprendre l'évolution des groupes humains, il va de soit de retrouver ces mots, qui expriment finalement l'objet d'étude en tant que tel. Pour les mots suivants, « population » et « groupe », ils représentent en quelque sorte les sujets d'études, c'est-à-dire que les chercheurs en archéologie étudient des objets, des témoins archéologiques, des sites, des monuments, et encore, afin de dresser des portraits de populations et de groupes lithiques. Nous remarquons ensuite le terme « structure », qui, à nos yeux, est assez intéressant et finalement, le dernier mot extrait, « période », est tout à fait pertinent pour cette classe, qui regroupe des résumés majoritairement archéologiques.

Dans un autre ordre d'idée, nous avons identifié l'archéologie comme étant la discipline au cœur de cette classe. Pour les autres disciplines qui s'y insèrent, nous constatons que ce sont surtout des résumés en ethnologie, mais nous retrouvons aussi quelques résumés en bioanthropologie et trois en ethnolinguistique. La liaison de ces documents aux résumés archéologiques est plutôt difficile à établir aux premiers abords, mais elle résulte nécessairement de la distribution des mots dans les textes.

Pour ce qu'il en est de la forme visuelle que prend la classe, nous remarquons que le graphe est assez allongé, avec un ensemble de documents plutôt rapproché dans le haut, alors que dans le bas du graphe, les nœuds sont plus distants les uns des autres. Si nous avions poursuivi le processus de classification, nous aurions sûrement obtenu deux classes pour cet ensemble de résumés.

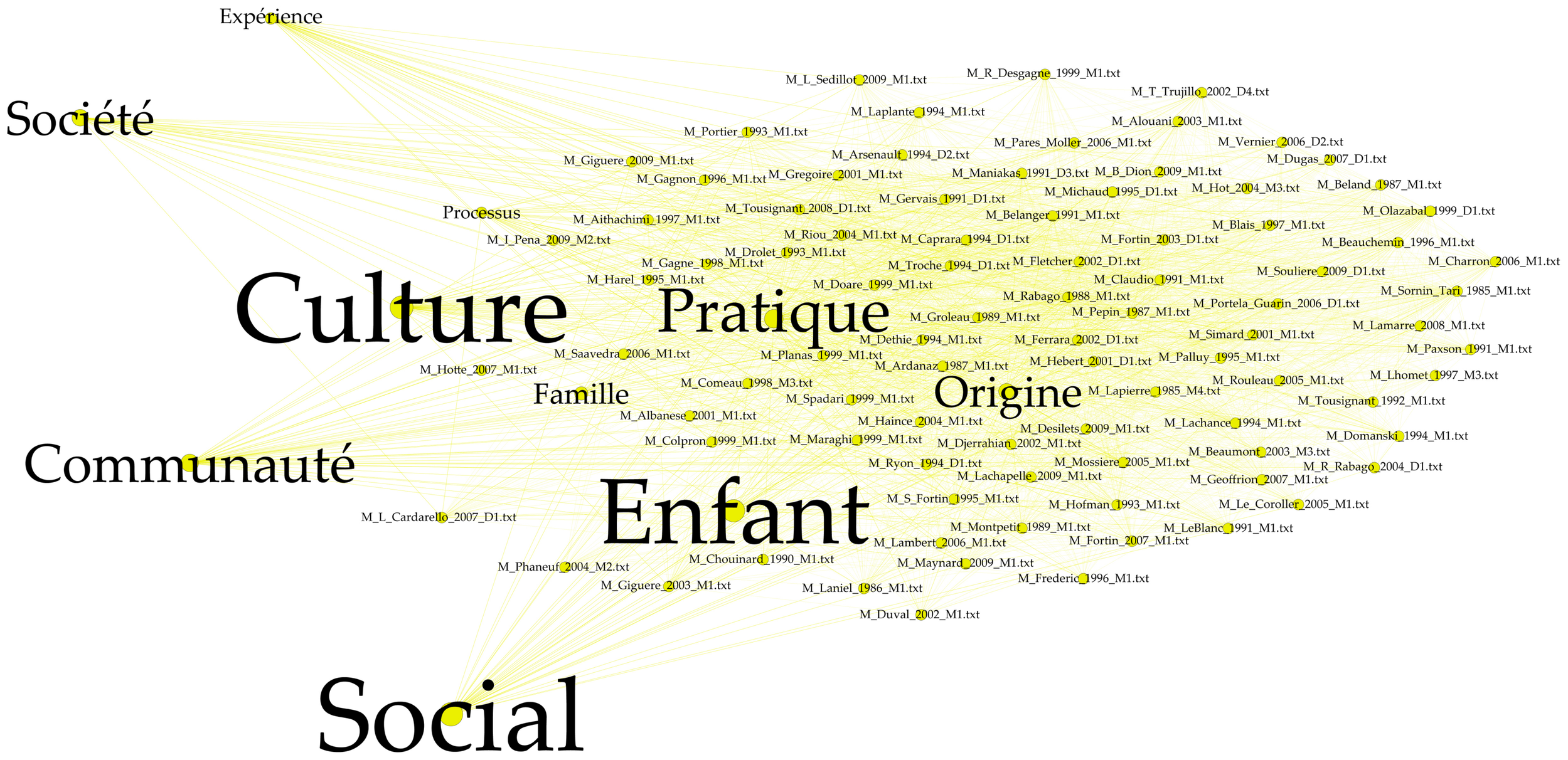


Figure 37. Partition graphique 2 (classe 2), département d'anthropologie de l'Université de Montréal.

Observations :

- Nombre de nœuds (résumés et mots-clés) : 99
- Nombre de liens : 2 239
- Occurrences de mots extraites et indices TF-IDF : « Social » 42.5, « culture » 41.8, « enfant » 41.0, « pratique » 36.1 « communauté » 34.7, « société » 33.5, « origine » 32.6, « famille » 29.2, « expérience» 27.0, « processus» 26.0.

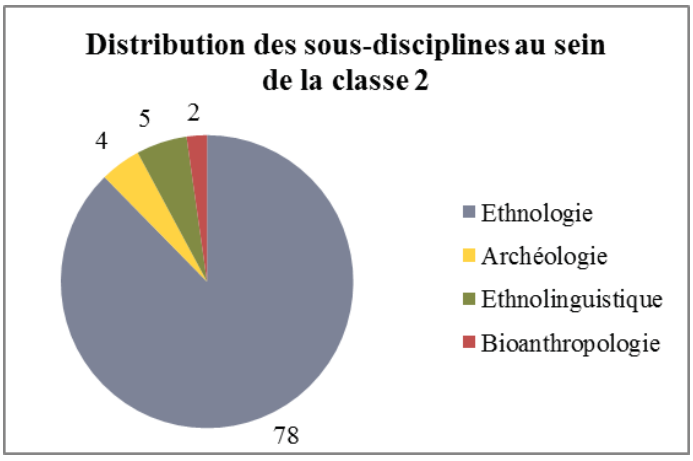


Figure 38. Répartition des résumés (valeurs réelles) en fonction des sous-disciplines, UdeM.

Pour cette deuxième classe, la tendance thématique est fortement axée vers l'ethnologie et les termes extraits s'orientent vers l'étude des phénomènes socioculturels. L'occurrence de mot la plus significative et discriminante de l'ensemble est « social », suivie de près par les mots « culture » et « enfant ». Nous croyons normal de retrouver en force les deux premiers mots puisque dès ses débuts, l'anthropologie s'intéressait avant tout aux aspects socioculturels des sociétés. Ainsi trouvons-nous intéressant d'obtenir le mot « enfant » en aussi grande force et nous comprenons qu'il est, pour cette classe, une thématique essentielle. En termes d'indice TF-IDF, viennent

ensuite les mots « pratique », « communauté », « société » et « origine ». Puis, les derniers termes ; « famille », « expérience » et « processus ». En bref, même si plusieurs occurrences de mots sont versatiles et pertinentes pour toutes les sous-disciplines anthropologiques, le schème thématique qui domine est indéniablement ethnologique.

En examinant la partition graphique et telle qu'observé par les mots extraits, nous constatons que la majorité des résumés se situent en ethnologie, mais nous remarquons néanmoins l'intrusion des quelques résumés issus des autres sous-disciplines ; il y a quatre résumés en archéologie, quatre autres résumés en ethnolinguistique et deux en bioanthropologie. Quant à la forme graphique que cette partition prend, les résumés sont assez proches les uns des autres, ne formant qu'un seul groupe assez homogène.

