UNIVERSITÉ DU QUÉBEC À MONTRÉAL

LA SÉMIOTIQUE COMPUTATIONNELLE ENTRE NARRATIVITÉ ET APPRENTISSAGE AUTOMATIQUE : UNE DÉMONSTRATION DE FAISABILITÉ SUR UN CORPUS JOURNALISTIQUE À PROPOS DU *PRINTEMPS ÉRABLE*

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN SÉMIOLOGIE

PAR

DAVIDE PULIZZOTTO

DÉCEMBRE 2019

REMERCIEMENTS

La thèse n'est pas seulement un travail de recherche présenté dans un texte argumentatif de plusieurs pages. C'est un chemin ponctué de succès et de défaites et, surtout, de vieillies et de nouvelles relations humaines. Autrement, cette thèse n'existerait pas.

Je tiens ainsi à remercier tous ceux qui ont rendu possible le franchissement de la ligne d'arrivée de ce parcours. Je veux avant tout commencer avec mes directeurs. D'abord, Jean-Guy Meunier est, sans doute, le meilleur mentor qu'un doctorant peut espérer, car avec lui, on apprend un métier. Sa générosité intellectuelle et sa manière de gérer les ressources humaines qui collaborent avec lui facilitent le développement des capacités, des compétences et des attitudes de chacun. Je ne finirai jamais de le remercier pour tout l'apprentissage que j'ai pu faire grâce à lui. Je lui dois une grande partie de la personne que je suis maintenant. Aussi, Louis Hébert a été un soutien exceptionnel dans plusieurs phases du travail de recherche, grâce à ses précieux conseils et ses réflexions. Son attention et son dévouement à son rôle de formateur est admirable. Il a été toujours répondu présent lorsque j'en avais besoin. Son travail professionnel et sa chaleur humaine ont été un guide nécessaire pour l'avancement de ma recherche. Puis, mes collègues du LANCI-UQAM (Laboratoire d'analyse cognitive de l'information) ont été une base fondamentale pour l'apprentissage et l'avancement de cette recherche. Je n'aurais jamais pu écrire cette thèse sans les interminables discussions et les collaborations avec Jean-François Chartier, dont sa générosité et sa qualité en tant que chercheur ont été déterminantes pour moi. Ce

projet n'aurait pu avancer sans la collaboration, le soutien et l'amitié de José López González, de Louis Chartrand, de Francis Lareau et de Maxime Sainte-Marie.

Il est également important pour moi de remercier les membres de ma famille, en commençant pas mes beaux-parents, Claudine et Guy, qui ont été un soutien formidable pour l'écriture de la thèse. Sans les heures passées à corriger la langue et à m'émettre leurs commentaires judicieux, le texte de cette thèse n'aurait pas pu être lisible. Leur générosité à mon égard est remarquable et je leur en suis très reconnaissant. Je remercie aussi mes parents, Pietro et Filippa, ainsi que mon frère et ma sœur, Alessio et Claudia, qui, même à distance, ont su me soutenir, surtout dans les moments plus difficiles.

Le plus grand remerciement va à ma conjointe, Hélène, qui a su conserver tenacement sa patience tout au long de ces années. Sa générosité, sa confiance envers moi, son dynamisme et sa détermination, tout deux contagieux, sa capacité d'écoute ainsi que son amour m'ont permis de continuer ce projet et d'arriver à la fin de ce parcours incroyable. Je ne serai pas Davide sans sa présence. Je t'aime.

TABLE DES MATIÈRES

LISTE DES FIGURES viii

LISTE DES TABLEAUX xiii

LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES xvi

RÉSUMÉ xix

INTRODUCTION 1

CHAPITRE I PROBLÉMATIQUE 12

1.1 La sémiotique computationelle 12

1.1.1 L'ordinateur comme « machine sémiotique » et *semiosis* computable 16

1.1.2 L'ordinateur comme artefact 19

1.1.3 L'ordinateur comme outil au service de l'analyse sémiotique 21

1.1.4 La vocation empirique de la sémiotique 25

1.2 Médias, texte et sémiotique 26

1.2.1 Les « macrostructures » de la production textuelle 32

1.3 L'analyse « quantitative » du texte journalistique 38

1.3.1 News mining 41

1.4 Un champ inexploré : le croisement de l'analyse sémiotique et du news mining 43

1.5 Question de recherche 46

1.5.1 Hypothèses 48

1.6 Les objectifs poursuivis 53

1.7 Le cas d'étude : le printemps érable 54

1.7.1 Description des événements principaux 54

1.7.2 La couverture médiatique 61
1.7.3 Quelques corpus sur le printemps érable 62

CHAPITRE II CADRE THÉORIQUE 69

2.1 L'approche sémio-computationnelle 70

2.1.1 Les macrostructures 71
2.1.2 Le cadre épistémologique d'intersection entre le modèle computationnel et le modèle sémiotique 79

2.2 Le paradigme sémio-narratif 83

2.2.1 Le *parcours génératif du sens* : un métalangage formel 86
2.2.2 Les structures sémio-narratives 88
2.2.3 L'analyse du texte au moyen du schéma actantiel 98
2.2.4 L'aspect cognitif du modèle sémio-narratif 101

2.3 Le paradigme computationnel 104

2.3.1 La sémantique vectorielle 106
2.3.2 L'apprentissage automatique et le *clustering* 111

CHAPITRE III MÉTHODE 125

3.1 Définition de corpus 129

3.2 Critères de constitution du corpus 135

3.2.1 Orientation 135
3.2.2 Pertinence 138
3.2.3 Cohérence ou homogénéité 141
3.2.4 Hétérogénéité 142
3.2.5 Exhaustivité 143
3.2.6 Réprésentativité 146

3.3 Prétraitement d'un corpus 150

3.3.1 Annotation automatique des textes 152
3.3.2 Représentation vectorielle du texte 159

3.4 Filtrage 166

3.4.1 Mots 169
3.4.2 Documents 170

3.5 Définition du corpus de travail 174

3.6 Regroupement de documents similaires 176

3.6.1 K-moyennes 176
3.6.2 Évaluation des partitions 182

3.7 Annotation 186

3.7.1 Principes de base 187
3.7.2 Détails de la mise en place 189
3.7.3 Évaluation de la chaîne de traitement 190
3.7.4 Détermination de l'échantillon 191

CHAPITRE IV EXPÉRIMENTATIONS 195

4.1 Le corpus 196

4.1.1 Protocole de constitution du corpus 198
4.1.2 Moissonage 200
4.1.3 Description du corpus 201

4.2 Le prétraitement 210

4.2.1 Annotation automatique 210
4.2.2 Représentation vectorielle 217
4.2.3 Filtrage 220

4.3 La définition du corpus de travail 228

4.4 Le regroupement des documents similaires 228

4.4.1 Initialisation de l'algorithme 229
4.4.2 Choix du paramètre *k* 230

4.5 L'annotation 237

4.5.1 Sélection d'un échantillon représentatif 237
4.5.2 Annotation des échantillons 240
4.5.3 Résumé résultats annotations 241

CHAPITRE V RÉSULTATS 247

5.1 Groupes de macrostructures similaires 251

5.1.1 Négociations : théâtre de la polémique 252
5.1.2 Charest versus Marois : corps à corps 266
5.1.3 Loi spéciale n. 12 : L'objet magique 276
5.1.4 Manifestations : Police versus Étudiants 283
5.1.5 Élections : l'objet de valeur 292
5.1.6 Cégep : le retour en classe 297
5.1.7 Quelle est votre opinion? 304

5.2 Groupes de macrostructures distinctives 314
5.2.1 Métro 316
5.2.2 Journal de Montréal 321
5.2.3 Journal de Québec 327
5.2.4 La Presse 338
5.2.5 Le Soleil 343
5.2.6 Le Devoir 348
5.2.7 Radio-Canada 356

CHAPITRE VI DISCUSSION 359
6.1 Réflexions sur les résultats 360
6.2 La méthode et ses implications sémiotiques 364
6.2.1 Tokenisation 365
6.2.2 Lemmatisation 369
6.2.3 Segmentation 372
6.2.4 Le paradigme distributionnel 374
6.2.5 Le filtrage des caractéristiques 391
6.2.6 Le clustering 392
6.2.7 Annotation 395

CONCLUSION 401

ANNEXE A DÉTAILS SUR LE CADRE PROBABILISTE À LA BASE DE L'ETIQUETAGE MORPHOSYNTAXIQUE 410

ANNEXE B LISTE DES ANNOTATIONS MORPHOSYNTAXIQUES UTILISÉE PAR TREETAGGER 414

ANNEXE C DÉTAILS DE L'ALGORITHME DBSCAN 416

ANNEXE D DÉTAILS SUR L'IDENTIFIANT UNIQUE DES ARTICLES ANNOTÉS 417

RÉFÉRENCES 419

LISTE DES FIGURES

Figure		Page
1.1	Hexagone définissant l'interdisciplinarité de l'analyse de texte assistée par ordinateur	25
1.2	Évolution hebdomadaire par quotidien du contenu rédactionnel consacré à la crise	63
2.1	Loi de Zipf	81
2.2	Le parcours génératif du sens	89
2.3	Carré sémiotique	90
2.4	Le schéma narratif canonique	93
2.5	Schéma actantiel	95
2.6	Articulation de l'actant sur le carré sémiotique	98
2.7	Représentation bidimensionnelle d'une partition à trois clusters	116
2.8	Représentation graphique d'un clustering de type hiérarchique	119
3.1	Modèle de communication	163
3.2	Modèle de la communication et ses six fonctions	163
3.3	Détection de données aberrantes par clustering	167

3.4 L'algorithme DBscan 168

4.1 Évaluation de l'archive 192

4.2 Distribution des 9 603 articles dans les sept sources d'information 196

4.3 Évaluation du corpus de référence 203

4.4 Exemple arbre de décision pour *Treetagger* 204

4.5 Exemple de probabilités pour les suffixes de la langue anglaise 205

4.6 Procédure de filtrage des documents 216

4.7 Évaluation du corpus d'étude 220

4.8 Indices en soutien du choix du paramètre k pour le Journal de Montréal ... 227

4.9 Indices en soutien du choix du paramètre k pour le journal Le Soleil 227

4.10 Indices en soutien du choix du paramètre k pour le Journal de Québec 228

4.11 Indices en soutien du choix du paramètre k pour le journal La Presse 228

4.12 Indices en soutien du choix du paramètre k pour le journal Le Devoir 229

4.13 Indices en soutien du choix du paramètre *k* pour le site web *Radio-Canada* 229

4.14 Indices en soutien du choix du paramètre *k* pour le journal *Métro* 230

4.15 Résumé du nombre de cluster et d'articles retenus pour la phase d'annotation par source d'information 232

5.1 Réseau global de tous les clusters créés dans la phase de regroupement automatique des documents similaires 242

5.2 Représentation graphique de tous les clusters avec un filtre des liens fixé à 0,77 et l'application de l'algorithme Force Atlas 2 243

5.3 Détail de la figure 5.2. Encerclé en rouge les 11 clusters sélectionnés 245

5.4 Représentation graphique de l'agglomérat des 11 clusters centraux..... 246

5.5 Détail de la figure 5.2. Au centre, l'agglomérat traitant la confrontation entre M. Charest et M.me Marois..... 258

5.6 Représentation graphique de l'agglomérat traitant la confrontation entre M. Charest et M.me Marois..... 259

5.7 Représentation graphique de l'agglomérat traitant le projet de loi 12..... 268

5.8 Représentation graphique de l'agglomérat traitant les manifestations et les affrontements entre police et manifestant 275

5.9 Représentation graphique de l'agglomérat traitant les élections avec le seuil du filtre des liens fixé à 0,77 284

5.10 Représentation graphique de l'agglomérat traitant les élections avec le seuil du filtre des liens fixé à 0,76 284

5.11 Représentation graphique de l'agglomérat traitant les élections avec le seuil du filtre des liens fixé à 0,7 289

5.12 Représentation graphique de l'agglomérat traitant les élections avec le seuil du filtre des liens fixé à 0,69 290

5.13 Représentation graphique de l'agglomérat traitant les articles d'opinion ... 296

5.14 Représentation graphique des 117 clusters qui ont été annotés avec le seuil du filtre des liens fixé à 0.6 306

5.15 Détail de la figure 5.14 308

5.16 Représentation graphique des liens du cluster ME_6 309

5.17 Détail de la figure 5.14 313

5.18 Nouage de mots plus fréquents du cluster JM_43 314

5.19 Détail de la figure 5.14. Les clusters JQ_4 et JQ_6 sont représentés à gauche, le cluster JQ_14 à droite. La figure montre leur isolement 318

5.20 Nouage de mots plus fréquents du cluster JQ_4 319

5.21 Nouage de mots plus fréquents du cluster JQ_14 322

5.22 Nouage de mots plus fréquents du cluster JQ_6 325

5.23 Détails de la figure 5.14. Les clusters PR_41 (a), PR_21 (b), PR_23 (c) et PR_29 (d) sont représentés 329

5.24 Détail de la figure 5.14. Les clusters SO_19 et SO_28 sont représentés avec leurs connexions 334

5.25 Détail de la figure 5.14. Les clusters DE_4,DE_6et DE_46 sont représentés 339

5.26 Détail de la figure 5.14. Le cluster DE_4 et ses connexions 339

5.27 Détail de la figure 5.14. Les clustersRC_16 et RC_28 sont représentés 346

6.1 Expression et contenu selon Hjelmslev 358

6.2 L'exemple de la feuille de papier 368

6.3 Illustration de la dépendance intrasystémique de la valeur linguistique 369

6.4 Jeu de données à deux dimensions représentées dans un espace plan 375

6.5 Représentation d'un jeu de données en trois dimensions 376

6.6 Histogrammes des distances de toutes les paires possibles parmi 100 points échantillonnés de manière uniforme 377

6.7 Genèse narrative comme une structure récursive 386

6.8 Génération d'une trame narrative simple par un automate fini 386

LISTE DES TABLEAUX

Tableau		Page
2.1	Représentation du modèle actantiel sous forme de tableau	97
2.2	Les sous-catégories actantielles	98
3.1	Représentation vectorielle de l'énoncé 1	160
3.2	Exemple d'un programme narratif	182
3.3	Exemple d'un programme narratif complété par son opposé	183
4.1	Résumé des indices de dispersion	196
4.2	Résumé des fréquences des signatures par source d'information	197
4.3	Les premiers dix journalistes par source d'information en relation à la fréquence d'apparition	198
4.4	Résumé des catégories d'articles et leurs fréquences	199
4.5	Catégories d'articles pour chaque source d'information	201
4.6	Description des fréquences des mots dans les articles	201
4.7	Détails des parties du discours annotées	206
4.8	Nombre total par partie du discours en ordre décroissant	206

4.9 Totaux des annotations par source d'information 207

4.10 Résumé du ration token/type (annotation/lemme)..... 209

4.11 Les 20 mots les plus fréquents du corpus, en excluant le mot « étudiant » 210

4.12 Les 20 mots les plus fréquents par source d'information 211

4.13 Variabilité du vocabulaire..... 214

4.14 Résultats filtrage par règles 215

4.15 Résultats du test SVM 219

4.16 Valeur max_k pour chaque source d'information..... 224

4.17 Valeur k choisie pour chaque sous-corpus..... 226

4.18 Exemples de codes des annotations, avec leurs racines sémantiques, les suffixes syntaxiques et les rôles actantiel correspondant 233

4.19 Les 20 codes les plus fréquents du processus de lecture et d'analyse des articles..... 235

4.20 Les 20 codes les plus fréquents pour le journal *Metro* et le *Journal de Montréal* 236

4.21 Les 20 codes les plus fréquents pour le *Journal de Québec* et le site-web *Radio-Canada*..... 237

4.22 Les 20 codes les plus fréquents pour *Le Devoir* et *La Presse* 238

4.23 Les 20 codes les plus fréquents pour *Le Soleil*..... 238

5.1 Programme narratif PN1..... 246

5.2 Programme narratif PN2..... 247

5.3 Programme narratif PN3 259

5.4 Taille des clusters de l'agglomérat sur la loi 12 268

5.5 Programme narratif PN4 269

5.6 Programme narratif PN5 270

5.7 Programme narratif PN6 275

5.8 Programme narratif PN7 285

5.9 Programme narratif PN8 290

5.10 Taille de chaque cluster avec son rang dans la partition du journal correspondant 296

5.11 Exemples de titres similaires pour les articles pareils entre le *Journal de Québec* et le *Journal de Montréal* 312

6.1 Schéma actantiel de l'idéologie marxiste selon Greimas 351

6.2 Exemple d'une matrice binaire pour la phrase « La linguistique est enseignée à l'université », suivi d'une deuxième phrase 372

LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES

ASCII	American Standard Code for Information Interchange
ATO	Analyse de texte assistée par ordinateur
CADEUL	Confédération des associations d'étudiants et d'étudiantes de l'Université de Laval
CAQ	Coalition avenir Québec
CDJ-ADQ	Commission des jeunes du parti Action démocratique du Québec
CLASSE	Coalition large de l'Association pour une solidarité syndicale étudiante
CRÉPUQ	Conférence des recteurs et des principaux des universités du Québec
CSN	Confédération des syndicats nationaux
CSQ	Centrale des syndicats du Québec
DBscan	Density-Based Spatial Clustering of Applications with Noise
DE	Le Devoir
FECQ	Fédération Étudiante Collégiale du Québec
FEUQ	Fédération Étudiante Universitaire du Québec
FNEEQ	Fédération nationale des enseignantes et enseignants du Québec

FTQ	Fédération des travailleurs et travailleuses du Québec
JM	Le Journal de Montréal
JQ	Le Journal de Québec
ME	Métro (Montréal)
PLQ	Parti Libéral du Québec
PN	Programme narrative
PQ	Parti Québécois
PR	La Presse
QS	Québec solidaire
RC	Radio-Canada
SO	Le Soleil
SPVM	Service de police de la ville de Montréal
SPVQ	Service de police de la ville de Québec
SPVG	Service de police de la ville de Gatineau
SQ	Sûreté du Québec
SQL	Structured Query Language
SVM	Support-Vector Machines (algorithme machine à vecteur de support)
TACEQ	Table de concertation étudiante du Québec

TF-IDF	Term Frequency–Inverse Document Frequency
TVQ	Taxe de vente du Québec
UQAM	Université du Québec à Montréal
UTF-8	Unicode Transformation Format, 8 bit

RÉSUMÉ

La *révolution numérique* est en train de modifier profondément les pratiques des sciences humaines et sociales. Dans ce nouveau parcours, la sémiotique joue un rôle prépondérant en raison de ses spécificités qui la rendent compatible avec le cadre computationnel requis par le *numérique. La sémiotique computationnelle* est un champ de recherche qui contribue au développement des *humanités numériques*.

Mais que signifie numérique ou computationnel pour les *humanités*? Cette thèse présente une piste de réponse à ce genre de questionnement à l'aide d'une *démonstration de faisabilité*. Pour ce faire, une pratique pertinente dans ce contexte est identifiée, soit l'*analyse de texte assistée par ordinateur* (ATO). En effet, l'analyse de texte est l'une des pratiques les plus répandues en humanités et constitue également l'un des objets d'étude privilégiés de la sémiotique. Par la suite, un cas d'étude est identifié dans le domaine des médias et, plus particulièrement, dans le champ de l'analyse de presse. Le *printemps érable*, le mouvement estudiantin québécois de 2012, a été choisi en raison de sa large couverture journalistique, ce qui garantit la possibilité d'obtenir un grand nombre d'articles de presse traitant ce phénomène socio-politique.

L'analyse de ce corpus journalistique a été effectuée au moyen d'une *chaîne de traitement* construite avec le support d'outils issus du *traitement automatique du langage naturel* et de l'*apprentissage automatique*. Pour ce faire, une approche théorique d'intersection entre la sémiotique et l'intelligence artificielle a été mise au point et ceci, afin de fournir une assistance computationnelle à une théorie sémiotique spécifique, soit la « sémiotique narrative ». Ainsi, une technique de *clustering* a été choisie pour assister l'analyse narrative du texte journalistique. La chaîne de traitement est évaluée par une phase d'annotation qui constitue une partie importante de la méthode.

Les résultats montrent que la chaîne de traitement est pertinente pour l'exploration d'un corpus de grande taille. En effet, de grands agglomérats de macrostructures sémantiques similaires ont été identifiés de manière transversale aux différentes sources d'information. On identifie, par exemple, que le discours électoraliste est un

des plus fréquents du corpus, ainsi que les débat sur les négociations entre les étudiants et le gouvernement. On identifie aussi certaines spécificités selon le journal, comme la mise en valeur du concept d'éducation par le journal *Le Devoir,* ou des concepts économiques par *Le Journal de Montréal*. Chaque résultat obtenu constitue ainsi une piste de recherche pour l'approfondissement de la couverture journalistique sur le printemps érable.

Ce travail de recherche contribue au transfert de connaissance de l'informatique vers les sciences humaines et, plus particulièrement, vers la sémiotique. Le parcours de la thèse tente de faciliter la compréhension des différents algorithmes et outils de la chaîne de traitement. Elle contribue également à la compréhension d'un des volets de la sémiotique computationnelle, celui qui voit l'ordinateur comme « outils au service de l'analyse sémiotique ». Elle propose aussi une perspective particulière pour l'agencement de la sémiotique et de l'intelligence artificielle et, donc, pour le développement de la sémiotique computationnelle. De façon plus spécifique, la recherche souligne la nécessité de décomposer les pratiques de la sémiotique en tâches simples et formelles, afin que leur conversion en tâches computables soit réalisée.

Enfin, les humanités n'ont pas de raison de craindre la révolution numérique amenée par le développement de l'intelligence artificielle, mais elle devrait plutôt en profiter pour augmenter ses capacités d'analyse des phénomènes humains et sociaux.

Mots clés :

Humanités numériques, sémiotique computationnelle, analyse de texte assistée par ordinateur, ATO, traitement automatique du langage naturel, apprentissage automatique, fouille de texte, text mining, analyse des données, clustering, analyse narrative, texte journalistique, analyse de presse, printemps érable, mouvement étudiant québécois de 2012

INTRODUCTION

La sémiotique est une discipline intrinsèquement interdisciplinaire et son histoire le prouve. En effet, elle a commencé son parcours en concomitance avec une autre discipline, la linguistique. Pendant longtemps, les histoires de ces deux disciplines se sont entrecroisées alors que la linguistique englobait la sémiotique en tant que champ d'étude. Ce n'est qu'à partir des années 60 que la sémiotique est devenue une discipline autonome et distincte de la linguistique et ce, tout en conservant sa nature interdisciplinaire. Avec le temps, la sémiotique a redéfini ses approches, ses méthodologies et ses objets d'étude spécifiques, mais elle n'a pas cessé de produire des « contaminations disciplinaires » de toute sorte. La linguistique et la philosophie, par exemple, n'ont jamais cessé d'interagir avec les recherches sémiotiques. Les travaux d'Umberto Eco en témoignent. Ses premières œuvres, comme « *Opera aperta* » (Eco, 1962, trad. fra. *L'œuvre ouverte*, 1965), laissent sans aucun doute transparaître les études en esthétique réalisées pendant son doctorat. Ainsi, les disciplines comme l'anthropologie et les études littéraires ont contribué de manière importante au développement de la sémiotique, en précisant à celle-ci des champs d'application et en lui fournissant des réflexions théoriques et méthodologiques. De plus, lorsque la sémiotique a défini davantage ses méthodologies, elle est devenue « une méthode » utilisée par une panoplie d'études en communication, marketing, analyse du discours, littérature, musique, histoire de l'art, photographie, cinéma, etc. Dans ce contexte, elle a lié son développement à celui des disciplines dans lesquelles elle opère. Ainsi, aujourd'hui comme hier, son originalité et sa pertinence dépendent de plus en plus de sa capacité à être interdisciplinaire.

La « révolution numérique » qui est en train de modifier profondément les sciences humaines et sociales ne peut pas être ignorée par la sémiotique. De plus en plus, l'interaction entre les sciences humaines et sociales avec les domaines de l'informatique illustre la pertinence de définir un nouveau domaine d'étude, soit celui des *humanités numériques*. Ces dernières « recouvrent un ensemble de pratiques de recherche à l'intersection des technologies numériques et des différentes disciplines des sciences humaines » (Dacos et Mounier, 2015, p. 7). La sémiotique est partie intégrante de cette révolution et y contribue par le développement d'un nouveau champ d'étude, soit la *sémiotique computationnelle*. En réalité, ce genre d'études est plus ancien que la définition des humanités numériques (autour des années 2000) et compte déjà à son actif trois volets de recherche différents dont nous traitons au premier chapitre. La sémiotique computationnelle possède sa propre autonomie par rapport aux humanités numériques et leur impose sa pertinence. En effet, la sémiotique, discipline qui étudie les systèmes et les processus de la signification, travaille sur les artefacts sémiotiques qui constituent les objets d'étude des différentes *humanités*. Elle leur fournit ainsi des théories et des méthodes pour l'analyse et l'interprétation de leurs propres objets d'étude. Il ne fait aucun doute que la sémiotique computationnelle agit de la même manière et rejoint ainsi les objectifs de recherche spécifiques aux humanités numériques.

Le rôle de la sémiotique dans le domaine des humanités numériques est très important, car elle possède les compétences pour approfondir le croisement entre sciences humaines et informatique. Ceci est surtout dû au fait que la sémiotique constitue un

> cadre formel et logique qui ressemble aux mathématiques puisqu'elle contient les outils pour exécuter des opérations logiques sur elle-même (en tant que système de signes), ainsi que pour utiliser ces outils pour analyser le mouvement, le changement ou la perturbation dans d'autres systèmes. (P. K. Manning, 2004, p. 586) [Notre traduction]

En effet, comme les mathématiques, la sémiotique opère sur la construction d'un métalangage formel qui exécute des opérations sur les sémiotiques pour les décrire, les reproduire ou les interpréter. Sa pertinence dérive donc essentiellement de ses modèles formels et logiques, très bien adaptés au contexte interdisciplinaire mis en place par les humanités numériques. En d'autres termes, les métalangages ou les métasémiotiques qui sont élaborés pour décrire les sémiotiques ont une nature formelle qui se conforme bien au cadre théorique du numérique. Ce point sera approfondi au chapitre II.

Mais, qu'est-ce que le *numérique*? Qu'est-ce que l'adjectif «numérique» signifie dans la locution «humanités numériques»? Ces mêmes questions se posent pour la locution « sémiotique computationnelle » : qu'est-ce que le *computationnel* en sémiotique? Nous tenterons de répondre à ces questions tout au long de ce travail de recherche. Pour commencer, soulignons un premier aspect qui, bien qu'il puisse sembler trivial, mérite selon nous une plus large réflexion de la part des chercheurs en humanités. L'adjectif numérique ou computationnel ne peut pas se limiter à qualifier des pratiques de recherche qui utilisent une technologie numérique dans le cadre des sciences humaines et sociales. L'interaction avec le numérique est un phénomène plus profond, qui concerne les racines épistémologiques des sciences humaines. Pour explorer davantage ce que le numérique implique, nous proposons de l'appréhender par le biais d'un terme qui a été défini dans le « *Dictionnaire raisonné de la théorie du langage* » (Greimas et Courtés, 1979) et qui appartient au groupe des concepts les plus négligés par la sémiotique, soit celui d'*a lgorithme*. En effet, ce concept constitue une porte d'entrée pour une compréhension approfondie du numérique et du computationnel. Ces auteurs nous l'offrent déjà prêt à l'emploi, dans un contexte de sciences humaines et, plus spécifiquement, en sémiotique.

> Par **algorithme**, on entend la prescription d'un ordre déterminé dans l'exécution d'un ensemble d'instructions explicites en vue de la solution d'un

> certain type de problème donné. Dans la métasémiotique scientifique, qui se donne pour tâche de représenter le fonctionnement d'une sémiotique sous la forme d'un système de règles, l'algorithme correspond à un savoir-faire syntagmatique, susceptible de programmer, sous forme d'instructions, l'application de règles appropriées. (Greimas et Courtés, 1979, p. 12)

Cette définition d'algorithme peut servir de base pour la définition des termes « numérique » et « computationnel » dans un contexte de recherche en sciences humaines, car elle élargit leurs divers sens en intégrant les pratiques de travail déjà en place en science humaines. En effet, cette définition met en relief un concept très important, soit celui de programme, considéré comme une suite d'opérations formelles, bien définies et ayant pour but d'accomplir une tâche précise. Pour la sémiotique, l'algorithme correspond aux procédures formelles ou au « savoir-faire syntagmatique », mis en place par les métasémiotiques scientifiques. Ainsi, les métalangages sont associés aux programmes informatiques puisqu'ils décrivent les sémiotiques par des suites de règles cohérentes, homogènes et surtout, formelles. En ce sens, la sémiotique comporte plusieurs éléments qui mènent vers cette perspective, comme la « sémiotique narrative » :

> Il est évident que la présentation algorithmique des suites de règles ne peut se faire que progressivement : l'organisation algorithmique ne peut être conférée d'abord qu'à certaines procédures d'analyse. Ainsi, en sémiotique narrative, les programmes narratifs complexes, par exemple, sont déjà susceptibles de recevoir une formulation algorithmique. C'est dans la même perspective que nous avons proposé de considérer comme un **algorithme de transformation** une suite ordonnée d'opérations permettant de passer de l'état initial à l'état final d'un récit fermé. (*ibid.*)

Plusieurs éléments intéressants se retrouvent dans cet extrait. D'abord, la conscience que l'organisation algorithmique d'un phénomène s'élabore progressivement et à partir de certaines procédures d'analyse. En d'autres termes, certaines opérations et procédures qui sont exécutées dans un contexte d'analyse peuvent commencer à être organisées selon une *suite de règles* ou *algorithme*. De plus, Greimas et Courtés

considèrent les metasémiotiques comme des représentations conceptuelles et organisées de phénomènes sémiotiques, qui peuvent être converties en procédures algorithmiques. En fait, ils identifient déjà un métalangage « susceptible de recevoir une formulation algorithmique », soit les programmes narratifs. Ainsi, un phénomène sémiotique est décrit par une métasémiotique laquelle, lorsqu'elle respecte un certain degré de formalisme en ressemblant le plus possible à une suite de règles, peut « recevoir une formulation algorithmique ». Sous cet angle, la sémiotique greimassienne peut déjà constituer un exemple de sémiotique computationelle.

Le numérique n'est donc pas seulement l'utilisation de la technologie, ou la description d'une technologie et de son usage (ex. l'ordinateur, Facebook, Twitter, etc.). Cette définition par Greimas de l'algorithme rapproche la sémiotique et les humanités à la discipline-mère de cette révolution numérique, soit l'*intelligence artificielle*. Cette discipline a beaucoup en commun avec les humanités, car sa mission est la *simulation* des compétences humaines. La métaphore de l'ordinateur-cerveau, surtout diffusée dans un des domaines qui ont participé au développement de l'intelligence artificielle, soit les sciences cognitives, est un symbole de cette perspective de simulation des compétences humaines. En effet, la définition la plus largement acceptée d'intelligence artificielle est la suivante : « Artificial Intelligence is the study of how to make computers do things at which, at the moment, people are better » (Rich *et al.*, 2009, p. 3). Elle souligne que c'est le transfert des fonctions cognitives vers l'ordinateur qui constitue le « Saint Graal » de cette discipline.

L'intelligence artificielle se base sur des phénomènes humains pour créer des modèles formels et computables. En faisant un certain saut théorique, on peut dire qu'elle requiert une représentation conceptuelle, ou une théorie, des fonctions cognitives qu'elle veut simuler avec une machine. D'une part, la sémiotique offre certainement des modèles pour la compréhension des dynamiques sémiotiques

puisqu'elle essaie de décrire par des métalangages formels et « susceptibles de recevoir une formulation algorithmique » les dynamiques sémiotiques. D'autre part, l'intelligence artificielle tente de faire la même chose avec des phénomènes sémio-cognitifs, en ajoutant un niveau de formalisme algorithmique et d'implémentation technologique. Ainsi, les métalangages formels de la sémiotique constituent la richesse même de sa relation avec l'intelligence artificielle. Certains auteurs ont explicitement affirmé, par exemple, que « la sémiotique a inventé des modèles théoriques qui peuvent servir à l'intelligence artificielle » (Thérien, 1989). Pour utiliser un terme du langage spécifique à l'informatique, on pourrait dire que la sémiotique a le rôle de mettre en place le *pseudo-code* qui constitue la représentation conceptuelle et formelle des fonctions sémio-cognitives.

Enfin, cette révolution numérique, alimentée par les développements de l'intelligence artificielle, est en train de bouleverser la société, le monde professionnel et la recherche académique. Deux derniers éléments doivent être introduits pour compléter cette description sommaire de cette révolution, soit le *big data* et l'*apprentissage automatique*. La locution *big data* désigne le phénomène d'accumulation massive de données qui, une fois analysées, peuvent fournir des connaissances précieuses. Le *big data* est lié au développement de la *science des données*, c'est-à-dire au développement des techniques d'analyse statistique en interaction avec l'apprentissage automatique. Ce dernier est un des programmes de l'intelligence artificielle et désigne des techniques qui simulent l'apprentissage humain afin de résoudre des tâches spécifiques. Ces techniques utilisent des données préalablement accumulées pour *apprendre à exécuter une tâche*. L'apprentissage automatique s'est développé parallèlement à l'apprentissage statistique et les deux sont désormais considérés comme des synonymes. En simplifiant, cette interaction entre statistique et informatique a donné lieu à la science des données. Ainsi, la révolution numérique est également liée à l'accroissement des possibilités d'accumulation de données, comme

la numérisation massive de textes, et des possibilités de leur analyse par l'apprentissage automatique.

La science des données apporte donc de nouveaux éléments aux humanités, et plus particulièrement, l'emploi d'une nouvelle approche méthodologique. Cette dernière est celle des études axées sur les données ou *data-driven.* Cette approche méthodologique mène à des réflexions épistémologiques qui tendent vers un réexamen de l'approche empirique. En effet, le *big data* et l'approche *data-driven* qu'il met en place inaugurent une nouvelle ère de l'empirisme, puisque

> le volume des données accompagné par des techniques qui peuvent en révéler la connaissance intrinsèque, donne la possibilité aux données de parler par elles-mêmes sans les contraintes d'une théorie qui leur est imposée. (Kitchin, 2014, p. 3) [Notre traduction]

L'approche axée sur les données comporte ses propres spécificités, qui la caractérisent comme un type d'approche empiriste qui se combine mieux avec les humanités. La principale caractéristique qui facilite cette combinaison est sa capacité à développer un *bricolage* d'approche et de méthodologies :

> Contrairement aux nouvelles formes d'empirisme, les approches axées sur les données cherchent à respecter les principes de la méthode scientifique, mais elles sont plus ouvertes à l'utilisation d'une combinaison hybride d'approches abductives, inductives et déductives pour faire progresser la compréhension d'un phénomène. (Kitchin, 2014, p. 5) [Notre traduction]

En ce sens, les sciences des données et les approches axées sur les données sont compatibles avec les humanités, car elles utilisent souvent un mélange d'approches pour avancer dans la compréhension d'un phénomène.

Les éléments méthodologiques des humanités numériques et de la sémiotique computationnelle se clarifient davantage avec la description de la révolution

numérique en cours. Dans ce contexte, il manque toutefois une définition très importante, soit celle de *données pour les humanités*. Le mot « donnée » vient du latin *datum*, qui signifie « ce qui est donné », en renvoyant essentiellement à l'expérience perceptible. En effet, les données sont les éléments d'un phénomène qui peuvent être perçus ou décrits, permettant ainsi l'étude même du phénomène. Il est possible, par exemple, d'obtenir des données sous forme de textes, d'images ou d'un ensemble de faits observés, ou encore décrits, comme dans le cas des données psychosociales ou économiques. En ce sens, les humanités ont toujours manipulé des « données ». Les livres, les peintures, les films, en sont des exemples. Les méthodes et les pratiques des humanités sont souvent liées à l'analyse de données (Schöch, 2013). La numérisation massive qui est en cours produit des jeux de données de plus en plus intéressants pour les humanités, ce qui alimente la révolution numérique et souligne encore plus le fait que les humanités sont très compatibles avec la science des données. La révolution numérique est surtout constituée de l'accumulation de données et de l'arrivée de nouvelles techniques d'analyse dont certaines, il faut le dire, ne sont pas si récentes. Ceci apporte certainement de nouvelles réflexions épistémologiques et méthodologiques au sein des sciences humaines. Cependant, nous sommes d'avis qu'il n'est pas nécessaire de repenser la nature même des humanités ou de révolutionner toutes ses pratiques. Il suffit de mettre en valeur ce qui est déjà compatible avec cette « révolution ». Il est certain que des bouleversements plus profonds peuvent se produire, comme le suggèrent Boyd et Crawford (2012):

> Le big data recadre des questions clés sur la constitution des connaissances, les processus de recherche, la manière dont nous devons interagir avec les informations, la nature et la catégorisation de la réalité. . . Le big data délimite de nouveaux objets, méthodes de connaissance et définitions de la vie sociale. (Boyd et Crawford, 2012, p. 663) [Notre traduction]

Toutefois, nous estimons que le changement est surtout lié à l'augmentation des possibilités d'analyse avec le big data et les sciences des données. Le paradigme du

numérique introduit d'abord des nouvelles possibilités d'extraction de connaissances et ceci à plus grande échelle, ce qui permet de découvrir de « nouveaux observables » (Rastier, 2011). Il faut donc saisir la possibilité que la science des données offre aux humanités pour la découverte d'éléments qu'on n'aurait pas pu identifier autrement.

Ce n'est donc pas toute la recherche en sciences humaine qui doit évoluer. Dans la majorité des cas, la science des données peut fournir une *assistance* aux pratiques de recherche plus traditionnelles des humanités. L'*analyse de texte assistée par ordinateur* en est le meilleur exemple. Dans ce cadre, l'analyse de texte reçoit une assistance des techniques issues de la statistique et de l'informatique pour des études en sciences humaines et sociales. Elle constitue également un exemple de sémiotique computationnelle, car l'analyse de texte est un de ses objets privilégiés. L'analyse de texte assistée par ordinateur constitue le cœur de cette thèse et elle sera précisée dans le premier chapitre.

Dans un contexte de mise en valeur de la matérialité observable ou descriptible, du plan d'immanence d'un phénomène, du concept de données et de méthode empirique, la sémiotique prend toute sa place. Comme l'affirme Fabbri (2008) en reprenant les mots de Greimas, la sémiotique est une discipline à *vocation empirique*. Chaque projet sémiotique est composé de quatre niveaux qui, en principe, devraient être cohérents entre eux, soit les niveaux épistémologique, théorique, méthodologique et empirique. Ce dernier niveau constitue le matériau à partir duquel la sémiotique commence son projet de recherche. Avec la révolution numérique, la sémiotique doit affronter de nouveaux défis introduits par les nouvelles techniques offertes par la science des données, l'apprentissage automatique et l'accumulation massive de données. Ses niveaux sont donc affectés par cette révolution et le principal défi est de comprendre de quelle manière la sémiotique peut intégrer l'intelligence artificielle dans ses pratiques. Il faut donc établir le chemin de la sémiotique computationnelle.

Dans le cadre décrit précédemment, une série de questions surgissent. Dans le contexte de la sémiotique computationnelle, quelles sont les théories sémiotiques formelles qui peuvent recevoir une assistance par ordinateur? Comment cette assistance peut-elle se réaliser? Quel est son cadre épistémologique? Et quelle est son approche théorique? Enfin, quels sont les outils computationnels qui peuvent être mis au service de l'analyse sémiotique? Et, une fois ce travail fait, obtenons-nous des résultats intéressants avec l'utilisation de ces outils?

Pour répondre à ce genre de questions, la présente recherche propose un cas d'étude spécifique et la construction d'une chaîne de traitement pour l'analyse d'un type particulier de texte, soit le texte journalistique. Le cas d'étude choisi est le « printemps érable », le mouvement estudiantin québécois de 2012. La chaîne de traitement est construite avec des outils venant du traitement automatique du langage naturel et de l'apprentissage automatique. Elle constitue ainsi un exemple de méthode appartenant à la sémiotique computationnelle. L'approche théorique est composée de deux paradigmes, le computationnel et le sémiotique, que nous essayons d'intégrer dans la présente thèse. Le but est ainsi de proposer une *démonstration de faisabilité* d'une méthode de sémiotique computationnelle et de réfléchir sur la façon dont la sémiotique peut contribuer au développement de l'analyse de texte assistée par ordinateur.

La thèse est divisée en six chapitres. Le premier chapitre explicite la problématique. Le champ de la sémiotique computationnelle y est décrit dans les premiers paragraphes. Par la suite, le contexte spécifique de cette recherche, soit l'analyse du texte journalistique, est présenté. Le chapitre se conclut avec la formulation de la question de recherche et des hypothèses de recherche. Le deuxième chapitre est quant à lui consacré au cadre théorique. Celui-ci présente trois composantes : l'approche sémio-computationnelle, qui définit la manière dont l'intersection entre sémiotique et

computationnel est abordée dans ce travail, le paradigme sémiotique et le paradigme computationnel. Le troisième chapitre se concentre de son côté sur la description de la méthode. La première partie est ainsi dédiée à l'exemplification des données qui sont utilisées dans un cadre d'analyse de texte assistée par ordinateur, c'est-à-dire la constitution du corpus. Pour la sémiotique computationnelle, cette étape est cruciale. Les autres parties du chapitre définissent la chaîne de traitement qui a été construite pour cette étude. Elles comportent une phase de prétraitement du corpus, à l'aide des outils issus du traitement automatique du langage naturel, une phase de filtrage et de regroupement automatique, à l'aide des outils de l'apprentissage automatique; et enfin, une étape d'annotation à l'aide d'un outil informatique qui permet la production d'une base de données relationnelle pour maintenir une trace de la lecture et de l'analyse effectuée. Le quatrième chapitre décrit en détails les expérimentations et aborde pas à pas les différentes étapes de la méthode, en présentant les résultats obtenus pour chacune d'elles. Le cinquième chapitre présente les résultats obtenus sur les corpus. Les caractéristiques similaires et distinctives entre les différentes sources d'information retenues pour cette étude y sont mises en évidence. Le sixième et dernier chapitre soumet une série de réflexions sur les résultats obtenus et surtout, sur les aspects méthodologiques de cette étude.

CHAPITRE I

PROBLÉMATIQUE

Le premier chapitre explicite la problématique. D'abord, la sémiotique computationnelle est décrite plus en profondeur avec ses trois axes de recherche. Elle constitue le principal cadre de référence de la thèse. Par la suite, nous introduisons brièvement le champ d'application du présent travail, qui est l'étude des médias au moyen de l'analyse du texte journalistique. À l'intérieur de ce champ d'application, nous introduisons également le cadre méthodologique de cette thèse, qui est essentiellement celui de l'analyse quantitative du texte journalistique et de l'analyse de texte assistée par ordinateur. Ainsi, nous soulignons les croisements inexplorés entre la sémiotique et le *news mining*. Le chapitre continue avec l'explicitation de la question de recherche, des hypothèses et des objectifs poursuivis. Il se conclut avec une description du cas d'étude qui a été choisi pour ce travail, qui est le printemps érable, ainsi qu'avec une brève revue d'autres corpus qui ont fait l'objet du même type d'analyse.

1.1 La sémiotique computationelle

Dans cette thèse, deux disciplines s'entrecroisent, soit l'intelligence artificielle et la sémiotique. La méthode proposée met de l'avant une vision de partage et de transfert de connaissances entre les deux. Plus particulièrement, elle explore les potentialités

computationnelles de l'analyse sémiotique en identifiant certaines opérations cognitives « computables » et en proposant des solutions algorithmiques qui peuvent les exécuter. La possibilité de réaliser des opérations cognitives par une machine constitue la problématique de base de l'intelligence artificielle. La naissance de cette dernière, comme celle du cognitivisme, est enracinée dans le mouvement cybernétique des années 1940 et 1950 (Dupuy, 2009). C'est dans ce contexte que l'intérêt pour les mécanismes de la pensée humaine et la possibilité de les simuler à l'aide d'un ordinateur a émergé. La définition d'intelligence artificielle actuellement la plus répandue (Ertel, 2011) est celle de Elaine Rich (Rich *et al.*, 2009), qui souligne comment *le transfert des fonctions cognitives vers l'ordinateur* est effectivement le principal but de cette discipline :

> L'intelligence artificielle est l'étude des manières de permettre aux ordinateurs de faire des choses pour lesquelles, à l'heure actuelle, l'être humain est meilleur. (Rich *et al.*, 2009, p. 3) [Notre traduction]

Du point de vue de la sémiotique, cette définition conduit à l'identification des aspects computables des dynamiques sémiotiques. Si la sémiotique a comme but de comprendre les manières utilisées par l'être humain pour mettre les « choses » *en condition de signifier* (Marrone, 2001), la sémiotique computationnelle devrait donc avoir comme but la « simulation » des « conditions de signification ». Le substantif « simulation » devrait ici être appréhendé dans une acception large, car il doit représenter des expériences de recherche très différentes les unes des autres, dont seule une portion utilise effectivement l'informatique pour implémenter des dynamiques sémiotiques.

De nos jours, il est de plus en plus évident que le développement des outils et des méthodes issus de l'intelligence artificielle a un impact important sur la recherche en sciences humaines. L'intelligence artificielle est de plus en plus souvent impliquée

dans des programmes de recherches interdisciplinaires, auxquels des disciplines comme les sciences du langage, la sociologie, la psychologie, la philosophie, la communication participent. La sémiotique est également une des disciplines qui, de façon directe ou indirecte, contribue au développement de ce type de programme de recherche. En effet, le champ de la *sémiotique computationnelle* est déjà une réalité de grande valeur.

Compte tenu de leur grand nombre et de leur diversité, il est très difficile d'identifier et de classer les études pouvant faire partie de la sémiotique computationnelle. Pour des raisons de simplicité, la sémiotique computationnelle est ici subdivisée en trois grandes branches : l'étude de *l'ordinateur comme machine sémiotique* (Andersen, 1997; Etxeberria et Ibáñez, 1999; Fetzer, 2011; Ketner, 1988; Meunier, 1989, 2014; Nadin, 1977; Queiroz et Merrell, 2008; Raber et Budd, 2003; Rapaport, 2012); l'étude de *l'ordinateur comme artefact* (De Souza, 2005; Nadin, 2011; Stamper, 1973; Tanaka-Ishii, 2010; Zemanek, 1966); et l'étude de *l'ordinateur comme outil au service de la recherche sémiotique* (Bernard et Bohet, 2017; Compagno, 2018; Lebart *et al.*, 2003; Lebart et Salem, 1994b; Meunier, 2009; Meunier et Forest, 2009; Moretti, 2013; Pêcheux, 1969; Rastier, 2018, 2011; Rieger, 1997; Tanaka-Ishii, 2015; Valette, 2018). Cette classification est basée sur la manière d'entendre la relation avec l'« ordinateur ». Le mot « ordinateur » doit être appréhendé comme une métaphore décrivant des éléments de nature différente mais qui ont tous en commun la possibilité d'être manipulables par un appareil informatique. Cette classification ne peut pas être considérée définitive puisque ce domaine est en évolution constante.

Le premier de ces trois axes de la sémiotique computationnelle considère l'ordinateur comme une machine sémiotique, c'est-à-dire une machine qui manipule des signes. Dans cet axe, on considère la *semiosis* comme un « fait computable ».

La deuxième approche considère l'ordinateur comme un artefact. Un artefact est un objet construit par l'homme et qui lui sert pour exécuter des tâches. Ce concept (artefact) prend une grande importance en sémiotique pour la définition du concept de *culture*. Pour Rastier, par exemple, les artefacts constituent la composante principale du niveau sémiotique de la culture, et ils regroupent les *objets culturels* :

> Les **artefacts** comprennent les objets culturels et les déchets : les premiers appellent l'interprétation qui fait de leur production une production de sens ; les seconds restent dans l'insignifiance. Les objets culturels [soit « tout résultat d'une objectivation, qui peut à ce titre participer d'une pratique sociale : ainsi, par exemple, d'une partition musicale. »] se divisent à leur tour en trois catégories : les **outils** et, plus complexes, les **instruments** (en comprenant par là aussi les instruments de communication comme les médias) ; les signes (linguistiques ou non : mots, symboles, chiffres, etc.) ; enfin les **œuvres**, qui sont issues d'une élaboration de signes, au moyen des outils. Entre les signes et les œuvres, on relève une différence de complexité : c'est l'action combinée des outils et des signes qui permet de produire les œuvres. [...] Elles sont l'aboutissement du mouvement propre de l'action humaine qui les produit, en créant les formations médiatrices entre le monde proximal et le monde distal : les arts, les religions et les sciences. (Rastier, 2010, p. 17-19)

Enfin, les artefacts regroupent plusieurs catégories d'objets culturels. Dans ce contexte, l'ordinateur peut être étudié comme un *instrument* (sous-catégorie des objets culturels), c'est-à-dire un objet construit par l'homme et qui interagit avec lui dans différentes situations.

Le troisième axe présente l'ordinateur comme un outil pour l'analyse sémiotique. Cet axe diffère du premier, même si à première vue il pourrait en constituer une sous-catégorie. En effet, le troisième axe étudie des phénomènes différents et qui ne sont pas liés à l'ordinateur et à son utilisation, comme par exemple pourrait l'être un travail sur l'utilisation du téléphone cellulaire dans la vie de l'être humain. Au contraire, dans le troisième axe, l'ordinateur est seulement un outil au service de la

sémiotique pour l'analyse de phénomènes sémiotiques. Ainsi, la principale différence réside dans le fait que l'objectif n'est pas la simulation de phénomène sémiotique par l'ordinateur ou l'analyse des aspects computables de la sémiotique, mais plutôt l'analyse de produits sémiotiques par le biais de l'ordinateur. Il est toutefois évident que la réflexion théorique du premier axe nourrit les applications pratiques du troisième.

1.1.1 L'ordinateur comme « machine sémiotique » et *semiosis* computable

Le premier champ de recherche est axé sur l'étude de modèles computationnels simulant des dynamiques sémiotiques, ce qui, métaphoriquement, peut être appelé *l'étude de l'ordinateur comme machine sémiotique*. Plusieurs travaux ont été effectués dans ce domaine et ont décrit plus ou moins explicitement le lien entre la sémiotique et l'intelligence artificielle. Une des questions fondamentales posées dans ce contexte est la manière de « modéliser » les processus de construction du *sens* dans des systèmes artificiels (Gudwin et Gomide, 1997). Répondre à cette question devient de plus en plus important si on considère les avancements dans le domaine de l'intelligence artificielle et l'émergence de travaux sur la reproduction de systèmes de sens dans un contexte artificiel. Certaines recherches ont abordé, par exemple, l'étude de l'évolution du langage par le biais d'une simulation artificielle (Cangelosi et Parisi, 2002). L'objectif est de simuler le processus d'évolution de la faculté du langage de la préhistoire à l'*homo sapiens*. Beaucoup d'autres études de ce genre ont été réalisées, contribuant ainsi à l'avancement des sciences cognitives (Loula *et al.*, 2007). L'effervescence de cet axe d'études a mené à la naissance de nouvelles disciplines, comme la linguistique computationnelle, dont la mission est basée sur la simulation artificielle de processus linguistiques et sur le développement d'outils pour la détection automatique des structures syntactiques ou d'outils pour la désambiguïsation sémantique automatique. Dans ce contexte, le choix du modèle sémiotique utilisé influence grandement la simulation elle-même (Queiroz et Merrell,

2008) et détermine le succès ou l'échec de certains outils. Dans les années 1960 par exemple, le modèle de Chomsky a permis le développement de plusieurs méthodes qui constituent aujourd'hui des pierres angulaires de la linguistique computationnelle.

La sémiotique joue un rôle très important dans ce genre de recherches et plusieurs modèles sémiotiques ont été proposés pour la simulation des systèmes et des processus du sens. Dans Queiroz et Merrell (2008) par exemple, la sémiotique de Peirce est étudiée comme cadre conceptuel à partir duquel il est possible de comprendre la *semiosis* et de construire un modèle formel pour sa simulation par l'ordinateur. La semiosis est entendue comme un processus qui s'auto-organise comme le font les organismes biologiques. Cette considération a mené à envisager l'utilisation de mécanismes mathématiques dits *automates cellulaires* à des fins de formalisation de certains processus sémiotiques (Etxeberria et Ibáñez, 1999).

En raison de la nature de la problématique, c'est-à-dire la simulation de processus cognitifs dans un contexte artificiel, les approches qui se développent dans ce cadre ne sont pas exemptes de critiques qui tendent à souligner les limites intrinsèques des machines face à la complexité des facultés cognitives :

> If minds were machines, then it would be very plausible to suppose that the computational model of the mind—according to which software is to hardware as minds are to brains—could be sustained and simulations of actions would have no special problems to overcome (Fetzer, 2011, p. 45)

La complexité d'un tel processus de simulation est due selon certains au cadre non-déterministe des processus cognitifs et des facteurs émotionnels jouant un rôle important dans la vie cognitive de l'être humain. Ces éléments constituent des problèmes majeurs pour les modèles artificiels (Fetzer, 2011) qui ne peuvent qu'être construits dans un cadre déterministe.

Le débat sur la nature sémiotique des ordinateurs et leur capacité à simuler des processus cognitifs demeure ouvert et plusieurs pistes de recherche sont prometteuses (Rapaport, 2012). Par exemple, l'utilisation des travaux de Peirce pour la compréhension des ordinateurs comme machines sémiotiques est très répandue (Ketner, 1988), probablement en raison de la relative facilité avec laquelle on peut considérer sa théorie du point de vue mathématique (Nadin, 1977).

Cette typologie de travaux a aussi ouvert les portes à une sémiotique computationnelle différente, dont les études portent plutôt sur l'importance de la computation en sémiotique. Ceci constitue un axe de recherche opposé à celui des études précédentes qui portent plutôt sur l'importance de la sémiotique pour la computation. Les travaux de Zipf, par exemple, pourraient appartenir à la sémiotique computationnelle. Ils ont mené à la définition de la loi de Zipf (1949) qui est un modèle mathématique décrivant un phénomène lexical. Cette recherche, bien qu'elle parait atypique, illustre le type de travail qu'il est possible de considérer dans un contexte de sémiotique computationnelle.

Toutefois, les études en sémiotique qui ont porté sur le lien entre les dynamiques sémiotiques et la computation demeurent limitées (Meunier, 2018). Les travaux qui ont le plus contribué à l'approfondissement de ces réflexions sont ceux qui mettent *en relation les caractéristiques formelles des modèles sémiotiques avec les caractéristiques formelles des modèles computationnels* (Meunier, 1989, 2014, 2017a, 2017b, 2019b, 2019a). Dans ces travaux, le point de départ est une considération sur la nature des théories scientifiques et sur les modèles qui les constituent. Ces modèles sont de différents types et permettent d'aborder un phénomène sous divers angles et perspectives. Par exemple, un phénomène quelconque, tel que le jeu d'échec, peut être représenté par un modèle symbolique où il est décrit comme une bataille médiévale, par un modèle formel où toutes les règles du jeu sont expliquées, par un

modèle formel mathématique ou par un modèle informatique (Meunier, 2018, p. 81). La sémiotique computationnelle devra alors identifier la portion de théorie sémiotique qui peut être formalisée. Il est aussi possible d'analyser les potentialités computationnelles de modèles sémiotiques formalisés. En d'autres termes, la sémiotique computationnelle doit répondre à la question « existe-t-il des phénomènes sémiotiques qui peuvent être modélisés formellement et qui sont aussi computables? »

1.1.2 L'ordinateur comme artefact

Le deuxième domaine de recherche est très varié et comporte différents programmes de recherche. Leur dénominateur commun est constitué par la vision de *l'ordinateur comme un artefact*, c'est-à-dire un objet façonné par l'homme et qui est fonctionnel dans la plupart des cas. Un artefact peut être étudié sous différents angles mais il demeure un objet culturel construit par et pour l'homme, qui interagit avec lui dans différentes situations possédant chacune ses propres caractéristiques sémiotiques. Dans ce cadre, la sémiotique est souvent utilisée pour ses avantages en termes de design et de conception des systèmes informatiques et pour améliorer l'interaction entre l'homme et l'ordinateur.

En simplifiant, on peut subdiviser cet axe de la sémiotique computationnelle en trois sous-groupes, soit les études sur l'interaction homme-machine (IHM), les études sur les systèmes d'information comme systèmes sémiotiques et les études sémiotiques de la programmation et de ses langages. Dans le premier sous-groupe, la sémiotique offre un cadre théorique à partir duquel le design et les fonctions des logiciels, des interfaces et de tout autre type de modèle d'interaction entre l'homme et la machine sont étudiés. Les travaux réalisés en *ingénierie sémiotique* (De Souza, 2005) constituent une référence importante, où les modèles sémiotiques sont au service du design informatique et de l'organisation fonctionnelle des interfaces d'interaction

homme-machine. Les réflexions théoriques dans ce contexte ont été menées sur la base du modèle de la communication de Jakobson qui permet l'analyse des interactions homme-machine sous l'angle des actes communicatifs.

Le deuxième sous-groupe est ancré dans la *sémiotique organisationnelle* qui a été inaugurée dans les années 1970 (Stamper, 1973) et qui est surtout enracinée dans la théorie des signes de Morris. Celui-ci les distinguait au niveau de leurs relations syntaxiques, sémantiques et pragmatiques. Ces études sont définies ainsi:

> l'étude de l'organisation en utilisant les concepts et les méthodes de la sémiotique [...][car] chaque comportement organisé est effectué au moyen de la communication et l'interprétation de signes entre humains, individuellement ou en groupe. (Liu K. , 2000, p. 19) [Notre traduction]

Les champs d'application sont très variés : le marketing, la gestion des ressources humaines, le droit des affaires, l'éthique des affaires, l'étude de l'évolution historique des organes gouvernementaux, religieux, politiques, juridiques, sociaux, et médicaux et enfin, les affaires publiques et privées. Toutefois, une grande partie de ces travaux est surtout liée aux systèmes d'information et concerne donc l'informatique. Le cadre théorique développé dans ce contexte est appelé *semiotic ladder* (Stamper, 1973) et il est formé de six niveaux d'analyse, soit les niveaux physique, empirique, syntaxique, sémantique, pragmatique et social.

Le troisième sous-groupe est strictement lié à *l'étude sémiotique des langages de programmation*, un projet qui remonte aux années 1960 (Zemanek, 1966) et qui met en valeur un aspect fondamental du design des ordinateurs, c'est-à-dire le système de codes et de signes qui organisent leur comportement. Dans ces travaux, les structures conceptuelles de la programmation sont perçues comme systèmes sémiotiques à l'intérieur desquels l'humain organise des codes et des signes pour *communiquer de manière non ambiguë* avec un système physique afin qu'il exécute des opérations

logiques et computables. La programmation est un système de contrôle des appareils physiques. Les théories qui ont été utilisées dans ce contexte sont surtout celles de Peirce (Gazoni, 2018) où le fonctionnement de l'ordinateur est associé aux phénomènes de *secondéité* (en anglais, *secondness*) puisque les opérations qu'un ordinateur exécute ne peuvent qu'être définies en amont à travers un système non ambigu de codes et signes. L'intérêt pour ce genre de recherches est revenu à la surface avec des études plus larges (Tanaka-Ishii, 2010), où, en suivant les deux grands courants de la sémiotique, le peircien et le saussurien, une catégorisation des modèles, des types et des systèmes de signes des langages de programmation a été proposée.

1.1.3 L'ordinateur comme outil au service de l'analyse sémiotique

Le dernier axe de recherche de la sémiotique computationnelle nous concerne de plus près. Dans cet axe, l'ordinateur est un outil au service des analyses sémiotiques. Le mot « ordinateur » n'est alors qu'une métaphore pour indiquer plusieurs éléments de nature mathématique, statistique, formelle, computationnelle et informatique. Pour le dire autrement, les travaux de cet axe continuent à étudier des phénomènes sémiotiques et ses produits, mais par le biais de méthodes hybrides d'analyse sémiotique, qui se servent de l'assistance informatique. Cet axe diffère donc des deux autres, car l'objet de recherche n'est pas la détermination de modèles computationnels pour la simulation sémiotique, ni les aspects computables de la signification et ni l'ordinateur comme artefact. L'objectif est de répondre à des questions traditionnelles en sémiotique, comme l'analyse d'un corpus de textes et de le faire avec une assistance informatique. Cependant, le développement ou l'adaptation de méthodes computationnelles à l'analyse sémiotique ne peut qu'alimenter la réflexion théorique sur la nature computationnelle de la signification. En ce sens, cet axe et le premier sont étroitement liés et interdépendants. La recherche

future en sémiotique computationnelle devra clarifier le plus possible la nature de cette relation.

La manipulation des produits sémiotiques par des moyens computationnels fait déjà partie des projets de recherche les plus importants de l'intelligence artificielle et ce, depuis la naissance de cette discipline. En particulier, le texte constitue le matériel sur lequel les méthodes qui manipulent ces produits sémiotiques est né et s'est développé. Ainsi, il existe plusieurs travaux qui appartiennent au champ de recherche de l'analyse de texte assistée par ordinateur et dont nous parlerons au prochain paragraphe (cfr 1.1.3.1). Ces travaux tirent profit d'un certain nombre de disciplines de l'intelligence artificielle: sémantique vectorielle et/ou sémantique distributionnelle (Benzécri, 1966a; Fabre et Lenci, 2015; Lebart et Salem, 1994b; Sahlgren, 2006; Salton *et al.*, 1975), apprentissage automatique (Aggarwal, 2018; Bishop, 2006; Hastie *et al.*, 2013; James *et al.*, 2013), recherche d'information (Baeza-Yates et Ribeiro-Neto, 2011; Kraft et Colvin, 2017; Le Priol *et al.*, 2009; Manning *et al.*, 2009; Salton, 1988; Salton et McGill, 1983; Van Rijsbergen, 2004; Zhai, 2009), traitement automatique des langues (Bouillon et Vandooren, 1998; Jacquemin, 2001; Manning et Schütze, 1999), fouille de texte (Aggarwal et Zhai, 2012) et exploration de données (Aggarwal, 2015; Attewell et Monaghan, 2015; Bouroche et Saporta, 1992; Bramer, 2007; Han *et al.*, 2011; Kantardzic, 2011; Larose, 2006; Linoff et Berry, 2011; Saporta, 2006; Tufféry, 2017). L'ensemble de ces disciplines est pertinent pour le traitement et l'analyse du langage naturel et, plus particulièrement, pour l'*analyse de texte*. En analyse du contenu par exemple, plusieurs recherches sont conduites à l'aide de l'ordinateur pour l'analyse de plusieurs typologies de textes, qu'ils soient journalistiques ou publicitaires (Anderson *et al.*, 2006; Bell et Milic, 2002; Glasgow University Media Group, 1980; Kordjazi, 2012; Leiss *et al.*, 1990; McQuarrie et Mick, 1992). Parfois, la sémiotique est amenée à jouer un rôle d'assistance dans l'interprétation des données ou sert de cadre théorique pour l'annotation.

Malgré le grand développement technologique et l'application toujours plus répandue de ces méthodes aux recherches en sciences humaines et sociales, *l'intérêt pour l'analyse sémiotique par ces moyens ne s'est pas diffusé au sein de la communauté de sémioticiens* (Meunier, 2018; Tanaka-Ishii, 2015). Mais les nouvelles technologies de traitement du texte obligent les chercheurs en sémiotique à s'interroger sur l'utilisation des modèles computationnels dans leur recherche. La compréhension sémiotique de ces modèles et leur exploitation dans un contexte d'analyse des produits sémiotiques constitue la perspective adoptée par cette thèse pour se pencher sur la sémiotique computationnelle.

1.1.3.1 L'analyse du texte assistée par ordinateur

Le troisième axe de recherche de la sémiotique computationnelle regroupe certainement les travaux déjà accomplis en analyse de texte assistée par ordinateur, dont l'objet de recherche est à la fois le développement de méthodes d'analyse de texte à l'aide d'outils statistiques, mathématiques et informatiques, et l'analyse de corpus textuels. Ce domaine est vaste et regroupe des travaux qui naissent dans différents contextes, comme l'analyse de contenu (Berelson, 1952; Carley, 1990; Drisko et Maschi, 2016; Franzosi, 2008, 2011; Krippendorff, 2004; Neuendorf, 2002; Weber, 1990, 1990), l'analyse du discours (Adam, 2011; Adam et Heidmann, 2005; Charaudeau *et al.*, 2002; Chartier, 2003; Dobre, 2013; Greimas, 1979; Z. S. Harris et Dubois-Charlier, 1969; Maingueneau, 1997; Née, 2017; Rizkallah, 2013; Rizkallah et Della Faille, 2014), la statistique textuelle et la linguistique de corpus (Bilger, 2000; Bolasco, 2005; Heylen et Bertels, 2016; Lebart et Salem, 1994b; Mayaffre, 2002; Nazarenko *et al.*, 1997; Poudat et Landragin, 2017; Rastier, 2005a, 2009, 2011; Valette, 2018), etc. Il n'est donc pas simple de définir ce champ de recherche de manière univoque et l'expression « analyse de texte assistée par ordinateur » ne fait pas l'unanimité d'autant plus qu'elle n'est pas la seule utilisée. Il demeure toutefois

possible d'esquisser un cadre général des disciplines qui contribuent aux avancées dans ce champ de recherche.