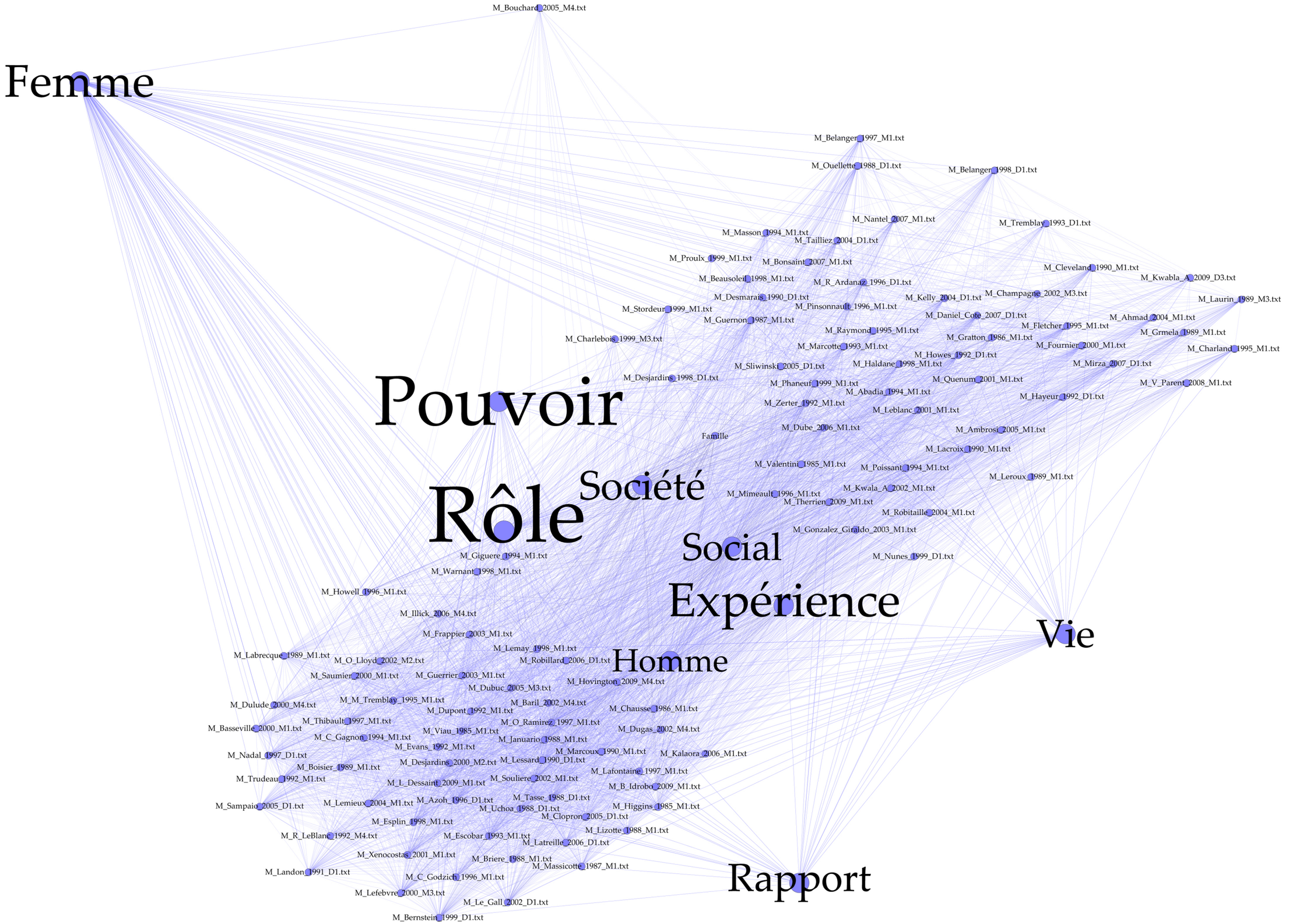


Figure 39. Partition graphique 3 (classe 3), département d'anthropologie de l'Université de Montréal.

Observations :

- Nombre de nœuds (résumés et mots-clés) : 122
- Nombre de liens : 4 033
- Occurrences de mots extraites et indices TF-IDF : « Rôle » 99.0, « pouvoir » 87.0, « expérience » 56.3, « femme » 53.4, « société » 46.6, « vie » 44.7, « famille » 44.6, « rapport » 44.5, « social » 43.1, « homme» 34.1.

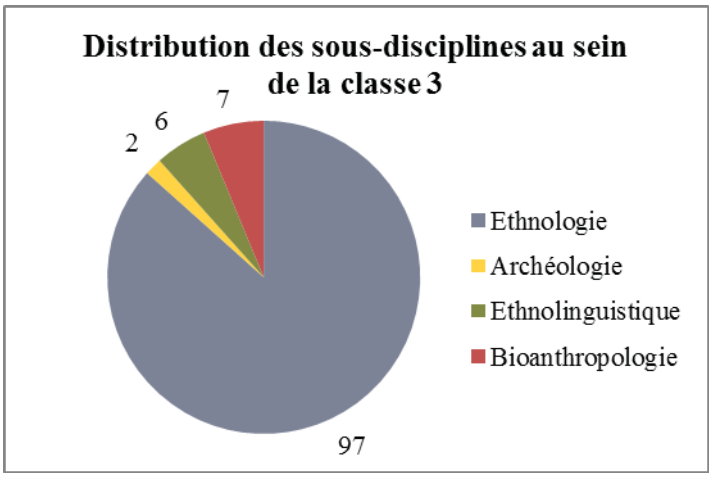


Figure 40. Répartition des résumés (valeurs réelles) en fonction des sous-disciplines, UdeM.

Les termes extraits de cette classe sont une fois de plus orientés vers l'ethnologie et l'étude des phénomènes socioculturels. Les mots qui présentent les plus grands indices TF-IDF sont « rôle », « pouvoir », « expérience » et « femme ». Ensuite viennent les termes suivants ; « société », « vie », « famille » et « social ». Puis, ayant un indice moins élevé que les autres, le terme « homme » vient clore l'ensemble des mots extraits pour cette classe. En somme, les thématiques qui ressortent tendent vers les relations et rôles entre femmes et hommes, les rôles sociaux, de pouvoir, familiaux. L'expérientiel semble également assez important pour cet ensemble de résumés.

Au niveau de la structure visuelle que prend cette partition graphique, nous remarquons la formation de deux groupes assez distants. Tout comme pour la première partition, cela nous porte à croire que nous aurions pu poursuivre l'algorithme de classification et obtenir deux classes distinctes à partir de cet ensemble de résumés. Toutefois, nous pouvons remarquer que puisque nous avons divisé notre graphe initial en cinq partitions, nous ne pouvons voir l'impact des liens interclasses (mot ou document lié à d'autres mots ou documents appartenant à d'autres classes), qui ont pu jouer un rôle important sur les structures visuelles des graphes.

Puis, pour ce qu'il en est de la distribution disciplinaire, mis à part l'ethnologie, nous remarquons l'intrusion de quelques résumés issus de d'autres sous-disciplines : nous avons cinq résumés en ethnolinguistique, cinq autres en bioanthropologie, et deux résumés en archéologie.

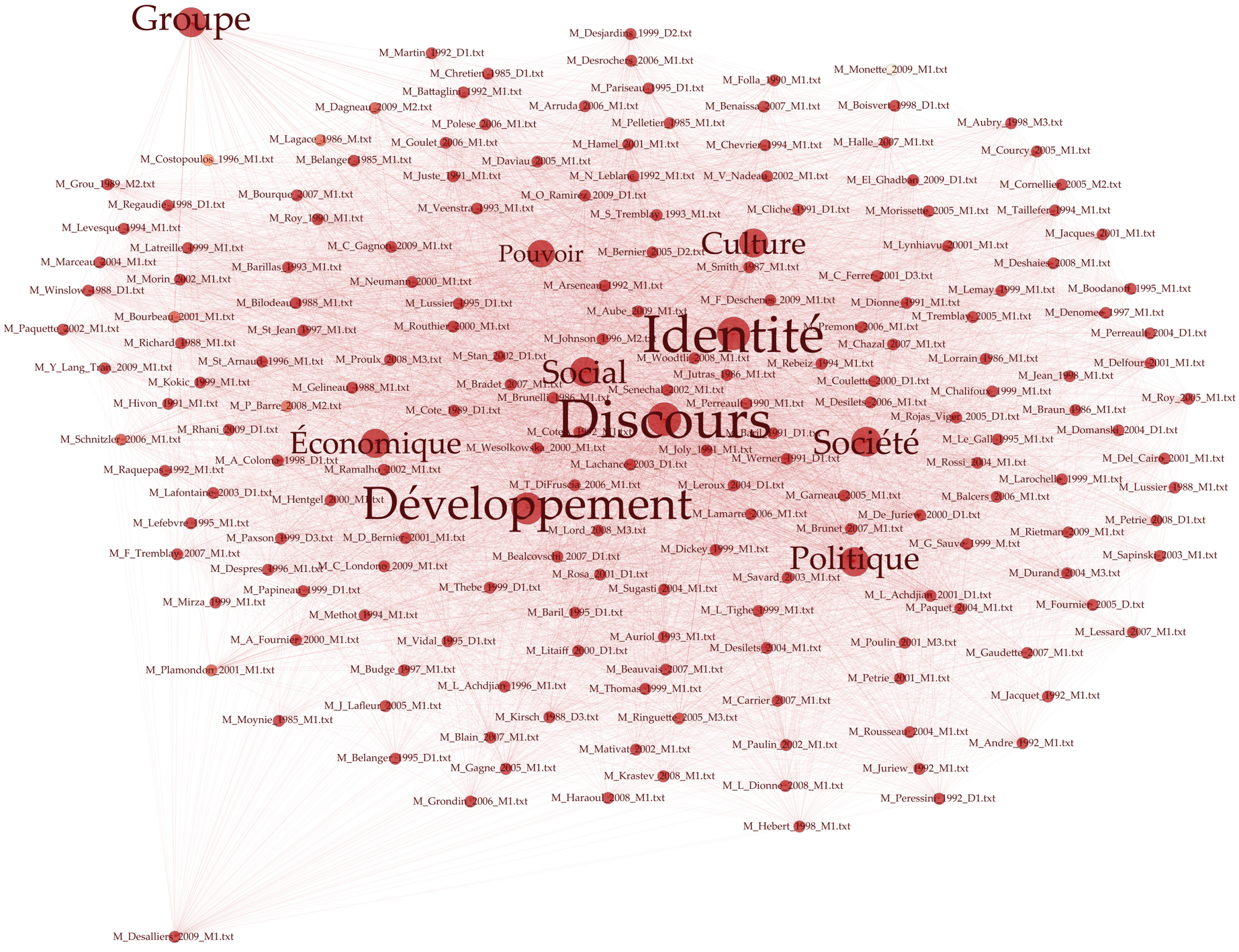


Figure 41. Partition graphique 4 (classe 4), département d'anthropologie de l'Université de Montréal.

Observations :

- Nombre de nœuds (résumés et mots-clés) : 194
- Nombre de liens : 8 660
- Occurrences de mots extraites et indices TF-IDF : « Discours » 89.2, « identité » 87.0, « développement » 78.8, « groupe » 63.7, « société » 62.2, « social » 59.9, « politique » 59.6, « économique » 57.5, « culture » 57.2, « pouvoir» 46.5.

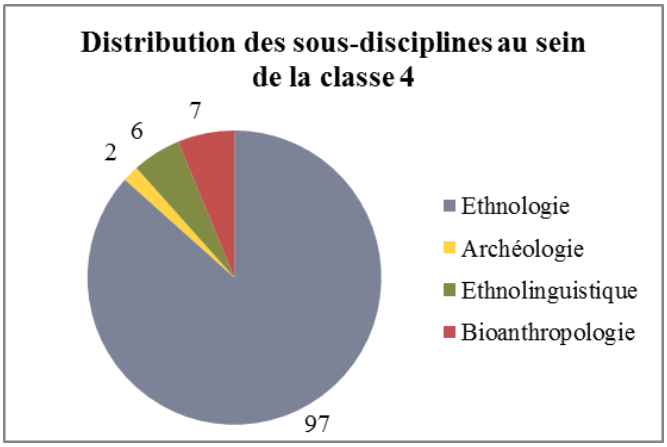


Figure 42. Répartition des résumés (valeurs réelles) en fonction des sous-disciplines, UdeM.

D'emblée, cette classe de documents est singulièrement ethnologique, avec que quelques résumés en ethnolinguistique (quatre) et en archéologie (quatre autres). Toujours en fonction de l'indice TF-IDF, les occurrences de mots se trouvant au cœur de cette classe sont « discours », « identité », « développement » et « groupe ». Cela nous porte à croire que plusieurs résumés abordent la construction de l'identité culturelle et des discours identitaires. Les mots qui suivent ; « société », « social », « politique », « économique » et « culture », enrichissent les thématiques socioculturelles. De ces derniers termes extraits, nous pouvons remarquer qu'ils s'expriment à un niveau macrostructurel ; c'est-à-dire que les résumés réfèrent assez fortement à une perspective sociétale et de groupe, en comparaison à l'individu pris

comme objet d'étude (niveau micro-structurel). Finalement, ayant un indice TF-IDF moins élevé, nous retrouvons l'occurrence de mot « pouvoir ».

Pour ce qu'il en est de la forme graphique que cette partition a prise, celle-ci est très uniforme et ronde, avec peu de distance entre les résumés. Cela donne l'impression que les thématiques présentes dans les résumés sont fortement liées entre elles, rendant un caractère compact à cette classe.