Puisque, à notre connaissance, il n'existe aucune tentative de décrire l'interdisciplinarité de l'analyse de texte assistée par ordinateur, le cadre (figure 1.1) ici proposé ne peut qu'être considéré provisoire et non exhaustif. Des disciplines comme l'analyse de contenu et l'analyse du discours ont utilisé cette méthode dans plusieurs domaines, comme la sociologie et la communication, ce qui a mené à des applications variées (Barcus, 1961; Bell et Milic, 2002; Carley, 1993; Drisko et Maschi, 2016; Fan, 1988). Ces disciplines ont contribué à la diffusion de méthodes mathématiques, comme l'analyse en composantes principales, pour l'exploration du contenu d'un corpus (Andsager et Powers, 1999). La statistique textuelle et la linguistique de corpus regroupent un certain nombre de travaux s'inspirant du statisticien Benzécri (1966b, 1982, 1981) et l'analyse factorielle de correspondances a été utilisée pour l'analyse de texte (Benzécri et Benzécri, 1980; Cibois, 1933; Escofier-Cordier, 1969; Lebart et Salem, 1994a; Salem, 1982). Un autre champ de recherche est celui du traitement automatique du langage naturel (TALN) et de la linguistique computationnelle (Allen, 1995; Bouillon et Vandooren, 1998; Clark *et al.*, 2010; Dale *et al.*, 2000; Fuchs *et al.*, 1993; Jacquemin, 2001; Kao et Poteet, 2007). Dans ce contexte, plusieurs outils de manipulation et de traitement de données textuelles ont été développés, augmentant considérablement les potentialités de l'analyse de texte. L'analyse de données regroupe toute une série de méthodes développées dans plusieurs champs de l'intelligence artificielle, comme l'apprentissage automatique (Adam, 2016; Amini et Bach, 2015; Cornuéjols et Miclet, 2003; Dreyfus, 2008). Cette dernière, par exemple, fournit des méthodes avancées pour l'analyse de données. Plusieurs autres disciplines des sciences humaines comme la linguistique, la sémiotique et la philosophie, contribuent également au développement de l'analyse de texte assistée par ordinateur. Le plus souvent ces

disciplines fournissent un cadre théorique pour l'analyse du texte mais elles peuvent aussi constituer des applications originales de ces méthodes, comme l'analyse conceptuelle en philosophie, ce qui enrichit la réflexion méthodologique et théorique.

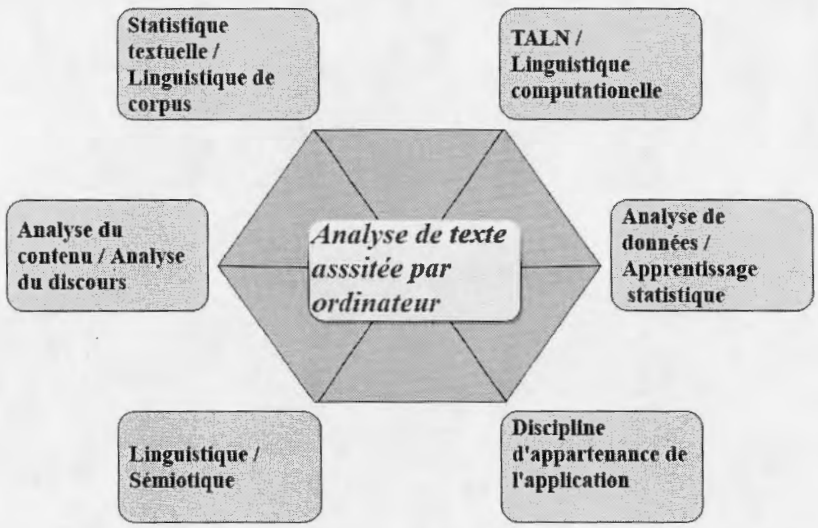


Figure 1.1 Hexagone définissant l'interdisciplinarité de l'analyse de texte assistée par ordinateur.

1.1.4 La vocation empirique de la sémiotique

Le premier à avoir esquissé un parcours scientifique du projet sémiotique a été A. J. Greimas, dans son œuvre *Sémantique structurale* (Greimas, 1966). Pour cet auteur, la sémantique doit agir en « vue de la scientificité » et construire un métalangage scientifique pour la description sémantique (Greimas, 1966, p. 13-20). Or, la notion de « scientificité » doit être appréhendée à l'aide de celle d'« empirie »:

> Si la sémiotique a [...] une vocation scientifique, elle a avant tout une « vocation empirique ». [...] Elle a cependant également le devoir d'entrer en contact avec [...] [les] pratiques de signification complexes dont on peut expliciter des fonctionnements du sens.[...] La vocation empirique de la sémiotique nous porte également et surtout à nous poser la question de savoir si, éventuellement, dans la pratique, par exemple dans la peinture de l'époque de

> Spinoza, par hasard il y eut une quelconque idée du signe implicite, qu'avec les instruments à notre disposition nous pouvons expliciter des tableaux que cette peinture a produits. (Fabbri, 2008, p. 74)

Pour Fabbri, la scientificité de la sémiotique dépend de la valeur qu'elle donne à l'empirie et aux analyses empiriques des phénomènes sémiotiques, mettant ainsi en lumière le rôle plus important de la sémiotique, c'est-à-dire l'analyse des produits sémiotiques.

La « vocation empirique » de la sémiotique est certainement mise en valeur par le troisième axe de la sémiotique computationnelle, qui découvre des nouvelles méthodes pour l'explicitation des fonctionnements du sens.

1.2 Médias, texte et sémiotique

Les *médias* et les produits sémiotiques qu'ils construisent, comme le texte journalistique, sont un des objets d'étude typiques de la sémiotique. L'étude du texte journalistique par assistance informatique s'insère dans le troisième axe de la sémiotique computationnelle. Le terme *média* désigne tout moyen de diffusion d'un *message*, mais il est le plus souvent entendu comme l'ensemble des sources d'information de masse, comme la radio, la télévision, le cinéma, etc. L'importance que les organes d'information ont sur la vie quotidienne est considérable.

> Village global, économie de l'information, société de communication, médiacratie, culture postmoderne : [...] tous [ces clichés] évoquent d'une manière ou d'une autre la place prépondérante de la communication dans la vie quotidienne des individus et le fonctionnement de la société. Les médias nous parviennent « en continu », sur « larges bandes », et notre expérience du monde se réduit souvent aux messages dont ils nous inondent (Bonville, 2006, p. 5)

Les médias jouent en effet un rôle très important dans la société, soit celui *d'informer le citoyen.* La transmission d'informations pertinentes pour la vie sociale et politique

des citoyens constitue le rôle principal des média et cela inclut également les motivations du contrat de lecture avec son public. Les attentes du lecteur envers une source information peuvent varier, mais elles dépendent de la véridicité de l'information. Lorsque cette confiance envers la validité de l'information diminue, le pacte entre le lecteur et une source d'information est brisé et le journal perd inévitablement son public. En raison de l'importance de leur rôle, les médias sont souvent au centre des pressions du monde politique qui, connaissant son pouvoir, essaie de les manipuler. Toutefois, ce jeu de manipulation n'est pas à sens unique. Les médias « manipulent autant qu'ils sont manipulés » (Charaudeau, 2013, p. 11), car d'un côté ils sont conscients de leur force manipulatrice sur leur public et ils l'utilisent, et de l'autre, ils sont utilisés par les politiciens pour communiquer des idées et faire passer des messages particuliers à la population ou, le plus souvent, à d'autres acteurs politiques. C'est pour ces raisons que les médias reçoivent une grande attention de la part des citoyens et des agences de contrôle non gouvernementales, comme « Reporters Sans Frontières » et autres. L'importance d'un tel contrôle est attestée par le fait que la liberté de presse d'un pays représente un des éléments les plus importants de la santé des systèmes démocratiques. L'information et les médias sont au cœur de la démocratie parce qu'ils tendent à assumer « une logique symbolique qui fait que tout organe d'information se donne pour vocation de participer à la construction de l'opinion publique ». (Charaudeau, 2013, p. 13). Dans ce contexte, le monde académique joue aussi un rôle important à travers ses analyses, le développement de méthodes d'analyse et la surveillance constante du système politique et financier des médias.

Les sciences humaines et sociales se sont penchées sur l'analyse de médias depuis leur naissance et ceci, pour toutes sortes d'objectifs de recherche. Depuis le début du XX^{e} siècle, l'analyse du texte journalistique est une pratique très répandue qui est le plus souvent utilisée pour étudier différents types de phénomènes culturels et sociaux

(Ringoot, 2014, p. 3). Il existe de nombreuses méthodes pour l'étude des médias d'information et elles varient selon le type de média, l'objet de l'analyse, les objectifs de la recherche, etc. Parmi elles, l'étude du texte journalistique dans sa forme écrite est sans doute la pratique la plus répandue. Par exemple, en sémiotique et en sociologie, il est très fréquent d'adopter une posture empirico-déductive (ou, plus rarement, empirico-inductive) à travers laquelle, « partant d'une théorie du découpage de l'objet empirique » (corpus), on se dote « d'instruments d'analyse permettant de rendre compte des effets de signifiance que cet objet produit en situation d'échange social » (Charaudeau, 2013, p. 14).

Le texte est un produit sémiotique contenant des informations, c'est-à-dire un *message* construit par un destinateur ou un *émetteur* et pour un destinataire ou *récepteur*, par exemple les lecteurs du journal *La Presse*. Il est ainsi transmis par un *canal* de communication, par exemple l'article de journal ou l'article destiné à la version numérique du journal. Le texte est la manifestation empirique d'un acte d'énonciation complexe. Fontanille le définit ainsi:

> un « texte-énoncé » est un ensemble de figures sémiotiques organisées en un ensemble homogène grâce à leur disposition sur un même support ou véhicule (uni-, bi-, ou tridimensionnel) : le discours oral est unidimensionnel, les textes écrits et les images, bidimensionnels, et la langue des signes, tridimensionnelle (Fontanille, 2008, p. 6)

Le texte est défini comme un ensemble homogène d'éléments et sa caractéristique principale est la cohérence. Les éléments mis ensemble dans le tissu du texte sont liés par une « force invisible » qui les rendent cohérents entre eux. Ces réflexions ont guidé la construction des méthodes et théories du texte menant à ce qu'on peut appeler la « grammaire textuelle »:

> Le discours était considéré ou bien comme une longue phrase dérivable par des enchâssements et des concaténations répétés [...] et non pas comme unité maximale formelle d'une grammaire. Il est apparu cependant que ces

> conclusions sont erronées. D'abord, les relations ont un caractère purement formel et grammatical et ne dépendent pas seulement de facteurs contextuels. [...]. Une description adéquate de ces phénomènes linguistiques [...] ne peut être donnée que dans une grammaire spécifiant explicitement des séquences de phrases. Tout porte à croire que le sujet parlant connaît les règles qui sous-tendent ces relations. Sans cela il lui serait impossible de produire des énoncés cohérents.[...] [S]a compétence est nécessairement une compétence textuelle. (Chabrol, 1973, p. 184)

Dans ce long extrait, on souligne que les individus se servent de la *compétence textuelle* pour produire des énoncés cohérents. Toutefois, le texte et sa cohérence ne sont pas acquis à l'avance, mais plutôt construits. Ainsi, le texte journalistique est le produit final d'un *mécanisme de construction du sens* qui suit un double processus de sémiotisation (Charaudeau, 2013, p. 30), soit celui de transformation et celui de transaction. Le *processus de transformation* consiste à donner une forme à la substance de l'information. Il rassemble les éléments qui se réfèrent à la réalité que le journaliste veut raconter. C'est un processus d'identification des éléments qui seront inclus dans le texte et qui en formeront un ensemble cohérent. En d'autres termes, il s'agit de nommer les éléments qui transforment l'information brute en information formée, comme les acteurs de l'action, leurs caractéristiques, leurs désirs et intentions, leurs actions, le temps de l'exécution, etc. De temps en temps, l'information formée et racontée nécessite une étape de *sanction*, c'est-à-dire d'évaluation, ce qui mène généralement à exprimer des opinions sur les évènements racontés. L'acte d'informer s'inscrit donc dans un processus complexe où l'émetteur raconte, explique, évalue, etc. Le *processus de transaction*, de son côté, regroupe les stratégies de « mise en scène de l'information », qui sont déterminées par la nature du récepteur ou destinataire du message. Les informations que le journaliste veut communiquer doivent respecter les attentes du destinataire et le contrat de lecteur que le journal a « signé » avec son public. Ainsi, tout discours journalistique témoigne d'une relation intime entre auteur du message et récepteur, ce qui modifie la relation entre information et réalité. L'information n'est jamais fournie telle quelle, elle est

construite de telle manière qu'elle répond aux intérêts, opinions et attentes de ses lecteurs.

Le processus de transaction conduit à réfléchir sur la relation entre *information, réalité et processus d'écriture*. L'information « n'est pas une marque impersonnelle des événements; elle est sélection, organisation, interprétation opérée sur les données de départ » (Volli, 2008, p. 231) [Notre traduction]. En ce sens, l'objectivité est un indicateur de la nature de cette relation. Plus un texte est objectif, plus la nature de cette relation est caractérisée par la véridicité. Un texte objectif est généralement un texte qui raconte les évènements de manière fidèle. Toutefois, en raison de stratégies de « mise en scène de l'information », l'objectivité est construite à travers une stratégie d'énonciation et elle ne peut être atteinte que partiellement. Paradoxalement, l'objectivité joue un rôle important dans la solidité du contrat de lecture entre le journaliste et ses lecteurs, mais elle n'est qu'une « utopie irréalisable ». Le pacte entre les citoyens et les organes d'information est basé sur cette relation entre information, réalité et processus d'écriture. En effet, on lit un journal si on peut croire avec confiance que les informations qu'on en tire sont véridiques et fidèles à la réalité. Toutefois, raconter des évènements implique toujours une « sélection, organisation [et] interprétation » de la réalité, ce qui met en danger constamment la solidité du contrat de lecture.

La prise de conscience de la difficulté de produire une vision objective de la réalité a souvent concentré l'attention des sémioticiens sur la « mise en scène de l'information » (Delforce, 1985; Jamet et Jannet, 1999) et sur les stratégies de présentation des nouvelles (Marrone, 2001), c'est-à-dire sur les processus de transaction. Toutefois, les études sur le texte journalistique en sémiotique sont très variées, passant de l'analyse du moyen de communication comme tel, par exemple, la presse écrite (Eco, 1997), à des études plus spécifiques, comme l'analyse du rapport

texte-image (Lambert, 1986) ou l'analyse du processus d'informatisation des quotidiens (Cotte, 2001). Une des approches les plus utilisées pour l'étude du texte journalistique est celle de la *sémiotique textuelle*. Cette approche sémiotique se distingue des autres par sa conception du texte qui est perçu comme matériel empirique privilégié pour l'enquête sémiotique (Volli, 2014). Le texte est alors « ce qui se donne à appréhender, l'ensemble des faits et des phénomènes que [le sémioticien] s'apprête à analyser » (Fontanille, 1998, p. 79). Cette approche a influencé l'étude sémiotique du texte journalistique, pour laquelle cinq niveaux d'analyse ont été identifiés, soit le plan de l'expression, la dimension énonciative, la dimension narrative, la dimension cognitive et la dimension passionnelle (Lorusso et Violi, 2004, p. 5; Pozzato, 2005). En général, l'analyse du plan de l'expression tient compte entre autres de la structure du journal, des formats possibles, de la mise en page, des titres, de la construction de la première page, etc. La dimension énonciative met de l'avant l'étude de l'acte énonciatif et donc le rôle joué par les marques de l'énonciation, comme celles du journaliste ou du journal. L'analyse de la dimension narrative se penche sur l'extraction des structures narratives des textes, en portant attention aux temps, aux lieux, aux actants, aux acteurs, aux thématiques, aux valeurs, aux figures, etc. La dimension cognitive souligne le rôle informatif du journal et, enfin, le niveau pathémique identifie les euphories et les dysphories, la tension et le rythme, etc. Chacun de ces niveaux peut être analysé séparément. Toutefois, les niveaux se chevauchent dans le texte et forment un réseau de dépendance. Par exemple, il n'est pas possible de faire l'analyse du niveau pathémique sans avoir un aperçu des dynamiques de l'énonciation; de même, le plan de l'expression est indispensable pour obtenir des informations pertinentes au niveau du contenu.

La dimension narrative du texte journalistique joue un rôle de connexion entre le plan de l'expression, les dynamiques de surface et les structures profondes de mise en place de la signification. Cette dimension n'est pas exclusivement identifiable dans

certains types d'articles de journaux, par exemple les nouvelles. L'organisation narrative du texte émerge aussi dans les éditoriaux ou les articles d'opinion, où « la fonction argumentative semble donc être la condition même de la production narrative » (Revaz, 1997, p. 27). Cette position sur le rôle de la narrativité dans la production textuelle est partagée par plusieurs auteurs. Pour Rastier par exemple, la dimension narrative demeure un élément pertinent à l'intérieur de l'analyse et ce particulièrement pour le plan dialectique (Rastier, 2011, p. 181).

Ces catégories sont certainement discutables et d'autres auteurs ont énuméré d'autres composantes pour le plan de l'expression et du contenu : la thématique, la dialectique, la dialogique et la tactique (Hébert, 2016, p. 44). Cependant, certaines d'entre elles, comme celle dialectique et thématique, s'agencent avec le plan de l'expression et les dimensions (narrative, cognitive, etc.) qui composent le plan du contenu.

1.2.1 Les « macrostructures » de la production textuelle

Lorsque l'analyse du texte journalistique se concentre sur les processus de transformation, c'est-à-dire sur les procédés de mise en forme des éléments qui composent l'ensemble cohérent du texte, les *macrostructures* qui soutiennent les textes et leur cohérence surgissent. Un texte est cohérent lorsque les parties qui le composent comportent une liaison logique entre elles. Dans le cas des articles de presse, un texte n'est pas cohérent s'il traite de deux événements très différents entre eux sans aucune motivation logique. Si dans un texte on lit les phrases « Mary est sortie dans la rue. Elle possède un chat. », alors le lecteur cherchera dans le texte la structure logique qui motive l'apparition de ces deux phrases. Par exemple, il est possible que le journaliste raconte l'histoire d'une femme qui est sortie dans la rue pour chercher son chat. Ceci implique l'existence d'une sorte de structure sémantique qui fait le lien entre les différents éléments qui composent l'article. Ces structures,

dites *macrostructures*, émergent parallèlement à la naissance de la « grammaire du texte »:

> On a réfléchi rarement jusqu'ici sur les structures plus globales de textes entiers : les *macro-structures*. C'est qu'il n'est pas du tout sûr, et même improbable, qu'un texte soit simplement une séquence de phrases (ou de couples, triples... n-tuple de phrases), pas plus qu'une phrase elle-même n'est pas une suite simplement linéaire de mots. [...] [L]a notion fondamentale de « cohérence » ne peut pas être définie seulement à ce niveau superficiel de concaténations phrastiques. (Chabrol, 1973, p. 184-185)

La *cohérence est une propriété du texte*, alors que la *macrostructure est l'apparat logique, conceptuel et sémantique* qui permet d'organiser les éléments d'un texte de manière cohérente. Cette structure sous-jacente aux textes a été étudiée avec différents approches. L'une des plus répandues en sémiotique est certainement l'analyse des isotopies du texte. L'isotopie est la récurrence des unités linguistique d'un texte (Rastier, 1972) ou, plus précisément, l'*isotopie sémantique* est l' « effet de la récurrence d'un même sème, [ce qui implique que] [l]es relations d'identité entre les occurrences du sème isotopant induisent des relations d'équivalence entre les sémèmes qui l'incluent. » (Rastier, 2001a, p. 299) Enfin, une isotopie est déterminée par la redondance sémique, c'est-à-dire d'un sème qui est l'unité minimale de la sémantique. En d'autres termes, l'isotopie se constitue sur le plan du contenu, elle génère ses manifestations du plan de l'expression et, ainsi, elle régit la cohérence textuelle (Pessoa de Barros, 1983).

Ainsi, le niveau macrostructurel spécifie le contenu « global » du texte en utilisant les unités qui le composent. Sans une telle macrostructure et les règles qui la sous-tendent, « la cohérence du texte serait seulement superficielle » (Chabrol, 1973, p. 189). Il s'agit du même principe énoncé par Rastier du *global qui détermine le local* (Rastier, 1997; Rastier, 2011, p. 29-30) Les unités linguistiques du texte et les structures sémantiques qu'elles représentent sont « interprétées dans des relations

plus globales caractérisant le texte entier » (*ibid.*). Par exemple, les différentes caractéristiques d'un personnage sont distribuées à travers le texte en entier, mais « intuitivement ils constituent une seule macro-catégorie, dénotant par exemple le « caractère » de ce personnage » (*ibid.*). Du point de vue cognitif, l'émetteur du texte ainsi que le récepteur sont « obligé[s] d'organiser son texte [...] sur la base de telles macro-structures » (*ibid.*). Ces structures abstraites garantissent la cohérence textuelle, qui est une des plus importantes caractéristiques du texte.

L'hypothèse de la macrostructure peut être étudiée selon diverses perspectives, mais elle demeure surtout un phénomène sémantique :

> Une série d'arguments décisifs, linguistiques et psychologiques, peuvent être apportés pour soutenir l'hypothèse selon laquelle la cohérence textuelle est définie aussi à un niveau *macro-structurel*. Ce niveau, qu'on peut aussi identifier avec la *structure profonde* d'un texte, spécifie un contenu « global » du texte déterminant la formation globale des représentations sémantiques des phrases successives. Sans une telle macro-structure et les règles qui la sous-tendent, la cohérence du texte serait seulement superficielle et linéaire. (Chabrol, 1973, p. 189)

Cette dimension sémantique se situe à un niveau profond du texte. En sémiotique textuelle, l'analyse de la dimension narrative a souvent été utilisée pour l'identification de structures subjacentes au texte. Comme l'analyse isotopique, l'analyse narrative permet de dégager une structure sur laquelle la cohérence textuelle s'appuie. Les composantes narratives d'un texte sont identifiées par un modèle narratif, celui-ci étant une hypothèse sur la forme de la structure qui régit la cohérence textuelle. Les théories et les méthodes sont multiples, mais l'approche est généralement similaire. L'idée de base est d'identifier la structure logico-sémantique qui est sous-jacente au texte et d'en dégager les composantes narratives. Par exemple, il est possible d'identifier les actions, les personnages, les temps des actions etc. Les travaux qui ont utilisé ce genre d'approche pour l'étude de la presse écrite sont

nombreux et très variés. Par exemple, elle a été utilisée dans des travaux socio-sémiotiques (Imbert, 1988) ou pour l'analyse des structures des mythes et des stéréotypes au sujet de l'URSS dans la presse française dans les années 1980 (Sueur, 1994). D'autres travaux ont dégagé ces structures à partir d'une étude sur les processus d'intitulation de la presse (Tahar, 2006). Certains auteurs ont étudié le phénomène de vedettisation, c'est-à-dire la tendance de la presse à accorder une grande visibilité à la vie privée des célébrités ou des personnages connus, en soulignant le rôle des structures narratives dans ce processus (Goepfert, 2010a, 2010b).

Un autre groupe d'études s'est penché sur l'examen de la presse avec une approche issue de l'école de Paris et en utilisant une théorie et une méthode spécifique, soit celle de Greimas. La variété de ces études est également très riche. Certaines se sont servies des théories de Greimas pour l'étude du discours journalistique français autour du conflit israélo-palestinien et tchétchène (Kervella, 2008), en dégageant plus spécifiquement les structures actancielles qui décrivent ces conflits. Ces schémas ont été ensuite associés et comparés à un schéma commun, soit la « guerre contre le terrorisme ». D'autres travaux ont étudié les stratégies déployées par la presse pour sensibiliser les lecteurs au don d'organes et ceci, dans l'objectif d'en souligner la logique structurante des ces stratégies. Dans ce cas, un des outils utilisés est le schéma actantiel (Hammer et Amey, 2009). D'autres travaux ont analysé le traitement journalistique dans la presse française du phénomène d'antibiorésistance, c'est-à-dire le développement des bactéries résistantes aux antibiotiques (Arquembourg, 2016). Certes, la sémiotique a contribué à la définition d'un cadre d'analyse globale pour le texte journalistique et à la détermination d'approches d'analyse mettant en valeur les éléments structuraux des textes. Elle s'est largement diffusée dans les autres disciplines, comme en analyse du discours (Barry, 2002, p. 32), où les théories

sémiotiques de l'école de Paris ont constitué ledites « approche sémiotique » (Delfosse, 1979).

Cependant, certains auteurs ont fortement critiqué le principe du « tout narratif », c'est-à-dire l'idée que tout texte journalistique peut être décrit à travers des structures narratives, ceci en soulignant les différences structurelles entre genres journalistiques (Adam, 1997). Ces critiques ont sûrement contribué à illustrer le phénomène des macrostructures qui soutiennent la cohérence textuelle.

Néanmoins, malgré les critiques se faisant grandissantes depuis les années 1980, le courant structuraliste a été amplement employé pour l'étude du texte journalistique (Van Dijk, 1988b, p. 18). Cette approche se base sur l'idée que les structures profondes du texte organisent l'émergence du « sens » du texte et que leur identification permet de fixer l'essentiel du contenu du texte. Puisque le texte journalistique possède ses propres spécificités, l'analyse de ses structures profondes présente des caractéristiques particulières. Le cadre conceptuel qui illustre le mieux les spécificités du texte journalistique est celui du *framing* (Scheufele, 1999), qui remonte aux concepts développés en intelligence artificielle et en psychologie cognitive (Minsky, 1988; Schank et Abelson, 1977). La théorie du framing concerne plusieurs aspects du phénomène des médias, dont certains s'éloignent de l'analyse du texte journalistique. Toutefois, si on se limite aux procédés de la production textuelle, le framing peut être appréhendé à travers le concept de frame ou de *schéma*. Un schéma est une structure sémantique qui traverse le texte et qui le rend cohérent. Le schéma est en place à la fois lors de la construction du texte et lors de sa réception. Plusieurs définitions du frame sont regroupées par Scheufele (Scheufele, 1999) :

> [Frame is] a central organizing idea or story line that provides meaning to an unfolding strip of events.(Gamson, 1989, p. 157)

> The news frame organizes everyday reality and the news frame is part and parcel of everyday reality .[It] is an essential feature of news (Tuchman, 1980, p. 193)

> Media frames also serve as working routines for journalists that allow the journalists to quickly identify and classify information and to package it for efficient relay to their audiences (Scheufele, 1999, p. 106)

> To frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation (Entman, 1993, p. 52).

Ces définitions reprennent l'idée de la compétence textuelle de l'être humain qui, pour produire des énoncés cohérents, s'appuie sur une structure sémantique qui, dans le contexte de l'étude des médias, est appelée *frame*. Ces schémas sont en fonction lors de la réception du texte journalistique:

> [Frames are] mentally stored clusters of ideas that guide individuals' processing of information (Entman, 1993, p. 53).

> [Frames are] as cognitive devices that operate as non-hierarchical categories that serve as forms of major headings into which any future news content can be filed (Scheufele, 1999, p. 107)

Le *frame* devient aussi une entité cognitive qui soutient la compréhension des textes journalistiques. Qu'il soit explicitement ou implicitement évoqué par les journalistes, le *frame* constitue un filtre de lecture. En effet, un *frame* particulier peut être évoqué à travers de figures rhétoriques ce qui constitue souvent un procédé sémiotique agissant pour la construction du sens et l'interprétation textuelle. En général, toute figure rhétorique peut affecter le contenu sémantique d'un texte et en déterminer la réception. Par exemple, Lakoff a étudié les relations entre certaines métaphores et les domaines conceptuels qu'elles évoquent et il en a dégagé la manière à travers laquelle elles influencent la formation des opinions politiques (Lakoff, 2008).

La théorie du *framing* est liée à l'hypothèse des macrostructures textuelles[1]. Puisque le framing est une théorie de la réception du discours journalistique et de la formation des opinions, l'identification d'éléments macrostructurels, comme les thématiques majoritairement abordées, les acteurs impliqués ou les évènements les plus marquants, contribue à la détection des éléments qui caractérisent la formation de l'opinion publique (Reese *et al.*, 2001, p. 79). Dans ce contexte, l'analyse de phénomènes médiatiques se base généralement sur l'étude d'un *corpus* de textes journalistiques, lequel est construit pour répondre à une question précise (Rastier, 2011, p. 6). Les méthodes pour l'étude d'un corpus journalistique sont très nombreuses et il n'est pas pertinent d'en faire ici une synthèse.

1.3 L'analyse « quantitative » du texte journalistique

L'étude des médias est certainement un des champs d'application le plus important pour le troisième axe de la sémiotique computationnelle. Le texte journalistique a souvent constitué un banc d'essais pour des expérimentations méthodologiques. En effet, l'utilisation d'outils mathématiques et informatiques pour l'analyse textuelle s'est répandue depuis longtemps dans le domaine de l'étude des médias. Une des premières expériences est l'application d'outils « quantitatifs » à l'analyse de contenu. La notion de quantitatif, opposée à celle de « qualitatif », est largement débattue au sein des études en communication. En général, il n'est pas possible de séparer les analyses en études quantitatives et études qualitatives puisque chacune d'entre elles contient une portion de l'une et de l'autre. Krippendorff le souligne à plusieurs reprises dans ses travaux:

> Ultimately, all reading of texts is qualitative, even when certain characteristics of a text are later converted into numbers. The fact that computers process great

[1] Voir aussi le concept de *topoï narratif* dans Hébert (2016, p. 109-113).

> volumes of text in a very short time does not take away from the qualitative nature of their algorithms: On the most basic level, they recognize zeros and ones and change them, proceeding one step at a time. Nevertheless, what their proponents call qualitative approaches to content analysis offer some alternative protocols for exploring texts systematically (Krippendorff, 2004, p. 16).

L'ambigüité de la distinction entre méthodes « quantitatives » et « qualitatives » est mise de l'avant dans les travaux de Van Dijk. Ses études sur l'analyse du discours et les structures du texte journalistique (Van Dijk, 1983, 1985b, 1985a, 1988a, 1988b) constituent des cas emblématiques de l'utilisation conjointe des méthodes quantitatives et qualitatives. L'auteur propose une approche dans laquelle les méthodes quantitatives offrent au chercheur des éléments statistiques qui vont ensuite alimenter l'analyse qualitative. Ainsi, les méthodes quantitatives et qualitatives deviennent les deux facettes d'une même étude (Van Dijk, 1988a, p. X). La linguistique de corpus (Baker, 2006, p. 1; Rastier, 2011, p. 51) et la statistique textuelle (Lebart et Salem, 1994b), issues du croisement de plusieurs disciplines, comme la linguistique, l'analyse du discours, la statistique, l'informatique et le traitement des enquêtes, ont contribué à la diffusion des méthodes « quantitatives » pour l'analyse des médias.

Une des premières analyses quantitatives de la presse remonte à 1893, où l'auteur (Speed, 1893) a illustré la tendance de certains journaux américains à mettre de plus en plus de l'avant les nouvelles à caractère sportif et scandaleux aux dépens du contenu scientifique, littéraire et religieux. Dès la fin du XIXe siècle, la diffusion des méthodologies quantitatives destinées à l'analyse du texte journalistique s'est produite parallèlement aux développements de l'analyse du discours et de l'analyse du contenu. Il existe plusieurs études qui démontrent comment les résultats d'une « analyse quantitative » du texte journalistique sont pertinents pour les disciplines en sciences humaines. En sociologie par exemple, on a étudié les représentations sociales de la féminité et de la masculinité dans la presse québécoise pour adolescent

(Lebreton, 2008). En science politique, on a confronté la hiérarchisation de l'information de plusieurs journaux et son impact sur l'agenda politique national (Hubé, 2007). En sciences du langage, l'étude d'un corpus journalistique a été utilisée pour décrire l'émergence et l'emploi de formules comme « purification ethnique » dans la presse française (Krieg, 2001). En communication, l'analyse de la campagne présidentielle française de 2012 a été supportée par l'analyse du texte journalistique (Maarek, 2014).

L'analyse du contenu est un des champs de recherche en sciences humaines qui est, plus que d'autres, liée à l'utilisation de méthodes quantitatives pour l'étude du texte journalistique (Krippendorff, 2004, p. 5-6). En effet, l'intérêt pour l'application de modèles mathématiques et l'assistance informatique à l'analyse du texte s'est majoritairement développé dans les années 50 (Krippendorff, 2004, p. 12), parallèlement à la naissance de l'analyse du contenu. Une des principales raisons de cette diffusion est le besoin grandissant de manipuler des corpus de grande taille, ce qui suppose des exigences qui ne peuvent être comblées que par une certaine forme d'automatisation, supportée par l'informatique. Ceci a rapidement amené le transfert de certains concepts du domaine des mathématiques à celui des sciences humaines. Dans ce domaine, le concept de « corrélation » est devenu un outil très répandu pour l'analyse de contenu, de la même manière que les concepts de « reproductibilité » et de « généralisation » des résultats sont devenus des exigences méthodologiques (Krippendorff, 2004, p. 31-32, 37, 106). Parmi les différentes techniques, l'analyse factorielle est une de celles qui s'est répandue le plus rapidement. L'analyse factorielle a permis par exemple d'étudier le traitement du thème du cancer dans les magazines pour femmes, soulignant les différentes orientations entre les médias. Dans une étude spécifique (Andsager et Powers, 1999), une différence de traitement du discours sur le cancer du sein entre les revues généralistes d'information et les revues spécialisées pour femmes a été identifiée. Pour certains médias, ce sujet a été

abordé par le biais des thématiques « auto-examens », « implants », « malignité » et « diagnostic », alors que pour les revues généralistes, le discours s'est concentré plutôt sur des questions économiques.

Certains auteurs considèrent l'analyse du contenu comme le précurseur de l'actuelle fouille de textes (text mining) (Yu *et al.*, 2011, p. 733).

1.3.1 News mining

La richesse du domaine de la science des données a permis l'éclosion d'une discipline comme la fouille de textes (*text mining*). Dans ce champ de recherche, l'étude du texte journalistique représente un terrain privilégié d'expérimentation. Toute une série de travaux peuvent être ainsi réunis sous l'étiquette de *news mining*. Les premières études dans ce domaine ont utilisé l'analyse de texte pour scruter les opinions et les thèmes les plus abordées dans les textes des journaux économiques, et ce, à des fins de prédiction des tendances boursières et des réponses du marché (Chowdhury *et al.*, 2014; Fung *et al.*, 2003; Hariharan, 2012; Ingvaldsen *et al.*, 2006; Mahajan *et al.*, 2008; Mittermayer et Knolmayer, 2006; Tang *et al.*, 2009). Le fil conducteur de ces travaux est l'hypothèse que le texte journalistique est une ressource importante pour la prédiction et l'analyse des mouvements boursiers. L'utilisation de certaines techniques, comme la détection des sentiments et des opinions (*sentiment analysis* et *opinion analysis*), en est aussi un thème récurrent. Quelques auteurs ont discuté des problématiques introduites par l'utilisation de ces outils lors de l'analyse du texte journalistique. Néanmoins, ils ont conclu que, même si les réponses aux problèmes demeurent partielles, les avantages de ces outils sont si grands au plan des résultats qu'ils ne peuvent être ignorés (Balahur *et al.*, 2013).

Plusieurs de ces recherches utilisent des techniques en apprentissage automatique (*machine learning*) et fouille de données (*data mining*) dans un contexte d'analyse du

texte journalistique. L'ensemble de ces travaux illustre comment l'adaptation de méthodes de l'analyse de données constitue un des premiers objectifs du champ de recherche en *news mining* (Sayyadi *et al.*, 2007), et ce, d'autant plus que l'augmentation exponentielle des données à caractère informatif (quotidiens en ligne, blogs d'information, base de données, etc.) entraîne la nécessité de disposer de techniques efficaces pour leur catégorisation. Certains chercheurs ont expérimenté les modèles topiques (topic modelling) (Juanzi *et al.*, 2009; Shah et ElBahesh, 2004) dans une tâche de classification des articles de presse; d'autres ont adapté des techniques de regroupement (*clustering*) pour l'analyse du texte journalistique (Bouras et Tsogkas, 2010; US9361369 B1, 2016). La théorie de la logique floue a été utilisée pour l'analyse de l'information journalistique circulant sur la toile (Iglesias *et al.*, 2015) afin d'établir de systèmes de classification des nouvelles en temps réel. D'autres chercheurs se sont concentrés plutôt sur l'analyse de blogs ou microblogs (Twitter) (Berendt, 2016) dans le but d'en extraire des opinions et des informations pertinentes sur la base d'un même thème abordé. Certaines recherches se sont limitées à tester des méthodes d'indexation pour la catégorisation des articles de journaux (Hsu, 2010). Enfin, le texte journalistique a servi aussi de banc d'essai pour la conception de techniques de résumé automatique (Malhotra et Dixit, 2013; McKeown et Radev, 1995). Dans ce dernier contexte, le texte journalistique constitue le cas idéal pour le développement de méthodes produisant des résultats personnalisés et adaptés aux intérêts et aux goûts de chaque utilisateur. Un autre type de travaux a exploité la nature même du texte journalistique pour développer des méthodes aptes à extraire automatiquement les événements qui y sont « racontés » (Hou *et al.*, 2015; Sayyadi *et al.*, 2008). Enfin, les techniques de text mining ont été utilisées pour déterminer de nouveaux indicateurs sociaux ou économiques qui sont utilisés pour la prise de décision au niveau politique. Par exemple, certains travaux, ont fourni des informations pertinentes à des agences gouvernementales comme *Frontex* (Piskorski

et Atkinson, 2011) et à des agences nationales dont les mandats sont le contrôle de l'immigration et la sécurité des frontières (Atkinson *et al.*, 2010).

1.4 Un champ inexploré : le croisement de l'analyse sémiotique et du news mining

Le texte journalistique est analysé par plusieurs moyens dans le but de répondre à plusieurs types de questions de recherche, et ceci, à l'intérieur de plusieurs disciplines appartenant aux sciences humaines. En informatique et en science des données, l'analyse des médias constitue un banc d'essai pour le développement de nouveaux algorithmes et méthodes. En général, l'analyse du texte journalistique revêt une grande importance pour les sciences. De plus, avec les développements des nouvelles technologies de l'information et de la communication, l'accumulation massive de textes journalistiques numérisés soulève des questions d'analyse et d'organisation qui ne peuvent qu'être traitées par des méthodes de l'intelligence artificielle. Dans ce contexte, la sémiotique computationnelle joue son propre rôle, à la fois dans la phase de développement des méthodes et dans la phase d'application et d'adaptation d'algorithmes et de méthodes dans un contexte d'analyse en sciences humaines.

Compte tenu de la vitesse avec laquelle les technologies se développent, le lien entre les méthodes d'analyse de la sémiotique et leur adaptation à un contexte computationnel reste un champ largement inexploré, surtout à l'intérieur de la communauté des sémioticiens. Pourtant, les théories linguistiques et sémantiques sur lesquelles reposent les méthodes actuelles en fouille de textes (text mining) ont une origine similaire aux théories sémiotiques et, en particulier, à celles de l'école de Paris. En effet, le courant structuraliste ouvert en linguistique par Saussure, est à l'origine de ces deux champs de recherche qui semblent, à première vue, très éloignés. Ce point de contact avait déjà été souligné par Van Dijk dans un article qui a été publié dans un ouvrage collectif en 1973 (Chabrol, 1973). Dans ce texte, les travaux

de Harris et ceux de la sémiotique textuelle et narrative se retrouvent l'un à côté de l'autre dans une entreprise de recherche qui essaie d'identifier la manière de passer d'une linguistique générative d'inspiration chomskienne, essentiellement liée à l'analyse des structures phrastiques, à une linguistique textuelle relié à l'analyse du texte et du discours. Ce problème était au cœur du recueil de textes cité mais aussi d'un des travaux de Harris (1952) où il définit le problème de l'analyse du discours en le reliant à la nécessité de trouver un cadre d'analyse linguistique qui va au-delà des bornes de la phrase.

> The first is the problem of continuing descriptive linguistics beyond the limits of a single sentence at a time. The other is the question of correlating 'culture' and language (i.e. non-linguistic and linguistic behavior). The first problem arises because descriptive linguistics generally stops at sentence boundaries. This is not due to any prior decision. The techniques of linguistics were constructed to study any stretch of speech, of whatever length. But in every language it turns out that almost all the results lie within a relatively short stretch, which we may call a sentence. (Harris, 1952, p. 1-2)

Harris identifie aussi un problème de dépassement des méthodes et des techniques ne permettant pas l'analyse inter-phrastique. Dans sa définition du problème de l'analyse du discours, apparaît aussi le besoin de lier l'analyse du texte à l'analyse des phénomènes culturels, ce qui deviendra une des missions les plus importantes de l'analyse du discours et de la sémiotique. Pour van Dijk (1973), le problème de Harris est identique à celui d'une grammaire textuelle se basant sur la découverte des structures narratives du texte, c'est-à-dire la construction d'une méthode pour l'analyse de textes. Les deux approches de l'analyse du discours ont suivi un chemin différent, s'éloignant de leur contexte d'origine. Toutefois, le développement de méthodes différentes basées sur des hypothèses différentes n'exclut pas la possibilité d'en explorer les points de contact. En particulier, la relation entre la grammaire textuelle de type narratif, au cœur de l'école de Paris, et l'hypothèse de la sémantique distributionnelle, développée par Harris entre autres, reste actuellement inexplorée.

En effet, explorer la manière à travers laquelle la sémiotique, l'analyse narrative et la fouille de textes peuvent être employées pour l'analyse de texte assistée par ordinateur représente un objet de recherche très pertinent pour la sémiotique computationnelle. Si l'intérêt pour le développement de l'intelligence artificielle et les méthodes de l'analyse des données n'est apparu que très récemment au sein de la communauté sémiotique (Compagno, 2018; EC-AISS, 2018), dans certains champs de recherche de l'informatique, comme celui des ontologies, les réflexions menées en sémiotique et en narratologie sur la « compétence narrative » ont constitué un cadre théorique important pour le développement des modèles formels de la narration. En effet, de nombreux chercheurs ont utilisé les travaux de Greimas ou de Propp pour la définition d'un modèle computationnel de la narrativité ou du drame (Akimoto et Ogata, 2012; Battaglino *et al.*, 2014; Ciotti, 2016; Damiano *et al.*, 2005; Gervás, 2013a; Lieto et Damiano, 2014; Lombardo et Pizzo, 2014; Mani, 2012). D'autres chercheurs ont aussi utilisé des modèles formels de la narration pour la génération du récit (Fararo, 1993; Porteous *et al.*, 2010; Porteous et Cavazza, 2009; Skvoretz, 1993) ou pour l'analyse des chaînes narratives par des méthodes non supervisées (Chambers et Jurafsky, 2008).

À notre connaissance, peu de travaux ont exploité une approche computationnelle pour l'analyse narrative de texte. Le seul travail pertinent de ce type (Sudhahar *et al.*, 2011) a été conduit sur un corpus journalistique. Ses auteurs ont développé une méthode pour identifier les acteurs (personnages) et les actions en se basant sur la structure de la phrase : sujet-verbe-objet (SVO). Malgré les difficultés de cet ambitieux projet qui a porté sur 100 000 articles, les auteurs illustrent bien comment cette approche peut produire des résultats intéressants pour l'analyse de grands corpus. Par exemple, la recherche a confirmé de manière empirique une opinion largement répandue dans la société, soit que les auteurs de crimes sont généralement des hommes alors que les femmes et les enfants en sont les victimes les plus fréquentes.

Toutefois, cette approche a été appliquée sur les phrases et a démontré des limites quant à l'identification des structures globales du texte.

Exécuter une analyse narrative assistée par ordinateur à des fins d'analyse de presse ne peut pas se limiter à l'analyse des structures des phrases. Ainsi, adopter une perspective d'analyse qui implique la notion de texte comporte des choix de formalisation et de modélisation différents, qui englobent des structures inter-phrastiques. Les outils de l'apprentissage automatique (*machine learning*), de la fouille de textes (*text mining*) et du traitement automatique du langage naturel (*natural language processing*) constituent déjà un groupe varié d'algorithmes et méthodes qui peuvent être utilisés dans une chaîne de traitement pour l'analyse de texte. Dans ce contexte, la sémiotique computationnelle peut étudier le lien entre les méthodes et les théories développées en sémiotique et les possibilités qu'une analyse assistée par ordinateur offre aux sciences humaines et, plus particulièrement, à la sémiotique.

1.5 Question de recherche

Dans le cadre décrit précédemment, une série de questions surgissent. Dans le nouveau contexte de la sémiotique computationnelle, est-il possible d'assister l'analyse narrative du texte journalistique? Quels pourraient être les outils à insérer dans une chaîne de traitement ayant pour but d'assister l'analyse narrative du texte journalistique? Pour répondre à ce genre de questions, la présente recherche propose de construire une chaîne de traitement pour l'analyse du texte journalistique et de l'évaluer dans le cadre de la sémiotique computationnelle. La méthode proposée ici doit être considérée comme une *démonstration de faisabilité*, c'est-à-dire une étude de faisabilité basée sur la mise en place d'une méthode mise à l'épreuve dans un contexte d'analyse sémiotique et pour un cas d'étude réel. La structure argumentative

de la thèse se base donc sur la construction d'une chaîne de traitement pour l'analyse de textes assistée par ordinateur qui est appliquée à un cas d'étude spécifique et pour lequel on évalue sa pertinence dans un contexte d'analyse sémiotique. Le cas d'étude choisi est le traitement journalistique de la grève étudiante québécoise de 2012 (cfr. 1.7). À cette fin, *un corpus de textes journalistiques* est construit dans un objectif d'analyse.

La thèse ainsi construite suscite des *questions de recherche* qui se situent sur deux plans. L'un regroupe des *questions d'ordre méthodologique* qui mettent de l'avant l'intérêt d'une méthode hybride entre l'informatique et la sémiotique: Est-il possible d'utiliser quelques méthodes et algorithmes pour la découverte des macrostructures textuelles de type narratif? Quel est le point d'ancrage sur lequel construire le lien entre les théories sémiotiques des macrostructures et les méthodes computationnelles? Quels sont les avantages de l'assistance informatique pour l'analyse sémiotique? Toutes ces questions ne peuvent être répondues ou soulever de pistes de réflexion que par le biais d'une étude de faisabilité. L'exploration des méthodes doit alors se faire dans un cadre de découverte empirique, à travers l'analyse d'un corpus de texte.

Une deuxième série de *questions* porte sur la *dimension cognitive* d'une étude de cas, c'est-à-dire des questions typiquement sémiotiques mettant de l'avant l'intérêt de la recherche à découvrir des informations sur le cas à l'étude. Quelles sont les caractéristiques principales de la couverture journalistique sur le mouvement de grève étudiante? Quelles sont les thématiques les plus abordées ou les événements les plus racontés et comment ont-ils été traités? Quelles sont les macrostructures les plus importantes du traitement journalistique du mouvement de grève? Répondre à ce genre de questions requiert la construction d'un corpus de textes journalistiques et la description de leurs principales caractéristiques, ce qui constitue une étude typique en science humaines et plus particulièrement, en sémiotique.

Ces deux types de questions sont complémentaires. L'étude de cas permet ainsi d'évaluer la pertinence d'une nouvelle méthode d'analyse pour répondre à des questions de recherche typiques des sciences humaines, comme par exemple, décrire le traitement journalistique d'un phénomène socio-politique.

Les *questions de recherche* peuvent alors se résumer ainsi:

Question de recherche A) compte tenu de l'existence des macrostructures comme dispositif sémiotique qui garantit la cohérence textuelle et discursive, *quelles sont les macrostructures les plus importantes de la couverture journalistique du printemps érable?*

Question de recherche B) compte tenu de l'importance de l'analyse narrative pour l'étude du texte journalistique, *comment peut-on assister computationnellement l'analyse narrative du texte journalistique?*

Ces questionnements nous mènent aux *objectifs de recherche* suivants:

Objectif A) explorer la couverture médiatique du cas d'étude en se basant sur le concept de macrostructure

Objectif B) exécuter cette analyse au moyen d'une méthode d'analyse de texte assistée par ordinateur qui associe un outil computationnel spécifique à l'analyse narrative du texte journalistique.

1.5.1 Hypothèses

Les postulats et les hypothèses qui régissent le raisonnement et la structure argumentative de la thèse sont séparés en deux catégories, soit sémiotique et computationnelle. D'une part, les hypothèses sémiotiques identifient un procédé sémiotique qui fonctionne comme un ancrage à l'analyse des macrostructures d'un

corpus. D'autre part, les hypothèses computationnelles identifient la nature computable du procédé sémiotique sélectionné et proposent le transfert d'une tâche d'analyse sémiotique vers un modèle formel et computable. L'analyse narrative assistée par ordinateur doit alors être composée d'une hypothèse sémiotique sur les mécanismes de la signification et d'une hypothèse sur l'utilisation d'un outil computationnel capable d'exécuter une tâche de l'analyse dans un contexte computable, soit l'assistance par ordinateur.

1.5.1.1 Hypothèses sémiotiques

Une des manières d'aborder l'analyse de textes est de postuler l'existence de macrostructures qui en soutiennent le sens global et permettent de maintenir sa cohérence. Ces macrostructures ont une nature sémantique et peuvent être étudiées sous différents angles et perspectives. La détection des structures narratives est une des manières possibles pour étudier les macrostructures d'un texte. L'importance de ce postulat pour l'analyse du texte journalistique a été décrite dans les paragraphes précédent (cfr. 1.5). Le défi de la présente recherche est donc d'identifier une façon d'exécuter une analyse narrative du texte journalistique à l'aide de l'ordinateur. Dans ce contexte, il est nécessaire de trouver une hypothèse de recherche qui fasse le lien entre sémiotique et computation. En ce sens, une *hypothèse de type sémiotique* doit être énoncée pour permettre l'identification d'une tâche d'analyse qui puisse être transférée dans un procédé algorithmique et formel tout en restant cohérente avec l'hypothèse sémiotique de base. En d'autres termes, il faut énoncer une hypothèse sémiotique qui *identifie un procédé sémiotique de nature computable*. L'*hypothèse sémiotique* explorée dans ce travail de recherche est la suivante:

Hypothèse A) lors de l'analyse d'un corpus de textes journalistique, *les articles qui partagent les mêmes macrostructures sont identifiables par la détection de schémas récurrents de nature lexicale*, ce qui

> implique que *ces articles sont similaires du point de vue sémantique.*

Cette hypothèse s'appuie sur l'idée que la similarité sémantique entre textes est déterminée par une *régularité lexicale*, c'est-à-dire par la répétition, dans les textes similaires, des mêmes éléments lexicaux. La récurrence de ces éléments lexicaux identifie un *schéma récurrent*, c'est-à-dire une structure lexicale. Ainsi, puisque chaque unité lexicale possède son contenu sémantique, un *schéma lexical correspond à un schéma sémantique* qui est appréhendé comme la macrostructure sémantique partagée. *La régularité lexicale permet ainsi d'identifier la macrostructure sémantique qui est partagée par les articles similaires.* Les schémas récurrents se forment alors en fonction des régularités sémantiques communes aux textes similaires. En d'autres termes, les articles qui traitent des mêmes événements utilisent un vocabulaire similaire et sont analogues du point de vue sémantique. Le concept de macrostructure sera plus largement traité dans le prochain chapitre, car il constitue un élément de l'approche théorique.

À cette hypothèse s'ajoute une seule et unique *sous-hypothèse*:

> Sous-hypothèse A.1) *les schémas récurrents*, qui constituent un phénomène lexical, *correspondent à des structures sémantiques* qui peuvent assumer forme et nature différente. Dans le cadre de cette thèse, nous faisons l'hypothèse que ces structures sémantiques *correspondent à des structures narratives* ayant la forme du *schéma actantiel.*

L'analyse narrative du texte peut être exécutée à travers diverses théories et méthodes. Une des méthodes les plus répandues en sémiotique est celle développée par Greimas. Dans le métalangage qu'il a développé pour la description des phénomènes sémiotiques, Greimas a souligné l'importance de regrouper les rôles narratifs d'un récit dans un modèle, soit le schéma actantiel. Les actants relèvent d'une syntaxe

narrative et permettent d'identifier les éléments qui organisent la structure sémantique d'un texte. La présente sous-hypothèse implique donc l'identification du schéma actantiel partagé par des articles similaires.

En résumé, le schéma récurrent est un concept opératoire qui désigne le phénomène de répétition des mêmes éléments lexicaux. Ce schéma correspond à l'émergence d'une macrostructure sémantique, qui apparaît dans plusieurs textes similaires du point de vue sémantique. Un schéma récurrent doit donc comprendre des composantes lexicales identifiables qui doivent se retrouver dans plusieurs textes. Ainsi, l'hypothèse sémiotique principale est que les textes similaires partagent un même schéma récurrent, ce dernier étant identifiable par un schéma actantiel. L'hypothèse s'appuie sur le postulat de l'existence des macrostructures de nature narrative. L'*identification d'articles similaires* du point de vue sémantique constitue une étape fondamentale de l'analyse et de l'*ancrage sémiotique* pour l'utilisation d'outils computationnels.

1.5.1.2 Hypothèses computationnelles

À ce stade, il est nécessaire d'identifier un outil computationnel qui puisse soutenir la tâche d'identification de macrostructures. Si ces structures sont identifiables par des schémas lexicaux récurrents dans un groupe d'articles similaires, la méthode computationnelle peut se limiter à *repérer les articles similaires sur la base de la dimension lexicale des textes*. Un algorithme appartenant à la fouille de données et à l'apprentissage automatique devra alors être identifié pour exécuter cette tâche de repérage ou de *regroupement d'articles similaires*. Une *phase de lecture et d'annotation* pourra ensuite être exécutée pour déterminer les macrostructures sémantiques qui correspondent aux schémas lexicaux de chaque groupe d'articles similaires.

À partir de l'hypothèse sémiotique énoncée au paragraphe précédent, quelques postulats et une hypothèse computationnelle peuvent être énoncés. Le premier postulat affirme que :

Postulat A) un texte peut être représenté dans un contexte formel et computable par une *transformation vectorielle*.

La transformation d'un texte dans une matrice est un des modèles les plus utilisés en intelligence artificielle pour sa manipulation. La construction d'une matrice représentant un texte constitue ainsi le contexte computationnel à partir duquel la méthode proposée peut être développée. Cette thèse ne remet pas en cause la légitimité d'un tel modèle et le considère donc comme un postulat. Le modèle computable qui représente un texte est appelé *modèle sémantique vectoriel* et il est détaillé dans le prochain chapitre puisqu'il représente un élément de notre approche théorique.

Le deuxième postulat énonce que :

Postulat B) une famille particulière d'algorithmes, le *clustering*, peut être appliquée à un modèle sémantique vectoriel pour la *reconnaissance de groupes d'articles similaires*.

Le clustering a toujours été utilisé en fouille de textes pour le regroupement de documents similaires et sa légitimité n'est donc pas discutée dans cette thèse. Ces deux postulats amènent à considérer l'*hypothèse de recherche* qui suit:

Hypothèse B) *le clustering permet de repérer des groupes d'articles similaires qui partagent les mêmes macrostructures de type narratif.*

Cette hypothèse de recherche et complémentaire à l'hypothèse sémiotique A). Ces macrostructures correspondent ainsi à des schémas lexicaux récurrents qui contribuent à la détermination des groupes d'articles similaires dans un contexte

computationnel. L'application d'une telle méthode permet de réduire le temps d'analyse de corpus de grande taille et apporte des avantages considérables à la pratique d'analyse sémiotique des textes journalistiques.

1.6 Les objectifs poursuivis

Dans 1.5 nous avons défini les objectifs de recherche en relation aux quesitons de recherche. Ces objectifs sont surtout liés à la nature de cette thèse, qui a un caractère méthodologique. En effet, ils expriment des attentes en relation à des aspects de la méthode qui sera explorée par la chaîne de traitement. Toutefois, nous voulons aussi souligner deux objectifs plus généraux qui sont poursuivis par la présente thèse.

Le premier est le suivant :

> Objectif général A) exécuter une véritable analyse du cas d'étude afin de *dégager les macrostructures les plus importantes du traitement journalistique du mouvement étudiant québécois de 2012.*

Atteindre ce premier objectif signifie obtenir de résultats significatifs et valables quant au cas d'étude, ce qui constitue à coup sûr un élément de réussite de la présente recherche.

Le second objectif general est le suivant :

> Objectif général B) *explorer le clustering comme méthode d'assistance à l'analyse narrative de corpus de textes journalistiques.*

Pour y arriver, il est nécessaire de créer une chaîne de traitement constituant une *démonstration de faisabilité*. Pour ce faire, la chaîne de traitement est utilisée pour l'analyse d'un cas d'étude. Si les résultats de l'analyse sont valides, il sera alors possible d'évaluer et de discuter de la faisabilité d'une assistance informatique à

l'analyse narrative de textes journalistiques, en soulignant les avantages et les inconvénients de l'utilisation d'une telle méthode. Le choix du cas d'étude est donc déterminant pour la réussite de l'étude de faisabilité puisque les résultats de l'analyse constituent le matériel sur lequel la méthode est évaluée. La présente thèse s'attache donc, via l'analyse d'un cas d'étude, à présenter une méthodologie hybride pour l'analyse narrative de textes journalistiques.

Cette thèse contribue principalement au transfert de connaissance de l'informatique à la sémiotique et à la mise en place de pistes et hypothèses de recherche pour la sémiotique computationnelle.

1.7 Le cas d'étude : le printemps érable

L'expression imagée de printemps érable désigne une grève étudiante qui a eu lieu au Québec entre les mois de février et septembre 2012 et qui a mobilisé jusqu'à 175 000 étudiantes et étudiants québécois afin de protester contre la hausse des droits de scolarité prévue par le gouvernement libéral du premier ministre Jean Charest, lors de la 39e législature du Québec. Le printemps érable a ouvert un débat public sur le statut de l'éducation au Québec et a impliqué plusieurs dizaines de milliers de citoyens, ce qui de fait a transformé une grève étudiante en un véritable mouvement social. La couverture journalistique de ce phénomène socio-politique a été totale et, pour ces raisons, il constitue un cas d'étude idéal pour présenter une méthodologie nouvelle.

1.7.1 Description des événements principaux

À la fin de 2008, les élections générales québécoises amènent au pouvoir un gouvernement composé d'une grande majorité de députés du Parti Libéral du Québec (PLQ) dirigé par Jean Charest, qui est ainsi devenu premier ministre du Québec pour

la troisième fois de suite. Le ministre de l'éducation est alors Line Beauchamp qui démissionne pendant la grève. Elle est remplacée par Michelle Courchesne.

Au début de 2010, le gouvernement annonce son intention de hausser les droits de scolarité universitaires de 75% et d'étaler cette hausse sur une période de cinq ans à compter de 2012, ce qui correspond à une augmentation de 1 625 $ par année par étudiant. Le gouvernement considère la hausse comme une source supplémentaire de financement du système d'éducation au Québec. En particulier, les universités québécoises souffrent d'un déficit budgétaire important ce qui, selon le gouvernement, les empêche d'être compétitives au niveau national et international.

Le choix de financer le système d'éducation par une manœuvre touchant surtout les étudiants est justifié par le gouvernement par le constat que les étudiants ne contribuent pas assez au soutien financier de l'éducation et ce en comparaison avec les étudiants des autres provinces canadiennes. Cette contribution est demandée comme « une juste part » des étudiants à la formation universitaire qui aurait un impact favorable en terme de « rentabilité économique privée ». En effet, selon le gouvernement, un travailleur détenteur d'un diplôme d'étude secondaire ajouterait près de 600 000 $ au cours de sa vie active s'il possédait un diplôme universitaire.

Dès le début, les fédérations étudiantes québécoises manifestent leur désaccord avec cette proposition au moyen de différentes actions, comme une pétition réunissant 30 000 signatures, plusieurs manifestations mobilisant quelque dizaines de milliers d'étudiants et des occupations de lieux à l'intérieur des campus universitaires. Ces actions ne convainquent pas le gouvernement d'amorcer des discussions avec les fédérations étudiantes.

Le 13 février 2012, en raison de l'absence de réponses gouvernementales, plusieurs associations étudiantes déclenchent une grève illimitée et ce, avec l'appui de la

majorité de leurs membres. Dans son moment le plus intense, la grève rejoint 175 000 étudiants, plus de la moitié des 342 000 étudiants québécois.

Les principales associations étudiantes qui représentent les grévistes sont : la Fédération Étudiante Universitaire du Québec (FEUQ) dont le président est Martine Desjardins, la Fédération Étudiante Collégiale du Québec (FECQ) dont le président est Léo Bureau-Blouin et la Coalition large de l'Association pour une solidarité syndicale étudiante (CLASSE), qui est un regroupement temporaire créé spécifiquement par l'Association pour une solidarité syndicale étudiante (ASSÉ) pour lutter contre la hausse des droits de scolarité et qui comporte deux porte-paroles, Gabriel Nadeau-Dubois et Jeanne Reynold. La CLASSE n'a pas de président. Les trois associations ne partagent pas la même plateforme de revendications car la CLASSE, considérée comme l'association la plus radicale, se distingue des deux autres par la revendication de la gratuité scolaire des études post-secondaires. Au contraire, la FEUQ et la FECQ proposent pour leur part des solutions moins radicales, comme un moratoire de la hausse de droits de scolarité.

Le principal élément du débat dans les associations étudiantes et entre celles-ci et le gouvernement est la vision de l'éducation publique au Québec. Si le gouvernement prône l'idée d'avoir un système éducatif compétitif et de qualité grâce à un financement supplémentaire auquel devraient contribuer principalement ses principaux bénéficiaires, les étudiants de leur part mettent de l'avant l'idée d'un système éducatif accessible et inclusif rendu possible par la réduction des coûts de scolarité.

De février à avril 2012, plusieurs manifestations, piquetages et occupations représentent le cœur des activités des grévistes. Le 22 mars, une manifestation de 200 000 personnes contre la hausse de droits de scolarité défile dans les rues de Montréal,

ce qui constitue la plus grande manifestation populaire enregistrée au Québec depuis celle du 2003 contre la guerre en Iraq. Pendant cette longue période, le gouvernement refuse toujours de rencontrer les étudiants et de mettre sur pied une table de concertation, ce qui augmente la tension et suscite la colère des étudiants. Les manifestations sont souvent le théâtre de durs affrontements entre les manifestants et la police. L'affrontement le plus violent a lieu à Victoriaville le 4 mai, en marge du conseil général du PLQ et deux manifestants y sont gravement blessés.

La stratégie principale des grévistes pour amener le gouvernement à discuter consiste à perturber le plus possible la vie citadine au moyen de manifestations et actions de différents types. Les actions de perturbation sont assez diversifiées et vont de la manifestation pacifique au blocage de pont ou aux actes de vandalisme contre les symboles du pouvoir libéral, comme la fenêtre du bureau du ministre de l'éducation, ou du pouvoir économique, comme les vitrine des banques.

L'usage de la force policière pour réprimer les manifestations est très répandu et suscite de vives critiques. Pendant la grève des étudiants, selon le rapport de la commission sur les événements du printemps 2012 du gouvernement du Québec, la police de Montréal (SPVM) intervient dans 532 cas et exécute 12 opérations d'arrestation de masse (CSEP 2012, 2014), la police de Québec (SPVQ) intervient dans environ 200 situations avec huit arrestations de masse, la police de Gatineau (SPVG) dans 163 manifestations avec deux arrestation de masse, et enfin, la Sûreté du Québec (SQ) intervient dans 473 situation, avec une opération d'arrestation de masse. La police fait large usage d'outils de multiplication de la force, comme le poivre de Cayenne, les grenades assourdissantes, le gaz lacrymogène, les bâtons et les balles en plastique. Le 7 mars, lors d'une manifestation devant les bureaux de la conférence des recteurs et des principaux des universités du Québec (CRÉPUQ), le

manifestant Francis Grenier perd l'usage d'un œil à cause d'une grenade assourdissante.

Cette longue période de grève et de manifestations provoque également la suspension de plusieurs cours universitaires et collégiaux. Le blocage de cours amène les tribunaux à émettre des injonctions forçant les professeurs à donner leurs cours et interdisant les piquetages ou les grands rassemblements sur les campus Les établissements d'enseignement opposent une vive résistance aux actions perturbatrices posées par les étudiants et utilisent à plusieurs reprises leurs forces de sécurité pour les contrer. La police est également appelée de temps en temps pour rétablir l'ordre dans les campus universitaires.

À la fin du mois d'avril, le gouvernement décide d'entamer des discussions avec les associations étudiantes afin de trouver une sortie de crise. La ministre Beauchamp exclut la CLASSE de la première rencontre parce qu'elle la tient responsable des actes de violence commis à plusieurs reprises pendant les manifestations, reproduisant ainsi la même stratégie adoptée pendant la grève étudiante de 2005. Les deux fédérations étudiantes, FEUQ et FECQ, refusent alors de rencontrer le gouvernement en absence de la CLASSE, démontrant ainsi une grande solidarité et une union solide entre les étudiants en grève. La ministre Beauchamp décide ensuite de rencontrer les étudiants pour discuter du système de prêts et bourses et non pas de la hausse de droits de scolarité. Le 23 avril, les travaux de la table de concertation s'amorcent. La CLASSE est exclue à nouveau après 24 heures, car elle est tenue responsable d'une action militante du Black Bloc. À la fin des travaux de la table de concertation, le PLQ annonce un nouveau plan qui prévoit une augmentation de 82% sur sept ans, plutôt que 75% sur 5 ans. L'offre est considérée inacceptable par les associations étudiantes qui préparent une contre-offre et demandent un moratoire.

Le 5 mai, le débat se déplace vers la gestion des universités et le gaspillage de ressources. À la fin des discussions de la table de concertation, la proposition du gouvernement porte sur la création d'un comité qui devra étudier les lacunes de gestion des universités afin de trouver les ressources financières équivalentes à la hausse proposée, dans le but de la réduire, voire même de l'annuler. Le comité serait composé de 19 personnes dont quatre représentants des étudiants. Après une consultation auprès de leurs membres, les associations étudiantes refusent la proposition.

Le refus de la proposition du 5 mai conduit à la démission de la ministre Line Beauchamp qui est remplacée par Michelle Courchesne. L'ex-ministre accuse les étudiants de s'être campés sur leurs positions et admet son incapacité à résoudre le conflit étudiant par le dialogue. Le 16 mai, le gouvernement annonce le dépôt d'une loi spéciale ayant pour objectif de mettre fin au conflit étudiant.