3.4.6. Classe 5, département d'anthropologie de l'UdeM

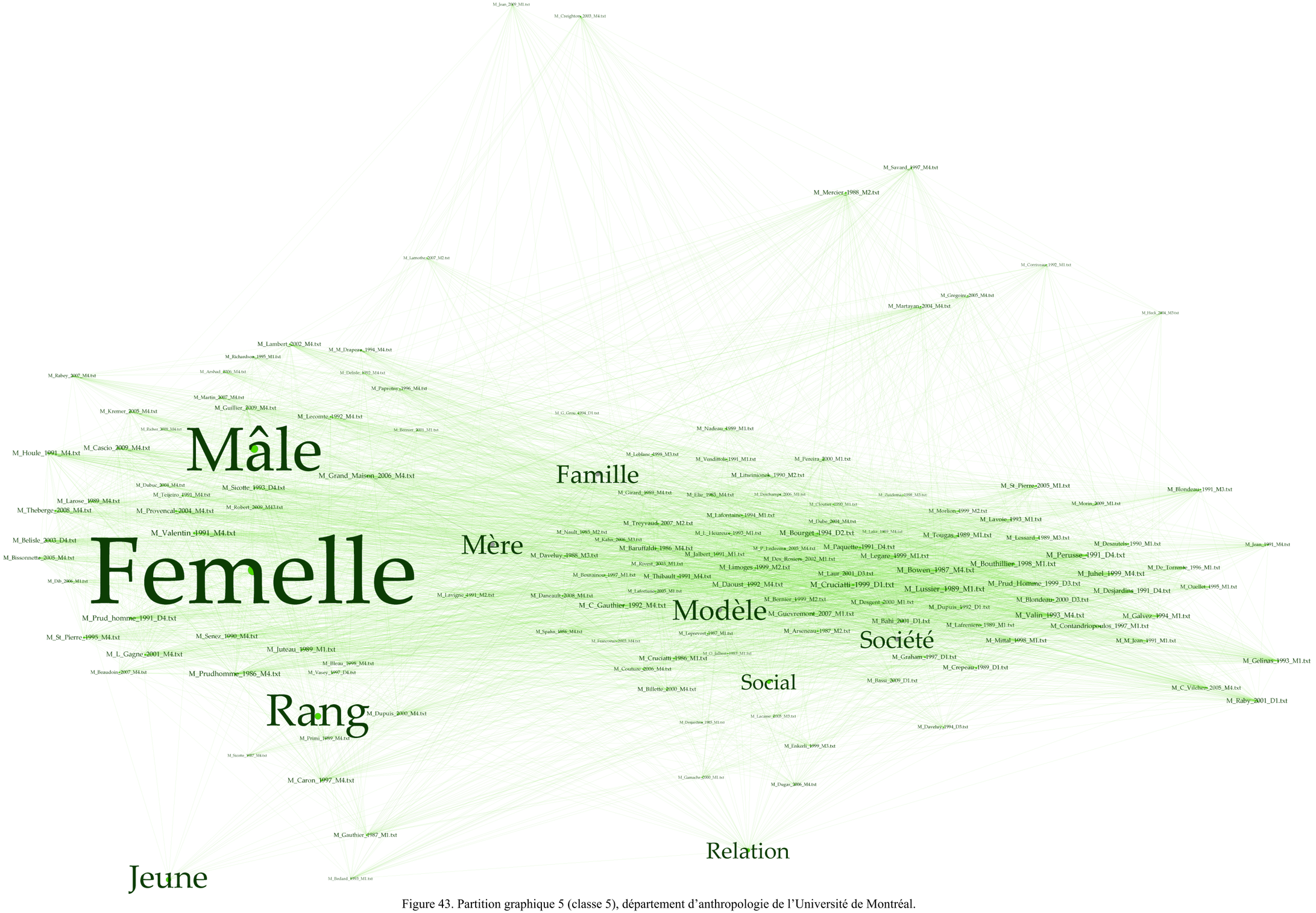


Figure 43. Partition graphique 5 (classe 5), département d'anthropologie de l'Université de Montréal.

Observations :

- Nombre de nœuds (résumés et mots-clés) : 147
- Nombre de liens : 3 950
- Occurrences de mots extraites et indices TF-IDF : « Femelle » 134.1, « mâle » 104.2, « rang » 86.4, « jeune » 70.1, « modèle » 64.9, « mère » 63.1, « famille » 60.6, « société » 58.7, « relation » 56.7, « social» 54.6.

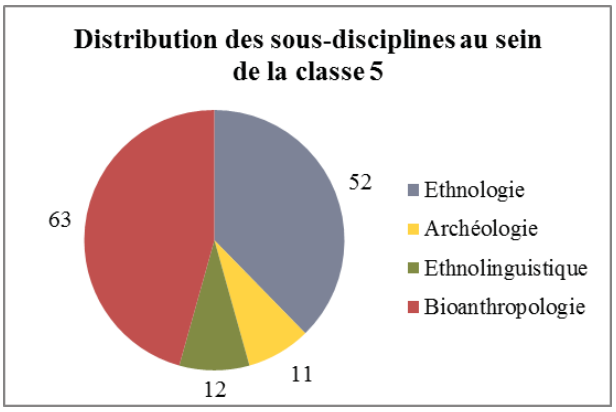


Figure 44. Répartition des résumés (valeurs réelles) en fonction des sous-disciplines, UdeM.

Pour cette dernière partition graphique, les mots extraits s'axent vers deux principales perspectives, la primatologie et l'ethnologie. Les termes qui possèdent les plus grandes valeurs de l'indice TF-IDF tendent assez fortement vers la primatologie : « femelle », « mâle », « rang » (rang social) et « jeune ». Cette dernière occurrence de mot peut par contre s'exprimer tant dans la l'étude des groupes humains que des groupes de primates. Toujours en termes d'indice TF-IDF, les termes extraits qui suivent sont « modèle », « mère » et « famille ». Tout comme pour les mots précédents, ceux-ci peuvent convenir à l'ethnologie et à l'étude des groupes humains, à la primatologie, mais aussi aux autres sous-disciplines, dans des perspectives biologiques, présentes comme passés. Néanmoins, il est intéressant de relever le thème de la famille, pris en tant qu'objet d'étude. Les dernières occurrences de mots qui furent extraites sont

« société », « relation » et « social », renforçant la thématique ethnologique et socioculturelle.

Quant à la forme graphique de cette classe, nous observons que le graphe est long, avec deux groupes que se dessinent vaguement : à gauche, ce sont des résumés en bioanthropologie (rappelons que la primatologie est une sous-discipline de la bioanthropologie) et à droite, ce sont des résumés en ethnologie, en ethnolinguistique et en archéologie. Il y a aussi quelques résumés qui sont plus distants, se divisant par rapport à l'ensemble.

3.5. Discussion

3.5.1. Interprétation globale

Globalement, plusieurs constats quant à l'interprétation de nos résultats s'imposent. Commençons par les occurrences de mots les plus fréquentes et discriminantes par rapport à l'ensemble du corpus :

Occurences de mots les plus fréquents et de ceux ayant les plus grands indices tf-idf							
Mots	Fréquences	Mots	Fréquences	Mots	tf-idf	Mots	tf-idf
Social	1167	Travail	699	Femme	609,6	Culturel	387,0
Groupe	849	Politique	622	Groupe	438,0	Pratique	374,7
Femme	816	Contexte	618	Social	429,8	Discours	364,2
Culturel	727	Pratique	577	Politique	394,4	Travail	357,4
Société	726	Rapport	564	Société	393,5	Communauté	346,9

Tableau 6. Occurrences extraites sur l'ensemble des données, 1240 résumés.

Premièrement, le mot « social » apparaît dans toutes les expérimentations, sans exception. Cela n'est pas vraiment surprenant puisque les phénomènes ou aspects sociaux sont très étudiés en ethnologie, en archéologie, en ethnolinguistique, et même en bioanthropologie (nous pensons par exemple aux comportements sociaux étudiés en primatologie). La deuxième occurrence de mot qui est très présente dans les différentes expérimentations est « groupe ». Il semble aller de soi que cette occurrence est importante pour le corpus, étant donné que les chercheurs étudient fréquemment des groupes, que ce soient des individus, des animaux ou des objets. Quant au mot suivant, « femme », celui-ci revêt une importance indéniable pour notre corpus étant donné son indice TF-IDF, qui est le plus élevé (609.6). Cela est intéressant et nous comprenons que la thématique de la femme aura été abordée sous différentes perspectives : culturellement, historiquement, archéologiquement et biologiquement. Vient ensuite le mot « culturel » qui, remarquons-nous, est nettement plus fréquent que sa racine ;

« culture ». Tout comme l'occurrence de mot « social », le terme « société » est très important dans notre corpus et est largement abordé en ethnologie et en ethnolinguistique, mais également dans les autres sous-disciplines de l'anthropologie, où les archéologues étudient fréquemment les sociétés du passé, tout comme les paléoanthropologues.

Pour ce qu'il en est du terme « travail », tel que précédemment expliqué, nous devons le rejeter, car trop sujet à la polysémie. Dans environ le tiers des cas, les auteurs utilisaient ce terme ainsi : « Dans ce travail, il est question de… ». À cause de cette problématique, cette occurrence ne peut être véritablement significative, malgré l'importance que peut avoir cette thématique en anthropologie. Les occurrences suivantes, par contre, nous semblent très significatives. Les termes « politique » et « pratique » ont en effet une grande portée pour notre corpus avec des indices TF-IDF élevés (394,4 dans le premier cas et 374,7 pour le second). En bref, ces thématiques ne sont pas négligeables quant à l'interprétation globale de nos résultats.

Du point de vue fréquentiel, les occurrences subséquentes sont aussi importantes : « contexte » et « rapport ». Peu importe le champ disciplinaire, ce sont des concepts-clés, permettant de contextualiser des objets d'études, quels qu'ils soient.

3.5.2. Les principaux schèmes thématiques

En regard à nos résultats de recherche, nous remarquons quelques grands schèmes thématiques qui ressortent. D'abord, comme nous l'avons déjà mentionné, les thématiques utilisées exprimant l'ethnologie sont nombreuses, variées et semblent moins circonscrites par rapport aux thématiques des autres branches disciplinaires de l'anthropologie (mise à part l'étude de la femme). Cette prépondérance des thématiques socioculturelles reflète plutôt bien notre corpus, constitué à 70 % de résumés en anthropologie sociale et culturelle. Aussi, à travers nos expérimentations, nous avons pu remarquer des classes de documents abordant des thématiques de l'ethnologie assez

précises dont, notamment, l'étude autochtone, l'étude du patrimoine québécois, des phénomènes religieux, migratoires, de l'étude de la famille, des communautés, des identités culturelles, etc.

Pour expliquer la dominance thématique de l'ethnologie et des thématiques socioculturelles et en prenant un point de vue historique, nous remarquons que les départements concernés furent créés à partir des années 1960 (UlavalA; 1970, UlavalH; 1971, UdeM; 1961), résultant en grande partie de volontés d'autonomie (par rapport à la sociologie et à la géographie, dans le cas du département d'histoire) et de modernisations sociétaires (par exemple, la Révolution tranquille fut un instrument de modernisation). Il y eu donc dans la société et au sein des institutions des besoins pour comprendre la mouvance sociétaire et étudier notre culture en tant qu'objet évolutif. Aussi, dès les débuts des départements d'anthropologie, l'ethnologie était essentielle quant à la définition de l'anthropologie.

> « De manière significative, il [Dubreuil, fondateur du département d'anthropologie de l'UdeM] définit l'ethnologie en s'appuyant sur un texte de Claude Lévi-Strauss posant que la comparaison des cultures contemporaines implique la recherche de « certaines propriétés générales de la vie sociale ». À l'Université Laval, dès sa création, le département d'anthropologie est orienté essentiellement vers l'enseignement de l'anthropologie sociale et culturelle et concentre ses activités de recherche sur des aires culturelles rapprochées : Canada français, Amérindiens et Inuit » (Crépeau : 2004, 387).

Ensuite, au fil des ans, les départements se sont développés, ont accueilli de nouveaux enseignants venus de différents milieux académiques, permettant la diversification des approches théoriques et pratiques. Ainsi, l'ethnologie est un pilier important de l'anthropologie québécoise et cela transparaît dans nos résultats. Par contre, les patrons thématiques ne sont pas nécessairement faciles à cerner, c'est-à-dire que nous ressentons une certaine confusion thématique.

Par ailleurs, la force de l'occurrence « femme » représente, selon nous, notre plus grande découverte découlant de nos expérimentations. Comme précédemment

mentionné, nous comprenons que cette thématique est très significative de l'anthropologie québécoise et qu'elle a été abordée sous diverses perspectives, tant biologique que culturelle et contemporaine que passé. Nous avançons également que les mouvements féministes du XX^{e} siècle, tant dans les milieux scientifiques et dans les sociétés, ont pu jouer un important rôle pour la prépondérance de cette thématique. « *Le Département d'anthropologie demeure en pleine expansion tout au long de la décennie 1980. […]Le féminisme, comme approche théorique, et la question des rapports sociaux de sexe, au plan thématique, s'implantent.* » (Beaudoin : 2010).

Un autre grand constat que nous faisons concerne la bioanthropologie et, plus spécialement, la primatologie. En fait, les classes générées présentant des mots propres à cette sous-discipline se perçoivent clairement et les algorithmes de classification ont donc pu regrouper facilement les résumés portant sur la primatologie, même si cette sous-discipline n'est enseignée qu'au département d'anthropologie de l'Université de Montréal. Cela signifie que la primatologie est un champ disciplinaire qui a su se démarquer, au Québec, durant le dernier quart de siècle et que les occurrences se rapportant à la primatologie sont suffisamment discriminantes pour se dissocier des autres thématiques de l'anthropologie. Remarquons finalement que lorsque des résumés issus d'autres sous-disciplines anthropologiques étaient associés aux classes de la primatologie, c'étaient généralement des résumés en ethnologie. Nous expliquons ces cooccurrences disciplinaires par les approches conceptuelles et méthodologiques souvent similaires ; les primatologues étudient généralement des comportements chez les primates afin de comprendre les origines de l'homme, alors que les ethnologues étudient fréquemment des comportements chez l'homme et dans les sociétés.

Du point de vue archéologique, les classes de résumés ont tendance à ressortir assez nettement dans nos résultats par rapport aux autres sous-disciplines, sans toutefois présenter de grandes démarcations thématiques entre les départements d'anthropologie de l'Université de Montréal et le département d'histoire de l'Université Laval. S'il y a des distinctions thématiques à faire, elles se trouvent certainement dans les aires

géoculturelles abordées et les périodes historiques. À l'UdeM, l'étude de la culture « Mochicas » (Pérou précolombien) ressort fortement des analyses, alors qu'à Ulaval, les mots extraits tendent vers le Québec, l'archéologie de la ville et le XX^{e} siècle.

Finalement, les résumés en ethnolinguistique semblent moins visibles, souvent associés avec des classes présentant des mots ethnologiques. Cela ne nous semble pas problématique puisque l'ethnologie et l'ethnolinguistique sont des sous-disciplines proches. Voici toutes les langues et les concepts clés de l'ethnolinguistique, qui ont été extraits de nos expérimentations : Grec, Afrikaans (langue africaine), québécois, francophone, pidgin (langues véhiculaires), Kabyle (langue berbère), symbolique et identité culturelle.

3.5.3. L'approche interprétative culture/biologique

À la suite de ce que nous venons de voir, il nous apparaît que l'approche disciplinaire nord-américaine, introduite par Franz Boas (1858-1942) au début du XX^{e} siècle (séparation de l'anthropologie en quatre branches disciplinaires), n'est peut-être pas une approche tout à fait appropriée pour interpréter nos résultats. En fait, en entreprenant ce projet de recherche, nous étions fortement imprégnés par cette approche et nous comprenons, finalement, que les thématiques sont inter-reliées et assez communes dans les différentes sous-disciplines. Ainsi, peu importe les champs disciplinaires, les chercheurs s'empruntent des théories, des méthodes, des approches et, bien sûr, des thématiques. Donc, il en ressort un certain métissage thématique et scientifique, faisant en sorte que l'anthropologie québécoise, à travers nos données, ne se compartimente pas toujours facilement en branches de disciplines, mais se complète plutôt par la variété des approches qu'elle témoigne. Ainsi, pour simplifier l'approche interprétative des résultats, peut-être est-il plus juste de percevoir l'anthropologie au Québec par la dichotomie culture/biologie. Donc, d'un côté, nous retrouvons les sous-disciplines qui abordent les phénomènes culturaux, présents comme passés, alors que de

l'autre côté, ce sont les sous-disciplines qui étudient les phénomènes biologiques, des points de vue préhistoriques, historiques et contemporains.

	Culture	**Biologie**
Présent	Anthropologie sociale et culturelle	Bioanthropologie
Passé	Archéologie	Paléoanthropologie

Tableau 7. Représentation disciplinaire de l'anthropologie selon la dichotomie culture/biologie. Tirée de St-Denis : 2006, 3.

3.5.4. Limites rencontrées et pistes

Lors de notre recherche, nous avons fait face à plusieurs difficultés, nous amenant à chercher des solutions pour les contrer.

La première limite concerne la durée du processus méthodologique. Certaines étapes ont en effet été longues à réaliser, et nous pensons précisément à la numérisation des textes et au prétraitement documentaire, ainsi qu'à la visualisation des résultats. Dans le cas de notre collecte des données, nous avons dû numériser plus de 60 % de nos textes et ainsi, cette étape s'étala sur plusieurs mois. Nous pensons que cette situation s'estompera dans l'avenir compte tenu de l'omniprésence du numérique dans nos sociétés. Quant à l'étape de prétraitement des données (normalisation et filtrage), nous sommes d'avis que celle-ci est indispensable en fouille de textes et qu'il est important de s'y attarder afin de construire des filtres adaptés aux corpus étudiés. Cette étape est incontournable, mais pourrait être allégée par l'utilisation de dictionnaires de mots fonctionnels et de lemmatisation déjà existants.

La deuxième limite rencontrée réside dans le processus de classification, où il arrivait que certaines classes générées soient disproportionnées par rapport aux autres classes. En fait, puisque le processus algorithmique impliquait l'ajustement des

paramètres (par exemple, dans les paramètres, nous rejetions les mots qui apparaissaient dans plus de 65 % des cas) pour ensuite lancer l'algorithme de classification, nous avancions par essais-erreurs, en observant les résultats. Et pour ajouter à cette difficulté, il n'était pas possible de prédéfinir *a priori* le nombre de documents par classe. Donc, l'algorithme créait automatiquement des classes grâce au paramétrage et lorsque les résultats nous semblaient satisfaisants, nous conservions les fichiers de classification. Seulement, pour certaines expérimentations, même après de nombreux essais de paramétrage, nous étions incapables d'obtenir des classes proportionnelles. Nous n'avons pas trouvé de solution applicable à notre projet. Peut-être en fait que cette problématique n'en est pas vraiment une et que la disproportion rencontrée reflète en quelque sorte la réalité, où les grandes classes représentent des thématiques fortes, souvent étudiées, alors que les petites classes exposent des thématiques qui se distinguent et se dissocient davantage. Par ailleurs, l'algorithme *Modularity class*, de Gephi, nous a permis de générer des classes beaucoup plus proportionnelles, mais en contrepartie, nous avions un accès limité aux paramètres, amenant une certaine imprécision de notre processus de classification. En bref, nous pensons qu'il est essentiel d'étudier et de bien choisir les approches et les méthodes employées pour réaliser tout projet de fouille de textes. Nous ne croyons pas nous être trompés lorsque nous avons choisi la classification hiérarchique ascendante, mais nous reconnaissons tout de même que d'autres méthodes ou d'autres systèmes pourraient peut-être surpasser la problématique de la disproportion des classes.