Le dépôt du projet de loi spéciale (projet de loi 78 qui deviendra la «loi 12») est un tournant important dans la crise et mène à une confrontation encore plus dure entre le gouvernement et les étudiants. La loi prévoit la suspension du trimestre d'hiver dans les universités et les cégeps, l'interdiction de faire du piquetage ou de bloquer l'accès aux cours et une série de restrictions aux droits de manifester et de se rassembler. Des amendes allant de quelques centaines de dollars à 125 000 $ sont prévues dans différentes situations. De plus, la loi permet aux universités de cesser la perception de la cotisation étudiante pour les associations considérées responsables d'actes de violence. Le 18 mai, le projet de loi 78 est adopté. Le même jour, la ville de Montréal introduit le règlement P-6 qui interdit le port du masque en public pendant les manifestations afin de faciliter l'identification des manifestants violents.

Le jour de l'adoption du projet de loi 78, une dizaine de manifestations spontanées ont lieu à Montréal, démontrant ainsi un grand appui populaire à la cause des étudiants et, surtout, un rejet de la loi 12. À partir de ce moment, des milliers de personnes défilent dans les rues de Montréal avec des casseroles à l'image des « Cacerolazo » de l'Amérique du sud. Cette forme de protestation est caractérisée par l'utilisation des casseroles pour faire du bruit pendant les manifestations. Le mouvement étudiant se répand donc définitivement à toute la société québécoise.

La loi 12 a conduit de plus à une criminalisation du mouvement étudiant et à plusieurs arrestations de masse. La CLASSE annonce qu'elle ne respectera pas la loi spéciale et promeut la désobéissance civile. En vertu des dispositions de la loi 12, les manifestations sont souvent déclarées illégales par la police qui utilise la force pour les réprimer et disperser les foules.

Le 28 mai, une nouvelle table de concertation est mise sur pied puisque la loi spéciale ne permet pas de diminuer la mobilisation étudiante. Les discussions se poursuivirent pendant quatre jours et le débat se concentre, pour la première fois, sur la hausse de droits de scolarité. Cependant, la table de concertation ne produit aucun résultat satisfaisant. Elle est la dernière possibilité de résoudre la crise par le dialogue.

Malgré l'arrivée de l'été, la mobilisation se poursuit. Les divers festivals d'été de Montréal et le Grand Prix risquent d'être perturbés par la crise étudiante en cours. Les affrontements entre policiers et étudiants se poursuivent régulièrement et les présidents de festivals et d'autres personnages publics interviennent dans le débat sur afin de sauver la saison touristique de Montréal.

Le premier août, le gouvernement Charest décide de déclencher des élections, parce que selon le ministre des Finances Raymond Bachand, elles sont désormais le seul moyen permettant de résoudre le conflit. Les élections ont lieu le 4 septembre 2012 et

un gouvernement minoritaire mené par le PQ de Pauline Marois est élu. Jean Charest, après plus de vingt ans comme député, est défait dans sa propre circonscription et le PLQ perd plus de 10% des électeurs qui l'avaient soutenu en 2008.

1.7.2 La couverture médiatique

Le printemps érable a fait l'objet d'une couverture médiatique complète et ce, dans plusieurs types de médias. La presse écrite a été une des principales sources d'information et a constitué un des moyens majeurs d'expression des acteurs impliqués dans le débat, ce qui en fait une composante importante pour la construction de l'opinion publique. Par exemple, selon une étude publiée par Influence Communication, un courtier en nouvelles, *Le Devoir*, *La Presse* et *The Gazzette* ont traité le conflit étudiant à la une dans 73,5% des cas et 42% pour le *Journal de Montréal* (Influence Communication, 2012b). La même firme a aussi constaté dans son bilan annuel que le conflit étudiant a été une des nouvelles le plus longuement traitées dans l'histoire des média québécois (Influence Communication, 2012a).

Selon un rapport (Giroux et Charlton, 2014) du Centre d'étude sur les médias (Université de Laval), *La Presse* est le journal qui a consacré le plus d'espace au conflit étudiant, avec 31% de la couverture totale des quatre journaux les plus importants de Montréal. *Le Devoir* a une couverture de 27%, le *Journal de Montréal* 22% et *The Gazette* 20%. L'attention de ces quotidiens s'est accentuée au fur et à mesure de l'ampleur prise par la grève , atteignant son apogée entre le 20 et le 26 mai, la semaine de l'adoption du projet de loi 78. Plus de 80% des articles ont pris la forme de nouvelles, de chroniques et de lettres d'opinion. La part la plus importante va à la catégorie nouvelle, qui constitue plus de 50% du matériel publié. Plus de 80% du matériel de *La Presse* a été signé par un journaliste alors que cette proportion s'élève à 75% pour *The Gazette*, 70% pour le Journal de Montréal et 63% pour le

Devoir. Enfin, les médias et leurs chroniqueurs ont pris une part active dans le débat et leur influence dans la formation de l'opinion publique au sujet du printemps érable a suscité plusieurs questionnements d'ordre éthique et politique (Tremblay *et al.*, 2015). De nombreux ouvrages traitant du printemps érable ont été publiés et il est extrêmement difficile d'en faire une synthèse.

1.7.3 Quelques corpus sur le printemps érable

Dans le but d'obtenir un cadre de comparaison, des travaux ayant étudié le traitement journalistique du printemps érable en se basant sur un corpus de texte sont présentés. Nous avons sélectionné ceux qui ont été constitués à des fins similaires à celles de la présente étude. Ces corpus ne constituent pas une liste exhaustive et toute comparaison ne peut donc pas être complète. Toutefois, cet ensemble de corpus et d'études sur le printemps érable permet de positionner partiellement la présente étude dans le cadre des recherches en sciences humaines.

Le premier corpus fait partie d'une étude conduite par *Influence Communication*, un « courtier en information média spécialisé dans la surveillance, la synthèse et l'analyse de contenus de médias imprimés et électroniques » (Influence Communication, 2018). Cette étude (Influence Communication, 2012b) a analysé les pages de couverture (les unes) de quatre quotidiens de Montréal, soit *La Presse*, *Le Devoir*, *Le Journal de Montréal* et *The Gazette*. La période couverte va du 15 février 2012 au 9 juin 2012, c'est-à-dire des premiers votes de grève à une semaine après la rupture des négociations entre les associations étudiantes et la ministre Michelle Courchesne. L'objectif de recherche était « d'évaluer l'ampleur de la couverture du conflit étudiant sur la première page des quatre quotidiens ». Le corpus comptait 396 pages. L'étude présente un certain nombre de conclusions, tel le fait que 63,14 % des quatre quotidiens ont fait mention du conflit étudiant pendant la période analysée ou que *Le Journal de Montréal* est celui qui a affiché le moins d'évènements liés au

printemps érable dans ses premières pages, ou encore que *Le Devoir* a montré le plus souvent des photos de manifestations pacifiques alors que *Le Journal de Montréal* a montré le plus souvent des photos de manifestations violentes.

Un rapport du *Centre d'études sur les médias* de l'Université Laval publié en février 2014 (Giroux et Charlton, 2014), a conduit une étude sur les mêmes sources d'information, soit *La Presse*, *Le Devoir*, *Le Journal de Montréal* et *The Gazette*. La période de couverture de cette étude est du 13 février au 23 juin 2012 ce qui, selon les auteurs, couvre « les évènements les plus importants » de ce qu'on appellera le printemps érable. L'information sur la taille du corpus et sur la méthode de collecte des données n'a pas été dévoilée par les auteurs, mais on connaît la taille d'un de leurs sous-corpus, formé de 185 unités. Toutefois, il est fort probable que l'étude ait été basée sur plus de 1 000 articles. L'objectif de la recherche était d'étudier le traitement journalistique du conflit étudiant et, plus particulièrement, de « décrire les caractéristiques des productions journalistiques telles qu'elles peuvent être perçues par les lecteurs » (*ibid.*). Pour ce faire, une méthode spécifique a été mise au point, sur la base d'une grille d'analyse déterminée en amont et sur une équipe de « codeurs-analystes ». L'étude est large et composée de plusieurs sous-analyses, dont nous résumons brièvement quelques conclusions du point de vue quantitatif. Le rapport affirme que *La Presse* est le quotidien qui a consacré le plus d'espace au conflit étudiant, atteignant 31 % de la couverture totale des quatre quotidiens[2]. Le journal *Le Devoir* est celui qui a consacré le plus grand contenu rédactionnel au conflit (11 %). La figure 1.2 montre l'évolution de la couverture tout au long des semaines.

---

[2] *Le Devoir* a réservé 27 % de son contenu rédactionnel, *Le Journal de Montréal* 22 % et *The Gazzette* 20 %.

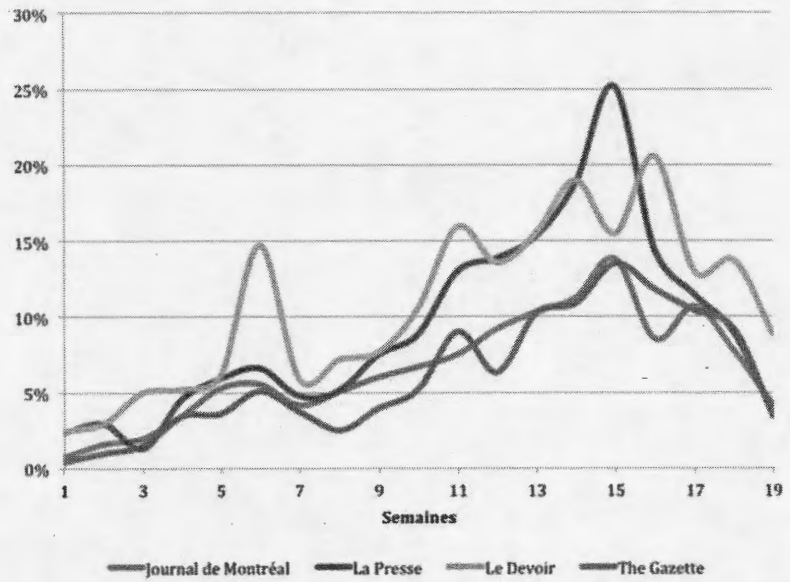


Figure 1.2 Évolution hebdomadaire par quotidien du contenu rédactionnel consacré à la crise (Giroux et Charlton, 2014, p. 8)

Selon cette même étude, plus de 80 % des textes portant sur le printemps érable ont pris la forme de nouvelles, de chroniques et de lettres d'opinion, les nouvelles, c'est-à-dire la catégorie « Actualité », constituant le genre journalistique dominant dans tous les quotidiens. Ces considérations coïncident avec les caractéristiques générales de notre corpus (cfr. 4.1.3), sauf pour *Le Devoir*, où la distinction des catégories est sensiblement différente des autres journaux. La majorité des autres informations quantitatives exposées par l'étude de l'Université Laval coïncident avec celles de notre présent corpus. D'autres résultats sur les thématiques les plus abordées par les quatre quotidiens constituent également une contribution importante pour la compréhension du traitement journalistique du printemps érable. Bien que le corpus de cette étude ne peut pas être évalué, car ni sa taille ni sa méthode de collecte ne sont connues, ses résultats sur les thématiques les plus importantes seront repris dans le chapitre cinq à des fins de comparaison. Cette étude est celle dont les objectifs coïncident le plus avec ceux de cette thèse.

Plusieurs autres travaux académiques sur le traitement journalistique du printemps érable ont été produits. Les méthodes et les approches sont très différentes les unes des autres et les études ne sont pas toutes comparables. Une première étude (Cléroux, 2015) répond à la question de recherche suivante : « comment le printemps érable a-t-il été couvert par les journalistes québécois? » À la différence des autres, ce travail s'est majoritairement concentré sur les aspects idéologiques et les « logiques » extrajournalistiques qui ont influencé le travail des journalistes. Le corpus est composé d'entretiens semi-dirigés auprès des professionnels de l'information. Toutefois, une analyse de contenus a également été conduite et représente un élément fondamental de l'argumentation de l'étude. Cette analyse a été réalisée sur un corpus très mince, composé de quelques articles publiés au lendemain des plus grandes manifestations étudiantes qui se sont tenues chaque 22e jour du mois, de février à août. Un tel choix est basé sur le fait que ces manifestations représentent des événements importants du printemps érable et qu'elles « sont passées à l'Histoire en raison du nombre de gens qu'elles ont mobilisés ». Les sources d'information retenues sont les mêmes que dans les deux autres études, soit *La Presse*, *Le Devoir*, *Le Journal de Montréal* et *The Gazette*. Ce choix est justifié, selon les auteurs, par le fait que ces journaux sont les quotidiens payants les plus lus au Québec et les plus cités à la télévision ainsi qu'à la radio, « ce qui augmente considérablement leur impact potentiel sur l'opinion publique » (*ibid.* p. 44). Les articles d'opinion ont été écartés et seules les « Actualités » ont été retenues. Positionnement de l'article, mise en page et photos n'ont pas été retenus non plus dans l'analyse. Enfin, le corpus est composé de 30 articles appartenant à la catégorie « Actualités ». Chaque texte analysé a été divisé en énoncés, chacun étant situé à l'intérieur d'une catégorie et d'une grille d'analyse particulière. Un corpus ainsi construit réduit la représentativité et le potentiel de généralisation des résultats à l'ensemble de la couverture journalistique du printemps érable, ce qui est reconnu par les auteurs eux-mêmes, lesquels confirment le risque d'un biais subjectif de la méthode (*ibid.* p. 49-50). Cette étude

montre en effet l'importance d'avoir un corpus de grande taille et, surtout, de le construire en suivant des critères bien précis et mettant en valeur l'orientation du corpus et sa représentativité. Cette étude illustre aussi la difficulté de construire et d'analyser un corpus de grande taille dans le cadre de recherches en sciences humaines et sociales.

Un deuxième travail (Hébert È.-L., 2017) a utilisé une méthode expérimentale (Grelley, 2012) pour « déceler les effets de la couverture médiatique de la grève étudiante de 2012 au Québec ». Pour ce faire, les chercheurs ont rassemblé 166 étudiants de l'Université du Québec à Montréal, selon une méthode de recrutement occasionnelle et aléatoire, ce qui mène à des résultats non généralisables à l'ensemble de la population québécoise (*ibid.* p. 43). Les raisons d'un tel échantillon sont essentiellement liées à la méthode utilisée qui ne peut pas être appliquée à des milliers d'individus sans la mise en place de ressources adéquates (*ibid.* p. 44). Chacun des étudiants a dû répondre à deux questionnaires qui visaient à la génération des données sur les effets du traitement médiatique du printemps érable, ce qui a constitué le matériel empirique sur lequel l'étude a été menée. Les données ont été analysées par des modèles de régression linéaire. Les auteurs concluent en confirmant que l'idéologie de chaque individu de l'échantillon a un effet « direct et significatif sur [les] trois variables dépendantes : les attitudes, les émotions et la propension à participer » (*ibid.* p. 80), et que les principaux changements de position idéologique ont concerné les étudiants ayant une idéologie de droite ou de centre.

Une autre étude (Dussault-Brodeur, 2015) a analysé le discours afin de dégager les caractéristiques du débat autour de la notion de « violence » présente pendant la crise étudiante de 2012. Le corpus est composé d'énoncés prononcés par les autorités politiques et publiés dans la presse écrite francophone montréalaise. La collecte a été effectuée avec la base de données *Eureka* et les sources d'information retenues sont

La Presse, *Le Devoir* et *Le Journal de Montréal*. Les articles ont été sélectionnés sur la base d'un certain nombre de mots clés[3] pour la période qui va du 1er février au 1er novembre 2012[4]. Malgré ses limites, les auteurs ont construit un corpus bien orienté sur la question de recherche qui vise à analyser « le niveau de criminalisation de la contestation ». En effet, la majorité des mots clés utilisés pour la collecte cible volontairement le vocabulaire spécifique utilisé pour nommer les « *manifestants violents* », ce qui facilite une collecte de données orientée et pertinente. Enfin, selon le critère de pertinence, les auteurs ont retenu 131 articles pour composer le corpus de référence. Leur corpus d'étude est constitué seulement des articles contenant des déclarations des autorités politiques, soit 42 articles. L'étude a décelé la manière à travers laquelle le discours des autorités politiques sur la violence dans les manifestations était au service d'une stratégie politique précise. Ce discours menait le plus souvent à la criminalisation de la grève étudiante et, plus particulièrement, au refus de reconnaître politiquement le « mouvement qui contestait l'ordre ».

Une autre étude a analysé le traitement du concept de « désobéissance civile » dans la presse au moyen d'une analyse du discours menée sur un corpus comprenant 176 documents, soit 107 commentaires d'opinion, 46 nouvelles, 9 interviews, 2 enquêtes sous forme d'interview et 12 analyses. La période de couverture va du 17 février au 16 novembre 2012 et les sources d'information sont les suivantes : *La Presse*, *Le Devoir* et *Le Journal de Montréal*. L'étude contribue à déceler un élément important du débat qui a eu lieu lors de la crise étudiante, débat lié à la vision que les différents acteurs avaient de la démocratie.

---

[3] Mots clés : « grève étudiante » et « ministre », « grève étudiante » et « violence », « grève étudiante » et « casseurs », « grève étudiante » et « criminels », et uniquement « criminels ».

[4] Les dates ont été choisies de manière approximative afin de couvrir l'entrée et la grève étudiante pour les différentes associations étudiantes.

Enfin, d'autres types d'étude ont été conduits sur une seule source d'information. Un exemple en est une étude de cas sur le journal étudiant *Montréal Campus*, lequel, selon ses auteurs (Desrochers, 2016) est caractérisé par sa neutralité. Afin de « mesurer » le niveau de neutralité de ce journal, les auteurs ont regroupé 87 articles allant du 30 mars 2010, date de l'annonce de la hausse des droits de scolarité, jusqu'en février 2013, lors du Sommet sur l'enseignement supérieur, point final du mouvement étudiant.

Toutes ces études démontrent comment l'analyse d'un corpus, par le biais de différentes méthodes et cadres théoriques, peut répondre à plusieurs types de questions de recherche. Chacun de ces corpus est orienté sur un objectif de recherche bien précis et possède des forces et des faiblesses. En général, *on remarque la difficulté à utiliser*, dans un cadre de recherche en sciences humaines et sociales, *un corpus de grande taille*. En effet, la majorité des corpus présentés ne dépassent pas les quelques dizaines de documents. Le développement de méthodes d'analyse de textes assistée par ordinateur peut sans doute contribuer à la réalisation d'analyse de corpus de plus grande taille, ce qui pourrait mener au renouvellement des méthodes en sciences humaines et sociales et, comme l'affirme Rastier à plusieurs reprises, ce qui pourrait ouvrir la porte à la découverte de « nouveaux observables ».

CHAPITRE II

CADRE THÉORIQUE

Ce chapitre présente en trois parties le cadre théorique de notre étude. La première partie est dédiée à l'approche sémio-computationnelle qui forme la base sur laquelle la relation entre l'approche sémiotique et l'approche computationnelle s'édifie. Elle est constituée de deux éléments. Le premier correspond à la théorie de la *macrostructure* qui permet d'avoir une vision globale du texte et de le traiter comme un produit sémiotique fondé sur le principe de *cohérence textuelle* soutenu par une quelconque forme de *régularité interne*. Le deuxième élément constitue un cadre pour l'identification des procédés sémiotiques de nature computable. Un des défis de la thèse est d'identifier un moyen pour mettre en relation les méthodes d'analyse des sciences humaines avec celles de l'informatique et de l'analyse de données. Pour ce faire, il est nécessaire de disposer d'un cadre qui systématise la transformation d'opérations sémiotiques en opérations formelles et computables.

La deuxième partie de ce cadre présente l'*approche sémio-narrative* qui met en place le cadre théorique pour appréhender l'analyse de textes. Cet encadrement dépend des considérations générales reliées à la théorie des macrostructures. Si cette dernière offre une vision générale du texte, l'approche sémio-narrative en permet une qui est spécifique à l'analyse de texte. Enfin, l'*approche computationnelle* est présentée. Elle constitue le cadre théorique des outils informatiques utilisés dans ce travail de recherche.

2.1 L'approche sémio-computationnelle

Nous présentons en premier lieu l'approche qui met en relation la sémiotique et certaines disciplines qui appartiennent à l'intelligence artificielle. Cette approche s'inscrit à l'intérieur du troisième axe de recherche de la sémiotique computationnelle, soit celui qui considère l'ordinateur comme un outil au service de l'*analyse des produits sémiotiques*. Une telle approche suppose d'abord deux types de postulats, l'un sur *les dynamiques globales de la production sémiotique* et l'autre sur *les processus de son analyse*.

Le premier postulat introduit des réflexions générales sur la nature du produit sémiotique qui est analysé et ce afin d'esquisser un profil adéquat de son fonctionnement. Dans cette thèse, la théorie des macrostructures constitue le premier postulat. Les macrostructures régissent les dynamiques de production textuelle et sont principalement à la base de la *compétence textuelle*, laquelle permet de produire des textes cohérents. En d'autres termes, la *cohérence* est une des propriétés les plus importantes d'un texte et elle est garantie par l'existence de macrostructures. Ce dernier concept est approfondi au paragraphe 2.1.1.

Le deuxième postulat esquisse le cadre à partir duquel une analyse de texte assistée par ordinateur peut être appréhendée. L'assistance informatique est encadrée par un système qui se base sur le principe suivant : *permettre le passage de tâches de l'analyse sémiotique vers l'ordinateur*. Afin de permettre ce passage, le système conçoit plusieurs niveaux de description d'un phénomène. Lorsque les niveaux sont cohérents, alors le passage d'une fonction cognitive vers une fonction computable devient possible. Ce système est approfondi au paragraphe 2.1.2.

2.1.1 Les macrostructures

Le concept de macrostructure retenu dans le présent travail est celui développé par le linguiste Teun Adrianus van Dijk, un des pères putatifs de l'analyse du discours. La macrostructure se définit comme une « structure de signification globale d'un texte » (Kintsch et Van Dijk, 1975, p. 101). Elle possède une nature sémantique :

> This means that this kind of notion should be made explicit in semantic terms; to distinguish them from other kinds of global structures, we talk about semantic global structures. It is this type of structure that we try to make explicit in this book in terms of (semantic) macrostructures (Van Dijk, 1980, p. 5)

> It is used to account for the various notions of global meaning, such as topic, theme, or gist. This implies that macrostructures in discourse are semantic objects. (Van Dijk, 1980, p. 10)

La macrostructure demeure toutefois une notion ambiguë utilisée pour se référer à différents concepts. Ces derniers partagent un même dénominateur commun, les renvoyant tous aux notions de « signification globale » et de « structure globale ». En quelque sorte, la macrostructure organise le sens à un niveau global :

> The first function of macrostructures is to *organize* complex (micro)-information. Without them we would only be able to have a large number of links between information units at the local level and not be able to form larger chunks that have their proper meaning and function (Van Dijk, 1980, p. 14)

Quelques éléments supplémentaires cernent davantage la nature du concept de macrostructure, soit la *cohérence textuelle*, le *thème*, le *cœur informatif* et la *structure cognitive et sociale de la représentation de la réalité*. Nous allons les décrire. D'abord, une macrostructure constitue la forme latente qui structure un texte, garantissant ainsi sa *cohérence sémantique*.

> [...] macrostructures in a theory of discourse are necessary to account for the intuitive notion of coherence: A discourse is coherent not only at the local level (e.g., by pairwise connections between sentences) but also at the global level. Notions such as global meaning, global reference, topic, or theme are intimately related, and macrostructures are needed to make these relations explicit. (Van Dijk, 1980, p. 10)

La macrostructure organise donc le sens et permet de rendre cohérent le message qui est construit. Sans la notion de cohérence sémantique, il ne serait pas possible de distinguer des discours différents :

> A particular and important case of this kind of semantic organization is (global) *coherence*: Due to macro structures, discourses, conversations, and action sequences are planned and understood as coherent wholes and hence as a unit that may as such be identified and distinguished from other, similar, objects. Without the macrostructurally formulated notion of coherence, it would not be possible to distinguish one discourse from a following discourse nor one action sequence from another action sequence. (Van Dijk, 1980, p. 14)

Le rôle principal de la macrostructure est celui de liaison entre les éléments d'un texte pour former une structure globale et organisatrice. La macrostructure est une forme d'*organisation textuelle profonde* ayant une forte composante sémantique, comme le dit van Dijk dans le passage suivant : « These macrostructures may be identified with global semantic representations or deep structures of texts ». (Van Dijk, 1972, p. 307). Ainsi, la cohérence textuelle ne peut être que de nature sémantique.

Le lien entre cohérence et sémantique est encore plus évident lorsqu'on associe la macrostructure à la notion de *thème*. En effet, la notion théorique de macrostructure peut être couplée à différentes formes d'organisation sémantique du texte, comme celle de thème. Dans ce dernier cas, la macrostructure correspond à l'*accumulation isotopique* [5], c'est-à-dire à la *répétition régulière et systématique d'éléments*

[5] Nous nous référons au processus d'accumulation de lexèmes, sèmes ou autres qui forment l'isotopie dans un texte.

sémantiques qui constituent le thème central d'un texte. Cette répétition est une forme de révélation d'une macrostructure :

> The theoretical term semantic macrostructure was introduced to capture that important aspect of discourse and discourse processing: It makes explicit the overall topics or themes of a text and at the same time defines what we could call the overall coherence of a text as well as its upshot or gist (Van Dijk, 1988b, p. 13).

Chez van Dijk, la macrostructure constitue la structure qui donne une *forme* à la *substance sémantique la plus importante* d'un texte. Cette forme est cernée à partir d'*éléments qui sont distribués* dans le texte et est souvent associée au concept de thème et de cohérence textuelle

> [...] the global meaning structures for which notions such as "theme, " "topic," or "gist" are used, [...]. (Van Dijk, 1980, p. 6)

> Such themes or topics are accounted for theoretically in terms of so-called "semantic macrostructures." Thus, a fragment of a discourse or a whole discourse is considered to be globally coherent if a topic (represented by a macroproposition) can be derived from such a fragment. (Van Dijk, 1983, p. 25)

Pour résumer, la macrostructure organise la structure sémantique globale d'un texte et, sa forme est souvent identifiée par celle du thème. Pour van Dijk, la macrostructure ainsi définie est également reliée à la vie cognitive de l'être humain et plus particulièrement, aux aspects cognitifs de la *mémorisation* et de la *compréhension* d'un texte. La notion qui met ces aspects cognitifs de l'avant est celle de *cœur informatif* d'un texte, c'est-à-dire les informations centrales et cruciales d'un texte, sans lesquelles la compréhension du texte est compromise, voire impossible. Le cœur informatif constitue une ressource cognitive de l'être humain qui, pour des raisons d'économie cognitive, *sélectionne* ce qui est crucial pour la mémorisation et la compréhension :

> When people talk about such a theme or topic, [...] At the same time these notions intuitively associate with that of relevance or importance: The point is the more relevant, important, central, prominent, or crucial aspect of what was said. (Van Dijk, 1980, p. 5)

> The macrostructure thus has to represent what is the major, more relevant, more general information out of complex information as represented at the more concrete microlevel. Both the notion of description level and that of representation level appears to play an important role in discourse, cognition, and action. (Van Dijk, 1980, p. 13)

Cette notion de prépondérance de l'information ou de cœur informatif possède un substrat cognitif et social, qui est résumé par l'idée de *représentation sociale de la réalité*. Pour van Dijk, l'être humain interprète les textes et les produits sémiotiques par le biais de *conventions sociales* qui forment des structures cognitives de l'expérience et de la connaissance :

> Whatever the more specific linguistic or sociological concepts of global structures may be, we shall assume that they have a cognitive basis. [...] In other words, our social behavior including our communicative verbal interaction-is determined by our interpretations and representations of social "reality." Later chapters show that global structures are the result of very fundamental cognitive principles operating in the ways we process this kind of highly complex information from the social situation. (Van Dijk, 1980, p. 2)

> All this has of course important cognitive implications: Complex information from discourse, episodes, action sequences, etc., may be organized in memory due to macrostructural information. Without this kind of global organization in memory, retrieval and hence use of complex information would be unthinkable. (Van Dijk, 1980, p. 14)

Enfin, le modèle cognitif sur lequel est bâtie la représentation de l'information et de la connaissance joue un rôle prédominant dans la définition du concept de macrostructure. En effet, celle-ci présente, des correspondances avec la structure cognitive de l'être humain, car c'est par le biais des macrostructures que la communication entre êtres humains peut s'établir en se référant au même cœur

informatif. C'est alors qu'il devient possible de mémoriser et de comprendre les messages véhiculés. Le lien entre macrostructure et structures cognitives de la représentation de l'expérience et de la réalité évoque le concept de *frame*. Formulée à l'origine dans le gestaltisme à partir des années 1920, la théorie du *schéma* ou *frame* — appelée ainsi surtout après la recherche sur l'intelligence artificielle de Marvin Minsky — repose sur la conviction que toute notre expérience est basée sur une comparaison avec un modèle stéréotypé, dérivé d'*expériences similaires enregistrées en mémoire*. Ainsi, chaque nouvelle expérience est évaluée sur la base de sa conformité ou de sa divergence par rapport à un *frame* précédent (Calabrese, 2014). Dans ce contexte, l'utilisation d'un *frame* est une condition cognitive préalable à la lisibilité de chaque événement vécu par des individus. Ainsi, la compréhension du texte est rendue possible grâce à une comparaison continue avec les *frames* qui constituent l'*encyclopédie* du lecteur (Eco, 1979). Le même mécanisme se produit, consciemment ou inconsciemment, lors de la production textuelle. Pour ces caractéristiques, le frame est complémentaire à la fonction cognitive de la *catégorisation*, puisque chaque objet perçu dans le monde (qui est une occurrence) est rapporté à une catégorie (qui est le type).

La théorie de van Dijk a été appliquée dans différents contextes et sa contribution au développement de l'analyse de textes représente un élément fondamental pour la sémiotique (Adam, 2016, p. 8). Le domaine dans lequel elle a été utilisée le plus fréquemment est celui de l'analyse du discours (Maingueneau, 1979), avec un champ d'application très large, allant du domaine de la médecine (Bloom *et al.*, 1994), de la linguistique (Patry et Nespoulous, 1990), des sciences politiques (Le, 2000) à celui du journalisme et de la communication (Van Dijk, 1983, 1988a, 2008a, 2008b). Cette théorie a aussi été utilisée dans le domaine du *text mining*. Certains travaux ont permis de développer une méthode pour le résumé automatique, en exploitant un autre aspect de la théorie de van Dijk, soit les *macrorègles* (Aluísio *et al.*, 2008;

Brown et Day, 1983; Lemaire *et al.*, 2005). D'autres travaux l'ont employée pour l'analyse thématique au moyen des modèles topiques (LSA) (Kintsch, 2002) et de *clustering* (Forest, 2006). Enfin, par l'intermédiaire de la perspective développée en logique (Freeman, 2011, 1991), le concept de macrostructure a été utilisé pour la fouille des arguments (*argument mining*) (Peldszus et Stede, 2013).

2.1.1.1 Macrostructure et narration

Le concept de macrostructure et de « structure de signification globale d'un texte » n'était pas nouveau à l'époque où Van Dijk a développé sa théorie. La genèse d'une telle structure correspond à celle d'une partie de la sémiotique (Fontanille, 2007, p. 1). Par exemple, la structure textuelle comme un *tout globalisant* et *signifiant* a déjà été décrite par Rastier dans son approche de l'analyse thématique :

> Rappelons donc qu'un texte peut être analysé à trois paliers principaux : micro -, méso -, et macrosémantique, qui correspondent au sémème, au contenu de la période, et à la structure textuelle (Rastier, 1995, p. 223-249)

Comme chez van Dijk, la définition de thème est ici décrite comme une structure globale. Le thème se réalise par *accumulation isotopique*[6] ou, pour utiliser ses termes, au moyen d'une structure de traits sémantiques (ou *sèmes*), qui est récurrente dans un corpus (Rastier, 1996). La différence entre ces deux auteurs est principalement de nature théorique, car pour Rastier, le thème est un phénomène exclusivement linguistique alors pour van Dijk, il est aussi cognitif.

D'autres sémioticiens, tels Propp, Bremond et Greimas, ou des linguistes, comme Labov, ont abordé, directement et indirectement, le concept de macrostructure. Pour eux, par contre, la macrostructure assume les contours d'une *structure de type*

[6] Nous nous référons au processus d'accumulation de lexèmes, sèmes ou autres qui forment l'isotopie dans un texte

narratif. C'est à cette tradition de la narratologie issue des travaux de la sémiotique structurale que van Dijk et Rastier sont associés et sans laquelle ces réflexions sur la macrostructure n'auraient pas pu avoir lieu (Maingueneau, 1979, p. 23). Dans ce contexte, la macrostructure maintient toujours ses propriétés de « collant textuel » garantissant la cohérence des textes, mais elle propose aussi une syntaxe de l'organisation textuelle plus complexe que celle correspondant au concept de thème.

Les trente et une fonctions de Propp, regroupées dans sept sphères d'action, correspondent à une syntaxe qui détermine un type particulier de récit, soit le conte de fées russe (Franzosi, 2010, p. 54, 145; Nöth, 1995, p. 372). Pour sa part, Bremond définit un concept de macrostructure dans lequel la morphologie du conte de fées caractérise la diégèse comme une séquence macrostructurelle unique et invariante. Pour lui, c'est plutôt la séquence de microstructures invariantes qui forme les macrostructures narratives (Bremond, 1966). Dans la *logique des possibles narratifs*, Bremond propose un modèle du récit composé de trois éléments : une situation initiale qui mène vers une potentialité, l'actualisation de cette potentialité (ou son contraire, l'absence d'action) et les résultats de cette action (succès ou faillite). Greimas a développé un métalangage sémiotique qui possède une nature tout à fait narrative. Deux des principaux modèles de sa théorie, le *schéma actanciel* et le *schéma narratif canonique*, sont des exemples de macrostructure textuelle (Nöth, 1995, p. 373). Enfin, Labov propose un modèle cyclique ancré sur les macrostructures des narrations orales (Nöth, 1995, p. 373). Ce dernier modèle est constitué de six phases qui seront ensuite réutilisées par Van Dijk (Van Dijk, 1980, p. 113) pour la définition de *superstructure* : le résumé, qui démarre le récit; l'*orientation*, qui décrit le contexte de départ; l'*intrigue* (*complicating action*) soit la description de la séquence d'actions; l'*évaluation* qui met en jeu le narrateur et la manière dont le récit est présenté, soulignant ainsi les points de vue déployés; la

résolution, qui récapitule les actions et met fin à la séquence narrative; la *coda* qui marque idéologiquement le récit, amenant donc un signifié plus profond au récepteur.

Pour résumer, la macrostructure fournit une vision sémantico-cognitive de la production textuelle qui se base sur la cohérence textuelle et sur une forme d'organisation latente. La macrostructure est donc un procédé sémiocognitif qui met en forme la substance signifiante d'un texte. Cette forme peut être interprétée de manières différentes. L'une est certainement de la considérer comme le procédé de constitution du thème principal d'un texte. Une autre possibilité est de l'identifier à une structure syntactico-sémantique plus complexe, qui peut être résumée par le concept de *récit*. Ainsi, la structure typique d'un récit correspond à la macrostructure qui régit la signification d'un texte. La présente thèse adopte cette dernière vision qui sera explicitée dans la description de l'approche sémiocognitive (cfr. 2.2.4). Toutefois, le concept de macrostructure peut assumer différentes formes, comme par exemple celle de *macroproposition* (Rastier, 2001b; Ruwet, 2009; Van Dijk, 1980).

Au-delà de la forme qu'on veut donner à la macrostructure, celle-ci révèle toujours que la production textuelle, ainsi que sa lecture et sa compréhension, sont fondées sur un phénomène de répétition et de *redondance*. Ce dernier est un *mécanisme de protection* qui préserve le message et l'information des phénomènes de *bruit* et de perte d'information lors de son transfert par un *canal* de l'*émetteur* au *récepteur* (Klinkenberg, 2000, p. 75). Cette redondance se concrétise par une sorte de *régularité interne* qui constitue le texte comme produit sémiotique. En effet, sans cette régularité interne, le texte ne serait pas cohérent et tout autre procédé de signification serait destiné à échouer. Enfin, la théorie de la macrostructure propose une hypothèse sur la forme que ce mécanisme de protection prend dans un texte.

2.1.2 Le cadre épistémologique d'intersection entre le modèle computationnel et le modèle sémiotique

L'utilisation de l'assistance informatique pour l'analyse sémiotique est au cœur du présent travail. L'intégration de l'informatique à l'analyse sémiotique doit être encadrée par un système cohérent qui tient compte des *spécificités de l'analyse sémiotique* et de celles *de l'informatique*. En d'autres termes, l'utilisation de l'informatique en sémiotique oblige à s'interroger sur le *type d'assistance* possible et sur la *manière* de la mettre en place. Dans cette thèse, ces réflexions sont guidées par un principe de base : *identifier à l'intérieur des pratiques d'analyse sémiotique les tâches qui peuvent être assistées ou carrément exécutées par un ordinateur*.

Suivre un tel principe correspond à identifier un cadre épistémologique pour le transfert d'une opération de la sémiotique à l'informatique. Dans le cas de cette thèse, ce cadre doit permettre une *rencontre entre sémiotique et informatique* et ceci, *spécifiquement pour l'analyse de texte*. Le cadre identifié et que nous prenons en compte a été proposé pour la définition des pratiques en humanités numériques (Meunier, 2019a). Celui-ci est composé de quatre modèles, ces derniers constituant des moyens différents pour « expliquer la nature d'une théorie scientifique » :

> Tout projet de recherche qui en appellera à l'informatique ne sera une démarche scientifique que s'il construit une théorie qui contient au moins ces quatre différents types de modèles à savoir :
> a) Un modèle **conceptuel** qui identifie les propriétés (*features*) et les relations des objets d'étude selon des points de vue spécifiques et qui les exprime dans une langue naturelle.
> b) Un modèle **formel** qui identifie dans l'objet de recherche une structuration de ces propriétés et de ces relations en les exprimant formellement (mathématique, logique, etc.). Certaines de ces relations seront plus intéressantes que d'autres, les relations de dépendances fonctionnelles par exemple.

> c) Un modèle **computationnel** traduira les relations fonctionnelles récursives en fonctions computables, c'est-à-dire en un langage algorithmique.
> d) Un modèle **physique** construira une architecture matérielle (électronique, mécanique, etc.) qui permettra de *computer* effectivement les modèles formels et computationnels. (Jean-Guy Meunier, 2019a, p. 6)

Un phénomène est donc décrit par ces quatre modèles. Par exemple, il est possible de définir le jeu d'échecs d'abord comme une bataille médiévale (modèle conceptuel), deuxièmement comme un ensemble de règles (modèle formel), ensuite comme un système de fonctions computables (modèle computationnel) et enfin comme un plateau et des pièces de jeu, par exemple un échiquier en bois (modèle physique).

Ce système permet d'élaborer un cadre pour le transfert d'une opération sémiotique vers une opération computable. Lors de son application à la pratique de l'analyse de textes, il met en place quatre étapes à suivre:

1- Décrire une tâche ou une étape de l'analyse sémiotique de texte. Par exemple, décrire la valeur de l'identification de textes similaires.
2- Structurer formellement cette étape, ce qui implique de décrire formellement la tâche d'identification de textes similaires.
3- Identifier les fonctions computables qui rendent opérationnel le modèle formel. Par exemple, identifier la fonction qui permet d'établir le degré de similarité entre deux textes.
4- Identifier l'implémentation physique nécessaire pour la computation. Par exemple, un ordinateur ou un autre calculateur.

Lorsqu'une tâche de l'analyse sémiotique peut être décrite par les quatre modèles, elle est exécutable par une machine. Par contre, *seule une tâche simple pourrait être décrite par les quatre étapes*. Par exemple, la tâche de lecture d'un texte est trop complexe pour être décrite d'un point de vue formel. En effet, la « lecture » est une opération cognitive complexe, et cette opération devra être décomposée en tâches plus petites. L'assistance informatique force ainsi la sémiotique à *décomposer sa*

pratique d'analyse en étapes ou tâches plus petites et à se questionner sur la valeur cognitive de chacune d'elles. Cette décomposition exécutée en amont facilite ensuite le transfert d'une opération simple du modèle conceptuel au modèle computable et physique.

Il se dessine donc une interaction entre un modèle conceptuel, son substrat cognitif[7] et une *représentation formelle* de celui-ci. En d'autres termes, chaque tâche ou opération de l'analyse de texte correspond à une opération cognitive et certaines d'entre elles peuvent faire l'objet d'une représentation formelle. Illustrons ce mécanisme par un exemple. La *loi de Zipf* (Zipf, 1949) est une loi empirique qui a été formalisée par le linguiste Zipf en observant la fréquence des mots dans un ensemble de textes. Elle décrit un phénomène linguistique au moyen d'une loi mathématique et statistique. Cette loi qui décrit la relation entre la fréquence d'apparition d'un mot et son rang à l'intérieur de l'ensemble de texte sur lesquels la fréquence a été calculée. Cette relation montre que la fréquence dépend du rang du mot et que cette fréquence diminue sensiblement en même temps que la diminution de son rang, ce qui implique que *les mots très fréquents sont peu nombreux alors que les mots les moins fréquents sont très nombreux.*

De plus, cette loi comporte une *hypothèse explicative de type cognitif.* Les éléments cognitifs qui déterminent ce phénomène sont liés, d'une part, aux ressources limitées de l'être humain relativement à la *mémoire* et à la *capacité d'attention* (Simone, 2005, p. 83) et, d'autre part, au *principe du moindre effort*, ce qui constitue la thèse même de Zipf (Zipf, 1949, p. 19). Cette relation a été étudiée par plusieurs disciplines du langage et ce, depuis les années 50. Shannon, par exemple, propose :

---

[7] Le modèle décrit par Meunier (2019) a été simplifié, mais il est opportun de rappeler que le substrat cognitif est accompagné par d'autres dimensions, comme celles sociale, linguistique, politiques, etc.

> d'utiliser les analyses statistiques de Zipf à des fins linguistiques. Se fondant sur l'idée de la redondance de la langue anglaise, il montre qu'il est possible de calculer l'entropie d'une langue à partir d'études statistiques sur la fréquence avec laquelle les éléments constitutifs d'une phrase sont sélectionnés (Léon, 2008, p. 943).

Si on tient compte du système décrit précédemment, la loi de Zipf peut être décrite par les différents modèles. Cette loi possède ainsi une description conceptuelle et sémiotique, qui est liée à la *redondance de l'information* observée dans les phénomènes linguistiques (Simone, 2005). La redondance de l'information est un *mécanisme de protection* des procédés sémiotiques qui garantit le transfert de l'information (Klinkenberg, 2000, p. 75). Alors, le fait que *peu de mots* sont *très fréquents* et que *beaucoup de mots* sont *peu fréquents* coïncide avec le principe de redondance et d'économie des ressources. En d'autres termes, on peut dire que la loi de Zipf est un indice du phénomène de redondance des langues naturelles. Cette description conceptuelle est formalisée par la formule mathématique suivante : $f * r = C$, où f est la fréquence et r le rang. Ainsi, la loi de Zipf formalisée spécifie que la fréquence d'un mot est inversement proportionnelle à son rang dans la liste des mots qui font partie d'un corpus. Cette formule est représentée par la courbe de la figure 2.1.

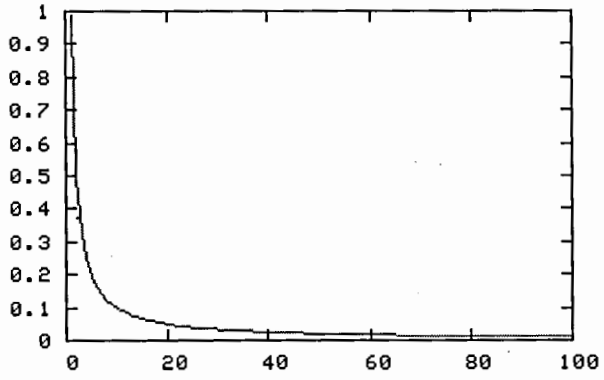


Figure 2.1 Loi de Zipf. L'axe des y correspond à la fréquence du mot; l'axe des x correspond rang du mot

Nous avons donc là un exemple de modélisation formelle qui démontre comment on peut passer du cognitif à une fonction « computable ».

2.2 Le paradigme sémio-narratif

L'analyse de texte est un des principaux objets de recherche de la sémiotique. En particulier, la sémiotique textuelle, qui est née au sein de l'école de Paris, présente des caractéristiques spécifiques qui constituent l'approche sémiotique adoptée par la thèse. Une de ses caractéristiques est de considérer le *texte comme l'objet privilégié de l'étude sémiotique* (Volli, 2014). Le texte est ainsi le matériel à partir duquel la sémiotique développe son métalangage descriptif. L'analyse sémiotique doit alors être *immanente*, c'est-à-dire que les conditions internes de la signification sont suffisantes pour l'analyse du texte (Fontanille, 2005, 2006). Le débat sémiotique sur les limites de l'analyse immanente, en dépit de l'utilisation d'aspects extratextuels pour sa compréhension, n'est pas pertinent dans le cadre de cette thèse qui présente *une méthode qui ne peut pas opérer sans le principe d'immanence*.

Un autre principe important pour la présente thèse est celui qui est à la base de la sémiotique structuraliste : « il n'y a de sens que par et dans la différence » (Groupe d'Entrevernes, 1984). Ceci implique que les conditions de signification du texte sont reliées par l'idée que le système qui les soutient est un « système structuré de relations ». Les éléments qui composent ce système se distinguent par des relations d'opposition, de complémentarité ou d'autres types, chacune ne faisant que souligner les *différences* entre ces éléments. Cet aspect définit les procédés de signification d'un texte puisque « les éléments d'un texte ne tiennent leur signification et ne peuvent être reconnus signifiants que par le jeu des relations qu'ils entretiennent » (*ibid.*). Enfin, dans ce contexte, l'analyse de texte vise à dégager les dynamiques sémiotiques qui organisent le contenu du texte.

Le texte est ainsi le « résultat d'un dispositif structuré de règles et de relations » où le métalangage de Greimas constitue une approche sémiotique pertinente. Dans ses travaux, les produits sémiotiques sont considérés des objets qui se construisent à travers des dynamiques se situant à plusieurs niveaux. En particulier, l'analyse du texte doit distinguer le *niveau superficiel*, où sont situées les composantes narratives et discursives, et le *niveau profond*, où d'autres relations et opérations organisent le sens « profond » d'un produit sémiotique.

En raison de la problématique décrite dans le chapitre précédent et de l'approche sémiocomputationelle adoptée (cfr. 2.1), la composante narrative est le seul niveau d'analyse qui est pris en considération dans ce travail. En effet, *l'analyse de la composante narrative correspond à l'identification de la macrostructure qui régit le texte*, ce qui répond aux attentes posées par la question de recherche de ce travail. La narrativité est considérée comme une des propriétés principales de l'organisation du sens et ceci vaut aussi pour le texte journalistique (Lorusso et Violi, 2004), qui constitue notre matériel empirique de la thèse. L'analyse de la composante narrative

consiste dans l'étude d'éléments qui organisent le contenu du texte à travers des *dispositifs* suivant une logique de type narratif. Par exemple, un texte peut être analysé comme une succession d'états et de transformations dans lesquels des personnages agissent. Cette structure est analysée à travers un métalangage spécifique qui décrit le fonctionnement des dispositifs de la narration.

Ce type de métalangage permet de construire un cadre pour l'analyse du contenu du texte. Décrire le contenu d'un article de journal correspond ainsi à l'opération d'identification des éléments narratifs qui sont en jeu, par exemple la présence d'un « héros », ou d'un « antihéros », de l'« objet de valeur » et d'« actions » accomplies et des « plans » narratifs recherchés. Même si le journaliste ne se réfère pas consciemment à un modèle narratif pendant la rédaction, celui-ci constitue une des modalités cognitives les plus importantes qui régissent le travail d'écriture d'un texte journalistique. La narrativité est donc une *structure sous-jacente* à la vie cognitivo-sémiotique de l'être humain et elle laisse ses traces au niveau de l'organisation superficielle du texte. En d'autres termes, cette structure sous-jacente coïncide avec la régularité interne d'un texte, cette dernière prenant une forme spécifique, celle de la narration.

En sémiotique, la littérature traitant de la narrativité est très abondante. Des auteurs tels que Propp, Barthes, Genette, Ricœur, Todorov, Greimas, Bremond, Eco et Kristeva ont procédé à une réflexion sur les structures narratives de textes littéraires. Notre travail s'appuie sur cette tradition et adopte un cadre prédéfini, qui est celui établi par Greimas dans *Le sens* (Greimas, 1983) et dans *Sémiotique. Dictionnaire raisonné de la théorie du langage* (Greimas et Courtés, 1979). Un approfondissement de certains de ces modèles est effectué dans les travaux de Hébert (2016), ce qui a permis de combler certaines de leurs lacunes. Ces améliorations sont également prises

en considération dans ce travail. Dans les prochains paragraphes, le métalangage spécifique utilisé pour l'analyse sémiotique est présenté.

Ces choix sont motivés par les points suivants. D'abord, Greimas fonde son approche sémiotique sur la notion de texte comme une *réalité empirique contenant ses conditions de production.* Dans ce travail, le texte est le seul matériel pris en considération pour l'analyse de contenu. Deuxièmement, Greimas développe un *métalangage sémiotique formel* pour la description des *sémiotiques-objets*[8]. Le développement d'une méthode assistée par ordinateur est facilité par un modèle sémiotique de type formel, car les outils computationnels nécessitent un ancrage formel pour exécuter des tâches sémiotiques. L'aspect computationnel de l'analyse doit ainsi nécessairement passer par une représentation formelle des dynamiques sémiotiques. Troisièmement, Greimas offre un modèle narratif qui se détache de plus en plus du genre littéraire et narratif pour construire une syntaxe générale du discours qui peut aussi être appliquée à l'analyse des textes non narratifs (Bertrand D., 2000). L'approche sémiotique choisie doit pouvoir être appliquée aux textes journalistiques. Enfin, certains de ces modèles, comme le schéma actantiel, sont très simples et intuitifs, facilitant ainsi leur utilisation dans un contexte computationnel. En raison de l'aspect expérimental du modèle développé, il serait très difficile d'implémenter ou d'explorer les données textuelles au moyen d'un outil trop complexe.

2.2.1 Le *parcours génératif du sens* : un métalangage formel

À partir de ses études de sémantique jusqu'au dictionnaire, Greimas a su interpréter le structuralisme et le formalisme des années 1920 grâce à une clé tout à fait nouvelle pour l'époque. Son projet de recherche initial était d'identifier une structure universelle valide pour chaque processus de *sémiosis*. La base de son projet était la

[8] Systèmes sémiotiques qui sont décrits par des métalangages sémiotiques (Barthes, 1964).

construction d'une sémiotique formelle, qui aurait pu « dire quelque chose sensé sur le sens » (Magli et Pozzato, 1985, p. I) et décrire les modalités de sa manifestation et de sa transformation. Son œuvre a été influencée par plusieurs structuralistes comme Lévi-Strauss, Dumézil, Propp et Hjelmslev. En contre-partie, son travail a aussi influencé plusieurs générations de sémioticiens, pour lesquels traiter de narrativité signifiait appliquer ou continuer à développer le modèle du *parcours génératif du sens*, ce qui a constitué un des objets d'études les plus importants de Greimas.

À plusieurs reprises, la communauté de sémioticiens a considéré le projet du parcours du sens comme trop ambitieux. Toutefois, le métalangage sémiotique de Greimas a été le dénominateur commun de plusieurs courants de pensée en sémiotique. Son succès est probablement dû au fait que sa sémiotique offre un *ensemble de concepts inter-définis* qui constitue une théorie du sens et une méthodologie d'analyse cohérente, même si elles sont incomplètes. Ses outils ont été et sont toujours utilisés dans plusieurs domaines de recherche et plusieurs champs d'application, du marketing à l'analyse littéraire. Cet ensemble de concepts est un *métalangage* et il représente son véritable *projet sémiotique* :

> L'homme vit dans un monde signifiant. Pour lui, le problème du sens ne se pose pas, le sens est posé, il s'impose comme une évidence, comme un « sentiment de comprendre » tout naturel. Dans un univers « blanc » où le langage serait pure dénotation des choses et des gestes, il ne serait pas possible de s'interroger sur le sens : toute interrogation est *métalinguistique*. [...] La signification n'est donc que cette transposition d'un niveau de langage dans un autre, d'un langage dans un langage différent, et le sens n'est que cette possibilité de *transcodage*. [...] La transcription sémiotique de la signification est, par conséquent, la construction d'un langage artificiel adéquat (Greimas, 1970, p. 12-14)

Cet extrait illustre bien la nécessité d'un métalangage pour la description des phénomènes sémiotiques. Pour décrire les « faits sémiotiques », c'est-à-dire les systèmes et les processus de la signification, il est nécessaire de formuler une

sémiotique comme métalangage. Le « sens » est quelque chose qui apparaît dans le langage et qui doit être décrit par le même langage que celui où il apparaît. Ainsi, la nécessité de s'outiller de concepts inter-définis permettant la description des « faits sémiotique » émerge. La *sémiotique-objet* doit donc être décrite par une *sémiotique-métalangage* ou une méta-sémiotique.

Pour Greimas, la méta-sémiotique ne peut qu'être un *métalangage formel*. Pourquoi? Parce que les intentions de Greimas sont de dégager des dispositifs qui généralisent des processus sémiotiques. Le projet d'une sémiotique formelle est un projet de sémiotique universelle qui veut décrire les mécanismes généraux de production du sens. Pour ce faire, Greimas postule la nécessité de disposer d'une sémiotique formelle, car elle possède des propriétés de *récursivité* et de *reproductibilité* (Galofaro, 2013). En d'autres termes, un dispositif explicatif des procédés généraux de production du sens est un dispositif qui fonctionne dans plusieurs « faits sémiotiques » différents. Pour l'obtenir, Greimas propose un métalangage de nature formelle[9]. Le débat sur la nécessité de disposer d'une telle sémiotique-métalangage a été tenu à plusieurs reprises (Fossali *et al.*, 2013) et il n'est pas pertinent de le reprendre.

2.2.2 Les structures sémio-narratives

Le *parcours génératif du sens* est un système syntactico-sémantique organisé en niveaux de profondeur que Greimas fait correspondre à un mécanisme de génération du sens (Traini, 2013, p. 55). L'*objet principal* de la théorie sémiotique de Greimas est la *détermination des conditions de production et de compréhension du sens* (Greimas, 1983). Ainsi, la théorie de Greimas tente de définir le sens à travers la description de ce processus de production. Dans ce contexte, les sémiotiques-objets

[9] Pour compléter ce raisonnement par une analogie, voir la définition du concept d'algorithme dans l'introduction du présent travail et dans Greimas et Courtes (1979, p.12).

sont constituées par des niveaux différents. Ces niveaux sont reliés entre eux par des *règles de conversion* qui construisent la signification, du niveau le plus profond au plus superficiel. Le principe théorique de Greimas est la « simulation » du chemin que le sens parcourt pour se construire, niveau après niveau. La manifestation du sens dans les langues ou tout autre système de signes est donc observée grâce à une hypothèse sur sa constitution. Le « parcours » de Greimas est ainsi *génératif* parce qu'il propose une théorie sur la *génération du sens*. En effet, Greimas « simule » son parcours de génération, qui est déclenché par l'apparition d'éléments logico-sémantiques élémentaires se situant au niveau profond, lesquels sont ensuite convertis en éléments *syntaxico-sémantiques* à un niveau plus superficiel, aboutissant enfin au niveau discursif à travers les mécanismes de l'énonciation. En d'autres termes, le niveau profond est l'endroit où la *substance* et la *forme* primordiale du sens apparaissent et elles possèdent une nature logico-sémantique. Les niveaux plus superficiels transforment la forme élémentaire du sens vers des formes plus complexes, jusqu'à sa mise en forme définitive au niveau discursif.

Le « parcours » de Greimas découle de la théorie du signe de Hjelmslev. Sa sémiotique est ainsi articulée dans un *plan du contenu* et un *plan de l'expression*, où la substance de la matière signifiante est organisée par la forme de l'expression et celle du contenu. En d'autres termes, il existe un plan du contenu où les aspects sémantiques se condensent et un plan de l'expression où les formes expressives s'organisent pour véhiculer des éléments du plan du contenu. Par exemple, dans le cas des langues naturelles, le signifié d'un mot est véhiculé par une séquence de syllabes. Le « parcours » prévoit ainsi des dispositifs d'organisation et de structuration du contenu et un niveau discursif dominé par l'acte d'énonciation. Toutefois, les travaux de Greimas accordent plus d'attention aux processus qui organisent le plan du contenu.

Tel que déjà mentionné, le « parcours » est composé de plusieurs niveaux, de l'élémentaire et profond aux plus complexes et superficiels. Chaque niveau contient une composante syntaxique et une composante sémantique (figure 2.2). *Les dispositifs qui sont au cœur du « parcours » se caractérisent par des propriétés de type narratif.* Considérer les structures narratives comme des dispositifs généraux de la vie cognitive de l'être humain et donc de sa production sémiotique constitue une hypothèse forte de Greimas. Cet élément sera approfondi au point 2.1.3 et il est le point de départ des expérimentations et de la structure argumentative de la thèse.

Ainsi, les niveaux du « parcours génératif » peuvent être répartis en deux catégories, soit les *structures sémio-narratives*, les structures les plus profondes et les *structures discursives*, qui sont les plus superficielles. Les structures semio-narratives constituent deux niveaux, soit le niveau profond, où sont situées la *syntaxe* et la *sémantique fondamentale*, et le niveau superficiel, où la *syntaxe narrative de surface* et la *sémantique narrative* opèrent. Les structures discursives constituent le dernier niveau, lequel est responsable de l'accomplissement de la dernière étape du « parcours du sens » c'est-à-dire la mise en forme et l'instanciation discursive des structures syntaxiques et sémantiques construites auparavant. À l'origine de ce niveau, se situe un véritable un acte d'énonciation. Ainsi, il est possible de retrouver dans le texte les marques des différents processus énonciatifs, soit l'*actorisation*, la *temporalisation*, la *spatialisation*, la *thématisation* et la *figurativisation*.

<table>
<tr><th colspan="4">PARCOURS GÉNÉRATIF</th></tr>
<tr><td></td><td colspan="2">composante syntaxique</td><td>composante sémantique</td></tr>
<tr><td rowspan="2">Structures sémio-narratives</td><td>niveau profond</td><td>SYNTAXE FONDAMENTALE</td><td>SÉMANTIQUE FONDAMENTALE</td></tr>
<tr><td>niveau de surface</td><td>SYNTAXE NARRATIVE DE SURFACE</td><td>SÉMANTIQUE NARRATIVE</td></tr>
<tr><td>Structures discursives</td><td colspan="2">SYNTAXE DISCURSIVE
Discursivisation
actorisation
temporalisation
spatialisation</td><td>SÉMANTIQUE DISCURSIVE
Thématisation
Figurativisation</td></tr>
</table>

Figure 2.2 Le parcours génératif du sens[10]

Les structures sémio-narratives constituent le cœur de la théorie de Greimas. La syntaxe fondamentale qui en fait partie correspond au *carré sémiotique*, organisant et structurant une catégorie sémantique. Le principe de base du carré correspond à un des principes les plus importants du structuralisme instauré par Saussure, c'est-à-dire l'affirmation que *quelque chose signifie, car elle fait partie d'un système de relations*. Le carré sémiotique reprend les relations établies par le carré logique d'Aristote qui définissent les oppositions logiques des propositions[11]. Par exemple, ainsi, la notion de **blanc**, existe par l'articulation de trois relations d'opposition. La première est la relation de *contradiction*, qui oppose le **blanc** au **non-blanc**. La deuxième est la

[10] Cette image est inspirée à celle proposée par Patrizia Magli et Maria Pia Pozzato dans l'avant-propos de la traductione italienne de *Le sens 2* de Greimas (Magli et Pozzato, 1985, p. IV).

[11] Le carré logique exprime les quatre modalités fondamentales d'une proposition, lesquelles entretiennent des relations logiques. Ce sont les suivantes : **l'universelle affirmative** (ex. « Tous les x sont P »), **l'universelle négative** (ex. « Aucun x n'est P »), qui est *contraire* à la première proposition, la **particulière affirmative** (ex. « Quelques x sont P »), qui est *subalterne* par rapport à la première proposition et *contradictoire* par rapport à la deuxième, et la **particulière négative** (ex. « Quelques x ne sont pas P »), qui est *contradictoire* par rapport la première, *subalterne* par rapport à la deuxième et *subcontraire* par rapport à la troisième.

relation de *contrariété* qui oppose le **blanc** au **noir**. Ensuite la relation de *complémentarité*, qui oppose le **non-blanc** au **noir** (figure 2.3). Le carré sémiotique a donc une composante syntaxique, constituée par les relations logiques, et une composante sémantique, constituée par la catégorie sémantique qui est définie par ces relations.

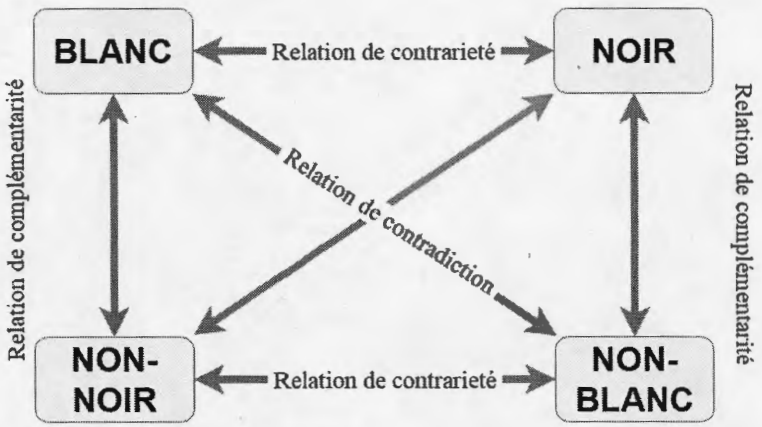


Figure 2.3 Carré sémiotique

En principe, chaque élément percevable comme signifiant dans un texte entreprend son parcours de constitution en donnant lieu à un élément sémantique qui est inséré dans un système de relations de contrariété, de contradiction et de complémentarité. Le sens se détermine dans ce système de différences et d'oppositions. L'étape suivante mène à la transformation de ces éléments dans un système de relations plus complexes qui se base sur une logique de type narratif. Ces relations sont définies par le *schéma actantiel* et par le *schéma narratif canonique*, ce qui constitue la syntaxe narrative de surface.

À la base de ces deux schémas se situe un processus sémiotique élémentaire, c'est-à-dire la constitution de l'*axe sujet — objet* (Greimas, 1973b). Cet axe est le cœur des structures sémio-narratives, car il représente le *programme narratif* d'un *sujet* qui *veut* se joindre à l'*objet*, ce dernier ayant une *valeur* pour le sujet. Pour cette raison,

cet axe est appelé l'*axe du vouloir*. Il est composé des *énoncés élémentaires*, qui sont les *énoncés de faire* et les *énoncés d'état*. Les premiers déterminent la *fonction de transformation*, qui est responsable du passage d'un état à l'autre, et les deuxièmes déterminent la *fonction de jonction* (ou disjonction), qui décrit un état. Le *programme narratif est basé sur la transformation d'un état de disjonction à un état de jonction*, puisque l'actant-sujet *planifie* de se joindre à l'objet de valeur. Le programme narratif peut être inversé lorsque l'actant-sujet programme une disjonction. La formule du programme narratif est alors la suivante :

$$PN = S_1 \rightarrow (S_2 \cap O) \tag{2.1}$$

$$\neg PN = S_1 \rightarrow (S_2 \cup O) \tag{2.2}$$

L'énoncé de faire est un *prédicat modal* lorsqu'il a comme objet un énoncé d'état, car il s'agit d'un énoncé qui en modélise un autre. Le terme entre parenthèses de la formule (2.1) est l'énoncé d'état $S_2 \cap O$, qui est transformé par le sujet S_1. La formule (2.2) indique son contraire, définissant ainsi l'*antisujet* S_2 (formule (2.3)) :

$$PN \cup \neg PN = S_1 \cup O \cap S_2 \tag{2.3}$$

Ces formules peuvent être décrites de manière plus détaillée en explicitant le passage d'état déterminé par l'énoncé de faire. Par exemple, pour la formule (2.1) :

$$PN = F\{S_1 \rightarrow [(S_2 \cup O) \rightarrow (S_2 \cap O)]\} \tag{2.4}$$

Dans la formule (2.4), F représente la fonction de transformation qui est à la base de ce changement d'état. L'énoncé de faire représente *l'acte qui produit un état* ce qui, en termes de modalité, peut être décrit par la jonction de *faire* et *être*, c'est-à-dire, *faire-être*. Les deux termes entre parenthèses sont deux énoncés d'état, le premier

étant l'état initial de disjonction et le deuxième, l'état final produit par la fonction de transformation qui représente un état de jonction. Cette modélisation du *faire-être* responsable de la transformation d'un état, *constitue le programme narratif.* À l'intérieur d'un texte, les programmes narratifs peuvent s'articuler de manière complexe et établir une hiérarchie ou un réseau. En effet, plusieurs programmes narratifs peuvent habiter un texte et être parallèles, opposés, complémentaires, superposés, emboîtés, etc.

Le programme narratif constitue le cœur de la syntaxe narrative greimasienne (Magli et Pozzato, 1985, p. XVIII-IX) et correspond à un « épisode » fondamental de tout récit, soit la *réalisation* du sujet. Dans la syntaxe de Greimas, cet « épisode » correspond à la *performance* du *schéma narratif canonique*. Ce dernier constitue une syntaxe narrative qui organise le parcours narratif du sujet. Il est composé de quatre séquences. La performance est le moment du *faire-être* et donc de la transformation d'un état de disjonction ou de jonction. Ainsi, le sujet *réalise* un acte qui le mène à la jonction ou à la disjonction avec un objet de valeur. La réalisation du sujet ou performance n'est qu'un des « épisodes » fondamentaux de tout récit. Les autres sont la *qualification* du sujet, où celui-ci obtient les qualités ou les compétences nécessaires pour accomplir son acte de jonction ou disjonction, ainsi que la *reconnaissance* qui est le moment où ses actions sont évaluées. Enfin, la *manipulation* représente la structure modale du « faire-faire » et c'est le moment où l'actant-destinateur *motive* l'actant-sujet dans la réalisation de son programme narratif. Ainsi, le récit s'articule en quatre séquences autonomes qui sont la *manipulation*, la *compétence*, la *performance* et la *sanction.*

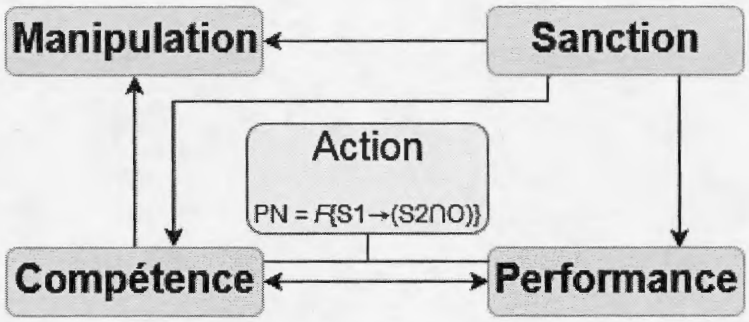


Figure 2.4 Le schéma narratif canonique.