Une autre limite que nous souhaitons mettre en évidence concerne l'exportation des fichiers d'analyse d'un logiciel à l'autre. Après l'étape de classification, réalisée dans WordStat, nous voulions exporter nos fichiers de classification dans Gephi pour débuter l'étape de visualisation des résultats. Seulement, dans le processus d'exportation des matrices vectorielles, les classes générées dans WordStat n'étaient pas reconnues par Gephi et nous avons dû réitérer le processus de classification. Puisque Gephi est un logiciel libre, en perpétuel développement, celui-ci possède des forces et des faiblesses (l'importation des fichiers en est une). Nous croyons que la problématique d'exportation

des fichiers mérite une réflexion plus poussée et qu'il sera possible, dans des études futures et grâce à des compétences en programmation, de corriger cette lacune.

Un autre problème dont nous souhaitons faire mention réside dans la difficulté de manipulation des graphes, causée par les grandes quantités d'information contenues dans les fichiers. Nous avions en effet trop de liens, ce qui rendait l'exploitation de notre graphe initial difficile, voire impossible à télécharger et à manipuler en formats PDF ou HTML. Pour résoudre cette problématique, nous avons partitionné notre graphe en cinq sous-graphes, ce qui est, à nos yeux, une bonne solution, car ainsi, nous décomposions et simplifions la visualisation graphique. Toutefois, cette procédure éliminait les liens interclasses, éléments pouvant être pertinents pour l'analyse interprétative. Nous croyons encore une fois que des compétences en programmation pourraient permettre de solutionner la problématique de surcharge informationnelle des fichiers informatiques.

Pour conclure les limites de notre recherche, nous aimerions aborder la question des corpus, et plus particulièrement de l'ampleur de ceux-ci. En fouille de textes, les grands corpus priment sur les petits et ces techniques gagnent ainsi en pertinence avec l'accroissement du nombre de textes à exploiter. Conséquemment, les 1240 résumés utilisés pour notre recherche représentent un petit corpus et nous croyons que nous obtiendrions des résultats plus percutants avec un corpus de milliers de documents. Imaginez en effet ce que nous pourrions découvrir avec plus de 100 000 textes scientifiques en anthropologie. La fouille de textes autant que la fouille épistémologique serraient alors des plus profitables.

Conclusion

À travers ce travail, nous prenions comme objectif premier l'exploration de méthodes de fouilles de textes afin d'évaluer leur pertinence pour le repérage et l'analyse thématique de grands corpus de textes. Nous souhaitions ainsi valider l'hypothèse selon laquelle la méthode de classification hiérarchique ascendante (CHA) peut permettre l'identification thématique et l'extraction de mots significatifs issus directement du lexique, menant ultimement à la visualisation thématique.

En première partie, nous avons d'abord présenté notre problématique, qui peut se résumer ainsi ; face au foisonnement technologique et à l'accroissement exponentiel des documents numériques, nous croyons qu'il est indispensable de développer des outils et des méthodes informatiques afin de favoriser l'exploitabilité de l'information numérique, tant sous une perspective de gestion documentaire qu'en regard aux aspects sémantiques et à l'analyse thématique de données non structurées (textes). De plus, pour construire notre cadre théorique, nous avons présenté les définitions de la fouille de textes, leurs origines, leurs évolutions et les applications qui en découlent. Nous avons également abordé les principales familles méthodologiques, c'est-à-dire la catégorisation et la classification automatiques de textes. Nous complétions par une revue de littérature qui, nous le reconnaissions, ne représentait qu'un bref survol des publications abordant la fouille de textes ou la visualisation graphique de l'information.

La deuxième partie du mémoire consistait à présenter notre cadre méthodologique, où nous commencions par la description de nos données, ainsi que par la définition de l'anthropologie québécoise. En bref, les données que nous avons utilisées représentent 1 240 résumés de mémoires et de thèses, octroyés durant les années 1985 à 2009, par les départements d'anthropologie de l'Université de Montréal et de l'Université Laval, ainsi que par le département d'histoire de l'Université Laval (pour les résumés en archéologie et en ethnologie). Remarquons que les sous-disciplines anthropologiques abordées à travers notre corpus sont l'ethnologie (y compris le folklore), l'archéologie, la bioanthropologie et l'ethnolinguistique. Quant à

l'anthropologie québécoise, nous considérons la définition du département d'anthropologie de l'Université de Montréal, voulant que cette discipline soit l'étude de la culture et de l'humanité sous toutes ses dimensions : biologique, historique, comportementale et communicative (Département d'anthropologie de l'Université de Montréal, Site Internet : Présentation du département). Toujours dans la deuxième partie, nous avons présenté les étapes de notre processus méthodologique ; la collecte et la numérisation des données, le prétraitement (normalisation des textes), le filtrage statistique et linguistique, la vectorisation, la classification automatique et l'extraction des termes discriminants, pour finalement compléter notre démarche par la visualisation (création de graphes à partir des fichiers d'analyse de la classification).

Dans la troisième partie, nous présentions en premier lieu les résultats de deux expérimentations, la classification hiérarchique ascendante à trois niveaux et la classification hiérarchique ascendante du département d'anthropologie de l'UdeM. Nous proposions à cet effet des cartes conceptuelles et les partitions graphiques obtenues du processus de visualisation, accompagnées de nos observations et de l'interprétation que nous en faisons. En deuxième lieu, nous proposions une discussion sur l'interprétation globale des résultats et sur des principaux schèmes thématiques qui s'en dégagent. Nous avons ainsi découvert que les thématiques se rapportant à la femme sont essentielles pour notre corpus, surpassant en termes d'indice TF-IDF les autres termes extraits dont « social », « culture », « société », « culturel », etc. Nous avons également reconnu la prédominance des thématiques socioculturelles, s'expliquant par la forte présence de l'ethnologie dans notre corpus (70% des résumés se rapportaient à l'ethnologie). Aussi, nous avons pu identifier facilement des classes ayant des thèmes de la primatologie, ou, encore, de l'archéologie. Les thématiques de la bioanthropologie étaient un peu moins visibles, mais néanmoins reconnaissables, alors que les thèmes de l'ethnolinguistique avaient plutôt tendance à s'entrecroiser aux thématiques de l'ethnologie. Finalement, nous sentions dans l'ensemble une confusion ou une désorganisation des thématiques ethnologiques et il était parfois difficile d'établir des patrons thématiques précis. Nous croyons que cela s'explique, en partie, par le

métissage scientifique, où les chercheurs issus de différents champs disciplinaires s'empruntent des théories, des méthodes et abordent des thématiques parentes, sous différentes perspectives.

Par ailleurs, notre réflexion sur l'interprétation globale des résultats nous a amené à aborder les approches interprétatives qu'il est possible d'adopter. En fait, dès les débuts de ce projet, nous concevions l'anthropologie québécoise selon l'approche nord-américaine, qui divise l'anthropologie en quatre branches disciplinaires. Seulement, plusieurs autres approches existent dont, notamment, l'approche européenne, où l'anthropologie se consacre aux phénomènes socioculturels et à l'humanité culturelle, laissant les aspects biologiques de l'homme aux disciplines de la biologie et de l'histoire, pour ce qui a trait au passé. Cela nous a conduit à la dichotomique « culture / biologie », qui pourrait possiblement se révéler plus pertinente quant à l'interprétation de nos résultats de recherche, principalement eu égard aux résumés purement ethnologiques. Poursuivant sur le sujet de l'interprétation des résultats de recherche, nous avons abordé la notion de navigation thématique, qui confère aux utilisateurs la possibilité de parcourir les structures thématiques du lexique. Autrement dit, nos résultats d'analyses peuvent être perçus comme un outil de navigation, permettant aux utilisateurs d'explorer et de découvrir des séries thématiques. Les interprétations peuvent donc être multiples et sont, jusqu'à une certaine limite, subjectives aux personnes qui thématisent.

Puis, comme tout projet, nous avons fait face à des limites méthodologiques, que nous exposions pour clore notre discussion. Ainsi, nous avons soulevé la grande durée du processus méthodologique, qui pourrait être réduit par des stratégies comme l'utilisation de dictionnaires de lemmatisation déjà existants. Nous avons également abordé la difficulté de classification de certains ensembles de documents (surtout les textes en ethnologie), produisant une disproportion des classes générées. Concernant cette difficulté, nous proposons l'exploration d'autres logiciels de classification ou l'emploi d'autres méthodes, telles que l'analyse réseau ou la catégorisation automatique.

L'exportation des fichiers d'analyse réalisés dans le logiciel WordStat vers le logiciel d'analyse réseau Gephi fut sans contre dit notre plus importante limite. En bref, le processus d'exportation des matrices vectorielles ne nous a pas permis de percevoir les classes générées par la classification une fois intégrées dans le logiciel Gephi. Pour contrer cette difficulté, nous avons réinitialisé la classification à l'aide de l'algorithme Modularity class, ce qui nous a permis d'obtenir des classes plus proportionnelles, présentant en majorité les mêmes termes extraits que dans l'expérimentation réalisée dans WordStat. Nous croyons que des études futures permettront de résoudre la difficulté de l'exportation et de l'importation des fichiers, par exemple grâce à d'autres méthodes d'intégration ou, encore, par la modification informatique des fichiers d'analyse (par exemple, ajout de balises).

Finalement, la complexité des fichiers (grande quantité d'information contenue dans les fichiers graphiques) est une limite qui nous aura hantée à de nombreuses reprises durant l'étape de visualisation des résultats. Notre fichier graphique initial possédait 729 nœuds (les résumés de l'UdeM accompagnés de 10 occurrences de mots pour chacune des classes), avec 33 538 liens. Nous avons alors décidé de partitionner notre graphe selon les cinq classes générées par la classification, nous amenant ainsi à décomposer et simplifier nos résultats graphiques, soustrayant toutefois tous les liens interclasses.

Les mots de la fin seront que ce projet fut enrichissant et passionnant, nous permettant d'atteindre plusieurs objectifs. Nous avons en effet constitué un corpus numérique qui n'existait pas auparavant, nous avons expérimenté des méthodes informatiques issues des domaines de la Lecture et de l'Analyse de Textes Assistées par Ordinateur (LATAO), de la Gestion Électronique des Documents (GÉD) et de la visualisation de l'information. Avec la prolifération des documents numériques et les innovations informatiques, nul doute que les techniques de fouilles de textes et, plus largement, les méthodes de gestion et d'analyses thématiques de grands corpus de textes représentent des voies prometteuses pour les chercheurs de diverses disciplines.

Bibliographie

Adobe Systems. Microsoft Illustrator. <http://tv.adobe.com/product/illustrator/> (consulté en 2010).

Archambeault, Jean. *Visualisation de l'évolution d'un domaine scientifique par l'analyse des résumés de publication à l'aide de réseaux neuronaux*. Mémoire de maîtrise, Montréal : Presses de l'Université de Montréal, 2002, 115 pages. <http://www.ebsi.umontreal.ca/rech/archambeault/memoire_jean_archambeault.pdf> (Consulté en octobre 2010).

Association des anthropologues du Québec. Projet *L'Hypo*^{Thèse}. 2008. <http://www.aanthq.qc.ca/Francais/Banque_theses/Theses.html> (consulté en 2011).

Batist, Zack. *Feasting in Bronze Age Greece : A network Analysis Approach*. Research project submitted for Bachelor of Arts with Honours. Institute of Interdisciplinary Studies, Carleton University, Ottawa, 2012, 57 pages.

Beaudoin, Samuel. *L'anthropologie à l'Université Laval et le département sur 40 ans. Histoire et mission*. Dans le cadre du 40^{e} anniversaire du Département d'anthropologie. 2010. <http://www.ant.ulaval.ca/?pid=1370> (Consulté en avril 2011).

Bibeau, Gilles et al. À l'âge des commencements : un récit de l'origine du département d'anthropologie de l'Université de Montréal. *Anthropolama*. Édition spéciale bilingue colloque AAA Annual Meeting Special Bilinguual Edition, novembre 2011, p. 3-6.

Blondel, Vincent D. et al. *Fast unfolding of communities in large networks*. Université catholique de Louvain. <http://arxiv.org/pdf/0803.0476v2.pdf> (Consulté en septembre 2011).

Bohannon, John. Google Books. Wikipedia, and the Future of Culturomics. *Science*, vol. 331, février 2011. <http://www.sciencemag.org/content/331/6014/135.full.pdf> (Consulté en mars 2012).

Bonte, Pierre et al. *Dictionnaire de l'ethnologie et de l'anthropologie*. Paris : Presses universitaires de France, 1991, 755 pages.

Borgatti, P. et al. Network Analyses in the Social Sciences. *Science*, vol. 323, février 2009, p. 892-895.

Bremond, C. 1985. Concept et thème. *Poétique*. No 64, p. 415-423.

Centre National de Ressources Textuelles et Lexicales (CNRTL). Portail lexical : Étymologie. <http://www.cnrtl.fr/etymologie/> (consulté en avril 2012).

Crépeau, Robert. Le couple nature-culture ou la chronique d'un divorce annoncé. Clermont, N. (éd.), *Anthropologie et Écologie*. Actes du troisième Colloque Annuel du Département d'Anthropologie, Université de Montréal, Montréal, 1997, p. 9-17.

Crépeau, Robert. La réception du structuralisme lévi-straussien au Québec. *Cahier de l'Herne*, vol 82, 2004, p. 387-395.

Delorme, Jean-Michel. *L'apport de la fouille de données dans l'analyse de texte. Examen probatoire*. Montpellier : Conservatoire National des Arts et Métiers, 2002, 31 pages.

Département d'histoire, Université Laval, présentation des disciplines ; ethnologie : <http://www.hst.ulaval.ca/le-departement/disciplines/ethnologie/> (consulté le 20 novembre 2012).

Département d'anthropologie, Université de Montréal. Historique du département. <http://anthropo.umontreal.ca/departement/historique-du-departement/> (Consulté en janvier 2010).

Dumonchel, Laurence. *L'étonnante relation entre l'archivistique et l'anthropologie*. Travail réalisé dans le cadre du cours ARV1050 – Introduction à l'archivistique, EBSI, Université de Montréal, hiver 2009, 12 pages.

Dumont, Micheline. *Du féminin au féminisme : L'exemple québécois. In Clio. Histoire, femmes et sociétés*. 1997, mis en ligne en janvier 2005, <http://clio.revues.org/388 ; DOI : 10.4000/clio.388> (consulté en février 2013).

Feldman, R. et al. *Text mining at the term level. Principles of Data Mining and Knowledge Discovery*. Second European symposium. Berlin : Springer-Verlag. 23 au 26 septembre 1998, p. 65-73.

Florida Institute for human & machine cognition. *C-Map*. <http://cmap.ihmc.us/> (consulté en avril 2012).

Forest, Dominic. *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 : Presses de l'Université du Québec à Montréal, 2002. <http://www.dominicforest.name/documents/MA.pdf> (Consulté en septembre 2010)

Forest, Dominic et Jean-Guy Meunier. *La classification mathématique des textes : un outil d'assistance à la lecture et à l'analyse de textes philosophiques*. Rajman, M. & Chappelier, J.-C. (eds.), *Actes des 5es Journées internationales d'Analyse statistique des Données Textuelles*, 9-11 mars 2000, Lausanne, Suisse : EPFL, vol. 1, p. 325-329.

Forest, Dominic et Jean-Guy Meunier. Classification et catégorisation automatiques : application à l'analyse thématique des données textuelles. *Le poids des mots.* Purnelle, G., Fairon, C. et Dister, A. (éditeurs). Actes des 7es Journées internationales d'analyse statistique des données textuelles. Louvain-la-Neuve : Presses Universitaires de l'Université Catholique de Louvain, vol. 1, 2004, p. 434 à 444.