Greimas ne considère pas ces séquences dans leur *dimension temporelle,* mais plutôt dans leur *dimension logique*. Il ne s'intéresse donc pas à l'ordre d'apparition mais à la présupposition logique entre ces éléments. Par conséquent, la performance présuppose la compétence, car le sujet obtient sa « qualification », c'est-à-dire les compétences pour agir, avant d'accomplir l'acte. Parallèlement, la sanction présuppose la performance car l'acte ne peut être « évalué » qu'après avoir été accompli. Enfin, la manipulation est présupposée à tout autre segment, car elle motive l'existence du sujet autour duquel le récit se construit.

Ces séquences qui fondent tout récit organisent des actants, lesquels relèvent de la syntaxe narrative visualisée par le *schéma actantiel*, et des *acteurs*, qui sont reconnaissables « dans le discours particulier où ils se trouvent manifestés » (Greimas, 1973a, p. 161). Par exemple, l'énoncé « Jean possède un pot plein d'écus d'or » est constitué d'un actant-sujet et d'un actant-objet. Ce dernier exprime la valeur sémantique de la richesse qui est associée au sujet et il est analysable à plusieurs niveaux :

1- *Niveau syntaxique* : Actant : objet
2- Niveau sémantique : Valeur : sème richesse
3- *Mode de manifestation* : Acteur : objet figuratif « pot plein d'écus »

L'exemple ci-dessus, qui a été présenté par Greimas (1973b, p. 17), illustre que le sème de la « richesse » a un *rôle syntaxique* à l'intérieur de l'énoncé, soit celui d'objet de valeur, et une *manifestation figurative*, soit la séquence de mots « pot plein d'écus ». Cette structure pourrait avoir un acteur différent, c'est-à-dire une manifestation figurative différente, mais conserver la même configuration d'actant et de valeur sémantique, comme dans le cas de l'énoncé « Jean a une grande fortune ». Cette dernière phrase a la même valeur sémantique que la première phrase, mais une manifestation figurative différente. Ainsi, la relation entre actant et acteur n'est pas nécessairement simple car elle peut être *multiple*. Ainsi, un actant peut être manifesté par plusieurs acteurs. Ce phénomène est appelé *syncrétisme actantiel* (Hébert, 2016, p.132). Le phénomène inverse est également possible. Ainsi, on appelle *syncrétisme actoriel* la relation entre un actant et plusieurs acteurs. Un exemple de ce phénomène est constitué par ces deux phrases, où l'actant-objet avec la valeur sémantique de la richesse est exprimé par deux manifestations figuratives différentes, soit le « pot plein d'écus » et « la fortune ».

Comme pour l'actant-objet, les autres actants du schéma actantiel représentent les *rôles narratifs généraux* qui sont organisés par les séquences narratives du schéma narratif canonique. Ils possèdent une fonction syntaxique et leurs relations sont représentées par le schéma actantiel (figure 2.5), où les trois axes qui le composent sont illustrés, soit l'*axe du vouloir*, l'*axe du pouvoir* et l'*axe de la transmission*.

Commençons par décrire l'*axe du vouloir*, qui correspond à la *performance*. Cet axe articule l'actant-*sujet* et l'actant-*objet de valeur*. Il constitue le programme narratif. En raison de la relation paradigmatique de contrariété, tout programme narratif a un contraire. Ainsi, l'actant-sujet a également son contraire, qui est l'*antisujet*, lequel est également organisé par l'axe du vouloir, mais avec des composantes inversées. L'*axe du pouvoir* correspond à la *compétence* et articule l'actant-*adjuvant* et l'actant-

opposant en relation avec l'actant-*sujet.* Ces actants influencent les actions du sujet, car ce sont les opposants qui nuisent à la réalisation de la jonction du sujet alors que les adjuvants la soutiennent en fournissant des compétences (Hébert, 2016, p. 131). Enfin, l'a*xe de la transmission* correspond à la *manipulation* ainsi qu'à la *sanction* et il articule l'actant-*destinateur*, celui qui demande la jonction au sujet, et l'actant-*destinataire*, celui pour lequel l'acte est accompli. Dans ce cadre, l'axe du vouloir assume une fonction prépondérante à l'intérieur des structures sémio-narratives, comme l'est aussi le rôle de la performance du schéma narratif canonique.

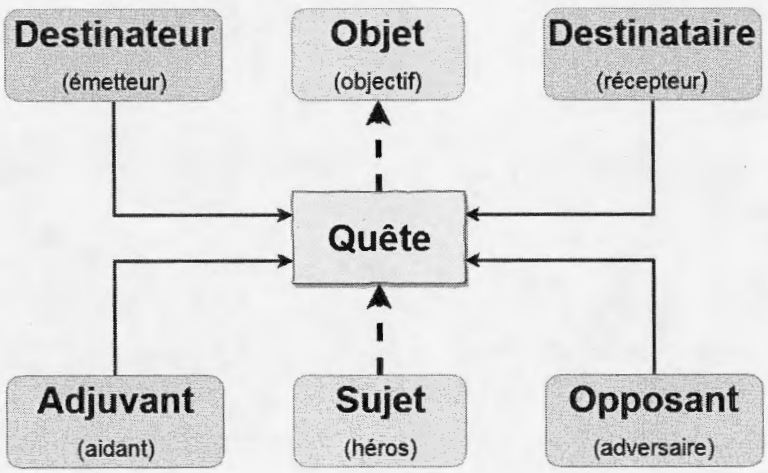


Figure 2.5 Schéma actantiel

Dans cette recherche, nous retenons de ce modèle la syntaxe narrative de surface (figure 2.2) puisqu'elle rassemble les dispositifs d'analyse les plus appropriés pour l'analyse de la composante narrative du texte (cfr. 2.2.3). À l'intérieur du parcours génératif, cette portion du modèle se situe à un niveau intermédiaire, entre celui plus profond constitué par le carré sémiotique et celui de surface, où sont situés les processus de *discursivisation.*

2.2.3 L'analyse du texte au moyen du schéma actantiel

Le modèle actantiel se définit comme « un dispositif permettant, en principe, d'analyser toute action réelle ou thématisée » (Hébert, 2016, p. 131). L'analyse au moyen de cet outil consiste donc à classer les manifestations figuratives percevables dans le texte (acteurs) dans les catégories actantielles déterminées par le modèle. Dans l'*axe du vouloir* se trouve un processus de jonction (ou de disjonction) entre un sujet et un objet de valeur. L'exemple classique est celui du prince qui *veut* sauver la princesse. L'*axe du pouvoir* identifie les adjuvants et les opposants, lesquels influencent positivement ou négativement le programme narratif du texte analysé. Par exemple, l'épée, un cheval ou le courage *peuvent* être des adjuvants pour le prince et, inversement, la sorcière, le dragon ou la peur *peuvent* tenir le rôle d'opposant. Enfin, c'est dans l'*axe de la transmission* qu'un destinateur, en s'adressant à un destinataire, investit le sujet d'une mission à accomplir. Par exemple, le roi envoie le prince sauver la princesse. Cette action est ensuite évaluée par le destinateur/destinataire, qui peut-être le roi lui-même qui a déclenché l'action, ou un autre acteur qualifié à le faire.

Les acteurs impliqués dans le schéma peuvent être des personnages, donc des acteurs anthropomorphes, comme c'est le cas du prince, de la princesse et du roi. Toutefois, des acteurs non anthropomorphes existent aussi, tels l'épée, le courage et le mauvais sort. Ils tiennent tous un rôle actantiel, sans pour autant être des personnages (Hébert, 2016, p. 34).

Ce modèle a été largement critiqué et amélioré. En général, les modifications ont ajouté du raffinement au modèle qui, dans sa version originale, tend à généraliser et simplifier fonctions et rôles actantiels qui ne peuvent appartenir qu'à une même classe. L'ambiguïté entre sujet, destinateur et destinataire en constitue un exemple. Ces actants, si l'on ne distingue pas leurs différentes fonctions, risquent de s'entremêler et de se superposer. Qui exécute l'action? Qui l'évalue? Pourquoi ces

trois fonctions peuvent-elles être accomplies par un seul acteur? De même, la relation entre destinateur, sujet et objet de valeur n'est pas claire. D'un côté, le destinateur assigne la quête de l'objet de valeur au sujet, mais en même temps, le sujet s'est investi lui-même de cette quête. L'objet de valeur est-il une assignation externe, qui vient du destinateur, ou un processus génératif qui est propre au sujet? (Ferraro, 2012, p. 46). Toutefois, il s'agit d'une fausse ambiguïté, puisque ce problème se produit seulement lorsqu'on veut assigner de manière univoque un acteur et un actant. En réalité, la relation acteur - actant est multiple, c'est-à-dire que plusieurs acteurs peuvent avoir plusieurs rôle actantiels et vice-versa. Il s'agit du phénomène de syncrétisme actantiel et actoriel (dont nous avons parlé à la page 96) et il est un argument suffisant pour répondre à cette critique.

Pour ces raisons, le modèle actantiel a été modifié et raffiné au cours des décennies. Il existe des versions très simples du modèle, comme c'est le cas du modèle « positionnel » (Bertrand D., 2000). Dans cette thèse, un modèle plus raffiné tel qu'illustré par le tableau 2.1 (Hébert, 2016, p. 133) est utilisé.

N°	temps	sujet observateur	élément actant	classe d'actant : s/o, deur/daire, adj/opp	sous-classe d'actant : factuel/possible	sous-classe d'actant : vrai/faux	autres sous-classes d'actant (par ex., actif/passif)
1							
2							
Etc.							

Tableau 2.1 Représentation du modèle actantiel sous forme de tableau (L. Hébert, 2016, p. 135)

De ce modèle, nous retenons surtout la possibilité de moduler chaque classe actantielle sur un carré sémiotique (Greimas et Courtés, 1979, p. 4). Ceci a progressivement mené à la construction d'autres sous-catégories actantielles qui demeurent tout de même ancrées dans le modèle original de Greimas. La

représentation abstraite de l'articulation d'un actant sur le carré sémiotique est fournie par la figure suivante :

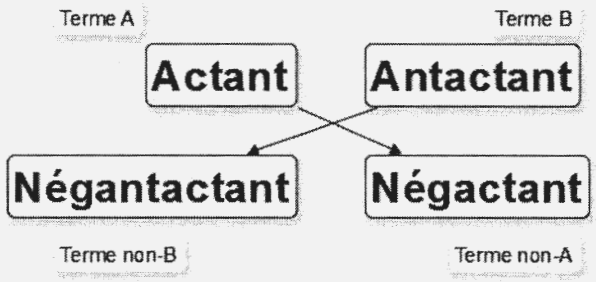


Figure 2.6 Articulation de l'actant sur le carré sémiotique

Chaque actant peut avoir ainsi un *antactant*, qui est son contraire, un *négactant*, qui est son opposé, et un *négantactant*, qui est son complémentaire. À partir de cette représentation, il est possible d'en dériver les sous-catégories actantielles suivantes : *actant/non-actant*, *actant possible/factuel*, *actif/passif*. Chacune de ces catégories est liée aux autres. Par exemple, l'axe actant/non-actant est lié à celui du possible/factuel. Prenons l'exemple de l'actant adjuvant. Si un actant adjuvant *n'aide pas* le sujet bien qu'il soit censé le faire, il ne peut pas être un adjuvant, ni un opposant. Son statut est mieux défini par la catégorie de non-actant qui, dans ce cas spécifique, est celle de non-adjuvant, mais également celle d'actant possible non devenu factuel. Une série d'exemples est présentée dans le tableau 2.2 suivant (Hébert, 2016, p. 137) :

Tableau 2.2 Les sous-catégories actantielles

		A	B	C	D
	exemple	un ami laisse un ami se noyer	un ami maintient la tête d'un ami sous l'eau	un policier laisse des voleurs voler	un policier tient le sac que les voleurs remplissent
1	actant	opposant	opposant	adjuvant	adjuvant
2	actants possible/factuel	adjuvant possible non devenu factuel	adjuvant possible non devenu factuel	opposant possible non devenu factuel	opposant possible non devenu factuel
3	actant/négactant	non-adjuvant	opposant	non-opposant	adjuvant
4	actants actif/passif	opposant passif	opposant actif	adjuvant passif	adjuvant actif

Par exemple, un individu qui laisse son ami se noyer est mieux défini si on explicite les sous-catégories actantielles. Il est alors un opposant, mais mieux défini comme un « adjuvant possible non devenu factuel », puisqu'il s'agit d'un ami qui devrait, par définition, être un adjuvant. À cette sous-catégorie on peut ajouter aussi les autres, soit « non-adjuvant » et « opposant passif ». Son rôle actantiel est le produit de toutes ces relations.

La dernière sous-catégorie, celle qui définit l'opposition actif/passif, est utilisée pour définir le type d'action que l'actant accomplit, laquelle peut être activement ou passivement dirigée vers un but. Cette catégorie permet donc de définir des actants selon leur type d'engagement dans l'action. Par exemple, un actant qui laisse mourir un ami est un opposant passif et un actant qui maintient la tête d'un ami sous l'eau est un opposant actif. Une nuance supplémentaire est aussi apportée par le fait que, dans les deux cas, il s'agit d'un adjuvant possible non devenu factuel, car l'actant est aussi l'ami de la victime et non son ennemi. Dans ce contexte, les différents actants peuvent être classés selon plusieurs sous-catégories qui spécifient davantage le rôle qu'ils ont dans le texte. Ceci permet donc de distinguer deux opposants qui ont deux rôles différents, comme dans l'exemple A et l'exemple B du tableau 2.2.

Le schéma actantiel peut ainsi être utilisé pour analyser différents énoncés ou textes. Cette version du modèle actantiel comporte aussi d'autres sous-catégories telles : actant conscient/non conscient, tout/partie, classe/élément, type/occurrence. Nous n'en tiendrons pas compte afin de maintenir la simplicité du modèle utilisé.

2.2.4 L'aspect cognitif du modèle sémio-narratif

Le récit est fondamental dans la vie culturelle. Dans un de ses textes les plus connus, Roland Barthes (1966) a souligné l'aspect omniprésent du récit : « il n'y a jamais eu nulle part aucun peuple sans récit ». D'autres auteurs suggèrent aussi que la narration

possède une valeur cognitive prépondérante et constitue un aspect nécessaire au développement de l'être humain (Young et Saver, 2001). Au cours des deux dernières décennies, cette vision a été corroborée par des études en sciences cognitives qui ont mis en évidence la nature narrative de notre manière de penser et de construire notre vision de la réalité ou notre souvenir du passé.

Plusieurs psychologues ont déjà mis de l'avant le rôle de la structure narrative comme support à la mémorisation et à la remémoration (Mandler, 1984; McAdams et McLean, 2013; Rumelhart, 1975). D'autres auteurs ont décrit comment certaines structures narratives « s'inscrivent dans les mécanismes cognitifs » (Herman, 2000; Ravoux Rallo, 1996) ou comment elles sont impliquées dans le processus de construction de l'identité personnelle (Bruner, 2006; Dennett, 1991; McAdams et McLean, 2013; Schaeffer, 2010). Dans ce genre de travaux, le récit a été abordé sous l'angle cognitif, mais aussi comme support à des opérations complexes, telle la résolution de problèmes (Herman, 2003a, 2003b). Enfin, la narrativité constitue une infrastructure sur laquelle la cognition humaine se base (Herman, 2013).

Ce type de recherche s'est aussi étendu à la neurologie, où on a commencé à identifier les régions cérébrales qui sont impliquées lors de la production ou la compréhension d'un récit. Ces recherches ont surtout souligné un partage de zones neuronales entre la compétence narrative et certaines fonctions cognitives : le système amygdale-hippocampe, où se construit la mémoire épisodique et autobiographique; la région péri-Sylvian gauche, où le langage est élaboré; et le lobe frontal, où des événements ou entités individuels sont conçus et traités comme des *frames* narratifs temporels. La relation de causalité entre cerveau et fonction cognitive de la narration est aussi confirmée par l'identification, dans certains troubles cognitifs consécutifs à un dommage neuronal localisé, d'un lien entre ce dommage et l'incapacité à réaliser certaines tâches dans lesquelles la compétence narrative est impliquée. Ces

pathologies sont appelées « dysnarrativia » (Young et Saver, 2001) et fournissent des indices pertinents sur la manière dont le cerveau organise l'expérience humaine :

> So inescapably bound are we to consciousness that we lose sight of how consciousness most often leads us to think. While we can be trained to think in geometrical shapes, patterns of sounds, poetry, movement, syllogisms, what predominates or fundamentally constitutes our consciousness is the understanding of self and world in story. (Young et Saver, 2001, p. 73)

Les courants de pensée dans la narratologie cognitive sont nombreux, mais ils s'accordent tous sur le fait que les fonctions cognitives associées à la compétence narrative se basent sur le *storyworld* (Herman, 2002), c'est-à-dire « une représentation mentale de l'univers diégétique » (Campion, 2015, p. 4).

Cette conception de la vie cognitive s'appuie sur les modèles de compréhension de texte (Van Dijk et Kintsch, 1983), sur la théorie des modèles mentaux (Johnson-Laird, 1983) et sur le concept de *pensée narrative* (Bruner, 1986). Ce dernier a joué un rôle particulier dans la définition et la diffusion du modèle cognitivo-narratif (Calabrese, 2014). Selon Bruner, la vie cognitive humaine est fondée sur deux types de pensée : la *pensée paradigmatique* (scientifique) et la *pensée narrative* (Bruner, 1991). Cette dernière est plus étroitement liée à une des fonctions les plus importantes de la vie cognitive humaine, soit le « *worldmaking* » (Bruner, 2004), dans laquelle la narration devient une compétence cognitive impliquée dans la construction de la réalité. Ce type d'approche, caractérisé par le concept de *worldmaking*, est adopté par plusieurs psychologues cognitifs et experts en narratologie. De plus, il est subsumé par le concept de *storyworld* :

> Narrative, in other words, is a basic human strategy for coming to terms with time, process, and change - a strategy that contrasts with, but is in no way inferior to, "scientific" modes of explanation that characterize phenomena as instances of general covering laws. Science explains the atmospheric processes that (all other things being equal) account for when precipitation will take the

> form of snow rather than rain; but it takes a story to convey what it was like to walk along a park trail in fresh-fallen snow as afternoon turned to evening in the late autumn of 2007 (Herman, 2011, p. 2).

Le concept du storyworld relie aussi les sciences cognitives et la sémiotique. Par exemple, Herman souligne l'importance du croisement entre les théories du langage, la narratologie et les sciences cognitives : « both language theory and narrative theory can be viewed as resources for—or modular components of—cognitive science » (Herman, 2002, p. 5).

Dans le contexte des sciences cognitives et du développement des théories cognitives de la narratologie, les théories structuralistes ont joué un rôle prépondérant (Herman, 2009b, p. 72, 2009a, p. 119, 2002, p. 3, 2013, p. 42)

Ces théories ne sont pas exemptes de critiques. En particulier, la nature spéculative de ce modèle narrato-cognitif et la difficulté à produire de véritables travaux empiriques ont déjà été soulignés (Campion, 2015, p. 3). Certains auteurs considèrent que les contributions concrètes des sciences cognitives à la narratologie sont très faibles, en soulignant que le chemin pour déterminer les fondements cognitifs de la narration sera encore long (Ryan, 2015). Le rôle que la narration joue dans la vie cognitive de l'être humain doit encore être prouvé empiriquement, mais l'hypothèse qu'il constitue une infrastructure cognitive fondamentale demeure très forte en sciences cognitives.

2.3 Le paradigme computationnel

Dans les dernières décennies, plusieurs disciplines s'intéressant aux données textuelles et linguistiques se sont développées, dont les principales sont : la fouille de texte (*Text Mining*), la recherche d'informations (*Information Retrievial*) et le traitement automatique des langues naturelles (*Natural Language Processing*). Ces

disciplines font partie de l'intelligence artificielle (IA) et explorent un de ses projets de recherche, c'est-à-dire de « simuler des opérations pour lesquelles l'être humain est meilleur »[12] (Rich *et al.*, 2009, p. 3).

Ces opérations sont des tâches réalisées par l'être humain. En effet, l'intelligence artificielle s'occupe de l'opérationnalisation de certaines de ces tâches par le biais d'un modèle formel et computationnel. Ce dernier sera ensuite implémenté dans une machine, pour reproduire ainsi l'opération effectuée par l'être humain. Toutefois, ce ne sont pas toutes les tâches qui peuvent être transférées directement dans un modèle computationnel. Tel que présenté au paragraphe précédent 2.1.2, les tâches complexes, lesquelles le plus souvent correspondent à des opérations cognitives complexes (par exemple, la lecture d'un texte), doivent être divisées en tâches et opérations plus petites. L'analyse de textes journalistiques assistée par ordinateur est une opération qui doit être décomposée en tâches et opérations. Certaines d'entre elles peuvent être sélectionnées pour les transférer dans un modèle formel et computationnel. Ainsi, une série de sous-opérations de lecture et d'analyse de texte peuvent être effectuées complètement ou partiellement par un programme informatique.

L'approche computationelle qui est utilisée dans cette thèse est constituée par un modèle de *représentation formelle et computable du texte* et par *une approche computationelle de la reconnaissance des formes récurrentes et des régularités*. En effet, le premier obstacle à surmonter est de transférer le texte dans un contexte computable. Ce n'est qu'après ce transfert que certaines opérations d'analyse peuvent être accomplies. La sémantique vectorielle offre un cadre théorique adéquat pour obtenir une représentation formelle et computable du texte. Le deuxième obstacle est

[12] [Notre traduction]

d'identifier l'outil computationnel qui « simule » l'opération cognitive qui est traitée dans cette thèse. Pour ce faire, une famille particulière d'outils, le *clustering*, qui fait partie du paradigme non supervisé de l'apprentissage automatique, constitue le cadre théorique pour l'opérationnalisation de la reconnaissance des formes récurrentes et des régularités internes aux textes.

2.3.1 La sémantique vectorielle

La sémantique vectorielle est une méthode computationnelle pour la représentation formelle et computable du langage naturel. Elle utilise les principes de l'algèbre linéaire pour construire cette représentation. Ainsi, le but principal du modèle est d'obtenir une représentation vectorielle de la sémantique des textes. Ce modèle partage les mêmes caractéristiques et propriétés que le modèle vectoriel, ce qui permet d'utiliser les outils mathématiques de l'algèbre linéaire.

Le modèle sémantique vectoriel a été développé par l'équipe de Salton (1968; 1971; 1975; 1983; 1988; 1988; 1991) dans les années 70. La base du modèle sémantique vectoriel est de représenter un document comme un point dans un espace géométrique, dont les coordonnés dépendent des caractéristiques lexicales et sémantiques (Sahlgren, 2006). Ainsi, la représentation vectorielle de plusieurs documents nous offre la possibilité de calculer (Schütze, 1993; Schütze et Pedersen, 1993) la similarité ou la dissemblance entre eux. Plus les textes sont proches, plus ils sont similaires. Chaque document est donc représenté par un vecteur sur lequel, grâce aux outils classiques de l'algèbre linéaire, plusieurs opérations peuvent être effectuées (Berry *et al.*, 1995).

2.3.1.1 Les fondements de l'application au langage des approches computationnelles

Le développement des méthodes computationnelles de traitement du langage naturel a été assujetti au débat entre rationalisme et empirisme (Brill et Mooney, 1997; Dale

et al., 2000, p. IV; Markie, 2017). Le sujet de la dispute entre ces deux positions est le rôle de l'expérience dans les processus d'apprentissage de l'être humain. En bref, les rationalistes affirment que la connaissance est acquise de manière indépendante de l'expérience. Au contraire, les empiristes affirment que l'expérience est la source ultime d'apprentissage et connaissance. Généralement, les rationalistes justifient leur vision par le biais de deux éléments. Le premier est le constat que la connaissance acquise dépasse les informations que l'expérience peut apporter. Le second est la description de la façon dont la raison fournit les informations additionnelles qui manquent dans l'expérience. Les empiristes présentent alors des argumentations opposées, en montrant comment l'expérience fournit assez d'informations pour l'apprentissage de l'être humain et en critiquant la raison comme source de connaissance.

L'application de l'approche rationaliste au langage a été dominée par la théorie de la *Grammaire Générative* de Noam Chomsky, qui définit le langage comme une faculté innée de l'être humain en développant une « conception rationaliste de la nature du langage » (Chomsky, 1967, p. 9). Sur la base de l'argument de la *pauvreté du stimulus*, Chomsky prend ses distances avec l'empirisme en démontrant un grand écart entre la pauvreté des informations que l'enfant a à sa disposition dans l'expérience et la complexité du langage qui est acquis. En d'autres termes, la position rationaliste affirme que les mécanismes d'apprentissage du langage contiennent des principes qui font du langage une compétence innée.

Dans le contexte du traitement automatique du langage naturel, ce genre d'approche a conduit au développement de l'*approche rationaliste* ou *symbolique* (Allen, 1995; Dale *et al.*, 2000). Le principe de cette approche est de construire un système de représentation de la connaissance accompagné par des *règles* d'associations ou de raisonnement qui sont *établies a priori* par le chercheur. Ce type de système, qui est

implémenté dans des programmes informatiques et est construit manuellement par le chercheur, est utilisé pour automatiser le processus humain de compréhension ou de production du langage. Par exemple, imaginons que nous voulions construire un programme informatique capable de désambiguïser les phrases qui contiennent le mot « hôte ». En français, ce terme indique à la fois la personne qui offre l'hospitalité et la personne qui est bénéficiaire de cet accueil. Le terme possède donc deux référents. Pour distinguer les deux sens, le rôle syntaxique du mot peut être utilisé de même que la sémantique du contexte dans lequel le mot apparaît. Ainsi, si le terme « hôte » est sujet du verbe « accueillir », le réfèrent est la personne qui accueille, alors que s'il est complément d'objet du même verbe, alors le référent est la personne qui est accueillie. Dans le contexte de l'approche rationnelle, toutes ces informations doivent faire partie de la base de connaissances à laquelle le programme informatique fait appel pour résoudre la tâche de désambiguïsation. Cette approche est donc liée à la construction d'*ontologies*, c'est-à-dire de représentations formelles des différents domaines du savoir à partir desquels un programme informatique dispose des informations pertinentes pour résoudre une tâche spécifique de compréhension ou de génération du langage. Le thésaurus *WordNet* est probablement l'ontologie la plus connue et la plus répandue dans le domaine du traitement automatique du langage naturel.

Toutefois, la création de tels systèmes « omniscients » est très difficile (Brill et Mooney, 1997) à cause de l'ambiguïté du langage et de la quantité de domaines spécifiques du savoir. Le plus souvent, ce genre de systèmes est utilisé pour répondre à des tâches très précises et dans des domaines très spécifiques. Ils peuvent être rarement utilisés dans des domaines différents de ceux pour lesquels ces systèmes ont été construits.

L'approche empiriste dans le domaine de l'étude du langage a été populaire dans les années 50 sous l'influence du béhaviorisme et des théories du psychologue Burrhus Skinner. Son principe est de mettre l'accent sur le rôle de l'expérience dans le processus d'acquisition du langage. L'apprentissage du langage est alors expliqué par des processus d'imitation et des mécanismes de renforcement du comportement et de conditionnement qui mènent l'être humain à associer un comportement linguistique à un stimulus particulier. Dans l'approche empiriste, l'analyse du langage se base sur la logique inductive et met de l'avant l'identification de règles générales à partir de l'observation de la réalité. Leonard Bloomfield, le père putatif du structuralisme américain, est un des représentants majeurs de l'utilisation de l'approche empiriste dans l'analyse du langage. Selon Bloomfield

> The only useful generalizations about language are inductive generalizations. Features which we think ought to be universal may be absent from the very next language that becomes accessible. (Bloomfield, 1933, p. 20)

Le structuralisme de Bloomfield est lié à la méthode scientifique de type inductif. Le système structural qui compose le langage n'est accessible qu'à travers l'observation de la réalité. Les successeurs de Bloomfield, avec leurs propres spécificités et distinctions, ont maintenu l'approche empirique dans l'étude du langage et sont ainsi considérés comme les inspirateurs de l'*approche empirique ou corpus-based/data-driven* du traitement automatique du langage naturel (Manning et Schütze, 1999). Cette approche est aussi celle qui a le plus contribué à l'utilisation de la sémantique vectorielle à des fins d'analyse sémantique.

2.3.1.2 L'hypothèse distributionnelle : du lexique à la sémantique

J.R. Firth a affirmé que l'objectif de la recherche en linguistique est d'*analyser le langage dans son usage*. Son expression « You shall know a word by the company it keeps » est devenue la citation principale des auteurs qui ont développé et utilisé la

sémantique vectorielle ou l'approche distributionnelle. Le principe qui est ici en cause est de construire une sémantique qui se définit exclusivement par des *relations de cooccurrences entre mots*. Par la suite, Zellig Harris a mis de l'avant le besoin de construire des procédures par lesquelles les structures du langage peuvent être découvertes automatiquement.

Le fondement de cette approche est l'*hypothèse distributionnelle* (Harris, 1954), qui se base sur l'idée que le sens des mots est distribué dans le contexte dans lequel il apparaît, reprenant ainsi les propos de Firth (1935, 1957). Il est possible de résumer cette hypothèse par la proposition suivante : *les éléments linguistiques ayant des distributions similaires ont des significations similaires*. Le paradigme distributionnel se base donc sur la notion que la *similarité entre entités sémantiques dépend de leurs propriétés distributionnelles*. Cela implique que la similarité du plan du contenu entre deux textes est corrélée à la similarité de leurs distributions lexicales. En d'autres termes, deux mots tendent à être similaires s'ils côtoient souvent les mêmes mots et par conséquent, les entités sémantiques qui partagent des contextes similaires tendent à être similaires.

Cette hypothèse a été entièrement élaborée par l'équipe de Salton, qui a développé le modèle sémantico-vectoriel. Ce modèle correspond à la construction d'une matrice *U*. Cette modélisation est représentée par la formule (2.5). Chaque ligne de cette matrice modélise un document sous la forme d'un vecteur $\overrightarrow{s_i} = (V_{i1}, V_{i2}, \ldots, V_{ij})$ dans lequel V_{ij} correspond à la *valeur de pondération* du $j^{ème}$ mot dans le $i^{ème}$ segment.

$$U = \begin{array}{|c|c|c|c|c|} \hline & mot_1 & mot_2 & \ldots & mot_j \\ \hline article_1 & V_{11} & V_{12} & \ldots & V_{ij} \\ \hline article_2 & V_{21} & V_{22} & \ldots & V_{ij} \\ \hline \vdots & \vdots & \vdots & \ddots & \vdots \\ \hline article_i & V_{i1} & V_{i2} & \ldots & V_{ij} \\ \hline \end{array} \qquad (2.5)$$

Dans ce cadre, chaque document est décrit par son lexique. Les textes qui partagent des entités lexicales similaires tendent à être similaires, ce qui implique que les vecteurs qui les représentent tendent à se rapprocher dans l'espace géométrique construit par le modèle sémantico-vectoriel.

Le postulat central de ce modèle est le *passage du lexique à la sémantique*. L'hypothèse distributionnelle garantit ce passage. En effet, chaque vecteur décrit un document sur la base de son lexique, lequel montre des relations avec les autres entités lexicales, c'est-à-dire des relations de cooccurrence, ce qui représente, selon l'hypothèse distributionnelle, un indice fiable et suffisant de la sémantique du mot et du texte.

La transformation du texte dans un modèle vectoriel a été légitimée par des expérimentations empiriques. En effet, les travaux en recherche d'information ont montré que le modèle sémantique est efficace dans une tâche de comparaison sémantique entre textes, ce qui justifie essentiellement le passage du niveau lexical au niveau sémantique. Les aspects techniques du modèle sont décrits au chapitre suivant (cfr. 3.3.2) puisque la transformation du texte dans un modèle vectoriel constitue une étape de la méthode.

2.3.2 L'apprentissage automatique et le *clustering*

L'apprentissage automatique est une discipline dont l'objectif est de développer des méthodes pour la découverte des caractéristiques d'une base de données (Bishop, 2006). Ces méthodes ont la capacité d'« apprendre » à partir d'une expérience et de résoudre certaines tâches complexes de manière automatique et performante (Mitchell, 1997).

Il existe deux modèles principaux d'apprentissage : le *modèle supervisé* et le *modèle non supervisé*. Le premier permet de reconnaître les objets qui appartiennent à une classe qui a été « apprise » précédemment. Par exemple, à partir d'une base d'apprentissage constituée d'exemples de pourriels et de bons courriels, le modèle est entraîné à identifier les pourriels parmi de nouveaux messages qui lui sont soumis (Drucker *et al.*, 1999). Au contraire, les modèles non supervisés n'apprennent pas les classes à partir d'exemples regroupés en catégories. Le modèle apprend à reconnaître les *motifs récurrents* ou les caractéristiques les plus importantes parmi des données non structurées et nullement classifiées et ce, en séparant les objets en groupes différents. Par exemple, à partir d'une base de données constituée d'articles de *Wikipédia* provenant de différents domaines de connaissance, un modèle non supervisé est capable de distinguer et de séparer les articles en fonction de leur similarité disciplinaire ou thématique (Huang *et al.*, 2008; Yan *et al.*, 2009; Yao *et al.*, 2012).

Au paradigme non supervisé, appartient une famille de méthodes qui est appelée *clustering*. La terminologie utilisée pour se référer au clustering n'est pas univoque, ce qui peut amener de la confusion. En effet, il existe plusieurs noms pour désigner cette famille de méthodes. En anglais, les termes les plus répandus sont *clustering*, *cluster analysis*, *segmentation*, *unsupervised classification*, alors qu'en français, ce sont *regroupement automatique*, *segmentation*, *classification*, *classification automatique*, *classification non supervisée*, *clustering*, *partitionnement*. Cette diversité terminologique est due au fait que ces méthodes ont été utilisées dans plusieurs domaines de recherche, chacun utilisant sa propre terminologie, qui est généralement justifiée par la nature de la discipline dans laquelle la recherche s'inscrit. Par exemple, dans le domaine du marketing, il est plus courant de trouver le terme segmentation, car le clustering s'intègre dans l'optimisation des processus de segmentation du marché ou des clients. Pour alléger ce travail, nous utilisons le terme

clustering qui demeure le plus utilisé, à la fois dans les pays anglophones et dans les pays francophones.

Les techniques de *clustering* se sont rapidement répandues dans le domaine de la fouille de texte. Deux hypothèses motivent leur utilisation dans ce cadre, soit celle de la contiguïté (*contiguity hypothesis*) et celle du cluster (*cluster hypothesis*). L'hypothèse de la contiguïté (Manning *et al.*, 2009, p. 289) stipule que les textes d'un même groupe (cluster) forment une région contiguë qui se distancie assez nettement des autres régions, alors que celle du cluster (Manning *et al.*, 2009, p. 350) affirme que les textes d'un même groupe partagent un contenu sémantique similaire.

2.3.2.1 Définition et principes de base

Le clustering est une famille de méthodes développée en informatique et en analyse de données qui exécute un regroupement automatique d'individus similaires. La majorité des définitions de clustering partagent un certain nombre de principes de base. Voici quelques définitions :

> Le clustering est l'art de trouver des groupes distincts dans les données (Kaufman et Rousseeuw, 2005, p. 1) [Notre traduction]

> [...] cluster analysis est probablement le meilleur terme pour indiquer les procédures qui cherchent à découvrir des groupes d'individus dans un jeu de données. (Everitt *et al.*, 2011, p. 5) [Notre traduction]

> [...] clustering veut dire trouver des groupes d'individus dans un jeu de données. La classification d'Aristote sur les entités vivantes est probablement le premier exemple de clustering, un clustering de type hiérarchique (Hennig *et al.*, 2016, p. 2) [Notre traduction]

Dans ces définitions, le clustering est présenté de manière fonctionnelle, en mettant en relief son objectif pratique qui est d'identifier des groupes (*clusters*) d'individus dans un jeu de données. Comme le souligne Hennig, le principe de base du clustering

fait partie de l'histoire de la pensée humaine et remonte à la classification des entités vivantes par Aristote. En effet, observer le monde qui nous entoure et classifier les entités ainsi observées en groupes distincts constituent une opération de clustering.

La notion intuitive de clustering est donc très simple. Au contraire, la notion technique de clustering exige quelques explications supplémentaires. Tout d'abord, le clustering exploite l'opposition entre *similarité* et *dissimilarité* :

> Le clustering est une des méthodes les plus importantes pour l'extraction des connaissances à partir de données multidimensionnelles. Le clustering a pour but d'identifier un schéma récurrent ou des groupes d'objets similaires dans un jeu de données. (Kassambara, 2017, p. 3) [Notre traduction]

La notion de similarité est donc fondamentale pour le clustering. Les objets qui sont regroupés ensemble doivent partager un certain nombre des caractéristiques communes afin que le concept de similarité entre individus soit significatif. Deux individus sont donc très similaires s'ils partagent beaucoup de caractéristiques. *La similarité est en effet une question de degré*. Si elle comporte des degrés, elle peut donc être mesurée :

> Le problème du clustering est défini comme étant celui de la recherche de groupes d'objets similaires dans les données. La similarité entre les objets est mesurée à l'aide d'une fonction de similarité. (Aggarwal et Zhai, 2012, p. 77-78) [Notre traduction]

Le mesurage de la similarité s'effectue donc à travers la *fonction de similarité*. Par exemple, la similarité entre un kangourou, une souris et un lion est mesurable à travers les caractéristiques qui les décrivent, comme le taxon, la taille, la façon de déambuler, etc. Dans cet exemple très simple, le nombre de caractéristiques partagées peut être alors transformé en une mesure qui établit le degré de similarité entre eux.

Toutefois, la notion de similarité est dépendante du contexte global dans lequel les données sont situées et analysées. Par exemple, la souris peut être considérée plus similaire au lion qu'au kangourou puisque les deux sont des mammifères placentaires au contraire du marsupial. Dans un contexte spécifique, par exemple dans une étude sur l'évolution des appareils reproducteurs du monde animal, cette information pourrait être suffisante pour situer la souris et le lion dans le même groupe et ce, même s'ils ne se ressemblent pas sous d'autres points de vue. *La similarité est donc relative au contexte* et aux caractéristiques qu'on décide de prendre pour décrire les individus.

En résumé, *la similarité est une question de degré, elle peut être mesurée et elle est ancrée dans le contexte*. De plus, dans le cadre du clustering, elle se définit en opposition avec une autre mesure, soit celle de la *dissimilarité* :

> Le clustering est la classification d'objets similaires en groups [*différents*] [...] (Abonyi et Feil, 2007, p. IX) [Notre traduction]

> Le clustering [...] est une méthode de création de groupes d'objets ou de clusters, de telle sorte que les objets d'un cluster sont très similaires entre eux et les objets de [*différents*] groupes sont assez [*distincts*] (Ma *et al.*, 2007, p. 3) [Notre traduction]

La notion technique de clustering est complétée donc par le concept de dissimilarité qui s'oppose à celui de similarité. Les groupes sont créés sur la base de la similarité entre individus de même groupe et de la dissimilarité entre les groupes.

La similarité est un concept qui s'applique aux individus d'un cluster alors que la dissimilarité concerne les clusters eux-mêmes. Cette relation entre similarité et dissimilarité constitue la base de toute méthode ou tout algorithme développés dans ce contexte. Ceci nous mène à une autre notion qui définit davantage les potentialités du clustering :

> Le clustering est une opération d'identification de groupes dans un jeu de données à analyser. Les groupes sont appelés cluster. Ces clusters sont des sous-groupes distincts du jeu de données, lesquels organisent les individus plus similaires dans le même sous-groupe, et les individus différents dans des sous-groupes différents. Le rôle du clustering est, donc, de découvrir une sorte de structure naturelle cachée dans le jeu de données. (Wierzchon et Klopotek, 2018, p. 9) [Notre traduction]

Dans ce passage, le clustering est défini comme une méthode de découverte d'une *structure latente et naturelle* présente dans les données. Cette structure naturelle est déterminée par les similarités et les dissimilarités des objets et constitue une organisation latente des données en catégories distinctes. En d'autres termes, le clustering identifie des groupes qui sont *naturellement* signifiants dans le jeu de données analysé :

> Le clustering subdivise les données dans des groupes différents (clusters) qui sont significatifs, utiles, ou les deux. Si l'objectif atteint est la découverte de groupes significatifs, alors les clusters devraient capturer la structure naturelle des données (Tan et Steinbach, 2019, p. 525) [Notre traduction]

Les groupes se définissent donc par leur utilité et leur significativité (au sens statistique), ou les deux, ce qui consiste à trouver cette structure sous-jacente au jeu de données. Les groupes doivent servir à quelque chose, par exemple décrire le jeu de données, et surgir naturellement pour des raisons de significativité statistique.

L'utilité d'une partition est fondamentale car les techniques de clustering visent un objectif plus général, l'*exploration* et la *génération d'hypothèses* :

> [...] Le clustering est utilisé dans plusieurs disciplines pour la génération d'hypothèses interprétatives qui sont utilisées dans l'analyse de jeux de données trop grands ou complexes pour être analysées à l'aide d'une inspection directe du chercheur. Ceci est accompli à l'aide de l'identification de la structure sous-jacente d'un jeu de données. [...] clustering est [donc] un outil pour la génération empirique d'hypothèses. Il identifie la structure latente d'un jeu de

> données et la prise de conscience de l'existence d'une telle structure est utilisée pour « dessiner » des inférences qui permettent de formuler une hypothèse. (Moisl, 2015, p. 15-18) [Notre traduction]

Cette nouvelle dimension est très pertinente pour le présent travail. Dans plusieurs disciplines, le clustering est utilisé pour explorer de grands jeux de données. Quand l'extraction des connaissances implique l'analyse d'un grand nombre d'individus, le clustering représente un support important dans une première phase d'exploration et mène le plus souvent à des hypothèses sur l'organisation ou la structure sous-jacente des données. Cette situation est fréquente dans l'analyse de presse, car une étape d'exploration préliminaire des données est nécessaire afin d'identifier les hypothèses d'interprétation qui peuvent être proposées. Même si elle est grossière, dégager une première structure sous-jacente aux données permet d'atteindre les objectifs d'une première phase d'exploration.

Dans un contexte d'analyse de textes, ces principes peuvent être appliqués et le clustering constitue sans aucun doute un outil important pour l'analyse et l'exploration. Sa définition peut être aisément adaptée comme suit :

> Le clustering de textes consiste à regrouper des documents similaires (ou associés) dans des classes. À cet égard, le regroupement de documents est [...] une opération de collecte de documents (Baeza-Yates et Ribeiro-Neto, 1999, p. 173) [Notre traduction]

Cet extrait identifie une des applications les plus répandues du clustering, soit le regroupement de documents, c'est-à-dire repérer les documents qui sont les plus similaires entre eux et les groupes qui sont les plus dissimilaires. Cette définition peut être explicitée d'un point de vue technique :

> Le clustering de textes: étant donné une collection D de documents, une méthode de clustering de textes sépare automatiquement ces documents en K

> clusters selon certains critères prédéfinis (Baeza-Yates et Ribeiro-Neto, 1999, p. 286) [Notre traduction]

Toutefois, il faut souligner que le regroupement de documents n'est qu'une des applications possibles de la méthode de clustering sur des données textuelles. Les autres applications ne sont pas pertinentes dans le cadre de ce travail et ne seront donc pas traitées.

Il est parfois possible de visualiser les résultats d'un clustering. La figure 2.7 montre une *partition* d'un jeu de données obtenu par un algorithme de clustering, lequel a été exécuté avec l'objectif de trouver trois clusters. Il s'agit alors d'une partition à trois clusters.

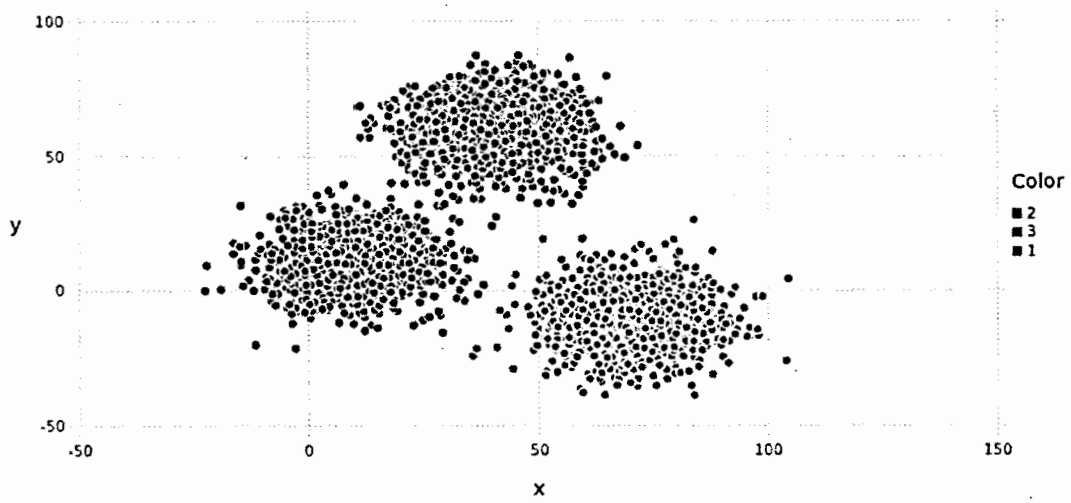


Figure 2.7 Représentation bidimensionnelle d'une partition à trois clusters

2.3.2.2 Le problème de la classification

Le paragraphe précédent peut être résumé ainsi : le clustering est l'opération qui consiste à réunir des objets (individus ou variables) en un nombre limité de groupes appelés clusters (ou segments ou classes) et possédant deux propriétés, l'*homogénéité*

interne ou *intracluster* et l'*hétérogénéité externe* ou *intercluster* (Tufféry, 2017). La partition effectuée par le clustering doit effectivement maximiser l'homogénéité interne des clusters, c'est-à-dire que *les individus doivent être le plus possible similaires entre eux*, et l'hétérogénéité externe des clusters, c'est-à-dire que *les clusters doivent être le plus possible différents entre eux*. Les clusters sont définis au cours de l'opération de clustering, ce qui en fait une opération de type descriptif (exploratoire) et non prédictif (confirmative).

Le problème posé par le clustering n'est pas trivial. Il s'agit d'analyser tous les individus du jeu de données et d'évaluer toutes les partitions possibles afin de maximiser une fonction qui tient compte de l'homogénéité intracluster et de l'hétérogénéité intercluster. En d'autres termes, ce problème mène à une *explosion combinatoire* qui correspond au *nombre de Bell*. Le nombre de Bell est le nombre de partitions possibles qui peuvent organiser un ensemble d'individus. Par exemple, si on considère un ensemble E composé de quatre individus $E = \{a, b, c, d\}$, le nombre de Bell répond à la question suivante : quelles sont les partitions possibles P que ces quatre individus peuvent former et combien y en a-t-il? La réponse est la suivante :

1- une partition composée d'un seul cluster : $P_1 = \{(a, b, c, d)\}$
2- sept partitions distinctes composées de deux clusters : $P_1 = \{(a, b) ; (c, d)\}$, $P_2 = \{(a, c); (b, d)\}$, $P_3 = \{(a, d); (b, c)\}$, $P_4 = \{(a); (b, c, d)\}$, $P_5 = \{(b); (a, c, d)\}$, $P_6 = \{(c); (a, b, d)\}$, $P_7 = \{(d); (a, b, c)\}$
3- six partitions distinctes composées de trois clusters : $P_1 = \{(a); (b); (c, d)\}$, $P_2 = \{(a); (c); (b, d)\}$, $P_3 = \{(a); (d); (b, c)\}$, $P_4 = \{(b); (c); (a, d)\}$, $P_5 = \{(b); (d); (a, c)\}$, $P_6 = \{(c); (d); (a, b)\}$
4- une partition composée de quatre clusters : $P_1 = \{(a); (b); (c); (d)\}$

Les combinaisons possibles explosent avec l'augmentation des individus contenus dans le jeu de données. Le clustering optimise le processus d'évaluation de la meilleure partition possible par le biais d'une heuristique qui permet de ne pas évaluer toutes les partitions possibles, réduisant ainsi les calculs à exécuter.

Cet aspect est important et pose des contraintes aux différents techniques de clustering. En effet, chaque algorithme de clustering doit déterminer ses propres hypothèses de classification et certains *paramètres à fixer en amont*, constituant ainsi sa propre heuristique. Ceci permet de trouver des solutions au problème de l'explosion combinatoire et de réduire les calculs à exécuter. Par exemple, une des approches possibles est de déterminer en amont le nombre de clusters de la partition finale. Le paramètre k est un paramètre très répandu parmi les algorithmes de clustering qui permet d'évaluer la meilleure partition avec un nombre de clusters fixé d'avance. Une autre approche est d'établir en amont la distance maximale entre les objets d'un même cluster, ce qui mène à n clusters dont les individus respectent le seuil de distance établi. Les algorithmes utilisent ces paramètres fixés d'avance comme une constante et se développent ainsi autour de cet ancrage.

2.3.2.3 Les approches

Les méthodes de clustering varient énormément et chacune tente de résoudre le problème de la classification avec des stratégies différentes possédant leurs propres avantages et désavantages. Faire un résumé de ces méthodes ne rentre pas dans les objectifs de ce travail. Cependant, il est pertinent de faire une brève description des caractéristiques principales qui permettent de distinguer les algorithmes. Ces caractéristiques sont présentées sous forme de dichotomies.

La première est la dichotomie entre les *algorithmes hiérarchiques* et les *algorithmes de partitionnement*. Les algorithmes hiérarchiques font l'hypothèse que les données peuvent être organisées comme un arbre hiérarchique, dont chaque branche regroupe des individus, des groupes d'individus ou les deux (figure 2.8). L'arbre est appelé *dendrogramme* et peut être construit de manière *ascendante* ou *descendante*. Le principe général est l'appariement de données (ou de groupes de données) par le biais d'une *notion de distance*, qui constitue l'*alter ego géométrique* de la notion de

similarité. Dans le cas ascendant, les algorithmes exécutent une première itération au cours de laquelle ils identifient les paires de données les plus proches, c'est-à-dire les plus similaires et ils créent ainsi les premiers groupes. La deuxième itération permet de réunir des groupes d'individus avec de nouveaux individus ou d'autres groupes et ainsi de suite, jusqu'à la dernière itération où tous les groupes sont reliés entre eux. L'arbre complet est alors généré. Les algorithmes descendants fonctionnent de la même manière, mais à l'inverse, donc en séparant les groupes et les individus les moins proches. Le groupe d'algorithmes de partitionnement ne produit pas d'arbres, mais sépare de manière horizontale les données entre elles (figure 2.7). Au contraire du partitionnement hiérarchique, le clustering par *partitionnement horizontal* ne fournit aucune information sur la relation entre les clusters. Son hypothèse de classification est basée de manière plus stricte sur le principe d'*homogénéité intraclasse* et d'*hétérogénéité interclasse* plutôt que sur la notion de distance. Dans ce contexte, la distance n'est pas le seul concept qui est exploité, car ce n'est pas important d'observer la distance minimale entre deux individus pour les regrouper, mais plutôt d'observer une optimisation globale des distances de chaque individu et de chaque cluster. Généralement, les itérations des algorithmes s'arrêtent lorsque la fonction de maximisation de l'homogénéité intraclasse totale est atteinte. La meilleure partition sera celle possédant des clusters très homogènes du point de vue intraclasse et très hétérogènes du point de vue interclasse.

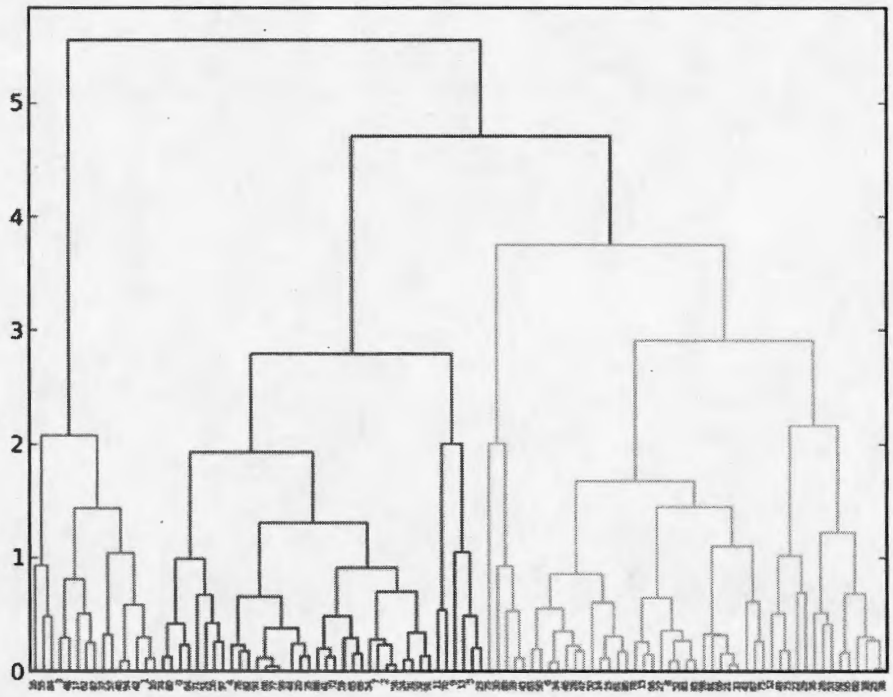


Figure 2.8 Représentation graphique d'un clustering de type hiérarchique

La deuxième dichotomie est celle entre les algorithmes à partitionnement *exclusif* versus ceux à partitionnement *flou*. Les algorithmes à partitionnement *exclusif* attribuent à chaque individu une appartenance à un seul cluster. Il existe de nombreuses situations dans lesquelles un individu peut raisonnablement être placé dans un seul cluster et cette stratégie de classification est alors sûrement la plus adaptée. Elle est très répandue dans toutes sortes d'applications en raison de la facilité d'interprétation des résultats puisqu'on génère moins d'informations à évaluer. Le désavantage de ce type de clustering est de ne pas repérer les individus placés aux limites entre deux ou plusieurs clusters. Au contraire, le partitionnement flou génère une partition dont les clusters se chevauchent et chaque individu peut avoir plus d'une assignation. Un individu peut donc appartenir à un ou plusieurs clusters, selon sa

position dans l'espace. En général, ces algorithmes déterminent un degré d'appartenance de chaque individu à chaque cluster, c'est-à-dire qu'ils assignent une valeur allant de 0 (ne fait absolument pas partie) à 1 (appartient absolument) pour chaque individu en correspondance avec chaque cluster.

La troisième et dernière dichotomie oppose les algorithmes à *partitionnement complet* aux algorithmes à *partitionnement partiel*. Le *partitionnement complet* force une assignation à chaque individu, afin que chacun d'eux ait son cluster d'appartenance. Au contraire, le partitionnement partiel ne force pas les assignations et laisse les individus éloignés sans appartenance aux clusters. Cette dernière stratégie de classification permet de regrouper les individus en clusters encore plus homogènes et d'exclure les individus qui ne présentent pas les caractéristiques appropriées pour appartenir à un cluster. Cette stratégie permet d'identifier en même temps les individus qui représentent des données aberrantes, réduisant ainsi le bruit.

Dans la chaîne de traitement présentée dans ce travail, un algorithme de type *hard clustering* est utilisé. Ce type d'algorithme génère une partition complète avec des clusters horizontaux de type exclusif. L'hypothèse de classification implique donc d'identifier une partition où *les articles* du corpus appartiennent à un *seul et unique cluster* et où chacun d'eux est situé *sur le même plan*. Les avantages de ce genre d'algorithme sont principalement liés à une plus grande interprétabilité des résultats, ce qui est généralement optimal dans une phase d'exploration d'un corpus textuel.

Cette hypothèse de classification est plus adaptée à l'objectif de cette recherche. D'abord, comme l'appartenance de chaque individu n'est pas ambiguë, nous avons la certitude que les articles regroupés ensemble partagent plus de caractéristiques qu'avec les articles des autres clusters, ce qui facilite la tâche d'identification des articles qui partagent sans ambiguïté les mêmes structures sémantiques. Le

désavantage majeur de cette approche est d'obtenir des clusters contenant un certain nombre d'articles qui se situent plutôt aux limites éloignées du cluster et qui peuvent donc constituer du bruit. Cependant, ceci ne compromet pas la réussite de notre recherche car son objectif est d'identifier les structures sémantiques les plus importantes et significatives, *les structures qui sont reconnaissables au-delà de tout doute et qui forment des macrostructures claires*. Dans ce cadre, disposer de clusters non ambigus est certainement un avantage. De plus, comme notre objectif n'est pas d'identifier les relations hiérarchiques entre les structures sémantiques, l'utilisation de la stratégie hiérarchique n'est pas justifiée.

CHAPITRE III

MÉTHODE

Le troisième chapitre se concentre sur la description de la méthode. La première partie est dédiée à l'exemplification des données qui sont utilisées dans un cadre d'analyse de texte assistée par ordinateur, c'est-à-dire la constitution du corpus. Pour la sémiotique computationnelle, cette étape est cruciale. C'est pourquoi nous nous attardons sur celle-ci avant d'initier concrètement le chapitre. Les autres parties du chapitre III définissent la chaîne de traitement qui a été construite pour cette étude. Elle comporte une phase de prétraitement du corpus, à l'aide des outils issus du traitement automatique du langage naturel, une phase de filtrage et une de regroupement automatique, à l'aide des outils de l'apprentissage automatique; et enfin, une étape d'annotation à l'aide d'un outil informatique qui permet la production d'une base de données relationnelle pour maintenir une trace de la lecture et de l'analyse.

La constitution du corpus est une opération fondamentale dans le champ de l'analyse de texte assistée par ordinateur (ATO). Le corpus constitue l'ensemble du matériel qui est utilisé pour l'analyse. Dans le contexte de l'analyse assistée par des algorithmes, cet ensemble devient un « univers clos », sur lequel les algorithmes sont exécutés. Appréhender le corpus comme un « univers clos » est important pour comprendre la nature des résultats produits par les différents outils computationnels utilisés. L'exemple de la chambre chinoise de Searle peut-être utilisé pour la

compréhension du corpus comme un « univers clos ». La chambre chinoise montre qu'un programme informatique produit des résultats sur la base des règles et des informations qui sont à sa disposition et , sans avoir « conscience » de ce qu'il fait. Ainsi, une personne entrainée à fournir une certaine réponse à un stimulus particulier peut exécuter la tâche sans réellement comprendre ce qu'elle fait. Par exemple, si on établit un schéma avec une liste d'idéogrammes chinois d'entrée et une autre d'idéogrammes chinois de sortie et qu'on confie à une personne qui ne connaît pas le chinois, la tâche de produire l'idéogramme de sortie correspondant à l'idéogramme d'entrée, on ne peut pas affirmer que la personne « comprend » ce qu'elle fait ou le message qu'elle envoie. La personne exécute une tâche selon une liste de règles. Le corpus n'est pas une liste de règles, mais c'est le matériel à partir duquel des algorithmes (ou des personnes) peuvent inférer des règles qui seront valables si on les considère dans « l'univers clos » que constitue le corpus. Le concept « d'univers clos » permet de situer le corpus à un niveau d'immanence qui est à la base de l'analyse du texte. Le sens est exprimé par les informations que le texte contient et qui peuvent, à l'occasion, être extraites à l'aide d'un programme informatique. Un programme informatique ne pourra pas extraire d'autres informations à partir du contexte ou d'une encyclopédie, sauf s'il est programmé pour le faire. Et même dans ce cas, son univers sera celui auquel le programmeur lui donne accès. Les algorithmes agissent donc dans des « univers clos » et pour cette raison, l'étape de constitution du corpus dans le cadre de l'analyse de texte assistée par ordinateur est fondamentale. Évidemment, elle l'est également dans une analyse « traditionnelle », mais en ATO, elle est plus critique.

Les disciplines qui utilisent une collection de documents comme objet d'étude ou comme matière première pour étudier un phénomène humain ou social sont nombreuses, comme la linguistique, les études littéraires, la sémiotique, la sociologie, l'anthropologie, les disciplines de la communication, les études culturelles, la

psychologie, l'éducation, etc. Généralement, un corpus peut être défini comme une collection de documents. Ces derniers peuvent être de différentes natures, par exemple des documents vidéo, audio, des textes écrits etc. Chaque document est collecté et rassemblé dans le corpus selon un certain nombre de critères ou principes, permettant ainsi de construire un corpus à des fins scientifiques, c'est-à-dire *pour répondre à une question de recherche*.

En général, en sémiotique, le corpus a une grande importance. Fabbri (2008) souligne que l'étude du corpus est un élément fondamental de la recherche sémiotique. Reprenant les propos de Greimas, Fabbri identifie quatre niveaux dans la recherche sémiotique : les niveaux *épistémologique*, *théorique*, *méthodologique* et *empirique*. En principe, ces quatre niveaux doivent être cohérents entre eux. Le niveau empirique (cfr. 1.1.4) est généralement fondé sur l'analyse de corpus. Fabbri souligne que la sémiotique est une discipline à *vocation scientifique*, car elle accorde une grande importance au point de vue empirique et donc, à l'analyse de corpus. En d'autres termes, la sémiotique est orientée vers l'analyse d'objets empiriquement observables et en revendiquant sa « vocation scientifique », elle a de plus en plus mis le corpus et une vision empirique de celui-ci au centre de ses recherches (Volli, 2014). Cette vision de la recherche sémiotique mène à considérer la valeur empirique du corpus sur la base de ses caractéristiques, dont nous traiterons dans les prochains paragraphes.

Dans un cadre d'ATO, la taille est un des éléments qui détermine la qualité du corpus. Le corpus utilisé en ATO est souvent très volumineux, car l'analyse d'un corpus de petite taille ne justifie pas l'utilisation d'outils informatiques. L'analyse traditionnelle en sémiotique, mais aussi dans toutes les autres disciplines qui font de l'analyse de corpus une de leurs méthodes, est réalisée sans l'assistance informatique et à travers l'observation directe des objets qui constituent le corpus. La méthode « traditionnelle » dépend donc d'une observation systématique et attentive de chaque

objet. En ATO, l'observation directe de chaque objet est le plus souvent exécutée par des algorithmes, laissant à l'humain le soin d'observer des échantillons plus petits. Ce transfert de compétence de l'humain à la machine conduit, selon Rastier, à la découverte de « nouveaux observables » (Rastier, 2011), c'est-à-dire des observables qui n'auraient pas pu être identifiés sans une assistance par ordinateur. Enfin, l'ATO apporte deux avantages : elle accélère l'analyse par rapport aux pratiques traditionnelles et elle donne accès à des nouveaux observables qui échapperaient à l'analyse manuelle.

Avant de passer à la définition de corpus et à la description de ses caractéristiques et propriétés, nous estimons nécessaire d'introduire les quatre niveaux du corpus, tels que définis par Rastier. La constitution du corpus doit tenir toujours compte de ces niveaux. L'*archive* correspond à « l'ensemble des documents accessibles pour une tâche de description ou une application ». Par définition, l'archive est un ensemble sans forme, non systématique, pour lequel il est difficile d'avoir une vue globale. Le *corpus de référence* est un ensemble de documents plus généraux, qui composent un fond au corpus qui sera matériellement étudié. Il situe déjà le corpus à l'intérieur d'une recherche particulière qui poursuit des objectifs spécifiques. Le *corpus d'étude* est l'ensemble de textes collectés sur lesquels l'analyse est effectuée. Dans ce cadre, le corpus d'étude est l'échantillon d'une population, soit le corpus de référence. Les *sous-corpus de travail* peuvent varier selon la recherche ou les différentes phases de l'étude. Ce sont de sous-groupes du corpus d'étude.

Dans les prochains paragraphes, nous définirons le concept de corpus, décrirons les critères les plus utilisés pour sa construction et les méthodes de collecte et d'échantillonnage. À notre avis, chacun de ces aspects doit faire partie de la constitution du corpus dans un cadre d'ATO.

3.1 Définition de corpus

Le corpus est un outil de recherche qui est utilisé par plusieurs disciplines et domaines de recherche. Synthétiser toutes les définitions existantes est une opération complexe et de plus, non pertinente dans le cadre de cette recherche. L'objectif de cette section est d'élaborer une définition de corpus qui soit cohérente avec notre étude, qui se situe dans le cadre de l'analyse de la presse, et avec notre méthode de recherche, qui est un hybride entre sémiotique et *text mining*.