Forest, Dominic. *Application de techniques de forage de textes de nature prédictive et exploratoire à des fins de gestion et d'analyse thématique de documents textuels non structurés.* Thèse de doctorat. Montréal : Presses de l'Université du Québec à Montréal. 2006. <http://www.dominicforest.name/documents/Forest_Thesis.pdf> (Consulté en avril 2010).

Forest, Dominic. Vers une nouvelle génération d'outils d'analyse et de recherche d'information. *Documentation et Bibliothèques.* Montréal : Éditions ASTED, Vol. 55, no. 2 (Avril-juin), 2009, p.77-89.

Fortunato, Santo. *Community detection in graphs*. Italy: Complex Networks and Systems Lagrange Laboratory, ISI Foundatio, 2010. <http://arxiv.org/pdf/0906.0612.pdf> (Consulté en juillet 2011).

Gephi. *Tutorial Layouts*. 2008. < https://gephi.org/> (consulté en mai 2011).

Gephi. *Open source software for exploring and manipulating networks*. 2008. < https://gephi.org/> (consulté en mai 2011).

Ghitalla, Franck. 2010. *Réseaux, Traces et Controverses. Projet de recherche. L'Atelier de Cartographie.* 2010. <http://www.slideshare.net/Ghitalla/reseaux-traces-controverses> (Consulté en avril 2011).

Grouin, Cyril et Dominic Forest. *Expérimentations et évaluations en fouille de textes : Un panorama des campagnes DEFT*. Paris : Lavoisier, 2012, 248 pages.

Hearst, M. *What is text mining* ? 2003. Document non-publié mais disponible à l'adresse <http://people.ischool.berkeley.edu/~hearst/text-mining.html> (Consulté en avril 2011).

Hébert, Louis. *La sémantique interprétative. Site Web Signo*. Université du Québec à Rimouski. 2006. <http://www.signosemio.com/rastier/semantique-interpretative.asp> (Consulté en mars 2012).

Ibekwe-SanJuan, Fidelia. *Fouille de textes : méthodes, outils et applications*. 1ère édition. Paris : Lavoisier, 2007, 352 pages.

Jacomy, Alexis. *GexfWalker* (open source web widget). 2011. <https://dhs.stanford.edu/spatial-humanities/walking-through-networks-with-gexfwalker/> (consulté en juin 2012).

Kodratoff, Y. Knowledge discovery in texts : A definition, and applications. *Foundation of Intelligent Systems*. Berlin : Springer-Verlag, 1999.

<http://www.lri.fr/LRI/ia/articles/yk/1999/kodratoff99a.pdf>. (Consulté en septembre 2010).

Labatut, Vincent et Jean-Michel Balasque. *Detection and Interpretation of Communities in Complex Networks*: Practical Methods and Application. Turquie : Galatasaray University, 2012, 33 pages. <http://www.academia.edu/551627/Detection_and_Interpretation_of_Communities_in_Complex_Networks_Practical_Methods_and_Application> (Consulté en octobre 2012).

Lerot, Jacques. *Précis de linguistique générale*. Paris : Minuit, 1993, 446 pages.

Louwerse, M. et Van Peer, W. *Thematics: Interdisciplinary Studies*. 1ère édition. Amsterdam : John Benjamins Publishing Company, 2002, 458 pages.

Lyons, John. *Linguistique générale: introduction à la linguistique théorique*. Paris : Librairie Larousses, 1970, 382 pages.

Martin, Éveline. Thème d'étude, étude de thème. *L'analyse thématique des données textuelles : l'exemple des sentiments*. Sous la dir. de Rastier, F. Paris : Didier érudition, 1995, p.13-24.

Meunier, J.-G. La lecture et l'analyse de texte assistée par ordinateur : La chaîne d'analyse. *Cahiers de recherche du Laboratoire d'ANnalyse Cognitive de l'Information*, 1995, vol. 6.

Michel, Jean-Baptiste et al. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*. vol. 331, no. 6014, 2011, p. 176-182.

NIST Information Technology Laboratory's (ITL) Retrieval Group of the Information Access Division (IAD). *Text REtrival Conference* (TREC). 2000. <http://trec.nist.gov/> (consulté en avril 2012).

Paranyus Kin, Dmitry. 2011. *Identifying the Pathways for Meaning Circulation using Text Network Analysis*. Berlin: Nodus Labs, 2011. < http://noduslabs.com/research/pathways-meaning-circulation-text-network-analysis/> (Consulté en juin 2011).

Prince, G. Thématiser. *Poétique*. No 64, 1985, p. 425-433.

Rafols, Ismael, Alan L. Proer et Loet Leydesdorff. *Science overlay maps : a new tool for research policy and library management*. England: Science and Technology Policy Research, University of Sussex, 2010.
< http://www.leydesdorff.net/overlaytoolkit/overlaytoolkit.htm> (Consulté en mars 2012).

Rastier, F. et al. *Sémantique pour l'analyse. De la linguistique à l'informatique*. Paris : Masson, 1994, 240 pages.

Rastier, François. Sciences de la culture et post-humanité. *Texto !* septembre 2004 [en ligne]. Disponible sur : <http://www.revue-texto.net/Inedits/Rastier/Rastier_Post-humanite.html>. (Consultée en septembre 2012).

Rossignol, M. et Sébillot, P. Combining statistical data analysis techniques to extract Topical keyword classes from corpora. *IDA. Intelligent Data Analysis*. Prague: IDA Intelligent Data Analysis Research Lab, vol. 9, no. 1, 2005, p. 105-127.

Rozenknop, A. *Modèles en Recherche d'Information*. Cours master en Recherche et extraction d'information, Université Paris 13, 2001. < http://lipn.univ-paris13.fr/~rozenknop/Cours/MICR_REI/Seance6/modeles-RI-1.pdf> (Consulté en janvier 2012).

Sebastiani, F. Machine learning in automated text categorization. *ACM Computing Survey*. London: Association for Computing Machinery (ACM), vol. 34, no 1, 2002, pages 1-47.

Shaban, Khaled. 2006. *A Semantic Graph Model for Text Representation and Matching in Document Mining*. Thèse de doctorat. Ontario: University of Waterloo Press, 2006. <http://www.collectionscanada.ca/obj/s4/f2/dsk3/OWTU/TC-OWTU-1049.pdf> (consulté en juin 2010).

Sprylogics International. *Cluuz beta* (search engine). 2008. <http://www.cluuz.com/> (consulté en mars 2011).

St-Denis, Karine. *Culture et diversité : initiation à l'anthropologie*. Québec : Éditions CEC, 2006, 196 pages.

Thèse Canada. Collection de Bibliothèque et Archives du Canada. Direction de la société et expression culturelle, Gatineau, Québec. <http://www.collectionscanada.gc.ca/thesescanada/index-f.html> (consulté en janvier 2010).

Trudel, François, Paul Carest et Yvan Breton (sous la dir.). 1995. *La construction de l'anthropologie québécoise : mélanges offerts à Marc-Adélard Tremblay à l'occasion du 25e anniversaire du département d'anthropologie de l'Université Laval*. Québec : Presses de l'Université Laval, 1995, 472 pages.

Tufféry, stéphane. *Data Mining et statistique décisionnelle: L'intelligence des données*. Paris : Technip, 2010, 705 pages.

Tuite, Kevin. Au-delà du Stammbaum: Théories modernes du changement linguistique. *Anthropologie et sociétés*, Volume 23, Numéro 3, 1999, p. 15-52.

WordStat 6: *Content Analysis Module for QDA Miner & SimStat*. User's guide. 1998-2010. Rovalis Research.
< http://www.provalisresearch.com/Documents/WordStat6.pdf> (consulté en janvier 2011)

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager. T. D. Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, vol. 8, no. 8, 2011, p. 665-670.
<http://pilab.colorado.edu/publications/Yarkoni_NatureMethods_2011.pdf> (Consulté en février 2012).

Yifan Hu. Efficient and High Quality Force-Directed Graph Drawing. *The Mathematica Journal*, vol. 10, no. 1, 2006, p. 37-71.

Zouaq, Amal. *Une approche d'ingénierie ontologique pour l'acquisition et l'exploitation des connaissances à partir de documents textuels : Vers des objets de connaissances et d'apprentissage*. Thèse de doctorat. Montréal : Presses de l'Université de Montréal, 2007, 259 pages.