Le point de départ est la définition suivante : le corpus est une collection de documents qui sont utilisés pour répondre à une question de recherche spécifique. Cette définition s'inspire de la notion de corpus en linguistique, où il est souvent défini en ces termes :

> [un corpus est un] ensemble d'énoncés écrits ou enregistrés dont on se sert pour la description linguistique. La méthode du corpus s'impose dans le domaine descriptif, car il est impossible de recueillir tous les énoncés d'une communauté linguistique à un moment donné, et dangereux de fabriquer ses exemples soi-même. Le linguiste limite la taille du corpus d'une manière plus ou moins arbitraire tout en essayant de le rendre représentatif de l'état de langue en question » (Mounin, 2004, p. 89)

Cette définition met de l'avant un certain nombre d'éléments pertinents dans le cadre de la présente recherche. Premièrement, le corpus y est défini comme étant la *matérialité* (au sens de physicalité[13]) à travers laquelle on étudie la réalité. En particulier, le corpus est un objet dont « on se sert » pour décrire un phénomène linguistique, par exemple l'utilisation du passé simple dans les textes du moyen-âge, ou un phénomène de nature différente. Le rôle du corpus est *fonctionnel*. C'est un

[13] Il est possible de l'interpréter comme l'ensemble des objets visibles et tangibles qu'on collection pour étudier un phénomène du monde

moyen pour décrire un phénomène humain ou social. Le corpus est le lieu où la réalité externe et l'univers du chercheur se rencontrent. La description d'un phénomène devient possible par la matérialité du corpus, qui est composé d'objets concrets et réels, c'est-à-dire des documents. Chaque document est une manifestation du phénomène à l'étude. La potentialité descriptive du corpus dépasse celle du simple document ou de la manifestation isolée. La collection de plusieurs manifestations[14] d'un phénomène augmente la potentialité descriptive du phénomène. Pour que la force descriptive du corpus soit effective, les documents récoltés doivent respecter un certain nombre de critères qui garantissent la sélection de manifestations effectives du phénomène étudié.

Deuxièmement, l'utilisation d'un corpus implique une *approche empirique*. En effet, le pratique de l'analyse d'un corpus suggère une posture épistémologique de type empirique, car la méthode pour atteindre la connaissance se base sur l'observation de la matérialité du corpus. Cette perspective peut être plus ou moins radicale, mais elle est toujours présente lorsqu'on analyse un corpus pour décrire un phénomène.

Troisièmement, la « méthode du corpus » exige aussi l'établissement de *règles de sélection*. Si un document doit respecter des critères précis afin de faire partie du corpus, la sélection des documents doit donc respecter des règles. La méthode de sélection de documents peut varier d'une discipline à l'autre et elle peut être plus ou moins systématique ou arbitraire, mais elle représente un élément très important de la démarche, car il n'existe pas de description possible sans une sélection pertinente. La sélection se présente aussi comme une question pratique, car il est effectivement

[14] Il ne faut pas voir les manifestations comme une répétition de la même manifestation identique. Ce sont plusieurs manifestations, non identiques, chacune avec sa particularité et son contexte de production, qui apportent richesse à la matérialité d'un corpus.

impossible d'obtenir l'extension complète des manifestations perceptibles d'un phénomène.

Enfin, le corpus est un *échantillon* de toutes les manifestations possibles d'un phénomène et sa taille est déterminée en suivant des règles plus ou moins arbitraires. Un échantillon d'un phénomène n'est pas le phénomène, mais plutôt un « univers clos » à partir duquel généraliser des règles de fonctionnement du phénomène. Toutefois, il demeure possible que le corpus ne soit pas adapté pour la découverte de toutes les propriétés d'un phénomène. Prendre conscience des caractéristiques du corpus et de ses potentialités descriptives est une étape de la « méthode du corpus », ce qui mène à l'idée qu'un corpus doit être *représentatif* du phénomène étudié ou d'une partie de ce dernier.

La propriété descriptive du corpus est au cœur de la « méthode du corpus » en linguistique et ceci en raison de caractéristiques propres à cette discipline et aux typologies de questions de recherche qui y sont posées. Le lien entre une question de recherche et un corpus est abordé dans l'extrait suivant :

> La grammaire descriptive d'une langue s'établit à partir d'un ensemble d'énoncés qu'on soumet à l'analyse et qui constitue le *corpus* de la recherche. Il est utile de distinguer le *corpus* des termes voisins désignant des ensembles d'énoncés : l'« univers » est l'ensemble des énoncés tenus dans une circonstance donnée, tant que le chercheur n'a pas décidé si ces énoncés entraient en totalité ou en partie dans la matière de sa recherche. [...] La totalité des énoncés recueillis est l'*univers du discours*. (Dubois, 2002, p. 123)

L'« univers clos » du corpus est construit à des fins de recherche spécifique et ceci détermine la collecte de documents et la pertinence d'un corpus. Les linguistes sont habitués à recueillir des énoncés « tenus dans une circonstance donnée » et qui *représentent* donc la dite circonstance. L'étude de la langue se prête bien à la construction de corpus qui peut répondre à une question de recherche spécifique, car

la langue, ou certaines de ses propriétés, est un phénomène qui se manifeste. Toutefois, la « méthode du corpus » s'applique à tout domaine du savoir. En effet, la représentativité des documents récoltés dépend de la typologie de recherche qui est menée et chaque corpus possède sa propre potentialité descriptive en relation avec une question de recherche et à l'intérieur d'un domaine du savoir.

La terminologie utilisée dans les différentes disciplines peut aussi varier et porter à confusion. Dubois (*ibid.*) distingue, par exemple, l'*univers* de l'*univers du discours*. L'univers est l'extension complète du phénomène[15]. Alors que l'univers du discours est la collection de documents, le corpus. Toutefois, les affirmations de Dubois contribuent à l'identification d'autres composantes de la notion de corpus, comme le concept d'échantillon et la méthode de sélection de documents :

> À partir de l'univers des énoncés réunis, le linguiste trie les énoncés qu'il va soumettre à l'analyse : dans le cas qui nous intéresse, ce pourra être l'ensemble des phrases, ou groupes de phrases, comprenant des mots présentant tel trait phonétique ou bien une terminaison ou une origine étrangère. Ce sont uniquement ces segments d'énoncés qui seront soumis à l'analyse et qui constitueront le corpus. On pourra aussi, sur des bases statistiques, délimiter soit dans l'univers, soit dans le corpus, des passages qui seront soumis à une analyse quantitative : par exemple, une page toutes les dix pages; les pages ainsi retenues constituent un échantillon du texte. Par hypothèse, on considérera comme échantillon toute partie représentative du tout. Le corpus peut évidemment, si le chercheur le juge utile ou nécessaire, être constitué par l'univers d'énoncés tout entier. De même, une analyse quantitative pourra fort bien se passer d'échantillonnage. Le corpus lui-même ne peut pas être considéré comme constituant la langue (il reflète le caractère de la situation artificielle dans laquelle il a été produit et enregistré), mais seulement comme un échantillon de la langue (Dubois, 2002, p. 124)

Dubois introduit l'idée de *distinguer différents statuts du corpus*, ce qui a aussi été élaboré par Rastier, avec les quatre niveaux de corpus décrit au début de ce chapitre.

---

[15] L'archive, selon Rastier, se situerait ainsi entre l'univers et le corpus.

Il existe un corpus composé de l'ensemble des énoncés réunis, mais il existe aussi un corpus soumis à l'analyse. La proposition de Dubois est donc de trier et de filtrer les documents pour sélectionner un *sous-corpus* qui sera soumis à l'analyse.

La définition de corpus qui suit est basée sur des éléments qui sont communément acceptés par la communauté scientifique et qui peuvent être rattachés à la relation entre corpus et questions de recherche, mais elle ajoute d'autres nuances :

> Les corpus sont ainsi des artefacts, c'est-à-dire des objets construits. Leur construction répond à un programme de recherches déterminé par un certain type d'usages langagiers dont ils sont censés n'offrir qu'une représentation partielle. Aucun corpus ne saurait en effet refléter la langue dans son ensemble, et se poser en référence universelle (Neveu, 2004, p. 103)

Le corpus est donc un artefact, car il est un objet construit. On ne le trouve pas naturellement dans le monde. Le processus de construction répond à des règles et plus particulièrement au besoin de répondre à une question de recherche spécifique. Toutefois, le corpus ne peut qu'être un échantillon partiel du phénomène et il ne peut aucunement prétendre à une représentation parfaite du phénomène. Par conséquent, il est important de bien choisir les documents qui font partie du corpus afin d'en améliorer la potentialité descriptive du corpus.

Ces concepts reviennent souvent dans les définitions de la notion de corpus, comme dans la suivante :

> Un *corpus* est une collection de textes d'une langue qui sont sélectionnés et classés selon des critères linguistiques explicites afin d'être utilisés comme échantillon de la langue. (Sinclair, 1996) [Notre traduction]

Bien que ces définitions soient tirées de la linguistique, les composantes qu'elles énoncent ont une valeur plus générale et s'appliquent tant aux études sémiotiques

qu'à d'autres domaines du savoir comme la sociologie. En général, les documents réunis peuvent avoir différentes formes et appartenir à toutes sortes de *productions sémiotiques*[16], car un corpus n'est pas d'emblée un corpus linguistique. Ceci dépend des propriétés spécifiques de ces productions, de la discipline dans laquelle l'étude s'inscrit et des questions de recherche qui sont posées. Il n'est pas rare qu'un corpus comporte des collections d'images photographiques, d'images picturales, de séquences vidéo, d'installations artistiques, etc. En anthropologie ou en sociologie, par exemple, il est plus fréquent de réunir des interviews réalisées en faisant du terrain. La *nature* du document peut être variée, mais elle peut aussi changer en fonction du *genre.* Par exemple, en études littéraires, il est fréquent d'étudier un genre littéraire plutôt qu'un autre, et donc de distinguer les documents sur la base d'une catégorie ontologique s'appuyant sur le genre. Les critères de sélections peuvent être différents et varier d'une discipline à l'autre ou d'une recherche à l'autre. Toutefois, les composantes qui ont été décrites précédemment demeurent importantes pour toute recherche qui adopte la « méthode du corpus », dont l'objectif est de *réunir des exemplaires d'un phénomène spécifique*.

Dans les sections précédentes, les principes généraux de constitution d'un corpus ont été énoncés. Toutefois, la notion de corpus nécessite d'être détaillée davantage. En particulier, les critères, les méthodes de collecte et d'échantillonnage qui s'appliquent au contexte de recherche de la présente étude doivent être précisés. Pour ce faire, nous utilisons comme point de départ, la définition de corpus donnée par Greimas :

> Constituer un corpus ne signifie donc pas simplement se préparer à la description, car de ce choix préalable dépend, en définitive, la valeur de la description, et, inversement, on ne pourra juger de la valeur du corpus qu'une fois la description achevée [...] Un certain nombre de précautions et de conseils pratiques doivent donc entourer ce choix, afin de réduire, autant que possible, la

[16] Rastier, par exemple, considère qu'un bon corpus doit être homogène au niveau du genre.

> part de subjectivité qui s'y manifeste. On dira qu'un corpus, pour être bien constitué, doit satisfaire à trois conditions : être représentatif, exhaustif et homogène (Greimas 1966, p. 142-143)

Ainsi, le corpus est décrit comme un objet qui doit essentiellement satisfaire trois conditions : *représentativité*, *exhaustivité* et *homogénéité*. Ces critères ou conditions constituent des éléments fondamentaux du processus de construction du corpus et en détermine sa qualité, laquelle influence la qualité des résultats qu'on peut obtenir. Dans les prochains paragraphes, nous allons raffiner la notion de corpus présentée plus haut en y ajoutant de nouveaux critères tout en reprenant ceux énoncés par Greimas.

3.2 Critères de constitution du corpus

3.2.1 Orientation

L'*orientation* est le premier critère de constitution d'un corpus. Son rôle est de déterminer le passage de l'archive au corpus de référence. Sa caractéristique principale est de rendre cohérent le corpus en fonction d'une question de recherche. Il est donc le critère qui assure le lien entre la nature de la recherche et la matérialité empirique mise en place pour l'analyse. Un corpus est donc *orienté* sur le projet de recherche par le biais de son corpus de référence.

Ce critère est identifié explicitement en linguistique et, plus particulièrement, dans la tradition linguistique anglo-saxonne. Par exemple, Wynne (2005) définit l'orientation comme le premier critère à adopter pour la construction d'un corpus, car il force le chercheur à s'orienter vers le langage ou la variété de langage qui doit être échantillonné. Les préoccupations de Wynne sont celles d'un sociolinguiste et dans ce cadre, le critère de l'orientation présente une valeur sociologique. Un phénomène linguistique qui est approché de manière sociale implique la construction d'un corpus

qui est orienté vers la communauté linguistique de référence. Dans ce contexte, l'orientation assure le lien entre l'aspect sociologique et l'aspect linguistique de la recherche. Ce critère peut, tout de même, être généralisé à toute construction de corpus. En effet, l'orientation ne doit pas toujours assurer un lien sociologique avec la matérialité à l'étude, mais elle doit surtout assurer *l'identification de la classe d'exemplaires d'un phénomène* et ceci, en fonction d'une recherche précise, y compris ses hypothèses et ses objectifs.

L'orientation fait donc le lien entre une question de recherche et le corpus, et ceci conduit à d'autres considérations :

> Il faut donc constamment garder à l'esprit que le corpus n'est peut-être pas l'échantillon le plus approprié pour toutes les études que l'on pourrait souhaiter mener. Plus l'objectif que nous avons en tête pour un corpus est spécifique, mieux sera la collecte de données (Clear, 1992, p. 26-27) [Notre traduction]

Dans ce passage, Clear souligne un élément déjà énoncé dans le paragraphe précédent : un corpus ne peut pas être un échantillon approprié pour toutes les études à réaliser. Pour cette raison, plus l'objectif que nous avons en tête pour un corpus est spécifique, plus la collecte de données sera efficace.

La première étape de construction d'un corpus est donc fondamentale pour la réussite du processus. L'orientation met en place les premiers éléments structurants de la conception générale du processus de constitution :

> Certains de ces éléments structurants de la construction d'un corpus concernent la conception générale : par exemple, les types de textes inclus, le nombre de textes, la sélection de textes particuliers, la sélection d'échantillons de texte et leur longueur (Biber, 1993, p. 243) [Notre traduction]

Ce genre de considérations préalables à la construction d'un corpus peut changer selon le type de recherche qui est menée. Ces considérations sont aussi importantes en sémiotique :

> Au sens restreint, il s'agit d'un produit ou d'un groupe de produits sémiotiques intégraux retenus sur la base de critères objectifs, conscients, explicites, rigoureux et pertinents pour l'application souhaitée (Hébert, 2016, p. 86)

Dans cet extrait, Hébert énumère un certain nombre de critères qui doivent être considérés lors de la construction d'un corpus et les relie tous à la question de recherche et à l'application du corpus qui sera faite dans le projet de recherche.

Les mêmes considérations entrent en jeu dans d'autres disciplines comme l'analyse du discours :

> le corpus n'est pas un simple recueil de textes qui serait à disposition et qu'il suffirait de compiler. Il est construit en fonction des questions et des hypothèses de recherche, voire même en fonction de la conception que l'on a de l'analyse du discours, ou encore des outils ou des catégories que l'on utilise (Née, 2017, p. 41)

Née souligne ici encore plus fortement le lien entre la recherche et la construction de corpus, en ajoutant l'approche, le cadre théorique et la méthodologie d'une recherche comme éléments d'orientation d'un corpus, ce qui est aussi très important dans un cadre d'analyse de texte assistée par ordinateur :

> Le choix d'un corpus présuppose... que ce corpus constitue bien un objet d'étude; c'est-à-dire, que l'analyste le perçoive comme une entité ou un objet dans l'univers référentiel qui l'intéresse. En définitive, même si ce n'est que de manière implicite, l'analyste fait des hypothèses sur les conditions d'existence de cet objet, sur ses lois de production, sur les paramètres qui le font reconnaître dans cet univers référentiel (Reinert, 1990, p. 27)

Dans ce dernier passage, on souligne à nouveau l'importance de construire un corpus pour l'étude d'un phénomène qui est abordé par la construction d'hypothèses et des questions de recherche précises.

Enfin, dans cette première phase de planification du processus de construction du corpus, des questions de nature administrative peuvent également apparaître :

> La phase de planification de la construction du corpus visera à spécifier des éléments de deux domaines connexes: la conception linguistique du corpus et les coûts d'administration du projet (Atkins *et al.*, 1992, p. 3) [Notre traduction]

Atkins souligne les aspects de nature administrative à prendre en compte dans la phase de planification d'un corpus, ce qui ajoute un autre élément au concept d'orientation. Ainsi le corpus doit respecter l'orientation du projet de recherche en tenant aussi compte des ressources prévues.

Pour l'étude de l'opinion publique par exemple, différents corpus peuvent être construits, selon le cadre de recherche et l'approche théorique ou l'objectif de recherche. Certains s'orienteront vers un corpus contenant des interviews téléphoniques réalisés auprès de citoyens, d'autres vers la collecte d'articles de journaux ou d'émissions radio. Ces choix impliquent que le corpus est orienté sur la recherche et toutes ses composantes, comme la problématique, la question de recherche, le cadre théorique, la méthodologie ainsi que la planification et des ressources, y compris celles budgétaires et de temps.

3.2.2 Pertinence

L'identification d'une archive et le passage au corpus de référence est donc la première étape de construction d'un corpus. Ce premier processus est encadré par le critère de l'orientation. D'autres critères sont importants pour déterminer le matériel

empirique qui sera effectivement utilisé dans l'analyse et donc, pour établir le corpus d'étude. Le premier est celui de la *pertinence* qui selon Rastier s'applique de manière transversale aux différents niveaux du corpus, tout en devenant opératoire lors de la construction du corpus d'étude. Lors du passage du corpus de référence au corpus d'étude, la pertinence du matériel utilisé joue un rôle prépondérant et limite de manière directe le processus de sélection et de collecte des exemplaires du phénomène à l'étude. Rastier affirme que « il faut s'assurer que le corpus traité ait été recueilli correctement et qu'il est entièrement pertinent pour la tâche à accomplir » (Rastier, 2005b, p. 75).

Le critère de pertinence *semble donc se fondre avec celui d'orientation*. Toutefois, la pertinence est un élément qui s'applique surtout lors de la récolte des documents et moins dans la phase de planification. Elle s'applique à chaque exemplaire qui est ajouté à la récolte et mène au corpus d'étude. Le processus de sélection des exemplaires doit donc respecter une règle de pertinence :

> *Règle de pertinence* : Les documents retenus doivent être adéquats comme source d'information pour correspondre à l'objectif qui suscite l'analyse (Bardin, 1977, p. 128)

Bardin souligne l'importance de ne retenir que les documents qui sont des sources d'information adéquates pour répondre à une question de recherche et pour atteindre un objectif précis. Le processus de sélection des éléments qui constitueront le corpus d'étude est exécuté sur le corpus orienté, c'est-à-dire sur le corpus de référence, et il doit respecter la règle de pertinence, ce qui assure le lien entre la matérialité empirique du phénomène et la question de recherche.

En d'autres termes, *la pertinence semble assumer un rôle similaire à celui de l'orientation, mais elle s'applique à un niveau différent de la construction du corpus*. Chaque document doit faire partie d'un corpus orienté, mais doit aussi répondre à des

questions de ce genre : est-ce que le document est un exemplaire du phénomène étudié et son analyse permet-elle d'atteindre l'objectif de la recherche? Pour que le document soit pertinent, le chercheur devra s'assurer qu'il est un exemplaire du phénomène à l'étude et qu'il contient les propriétés qui seront l'objet de la phase d'analyse du corpus.

Si on veut étudier le rôle des personnages féminins dans les productions de Walt Disney, quel est le corpus de référence et quel est le corpus d'étude? Et comment l'orientation et la pertinence se combinent-elles? Dans une première phase, le chercheur s'orientera vers une typologie de documents, sur la base de la nature de la recherche. Il est possible, par exemple, d'identifier le corpus de référence avec l'œuvre complète de Walt Disney et ainsi, de trouver un corpus orienté. Toutefois, ce ne sont pas tous les documents du corpus de référence qui sont pertinents pour atteindre l'objectif de recherche. Il faudra donc effectuer un choix pour construire le corpus d'étude. Il est possible, par exemple de définir comme pertinents tous les extraits des films de Walt Disney où les personnages masculins parlent à des personnages féminins, ou toutes les scènes dans lesquelles les personnages féminins prennent la parole ou toutes les scènes où les personnages féminins apparaissent, etc.

Enfin, la pertinence réduit le bruit, phénomène qui généralement diminue la qualité du corpus. Le bruit désigne le nombre de documents d'un corpus qui ne sont pas pertinents pour l'analyse. Il définit donc le phénomène de « fourvoiement» qu'un certain nombre d'éléments du corpus a sur l'analyse, car, bien que respectant quelques critères, ces documents ne sont pas pertinents pour une étude particulière. Le bruit est un phénomène spéculaire à celui du *silence*, qui est abordé dans le critère de l'exhaustivité.

3.2.3 Cohérence ou homogénéité

Un corpus doit aussi respecter des critères qui sont internes à la collection, comme la *cohérence* ou *homogénéité*. Un corpus est homogène si tous les éléments qui le composent partagent un certain nombre de caractéristiques. Ce critère contribue à améliorer la qualité de la collection et à bien cibler les documents sur la base des propriétés qu'ils contiennent. La cohérence ou homogénéité fait partie du groupe de critères fondamentaux pour la construction d'un corpus et elle est définie de manière non ambiguë :

> *Règle d'homogénéité* : les documents retenus doivent être homogènes, c'est-à-dire obéir à des critères de choix précis et ne pas présenter trop de singularité en dehors de ces critères de choix. Par exemple, des entretiens d'enquête, effectués sur un thème donné, doivent : être tous concernés par ce thème, avoir été obtenus par des techniques identiques, être le fait d'individus comparables. Cette règle est surtout utilisée lorsqu'on désire obtenir des résultats globaux ou comparer les résultats individuels entre eux (Bardin, 1977, p. 128)

Bardin souligne ici l'importance d'établir des critères de choix précis lors de la récolte des documents. La collection finale ne doit pas présenter de documents ayant des propriétés qui ne respectent pas totalement les critères de choix et de collecte. En particulier, les éléments doivent partager un certain nombre de propriétés pour que le corpus soit homogène, et c'est la présence ou l'absence de ces propriétés qui fait partie des critères de sélection des documents. Un corpus homogène améliore la qualité des résultats et augmente la fiabilité des analyses et des interprétations. Ainsi, l'homogénéité est souvent considérée comme un *indice de fiabilité*. Toutefois, le critère de l'homogénéité peut s'appliquer selon différents degrés, d'un niveau plus radical, qui mène à des corpus très homogènes, à un niveau plus souple, qui produit des corpus moins homogènes. Le degré d'application du critère d'homogénéité dépend aussi de la nature de la recherche et de la typologie de l'objet de recherche

Cependant, il demeure important de prendre conscience des impacts que ce critère a au niveau des résultats et des interprétations qu'on peut tirer de l'analyse du corpus.

Il est également important de distinguer l'homogénéité de l'*hétérogénéité*. Dans le contexte de la constitution du corpus, *ces deux critères ne sont pas opposés mais complémentaires* comme l'illustre le paragraphe suivant. L'adjectif qui s'oppose le mieux à « homogène » est *hétéroclite*, qui met de l'avant le concept d' « étrange ».

> Lorsque nous utilisons [le terme *corpus*], nous sous-entendons « corpus de documents homogènes», à savoir un ensemble de documents qui ne soit pas hétéroclite. Il ne s'agit pas de considérer n'importe quel ensemble de documents sans aucun rapport les uns avec les autres. Par exemple, un ensemble de brevets relatifs aux céramiques, un ensemble de publications mondiales sur l'intelligence artificielle constituent pour nous, des corpus homogènes (Chartron, 1988, p. 16)

Selon Chartron, l'homogénéité est une composante implicite de la notion de corpus. L'auteur souligne le rôle de l'homogénéité en opposition au concept d'ensemble hétéroclite. Un corpus n'est pas un ensemble des documents qui n'ont « aucun rapport les uns avec les autres », mais est plutôt composée de documents qui partagent des propriétés entre eux, ce qui, au final, constitue le fondement du concept de corpus lui-même.

3.2.4 Hétérogénéité

Le critère d'homogénéité essaie de limiter les impacts négatifs qu'un ensemble hétéroclite peut avoir sur la qualité de résultats et sur la qualité des interprétations. La qualité des résultats et la fiabilité des analyses dépendent aussi de la qualité des *sources d'information* qui doivent être *hétérogènes* le plus possible, ceci afin d'éviter les biais qu'une homogénéité des sources d'information peut apporter.

> Un corpus de référence n'est pas une collection de: matériel provenant de différents domaines spécialisés — technique, dialectal, juvénile, etc. Il est une collection de materiel homogène, mais qui est regrouppé à partir de sources variées ainsi que l'individualité de la source soit éclipsée, à moins que le chercheur ne veuille isoler un texte particulier. La diversité des sources est une protection essentielle, jusqu'à ce que des travaux tels que les catégorisations de Biber (1988) utilisant des critères linguistiques internes aient suffisamment avancé pour permettre un échantillonnage efficace (Sinclair, 1991, p. 17-18) [Notre traduction]

Selon Sinclair, le corpus doit donc refléter la variété des sources en lien avec le phénomène étudié, mais selon un degré de pertinence avec l'objectif de la recherche. Le critère de l'hétérogénéité est mieux respecté si la variété des sources d'information apporte des avantages en termes d'analyse du phénomène et de possibilité d'atteindre les objectifs. L'application de ce critère peut donc varier selon la nature de la recherche, le domaine du savoir ou l'objectif de la recherche. En effet, en sciences humaines, il n'est pas rare de voir des sources peu hétérogènes et, des fois, même uniques, comme dans des recherches qui portent sur l'analyse d'un seul journal, ou d'une seule œuvre d'un seul philosophe; ce qui n'affecte point la qualité de l'analyse. Dans le cas de l'étude de la tendance de vote en sciences politiques, au contraire, la qualité de résultats et la valeur prédictive de l'étude sont étroitement liées à la variété socio-culturelle des répondants. Le critère d'*hétérogénéité* ne s'applique donc pas à tout corpus, mais cela dépend fortement du type de questions de recherche qui sont posées.

3.2.5 Exhaustivité

Le critère d'exhaustivité est partiellement lié à celui d'hétérogénéité. Comme on l'a vu, la variété des sources d'information peut être déterminante pour l'étude de certains phénomènes. Quant à l'*exhaustivité*, elle met de l'avant la qualité de la *couverture du phénomène* par le corpus. Il faut donc bien évaluer les sources d'information et les documents qui doivent nécessairement être inclus dans le corpus

d'étude. En respectant toujours les limites qui sont imposées par les autres critères, l'exhaustivité assure qu'aucune source d'information ou document ne soit négligée pour des raisons non motivées.

> *Règle de l'exhaustivité* : une fois défini le champ du corpus (entretiens d'une enquête, réponses à un questionnaire, éditoriaux d'un quotidien de Paris entre telle et telle date, émissions de télévision concernant tel sujet, etc.), il faut prendre en compte tous les éléments de celui-ci. Autrement dit, il n'y a pas lieu d'ignorer un élément pour une raison quelconque (difficulté d'accès, impression de non-intérêt) non justifiable sur le plan de la rigueur. Cette règle est complétée par la règle de non-sélectivité. Par exemple, on réunit un matériel d'analyse des publicités pour automobiles parues dans la presse pendant une année. Toute annonce publicitaire répondant à ces critères doit être recensée. (Bardin, 1977, p. 127)

L'exhaustivité représente un *besoin de rigueur* dans la constitution du corpus. Sur la base de la question de recherche, le matériel empirique visé pour l'analyse doit être le plus possible complet et couvrir le plus possible l'étendue du phénomène à l'étude. L'exhaustivité détermine la qualité de l'étude et son niveau de fiabilité. L'importance d'avoir un corpus exhaustif est souvent mise de l'avant dans les recherches en sciences humaines :

> Le principe d'exhaustivité a été considéré, tout le long du XIXe siècle — et il l'est encore aujourd'hui —, comme la condition *sine qua non* de toute recherche humaniste. (Greimas, 1966, p. 143)

Greimas souligne que l'exhaustivité est une *condition sine qua non*. Mais, l'exhaustivité « pure » est impossible en général. Une étape de filtrage est donc requise. Ainsi, tous les éléments qui caractérisent un certain phénomène sont également importants, mais ceci est vrai si et seulement si ces éléments sont cohérents et pertinents pour la recherche :

> Le corpus doit avoir un niveau de détail adapté aux besoins de l'analyse : les adaptations nécessaires peuvent être soit de l'enrichir et de l'affiner, soit d'ajuster, par réduction, le niveau de discrétisation de la réalité à représenter réalisée à partir des données. (Bommier-Pincemin, 1999, p. 418)

Il faut donc prévoir que certains types d'adaptation sont requis pour respecter le critère d'exhaustivité tout en restant cohérent avec les objectifs de la recherche. En effet, si les documents collectés ne sont pas suffisants pour couvrir l'étendue du phénomène à étudier en fonction de la question de recherche, alors il devient nécessaire d'enrichir le corpus. Toutefois, le respect de l'exhaustivité peut aussi conduire à évaluer le « niveau de discrétisation de la réalité à représenter », ce qui implique à l'occasion une révision des objectifs de la recherche. L'exhaustivité amène le chercheur à réfléchir sur la disponibilité des données et, par conséquent, sur la nature de la recherche. En d'autres termes, il est possible que le respect de ce critère entraîne une adaptation de la question de recherche et une délimitation du phénomène étudié. Dans certains cas, ce dernier élément constitue la base même de la définition du critère d'exhaustivité :

> L'exhaustivité du corpus est [...] à concevoir comme l'adéquation du modèle à construire à la totalité de ses éléments implicitement contenus dans le corpus. (Greimas, 1966, p. 143)

Il existe alors l'exhaustivité du phénomène et l'exhaustivité d'un *échantillon*, lesquelles sont inévitablement reliées. Lorsqu'on a accès aux données, il faut se poser la question de leur exhaustivité par rapport à quel phénomène ou portion du phénomène. Ainsi défini, ce critère permet d'adapter la recherche aux données disponibles, ce qui est extrêmement important en sciences humaines, car il est très fréquent qu'on ne puisse atteindre l'étendue complète d'un phénomène.

Dans ce contexte, l'exhaustivité et la représentativité (dont nous parlerons dans le prochain paragraphe) sont liées aux méthodes d'échantillonnage :

> *[E]xhaustivité* : l'exhaustivité des données (qui assure à l'analyse une base intrinsèque [...]) peut, conformément au principe d'équivalence distributionnelle, être assurée par une partition [...], ou [par le] choix d'un échantillon fini (éventuellement stratifié [...]) sur un espace potentiel continu (Benzécri, 1973)

Dans cet extrait, Benzécri élabore la définition d'exhaustivité en relation étroite avec la construction d'un échantillon fini. Évidemment, la nature de l'échantillon dépend du type d'« espace potentiel continu » visé par la question de recherche. L'échantillon est un concept central dans le processus de constitution du corpus. Les méthodes d'échantillonnage doivent alors permettre d'équilibrer les différents critères de constitution du corpus, dont celui d'exhaustivité.

Enfin, le *silence* est un élément directement relié à l'exhaustivité. Le silence est le phénomène qui affecte négativement la qualité du corpus et qui définit, consciemment ou inconsciemment, tout ce qui a été mis à l'écart relativement au phénomène à l'étude. Une portion du phénomène non représentée dans le corpus constitue une portion de silence du phénomène. Le phénomène du silence est spéculaire à celui du *bruit*, qui est lié au critère de pertinence. À nouveau, le critère d'exhaustivité souligne l'importance de construire un corpus pour la recherche et pour répondre à des objectifs spécifiques et ceci, par le biais d'un processus d'adaptation interactif circulaire qui va du corpus au phénomène à l'étude et vice-versa, en ajustant au besoin le niveau de silence du phénomène.

3.2.6 Réprésentativité

La *représentativité* est l'un des critères les plus importants du processus de constitution d'un corpus et est associé à plusieurs autres critères, tel celui d'exhaustivité. Un corpus doit être représentatif de la réalité qu'on veut décrire ou pour laquelle on veut inférer des règles de fonctionnement. Pour être représentatif, un corpus doit refléter la réalité, c'est-à-dire qu'il doit regrouper un certain nombre de

cas qui expriment les propriétés du phénomène à l'étude. Les résultats d'une recherche et leur niveau d'interprétabilité sont particulièrement sensibles au niveau de représentativité des données. Les résultats peuvent être considérés comme significatifs si un degré minimal de représentativité est atteint. En linguistique, comme pour d'autres disciplines, la représentativité est fondamentale :

> Pour que [le corpus] soit représentatif du système, il faut qu'il manifeste tous les types de situations dans lesquelles le système est amené à fonctionner, autrement dit qu'il fasse apparaître la totalité du champ des signifiés (Martinet, 1975, p. 192)

Le lien entre l'exhaustivité et la représentativité est souligné. Pour qu'un corpus soit représentatif, il faut contrôler le niveau d'exhaustivité, car si une certaine portion du phénomène à l'étude n'est pas représentée par le corpus, alors il sera nécessaire d'augmenter à la fois l'exhaustivité et la représentativité. Tel qu'indiqué dans le paragraphe précédent, le lien entre ces deux critères est ancré dans le concept d'échantillonnage :

> Le corpus lui-même ne peut pas être considéré comme constituant la langue (il reflète le caractère de la situation artificielle dans laquelle il a été produit et enregistré), mais seulement comme un échantillon de la langue. Le corpus doit être *représentatif*, c'est-à-dire qu'il doit illustrer toute la gamme des caractéristiques structurelles. On pourrait penser que les difficultés sont levées si un corpus est *exhaustif*, c'est-à-dire s'il réunit tous les textes produits. En réalité, le nombre d'énoncés possibles étant indéfini, il n'y a pas d'exhaustivité véritable et, en outre, de grandes quantités de données inutiles ne peuvent que compliquer la recherche en l'alourdissant. Le linguiste doit donc chercher à obtenir un corpus réellement représentatif et écarter tout ce qui peut rendre son corpus non représentatif (méthode d'enquête choisie, anomalie que constitue l'intrusion du linguiste, préjugé sur la langue, etc.), en veillant à éviter tout ce qui conduit à un artefact (Dubois, 2002, p. 124)

L'échantillon permet donc d'établir un équilibre entre l'exhaustivité et la représentativité. L'exhaustivité complète étant souvent impossible à atteindre, il est

nécessaire de porter une très grande attention à la représentativité des documents qui peuvent être collectés ou aux sources d'information disponibles. Dubois précise comment déterminer si un corpus est représentatif :

> Un ensemble d'énoncés est *représentatif* quand il contient tous les traits concernés par la recherche et sur lesquels on veut formuler des conclusions; un corpus représentatif d'une langue comporte toutes les caractéristiques structurelles de cette langue impliquée par la recherche (Dubois, 2002, p. 410)

Dubois met de l'avant l'attention qui doit être portée aux propriétés du phénomène qui sont à l'étude lors de la collecte et de l'évaluation de la représentativité d'un document. Le corpus sera représentatif si tous les traits du phénomène ou caractéristiques structurales sont représentés par un certain nombre de documents et ce, peu importe la discipline dans laquelle la recherche s'inscrit. L'obtention d'un corpus représentatif implique la détermination de règles de sélection des éléments du corpus afin de gérer la redondance du matériel à l'étude :

> Le corpus n'est [...] jamais que partiel, et ce serait renoncer à la description que de chercher à assimiler, sans plus, l'idée de sa représentativité à celle de la totalité de la manifestation. Ce qui permet de soutenir que le corpus, tout en restant partiel, peut être représentatif, ce sont les traits fondamentaux du fonctionnement du discours retenus sous les noms de *redondance* et de *clôture*. Nous avons vu que toute manifestation est itérative, que le discours tend très vite à se fermer sur lui-même : autrement dit, la manière d'être du discours porte en elle-même les conditions de sa représentativité. (Greimas, 1966, p. 143)

Dans cet extrait, Greimas souligne la possibilité d'avoir un corpus exhaustif et représentatif et ce, sans que la totalité de l'étendue du phénomène à l'étude fasse partie du corpus. Un corpus partiel peut et doit être un corpus exhaustif et représentatif. Grâce à la redondance de certaines propriétés du phénomène. Pour que le corpus soit représentatif, le chercheur doit donc identifier la redondance du matériel récolté et sélectionner un nombre limité de documents garantissant la

représentativité du phénomène. Dans ce cadre, il est nécessaire d'être conscient des limites de l'échantillon :

> Tous les échantillons sont biaisés d'une manière ou d'une autre. En effet, le problème de l'échantillonnage est le fait qu'un corpus est inévitablement biaisé à certains égards. Les utilisateurs du corpus doivent continuellement évaluer les résultats tirés de leurs études et doivent rapporter les difficultés qui derivent des biais. (Atkins *et al.*, 1992, p. 5-7) [Notre traduction]

Atkins souligne que tout échantillon est biaisé et que le corpus parfait est pratiquement une utopie. Le chercheur doit alors déterminer, au fur et à mesure que sa recherche avance, les adaptations à faire sur le corpus, en améliorant le plus possible l'équilibre entre question de recherche et collecte de données. Cette relation entre échantillon et représentativité est au cœur du concept de corpus de plusieurs auteurs :

> L'unité d'échantillonnage doit être suffisamment grande pour bien représenter le phénomène étudié (Neuendorf, 2002, p. 73) [Notre traduction]

> Un échantillon est dit représentatif s'il possède la même « structure » que la population de référence. Cela signifie que les différents sous-groupes qui composent cet échantillon doivent représenter une part identique à la part qu'ils représentent dans la population [...] La notion de représentativité n'a pas de sens si on ne précise pas la population à laquelle l'échantillon se réfère et les critères ou variables dont les distributions sont respectées. Un échantillon n'est pas « représentatif », il peut simplement être « représentatif d'une population au sens d'une série de critères ». Quand on entend ou qu'on dit qu'un échantillon est représentatif, il faut se demander : 1°) représentatif de quoi (de quelle population) ? 2°) au sens de quels critères (quels sont les caractères ou les variables dont les distributions sont respectées) ? (Martin, 2009, p. 23)

Un nouveau concept émerge du dernier extrait, soit celui de population. La *population* est l'ensemble des occurrences d'un phénomène ou des sources d'informations pertinentes pour son étude. Dans ce contexte, un corpus est représentatif en relation avec la population de référence et il peut être assimilé au

corpus de référence défini par Rastier. Le concept d'échantillon, qui est synonyme de celui de corpus d'étude, prend forme à partir du concept de population :

> Un échantillon est représentatif d'une population si son étude conduit à des conclusions similaires à celles auxquelles on aboutirait en étudiant l'ensemble de la population (Krippendorff, 2004, p. 112) [Notre traduction]

> *Règle de représentativité* : On peut, lorsque le matériel s'y prête, effectuer l'analyse sur échantillon. L'échantillonnage est dit rigoureux si l'échantillon est une partie représentative de l'univers de départ. Dans ce cas, les résultats obtenus sur échantillon seront généralisables à tout l'ensemble (Bardin, 1977, p. 127)

Un échantillon représentatif doit donc posséder les mêmes propriétés que la population et, en principe, permettre à chaque exemplaire du phénomène de faire partie du corpus d'étude.

En conclusion, la constitution du corpus est un processus interactif qui doit tenir compte de l'application de différents critères et qui nécessite un échantillonnage afin que le corpus d'étude soit le plus représentatif possible du phénomène à l'étude. L'équilibre entre les différents critères dépend des objectifs de la recherche et de la discipline dans laquelle elle s'inscrit.

3.3 Prétraitement d'un corpus

Une fois le corpus créé, il doit être préparé pour le traitement informatique et statistique (Meyer *et al.*, 2008; Vijayarani *et al.*, 2015). La phase de prétraitement est une étape cruciale dans une chaîne de traitement pour l'analyse de texte assistée par ordinateur. Pour certains, le prétraitement implique un certain nombre d'opérations qui ont pour but de « nettoyer les textes et de réduire le phénomène du bruit » (Lebart et Salem, 1994b, p. 116). Le bruit, qui a été défini dans la section sur le critère de

pertinence, est le phénomène qui désigne les documents d'un corpus non pertinents pour l'analyse. Toutefois, dans le contexte du prétraitement, ce phénomène assume des contours plus larges, car il désigne aussi un certain nombre d'éléments à « nettoyer » à l'intérieur du texte. Un exemple de cette opération de « nettoyage » des textes est l'effacement des *mots fonctionnels*, c'est-à-dire les signes graphiques qui représentent des déterminants, des connecteurs ou tout autre mot qui le chercheur ne retient pas comme pertinent (Savoy, 1999). Un autre exemple est la racinisation des mots (Bullinaria et Levy, 2007; Porter, 2001), qui permet de réduire les différentes formes graphiques d'un mot dans une forme plus standard ou normalisée.

Dans la présentation de la méthode, nous omettrons de parler de l'étape d'obtention du *format numérique du texte*. Ainsi, nous présupposons que le texte soit déjà encodé dans un format qui est traitable par un ordinateur, par exemple le format ANSI ou le format UTF-8. Ces deux formats sont des typologies d'encodage du texte, qui transforment un caractère de la langue dans une séquence de bits. Si le corpus est constitué d'images en format PDF ou de véritables pages de papier, une étape de *numérisation* et ensuite de *reconnaissance optique de caractères* doit être accomplie en premier lieu. Les systèmes de reconnaissance des signes graphiques changent selon le système d'écriture. Différents outils ont été développés pour les systèmes logographiques, les systèmes syllabiques ou les systèmes alphabétiques. Une fois que le texte numérique a été obtenu et que les caractères qui le composent ont été reconnus et encodés dans un des formats traitables par un ordinateur, le chercheur passera aux étapes de prétraitement.

Les opérations décrites dans les prochains sections supposent aussi que le corpus ne soit pas un corpus de documents tirés de la langue parlée, mais de la *langue écrite*. En effet, les outils de prétraitement pour des transcriptions automatiques ou non de la

langue parlée, sont différents et requièrent des argumentations spécifiques au cas du langage oral, ce qui n'est pas pertinent dans le cadre de cette thèse.

La phase de prétraitement d'un corpus est composée de deux sous-phases. La première est l'*annotation automatique* et elle est composée de plusieurs opérations. La deuxième est principalement constituée d'une opération de *transformation vectorielle* des textes, c'est-à-dire une conversion des textes dans un modèle mathématique et computable, ce qui permettra d'utiliser des algorithmes et des computations pour effectuer des analyses dans le corpus.

3.3.1 Annotation automatique des textes

En général, la première opération du prétraitement consiste à effectuer un ensemble de procédures d'annotation automatique qui permettent d'ajouter aux textes des informations supplémentaires sur leur nature. Ces informations peuvent ensuite être utilisées pour « nettoyer » les textes avant l'opération de transformation dans un modèle mathématique. Les principales procédures sont les suivantes. La *tokenisation*, c'est-à-dire l'identification de chaque signe graphique constituant une unité lexicale. L'*étiquetage automatique des parties du discours* (en anglais, *pos-tagging*), soit une analyse morpho-syntaxique du texte. La *lemmatisation*, qui permet de réduire les différentes formes graphiques des mots à des formes normalisées, c'est-à-dire les lemmes, cette opération peut être remplacée par la *racinisation*, dans laquelle la forme normalisée des mots est appelée la racine. L'*identification des mots fonctionnels* (en anglais, *stopword*), qui sont des catégories de mots (ex. connecteurs, conjonctions, etc.) qui ne sont pas pertinents pour l'analyse. Enfin, la *segmentation par phrases* des textes, ce qui implique d'identifier les délimiteurs de phrase à l'intérieur des textes. Toutes ces sous-opérations sont exécutées au moyen des outils développés dans le cadre des recherches en traitement automatique du langage naturel, discipline qui est au cœur de la présente chaîne de traitement.

Les opérations sont présentées de manière linéaire pour des raisons de simplicité et de clarté de l'argumentation. En réalité, le prétraitement est une phase non linéaire et complexe, où les différentes opérations se retrouvent imbriquées les unes dans les autres et où le résultat d'une opération dépend des résultats d'une autre opération. Par exemple, la segmentation par phrases est souvent accomplie parallèlement à l'analyse des parties du discours et elle ne peut pas être accomplie sans cette analyse.

3.3.1.1 Tokenisation

Le mot *tokenisation* dérive du mot anglais *token* qui désigne « l'occurrence d'un signe » (Rey-Debove, 1979, p. 148). Peirce fut le premier à avoir introduit le concept de token dans un contexte de théorisation sémiotique. Ce concept est indissociablement lié à celui de *type* et il est possible de considérer un token comme une *instance* du type :

> In order that a Type may be used, it has to be embodied in a Token which shall be a sign of the Type, and thereby of the object the Type signifies. I propose to call such a Token of a Type an **Instance** of the Type. Thus, there may be twenty Instances of the Type "the" on a page. (Peirce, 1994, p. CP 4.537)

Tel qu'indiqué dans cet extrait, le type est le signe dans sa forme générale alors que le token est le signe dans sa forme *instanciée*, qui est communément appelé *occurrence*. La réflexion de Peirce élabore une relation du type et du token avec sa théorie du signe et, en particulier, avec les concepts de *légisigne* et de *sinsigne*. D'un côté, le sinsigne donne au token la matérialité dont il a besoin car il est un objet qui existe et qui est perceptible. De l'autre, le légisigne définit le type dans un contexte plus large, soit la définition des signes au moyen de la loi humaine ou des conventions. Le légisigne est un type général qui est signifiant si et seulement si il « s'incarne » dans un token.

> Thus, the word "the" will usually occur from fifteen to twenty-five times on a page. It is in all these occurrences one and the same word, the same legisign. Each single instance of it is a Replica. The Replica is a Sinsign. Thus, every Legisign requires Sinsigns. (Peirce, 1994, p. CP 2.246)

Peirce souligne ici comment chaque légisigne (type) a besoin d'un sinsigne (token) pour exister et se matérialiser. Un texte est donc composé seulement de sinsignes, c'est-à-dire de tokens. L'entité du type est une convention et elle regroupe toutes les formes que son instanciation peut prendre.

Dans un contexte informatique, la *tokenisation* est une procédure de *reconnaissance des signes graphiques* qui sont susceptibles de constituer un token (sinsigne). En linguistique et en lexicographie, cette procédure n'est généralement pas détaillée car elle ne représente pas un enjeu majeur. Au contraire, en traitement automatique des langues, la tokenisation est un problème non trivial, car c'est un « lexicographe computationnel » qui doit effectuer les choix de tokenisation et non un humain. Par nature, le langage est un système ambigu auquel les êtres humains sont habitués, mais un lexicographe computationel devra adopter des stratégies de désambiguïsation définies en amont et détaillées dans un programme informatique. Cette dichotomie entre l'expertise humaine et le transfert des connaissances dans un programme informatique est souvent présente lors du passage de la linguistique ou de la sémiotique aux systèmes informatiques de manipulation des signes. En effet, cette première opération de prétraitement peut comporter différents niveaux de complexité, selon le degré de justesse que le chercheur veut obtenir lors de la reconnaissance des unités lexicales.

3.3.1.2 Étiquetage automatique des parties du discours

L'étiquetage des parties du discours consiste à identifier l'« emploi grammatical » des tokens, c'est-à-dire « la classe syntaxique par laquelle le mot s'intègre à la phrase »

(Lehmann et Martin-Berthet, 2014, p. 15). En français, chaque mot (token) peut appartenir à une des *classes de mots* ou *parties du discours* suivantes: nom, verbe, adjectif, déterminant, pronom, adverbe, préposition et conjonction. Les parties du discours sont alors les constituants de la phrase et sont appelées catégories morphosyntaxiques. L'étiquetage des parties du discours peut se complexifier jusqu'à une véritable analyse morphosyntaxique, où on prend en compte aussi d'autres catégories qui définissent chaque token, comme le genre, le nombre, le mode du verbe ou le temps, les interjections, etc.

Cette opération est exécutable une fois que le processus de tokenisation est terminé. L'étiquetage des parties du discours peut être utilisé pour plusieurs tâches, qu'elles soient simples ou plus complexes. Dans la traduction automatique de l'anglais au français par exemple, il est fondamental de connaître la catégorie grammaticale d'un mot, car cela permet de désambiguïser des cas comme « record », qui peut être traduit en français de deux façons « disque » ou « enregistrer » s'il s'agit d'un verbe. Dans des tâches plus simples, comme pour des étapes de filtrage des mots fonctionnels, il est utile de connaître quels sont les déterminants, par exemple des mots comme « de », « il » « la » (en français). La tâche planifiée par le chercheur et pour laquelle l'étiquetage sera utilisé détermine aussi le niveau de justesse de l'étiquetage automatique. Le plus souvent, l'étiquetage ne prend pas en compte tous les éléments pertinents d'une analyse morpho-syntaxique, tels le mode et le temps du verbe. Le meilleur étiquetage est celui qui offre un bon compromis entre la justesse ou le détail de l'information et l'utilité des informations pour la tâche à accomplir. Pour chaque chaîne de traitement, il faut donc estimer les ressources nécessaires pour développer des méthodes plus sophistiquées d'étiquetage et évaluer si cela en vaut la peine.

Il existe deux phénomènes qui font du processus d'étiquetage morphosyntaxique une opération complexe (Dale *et al.*, 2000). Le premier est le fait que chaque mot peut

avoir différentes étiquettes morphosyntaxiques, ce qui implique qu'il n'est pas possible d'utiliser un dictionnaire de mots auxquels on assigne une étiquette. Un mot peut être un nom ou un verbe, comme c'est le cas du mot « record » en anglais et seul le contexte dans lequel le mot est inséré permet d'identifier la bonne étiquette. Le deuxième phénomène est la présence des mots qui sont spécifiques à un domaine du savoir et donc très rares. Dans ce cas, il est encore plus difficile de fournir les connaissances adéquates aux algorithmes d'étiquetage.

Une façon de résoudre ce type de problème est d'inférer l'étiquette morphosyntaxique à partir du contexte dans lequel le mot apparaît et à partir des informations morphologiques de chaque mot. Le contexte dans lequel chaque mot est inséré apporte des informations pertinentes à l'identification des parties du discours de chaque mot. En anglais par exemple, il est très rare que des phrases commencent par l'objet et plus fréquent qu'elles débutent par un nom (ou un pronom) et suivis par un verbe. Les informations morphologiques du mot, surtout pour les langues flexionnelles, sont également importantes. En français par exemple, le suffixe « -ment » permet d'identifier un adverbe.

Il existe plusieurs méthodes d'étiquetage des parties du discours et il est possible de les regrouper en deux grandes familles : les approches supervisées et les approches non supervisées. Les approches supervisées sont basées sur des ressources préalablement construites par des annotateurs humains. Le *Penn Treebank* (Marcus *et al.*, 1993) est le corpus le plus connu pour la langue anglaise. Il contient des millions de mots qui ont été étiquetés manuellement. Le *British National Corpus*, qui contient du texte en anglais britannique, et le *French Treebank* pour la langue française constituent d'autres ressources sur lesquelles peuvent s'appuyer les approches supervisées pour l'étiquetage automatique de nouveaux textes.

Les *processus de Markov* ont été largement utilisés dès les années 1980 pour l'annotation automatique des parties du discours dans un contexte supervisé (Church, 1989; Derouault et Jelinek, 1984).

3.3.1.3 Lemmatisation

La lemmatisation est l'opération qui identifie le *morphème porteur du signifié* pour les unités lexicales des langues flexionnelles. Les tokens qui ont été identifiés dans l'étape de tokenisation sont des occurrences de différents mots. Chacun de ces tokens est porteur d'un signifié. Du point de vue sémantique, ce ne sont pas tous les caractères qui composent l'occurrence d'un mot qui sont pertinents. Le mot « mangeait » par exemple, est composé de deux morphèmes, le radical qui renvoie à l'unité lexicale abstraite d'un mot et qui est le signifiant qui véhicule le signifié du mot (Polguère, 2016) et les affixes, qui sont utilisés pour conférer un rôle syntaxique au mot.

La lemmatisation peut être remplacée à l'occasion par la *racinisation*. Ce processus vise les mêmes objectifs que la lemmatisation mais le type ne correspond pas au lemme mais à la racine du mot. Il existe différents algorithmes pour optimiser ce processus. Un des plus répandus est l'*algorithme de Porter* qui analyse la suite de caractères du mot et, sur la base d'une série de règles, identifie les suffixes et, par conséquent, les racines. Ce processus est appelé également désuffixation. Par exemple, en français, il existe une règle d'identification des suffixes d'adverbes comme « quotidiennement », « grossièrement », « facilement », qui consiste dans le suffixe «-ment».

Le but principal de ces deux processus substituables est de *mettre en valeur les aspects sémantiques* du texte, en simplifiant ainsi les formes et les flexions que chaque mot peut prendre.

3.3.1.4 Identification des mots vides

Les *mots vides* ou *mots fonctionnels* sont des mots « qui n'ont pas de sens par eux-mêmes » (Lehmann et Martin-Berthet, 2014, p. 217). Ils s'opposent aux *mots pleins*, qui véhiculent du « sens ». En général, les mots vides sont des mots ayant un rôle syntaxique et fonctionnel et non sémantique, par exemple les mots qui appartiennent à des parties du discours, comme les prépositions, les conjonctions, les déterminants, les pronoms. Dans ce sens, les mots pleins font partie de catégories comme le nom, l'adjectif, le verbe ou l'adverbe.

Cette opération d'identification des mots vides joue le même rôle que la lemmatisation, c'est-à-dire *mettre en valeur les aspects sémantiques plutôt que syntaxiques*. L'idée est donc de repérer les mots vides et de les éliminer afin d'éviter d'introduire du bruit dans l'analyse. En d'autres termes, les mots fonctionnels attirent l'attention vers des éléments peu pertinents du point de vue sémantique.

L'identification des mots vides est exécutée à l'aide d'un vocabulaire ou d'une liste de mots. Ces listes peuvent être plus ou moins « agressives ». Des fois, elles peuvent aussi contenir les auxiliaires (être, avoir), les modaux (pouvoir, vouloir, savoir, devoir), les adjectifs possessifs (leur, notre, etc.) ou des mots pleins qui sont considérés peu pertinents par le chercheur. Il n'est pas rare d'appliquer des listes de mots vides très agressives dans le but de faire ressortir davantage le « cœur sémantique » d'un texte.

L'impact positif de l'élimination de ce type de mots sur les analyses orientées sur les aspects sémantiques a été amplement démontré dans le domaine de la recherche d'information où le bruit que ces mots introduisent est déterminant pour les performances d'un moteur de recherche. Dans ce contexte, le cadre d'analyse est également de type sémantique, car le but est de trouver les documents qui répondent

le mieux à une requête de type sémantique et les mots vides produisent un bruit qui réduit la qualité des résultats.

3.3.1.5 Segmentation en phrases ou en d'autres unités syntagmatiques élémentaires

La segmentation en unités syntagmatiques élémentaires coïncide généralement avec la segmentation en phrases, c'est-à-dire l'identification des phrases qui composent un texte. Dans les langues flexionnelles les plus répandues du monde occidental, la fin d'une phrase est explicitée par un signe de ponctuation. Une des premières sous-opérations de la segmentation en phrases est d'identifier les marqueurs de fin de phrase. Pour le français par exemple, le délimiteur de phrase le plus répandu est le point. Toutefois, les règles d'utilisation de la ponctuation ne sont pas toujours définies de manière univoque. Pour exécuter de manière efficace la segmentation en phrases, il est nécessaire de connaître les règles de ponctuation de la langue traitée. Ensuite, le défi principal consiste à établir des règles de désambiguïsation des occurrences de la ponctuation. Identifier les différents rôles qu'un point peut assumer est une opération parallèle au processus de tokenisation. Par exemple, l'identification des abréviations permet en même temps d'identifier des occurrences du point dont le rôle n'est pas celui de délimiteur de phrase.

3.3.2 Représentation vectorielle du texte

Cette étape consiste à transformer le corpus dans une matrice *U*. Cette modélisation est représentée par la figure 3.1. Chaque ligne de cette matrice modélise un article du corpus sous la forme d'un vecteur $\overrightarrow{S_i} = (V_{i1}, \dots, V_{ij})$ dans lequel V_{ij} correspond à la *valeur de pondération* du $j^{ème}$ mot dans le $i^{ème}$ segment (cfr. formule mathématique 2.5 dans la section 2.3.1.2).

Ce modèle est appelé *modèle sémantique vectoriel* et il est construit au moyen de trois éléments, soit les paramètres *d* et *v* et la fonction de pondération *f*. Le paramètre

document *D* correspond au corpus et est composé des documents *d*. Les documents *d* doivent être interprétés comme les unités syntagmatiques élémentaires que le chercheur a retenues. Pour simplifier, nous appellerons *d* le document. Le paramètre vocabulaire *V* qui est composé par les caractéristiques des documents que le chercheur a considéré comme pertinentes pour sa recherche. Généralement, le vocabulaire *V* correspond aux lemmes. Enfin, la fonction de pondération *f*, évalue le poids de chaque caractéristique pour chaque document.

Les postulats, les hypothèses et toute autre implication sémiotique du modèle vectoriel ont été décrits dans le cadre théorique. Dans les prochains sections, nous détaillerons le modèle du point de vue technique, en décrivant davantage les trois paramètres qui le composent et en soulignant les aspects les plus pertinents pour le présent travail.

3.3.2.1 Le paramètre *D* : Les documents

Le corpus *D* est composé des éléments qui, dans la littérature, sont appelés *documents*. En réalité, les documents correspondent aux unités syntagmatiques élémentaires que le chercheur a choisies d'utiliser. Les documents peuvent être de différente nature en fonction des objectifs de la recherche. Le document *d* peut être une proposition, une phrase, un groupe de phrases ou de propositions, un paragraphe, une section ou une concordance. Le choix est généralement déterminé par l'objectif de la recherche, mais il n'est pas rare d'identifier la meilleure segmentation par une enquête empirique. En effet, il est fréquent de choisir l'unité syntagmatique élémentaire à utiliser après avoir testé plusieurs typologies de documents, ce qui implique que les différentes typologies de documents seront utilisées pour l'analyse et, ensuite, le chercheur choisira la typologie de documents qui répond le mieux à ses attentes.

Généralement, un document plus long, par exemple le paragraphe ou une section, implique de considérer l'hypothèse distributionnelle et le principe de cooccurrence de mots dans un contexte plus grand et donc avec plus de mots cooccurrents. Le mot « étudiant » par exemple, peut être cooccurrent avec le mot « grève » et ceci, un certain nombre de fois. Si l'unité syntagmatique élémentaire est l'article au complet, le mot « étudiant » sera cooccurrent avec le mot « grève », mais aussi avec plusieurs autres mots. La distinction de l'usage du mot « étudiant » du point de vue de la sémantique distributionnelle est effectuée en considérant toutes les cooccurrences du mot. En d'autres termes, plus les unités syntagmatiques élémentaires sont larges, plus nombreux sont les mots impliqués dans la détermination de l'usage d'un mot.

3.3.2.2 Le paramètre V : Les caractéristiques

Le choix du type de document détermine le cadre d'évaluation de l'usage du mot. Plus un contexte est large, plus les mots sont considérés comme cooccurrents. Toutefois, le mot tel quel n'est presque jamais considéré comme une caractéristique pertinente dans un modèle sémantique vectoriel. Tel que décrit auparavant, le lemme est le morphème le plus pertinent dans une analyse orientée sur les aspects sémantiques du langage. Les lemmes ou les racines sont une caractéristique du document plus intéressante que le mot simple. Deux documents seront considérés similaires s'ils partagent un certain nombre de caractéristiques et donc, s'ils partagent les mêmes lemmes. Toutefois, il existe d'autres caractéristiques d'un document qui peuvent être pertinentes. Par exemple, il est possible d'utiliser les parties du discours comme étant une caractéristique pertinente pour la description du segment. La suite de caractères suivante « Étudiant_NOM » peut être une caractéristique d'un document qui désigne l'occurrence du nom « étudiant ». De la même manière, la suite de caractères « Étudiant_ADJ » désigne l'occurrence de l'adjectif « étudiant » et ainsi de suite. En utilisant un analyseur syntaxique plus complexe, il est aussi possible d'utiliser les catégories grammaticales du sujet, du verbe et du complément. En

principe, tout ce qui peut décrire le contenu d'une unité syntagmatique est susceptible d'être considéré comme une caractéristique pertinente à ajouter au modèle sémantique vectoriel.

Le modèle sémantique vectoriel est généralement bâti selon une approche qu'on appelle *sac de mots.* Le modèle sémantique à sac de mots consiste à créer une matrice *U*, ayant *D* et *V* comme paramètres, dont *V* est la liste de mots qui apparaissent dans *D*. Généralement, les mots correspondent aux lemmes et aux racines. Dans ce contexte, chaque document est décrit sur la base des lemmes qu'il contient. Cette même approche peut être appliquée à un certain nombre de caractéristiques, par exemple à la combinaison lemme/partie du discours. Le principe du modèle à sac de mots consiste à assumer qu'un document peut être décrit exclusivement par les éléments qu'il contient et ceci, sans tenir compte de leur ordre d'apparition. Le document est un contenant dans lequel les éléments ne sont pas ordonnés, comme un sac de jetons numérotés du bingo. Dans ce contexte, ce qui est retenu comme important est le contenu du sac et non son ordre.

L'approche appelée à *n-gramme*, au contraire du modèle à sac de mots, assume que l'ordre d'apparition des caractéristiques est pertinent pour la description du document. Le principe est similaire au modèle de langage probabiliste décrit dans le paragraphe sur l'annotation des parties du discours. Cette approche décrit les documents en considérant toutes les suites de caractéristiques possibles. La grandeur de la suite est déterminée par le type de modèle choisi, par exemple un modèle bigramme, qui correspond à une suite de deux éléments ou un modèle trigramme, qui correspond à une suite de trois éléments et ainsi de suite. Par exemple, dans un modèle bigramme, la phrase « La linguistique est enseignée à l'université » sera décrite par une série de suites de bigramme : « la linguistique », « linguistique est », « est enseignée », « enseignée à », « à l'université ». Si nous avions effectué une lemmatisation, la

phrase aurait été transformée ainsi, « le linguistique être enseigner à le université » et les séquences de bigramme seraient les suivantes : « le linguistique », « linguistique être », « être enseigner », « enseigner à », « à le », « le université ». Si nous avions aussi supprimé les mots vides, la phrase serait alors transformée ainsi : « linguistique enseigner université » et les suites de bigramme seraient les suivantes : « linguistique enseigner », « enseigner université ». Dans ces genres de modèle, le début de la phrase peut aussi être considéré comme étant une caractéristique, ce qui implique que la séquence de bigramme change ainsi : « DÉBUT linguistique », « linguistique enseigner », « enseigner université ». Utiliser un modèle n-gramme dans le contexte vectoriel est complexe et n'apporte pas toujours des avantages. Pour cette raison, le modèle à sac de mots est le plus souvent choisi.

3.3.2.3 La fonction de pondération

Après avoir identifié les documents du corpus et leurs caractéristiques il est possible de construire le véritable modèle sémantique vectoriel, lequel se concrétise dans la construction d'une matrice. Cette matrice est la représentation mathématique des documents. Le passage d'une liste de caractéristiques par document à la matrice est rendu possible par la fonction de *pondération*, à travers laquelle on attribue une valeur numérique à chaque cellule de la matrice, c'est-à-dire à chaque caractéristique de chaque document. En d'autres termes, la représentation vectorielle d'un texte est une matrice U avec les documents d_i comme lignes, qui font partie du corpus D. Chaque document est décrit par un certain nombre de caractéristiques v_i, qui font partie du vocabulaire V. La matrice U a été montrée avec la formule 2.5 au paragraphe 2.3.1.2.

La valeur de chaque cellule correspond au « poids » d'une caractéristique dans un contexte. Le « poids » est établi par la fonction de pondération. La fonction la plus

simple correspond au décompte des fréquences observées de chaque caractéristique pour chaque document. Par exemple, la phrase suivante (énoncé 1)

Énoncé 1 - La linguistique est enseignée à l'université

est ainsi représentée dans la ligne « Doc 1 » de la matrice *U*, qui a été transformée pour représenter l'énoncé 1. La matrice est présentée sous-forme de tableau (tableau 3.1).

Tableau 3.1 Représentation vectorielle de l'énoncé 1

Doc/ Lemmes	***aller***	***apprendr***	***classe***	***enseigner***	***étudiant***	***linguistiq***	***philosoph***	***professeu***	***salle***	***sémiotiqu***	***université***	***n lemme***
Doc 1	0	0	0	1	0	1	0	0	0	0	1	…
Doc 2	1	1	1	0	1	0	1	0	0	0	0	…
n doc	…	…	…	…	…	…	…	…	…	…	…	…

Les cellules correspondant aux colonnes v_i du vocabulaire *V* « linguistique », » enseigner », « université » et à la ligne d_i du corpus *D* « Doc 1 », ont une valeur égale à 1, car ces lemmes sont observés une seule fois dans le document « Doc 1 ». Les autres cellules de la ligne sont remplies par des zéros. Le « Doc 2 », au contraire, représente une phrase dans laquelle les mots « aller », « apprendre », « classe », « étudiant » et « philosophie » apparaissent qu'une fois. Si on voulait déduire la phrase originale, on pourrait imaginer qu'elle soit « l'étudiant va en classe pour apprendre la philosophie », ou « l'étudiant de la classe va apprendre la philosophie ».

Il existe différentes méthodes de pondération (Lan *et al.*, 2009; Salton, 1971; Salton et Buckley, 1988) et chacune d'elles évalue différemment les poids des

caractéristiques dans un document. Les principales fonctions sont : la fonction binaire, la fonction de la fréquence et la fonction *Tf-Idf*, de l'anglais *Term frequency-Inverse document frequency*. Imaginons que les caractéristiques correspondent exclusivement aux lemmes. La fonction binaire, comme son nom l'indique, assigne seulement des valeurs binaires (1,0), signalant ainsi la présence ou l'absence d'un lemme. Quant à elle, la fonction de la fréquence assigne à chaque lemme sa fréquence pour chaque document. Si un lemme est répété plus d'une fois, la valeur de la cellule ne sera pas un, mais un chiffre correspondant au nombre d'apparitions du lemme dans le document. Ces deux fonctions de pondération ont des limites. La première ne fournit pas assez d'information sur la valeur du lemme dans un document. En effet, il suffit d'avoir une seule occurrence d'un lemme pour qu'il soit signalé. Chaque lemme a donc la même valeur qu'il apparaisse une ou plusieurs fois. Par ailleurs, la méthode de la fréquence du mot, exacerbe la distribution lexicale décrite par la loi de Zipf (cfr. 2.1.2), ce qui implique que seule une petite portion du vocabulaire V correspondant aux lemmes ayant la fréquence la plus grande sera déterminante pour le calcul des similarités entre documents.

La dernière fonction, le *Tf-Idf*, est la plus utilisée par les différentes communautés, et il en existe différentes versions. Sa puissance repose sur la présupposition que le poids d'une caractéristique (ex. un lemme) est *proportionnel à sa fréquence* dans chaque document et *inversement proportionnel à sa fréquence documentaire*. Pour exécuter cette fonction, on calcule d'abord la fréquence documentaire, c'est-à-dire le nombre de documents dans lesquels les caractéristiques apparaissent. Par exemple, si le lemme « étudiant » apparaît dans 232 documents, alors 232 sera la fréquence documentaire du lemme « étudiant ». La fréquence documentaire est ensuite convertie en fréquence documentaire inverse comme suit :

$$Idf = \frac{n}{f(d_j)} \tag{3.1}$$

où n est égal au nombre de document de D et $f(d_j)$ est la fréquence documentaire du lemme j. À laquelle on ajoute la normalisation logarithmique $\log idf + 1$:

$$Idf = \log\frac{n}{f(d_j)} + 1 \tag{3.2}$$

La composante inverse de la fréquence documentaire est donc représentée par la fraction entre le nombre total n de documents du corpus D et la fréquence documentaire $f(d_j)$ du lemme. La fréquence documentaire inverse est ensuite multipliée par la fréquence du lemme, ce qui donne la formule suivante :

$$TfIdf = f(j) * \log\frac{n}{f(d_j)} + 1 \tag{3.3}$$

Ainsi construite, la valeur de la fonction Tf-Idf est plus élevée lors de l'augmentation de la fréquence du terme et elle est moins élevée lors de l'augmentation du nombre de documents qui contiennent le lemme. Donc, si un lemme est très fréquent et très répandu dans le corpus, il obtient une valeur Tf-Idf plus basse, tandis qu'un lemme peu fréquent en général, mais très fréquent dans un groupe de documents, obtient une valeur Tf-Idf élevée. L'objectif de cette fonction est d'identifier les lemmes les plus discriminants.

3.4 Filtrage

Le filtrage est une opération générale qui s'effectue tant au niveau des caractéristiques qui décrivent les documents, c'est-à-dire les colonnes de la matrice, qu'au niveau des documents. Son objectif principal est de réduire le *bruit* et

d'augmenter le « pouvoir prédictif » du modèle. En d'autres termes, le filtrage réduit les dimensions et les contextes peu pertinents, afin d'accroître les possibilités d'identifier des caractéristiques signifiantes dans le corpus. La notion de bruit a été abordée à plusieurs reprises au cours du présent chapitre et elle demeure centrale pour l'opération de filtrage. Sans entrer dans les détails mathématiques, la notion de bruit peut être appréhendée par le schéma d'un système de communication tel qu'élaboré par Shannon et Weaver.

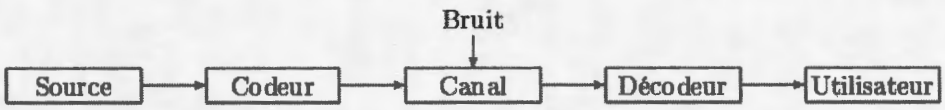


Figure 3.1 Modèle de communication

Dans ce schéma, est illustrée la manière dont un message est transféré d'une *source* à l'utilisateur, en passant par un canal. En reprenant la terminologie de Jakobson un message se transfère d'un destinateur au destinataire par le biais d'un contexte, d'un code et d'un contact.

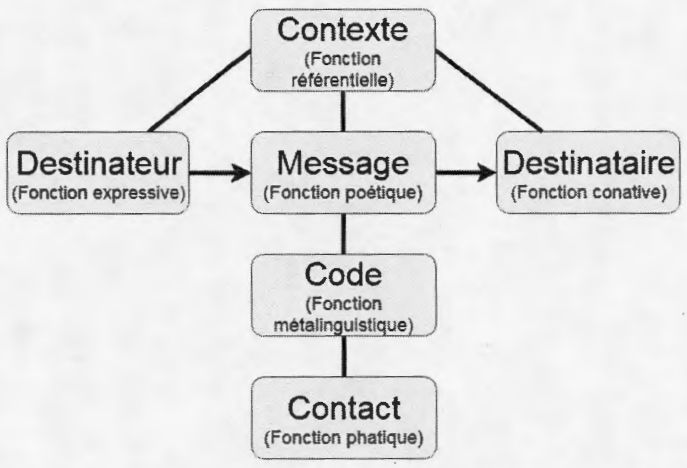


Figure 3.2 Modèle de la communication et ses six fonctions (Jakobson)

Shannon prévoit dans son modèle l'impact que le bruit a sur le transfert du message de la source vers le destinataire et il démontre comment il est possible de calculer la probabilité que l'information soit transférée, connaissant le bruit potentiel. Le bruit est tout obstacle qui interfère dans le transfert du message. Il est possible d'élargir la définition : le bruit correspond à l'écart entre l'*intention* de l'auteur (destinateur) et le *marquage* de l'artefact sémiotique produit (message), auquel il s'ajoute l'écart entre le marquage et l'*interprétation* du récepteur (destinataire). Cette définition est aussi valable en sémiotique. Chaque communication humaine, peu importe le canal, comporte toujours un potentiel d'interférence. Eco a affirmé que la communication humaine n'existe pas et qu'elle est toujours un malentendu (Eco, 1979).

En sémiotique, le bruit peut être de différentes natures selon qu'il interfère directement avec le message, avec le code, avec le contexte ou avec le contact. La notion de bruit qui est utilisée dans le contexte du traitement automatique du texte peut également être de différentes natures. En ce qui concerne l'opération de filtrage, la nature du bruit est surtout liée au message. Chaque message peut avoir une portion d'information plus pertinente que d'autres. Certaines portions du message peuvent nuire au transfert de l'information et, ainsi, nuire à l'interprétation du message. La communication peut échouer si les informations contenues dans le message « distraient » le destinataire de l'information la plus pertinente.

L'opération de filtrage tend à réduire le bruit du message en agissant au niveau des mots et des documents. Réduire les signes graphiques qui composent les messages afin de ne conserver que les plus pertinents est un exemple de filtrage au niveau des mots. Ce même principe s'applique aussi au niveau du document, surtout lors de l'analyse d'un corpus puisque cette opération augmente ainsi l'indice de pertinence du corpus.

3.4.1 Mots

La sélection des caractéristiques des données non catégorisées peut toutefois être exécutée par une opération de filtrage moins complexe. Dans ce travail, le filtrage des caractéristiques est appréhendé selon deux types d'approches : l'utilisation de classes de caractéristiques construites à priori et le lissage.

La première approche de filtrage des mots est celle qui se base sur des classes préalablement constituées. Par exemple, les mots vides constituent une classe de mots à filtrer préventivement construite. En général, les classes de mots à filtrer sont déterminées par deux typologies d'objets : de véritables listes de mots, comme pour les mots vides, et l'annotation des parties du discours qui permet de regrouper les mots selon leur catégorie morphosyntaxique. Dans ce contexte, il est donc possible de filtrer les mots qui font partie de classes comme les pronoms, les connecteurs, les chiffres, etc. Ce type de filtrage est donc appliqué avant que les textes ne soient transformés dans une matrice.

L'approche par lissage est basée sur les fréquences d'apparition des mots et elle est appliquée après que les textes aient été transformés en matrice. Il existe différentes méthodes pour filtrer les mots les moins significatifs de la matrice. Afin de simplifier, nous soulignons la méthode *basée sur la fréquence d'apparition* et la méthode *basée sur la variance*. La première établit un seuil qui est appliqué à la fréquence d'apparition du terme ou à la fréquence documentaire. Par exemple, il est possible de filtrer les mots qui n'apparaissent pas plus qu'un certain seuil (ex. 10, 15, 20, etc.) ou qui ne dépassent pas une certaine fréquence documentaire. Dans cette approche, le seuil appliqué à la fréquence documentaire est davantage utilisé car les résultats sont plus facilement interprétables. En général, ces méthodes tendent à filtrer les mots qui correspondent aux hapax (Lehmann et Martin-Berthet, 2014), c'est-à-dire aux mots très rares ou avec une seule occurrence. La deuxième méthode établit un seuil qui

s'applique à la variance. Le but est d'éliminer les caractéristiques qui ont une variance basse. Plus la variance d'une caractéristique s'approche de zéro, moins elle apporte d'information. Par exemple, si une caractéristique d'un ensemble de données a toujours la même valeur, sa variance est égale à zéro. Dans notre cas, si un mot a la même valeur Tf-Idf dans la majorité des articles, sa variance est forcément moins élevée et peut donc être retirée.

3.4.2 Documents

La réduction du bruit est également appliquée au niveau du document afin d'identifier les *données aberrantes*, c'est-à-dire les données non conformes aux autres ou éloignées d'un centre d'agrégation. Les données aberrantes constituent fréquemment des erreurs ou du bruit qu'il est préférable de filtrer. En général, les données aberrantes sont des observations inusuelles. Dans le cas du texte journalistique, la détection d'articles inusuels est une opération essentielle car elle permet de maintenir le corpus homogène et pertinent (Kannan *et al.*, 2017).

En statistiques, il existe une méthode très répandue pour identifier les données aberrantes qui est basée sur les *quartiles*. Quand on dispose d'une suite de valeurs numériques suivant une distribution gaussienne, les valeurs aberrantes sont les chiffres qui se situent à un saut égal à une fois et demie à la longueur de l'étendue interquartile (Bertrand R., 1986; Tukey, 1977). Toutefois, le problème de données aberrantes se complexifie dans un espace multidimensionnel. Dans cet espace, les données aberrantes ne correspondent pas à une valeur, mais à un vecteur multidimensionnel. En général, il est possible d'aborder ce problème dans un espace sémantique avec trois types d'approche : non-supervisée, supervisée et hybride.

Dans la plupart des cas, la détection de données aberrantes a été interprétée comme un problème de type non supervisé, car il est rare de connaître a priori les

caractéristiques des observations inusuelles. Dans ce contexte, l'approche principale s'appuie sur le concept de distance entre points. En simplifiant, il est possible de distinguer des méthodes basées sur la *détection de clusters*, de méthodes basées sur des *seuils de distance* et des méthodes basées *sur la densité des points*. Toutefois, cette distinction n'est pas nette car les méthodes par clusters sont généralement très proches de celles basées sur la densité ou sur la distance. Ceci est dû au fait que les notions de distance, de cluster et de densité sont étroitement liées et interdépendantes. La première classe de méthodes, soit celle basée sur la détection de clusters, exécute une analyse des données aberrantes parallèlement à un clustering. Si le clustering a comme but d'identifier les classes d'objets similaires (cfr. 2.3.2), l'analyse de données aberrantes vise à identifier les objets qui sont trop inusuels pour appartenir à une classe particulière. Les méthodes basées sur le seuil de distance établissent des heuristiques pour repérer les données aberrantes. Cette approche est basée sur le fait que les données aberrantes sont à une distance importante d'un ou plusieurs centres de gravité autour desquels la majorité de données se situent. Il est alors suffisant d'établir un seuil de distance et les centres de gravité à utiliser comme référence. Un exemple en est fourni par la figure 3.3 qui montre les résultats obtenus par un algorithme de clustering et un seuil de distance.

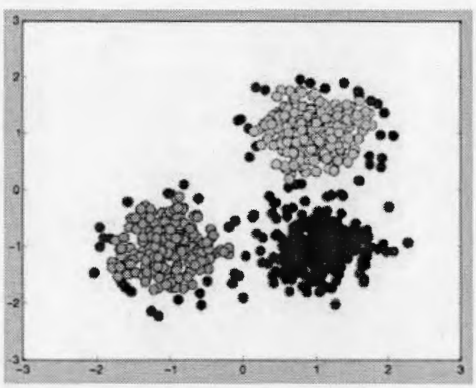


Figure 3.3 Détection de données aberrantes par clustering. Les points plus foncés sont loin des trois clusters et, ainsi, sont classifiés comme bruit.

Les méthodes basées sur la densité utilisent le nombre de points situés dans des régions spécifiques de l'espace. Il existe plusieurs algorithmes de clustering qui se basent sur la densité des points qui sont également adaptés à la détection des données aberrantes. Par exemple, *DBscan* (cfr. Anexe C) produit une classe de données aberrantes si l'algorithme n'est pas capable de les regrouper dans une région assez dense. Dans la figure 3.4, l'algorithme identifie une région assez dense dans un certain nombre de points proches (les « border », les points en forme de triangle) et élimine par la suite les points qui sont trop distants de cette région (le « noise », les points en forme carré) et qui constituent le bruit.

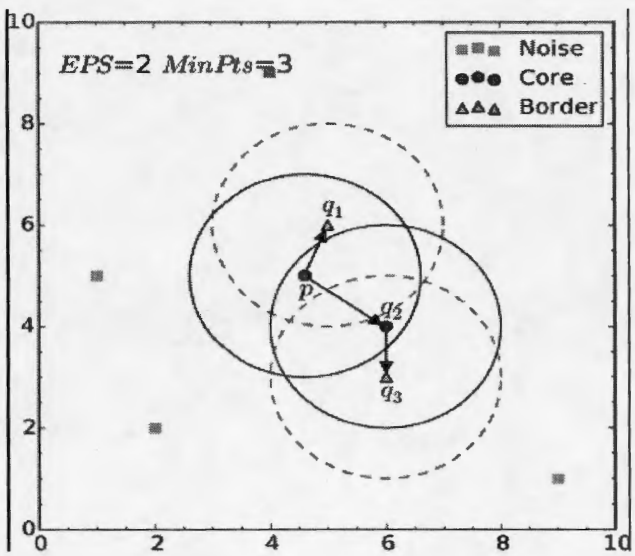


Figure 3.4 L'algorithme DBscan. *Eps* et *MinPts* sont les deux paramètres, respectivement l'un pour le seuil de distance et l'autre pour le nombre de points minimaux pour construire un cluster. Les cercles correspondent aux centroïdes des clusters, les triangles correspondent aux points constituant la limite du cluster et les carrés correspondent aux points classifiés comme bruit (cfr. Annexe C).

Lorsqu'il est possible de disposer d'exemples de données aberrantes qui peuvent être utilisés pour apprendre à un algorithme à les reconnaître, la détection des données aberrantes peut être transposée dans un contexte supervisé et peut donc être considérée comme un problème de classification (Aggarwal, 2016). Il s'agit toutefois

d'un contexte particulier de classification. En général, une classification pour la détection de données aberrantes est caractérisée par une série de problèmes, dont le plus important est le *déséquilibre de l'échantillon d'apprentissage* (en anglais, *class imbalance*). Puisque les données aberrantes sont des observations inusuelles dans un jeu de données, il est normal que la taille des deux classes, c'est-à-dire celle de données aberrantes et celle de données « normales », soit différente. Évidemment, il existe beaucoup plus de données normales que de données aberrantes. Une manière de résoudre ce problème est la *configuration des paramètres* des algorithmes pour les rendre plus « sensibles » à la différence de taille entre les deux classes. L'exécution d'un *re-échantillonnage* de données aberrantes, afin d'égaliser les proportions entre les deux classes constitue une autre solution intéressante. La contamination des données «normales» par des données aberrantes est un problème fréquent et important car il génère de la confusion dans l'apprentissage. Ce phénomène est dû au fait qu'il est difficile d'obtenir un échantillon d'apprentissage suffisamment exhaustif et contenant chaque exemplaire de données aberrantes. En principe, la possibilité de rencontrer des nouveaux cas de données aberrantes est toujours présente.

Les méthodes hybrides sont reconductibles dans les méthodes d'apprentissage automatique de type semi-supervisé. Ainsi, l'approche non-supervisée et l'approche supervisée sont combinées dans une même chaîne de traitement. Généralement, ces chaînes de traitement identifient d'abord un petit nombre d'exemplaires de documents faisant partie de la classe des données « normales » et un autre de la classe des données « aberrantes ». Cette première étape, au cours de laquelle l'intervention du chercheur et l'évaluation humaine sont occasionnellement prévues, prépare l'échantillon d'apprentissage pour l'étape supervisée de détection des données aberrantes (Aggarwal, 2016, p. 232).

L'utilisation des différentes méthodes dépend du type de données et des objectifs de la recherche. Dans le cas du présent travail et pour toute l'analyse de corpus, la détection de données aberrantes constitue une étape importante pour augmenter la pertinence et l'homogénéité du corpus puisque l'identification du bruit du corpus au niveau du document permet d'éliminer les documents non pertinents pour la recherche. Son application est donc très importante pour la définition d'un corpus de travail répondant à un maximum de critères de sa constitution.

3.5 Définition du corpus de travail

Cette étape de la méthode vise à identifier définitivement le corpus de travail, c'est-à-dire l'ensemble des documents sur lesquels les analyses seront effectuées. La majorité des étapes préalablement décrites préparent le terrain au traitement informatique. Ces opérations sont appliquées à plusieurs typologies de corpus textuel et cela, indépendamment des objectifs de la recherche. Au contraire, la définition du corpus de travail varie sensiblement selon la typologie de texte, la typologie d'analyse ou les objectifs de la recherche. Dans cette étape, le chercheur synthétise tous les résultats de la constitution du corpus, du prétraitement et du filtrage afin de déterminer de manière définitive l'ensemble de documents qui seront analysés.

Dans le cas de l'analyse d'un corpus comportant des textes journalistiques, certaines métadonnées peuvent être utilisées pour *créer des sous-corpus de travail*. L'organisation du corpus de travail en sous-corpus dépend de la question de recherche. Par exemple, si une analyse diachronique est prévue, la division du corpus de travail en périodes de temps est requise. Si l'objectif est de comparer le discours de différents journaux, il devient pertinent d'organiser le corpus de travail par sources d'information. Les sous-corpus de travail doivent donc être orientés. Il existe deux approches pour créer des sous-corpus de travail dans un contexte d'analyse du

discours de presse. La première exploite des *éléments extratextuels* récoltés lors de la constitution du corpus. En principe, chaque article journalistique peut contenir une série de métadonnées qui le décrivent, comme la date, l'auteur, le journal de publication, la catégorie du texte, etc. Les articles peuvent alors être séparés en amont sur la base de ces caractéristiques et il est possible par exemple, de regrouper les articles par journal ou par auteur si cela est pertinent pour la recherche. Cette subdivision est généralement opérée lors d'une analyse comparative pour laquelle il est nécessaire d'appliquer les protocoles d'analyse de manière indépendante sur chaque sous-corpus. En effet, il est possible d'analyser le corpus globalement et d'examiner les résultats sous l'angle de ces éléments extratextuels, mais ce choix dépend du type d'analyse et de la possibilité de limiter les « contaminations » qu'une analyse effectuée globalement peut apporter. La seconde approche exécute une subdivision en sous-corpus sur la base d'*éléments intratextuels*. Cette approche est plus complexe et requiert une analyse de chaque article, à l'intérieur desquels des catégories particulières peuvent être identifiées. Les catégories à repérer dans chaque article peuvent varier selon le cadre théorique adopté par le chercheur. Par exemple, il est possible de distinguer la partie de l'article qui « raconte les évènements » de celle qui « commente les événements » et créer ainsi deux sous-corpus distincts. Cette dernière approche peut être employée quand une étape d'annotation préalable est prévue. Dans ce contexte, les annotations ou *codes* sont utilisés pour enrichir les analyses successives.

Dans la phase analytique, les sous-corpus peuvent être soumis aux mêmes algorithmes de *manière individuelle* ou *en groupe*. Si les sous-corpus sont soumis un par un aux algorithmes, les résultats de l'analyse sont indépendants, car une distinction en amont est effectuée. En effet, les algorithmes analysent un jeu de données plus petit, les impacts de la dispersion de l'information sont réduits et les impacts d'autres corpus sur l'analyse sont annulés. Si le sous-corpus est soumis en

groupe, les résultats ne sont pas indépendants car chaque sous-corpus influence les résultats des autres sous-corpus et les spécificités de chacun d'entre eux doivent être identifiées a posteriori. Toutefois dans ce cas, les analyses offrent une vision plus globale et moins relative de l'ensemble du corpus, ce qui peut être également pertinent.

3.6 Regroupement de documents similaires

Le regroupement des documents similaires s'applique au corpus de travail ou aux sous-corpus de travail. Il s'agit d'une étape d'analyse qui a comme but la production de groupes d'articles qui partagent les mêmes caractéristiques. Ces groupes seront appelés clusters parce que la technique utilisée fait partie de la famille de méthodes non supervisées appelée *clustering* (cfr. 2.3.2).

3.6.1 K-moyennes

Le *k-moyennes* est l'algorithme utilisé dans le présent travail pour identifier les groupes d'articles similaires. Cet algorithme est un des plus vieux algorithmes développés dans le cadre de l'analyse de données et il demeure encore une des pierres angulaires du domaine (Bouroche et Saporta, 1992; Jain, 2010). Son succès est essentiellement dû à ses performances et à sa simplicité. Le k-moyennes, aussi appelé l'algorithme de centres mobiles, est un algorithme de partitionnement de type complet et exclusif. Il en existe toutefois des versions floues. Son principe est d'identifier les clusters d'une partition au moyen de centres mobiles, appellés *centroïdes*. Ces derniers se déplacent dans l'espace et, à chaque itération, forment des clusters sur la base de la *distance* entre individus et centroïdes. Lorsque l'inertie intraclasse ne diminue plus, l'algorithme converge (c'est-à-dire qu'il s'arrête) et il retourne la dernière partition obtenue. Le principal paramètre de l'algorithme est k,

qui détermine le nombre de centroïdes qui sont utilisés et donc, les nombres de clusters qui seront construits.

Cet algorithme est basé sur l'*inertie intraclasse* et l'*inertie interclasse* qui correspondent aux concepts d'homogénéité intraclasse et d'hétérogénéité interclasse. En ce sens, l'inertie est un indice d'homogénéité ou d'hétérogénéité. Pour comprendre ces deux concepts, il faut commencer par l'*inertie totale* d'un nuage, qui est l'espace des individus représenté en n dimensions. L'inertie totale d'un nuage est la somme pondérée des carrés des distances des individus x_i à leur centre de gravité.

$$I(A) = \frac{1}{n}\left[d^2(x_1,g) + d^2(x_2,g) + \cdots + d^2(x_i,g)\right] \tag{3.4}$$

dans laquelle d est la distance euclidienne. La formule peut être simplifiée ainsi :

$$\sum_{i=1}^{n} p_i d^2(g, x_i) \tag{3.5}$$

dans laquelle p est le poids de chaque individu (par exemple $\frac{1}{n}$) et d est la distance euclidienne de chaque individu x_i au centre de gravité g du nuage. L'inertie mesure la dispersion du nuage; elle sera grande pour un nuage très dispersé et petite lorsque le nuage est constitué de points bien regroupés.

Ainsi, le concept de dispersion constitue la base de l'algorithme k-moyennes et du concept d'inertie intraclasse et d'inertie interclasse. En effet, l'inertie intraclasse mesure la dispersion des individus autour d'un sous-nuage, soit le cluster. La formule est donc la même que la précédente, sauf qu'on remplace le centre de gravité g, qui

est le centre du nuage entier, par les centroïdes k_i et le poids p, qui sont relatifs à chaque cluster. Dans ce contexte, un cluster peut ainsi être interprété comme un sous-nuage. Au contraire, l'inertie interclasse mesure la dispersion des centroïdes k_i en relation avec le centre de gravité g du nuage et il mesure donc le degré d'éloignement entre clusters. Dans la formule (3.5), on substitue ainsi les points du nuage x_i pour le centroides k_i et on réintroduit le centre du nuage total g. La somme de l'inertie intraclasse et de l'inertie interclasse correspond à *l'inertie totale* et, à travers ce qu'on appelle la *décomposition de Huygens*, il est possible d'énoncer le théorème suivant :

> *l'inertie totale d'un nuage de point, à l'intérieur duquel existent des sous-nuages distincts est la somme de son inertie intraclasse et de son inertie interclasse.*

Ce théorème mène à la formule suivante :

$$I(A) = I_{intra} + I_{inter} \tag{3.6}$$

ce qui implique que, si l'inertie intraclasse diminue, alors l'inertie interclasse doit augmenter et vice versa. Les deux inerties sont alors dépendantes l'une de l'autre. À travers la décomposition de Huygens, on explicite le principe de classification du k-moyennes, qui veut que la partition d'un nuage en k clusters, où k est fixé d'avance, est d'autant *meilleure* que son inertie intraclasse[17] est petite ou que son inertie interclasses est grande. L'algorithme k-moyennes tente de trouver la meilleure partition en minimisant le plus possible l'inertie intraclasse ce qui correspond à la maximisation de l'inertie interclasse. Ces notions seront reprises dans le prochain

[17] L'inertie intraclasse dans ce contexte correspond à la somme des inerties intraclasse de chaque cluster.

paragraphe où seront traités l'évaluation de la partition et le choix du meilleur k pour un jeu de données.

Les présupposés de l'algorithme k-moyennes correspondent exactement aux principes d'homogénéité intraclasse et d'hétérogénéité interclasse qui sont à la base du problème de la classification. Il existe plusieurs versions de l'algorithme k-moyennes de type *hard clustering* (Lloyd-Forty, MacQueen, Hartigan & Wong, et d'autres). Chacune d'entre elles exécute les mêmes opérations algorithmiques, mais avec quelque petite variation. En général, l'algorithme peut être décrit par les opérations suivantes :

Algorithme 3.1 K-moyennes
Adapté de Hartigan *et al.*, (1979)

K-moyennes – adapté de Hartigan *et al.*, (1979)

Input : Une matrice U où chaque ligne correspond à une observation
Output : Une partition de U en K clusters

```
1   Choisir le nombre n du paramètre K, qui détermine le nombre de cluster à produire
2   Choisir la distance d (généralement euclidienne ou cosinus)
3   Choisir la méthode d'initialisation des k_n
4       Jusqu'à la convergence :
5           Pour chaque x_i :
6               Déterminer le centroïde le plus proche de x_i
7               Assigner x_i au cluster k généré par le centroïde le plus proche
8           Pour chaque cluster k_n :
9               Calculer la moyenne des points x_i qui lui appartient
10          Si les moyennes x̄_k ne sont pas égaux à k_n :
11              Les moyennes x̄_k constituent les nouveaux centroïdes k_n
12              On exécute à nouveau l'étape 5 avec les nouveaux centroïdes k_n
```

Ceci est la procédure algorithmique la plus classique du k-moyennes et elle garantit la convergence vers une partition qui minimise l'inertie intraclasse et maximise l'inertie interclasse. Son principal paramètre est *K*, mais d'autres paramètres en font partie et ils influencent également la qualité de la partition. Un premier paramètre est *la méthode d'initialisation* (étape 3 de l'algorithme 3.1). Les centroïdes k_n sont des vecteurs ayant la même dimension que la matrice *U* qui a été construite dans la phase de vectorisation des textes. Dans la version classique de l'algorithme, soit la version de Lloyd, ces vecteurs sont créés avec des valeurs aléatoires. Ils sont ainsi projetés dans l'espace où ils constituent des points aléatoires. Il s'agit donc d'une méthode d'initialisation aléatoire. Toutefois, *la partition finale est très sensible à la position initiale des centroïdes*. Il existe alors des méthodes différentes qui permettent de mitiger le rôle de l'initialisation dans la détermination de la partition. Un exemple est la méthode *k-means++* (Arthur et Vassilvitskii, 2007) qui permet de choisir les centroïdes de manière aléatoire mais par le biais d'une fonction qui équilibre les poids en fonction du jeu des données analysées. Un autre paramètre important est la *fonction de distance* (étape 2 de l'algorithme 3.1). Généralement, la distance euclidienne est utilisée. Toutefois, cette distance n'est pas toujours adéquate, surtout lors du traitement de données textuelles. En effet, cette distance ne prend pas en compte la longueur du texte et ainsi, des textes de diverses grandeurs risquent d'être traités différemment. Pour cette raison, *en fouille de texte, la distance cosinus est plus adaptée* et répandue, car elle normalise la longueur du vecteur et se concentre davantage sur l'angle entre les vecteurs plutôt que sur leur position dans l'espace. Le k-moyennes appliqué à des données textuelles *exige l'utilisation de la distance cosinus*. Elle est obtenue à partir de la similarité cosinus et correspond à la fonction suivante :

$$d_{cos} = 1 - \cos\theta \tag{3.7}$$

la fonction de similarité cosinus correspond à l'angle de deux vecteurs et elle sera utilisée pour l'exécution de l'algorithme k-moyenne et pour d'autres tâches :

$$\cos\theta = \frac{A \cdot B}{\|A\| \cdot \|B\|} \tag{3.8}$$

En effet, cette fonction permet de *mesure la similarité sémantique entre deux documents* et, en raisons de ses propriétés, elle est la fonction de similarité la plus utilisée en recherche d'information.

3.6.2 Évaluation des partitions

Le k-moyennes possède un paramètre principal, le paramètre *k*, qui détermine le nombre de centroïdes à utiliser et donc le nombre de clusters qui sont formés. Fixer *k* permet de créer une procédure algorithmique de découverte de la meilleure partition basée sur l'inertie intraclasse (homogénéité) et l'inertie interclasse (hétérogénéité). Ne pas fixer ce paramètre à l'avance conduit à l'impossibilité de faire converger ce type d'algorithmes car *l'inertie intraclasse diminue constamment lors de l'augmentation de k.* Il s'agit d'un aspect déterminant pour comprendre le contexte d'évaluation de l'algorithme k-moyennes. En effet, le théorème de Huygens, qui est à la base de plusieurs autres algorithmes de partitionnement, met en évidence l'impossibilité de comparer deux partitions ayant deux *k* différents. En se basant sur la minimisation de l'inertie intracalsse, une partition $k+1$ sera toujours meilleure qu'une partition *k*. À la limite, la meilleure partition possible est celle où chaque individu constitue un cluster, car dans ce cas l'inertie intraclasse est égale à zéro et l'inertie interclasse sera égale à l'inertie totale du nuage.

Ceci introduit les problèmes d'évaluation d'une partition dans un contexte non supervisé. Ce type de problème n'est pas limité à l'algorithme k-moyennes mais concerne la majorité des algorithmes de partitionnement. Pourtant, la phase

d'évaluation demeure importante pour disposer d'indices sur la qualité des résultats. Car malgré les difficultés, cette phase est effectuée pour comprendre la qualité de résultats et surtout pour fixer les paramètres des algorithmes. Dans le cas du k-moyennes, le principal paramètre à fixer est k et donc, le nombre de clusters à produire. Lors de l'application du k-moyennes à un corpus textuel, le but est d'identifier les clusters qui regroupent les documents en fonction de leur similarité. Déterminer le nombre de clusters est très important car les résultats changent avec chaque modification de k. Dans ce contexte, les difficultés de l'évaluation mènent à des procédures nécessitant une intervention humaine. Le contexte d'évaluation demeure toutefois le même et conduit à un protocole de ce genre : *le clustering est appliqué au corpus de travail ou aux sous-corpus de travail n fois afin d'identifier n différentes partitions, qui sont ensuite évaluées pour fixer le paramètre k.*

Il existe des indices qui proposent des fonctions alternatives afin de mitiger le problème du $k+1$ de l'inertie intraclasse. Pour simplifier, il est possible de séparer les indices d'évaluation du clustering en deux approches : une approche par *critères internes* et une approche par *critères externes*. Les premiers établissent une valeur pour la partition sur la base d'informations qui sont internes aux données et à la partition générée, ce qui se traduit essentiellement par l'évaluation du niveau d'homogénéité intraclasse et du niveau d'hétérogénéité interclasse. Cette typologie d'indices est la plus utilisée dans des applications réelles du clustering (Rendón *et al.*, 2011). Les approches par validation externe se basent sur le concept de précision et rappel et exigent de déterminer en amont quelle est la meilleure partition. Pour cette dernière raison, les critères externes sont le plus souvent utilisés dans un contexte de développement et d'optimisation des algorithmes car leur but est d'évaluer les performances de l'algorithme.

Les critères internes se basent sur l'homogénéité et l'hétérogénéité, aussi nommés niveau de *compacité* et niveau de *séparabilité* (Liu Y. *et al.*, 2010). Mesurer le niveau de compacité signifie qu'il faut établir jusqu'à quel point les objets d'un cluster sont similaires. Certains indices se basent sur le calcul de la variance. En effet, moins la variance est élevée, plus compact sera le cluster, car les objets s'éloignent moins de la moyenne. D'autres indices se basent sur le calcul des distances entre les objets. Ainsi, moins la somme des distances entre toutes les paires d'individus d'un cluster est grande, plus son niveau de compacité est élevé. Mesurer le niveau de séparabilité signifie qu'il faut établir jusqu'à quel point les clusters sont distincts. Plus la distance entre clusters est élevée, plus grand est le niveau de séparabilité. La majorité des indices évaluent les partitions en combinant ces deux critères, la compacité et la séparabilité, comme le fait, par exemple, l'indice de Calinski-Harabasz ou l'indice silhouette. Malgré tous les efforts pour réduire l'impact du phénomène du k +1, il n'existe pas encore de méthode d'évaluation fonctionnelle et largement acceptée pour la détermination du meilleur k dans un contexte de fouille de textes. Généralement, les indices sont utilisés pour guider le chercheur dans le choix des partitions à évaluer manuellement. La meilleure partition dépend alors de la *granularité* de l'analyse. En effet, les différentes partitions présentent des résultats qui sont plus détaillés avec l'augmentation de k. Choisir un k plus ou moins élevé dépend aussi des objectifs de la recherche et de la granularité de l'analyse à conduire.

Les indices d'évaluation basés sur des critères externes ont besoin préalablement d'informations sur la meilleure partition. L'évaluation est alors exécutée dans le même cadre que celui de la recherche d'informations, où les mesures principales sont la précision et le rappel. L'indice d'évaluation le plus simple est la *F-mesure* qui fournit une évaluation combinée de ces deux mesures. La meilleure partition est alors celle qui se rapproche le plus de la partition réelle, qui est connue préalablement. La précision indique la proportion d'individus du cluster en faisant réellement partie et le

rappel mesure la quantité d'objets qui n'ont pas été classés dans le bon cluster. Un grand niveau de précision et un grand niveau de rappel génèrent une grande valeur de la F-mesure.

Enfin, l'évaluation des partitions obtenues par un clustering est une opération fondamentale pour garantir le développement de cette famille de méthodes. Elle demeure toutefois une opération extrêmement difficile dans des contextes non-supervisés et lors des applications réelles (Aggarwal et Chandan, 2014, p. 571). En ce qui concerne la fouille de textes, la procédure d'évaluation prévoit l'intervention du chercheur pour déterminer la partition à utiliser dans les analyses. L'intervention humaine dans la procédure qui est adoptée dans cette chaîne de traitement prévoit l'évaluation de la distribution des documents dans les clusters. Cet élément constitue un outil supplémentaire pour guider le choix du meilleur k et a comme but l'élimination des partitions contenant des clusters trop grands. Ceci est un phénomène assez commun dans la pratique de la fouille de textes qui comporte des partitions contenant un ou plusieurs clusters auxquels la majorité des documents est assignée, ce qui implique de nombreux clusters avec peu de documents. En principe, moins la partition contient de grands clusters, plus les documents tendent à être distribués également dans les clusters. La meilleure partition est celle qui, compte tenu de l'indice interne et de la granularité de l'analyse visée, contient des clusters qui tendent à avoir le même nombre de documents. Évidemment, il est impossible d'obtenir n clusters comportant le même nombre de documents, mais on peut évaluer la tendance générale. Cette tendance peut être illustrée par l'écart type des distributions de documents dans les clusters. En calculant l'écart type, il est possible d'observer un indice de dispersion permettant d'identifier la différence entre les clusters en termes de nombre de documents qui leur appartiennent. Moins l'écart type d'une partition est grand, plus les chances sont élevées d'obtenir une partition où les clusters tendent à avoir le même nombre de documents. *L'intervention humaine*

implique donc de choisir le meilleur k sur la base d'un certain nombre d'éléments : l'indice interne, la granularité de l'analyse visée, l'écart type des distributions de documents dans les clusters et enfin, le rapport entre le nombre de *k* et le nombre de documents du corpus. Ce dernier élément est utilisé lors de la définition de plusieurs sous-corpus de travail et il garantit le maintien du même niveau de granularité pour des sous-corpus de dimensions différentes. En effet, lors de la division du corpus de travail en sous-corpus, le clustering peut être effectué de manière indépendante sur les différents sous-corpus. Tenir compte de la grandeur de chaque sous-corpus lors du choix du *k* permet d'équilibrer les analyses comparatives. Si un sous-corpus très petit obtient un *k* égal à un sous-corpus plus grand, l'analyse risque d'être biaisée car la granularité sera différente pour les deux sous-corpus.

3.7 Annotation

À la fin de la chaîne de traitement, une série de groupes d'articles (*clusters*) sont identifiés en fonction de leur similarité sémantique (James *et al.*, 2013, p. 385). Ils sont ensuite analysés afin de détecter les motifs récurrents des clusters (Aggarwal et Chandan, 2014; Lapalut, 1995) et, en raison de notre hypothèse, afin d'associer à chacun d'entre eux des structures narratives. L'analyse se poursuit par *une étape d'annotation* qui est réalisée au moyen d'un *protocole d'annotation.* L'objectif principal de cette étape consiste à identifier, dans chaque article d'un échantillon spécifique, le ou les programmes narratifs principaux et les actants qui leur sont associés.

Selon le cadre théorique qui soutient notre hypothèse sémiotique (cfr. chapitre II), un texte journalistique est considéré comme une sorte de « plate-forme » de la manifestation des structures sémantiques régies par une syntaxe narrative. La syntaxe narrative organise des actants qui, dans le texte, se manifestent à travers des acteurs

puisque les actants passent toujours par un processus de conversion (cfr. chapitre II). La puissance d'un tel modèle est constituée par le fait que la sémantique d'un texte ne se définit pas que par la *présence* des éléments sémantique, mais aussi par les *relations* qu'ils entretiennent les uns avec les autres. Ainsi un acteur ayant le rôle actantiel du sujet ne peut être défini seulement par sa relation de jonction avec l'acteur ayant le rôle actantiel de l'objet de valeur. Les éléments sémantiques tendent alors à être identifiés à un rôle actantiel et à se définir par les relations qu'ils entretiennent avec les autres. Ces relations sont régies par une syntaxe narrative. Ceci permet de considérer l'analyse de texte comme l'analyse des relations narratives entre les éléments sémantiques, ce qui enrichit l'analyse et les interprétations.

L'hypothèse computationelle de notre thèse affirme qu'un algorithme de clustering permet de regrouper les articles qui partagent le même schéma récurrent, ce dernier étant l'instanciation d'une macrostructure. L'hypothèse sémiotique affirme que le schéma récurrent responsable de la constitution d'un cluster correspond à une structure narrative. Un schéma récurrent *se manifeste* alors par un certain nombre d'éléments sémantiques qui entretiennent des relations narratives, lesquels sont *récurrents* dans la plupart des articles que le cluster contient. L'identification de ces éléments et de leurs relations est effectuée grâce à *un protocole d'annotation* qui révèle une structure narrative. Pour ce faire, le protocole d'annotation s'inspire du *métalangage sémiotique de Greimas* et, plus particulièrement de la *syntaxe narrative* qui constitue le parcours génératif du sens.

3.7.1 Principes de base

Le protocole vise à identifier les *actants* qui sont en jeu dans le ou les principaux *programmes narratifs* de chaque article. Il est constitué de deux étapes principales : le repérage du ou des programmes narratifs principaux et la détection de leurs actants manifestés. Par exemple, si un article de chronique raconte les évènements de la

dernière manifestation étudiante, l'annotation identifie d'abord le programme narratif principal et ensuite les actants qui le concernent, ce qui correspond à l'annotation suivante à partir de la formule (2.1) (cfr. 2.2.2) :

Tableau 3.2 Exemple d'un programme narratif

PN*n*	*Sujet*	*Object de valeur*
	manifestants	**manifestation**

où l'actant *sujet* est les **manifestants** et l'actant *objet de valeur* est la **manifestation.** Cette première annotation peut ensuite être enrichie par le repérage d'autres programmes narratifs ou d'autres actants impliqués. Un des premiers actants qui peut être ajouté est l'*antisujet*, qui fait partie du programme narratif contraire. En effet, chaque programme narratif a son contraire selon la règle définie dans la formule (2.3) (cfr. 2.2.2).

Tableau 3.3 Exemple d'un programme narratif complété par son opposé

PN*n*	*Sujet*	*Object de valeur*	*antisujet*
	manifestants	**manifestation**	**police**

Cette règle implique, par définition qu'à chaque programme narratif de jonction correspond un programme narratif de disjonction opposé où les rôles actantiels sont inversés. Par exemple, la **police** peut avoir un rôle de sujet ou d'antisujet selon que le programme narratif se manifeste par la jonction ou la disjonction. L'axe qui connecte le sujet à l'objet de valeur est appelé l'*axe du vouloir* et il constitue le pivot sur lequel la modélisation greimassienne se fonde. Enfin, le protocole d'annotation est centré sur le repérage de cet axe, qui constitue le programme narratif.

L'annotation est ensuite enrichie par les deux autres axes impliqués dans chaque programme narratif et qui organisent d'autres actants : l'*adjuvant* et l'*opposant* dans *l'axe du pouvoir*, le *destinateur* et le *destinataire* dans l'*axe de la transmission*. La figure 2.5 (cfr. 2.2.2) permet de visualiser des axes actantiels autour d'un programme narratif. La figure illustre ce qui est appelé le *schéma actantiel*, lequel met au centre la relation entre sujet et objet. Il constitue la base du protocole d'annotation, qui est orienté par les principaux éléments de la théorie greimasienne. L'axe du vouloir, qui correspond au programme narratif, explicite la relation entre un sujet et un objet.

Le protocole d'annotation se base alors sur le schéma actantiel. Le schéma actantiel a été préféré au schéma narratif canonique qui, en principe, représente une évolution du premier. Le choix du schéma actantiel est motivé par la simplicité du modèle, mais surtout par le fait qu'il est centré sur l'organisation actantielle plutôt que sur les segments logiques qui fondent un récit. Bien que le schéma narratif canonique présuppose le schéma actantiel, il met de l'avant l'acte du sujet plutôt que les relations actantielles. En fait, l'identification des actants et de leurs relations est une opération différente de l'identification des séquences narratives. De notre point de vue, le *modèle actantiel permet de cerner le cœur narratif de chaque texte en se concentrant sur les relations syntaxiques entre éléments sémantiques*. Le schéma actantiel permet ainsi de visualiser facilement les éléments communs et récurrents entre articles. Mettre en valeur les actants et leurs relations plutôt que les processus dans lesquels ils sont impliqués s'adapte à notre objectif de recherche et évite d'utiliser infructueusement un cadre d'analyse plus complexe.

3.7.2 Détails de la mise en place

D'un point de vue pratique, l'annotation sera assistée par un logiciel construit pour l'analyse qualitative, qui est basé sur R et qui s'appelle RQDA (Huang, 2018). Les raisons de l'utilisation de RQDA pour l'annotation sont la facilité avec laquelle le

logiciel permet d'associer des annotations à des passages des textes, mais c'est surtout la base de données que le logiciel construit qui représente sa véritable valeur ajoutée. La base de données organise les annotations et les passages de textes correspondants dans une base de données relationnelle qui peut être interrogée aisément à travers le langage SQL. La base de données est une forme instanciée de notre « trace de lecture » et RQDA permet de retrouver facilement ces « traces » et de les organiser de manière structurée.

3.7.3 Évaluation de la chaîne de traitement

La procédure d'annotation produit des résultats sur les caractéristiques sémantiques et structurelles des clusters et, de cette manière, elle permet d'évaluer la chaîne de traitement. En effet, pendant l'annotation, il est possible de valider si les clusters sont véritablement organisés autour des mêmes schémas récurrents. Les résultats de cette procédure doivent répondre aux attentes de nos hypothèses de recherche : celle computationelle qui affirme que le clustering regroupe des articles par similarité sémantique et celle sémiotique qui veut que les similarités sémantiques soient identifiables par un schéma récurrent de type narratif. En effet, si les hypothèses sont fondées, alors la procédure d'annotation devrait fournir des éléments de comparaison entre les articles d'un même cluster. En d'autres termes, si la phase d'annotation ne permet pas d'utiliser les mêmes annotations pour un cluster, alors la méthode ne répond pas, ou ne répond que partiellement, aux attentes qui découlent des hypothèses de la recherche. Au contraire, *si pendant la phase d'annotation il est possible d'identifier un certain nombre d'éléments similaires entre les articles d'un même cluster, alors la méthode répond positivement aux attentes de la recherche*. Enfin, la chaîne de traitement décrite dans la méthode représente une étude de faisabilité dont les résultats sont analysés et évalués à travers une analyse détaillée qui est essentiellement constituée par la présente procédure d'annotation. Si les résultats

répondent aux attentes de la recherche, alors l'étude de faisabilité peut être considérée comme une véritable nouvelle piste de recherche pour la sémiotique computationelle.

3.7.4 Détermination de l'échantillon

Dans la présente méthode, un échantillon est construit pour chaque cluster lors de la phase d'annotation. En analyse de presse, l'échantillon est un groupe restreint d'articles qui représente un groupe d'articles plus grand. En gros, l'échantillon doit répondre aux mêmes critères en jeu lors de la constitution du corpus afin de représenter une population plus large (cfr.3.2). Restreindre le nombre d'articles à analyser en détail fait partie des pratiques courantes en sciences humaines et sociales. Construire un échantillon d'une population est donc une étape importante de l'analyse de textes et de l'analyse de presse.

Dans le présent travail, toutefois, *l'échantillon ne se situe pas au point de départ de la méthode*, comme on pourrait s'y attendre, *mais plutôt au point d'arrivée*. Ceci est dû au fait que la méthode permet de traiter un corpus de grande taille et donc, que le besoin d'échantillonner au début de la chaîne de traitement est moins importante, ce qui fait déplacer la constitution d'échantillons à la fin de la chaîne de traitement. Dans ce contexte, l'échantillon assume un rôle différent, car il est constitué pour les phases d'annotation et d'évaluation et non pour la phase d'analyse elle-même. Enfin, *la population de l'échantillon correspond à la totalité des articles qui ont été regroupés dans un même cluster*. On peut ainsi parler de populations (au pluriel). *Le nombre d'échantillons à construire est alors égal au nombre de clusters* obtenus dans la phase de regroupement des articles similaires.

Dans les champs d'études liés à l'analyse de presse, comme l'analyse de contenu ou l'analyse du discours, plusieurs méthodes d'échantillonnage ont été développées pour *construire un ensemble de documents aptes à être analysés* tout en respectant les

critères d'orientation, de pertinence, d'homogénéité, d'hétérogénéité, d'exhaustivité et de représentativité. Les méthodes sont nombreuses et chacune d'elles présente des avantages et des désavantages. En simplifiant, il est possible de distinguer deux grandes catégories d'échantillons : les *échantillons aléatoires* ou « probabilistes », où généralement les observables sont choisis au hasard, et les *échantillons empiriques* ou « non probabilistes », où les observables sont choisis selon des principes non aléatoires. (Martin, 2009, p. 16).

Le premier groupe, soit celui des échantillons aléatoires, est basé sur des méthodes statistiques et sur la loi de grands nombres. Ces méthodes se justifient par des arguments mathématiques de nature probabiliste, mais aussi par le présupposé que chaque unité de la population possède le même poids informatif (Krippendorff, 2004, p. 113-114). Le deuxième groupe de méthodes, soit celui des échantillons empiriques, est très varié et les méthodes qui en font partie ne se basent pas sur des principes probabilistes.

Ces méthodes sont appliquées lors de la construction d'un échantillon qui est utilisé *pour la véritable analyse*. Dans le cadre de la présente étude, l'échantillonnage est exécuté sur chaque cluster qui a été créé et donc, *après l'étape d'analyse*. Dans notre cas, l'échantillon doit permettre d'obtenir un petit nombre d'articles pour compléter l'analyse et permettre l'évaluation de la méthode. En d'autres termes, les échantillons servent à analyser les résultats du processus automatique, et ils doivent être représentatifs des clusters.

Ainsi, l'échantillonnage est adapté à la chaîne de traitement et pour qu'il le soit, on exploite les caractéristiques de la méthode. Le clustering est l'opération qui nous a permis de regrouper des articles similaires. L'algorithme utilisé est le k-moyennes. La procédure algorithmique converge lorsque les centroïdes s'arrêtent sur des centres de

gravité de l'espace. Ces centres correspondent à des *sous-espaces dans lesquels des articles sont plus rapprochés* ce qui signifie qu'ils sont similaires entre eux. En d'autres termes, les centres de gravité sont les centres des sous-espaces les plus denses. Les centroïdes représentent alors une sorte d'*article prototypique du cluster*. Cette information n'est pas triviale et elle peut être exploitée pour *trier* et *échantillonner*. En effet, les articles d'un cluster peuvent être ordonnés selon la distance qui les sépare de leur centroïde, ce qui correspond à une mesure de similarité entre les articles et le prototype du cluster.

Afin de produire l'échantillon le plus représentatif du cluster, il est fondamental d'utiliser la mesure de similarité (formule 3.8) entre les articles et le prototype. Cette approche est justifiée par le fait qu'un cluster correspond à un sous-espace dont la densité augmente en fonction de la diminution des distances qui séparent les articles du sous-espace au centre de gravité. Il est donc raisonnable de sélectionner les individus qui sont proches du centre de gravité du sous-espace à des fins d'échantillonnage. En outre, surtout dans un contexte de fouille de textes, il est très fréquent que des documents se situent aux limites du sous-espace. La distance avec le centre de gravité permet d'exclure les documents plus éloignés, garantissant la sélection d'articles plus proches du prototype et donc *plus représentatifs du sous-espace*, qui est le cluster.

La méthode d'échantillonnage se base donc sur la distance entre les articles et les prototypes et sur un *seuil* choisi par le chercheur. En résumé, un échantillon représentatif est construit pour chaque cluster afin que la procédure d'annotation puisse être exécutée. La méthode d'échantillonnage exploite les centroïdes du clustering pour identifier *les articles les plus proches du centre de gravité de chaque cluster*. Ensuite, un *seuil doit être établi pour extraire les n articles les plus proches du centre de gravité*. Le seuil utilisé dans le présent travail est d'un tiers, ce qui

permet de retenir le 33 % des articles contenus dans chaque cluster. Ces articles seront annotés selon la procédure décrite dans le paragraphe précédent.

CHAPITRE IV

EXPÉRIMENTATIONS

La méthode de la présente étude est composée d'une chaîne de traitements qui utilise des outils informatiques et d'une phase d'annotation employant des outils de la sémiotique. La méthode est construite pour l'analyse d'un grand nombre d'articles de presse. Plus particulièrement, la chaîne de traitement vise à identifier des groupes d'articles similaires et à fournir des échantillons à la phase d'annotation afin de dégager des schémas récurrents mis en place par l'ensemble des articles. L'étude de ces schémas fournit des éléments importants pour l'analyse de la formation de l'opinion publique. La présente étude constitue une démonstration de faisabilité qui a été appliqué à un corpus qui regroupe des articles de presse sur le printemps érable.

Le présent chapitre décrit l'application de la méthode au corpus utilisé pour cette étude. L'application de la méthode implique une série de choix et décisions spécifiques à chaque différente étape de la méthode. Des *protocoles d'application* ont été ainsi générés et ils seront présentés dans ce chapitre. Les résultats obtenus dans l'application de chaque étape de la méthode seront aussi présentés.

Le chapitre est subdivisé en cinq sections : la première explicite les détails de la construction du corpus; la deuxième détaille les procédures de la phase de prétraitement du corpus; la troisième est la plus courte et explicite la nature du corpus

de travail; la quatrième présente les résultats de l'application du clustering au corpus de travail; enfin, la dernière présente les résultats de la phase d'annotation.

Le résultats obtenus par la chaîne de traitement, et plus particulièrement les résultats de l'étape de regroupemment automatique d'articles similaires, peuvent être consultés de manière autonome au moyen d'une application web[18] construite à cet effet. L'application est en version beta et elle doit être considerée exclusivement comme support optionel. Elle ne constitue aucunement en un prolongement officiel de cette thèse.

4.1 Le corpus

Un corpus est composé de différents niveaux[19] : l'archive, le corpus de référence, le corpus d'étude et les sous-corpus de travail. Bien qu'il n'existe pas de véritable archive ou de base de données organisée sur le printemps érable, l'archive sur le sujet est constituée par tous les articles de journaux, les émissions radiophoniques et télévisées, les publications scientifiques, les livres, les blogues et les microblogues d'information etc. qui ont traité, de manière directe ou indirecte, de la question du printemps érable. L'archive correspond à tous les énoncés qui ont été produits publiquement, tous médias confondus. Le corpus de référence est constitué de l'ensemble des articles, sous forme écrite, des principales sources d'information

[18] Le lien de l'application est : https://pulizzottodavide.shinyapps.io/these_phd_printemps_erable/

[19] L'*archive* est, pour Rastier, un ensemble de documents accessibles mais qui ne constitue pas un véritable corpus. Le *corpus de référence* est constitué par un ensemble de documents à partir duquel le corpus d'étude sera construit. Le passage de l'archive au corpus de référence se réalise à travers le critère d'orientation qui met en relief la nécessité d'avoir un ensemble de documents pour une tâche ou une application précise. Le *corpus d'étude* est l'ensemble de documents collectés sur lesquels l'analyse est effectuée. Ce corpus dérive du corpus de référence et il est construit pour les besoins spécifiques de l'application. Les *sous-corpus de travail* sont construits à partir du corpus d'étude et ils sont mis en place pour optimiser la phase d'analyse

québécoises qui ont traité, directement ou indirectement, de la question du printemps érable. Pour obtenir ce corpus de référence, un *protocole* a été construit et il est décrit dans la prochaine section. Le corpus d'étude est le résultat d'un processus de filtrage mis en place pour améliorer l'homogénéité et la pertinence du corpus. Enfin, les sous-corpus de travail sont constitués par sept ensembles d'articles divisés par source d'information.

Le corpus est soumis à plusieurs étapes de constitution qui permettent d'identifier les groupes de documents sur lesquels l'analyse sera effectivement exécutée. Toutes les étapes de constitution des différents niveaux du corpus respectent le plus possible les critères de constitution du corpus. Chaque niveau du corpus peut donc être évalué sur la base de ces critères. Dans ce contexte, il pourrait être pertinent d'évaluer un corpus en lui assignant une valeur allant de zéro (critère non respecté) à cinq (critère respecté pleinement) pour chacun de ces critères. Par exemple, un ensemble d'articles peut être évalué en considérant l'orientation et la pertinence, auxquelles deux valeurs de 0 à 5 y sont associées. Afin que ce genre d'évaluation puisse être considérée valide, un protocole complexe devrait être mis en place. En particulier, la plus grande partie de sa validité dérive principalement de l'*entente interjuge*, ce qui implique des ressources importantes. Dans le présent travail, cette opération n'a pas été réalisée à cause du manque de ressources adéquates.

Toutefois, des évaluations de ce genre seront présentées afin de fournir quelques indices sur la qualité du corpus. Ces évaluations ne suivent pas un protocole valide et ne sont présentées qu'à titre indicatif. Par exemple, la figure 4.1, illustre l'évaluation de l'archive et résume ses caractéristiques. En effet, l'archive est un corpus sans orientation car elle ne constitue pas un corpus pour une analyse spécifique. Par conséquent, sa pertinence en relation à une étude est nulle. L'archive est associée à une cohérence basse car, bien qu'elle regroupe tous les énoncés liés à un phénomène

spécifique, sa cohérence ne peut être élevée puisque les énoncés sont trop différente en terme de genre. En effet, dans un tel corpus, le genre textuel est très varié et, de plus, il est très probable qu'un très grand nombre d'articles ne traitent le phénomène à l'étude que de manière indirecte. Enfin, cette archive possède une très grande hétérogénéité et une très grande exhaustivité car elle regroupe tous les énoncés de tous médias. Un tel corpus, à la fois très hétérogène et très exhaustif, est par définition très représentatif. Le passage de l'archive au corpus d'étude et aux sous-corpus de travail doit augmenter le degré d'orientation, de pertinence et de cohérence sans diminuer excessivement l'hétérogénéité, l'exhaustivité et la représentativité.

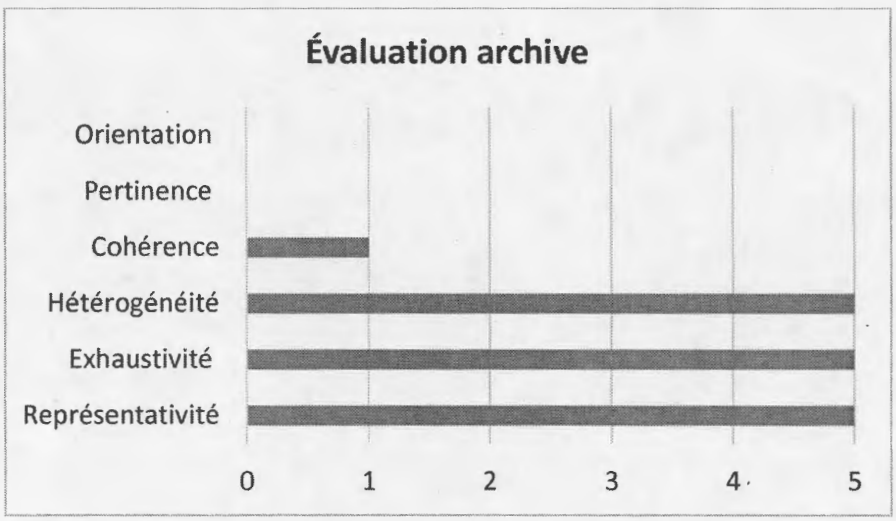


Figure 4.1 Évaluation de l'archive

4.1.1 Protocole de constitution du corpus

Le *corpus de référence* (Rastier, 2005a) est constitué par tous les articles de presse, sous forme écrite, publiés au sujet de la grève étudiante par les principales sources d'information. Il est principalement déterminé par le critère de l'orientation qui permet une première sélection sur la base d'une question de recherche. Dans le cadre de cette thèse, les émissions radiophoniques ou télévisées, les publications

scientifiques, les blogues et les microblogues d'information ne seront pas traités. Le cœur de notre matériel empirique est la *presse écrite* car notre question de recherche vise à identifier les macrostructures dégagées par la presse écrite québécoise. Plus spécifiquement, notre question de recherche vise l'analyse des *sources d'information les plus importantes en langue française*. Ces trois éléments, presse écrite, sources d'information les plus importantes et langue française, constituent les premiers *choix pour la sélection du matériel d'étude*.

Le choix de la presse écrite est déterminé par un élément fondamental de la problématique : étudier le traitement journalistique dans un cadre de recherche en sémiotique et à l'aide d'outils computationels. En effet, la majorité de ces outils ont été développés pour le traitement du langage naturel et, plus particulièrement, pour le texte écrit. Étudier d'autres types de média, comme les émissions télévisées, est un défi important pour la sémiotique computationelle appliquée dans un contexte empirique. L'analyse de corpus appartenant à des systèmes sémiotiques différents de celui de l'écriture requerrait l'utilisation d'outils très complexes, comme la reconnaissance vocale ou le traitement des images. De plus, ces outils devraient être adaptés pour la recherche en sciences humaines et sociales, ce qui représente une problématique qui, à notre connaissance, n'a jamais été abordée. Ces mêmes raisons déterminent également le choix d'un corpus mono-langue. Le traitement d'un corpus bilingue multiplie les procédures à exécuter et, surtout dans un cadre computationnel, complexifie le travail d'analyse. En effet, ce type d'analyse requerrait des outils de traduction automatique qui sont également très complexes.

Enfin, le choix de sources d'information retenues est basé sur l'analyse des plus importantes sources d'information québécoises. Ainsi, parmi les douze quotidiens imprimés en langue française au Québec, onze appartiennent à de grands conglomérats et ils se distribuent ainsi : *La Presse*, *Le Soleil*, *Le Nouvelliste*, *Le Droit*

(Ottawa/Gatineau), *Le Quotidien*, *La Tribune* et *La Voix de l'Est* appartiennent à Power Corporation; *Le Journal de Montréal*, *Le Journal de Québec* et *Montréal 24 heures* sont la propriété de Quebecor; *Journal Métro* appartient à TC Transcontinental, alors que le journal *Le Devoir* est indépendant. Compte tenu des informations obtenues sur le tirage, la taille du marché couvert en terme de population et la propriété, nous avons choisi six sources d'information considérées comme les plus importantes : *La Presse*, *Le Soleil, Le Journal de Montréal*, *Le Journal de Québec, Le Devoir* et *Journal Métro*. Les deux premiers sont les journaux les plus diffusés du groupe Power Corporation, les deux autres sont, en excluant le tabloïd *Montréal 24 heures*, les plus importants du groupe Quebecor, alors que *Le Devoir* est le seul journal indépendant. Ensemble, ces journaux atteignent une population supérieure à un million de personnes. Enfin, *Journal Métro* est un tabloïd gratuit, qui est très diffusé et qui n'appartient pas aux principaux conglomérats. Ainsi, parmi les sources choisies, les deux éditeurs les plus importants au Québec sont représentés de façon égale et le groupe d'éditeurs indépendants et le groupe de tabloïds distribués gratuitement nous apparaissent comme suffisamment représentés. À ces six quotidiens, une septième source d'information a été ajoutée : le site web de *Radio-Canada*, qui constitue une source d'information n'appartenant pas à des intérêts privés. Ces *sept sources d'information* atteignent la grande majorité des lecteurs québécois.

4.1.2 Moissonage

La sélection des articles a été réalisée à l'aide d'une procédure de recherche automatique sur la base de données *Eureka* mise à disposition par les bibliothèques de l'UQAM. Un script de moissonnage [20] a été développé pour effectuer les recherches et extraire les données textuelles. La procédure de recherche est basée sur

[20] Les modules *Selenium*, *Scrapy* et *Beautifulsoup4* de Python ont été utilisés.

trois types d'information : *mot clé*, *date de publication* et *source d'information*. Le mot clé a été utilisé pour sélectionner les articles en fonction du *critère de saillance* (Flament et Rouquette, 2003) selon lequel il existe des indicateurs lexicaux qui peuvent servir d'« ancrage empirique » au phénomène exploré. Le critère de saillance implique de choisir les mots qui, pour le printemps érable, peuvent remplir cette fonction d'ancrage empirique. Les termes « étudiant, étudiants, étudiante et étudiantes» ont été retenus à cet effet, car ils représentent les mots qui ont le plus de probabilité d'apparaître lorsqu'un article traite du printemps érable. La deuxième information utilisée par la procédure de recherche automatique est la date. La recherche a été limitée à une période qui va du 13 février 2012 au 10 septembre 2012, ce qui couvre l'entièreté de la période de grève jusqu'aux élections provinciales. Enfin, les articles qui appartiennent aux sept sources d'informations choisies ont été retenus. Pour le site web de Radio-Canada, les articles sélectionnés sont ceux qui sont apparus dans la page « Nouvelles », dans la page « Montréal » et dans la page « Québec ».

Il a été possible d'extraire de chaque article les informations suivantes : *source d'informations*, *date*, *titre*, *contenu de l'article*, *catégorie de l'article*, *numéro de page* et *auteur*. Les informations les plus importantes aux fins de la présente recherche sont contenues dans la source d'information, le titre et le contenu de l'article. À la fin du processus, 9 603 articles ont été extraits de la base de données.

4.1.3 Description du corpus

Les 9 603 articles extraits ne se distribuent pas également dans les sept sources d'information (figure 4.2). Pour *La Presse*, *Le Devoir*, *Le Journal de Montréal*, *Le Journal de Québec* et *Le Soleil*, la taille des documents qui les concernent est à peu près similaire. En effet, leur écart type (tableau 4.1) est plus petit que l'écart type total (tableau 4.2) et leur moyenne se rapproche beaucoup plus de leur médiane que de

celle des sept sources d'information. Le site web de *Radio-Canada* et le *Journal Métro* se distinguent donc des autres journaux en termes de taille car un nombre plus petit d'articles en proviennent. Le tabloïd distribué gratuitement affiche le moins grand nombre d'articles, ce qui est probablement dû à sa fréquence d'impression car il n'a pas d'édition de fin de semaine. Le nombre d'articles provenant de la source d'information publique n'est pas comparable avec les autres, car il s'agit d'un média différent, soit un site web, qui ne suit pas les mêmes règles. En effet, les sources d'information en ligne diffèrent des sources traditionnelles par leur capacité de traiter un événement en direct, ce qui influence grandement la fréquence de publication.

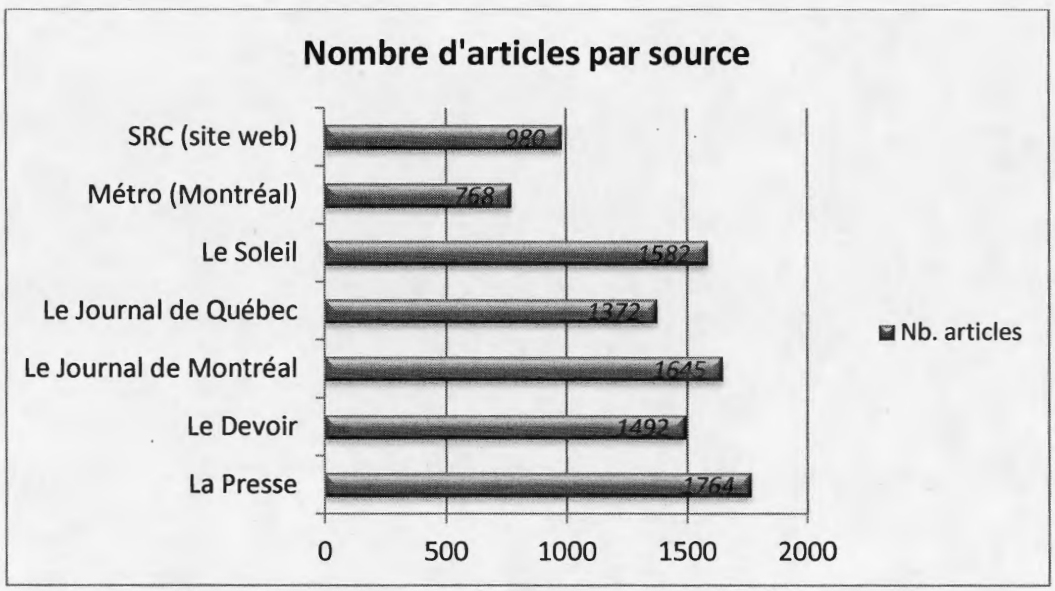


Figure 4.2 Distribution des 9 603 articles dans les sept sources d'information

Tableau 4.1 (a) Résumé des indices de dispersion. Indices de dispersion pour les sept sources d'information. (b) Indices de dispersion pour les cinq sources d'information contenant le plus grand nombre d'articles, soit *La Presse*, *Le Devoir*, *Le Journal de Montréal*, *Le Journal de Québec* et *Le Soleil*

(a) Nb. Sources information	7
Moyenne	1371.86
Écart type	366.33
Minimum	768
25e centile	1176
Médiane	1492
75e centile	1613.5
Maximum	1764

(b) Nb. Sources information	5
Moyenne	1571
Écart type	148.9
Minimum	1372
25e centile	1492
Médiane	1582
75e centile	1645
Maximum	1764

Les auteurs des articles sont très variés et il est préférable de les considérer comme des *signatures*. Les signatures sont très nombreuses et le ratio entre signature et nombre d'articles est très élevé. Pour *Le Devoir*, une nouvelle signature apparaît à presque chaque trois articles, à chaque 3,5 articles pour *Le Journal de Montréal* et pour *Le Journal de Québec*, à chaque 4,5 articles pour *La Presse*, à chaque huit pour *Métro*, à chaque neuf pour *Le Soleil* et à chaque 37 pour *Radio-Canada*. La signature la plus fréquente de chaque source d'information est celle Anonyme, sauf pour *Le Devoir* qui présente la journaliste Lisa-Marie Gervais comme la signature la plus fréquente.

Tableau 4.2 Résumé des fréquences des signatures par source d'information

Publication	Nombre articles	Signatures uniques	Signature la plus fréquente	Fréquence	Ratio art./signa.
Le Devoir	1492	534	Lisa-Marie Gervais	95	2,79
Le Journal de Montréal	1645	470	Anonyme	193	3,50
Le Journal de Québec	1372	390	Anonyme	164	3,52
Métro (Montréal)	768	96	Anonyme	458	8,00
La Presse	1764	393	Anonyme	229	4,49
SRC (site web)	980	26	Anonyme	854	37,69
Le Soleil	1582	172	Anonyme	494	9,20

Parmi les dix signatures les plus fréquentes par source d'information, on remarque dans *Le Devoir* deux typologies de signature qui correspondent à une signature anonyme, soit « Le Devoir » et « Anonyme »[21] pour un total de 80 signatures « Anonymes ». Ce journal présente aussi deux agences de presse (La Presse canadienne et l'Agence France-Presse) et d'autres journalistes comme Antoine Robitaille (44 signatures) ou Marie-Andrée Chouinard (37 signatures). Les journaux de Quebecor présentent des caractéristiques similaires affichant la signature anonyme et celle de l'agence de presse QMI, respectivement à la première et à la deuxième position. Les journalistes Sarah-Maude Lefebvre, Richard Martineau, Régys Caron, J. Jacques Samson et d'autres apparaissent les plus souvent dans les deux journaux. Pour le *Journal Métro*, Annabelle Blais (46 signatures) et Maxence Knepper (31 signatures) sont les journalistes qui signent le plus fréquemment alors que Pascale Breton (68 signatures) et Philippe Teisceira-Lessard (54 signatures) signent fréquemment pour *La Presse*. Le site web *Radio-Canada* n'affiche pas beaucoup de signatures et présente plusieurs versions de signatures anonymes. De plus, les données extraites de la base de données contiennent plusieurs erreurs de transcription ou de signatures doubles, ce qui requiert des outils de reconnaissance des entités nommées pour nettoyer et normaliser chaque entrée. Cette dernière opération n'a pas été effectuée, car l'étude de signature ne constitue pas un objectif de recherche. Le tableau 4.3 présente les données brutes telles que extraites de la base de données.

Tableau 4.3 Les premiers dix journalistes par source d'information en relation à la fréquence d'apparition

Le Devoir	**Freq**	**Le Journal de Montréal**	**Freq**
Lisa-Marie Gervais	95	Anonyme	193

[21] La valeur anonyme indique une absence d'information liée à la signature lors de la procédure de moissonnage.

La Presse canadienne	76	Agence QMI	163
Le Devoir	55	Sarah-Maude Lefebvre	56
Antoine Robitaille	44	Richard Marteneau	34
Marie-Andrée Chouinard	37	Régys Caron (bureau parlementaire)	26
Agence France-Presse	33	Marc-André Lemieux	24
Michel David	32	Michaël Nguyen	21
Robert Dutrisac	31	J. Jacques Samson	21
Anonyme	25	Francis A. Trudel	17
Christian Rioux	25	Jean-Luc Lavallée (bureau parlementaire)	16
Journal Métro (Montréal)	**Freq**	**La Presse**	**Freq**
Anonyme	458	Anonyme	229
Annabelle Blais	46	Pascale Breton	68
Maxence Knepper	31	Philippe Teisceira-Lessard	54
Mario Charette	17	Gabrielle Duchaine	48
Marc-Antoine Audette; Sébastien Trudel	15	Émilie Bilodeau	47
Les justiciers masq	10	Tommy Chouinard	38
annabelle blais	10	Denis Lessard	38
Mathias Marchal	8	Alain Dubuc	35
Natalia Wysocka	7	André Pratte	30
Roxane Léouzon	7	David Santerre	30
Le Soleil	**Freq**	**Le Journal de Québec**	**Freq**
Anonyme	494	Anonyme	164
Marc Allard	94	Agence QMI	90
Annie Mathieu	69	Richard Martineau (collaboration spéciale)	47
Michel Corbeil	62	Régys Caron (bureau parlementaire)	41
Gilbert Lavoie	54	J. Jacques Samson	31
Ian Bussières	45	Sarah-Maude Lefebvre	25
Samuel Auger	44	Jean-Luc Lavallée (bureau parlementaire)	23
Carl Thériault	38	Christian Dufour (collaboration spéciale)	21
Jean-Marc Salvet	37	Geneviève Lajoie (bureau parlementaire)	20
Simon Boivin	29	Pierre Gingras (collaboration spéciale)	20
SRC (site web)	**Freq**		
Anonyme	854		
Radio-Canada avec La Presse Canadienne	52		
Radio-Canada avec Agence France-Presse et Reuters	18		
Radio-Canada avec Agence France-Presse	11		
Sophie-Hélène Lebeuf ; Radio-Canada	10		

Lili Boisvert ; Radio-Canada	5
Radio-Canada avec Reuters, Agence France-Presse, CNN, La Presse Canadienne et Associated Press	4
Radio-Canada avec Associated Press et Reuters	3
Radio-Canada avec Agence France-Presse, Associated Press et Reuters	2
Radio-Canada avec Agence France-Presse, Associated Press, La Presse Canadienne, Reuters et CNN	2

Tableau 4.4 Résumé des catégories d'articles et leurs fréquences

Publication	**Nb. articles**	**Catégories uniques**	**Catégorie la plus fréquente**	**Fréquence**
Le Devoir	1492	53	Société	283
Le Journal de Montréal	1645	11	Nouvelles	1021
Le Journal de Québec	1372	16	Nouvelles	782
Journal Métro (Montréal)	768	27	Actualité	500
La Presse	1764	39	Actualités	788
SRC (site web)	980	1	Aucune	980
Le Soleil	1582	35	Actualités	875

Une étape de nettoyage a été effectuée seulement pour les données effectivement utilisées dans la présente étude, soit le titre de l'article, le contenu de l'article et la catégorie de l'article. Les deux premiers éléments font l'objet du prochain paragraphe. Un résumé de la métadonnée « Catégorie de l'article » est présenté dans le tableau 4.4 où on remarque une grande variété entre les différentes catégories d'article. Toutefois, les catégories qui contiennent la majorité des articles sont « Actualités » et « Opinion ». Cette variété est due à plusieurs erreurs de transcription qui font partie de la base de données mais aussi à des choix éditoriaux différents selon les sources d'information. Cependant, cette variété démontre également l'étendue de la couverture journalistique sur le printemps érable; ce sujet a été abordé en effet par plusieurs comités éditoriaux et dans plusieurs contenus journalistiques. Le tableau 4.5

affiche les dix catégories les plus fréquentes pour chaque journal[22]. Ces catégories ont été regroupées en trois grandes catégories, et sont : Actualité, Opinion, Autre.

Le nombre de mots par article indique également la présence de quelques erreurs contenues dans la base de données, comme des articles vides ou comportant moins de douze mots. En général, la moyenne des mots par article est plus grande pour *Le Devoir* et moins grande pour *Métro*, ce qui correspond à la nature même des sources d'information. Les deux journaux de Quebecor affichent les mêmes caractéristiques tandis que les deux journaux de Power Corporation présentent quelques différences. En effet, *Le Soleil* contient des articles moins longs que *La Presse*. Le tableau 4.6 présente par source d'information les résultats sur le nombre de mots contenus dans les articles[23].

Tableau 4.5 Catégories d'articles pour chaque source d'information

Le Devoir	**Nb. articles**	**Le Journal de Montréal**	**Nb. articles**	**Le Journal de Québec**	**Nb. articles**
Société	283	Nouvelles	1021	Nouvelles	782
Politique	252	Votre Opinion	259	Votre Opinion	191
Actualités	187	Arts Et Spectacles	125	Spectacles	96
Idées	182	Sports	72	Sports	84
Éditorial	124	Weekend	64	Opinions	77
Culture	113	Votre Argent	43	Weekend	49
Lettres	84	Votre Vie	40	Votre Argent	42
Économie	45	Autonet	8	Vie et Société	15
Libre opinion	40	Votre Maison	7	Votre Maison	10
International	34	Livres	5	News	9
Métro (Montréal)	**Nb. articles**	**La Presse**	**Nb. articles**	**Le Soleil**	**Nb. articles**
Actualité	500	Actualités	788	Actualités	875

[22] Le site web Radio-Canada n'a pas de catégorie et n'est pas inclus dans le tableau.

[23] Pour des questions de simplicité, le nombre d'espaces contenus dans les articles a été calculé afin de déduire le nombre de mots.

Monde	48	Débats	332	Éditorial	195
Opinions	46	Arts	121	Opinion	89
Carrières	43	La Presse Affaires	94	Sports	57
Culture	30	Politique	67	Arts et spectacles	51
Carrières	22	Monde	56	Nos régions	42
Éducation	15	CV	55	Politique	36
Week-end	10	Sports	43	Le samedi	34
Plus	6	Enjeux	28	Le Monde	31
Sports	6	Cinéma	21	Affaires	30

Tableau 4.6 Description des fréquences des mots dans les articles

Publication	**Moyenne**	**Écart type**	**Minimum**	**25e centile**	**Médiane**	**75e centile**	**Maximum**
Le Devoir	651.81	331.75	0	404	624	871.25	1925
Le Journal de Montréal	471.07	247.57	5	311	473	577	1897
Le Journal de Québec	481.23	239.02	11	345	477	559.25	1897
Métro (Montréal)	271.14	175.96	0	116	259	397	1177
La Presse	501.98	329.99	0	291.75	485.5	656.25	5580
SRC (site web)	423.59	313.79	2	212	344	550	2643
Le Soleil	414.11	342.29	0	189.25	393	565	4005

Comment peut-on évaluer cet ensemble d'articles construit grâce à cette procédure de recherche automatique? L'ensemble correspond au corpus de référence et possède un certain nombre de caractéristiques qui divergent de l'archive. D'abord, le corpus est certainement plus orienté, car il a été construit sur la base d'une problématique, le traitement journalistique du printemps érable, et d'une sélection d'articles tenant compte des dates, des sources d'information et d'un ancrage empirique par mot clé, selon le critère de saillance. La pertinence et la cohérence se situent à un niveau non optimal.

Le problème le plus important d'un corpus construit par moissonnage automatique et sur la base de mots clés est la portion de bruit qu'il inclut, c'est-à-dire la portion

d'articles qui, malgré qu'ils contiennent le mot clé, ne traitent pas du printemps érable et ne sont donc pas pertinents. Ces articles peuvent contenir le mot « étudiant » pour d'autres raisons. Un exemple d'article constituant du bruit est le cas « Magnotta », un jeune homme qui a tué un *étudiant* d'origine asiatique dans la même période qu'avait lieu le printemps érable. Les articles qui constituent du bruit sont nombreux et ils réduisent le niveau de pertinence, de cohérence et de représentativité du corpus. Pour ces raisons, dans la phase de filtrage de documents, le corpus de référence doit être analysé et nettoyé afin que la représentativité, la pertinence et la cohérence augmentent le plus possible (figure 4.3). Enfin, l'hétérogénéité et l'exhaustivité sont réduites par rapport à l'archive puisque certaines sources d'information ont été exclues. Toutefois, nous estimons que le fait d'avoir retenu les sources d'informations les plus importantes permet de maintenir un niveau élevé dans ces critères.

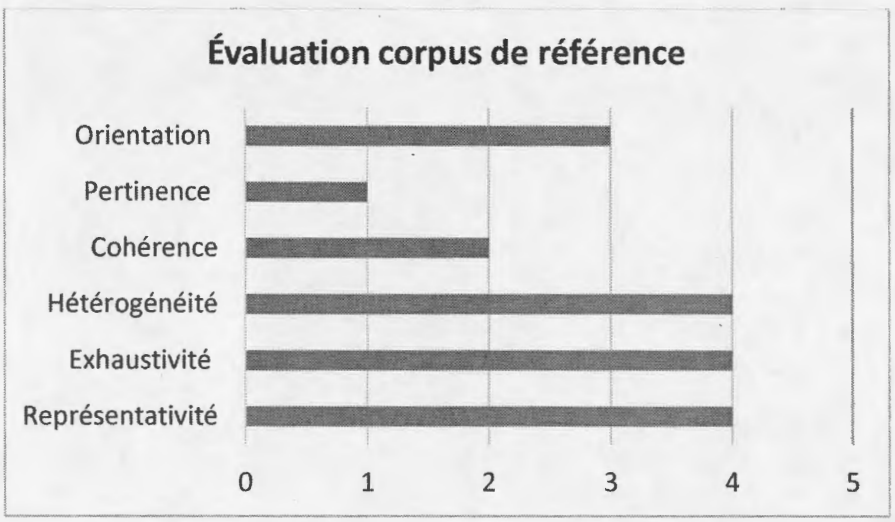


Figure 4.3 Évaluation du corpus de référence

4.2 Le prétraitement

Le prétraitement est exécuté sur les 9 603 articles qui composent le corpus. La première phase d'*annotation automatique* traite ces documents avec quelques outils issus du traitement automatique du langage naturel afin d'en dégager certaines caractéristiques proprement linguistiques. La *transformation vectorielle* produit une représentation vectorielle des documents, ce qui est fondamental pour le traitement computationnel et pour la réalisation des analyses au moyen des techniques de l'apprentissage automatique.

4.2.1 Annotation automatique

Les opérations de l'annotation automatique s'exécutent en parallèle. Pour ce faire, un outil développé dans un cadre probabiliste, tel que détaillé dans l'annexe A, est utilisé. L'outil s'appelle *TreeTagger* (Schmid, 1994, 1995) et il a été développé dans les années '90 afin d'améliorer les estimations des parties du discours réalisées avec des méthodes trigrammes. La spécificité de cette méthode demeure l'utilisation d'un *arbre de décision* pour l'estimation des probabilités de transition[24]. Ce dernier produit de règles plus complexes que les modèles bigrammes et trigrammes, comme par exemple la négation d'une occurrence (ex. $tag_{-1} = ADJ \wedge (tag_{-2} \neq ADJ) \wedge (tag_{-2} \neq DET)$). L'outil Treetagger s'effectue donc sur les résultats binaires de l'algorithme arbre de décision appliqué d'abord sur des corpus annotés. Une fois le modèle créé, l'algorithme applique, comme plusieurs annotateurs le font, l'algorithme de *viterbi* afin d'identifier une séquence d'annotations pour une séquence particulière

[24] Chaque état correspond à une étiquette morphosyntaxique et les probabilités de transition d'un état à l'autre sont les probabilités $P(t_i|t_j)$, et les probabilités d'émission d'un symbole sont $P(w_i|t_j)$, qui est la probabilité qu'un mot soit émis à partir d'une étiquette morphosyntaxique particulière. Consulter l'annexe A pour plus de détails.

de mots. Pour ce faire, l'algorithme utilise un vocabulaire qui contient pour chaque mot les probabilités *a priori* d'appartenance aux parties de discours. Le vocabulaire contient aussi les probabilités a priori pour chaque suffixe de mot (figure 4.4), ce qui représente une caractéristique importante pour identifier la partie du discours d'un mot.

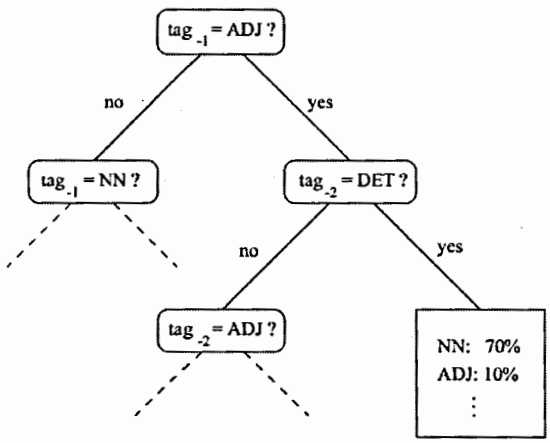


Figure 4.4 Exemple arbre de décision pour *Treetagger* (Schmid, 1994, p. 3)

La figure 4.4 montre un exemple de modèle construit par l'algorithme arbre de décision. La probabilité d'une annotation est déterminée en suivant le chemin le plus probable, dans ce cas défini par NN = 70 % à partir de deux annotations précédentes, composées par un adjectif à la position -1 et un déterminant à la position -2. Ceci signifie que, si l'algorithme viterbi trouve un « déterminant » suivi d'un « adjectif », il aura 70 % de chance d'annoter le prochain mot sous la catégorie « nom ».

Les différentes opérations exécutées pour identifier les parties du discours de chaque mot impliquent que la tokenisation et la lemmatisation aient également été accomplies. Ceci est possible grâce à un certain nombre de règles que Treetagger exécute et qui sont spécifiques à la langue traitée (figure 4.5). Plus particulièrement, l'outil utilise deux typologies d'information : une série de règles pour la tokenisation

et un vocabulaire qui contient les lemmes pour chaque mot. Si un mot n'est pas contenu dans le vocabulaire, il ne sera pas transformé en lemme.

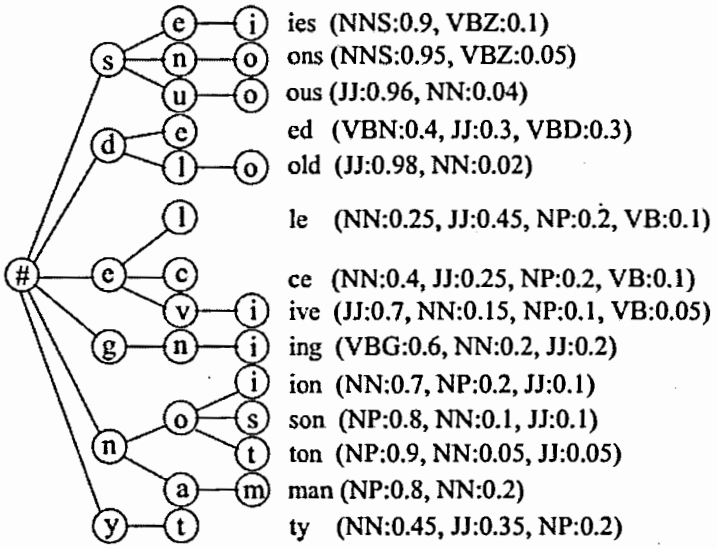


Figure 4.5 Exemple de probabilités pour les suffixes de la langue anglaise (Schmid, 1995, p. 5)

À la fin du processus, Treetagger retourne un certain nombre d'informations, dont la tokenisation, la lemmatisation et l'annotation des parties de discours. Dans l'annexe B, la liste des codes pour chaque annotation effectuée par la méthode est présentée.

4.2.1.1 Résultats

Les résultats de l'annotation sont résumés dans le présent paragraphe. La liste complète des annotations effectuées est contenue dans le tableau 4.7. Les différentes catégories ont été réunies dans de plus grands contenants, présentés dans le tableau 4.8. Par exemple, les six différentes typologies de pronoms ont été réunies en une seule catégorie, soit la macrocatégorie « PRO », correspondant aux pronoms de toute sorte. Pour l'explication des codes, consulter l'annexe B.

Tableau 4.7 Détails des parties du discours annotées

ABR	ADJ	ADV	DET: ART	DET: POS	INT	KON	NAM	NOM	NUM	PRO
PRO: DEM	PRO: IND	PRO: PER	PRO: POS	PRO: REL	PRP	PRP: det	PUN	PUN: cit	SENT	SYM
VER: cond	VER: futu	VER: impe	VER: impf	VER: infi	VER: pper	VER: ppre	VER: pres	VER: simp	VER: subi	VER: subp

Tableau 4.8 Nombre total par partie du discours en ordre décroissant

Partie du discours	**Moyennes par article**	**Totaux**	**% du total**
NOM	133,72	1283932	22,64%
PRP	90,77	871562	15,37%
VER	82,92	796203	14,04%
DET	64,34	617788	10,89%
PUN	46,22	443763	7,82%
PRO	45,28	434753	7,67%
ADJ	37,28	358007	6,31%
ADV	29,90	287069	5,06%
SENT	25,33	243190	4,29%
KON	21,76	208974	3,68%
NUM	11,14	106947	1,89%
ABR	1,03	9896	0,17%
SYM	0,61	5871	0,10%
INT	0,20	1956	0,03%
NAM	0,14	1312	0,02%

Le tableau 4.8 décrit des phénomènes attendus. Par exemple, les noms communs constituent la catégorie la plus fréquente, suivis par les prépositions, les verbes et les déterminants. La ponctuation et les pronoms apparaissent à un rythme similaire et les adjectifs sont plus nombreux que les adverbes. Une donnée particulière est celle des noms propres qui ne dépasse pas le 0,02 % ce qui est dû en partie aux erreurs de l'outil Treetagger. Les performances connues de cet outil ne dépassent pas le 97,53 % de F-mesure[25], mais ces prédictions sont optimistes pour le traitement de données

[25] Ce sont les résultats du module pour la langue allemande comme décrits dans

réelles. Enfin, on souligne qu'il y a 590,63 annotations pour chaque article en moyenne et un total de 5 671 223 annotations, correspondant au nombre de tokens identifiés. Le tableau 4.9 montre les résultats des annotations par journal. Pour chaque journal, on observe les mêmes proportions de la fréquence des catégories identifiées.

Tableau 4.9 Totaux des annotations par source d'information

Journal:	**PR**			**RC**			**SO**		
	moyennes	*totaux*	*% du total*	*moyennes*	*totaux*	*% du total*	*moyennes*	*totaux*	*% du total*
ABR	0,54	950	0,09%	1,56	1530	0,30%	0,94	1483	0,18%
ADJ	38,70	68259	6,14%	32,06	31422	6,17%	32,34	51130	6,23%
ADV	33,63	59322	5,34%	19,56	19173	3,76%	24,58	38865	4,73%
DET	68,52	120871	10,87%	58,93	57755	11,33%	56,14	88760	10,81%
INT	0,26	455	0,04%	0,08	83	0,02%	0,14	220	0,03%
KON	23,30	41093	3,70%	16,80	16461	3,23%	18,30	28933	3,52%
NAM	0,15	271	0,02%	0,13	123	0,02%	0,14	220	0,03%
NOM	140,35	247581	22,27%	128,68	126103	24,74%	120,84	191051	23,26%
NUM	10,98	19371	1,74%	11,12	10901	2,14%	12,26	19376	2,36%
PRO	49,69	87661	7,89%	30,69	30077	5,90%	36,50	57707	7,03%
PRP	95,31	168119	15,13%	88,02	86257	16,93%	80,70	127592	15,54%
PUN	49,14	86685	7,80%	40,94	40123	7,87%	41,61	65780	8,01%
SENT	28,78	50769	4,57%	19,31	18922	3,71%	23,69	37452	4,56%
SYM	0,87	1534	0,14%	0,44	427	0,08%	0,51	810	0,10%
VER	89,88	158541	14,26%	71,69	70255	13,79%	70,73	111819	13,62%
Journal	**JQ**			**DE**			**JM**		
	moyennes	*totaux*	*% du total*	*moyennes*	*totaux*	*% du total*	*moyennes*	*totaux*	*% du total*
ABR	1,20	1641	0,20%	1,22	1826	0,15%	1,05	1721	0,18%
ADJ	37,94	51984	6,30%	52,77	78728	6,63%	36,35	59797	6,28%
ADV	31,34	42969	5,21%	41,64	62124	5,24%	30,92	50862	5,34%
DET	64,09	87860	10,64%	88,76	132428	11,16%	61,60	101334	10,65%
INT	0,25	341	0,04%	0,21	318	0,03%	0,24	398	0,04%
KON	22,27	30533	3,70%	31,12	46429	3,91%	21,65	35610	3,74%
NAM	0,15	211	0,03%	0,16	235	0,02%	0,13	213	0,02%
NOM	135,01	185065	22,42%	176,58	263453	22,20%	128,48	211351	22,21%
NUM	13,15	18014	2,18%	11,02	16448	1,39%	10,70	17603	1,85%
PRO	48,48	66458	8,05%	62,45	93172	7,85%	48,03	79008	8,30%
PRP	89,41	122537	14,85%	124,33	185497	15,63%	85,75	141057	14,82%

Schmid (1995).

PUN	45,91	62991	7,63%	65,50	97720	8,24%	43,48	71532	7,52%
SENT	26,29	36041	4,37%	30,67	45757	3,86%	25,51	41959	4,41%
SYM	0,64	872	0,11%	0,70	1050	0,09%	0,58	946	0,10%
VER	86,02	117929	14,29%	108,20	161435	13,60%	84,13	138390	14,54%

Journal	**ME**		
	moyennes	*totaux*	*% du total*
ABR	0,97	743	0,28%
ADJ	21,64	16623	6,29%
ADV	17,86	13720	5,19%
DET	37,37	28702	10,86%
INT	0,18	141	0,05%
KON	12,89	9896	3,74%
NAM	0,05	39	0,01%
NOM	77,04	59166	22,38%
NUM	6,79	5211	1,97%
PRO	26,85	20618	7,80%
PRP	52,57	40372	15,27%
PUN	24,65	18928	7,16%
SENT	15,97	12263	4,64%
SYM	0,30	230	0,09%
VER	49,15	37749	14,28%

Le total de types annotés est 56 301[26], ce qui comprend toutes les catégories de mots. Pour chaque catégorie, la variété des lemmes est calculée et indiquée dans le tableau 4.10. Par exemple, les parties de discours « noms propres » sont celles qui ont une variété plus grande car le ratio annotation/lemme est le plus bas alors que les conjonctions et les signes de ponctuation présentent le moins de variété. Pour les adjectifs, la variété est plus grande que celle des adverbes.

Tableau 4.10 Résumé du ration token/type (annotation/lemme)

Partie du discours	**Nombre Lemmes différents**	**Nombre annotations**	**Ratio annotations/lemmes**
ABR	144	9896	68,72
ADJ	12883	358007	27,79
ADV	915	287069	313,74
DET	10	617788	61778,80

[26] Toutefois, si on ne considère pas la subdivision en parties du discours, les types sont 50 471, car un certain nombre de lemmes peuvent appartenir à plus qu'une catégorie.

INT	78	1956	25,08
KON	28	208974	7463,36
NAM	242	1312	5,42
NOM	34229	1283932	37,51
NUM	97	106947	1102,55
PRO	58	434753	7495,74
PRP	56	871562	15563,61
PUN	17	443763	26103,71
SENT	4	243190	60797,50
SYM	3	5871	1957,00
VER	7537	796203	105,64

La sélection des mots vides est réalisée par trois méthodes : à partir d'une liste préétablie de parties du discours, d'une liste préétablie de mots vides et de la fréquence documentaire. Les parties de discours qui correspondent aux adjectifs, aux adverbes, aux noms propres, aux noms communs et aux verbes sont retenues, tandis que les autres catégories ne le sont pas. Les tokens retenus totalisent 2 726 523 tokens, c'est-à-dire 48 % des occurrences totales. Bien que cette opération permette de diminuer de plus de la moitié des tokens, le nombre de types demeure presque invarié, soit 55 806[27] pour 99 % des types.

Par la suite, une liste préétablie comptant 466 mots est utilisée pour effacer des mots qui sont généralement considérés comme mots vides. Cette opération de filtrage mène à 2 229 145 tokens et à 49 950 types. Enfin, la fréquence documentaire est utilisée pour réduire les mots peu significatifs, mais elle est calculée pendant l'étape de représentation vectorielle des textes et permettra alors de réduire de presque 40 % le nombre de types.

[27] Sans tenir compte de la subdivision par catégorie, 50 144 types demeurent après le filtrage.

4.2.2 Représentation vectorielle

À la fin de l'annotation automatique, le corpus textuel est converti dans une matrice *Documents-Mots.* La matrice contient 9 602[28] documents et 29 615 types. Les types ayant une fréquence documentaire de moins de deux articles ont été supprimés. Ainsi la matrice *U* définie dans la formule (2.5) est construite. Dans cette matrice, les lignes correspondent aux articles, les colonnes aux lemmes, que dorénavant nous appellerons mots. À partir de cette matrice, il est possible de voir quels sont les mots les plus fréquents dans le corpus entier (tableau 4.11) et pour chaque journal (tableau 4.12).

Tableau 4.11 Les 20 mots les plus fréquents du corpus, en excluant le mot « étudiant »

Rang	Fréquence	Mot
1	25647	faire
2	18145	Québec
3	15423	pouvoir
4	11710	gouvernement
5	10839	devoir
6	10025	droit
7	9574	université
8	9262	aller
9	8515	Montréal
10	8177	an
11	7872	ministre
12	7859	grève
13	7323	vouloir
14	7189	Charest
15	7009	québécois
16	6822	dernier
17	6721	scolarité
18	6638	classe
19	6610	grand
20	6511	cour|cours

[28] Un document est vide et il a été supprimé.

Parmi les journaux, il y a peu de différence au niveau des mots les plus fréquents, ce qui ne révèle rien de spécifique ou caractéristique. En examinant les fréquences (tableau 4.12), on obtient une liste de mots les plus fréquents pour chaque journal. Ces listes sont très semblables et sont aussi similaires à celle qui correspond aux mots les plus fréquents du corpus en entier (tableau 4.11).

Tableau 4.12 Les 20 mots les plus fréquents par source d'information

RC

Rang	*Fréquence*	*Mot*
1	2630	Québec
2	1977	faire
3	1621	grève
4	1593	université
5	1517	Montréal
6	1428	droit
7	1299	manifestation
8	1281	gouvernement
9	1224	classe
10	1163	ministre
11	1105	pouvoir
12	1100	association
13	999	scolarité
14	989	cour|cours
15	974	devoir
16	957	loi
17	913	Charest
18	906	parti
19	897	cégep
20	845	hausse

PR

Rang	*Fréquence*	*Mot*
1	5192	faire
2	3088	pouvoir
3	2867	Québec
4	2379	gouvernement
5	2163	devoir
6	2145	droit
7	1842	an
8	1807	Montréal
9	1801	aller
10	1753	université
11	1505	vouloir
12	1489	grève
13	1417	grand
14	1379	voir
15	1342	dernier
16	1305	jeune
17	1296	classe
18	1280	cour|cours
19	1279	ministre
20	1252	année

JM

Rang	*Fréquence*	*Mot*
1	4624	faire
2	2601	pouvoir
3	2591	Québec
4	1785	gouvernement
5	1751	aller
6	1732	devoir
7	1471	Montréal
8	1383	vouloir
9	1380	an
10	1269	voir
11	1186	québécois

JQ

Rang	*Fréquence*	*Mot*
1	3992	faire
2	2714	Québec
3	2259	pouvoir
4	1649	gouvernement
5	1553	an
6	1534	devoir
7	1460	aller
8	1154	québécois
9	1154	vouloir
10	1113	ministre
11	1091	Charest

12	1182	université
13	1176	dernier
14	1124	Charest
15	1114	ministre
16	1095	prendre
17	1070	grand
18	1064	droit
19	1060	jeune
20	1028	hier

12	1034	Montréal
13	1023	voir
14	1011	université
15	986	grand
16	983	dernier
17	957	jeune
18	941	droit
19	923	année
20	920	prendre

SO

Rang	*Fréquence*	*Mot*
1	3386	faire
2	3308	Québec
3	2073	pouvoir
4	1598	gouvernement
5	1585	droit
6	1500	université
7	1382	aller
8	1353	devoir
9	1299	grève
10	1251	hier
11	1227	an
12	1184	Charest
13	1138	ministre
14	1052	cour|cours
15	1038	association
16	1034	scolarité
17	1022	vouloir
18	940	québécois
19	900	loi
20	896	prendre

DE

Rang	*Fréquence*	*Mot*
1	5189	faire
2	3551	pouvoir
3	3314	Québec
4	2569	devoir
5	2496	gouvernement
6	2398	droit
7	2068	université
8	1715	aller
9	1699	ministre
10	1593	québécois
11	1534	politique
12	1513	grand
13	1454	Montréal
14	1411	grève
15	1374	dernier
16	1358	loi
17	1331	vouloir
18	1314	Charest
19	1284	an
20	1265	scolarité

ME

Rang	*Fréquence*	*Mot*
1	1287	faire
2	745	pouvoir
3	720	Québec
4	519	gouvernement
5	514	devoir
6	472	Montréal
7	467	université
8	463	droit
9	385	aller
10	363	hier
11	362	ministre
12	349	scolarité

13	348	jeune
14	345	an
15	337	Charest
16	325	vouloir
17	312	grève
18	310	classe
19	301	cour|cours
20	297	grand

4.2.3 Filtrage

Les opérations de filtrage se divisent en deux : celles exécutées au niveau des mots et celles exécutées au niveau des documents. Généralement, les opérations sur les documents sont plus complexes que celles sur les mots, ce qui est le cas pour le présent travail. La plus grande complexité du filtrage des documents est due principalement aux problèmes liés à la manière dont le corpus est construit, ce qui introduit un niveau non négligeable de *bruit*. En effet, la recherche par mots clés ne peut pas distinguer les articles traitant du printemps érable de ceux qui traitent d'autres sujets, car, le mot « étudiant », qui a été utilisé comme mot clé pour la recherche automatique, peut être employé dans des articles couvrant différents sujets. Par exemple, le mot « étudiant » peut être utilisé dans des articles qui décrivent des événements criminels, comme le meurtre de l'étudiant chinois Jun Lin dans l'affaire « Magnotta ». La recherche par mots clés n'est donc pas suffisante pour discriminer les articles qui feront partie du corpus d'étude et une étape supplémentaire de filtrage est nécessaire. Une procédure complexe de filtrage de documents a donc été conçue pendant la phase des expérimentations.

4.2.3.1 Les mots

Le filtrage des mots est exécuté par le biais d'une opération simple, c'est-à-dire la fréquence documentaire est exécutée en même temps que l'étape de vectorisation. Ainsi, les mots qui n'apparaissent pas dans plus de deux articles ont été exclus, ce qui

mène à 29 615 types différents. À partir de cette nouvelle matrice, un indice de variabilité du vocabulaire entre les journaux est obtenu en considérant le ratio token/type de chaque journal. Dans le tableau 4.13, on remarque que le vocabulaire le plus varié est celui du *Journal Métro*, les autres ayant presque le même degré de variabilité.

Tableau 4.13 Variabilité du vocabulaire

Journal	Tokens	Type	Ratio tokens/types	Ratio types/tokens en %
RC	206572	9719	21,25445005	4,70%
PR	426398	19158	22,25691617	4,49%
DE	458271	18916	24,22663354	4,13%
JM	369420	21285	17,35588443	5,76%
JQ	321707	20603	15,61457069	6,40%
ME	102202	10595	9,64624823	10,37%
SO	319485	17206	18,56823201	5,39%

4.2.3.2 Les documents

Le filtrage sur les documents est exécuté en deux étapes différentes : une *procédure par règles* et une *procédure d'apprentissage semi-supervisé*. La première procédure a comme but l'élimination d'articles qui répondent à deux caractéristiques : articles doubles ou article contenant des mots clés spécifiques. Une série de règles basées sur des expressions régulières, sont créées afin d'éliminer les articles doubles. Un premier groupe de règles identifie les articles qui ont le même contenu, ceux qui ont le même titre et presque le même contenu et ceux qui sont vides. Le deuxième groupe de règles est constitué d'une seule règle qui vise à éliminer les articles qui contiennent le mot clé « Magnotta ». Cette procédure de type supervisé et construite sur une série de règles est efficace lorsqu'on connaît au préalable le type de bruit présent dans le corpus. Dans le cas du corpus de référence du présent travail, nous étions conscients du fait qu'il y avait des articles doubles et surtout du fait qu'un événement particulier concernant un *étudiant* avait reçu une grande attention de la

part des médias québécois dans la même période que le printemps érable, c'est-à-dire l'affaire « Magnotta ». Pour ces raisons, nous avons choisi d'effectuer ces deux procédures supervisées, ce que nous a permis de réduire le corpus de référence à 8929 articles. Les résultats de cette première étape de filtrage sont présentés dans le tableau 4.14.

Tableau 4.14 Résultats filtrage par règles

Source d'information	Contenant « Magnotta »	Double article	Double titre	Articles vide	Nombre d'articles retenus
DE	16	16	11	1	1448
JM	39	14	82	0	1510
JQ	28	12	83	0	1249
PR	33	10	57	0	1664
ME	15	30	2	0	721
SO	22	27	21	0	1512
RC	34	15	106	0	825

En principe, si toutes les typologies de bruit contenues dans le corpus étaient connues préalablement et si on pouvait construire des règles pour les identifier avec certitude, la procédure précédente serait suffisante pour nettoyer le corpus de référence. Bien que ce ne soit rarement possible, il est important d'identifier préalablement un certain nombre d'erreurs. Par la suite, d'autres types de procédures de nettoyage peuvent être accomplies, comme celle de type semi-supervisée qui a été utilisée dans le présent travail et dont l'algorithmique est représentée dans la figure 4.6. De manière générale, cette procédure est constituée de plusieurs étapes. La première est l'application d'un clustering basé sur la densité qui s'appelle *DBscan* dans le but d'identifier un certain nombre de documents constituant potentiellement du bruit et à analyser ses résultats afin de déterminer un véritable ensemble d'articles constituant du bruit. La deuxième étape est constituée d'une phase d'apprentissage, avec une sous-étape d'entraînement de test, afin de créer un modèle capable de distinguer les articles constituant du bruit des articles traitant du printemps érable. La dernière étape consiste à appliquer le modèle au corpus pour identifier le bruit et accomplir ainsi le filtrage de documents.

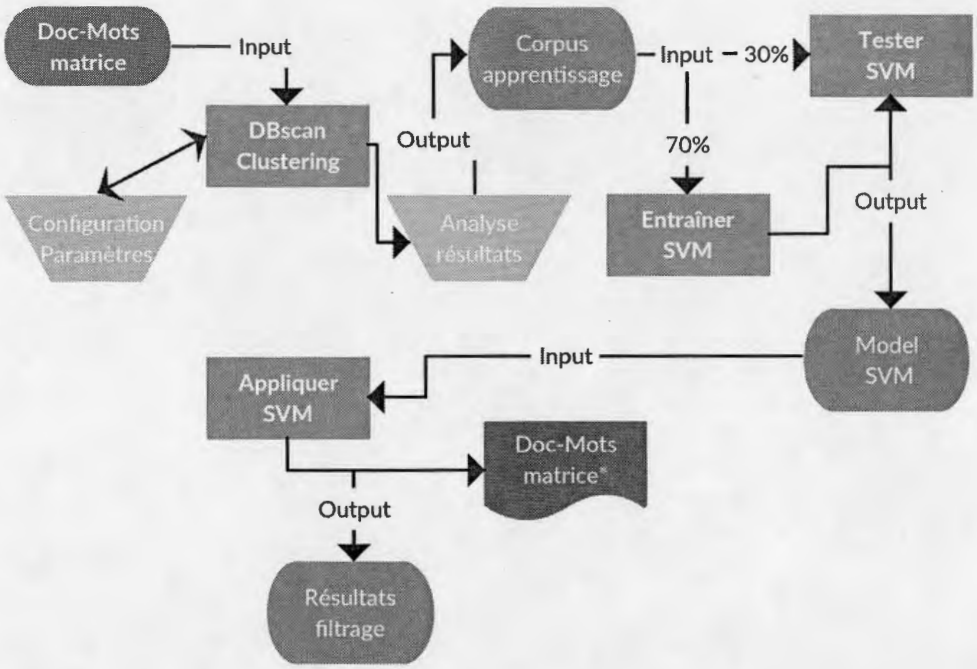


Figure 4.6 Procédure de filtrage des documents. Les *rectangles arrondis* constituent un bloc d'information d'**input** ou d'**output**. Le *rectangle* représente des **procédures algorithmiques**. Les *trapèzes inversés* constituent des **procédures non-automatiques** (intervention humaine). Le *rectangle à forme irrégulier* correspond à la **matrice *U*** (documents-mots) de laquelle le corpus d'apprentissage et le groupe de documents du filtrage par règles ont été soustraits.

De manière plus détaillée, la première étape (rectangle nommé « DBscan Clustering » dans la figure 4.6) consiste à utiliser un clustering basé sur la densité qui s'appelle *DBscan* (*Density-Based Spatial Clustering of Applications with Noise*), avec l'objectif d'identifier un premier groupe de documents constituant potentiellement du bruit. Consulter l'annexe C pour plus de détails. Pour atteindre l'objectif fixé, DBscan est exécuté plusieurs fois afin d'identifier la configuration des paramètres qui permet *d'obtenir un seul grand cluster* répondant aux paramètres de densité et *un cluster de bruit* « -1 » qui ne dépasse pas le tiers du corpus, mais qui soit assez volumineux. De cette manière, il est possible d'identifier un grand groupe d'articles assez proches entre eux et constituant 66 % du corpus. Le 33 % d'articles restants,

distants et éparpillés dans l'espace, ne répondent pas aux paramètres de densité et sont ainsi réunis sous le cluster du bruit « -1 ». Le cluster « -1 » constitue ainsi un groupe d'articles constituant potentiellement du bruit, alors que le cluster « 1 » constitue le groupe d'articles qui ont le plus de chance d'être des textes sur le printemps érable. Après plusieurs essais, les paramètres de densité ont été configurés ainsi : $\varepsilon = 0{,}8$ et *MinPts* = 1 320. La fonction de distance utilisée est le cosinus et la méthode des k plus proches voisins utilisés est basée sur une recherche par force brute[29]. À la fin du processus, le cluster « 1 » contient 6 204 articles et le cluster « -1 » contient 2 725 articles.

La deuxième étape (trapèze isocèle nommé « Analyse résultats » dans figure 4.6) est l'*analyse des résultats* de l'étape précédente et elle a comme but d'identifier un véritable groupe d'articles constituant du bruit, ce qui conduit à la construction du *corpus d'apprentissage*. Plus ou moins 10 % de chacun des deux clusters issus de l'application de DBscan a été sélectionné par une méthode d'échantillonnage aléatoire simple. Les tailles ont été arrondies pour obtenir deux groupes parfaitement proportionnés (1 versus 1/2) et nous obtenons ainsi 600 articles pour le cluster « 1 » et 300 articles pour le cluster du bruit « -1 ». La lecture et l'analyse de chacun de ces articles permettent de créer un corpus d'apprentissage constitué d'une classe « 1 », constitués d'articles qui traitent du printemps érable et d'une classe « -1 », comportant des articles qui ne traitent pas du printemps érable.

À la fin du processus d'analyse, nous avons obtenu une classe d'articles traitant du printemps érable contenant 550 documents et une classe d'articles constituant du bruit de taille 348. Deux articles ont été exclus du corpus d'apprentissage à cause de leur ambiguïté et 48 ont été transférés dans la classe « -1 ». Le corpus d'apprentissage,

[29] L'implémentation de l'algorithme utilisée est celle contenue dans le module scikit-learn de Python.

avec 989 articles, correspond environ à 10 % de la taille du corpus de référence, dont il est représentatif avec un intervalle de confiance de 3,1 % et un niveau de confiance de 95 %, c'est-à-dire que nous avons 5 % de chance de commettre une erreur avec un intervalle de 3,1 %.

La prochaine étape (rectangles nommés « Entrainer SVM » et « Tester SVM » dans la figure 4.6) consiste à entraîner et tester un modèle SVM linéaire (machine à vecteur de support), c'est-à-dire à lui faire « apprendre » à reconnaître le bruit et à le distinguer de la classe « 1 ». La technique SVM est utilisée pour des tâches de discrimination surtout dans des contextes de fouille de textes où son efficacité[30] a été démontrée. Le principe de base de cette technique consiste à identifier une fonction de discrimination linéaire f pour séparer de manière binaire les données, c'est-à-dire les distinguer en deux classes, la classe « 1 » et la classe « -1 ». La frontière de décision construite par l'algorithme est appelée l'hyperplan séparateur. Le but de l'algorithme est d'apprendre la fonction f et de construire l'hyperplan séparateur optimal pour le corpus d'apprentissage qui a été préalablement annoté avec les classes « 1 » et « -1 ». Il s'agit donc d'une méthode d'apprentissage supervisé permettant d'apprendre à reconnaître deux classes d'objets sur la base d'exemplaires de chacune de deux classes. Pour construire le modèle, le corpus d'apprentissage est divisé en deux parties : le 70 % est utilisé pour la phase d'apprentissage et 30 % pour la phase de test, c'est-à-dire 628 documents pour l'apprentissage et 270 documents pour le test. Le SVM a été appliqué plusieurs fois avec une méthode de validation croisée en dix sous-échantillons (10-fold-cross-validation). La validation croisée a été effectuée plusieurs fois pour déterminer la meilleure valeur du paramètre de pénalité C, qui a été fixé à une valeur de 6.25. Le test sur le modèle final a produit des résultats très satisfaisants, car la F-mesure (tableau 15) est de 97 %.

Tableau 4.15 Résultats du test SVM

	Précision	Rappel	F1-score	Nb. supports
Classe 1 (non-bruit)	0.97	0.99	0.98	152
Classe -1 (bruit)	0.98	0.96	0.97	118
Moy. ponderée	0.97	0.97	0.97	270

Le modèle final est utilisé dans la dernière étape (rectangle nommé « Appliquer SVM » figure 4.6) de cette procédure complexe de filtrage de documents, qui consiste à identifier dans le corpus de référence les articles qui, sur la base des résultats du SVM, correspondent à du bruit. Le SVM est appliqué à une Matrice Document-Mots (rectangle irrégulier de la figure 4.6) qui compte 8 031[31] lignes. À la fin de ce processus, on obtient les *résultats de la procédure de filtrage*, qui correspondent à une liste de valeurs binaires (1 ou -1) de la même taille que la matrice sur laquelle le modèle SVM est appliqué. Ce dernier processus identifie 2 228 articles constituant du bruit. Ces articles ont été ajoutés aux documents constituant du bruit utilisés dans le corpus d'apprentissage (348 articles), pour un total de 2 576 articles faisant partie de la classe « -1 », c'est-à-dire du bruit. Des 8 929 articles qui étaient restés après le filtrage par règle, on soustrait les 2 576 articles identifiés dans la phase de filtrage supervisé. Une dernière vérification manuelle sur le corpus a été accomplie,

---

[30] Dans le cadre de cette thèse, il n'est ni possible ni pertinent de présenter le fonctionnement de cette technique.

[31] De la matrice initiale *U*, qui comptait 9 603, le filtrage par règles a soustrait 674 articles. Par la suite, un corpus d'apprentissage de 898 articles a été construit pour l'entrainement de l'algorithme SVM. Ces documents ont été également soustraits de la matrice initiale, laquelle, pour l'application finale du SVM, compte 8 031 articles.

permettant d'identifier 77 articles supplémentaires comme étant du bruit. Les 6 276[32] *articles restants constituent notre corpus d'étude*.

Ainsi, l'étape de filtrage de documents se conclut avec la détermination du corpus d'étude, constitué de 6 276 articles traitant du printemps érable. L'évaluation de ce corpus est présentée à la figure 4.7. Évidemment, le critère d'orientation est celui qui augmente le plus, car le corpus d'étude, après le filtrage, répond très bien à l'objectif de la recherche. La pertinence et la cohérence augmentent également, car une grande partie de bruit issu de la méthode de collecte de données a été identifiée et retirée. La cohérence augmente moins que la pertinence, car il demeure une grande hétérogénéité de genre dans le corpus. De plus, la marge d'erreur du modèle SVM, autour de 3% (tableau 4.15), ne permet pas d'atteindre un niveau optimal. Le corpus demeure tout de même très hétérogène, exhaustif et représentatif par rapport à la question de recherche posée.

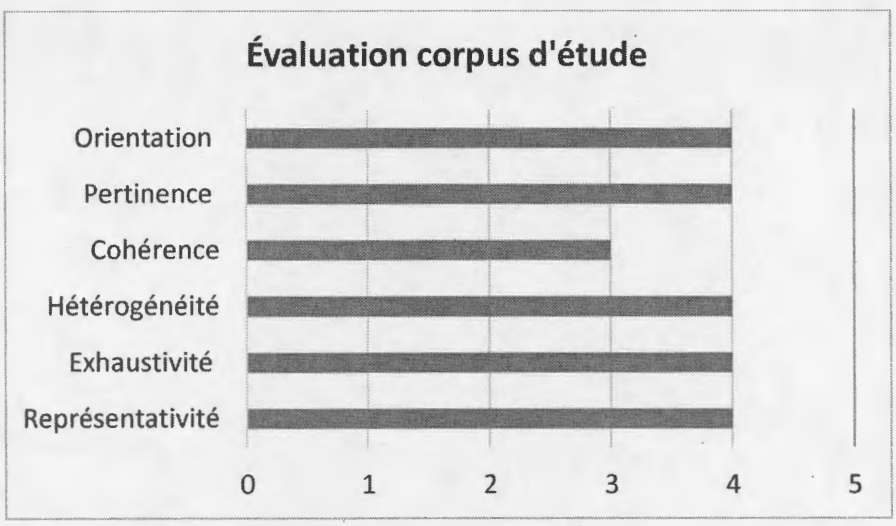


Figure 4.7 Évaluation du corpus d'étude

[32] Pour résumer, 9 603 – 674 (filtrage par règle) – 2 576 (filtrage supervisé) – 77 (vérification supplémentaire) = 6 276 articles constituant le corpus d'étude.

4.3 La définition du corpus de travail

Le corpus d'étude est soumis encore une fois au prétraitement afin de réduire le bruit résiduel au niveau des types propres aux documents filtrés. Essentiellement, la vectorisation est exécutée sur le corpus d'étude, obtenant une matrice de 6 276 lignes et 20 040 colonnes. Le corpus d'étude est composé ainsi de 6 276 articles, dont 1 110 de *La Presse*, 1 011 du *Le Devoir*, 1 072 du *Le Journal de Montréal*, 865 du *Le Journal de Québec*, 1 073 du *Le Soleil*, 460 du *Métro* et 685 du site web *Radio-Canada*.

La distinction par journal est utilisée pour obtenir *sept sous-corpus de travail différents*, ce qui mène à distinguer sept matrices différentes, c'est-à-dire une matrice par journal. Chaque sous-matrice possède un nombre de lignes égal au nombre d'articles contenus dans le journal et un nombre de colonnes égal au nombre de mots du corpus d'étude. Les étapes d'analyse seront alors effectuées sur les sept sous-corpus de travail de manière indépendante. Ce choix est basé sur l'idée que chaque source d'information possède ses spécificités dans le traitement des nouvelles et la comparaison entre ces sous-corpus est certainement facilitée si les analyses sont exécutées séparément.

4.4 Le regroupement des documents similaires

Le cœur du processus d'analyse est constitué par l'identification de groupes d'articles similaires. Cet objectif est atteint par une méthode qui appartient au paradigme non supervisé de l'apprentissage automatique, c'est-à-dire le *clustering*. Un algorithme spécifique, le *K-moyennes*, est utilisé pour atteindre les objectifs fixés. Tel que décrit dans le paragraphe 3.6.1, cet algorithme génère une partition contenant un nombre de clusters égal à la valeur du paramètre k. Son hypothèse de classification se base sur

l'optimisation de l'homogénéité intraclasse, ce qui implique de minimiser l'inertie intraclasse de chaque cluster. Concrètement, cet algorithme produit plusieurs groupes d'articles, chacun étant proche d'un *centroïde* c_k, ce dernier étant le centre géométrique du *cluster*.

Les résultats de ce type d'algorithme dépendent des paramètres de l'algorithme, soit la valeur du paramètre k, la fonction de distance d et de l'initialisation de l'algorithme. Ces trois paramètres correspondent aux trois premières étapes de l'algorithme (algorithme 3.1, dans paragraphe 3.6.1). La valeur k, étape 1, détermine le nombre de clusters à créer et elle doit être choisie avec attention. La fonction de distance, étape 2, correspond à la distance cosinus (formule (3.7)). L'initialisation, étape 3, détermine la composition des clusters et elle doit être exécutée de manière à limiter les biais de la méthode. La fonction de distance est le choix plus simple à faire. Pour les autres deux, il est important de déterminer un protocole. Le principe général de ce dernier est d'exécuter plusieurs versions de l'algorithme k-moyennes pour évaluer leurs partitions.

Pour des questions de lisibilité, nous présentons d'abord la méthode d'initialisation et ensuite le protocole pour choisir la valeur du paramètre k.

4.4.1 Initialisation de l'algorithme

La phase d'initialisation est exécutée par le biais d'une méthode appelée *k-means++* (Arthur et Vassilvitskii, 2007) qui a été développée pour résoudre partiellement les problèmes d'initialisation spécifiques à l'algorithme k-moyennes. Pour aborder ces problèmes, imaginons un jeu de données dont les clusters sont très bien identifiables, comme dans la figure 2.7. Dans un tel cas, il est suffisant d'utiliser une stratégie d'initialisation qui choisit les centroïdes de manière aléatoire, mais en les situant distants entre eux. Ainsi, les points c_k, soit les centroïdes, tendent à converger vers

les centres des clusters bien séparables. Dans un contexte où les points se distribuent de manière plus irrégulière et que des clusters bien séparables ne peuvent être construits, cette stratégie devient très sensible aux données aberrantes et elle est moins efficace.

Une autre approche pour l'initialisation est de déterminer une distribution des probabilités sur les possibles c_k initiaux et de sélectionner ceux qui obtiennent la plus grande valeur. La méthode *k-means++* étudie la probabilité que les points de l'espace soient sélectionnés comme prochain possible c_k. L'algorithme commence à créer un cluster de manière aléatoire et détermine ensuite au fur et à mesure les centroïdes qui sont les plus probables pour constituer un cluster. Par exemple, si dans un cluster il existe dix points et qu'un d'entre eux est vraiment loin des autres, alors les neuf points plus proches ont une plus basse probabilité d'être choisi comme prochains centroïdes. Sur cette base, la méthode exécute plusieurs tests et établit une liste de centroïdes les plus probables et détermine la meilleure initialisation sur la base des distances entre les points. En général, l'objectif de la méthode est d'éviter de situer plus d'un centroïde dans un groupe de points très condensé, réduisant ainsi le risque d'obtenir une partition de mauvaise qualité. L'implémentation algorithmique utilisée est contenue dans le module scikit-learn (Pedregosa *et al.*, 2011), utilisé par les concepteurs de la méthode.

4.4.2 Choix du paramètre *k*

Le paramètre *k* est choisi en évaluant plusieurs partitions obtenues avec une valeur *k* de 2 à max_k, ce dernier correspondant à la troncature à l'unité de $\frac{n_c}{10}$, où n_c est le nombre d'articles contenus dans chaque sous-corpus et le nombre entier du dénominateur est une constante. Cette dernière a été choisie en suivant le principe que, dans des conditions de distribution parfaitement symétrique des points, un cluster doit contenir au moins dix éléments. Donc, si on traite le sous-corpus de 1 011 articles du

Le Devoir, max_k correspondra à la troncature à l'unité de la fraction $\frac{1011}{10} = 101{,}1$, c'est-à-dire 101. Pour chacun des sous-corpus de travail, max_k a été ainsi calculé (tableau 4.16).

Tableau 4.16 Valeur max_k pour chaque source d'information

Source d'information	Valeur *max_k*
La Presse	111
Le Devoir	101
Le Journal de Montréal	107
Le Journal de Québec	86
Le Soleil	107
Métro	46
Radio-Canada (site-web)	68

Pour chaque sous-corpus, l'algorithme k-moyennes a été exécuté un nombre de fois égal à la quantité de nombres entiers contenus dans la séquence qui va de 2 à max_k. Pour *La Presse* par exemple, 109 applications de l'algorithme *k-moyennes* ont été exécutées et, à chaque application, la valeur du paramètre k a été augmentée d'une unité (ex. 2 + 1 = 3) et ce, pour la séquence qui va de 2 à 111. La première application a produit une partition avec deux clusters, la deuxième une partition de 3 clusters et ainsi de suite jusqu'à la dernière application qui a produit une partition de 111 clusters.

Pour chacune des partitions, un indice d'évaluation interne et un indice de distribution des articles par clusters a été calculé. *L'évaluation de la meilleure partition à retenir a été effectuée en tenant compte de ces deux indices*. D'un côté, l'*indice silhouette* (Rousseeuw, 1987) a été utilisé pour évaluer les partitions selon des critères de type statistique. Cet indice est un des plus utilisés parmi ceux qui basent l'évaluation de la qualité de la partition sur des critères internes. L'indice silhouette évalue la *cohésion interne* d'un cluster et sa *séparabilité* par rapport aux

autres, déterminant ainsi la qualité globale de la partition. Il est obtenu en calculant la moyenne de la distance entre un observable (article de journal) et tous les autres appartenant au même *cluster* et par la suite, la distance entre cet observable et ceux du cluster le plus proche. La formule de la fonction silhouette est la suivante $s = \frac{b-a}{\max(a,b)}$, où a est la moyenne de la distance entre un observable et tous les autres du même cluster et b, la moyenne de la distance entre le même observable et tous les points du cluster le plus proche. En d'autres termes, plus les clusters sont homogènes, compacts et bien séparables des autres, meilleure est la partition. Toutefois, l'indice silhouette, ainsi que les autres indices basés sur des critères d'évaluation interne, souffre du problème du *k+1*, tel que décrit dans le paragraphe (3.6.2) et il n'est pas possible de seulement s'y fier.

Pour déterminer la valeur *k* et donc la partition à retenir pour chaque sous-corpus, nous utilisons une heuristique basée sur le principe de *distribution symétrique* des articles dans chaque cluster. Selon ce principe, la meilleure partition est celle qui a un nombre d'articles similaire pour chacun de ses clusters, c'est-à-dire que les articles se distribuent de manière symétrique dans les clusters. En réalité, il est impossible d'obtenir une version parfaite de ce type de partition, car cela suppose que les éléments soient équidistants, ce qui n'est jamais le cas dans des situations réelles. Toutefois, le principe permet de mesurer la tendance des distributions et de choisir comme meilleure partition celle qui tend le plus vers une distribution symétrique des articles dans les clusters. Cette tendance peut être inférée à partir de l'écart type des distributions des articles dans les clusters d'une partition. En calculant l'écart type, il est possible d'obtenir un *indice de dispersion qui évalue la différence entre les clusters en termes de nombre de documents contenus*. Moins l'écart type d'une partition est grand, plus on a de chance d'obtenir une partition où ses clusters tendent à avoir le même nombre de documents et donc, tendent à avoir une taille similaire.

À la fin du processus (tableau 4.17), nous obtenons une partition de 34 clusters pour le tabloïd *Métro* (PubME), une partition de 54 clusters pour le Journal de Montréal (PubJM), une partition de 55 clusters pour le journal *Le Soleil* (PubSO), une partition de 54 clusters pour *Le Journal de Québec* (PubJQ), une partition de 50 clusters pour le site web *Radio-Canada* (PubRC), une partition de 55 clusters pour *La Presse* (PubPR) et, enfin, une partition de 55 clusters pour *Le Devoir* (PubDE).

Tableau 4.17 Valeur *k* choisie pour chaque sous-corpus.

Clustering	**Valeur K**
PubJM	54
PubSO	55
PubJQ	54
PubPR	55
PubDE	55
PubRC	50
PubME	34

Les figures qui suivent décrivent les résultats de l'indice silhouette et de l'indice de dispersion pour chaque source d'information. Pour chaque figure de 4.8 à 4.14, le **graphique à gauche** montre la valeur de l'indice silhouette (axe des y) pour chaque valeur *k* (axe des x). Les lignes verticales identifient quelques valeurs intéressantes. La ligne pointillée souligne la tendance de la fonction silhouette, qui a été calculée par une régression locale. Le **graphique à droite** montre la valeur de l'écart type (axe des y) pour chaque valeur *k* (axe des x). Le choix de la meilleure partition pour la phase d'annotation a été exécuté à l'aide des deux indices de qualité et en tenant compte de la nécessité d'obtenir des partitions comparables en termes de taille et de proportion.

Pour le Journal de Montréal, la partition à 54 clusters a été sélectionnée. Dans ce cas, l'indice de dispersion montre des valeurs très basses pour les partitions à 53 et à 54 clusters. L'indice silhouette présente une meilleure valeur pour celle à 54. Après

avoir analysé les résultats de ces deux partitions, celle à 54 a été choisie car les résultats sont plus significatifs[33].

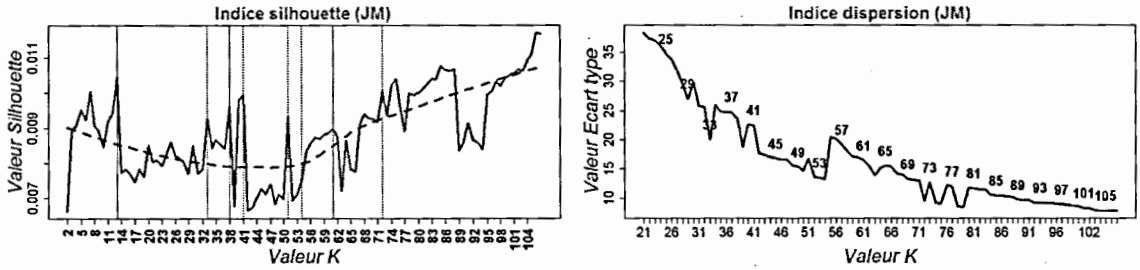


Figure 4.8 Indices en soutien du choix du paramètre k pour le Journal de Montréal.

Pour le journal *Le Soleil*, la partition à 55 clusters a été retenue. Pour ce journal, l'indice de dispersion montre des valeurs plus basses à partir de la partition à 55 clusters. L'indice silhouette indique des valeurs intéressantes autour de la partition à 51 clusters et autour de celle à 58 clusters. Après avoir étudié les résultats des partions à 51, à 55 et à 58 clusters, celle à 55 a été sélectionnée.

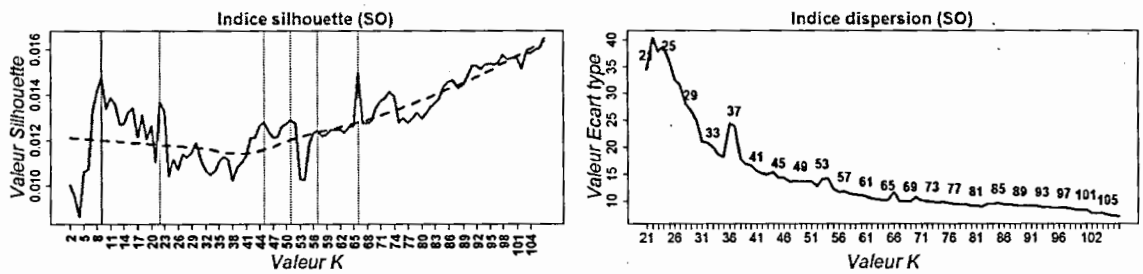


Figure 4.9 Indices en soutien du choix du paramètre k pour le journal Le Soleil

Pour le *Journal de Québec*, la partition à 54 clusters a été choisie. Pour ce journal, l'indice de dispersion indique la partition à 54 clusters parmi celles ayant une meilleure distribution des articles. L'indice silhouette présente des valeurs

[33] Dans ce contexte, ce terme n'est doit pas être entendu par son acception statistique, mais par celle plus commune.

intéressantes autour des partitions à 51 et à 55 clusters. Après l'examen des résultats de ces trois partitions (51, 54 et 55), celle à 54 clusters a été choisie.

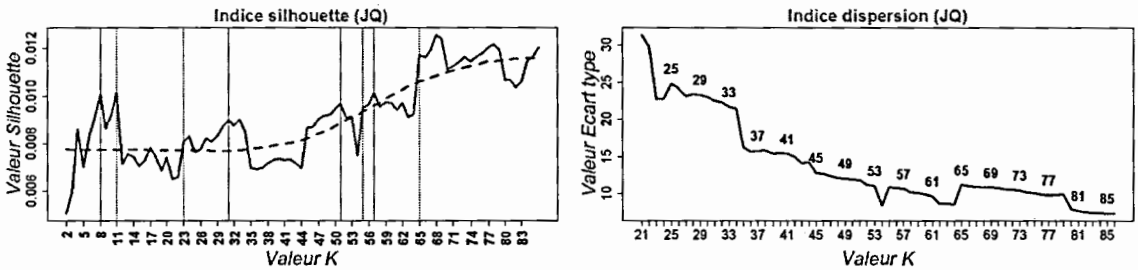


Figure 4.10 Indices en soutien du choix du paramètre k pour le Journal de Québec

Pour le *La Presse,* la partition à 55 clusters a été sélectionnée. Dans ce cas, l'indice de dispersion montre une chute de l'écart type pour la partition à 55 clusters, laquelle constitue une des meilleures partitions du point de vue de la distribution des articles dans les clusters. L'indice silhouette indique également une valeur haute pour cette partition, avec une valeur maximale pour la partition à 59 clusters. Après avoir analysé les résultats des partitions à 55 et à 59 clusters, la première a été retenue.

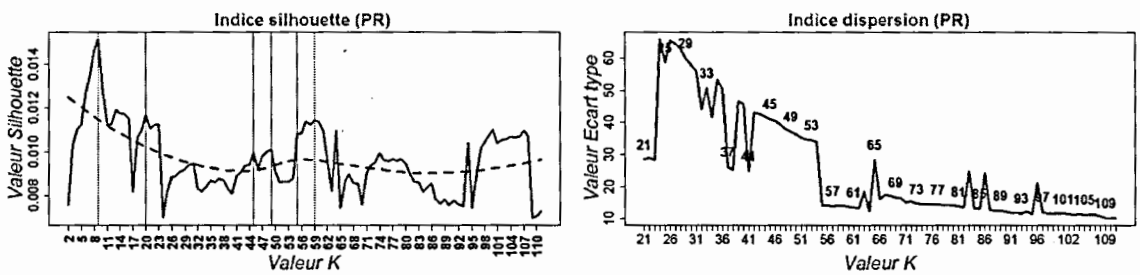


Figure 4.11 Indices en soutien du choix du paramètre k pour le journal La Presse

Pour le journal *Le Devoir,* la partition à 55 clusters a été choisie. Pour ce journal, l'indice de dispersion montre des valeurs plus basses à partir de la partition à 49 clusters. L'indice silhouette présente des valeurs plus grandes pour la partition à 35 cluster et à 55 clusters. Après l'examen des résultats des partitions à 35, 55 56 et 57, celle à 54 a été choisie car les résultats sont les plus significatifs.

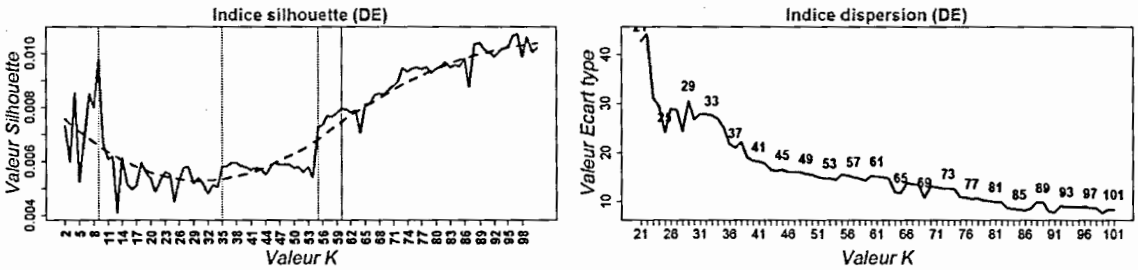


Figure 4.12 Indices en soutien du choix du paramètre k pour le journal Le Devoir

Pour le site web *Radio-Canada*, la partition à 50 clusters a été retenue. Pour cette source d'information, l'indice de dispersion montre des valeurs qui descendent progressivement et qui chutent autour de la partition à 46 clusters. L'indice silhouette indique une valeur montante au tour de la partition à 50 clusters. Après l'examen des résultats des partitions qui vont de 46 à 50 clusters, celle à 55 a été choisie car les résultats sont plus significatifs.

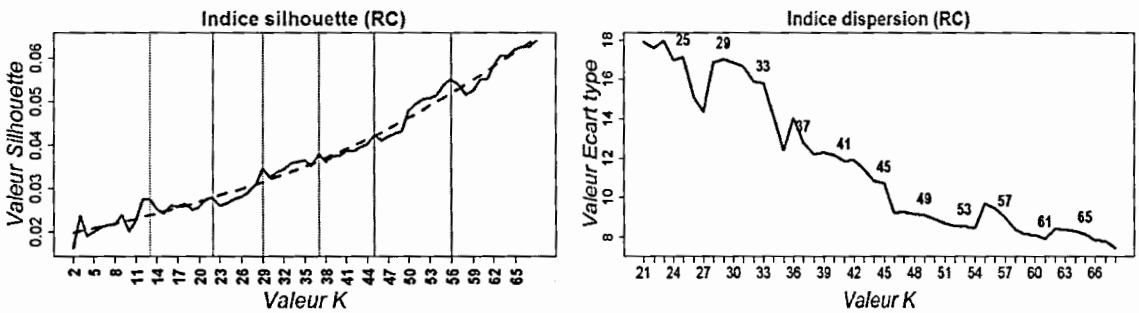


Figure 4.13 Indices en soutien du choix du paramètre *k* pour le site web *Radio-Canada*

Pour le journal *Métro*, la partition à 34 clusters a été sélectionnée. L'indice de dispersion présente des meilleures valeurs à partir de la partition à 30 clusters. L'indice silhouette présente deux valeurs intéressantes, une pour la partition à 32 et l'autre pour celle à 35. Après avoir étudié les résultats pour les partitions qui vont de 30 à 35, celle à 34 a été retenue pour l'analyse.

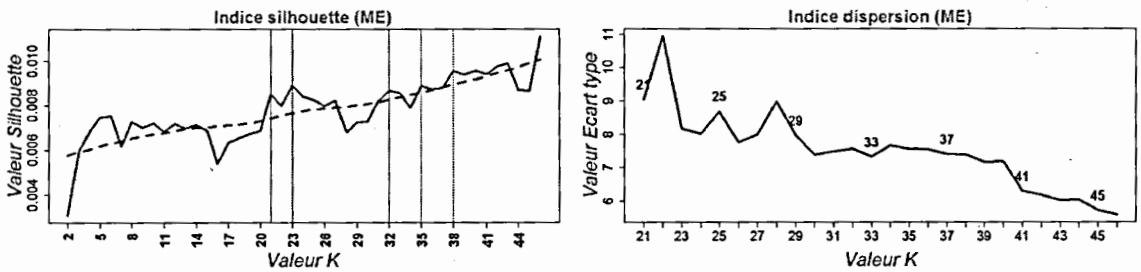


Figure 4.14 Indices en soutien du choix du paramètre *k* pour le journal *Métro*

4.5 L'annotation

L'analyse de chaque partition obtenue est réalisée à partir d'un protocole expérimental d'annotation. Cette étape vise à explorer une des différentes manières d'identifier les schémas récurrents des groupes d'articles (clusters) issus d'un clustering. Deux étapes sont réalisées, soit la sélection d'un échantillon représentatif pour chaque groupe et l'annotation de cet échantillon sur la base du schéma actanciel. La première étape constitue un élément important de la phase d'analyse et s'avère être aussi un élément crucial pour la justification des méthodes computationelle dans l'analyse sémiotique. En effet, la valeur ajoutée de l'utilisation de ces méthodes est la possibilité de construire des échantillons qui représentent des macrostructures du phénomène à l'étude, ce qui n'est pas réalisable avec des méthodes traditionnelles. La véritable annotation est donc exécutée sur un petit nombre d'articles et vise à identifier pour chacun d'entre eux le ou les programmes narratifs en jeu et les actants qui sont reliés. Pour chaque cluster, le niveau de cohérence entre les articles annotés sera alors évalué.

4.5.1 Sélection d'un échantillon représentatif

La chaîne de traitement développée permet de sélectionner un échantillon représentatif d'articles à analyser parmi les 6 276 retenus dans le corpus d'étude. Ceci est possible grâce aux caractéristiques spécifiques de l'algorithme de clustering choisi.

Un avantage majeur de l'utilisation du k-moyennes est la possibilité de considérer chaque centroïde c_k qui forme un cluster comme le centre d'un espace géométrique où un agglomérat d'articles est observé permettant ainsi de trouver un centre d'agrégation. Ainsi, les articles les plus proches du centre de l'espace seront ceux qui sont les plus similaires au document prototype du cluster, lequel est représenté par le vecteur du centroïde c_k . Lorsqu'on veut sélectionner un ensemble d'articles représentatif du cluster, il est possible de cibler l'ensemble des documents les plus proches du centroïde, ce qui correspond à un ensemble assez représentatif de l'agglomérat d'articles. Une simple heuristique des n documents les plus proches à c_k est alors utilisée.

Pour sélectionner cet ensemble d'articles, il est requis de calculer la similarité de chaque article par rapport au centroïde du cluster d'appartenance, ce qui est effectué par la fonction de similarité cosinus (formule (3.7)). À la fin du processus, les groupes d'articles sont triés selon leur proximité avec les centroïdes. Ensuite, on détermine une heuristique pour sélectionner l'ensemble le plus représentatif de l'agglomérat d'articles. Ainsi, nous avons choisi de suivre le principe des n documents les plus proches et de déterminer un seuil pour établir la taille de chaque échantillon. Le seuil est établi par la formule suivante : troncature à l'unité de $T_k = \frac{n_k}{3}$, où n_k correspond au nombre d'articles contenus dans chaque cluster et T_k est la taille de l'échantillon, ce qui implique que la taille de l'échantillon est fixée à un tiers des documents de chaque cluster. Pour chaque cluster k, on sélectionne les n articles les plus proches du centroïde c_k, où n est égal à T_k.

Ce processus est appliqué à chaque cluster de chaque partition, ce qui mène à un nombre assez important de clusters à analyser. En effet, le nombre total de clusters, toutes partitions confondues, est égal à 357. Afin de réduire ce nombre et de rendre les analyses plus efficaces, une heuristique est déterminée pour sélectionner la portion

des clusters la plus importante de chaque partition. Pour ce faire, on identifie une fonction de tri pour les clusters de chaque partition et une autre fonction qui détermine la taille de clusters à retenir dans la phase d'annotation. Les clusters sont donc d'abord triés selon le nombre de documents qui y sont contenus, c'est-à-dire que plus un cluster contient d'articles, plus il a de chance d'être retenu dans la phase d'annotation finale. La deuxième fonction établit le nombre de clusters à retenir. La formule est similaire à celle qui détermine l'échantillon de documents pour chaque cluster : troncature à l'unité de la fraction $T_c = \frac{n_c}{3}$, où n_c est le nombre de clusters dans la partition et T_c est la taille du groupe de clusters retenus. Cette formule implique donc la sélection d'un tiers des clusters de chaque partition. Enfin, pour chaque partition P, on retient les clusters les plus volumineux et ceci, pour un nombre de clusters égal à T_c.

Enfin, 117 clusters et 1 184 articles sont sélectionnés pour la phase d'annotation. Les détails sont présentés dans la figure 4.15.

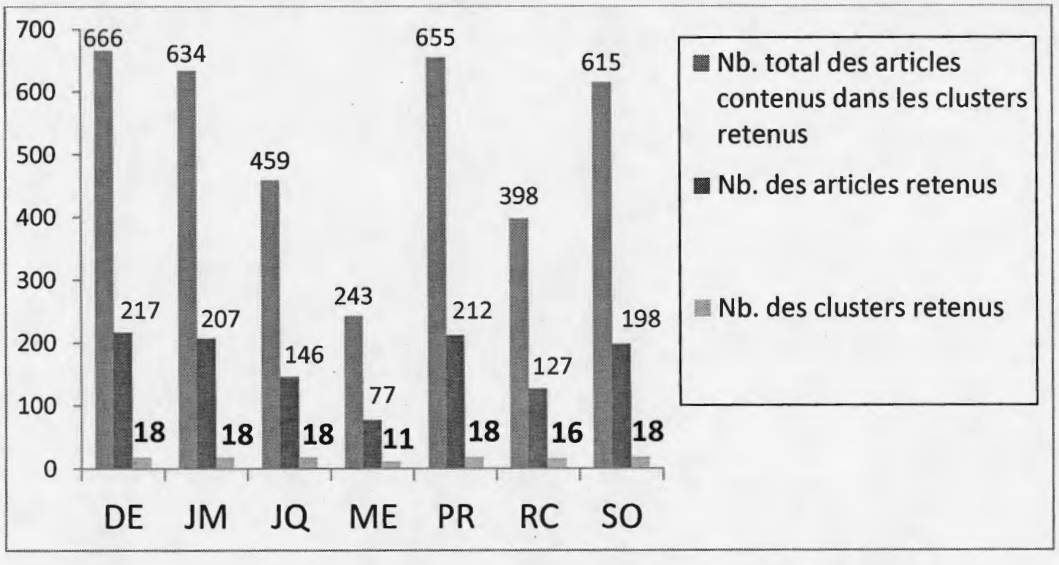


Figure 4.15 Résumé du nombre de cluster et d'articles retenus pour la phase d'annotation par source d'information

4.5.2 Annotation des échantillons

L'objectif principal de cette étape est de *déterminer un nombre minimal de schémas actantiels résumant un nombre maximal d'articles du même cluster*. Le modèle d'annotation retenu utilise des outils issus de la sémiotique greimasienne et se base sur l'hypothèse sémiotique déjà énoncée : *le niveau profond du sens dans le discours journalistique s'organise de manière narrative*. Les articles similaires partagent des structures narratives similaires. Sur cette base, un des modèles les plus simples de la théorie greimasienne est employé, soit le modèle actantiel. La procédure d'annotation, bien que baséee sur le schéma actantiel, est complétée par les catégories actantielles ajoutées par Hébert (2016). D'abord, l'annotation met en relief le ou les programmes narratifs de chaque article. Ensuite, les rôles actantiels qui y sont liés sont identifiés. De manière pratique, chaque article de l'échantillon est lu et annoté sur la base du modèle actantiel afin de relever les programmes narratifs en jeu et leurs actants et de souligner les similarités structurelles avec les autres articles. Le modèle de Greimas est utilisé comme ancrage théorique et méthodologique du protocole d'annotation et il balise l'interprétation des échantillons des *clusters*.

Chaque annotation est composée d'une *racine sémantique* et d'un *suffixe syntaxique*. La racine contient un mot ou une séquence de mots séparés par un tiret bas. Le suffixe contient une ou deux lettres en majuscule, séparées par un tiret bas. Le suffixe est le code qui identifie le rôle actantiel. Une liste des codes utilisés est présentée dans le tableau 4.18.

Tableau 4.18 Exemples de codes des annotations, avec leurs racines sémantiques, les suffixes syntaxiques et les rôles actantiel correspondant

Code Annotation	**Racine**	**Suffixe**	**Rôle actantiel**
élection_gagner_O	élection_gagner	O	Objet de valeur
gouvernement_S	gouvernement	S	Sujet
étudiant_AS	étudiant	AS	Anti-sujet
négociation_AD	négociation	AD	Adjuvant
violence_OP	violence	OP	Opposant

professeurs_AD_P	professeurs	AD_P	Adjuvant possible
éducation_DE	éducation	DE	Destinataire
troubler_paix_sociale_DR	troubler_paix_sociale	DR	Destinateur
universités_sous-financées_DR_P	universités_sous-financées	DR_P	Destinateur possible

4.5.3 Résumé résultats annotations

La phase d'annotation a été accomplie au moyen du « package » de R qui s'appelle RQDA (cfr. 3.7.2). Une base de données relationnelle a été créée pour archiver les passages de texte annotés et les codes utilisés pour l'annotation. Les tables qui ont été créées pendant la phase d'annotation et donc la structure de la base de données ne sont pas présentées. Nous nous limitons à souligner que chaque code possède un identifiant unique ainsi que chaque annotation, ce qui permet de retracer la totalité du processus de lecture et d'analyse. En général, il est possible de retrouver les segments de textes qui ont été annotés, l'article et la source d'information d'appartenance et le code de l'annotation. Il est ainsi possible de regrouper les annotations par source d'information ou catégorie de code.

Dans ce paragraphe, nous présentons quelques résumés des annotations par journal. Dans les tableaux qui suivent, les annotations correspondent aux segments de texte qui ont été concrètement annotés. Les codes correspondent à l'étiquette qui a été utilisée pour annoter le segment de texte. Le premier tableau présente les codes les plus fréquemment utilisés. Ensuite, les codes les plus fréquents par source d'information sont également illustrés.

Au total, les codes créés pendant la phase d'annotation sont au nombre de 259 et le total des annotations est de 4 311 pour 1 184 articles annotés. Le tableau 4.19 montre les codes les plus fréquents du processus complet de lecture. On y remarque que l'objet de valeur « gagner les élections » a été identifié avec le code « élection_gagner_O » et qu'il est le plus fréquent, suivi par le sujet

« gouvernement », l'objet de valeur « contre la hausse » et le sujet « associations étudiantes ».

Tableau 4.19 Les 20 codes les plus fréquents du processus de lecture et d'analyse des articles

Fréquence	Code
141	élection_gagner_O
140	gouvernement_S
133	hausse_contre_O
115	association_étudiant_S
107	étudiant_S
88	dialogue_O
86	grève_AD
81	manifestation_O
71	hausse_O
67	police_S

Fréquence	Code
63	manifestation_contrer_O
55	hausse_entente_O
52	négociation_AD
52	gouvernement_AS
51	P_Marois_S
50	crise_sortir_O
44	CLASSE_S
44	manifestation_AD
43	discréditer_adversaire_politique_O
43	manifestant_S

La fréquence des codes varie selon la source d'information. Pour le journal *Métro* (tableau 4.20), 688 annotations ont été exécutées avec 150 codes différents. Ce journal présente une plus grande majorité d'annotations que les autres journaux sur le « gouvernement » dans le rôle de sujet. Les deux autres codes les plus fréquents sont le sujet « association étudiante » et l'adjuvant « négociation ». Pour le *Journal de Montréal* (tableau 4.20), il y a 1 182[34] annotations exécutées avec 169 codes différents. À la différence des autres journaux, le *Journal de Montréal* présente quatre objets de valeur dans les quatre premiers rangs, soit « gagner les élections », « contre la hausse », « chercher le dialogue » et « éducation ».

[34] Avec le journal *Métro*, le *Journal de Montréal* a été le premier journal à être annoté. pour cette raison, ces deux journaux contiennent un plus grand nombre d'annotations par rapport aux autres. La réduction de l'attention et des performances des annotateurs au fur et à mesure que l'annotation procède est un phénomène bien connu dans le domaine de l'analyse du discours et de l'analyse du contenu. Nous ne traiterons pas ce sujet dans ce travail.

Tableau 4.20 Les 20 codes les plus fréquents pour le journal *Metro* (à gauche) et le *Journal de Montréal* (à droite)

Fréquence	Code pour *Métro*
41	gouvernement_S
35	association_étudiant_S
31	négotiation_AD
31	étudiant_S
29	police_S
25	hausse_contre_O
22	gouvernement_AS
22	manifestation_contrer_O
20	hausse_entente_O
18	manifestation_O
14	démocratie_DE
13	élection_gagner_O
12	carré_rouge_AD
12	CLASSE_S
11	dialogue_O
11	vandalisme_DR
10	étudiant_AS
9	grève_O
9	discréditer_adversaire_politique_O
9	arrestation_O

Fréquence	Code pour *Journal de Montréal*
48	élection_gagner_O
45	hausse_contre_O
43	dialogue_O
35	éducation_O
34	étudiant_S
31	gouvernement_S
31	manifestation_O
28	grève_OP
28	grève_AD
26	hausse_O
24	paix_O
23	association_étudiant_S
21	retour_en_classe_O
20	université_S
19	manifestation_contrer_O
19	cours_boycottage_OP
18	gouvernement_AS
18	mobilisation_O
18	cégep_S
17	session_annulation_AD

Le *Journal de Québec* (tableau 4.21) présente 460 annotations pour 101 codes. Au premier rang, on trouve l'objet de valeur « gagner les élections », suivi par trois sujets, dont une politicienne, Mme Pauline Marois. C'est le seul journal qui présente un acteur politique dans ses premiers rangs, c'est-à-dire un politicien spécifique. Le site web *Radio-Canada* (tableau 4.21) présente 378 annotations en 67 codes. L'objet de valeur « gagner les élections » est le code le plus fréquent, comme dans le cas des deux journaux du groupe Quebecor. Les sujets « associations étudiantes » et « étudiants » et l'objet de valeur « contre la hausse » le suivent. À la différence des autres journaux, le site de Radio-Canada semble accorder plus d'attention aux actions accomplies par les étudiants, vus comme un acteur politique généralisé, et les

associations étudiantes, considérées comme un acteur politique spécifique et identifiable.

Tableau 4.21 Les 20 codes les plus fréquents pour le *Journal de Québec* (à gauche) et le site-web *Radio-Canada* (à droite)

Fréquence	Code pour *Journal de Québec*	Fréquence	Code pour *Radio-Canada*
26	élection_gagner_O	28	élection_gagner_O
21	étudiant_S	25	association_étudiant_S
19	P_Marois_S	23	hausse_contre_O
18	gouvernement_S	23	étudiant_S
18	grève_AD	22	grève_AD
16	crise_sortir_O	19	manifestation_AD
12	hausse_contre_O	17	gouvernement_S
11	dialogue_O	13	retour_en_classe_O
10	association_étudiant_S	13	P_Marois_S
10	discréditer_adversaire_politique_O	12	dialogue_O
10	retour_en_classe_O	10	Police_S
9	manifestation_O	10	manifestation_contrer_O
8	paix_O	8	discréditer_adversaire_politique_O
8	police_S	8	J_Charest_S
8	étudiant_AS	8	arrestation_AD
8	outils_oppression_AD	8	cégep_S
8	PLQ_S	7	mobilisation_O
7	loi_78_OP	6	négotiation_AD
7	recours_tribunal_déposer_AD	6	gouvernement_AS
7	hausse_O	6	crise_sortir_O

Pour le *Devoir* (tableau 4.22), 469 annotations ont été accomplies en 92 codes différents. Le code le plus fréquent souligne la présence d'un article ou d'un segment dans l'article qui constitue une opinion et qui n'ont pas pu être annotés de la même manière que les autres articles ou segments. Nous parlerons de cet élément aux points 5.1.7 et 6.2.7. Le *Devoir* est également le journal qui possède le plus d'annotations sur le « débat politique et culturel » comme objet de valeur. La lecture et l'analyse des articles de *La Presse* (tableau 4.22) ont généré 529 annotations avec 92 codes. Ainsi comme pour le journal *Le Devoir*, La Presse comporte un grand nombre d'articles ou de segments qui constituent des opinions. C'est le journal qui contient,

par rapport aux autres, le plus grand nombre d'articles ou de segments sur les objets de valeur « manifestation », « sortir de la crise » et « retour en classe ».

4.22 Les 20 codes les plus fréquents pour *Le Devoir* (à gauche) et *La Presse* (à droite)

Fréquence	Code pour *Le Devoir*
47	opinion_difficile_à_annoter
20	hausse_contre_O
18	grève_AD
18	débat_politique_culturel_O
17	éducation_DE
16	crise_sortir_O
15	loi_78_contre_O
13	élection_gagner_O
13	retour_en_classe_O
12	association_étudiant_S
12	dialogue_O
12	police_S
12	manifestation_contrer_O
12	manifestation_O
11	cégep_S
10	gouvernement_S
10	société_DE
10	CLASSE_S
10	manifestation_AD
9	étudiant_S

Fréquence	Code pour *La Presse*
33	opinion_difficile_à_annoter
31	crise_sortir_O
29	manifestation_O
29	retour_en_classe_O
28	élection_gagner_O
19	manifestant_S
14	outils_oppression_AD
14	hausse_O
14	grève_OP
13	hausse_entente_O
13	gouvernement_S
13	J_Charest_S
12	étudiant_S
12	grève_AD
11	Police_S
11	manifestation_contrer_O
11	casseroles_AD
10	police_AS
10	cégep_S
9	hausse_contre_O

Pour le journal *Le Soleil* (tableau 4.23), 625 annotations ont été exécutées avec 112 codes différents. Les codes les plus fréquents sont les sujets « gouvernement » et « association étudiante », qui sont accompagnés par les deux objets de valeur qui répondent le plus aux attentes de la recherche, c'est-à-dire « pour la hausse » et « contre la hausse ».

4.23 Les 20 codes les plus fréquents pour *Le Soleil*

Fréquence	Code pour *Le Soleil*
40	gouvernement_S
31	hausse_contre_O

Fréquence	Code pour *Le Soleil*
15	dialogue_O
14	J_Charest_S

30	association_étudiant_S
28	hausse_O
25	élection_gagner_O
20	manifestation_O
20	grève_AD
17	crise_sortir_O
16	hausse_entente_O
16	manifestant_S

13	grève_O
13	loi_78_O
12	négotiation_AD
12	CLASSE_S
12	Beauchamp_S
10	loi_78_contre_O
10	prets_bourse_AD
9	P_Marois_AS

Globalement, la donnée qui a attiré le plus notre attention est la fréquence de l'objet de valeur « gagner les élections ». On peut remarquer en effet qu'il constitue un code très fréquent dans toutes les sources d'information et que pour certaines, il est le plus fréquent. Enfin, cet objet de valeur est le code le plus fréquent de manière absolue (tableau 4.19). Avant de commencer la phase de l'annotation, nous n'imaginons pas rencontrer autant d'articles ou de segments mettant de l'avant cet objet de valeur. En effet, le programme narratif lié à la course électorale est un des sujets plus répandus et se mêle de manière directe ou indirecte avec la question étudiante. Nous analyserons ces résultats dans le prochain chapitre.

CHAPITRE V

RÉSULTATS

L'analyse des résultats tire profit du travail accompli lors des phases de regroupement automatique et d'annotation. Dans la première, un algorithme spécifique a permis d'organiser les articles de chaque source d'information en groupe d'articles similaires. La phase d'annotation a ensuite permis d'annoter les articles les plus importants d'une sélection de clusters et ceci, afin d'identifier les macrostructures qui organisent leur contenu. En raison du nombre de clusters sélectionnés, soit 117, et du nombre d'articles annotés 1 184, la présentation et l'interprétation des résultats ne peuvent pas être exhaustives. Afin d'assurer la lisibilité des résultats obtenus, il est indispensable de disposer d'un *cadre synthétique*. Comme le précise Rastier lorsqu'il décrit le processus herméneutique : « toute interprétation suppose une stratégie d'analyse qui précise les tactiques à employer, et garantit la pertinence des éléments retenus » (Rastier *et al.*, 1994). Ainsi, nous présenterons les résultats en utilisant une stratégie de synthèse basée sur l'*analyse des réseaux*.

Toutefois, cette technique ne sera pas exploitée entièrement car l'analyse des réseaux est une technique complexe, dont il faudrait détailler le cadre théorique, soit la théorie des réseaux, et les algorithmes utilisés dans ce cadre (Ahuja *et al.*, 1993). Son application en analyse de textes constitue un fait acquis depuis plusieurs années (Franzosi, 2010). Dans ce contexte, nous *utiliserons cette technique de manière*

intuitive et en exploitant ses *propriétés de visualisation*. L'outil Gephi[35] (Bastian *et al.*, 2009) sera utilisé à cet effet.

La figure 5.1 illustre tous les clusters et leurs liens. Chaque nœud du graphique constitue un cluster et chaque lien correspond à la mesure de similarité cosinus entre la paire de clusters connectée. Les liens les plus forts ont une couleur plus foncée et les nœuds qui correspondent à des clusters plus volumineux sont, eux aussi, de couleur plus foncée. La position de chaque nœud dépend de la force de ses liens. Un tel graphique étant difficile à utiliser à des fins d'analyse et d'interprétation, nous avons calculé un nouveau graphique (figure 5.2) au moyen de l'algorithme Force Atlas 2 et ce, en fixant à 0,77 le seuil permettant de filtrer le poids des liens (Jacomy *et al.*, 2014)[36]. Il est possible d'y distinguer des agglomérats de clusters connectés entre eux. On y observe aussi que la majorité de ces agglomérats contient des clusters de grande taille (couleur du nœud plus foncée).

[35] Pour plus de détails, voir le site web officiel : https://gephi.org/

[36] Les paramètres par défaut de l'algorithme ont été utilisés avec l'ajout du mode *LinLog* qui permet d'amplifier la distance entre les nœuds qui n'ont aucun lien.

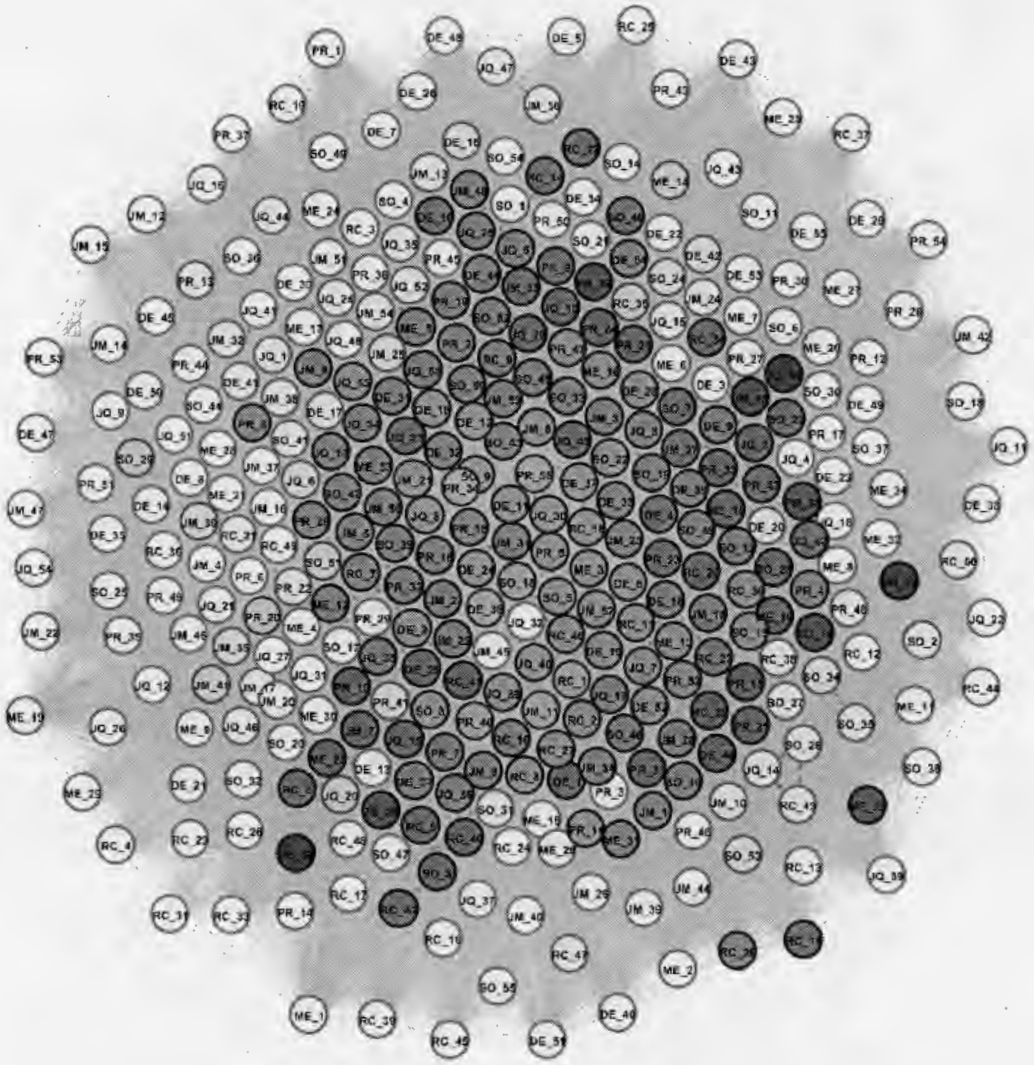


Figure 5.1 Réseau global de tous les clusters créés dans la phase de regroupement automatique des documents similaires. Chaque lien entre nœuds correspond à la valeur cosinus.

La première partie de ce chapitre est dédiée à la présentation des résultats qui se base sur le dernier graphique. En particulier, nous utiliserons des sous-espaces de ce graphique pour présenter de *grands agglomérats de clusters*. Ces derniers seront analysés et interprétés comme des macro-groupes de clusters identifiant une grande catégorie de macrostructures. Les macrostructures que ces agglomérats véhiculent

sont les plus importantes du corpus. En effet, d'un côté elles sont les macrostructures qui regroupent le plus de clusters à un seuil cosinus élevée. De l'autre, les clusters impliqués dans ces agglomérats sont parmi les plus volumineux. En d'autres termes, ces macro-groupes de cluster regroupent un nombre très grand d'articles traitant les mêmes macrostructures, ce qui nous mène à considérer ces macrostructures les plus importantes du corpus. Enfin, ces macrostructures peuvent être considérées comme constituant les *grandes narrations du traitement journalistique du printemps érable*.

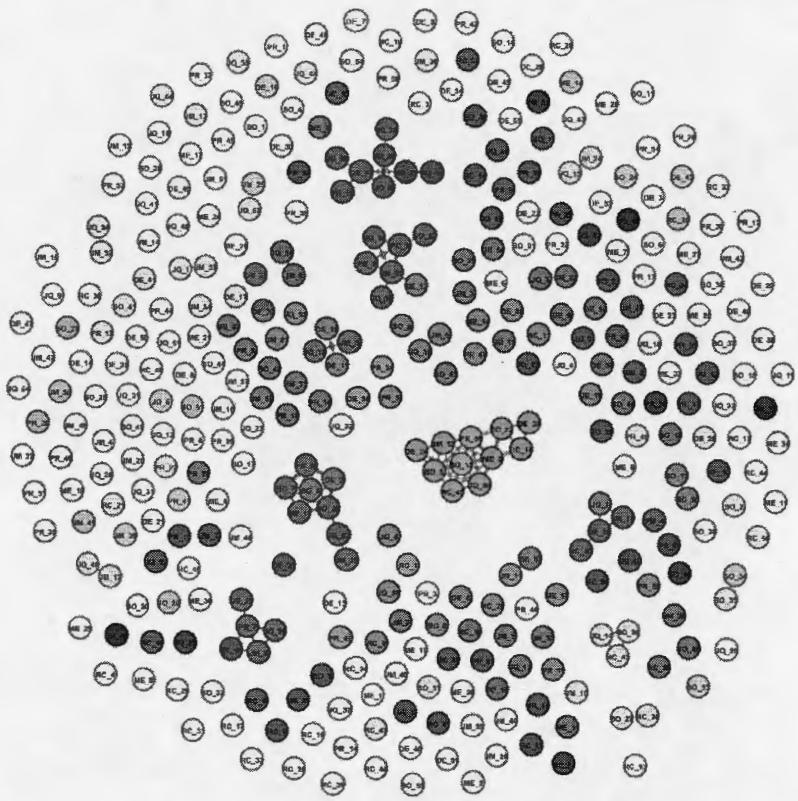


Figure 5.2 Représentation graphique de tous les clusters avec un filtre des liens fixé à 0,77 et l'application de l'algorithme Force Atlas 2

La deuxième partie de ce chapitre est dédiée, au contraire, aux clusters qui se situent aux marges du graphique et qui sont susceptibles de représenter des caractéristiques

uniques et distinctives de chaque journal. Pour présenter ces résultats de nouveaux graphiques seront calculés.

Nous utiliserons le lexique sémiotique appartenant au métalangage greimassien, tel qu'il a été présenté dans les précédentes sections. En général, les *actants* seront indiqués en *italique*, alors que les **acteurs** ou **figures** sont présentés **en gras**. Dans l'annexe D, le code identifiant de manière unique les articles est présenté. Plusieurs extraits d'articles seront présentés à l'intérieur de ce chapitre. Le chapitre est ainsi divisé en deux. La première partie est dédiée à l'analyse et à l'interprétation des grands agglomérats d'articles alors que la deuxième partie est dédiée aux traits distinctifs de chacun des journaux retenus dans cette étude.

5.1 Groupes de macrostructures similaires

La phase de regroupement automatique a été exécutée à l'aide du clustering. Les caractéristiques de l'algorithme utilisé, le k-moyennes, permettent d'identifier un *vecteur prototype* pour chaque cluster, soit le *centroïde*. Ainsi, les centroïdes ont été projetés dans un espace commun et la similarité cosinus a été calculée pour chaque couple de clusters. Les résultats ont été utilisés pour créer un graphique, où *chaque nœud représente un cluster* et *chaque lien représente la similarité entre deux nœuds*. La *force du lien est la valeur cosinus*. Ainsi, plus le lien entre deux nœuds est fort, plus les clusters correspondants sont similaires. Le lien le plus fort correspond à une valeur cosinus de 0,89, alors que le moins fort a une valeur de 0,004.

Il est donc possible de *filtrer les liens* selon leur force. Ceci signifie que ce ne sont pas tous les liens qui sont représentés dans le graphique mais seulement ceux qui ne dépassent pas un seuil établi à priori. Ceci permet de construire un graphique qui met en évidence les liens les plus importants tout en le rendant plus lisible et plus

facilement interprétable. Quand le seuil est élevé, les liens qui restent dans le graphique sont les plus forts, alors que, lorsqu'il est plus bas, les liens restant dans le graphique sont plus nombreux, affectant ainsi la lisibilité. À un seuil égal à zéro, chaque nœud est connecté à tous les autres nœuds.

Afin d'identifier des agglomérats de cluster, nous avons analysé des graphiques construits avec un filtrage des liens à des seuils différents, soit de 0,6 jusqu'à 0,77. Plus particulièrement, un certain nombre d'agglomérats de cluster a été identifié dans une échelle de 0,7 à 0,77. Les agglomérats identifiés sont au nombre de sept et seront présentés dans les prochains paragraphes. Ils représentent les macrostructures qui ont dominé le traitement journalistique du printemps érable.

5.1.1 Négociations : théâtre de la polémique

À partir du graphique de centroïdes avec un seuil de 0,77 (figures 5.3 et figure 5.4), le centre d'agrégation le plus peuplé est situé au milieu de l'espace. Dans cet agglomérat, 11 clusters sont liés entre eux avec une similarité cosinus entre 0,77 et 0,85. Dans ce sous-réseau, JQ_40, ME_3 et PR_55 possèdent un nombre plus grand de liens (7 par cluster). *Le Soleil* est présent avec trois clusters (SO_5, SO_15 et SO_22), *Le Devoir* avec deux (DE_24 et DE_37), *Radio-Canada* avec deux (RC_18 et RC_46) et les autres sources avec un seul (JM_52, JQ_40, ME_3 et PR_55).

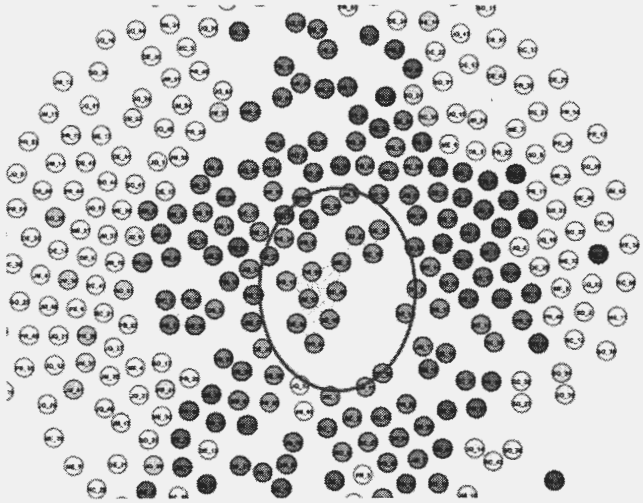


Figure 5.3 : Détail de la figure 5.2. Encerclé en rouge les 11 clusters sélectionnés

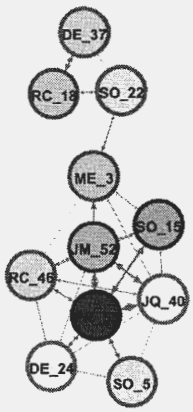


Figure 5.4 : Représentation graphique de l'agglomérat des 11 clusters centraux. Les clusters plus foncés désignent un cluster plus volumineux

Le cœur narratif de cet agglomérat est celui des *négociations entre les associations étudiantes et le gouvernement*. Parmi les différents programmes narratifs, le **PN1** (tableau 5.1) s'en dégage très nettement et constitue le pivot central autour duquel les autres se développent : les **associations étudiantes** (*actant-sujet*) veulent obtenir un **dialogue** (*actant-objet*) avec le **gouvernement** (*actant-antisujet*). Les trois actants de

l'axe du vouloir sont souvent représentés au niveau discursif par les acteurs **associations étudiantes**, **dialogue** et **gouvernement** mais d'autres acteurs peuvent apparaître également, comme **étudiants**, **FEUQ**, **CLASSE**, **ministre Beauchamp**, **Charest**, **négociations** etc.

Tableau 5.1 : Programme narratif PN1.

Sujet	*Objet de valeur*	*Antisujet*
associations étudiantes	**dialogue**	**gouvernement**

Ce programme narratif apparaît majoritairement dans la phase initiale de la grève. Dans les extraits qui suivent, on remarque également le rôle d'*antisujet* du **gouvernement** qui se montre fermé quant à la possibilité de mettre sur pied une table des **négociations** :

> La CLASSE, la FEUQ et la FECQ ont réitéré qu'elles étaient ouvertes au dialogue avec le gouvernement. (K03ME0322Act0137)

> Le gouvernement fermé. Alors que quelque 200 000 personnes ont signifié leur appui envers les étudiants en défilant dans les rues de Montréal, jeudi après-midi, aucune discussion n'a eu lieu entre les étudiants et le gouvernement, ont rappelé les associations étudiantes. Devant cette fermeture de Québec, les étudiants prévoient une escalade des moyens de pression dans les prochains jours, afin de forcer le gouvernement à entamer des négociations. (K52JM0324Nou0268)

> L'offre de Québec rejetée. La CLASSE, la FEUQ et la FECQ poursuivront la lutte contre la hausse. Pour la toute première fois, les trois grandes associations étudiantes se sont réunies, hier, afin d'annoncer qu'elles rejetaient à l'unanimité la mesure prise par Québec visant à mettre fin à la grève étudiante et qu'elles poursuivront leur lutte au cours des prochaines semaines. (K52JM0407Nou0393)

L'impasse créée par l'opposition du gouvernement est partiellement résolue à la fin du mois de mars 2012. Le gouvernement se dit « ouvert » à des discussions. Dans ce contexte, d'autres acteurs apparaissent dans le rôle d'*adjuvant-possible* et à l'intérieur d'un nouveau programme narratif (**PN2**) où le **gouvernement** est un *sujet* et l'**entente sur la hausse** un *objet de valeur*.

Tableau 5.2 : Programme narratif PN2

Sujet	*Objet de valeur*	*Antisujet*	*Adjuvant-possible*	*Opposant*
gouvernement	**entente sur la hausse**	**associations étudiantes**	**réaménagement du régime de prêts et bourses**	**gel des droits de scolarité**

Le principal acteur qui recouvre la fonction d'*adjuvant-possible* est le **réaménagement du régime de prêts et bourses** ou **gestion des universités**. Ce dernier s'oppose aux acteurs **gel des droits de scolarité** ou **gratuité scolaire**, qui sont des *objets* recherchés par le *sujet* **associations étudiantes**. Les actions accomplies par les étudiants pour obtenir le **gel de frais de scolarité** sont en antithèse avec le programme narratif **PN2** et représentent un *opposant actif* :

> Négocier... à une condition. DROITS DE SCOLARITÉ. QUÉBEC - Québec se dit prêt à discuter de « réaménagements » au régime de prêts et bourses si les étudiants renoncent à exiger le gel des droits de scolarité. « De la mauvaise foi », répondent les associations étudiantes, qui maintiennent leur principale revendication. La ministre de l'Éducation, Line Beauchamp, se dit ouverte à des discussions avec les étudiants « pour encore mieux assurer l'accessibilité des études », mais à une condition : ils doivent cesser de revendiquer le gel des droits de scolarité ou encore la gratuité scolaire. (K55PR0330Act0334)

> Pour la première fois en neuf semaines, la ministre de l'Éducation Line Beauchamp a accepté une proposition venant des étudiants : discuter de la gestion des universités. (K03ME0416Act0235)

> Légère ouverture. La ministre invite la FEUQ à discuter sur la gestion des universités. (K40JQ0416Nou0384)

L'*adjuvant-possible* associé à ce programme narratif comporte d'autres acteurs proposés par le sujet **gouvernement**, comme la **commission permanente sur la gestion des universités**. Dans les extraits qui suivent, on observe que le gouvernement propose une entente avec les étudiants basée sur la constitution de cette commission. Toutefois, la proposition est au service du programme narratif **PN1**, c'est-à-dire l'application de la hausse des droits de scolarité :

> Elle vise la création d'une commission permanente sur la gestion des universités à laquelle prendraient part des étudiants, des professeurs et des dirigeants d'institution. (K03ME0416Act0235)

> Le point de presse a eu lieu peu de temps avant l'annonce de la ministre de l'Éducation, Line Beauchamp, qui s'est dite prête à entamer des discussions avec la FEUQ sur la mise en place d'une « commission indépendante et permanente ». (K52JM0416Nou0470)

Un des éléments les plus présents de cet agglomérat est l'*opposant* du programme narratif **PN1**. Cet opposant est utilisé par l'*antisujet* **gouvernement** contre le *sujet* **associations étudiantes**. Le rôle d'*opposant* est couvert par l'association la **CLASSE**, dont les actions et les affirmations publiques sont utilisées par le gouvernement pour justifier sa position de fermeture envers les requêtes de dialogue provenant des associations étudiantes. Généralement, les articles sont dominés par l'ouverture de la table des négociations pour la FEUQ et la FECQ et l'exclusion de la CLASSE.

> Une ouverture, des conditions. GRÈVE ÉTUDIANTE. Bras de fer avec les étudiants sur la présence de la CLASSE aux discussions. Line Beauchamp voulait parler de gestion des universités sans la CLASSE. (K55PR0416Act0468)

> Guerre de mots. DROITS DE SCOLARITÉ. Charest condamne la violence, les étudiants poursuivent la contestation. Jean Charest exige que la Coalition large de l'Association pour une solidarité syndicale étudiante (CLASSE) condamne les actes de violence si elle veut participer à la table de discussion avec la ministre de l'Éducation et les fédérations étudiantes. (K55PR0417Act0475)
>
> Pas de dialogue sans la CLASSE. Il n'y aura pas de dialogue entre les associations étudiantes et la ministre de l'Éducation à moins que la CLASSE y participe et dénonce les actes de violence commis depuis le début du conflit sur la hausse des droits de scolarité. (K52JM0418Nou0496)
>
> Dialogue de sourds. PARTICIPATION DE LA CLASSE. « Le problème, c'est de demander à la CLASSE de dénoncer plus ouvertement les différents actes de vandalisme ou d'inviter les étudiants à poursuivre dans des actions qui sont pacifiques », a exposé hier la présidente de la Fédération étudiante universitaire du Québec (FEUQ), Martine Desjardins. (K40JQ0418Nou0400)

L'ouverture du dialogue se concrétise une fois que *la CLASSE a dénoncé les actes de violence*. Toutefois, ceci n'est pas suffisant aux yeux du **gouvernement** et elle est exclue à nouveau de la table des négociations.

> La CLASSE condamne la violence. (K55PR0423Act0542)
>
> La ministre rencontre deux des trois associations étudiantes dès aujourd'hui. La CLASSE a fini hier par condamner la violence physique, clé d'une place à la table des négociations proposée par la ministre de l'Éducation. (K52JM0423Nou0551)
>
> Retour à la case départ. LA CLASSE EXCLUE DES NÉGOCIATIONS. Les négociations entre le gouvernement Charest et les associations étudiantes sont tombées dans une impasse hier, dès que la ministre de l'Éducation, Line Beauchamp, eut annoncé qu'elle en excluait la CLASSE. (K40JQ0426Nou0473)
>
> Grève étudiante : Charest imperturbable. Le premier ministre Jean Charest défend la décision de sa ministre de l'Éducation, Line Beauchamp, d'expulser la Coalition large de l'Association pour une solidarité syndicale étudiante

> (CLASSE) des négociations visant à mettre fin à la grève étudiante en cours depuis la mi-février. (K46RC0426no_0292)

L'acteur **CLASSE** est très présent à l'intérieur de cet agglomérat et le discours qui met en valeur son rôle d'*opposant* dans le programme narratif **PN1** est fortement répandu. Le journal *Le Soleil* a un cluster complètement dédié à cette configuration actantielle (SO_5).

> Quant à la Coalition large de l'Association pour une solidarité syndicale étudiante (CLASSE), la ministre a émis des doutes sur son désir de discuter avec le gouvernement. « À ma connaissance, la CLASSE n'a jamais soumis aucune base de discussion. Et cette position assez extrême amène même son porte-parole à être incapable de dénoncer les gestes de vandalisme et de violence qui sont récemment survenus », a-t-elle noté, faisant vraisemblablement référence au saccage de son bureau de circonscription à Montréal-Nord vendredi matin par une quinzaine de manifestants. (K05SO0416Act0413)

> Le gouvernement a refusé, hier, d'inviter la Coalition large de l'Association pour une solidarité syndicale étudiante (CLASSE) à la commission indépendante sur la gestion des universités tant qu'elle ne condamnera pas explicitement les actes de violence commis pendant le conflit étudiant. (K05SO0417Act0422)

Dans ce contexte, un autre *adjuvant* important du programme narratif **PN1** est le **front commun des associations**, qui a contribué à l'obtention de l'ouverture du **dialogue** :

> Front commun « historique ». Mobilisation étudiante. MONTRÉAL - Le mouvement étudiant unit ses forces et rejette en bloc les mesures adoptées par le gouvernement du Québec. Les deux fédérations - collégiale et universitaire - et la CLASSE ont déclaré côte à côte qu'elles allaient poursuivre leur combat contre la hausse des droits de scolarité. (K15SO0407Act0342)

> Grève étudiante : La CLASSE lutte contre son isolement. La Coalition large de l'Association pour une solidarité syndicale étudiante (CLASSE) demande à la

> Fédération étudiante collégiale du Québec (FEUQ) et la Fédération étudiante collégiale du Québec (FECQ) de respecter le pacte conclu il y a deux semaines et de ne pas aller négocier avec la ministre de l'Éducation du Québec sans eux. (K46RC0416no_0235)
>
> GRÈVE ÉTUDIANTE. QUÉBEC - La ministre de l'Éducation, Line Beauchamp, a mis à l'épreuve l'unité du mouvement étudiant hier. Mais les fédérations étudiantes collégiale et universitaire, la FECQ et la FEUQ, ont finalement rejeté son invitation à une rencontre sans la Coalition large de l'association pour une solidarité syndicale étudiante (CLASSE). (K55PR0420Act0511)
>
> Pas de discussion avec la ministre sans la CLASSE. (K03ME0420Act0269)
>
> « On est prêts à aller s'asseoir avec Mme Beauchamp à condition que la CLASSE soit invitée », a statué la Fédération étudiante universitaire du Québec (FEUQ). (K52JM0420Nou0511)

Une fois que le dialogue est ouvert et que les deux parties ont entamé une session de rencontres, d'autres programmes narratifs apparaissent. Des positions différentes entre les associations apparaissent alors. Mais ce qui occupe la majorité de l'espace médiatique demeure encore la polémique entre les **étudiants** et le **gouvernement**.

Le programme narratif **PN2** est configuré de manière différente selon le *sujet*, c'est-à-dire les **associations étudiantes** et le **gouvernement**, puisque les premières ont pour but de bloquer la hausse (avec le « gel » ou la « gratuité scolaire ») alors que le deuxième poursuit un objectif inverse.

> Les négociations amorcées. Mobilisation étudiante. Après 11 semaines de grève, le gouvernement et quatre grandes associations étudiantes ont entamé hier après-midi des discussions pour s'entendre sur une sortie de crise. (K15SO0424Act0492)
>
> Québec et les étudiants poursuivent leurs discussions mardi. Les discussions visant à mettre fin au conflit étudiant en cours depuis maintenant 72 jours se

> poursuivent mardi, à Québec, entre la ministre de l'Éducation Line Beauchamp et les 11 représentants des grandes associations étudiantes. (K46RC0424no_0273)

> Les associations étudiantes ont entrepris ces négociations avec des attitudes différentes. La Fédération étudiante collégiale du Québec (FECQ) et la Fédération étudiante universitaire du Québec (FEUQ) se sont dites « en mode ouverture », prêtes à analyser toute position de la ministre qui pourrait leur paraître raisonnable. [...] Pour leur part, les quatre délégués de la CLASSE sont arrivés à Québec avec le mandat de « bloquer toute hausse des frais de scolarité », a confirmé Gabriel Nadeau- Dubois, porte-parole de l'organisme. [...] Mme Beauchamp a signifié qu'elle aurait des propositions à faire au sujet de l'accessibilité aux études postsecondaires et s'est montrée disposée à écouter les étudiants lui parler des droits de scolarité qu'elle prévoit toujours augmenter de 75 % au cours des cinq prochaines années. La FECQ et la FEUQ ont instamment accepté la condition posée par la ministre. La CLASSE s'est montrée plus « nuancée », laissant savoir qu'elle n'acceptait ni ne rejetait la trêve proposée par la ministre, précisant qu'aucune action de perturbation n'était prévue dans les 48 heures suivant l'appel de la ministre (K40JQ0424Nou0459)

> La ministre de l'Éducation, Line Beauchamp, se dit ouverte à des discussions avec les étudiants [...] mais à une condition : ils doivent cesser de revendiquer le gel des droits de scolarité ou encore la gratuité scolaire. (K55PR0330Act0334)

La confrontation entre les deux parties demeure toujours très houleuse et comporte plusieurs accusations brutales, ruptures et réouvertures du dialogue. En général, les deux parties ne semblent jamais être proches de la signature d'une entente et toutes les actions les ramènent toujours au point de départ. Ce qui représente le mieux cette dynamique infructueuse est l'exclusion répétitive de la CLASSE de la table des négociations, qui constitue toujours un obstacle insurmontable dans la négociation d'une entente.

> Retour à la case départ. LA CLASSE EXCLUE DES NÉGOCIATIONS. Les négociations entre le gouvernement Charest et les associations étudiantes sont tombées dans une impasse hier, dès que la ministre de l'Éducation, Line

> Beauchamp, eut annoncé qu'elle en excluait la CLASSE. (K40JQ0426Nou0473)
>
> Line Beauchamp se dit déçue de voir les étudiants « campés » sur le gel des droits de scolarité. « Ça ne peut pas aider à trouver une entente. » (K52JM0502Vot0657)
>
> Négociations rompues. QUÉBEC - Les discussions entre le gouvernement Charest et les associations étudiantes sont dans une impasse depuis que la ministre de l'Éducation Line Beauchamp a annoncé qu'elle excluait la CLASSE des négociations. Les fédérations étudiantes collégiale et universitaire du Québec (FECQ et FEUQ) ont annoncé qu'elles suspendaient les discussions avec le ministère de l'Éducation immédiatement après que la ministre eut éjecté l'Association pour une solidarité syndicale étudiante (CLASSE). (K52JM0426Nou0585)
>
> Québec décline l'invitation de la FECQ et de la FEUQ pour la reprise des négociations. Le gouvernement Charest oppose une fin de non-recevoir à l'invitation lancée par la Fédération étudiante collégiale (FECQ) et la Fédération étudiante universitaire du Québec (FEUQ) pour reprendre les négociations vendredi, à 14 h. (K46RC0426no_0290)
>
> Les négos déraillent. Mobilisation étudiante. Les négociations entre le gouvernement et les étudiants ont déraillé, hier, avant la fin de la trêve de 48 heures qui devait permettre de dénouer la crise qui secoue le Québec depuis 11 semaines autour de la hausse des droits de scolarité. La ministre de l'Éducation, Line Beauchamp, a annoncé hier en début d'après-midi qu'elle expulsait la Coalition large de l'Association pour une solidarité syndicale étudiante (CLASSE) de la table des négociations. (K15SO0426Act0506)

La polémique entre les deux parties continue après l'échec de la première ronde de négociation. À ce point, le **PN2** prend le dessus sur le **PN1**. D'autres *adjuvants* apparaissent, comme la **proposition** du gouvernement **d'étaler la hausse**. Cet objet proposé par le gouvernement veut être un objet graduel qui substitue l'objet catégoriel **gel de frais**. Généralement un objet graduel s'oppose au catégoriel et, au contraire de ce dernier, ouvre la porte au compromis. Toutefois, afin qu'il soit perçu comme tel, les éléments contextuels doivent véhiculer cette même signification.

> Québec propose d'étaler la hausse des droits de scolarité. Dans un geste visant à mettre un terme au conflit étudiant en cours depuis plus de deux mois, le gouvernement Charest a présenté l'offre globale faite aux associations étudiantes. Québec propose notamment d'étaler la hausse des droits de scolarité de 1625 $ sur sept ans plutôt que cinq. (K46RC0427no_0297)

Les axes du **PN2** restent en opposition, ce qui mène les **associations étudiantes** à un refus de la **proposition**, laquelle a été présentée comme la solution pour arriver à une entente :

> Vers un refus général de la « solution globale ». Mobilisation étudiante. La « solution globale » du gouvernement de Jean Charest pour mettre un terme au conflit étudiant se dirige tout droit vers un rejet massif de la part des étudiants. (K15SO0429Act0539)

L'échec de cette nouvelle tentative de **dialogue** est suivi par une contre-offre de la part des **associations étudiantes** qui visent une **entente** sans céder de terrain sur le **gel des droits de scolarité** :

> Un gel « à coût nul ». Contre-offre de la FEUQ et de la FECQ. Québec - Les fédérations étudiantes présenteront aujourd'hui une contre-offre qui prévoit un gel des droits de scolarité « à coût nul pour les contribuables ». (K55PR0501Act0638)

> La FEUQ et la FECQ ripostent. « Offre globale » du gouvernement Charest. Les fédérations étudiantes universitaires et collégiales du Québec vont riposter aujourd'hui à l'« offre globale » du gouvernement Charest en dévoilant une contre-proposition sans compromis sur le gel des droits de scolarité, mais qui encourage un recours à la médiation. (K15SO0501Act0564)

Les stratégies des diverses **associations étudiantes** demeurent toutefois différentes et variées :

> Contre-offre sans compromis. Mobilisation étudiante. La Coalition large de l'Association pour une solidarité syndicale étudiante (CLASSE) présentera à

> son tour aujourd'hui sa « contre-proposition » qui vise un retour aux droits imposés en 2007, alors que le gouvernement du Québec estime qu'il est illusoire de revenir à la table des négociations et s'en remet aux électeurs pour trancher de la question des droits de scolarité. (K15SO0503Act0596)

> « Nous sommes ouverts à un retour à une table de négociations, mais à la condition qu'on aborde enfin la question des droits de scolarité, fait valoir Gabriel Nadeau - Dubois, porte-parole de la Coalition large de l'Association pour une solidarité syndicale étudiante (CLASSE). On nous a parlé de bourses, de prêts, de gestion, il est temps qu'on s'occupe du noeud du problème. » (K03ME0514Act0373)

Un élément nouveau surgit au mois de mai : la **démission** de la ministre de l'Éducation, madame **Lyne Beauchamp,** et l'arrivée de madame **Michèle Courchesne** comme nouvelle ministre. Cet événement se retrouve à l'intérieur de cet agglomérat, puisque Courchesne intervient principalement pour garantir de nouvelles possibilités de **négociation** avec les étudiants.

Cette arrivée se configure principalement comme un *adjuvant* au programme narratif **PN2** et constitue un changement de stratégie du gouvernement. À la nouvelle ministre, le gouvernement confère aussi de pouvoirs spéciaux :

> Déclaration du premier ministre. Fonctions additionnelles de madame Michelle Courchesne à titre de vice-première ministre et ministre de l'Éducation, du Loisir et du Sport. (K37DE0515Pol0699)

Les clusters qui se concentrent davantage sur le changement de ministre de l'Éducation sont les suivants : SO_22 DE_37 et RC18. Voici quelques extraits qui résument ce changement de garde :

> Un départ surprise. Démission de Line Beauchamp. Si la FEUQ et la FECQ croient que le départ de Line Beauchamp peut favoriser une sortie de crise, la CLASSE rappelle que le problème n'est pas la titulaire au poste de ministre de

> l'Éducation, mais plutôt le refus du gouvernement de discuter des droits de scolarité. (K22SO0515Act0758)

> Les canaux de communication sont ouverts, disent les leaders étudiants. Au sortir de leur rencontre avec la nouvelle ministre de l'Éducation, Michelle Courchesne, mardi, les leaders étudiants des quatre principales associations étudiantes ont tous indiqué que les « canaux de communication » avec le gouvernement restaient ouverts. (K18RC0515no_0415)

> Nouvelles négos, moins de voix. CONFLIT ÉTUDIANT. Le gouvernement et les étudiants ont convenu hier de reprendre les négociations. Mais la Fédération étudiante universitaire du Québec (FEUQ) demande cette fois que les recteurs, la fédération des cégeps et les syndicats n'en fassent pas partie. (K22SO0524Act0897)

> Conflit étudiant - Courchesne se dit prête à reprendre les pourparlers. Les associations étudiantes affirment toujours attendre une proposition de rencontre. (K37DE0524Pol0801)

Avec l'arrivée de madame **Courchesne**, les négociations reprennent. Le changement de stratégie du gouvernement est aussi marqué par la position ambiguë du premier ministre Charest, qui vacille entre un *adjuvant* et un *opposant*. En effet, son **absence** de la table de concertation est considérée comme positive pour l'obtention d'une entente alors que sa **présence** serait négative :

> Reprise des négos aujourd'hui, sans Charest. CONFLIT ÉTUDIANT. Les quatre associations étudiantes sont conviées à une rencontre au sommet aujourd'hui à 14h avec la ministre de l'Éducation Michelle Courchesne et le négociateur gouvernemental Pierre Pilote. (K22SO0528Act0967)

> Jean Charest a participé aux négociations avec les étudiants. Le premier ministre du Québec, Jean Charest, s'est joint aux discussions visant à mettre fin au conflit étudiant qui ont réuni lundi sa ministre de l'Éducation, Michelle Courchesne, et les représentants des grandes associations étudiantes. (K18RC0529no_0526)

Après une période d'optimisme, les discussions entre les deux parties s'arrêtent à nouveau :

> Les leaders étudiants ont fait preuve d'un optimisme prudent à la sortie de leur courte rencontre avec la nouvelle ministre de l'Éducation, Michelle Courchesne, hier soir à Québec. (K03ME0516Act0386)

> Conflit étudiant. Les négociations ont tourné au vinaigre, hier, entre Québec et les leaders étudiants, avant de se conclure sur une nouvelle offre à l'issue incertaine que doit étudier la ministre de l'Éducation, Michelle Courchesne. (K22SO0531Act1009)

> Le gouvernement déplore un « important fossé »; les étudiants dénoncent sa « mauvaise foi ». Les négociations sont rompues. (K18RC0531no_0542)

Avec la ministre Courchesne, les négociations deviennent un véritable *adjuvant* pour la poursuite d'un nouvel *objet*, soit la **sortie de crise** et ceci, en raison du fait que la confrontation entre les étudiants et le gouvernement prend les contours d'une véritable **crise sociale** :

> Ce que les parties souhaitent, c'est une sortie de crise, a assuré la ministre. C'est sûr que si on prend le temps de préparation requis [pour choisir le moment de la rencontre], c'est parce que, de part et d'autre, on est très conscients du sérieux de la situation et qu'on veut mettre toutes les chances de notre côté pour arriver à une entente. (K03ME0525Act0446)

> La négociation pour sortir de la crise. La ministre Courchesne et les associations se rencontrent aujourd'hui à Québec. (K37DE0528Act0834)

> Québec prêt à reculer. Conflit étudiant. Après 16 semaines de refus catégorique, le gouvernement Charest accepte de revoir à la baisse son augmentation des droits de scolarité. (K22SO0530Act0989)

5.1.2 Charest versus Marois : corps à corps

Un autre centre d'agrégation important du réseau (figure 5.5 et 5.6) regroupe sept clusters qui sont connectés avec une valeur cosinus supérieure à 0,77. Parmi ces clusters, deux appartiennent au journal *Le Soleil* (SO_50 et SO_53). Tous les autres clusters appartiennent à des journaux différents (JM_52, JQ_50, RC_9, PR_2 et DE_12). Le journal *Metro* est exclu de cet agglomérat, puisqu'aucun de ses clusters n'est présent. Les clusters centraux sont RC_9 et JM_53 car ils comportent un plus grand nombre de liens. La connexion la plus forte est entre le *Journal de Montréal* et *Le Journal de Québec*.

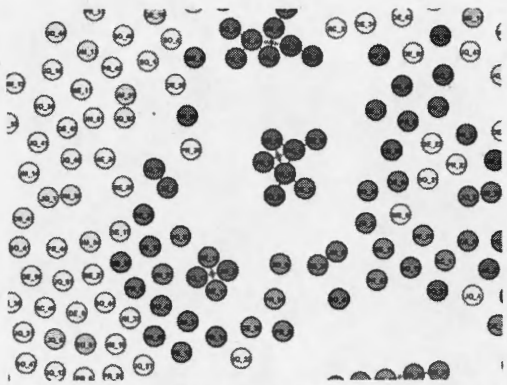


Figure 5.5 Détail de la figure 5.2. Au centre, l'agglomérat traitant la confrontation entre M. Charest et M.me Marois.

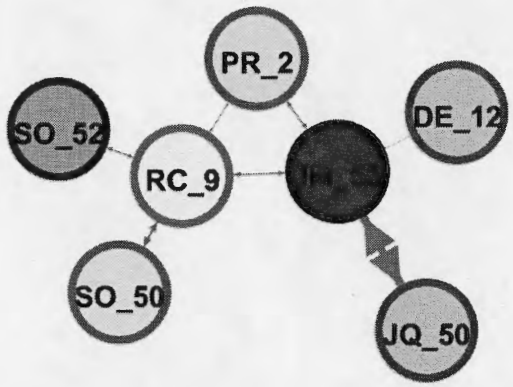


Figure 5.6 Représentation graphique de l'agglomérat traitant la confrontation entre M. Charest et M.me Marois. Les clusters plus foncés désignent un cluster plus volumineux

Le cœur narratif de cet agglomérat est la confrontation entre le chef du gouvernement, le premier ministre, et le chef de l'opposition. Le programme narratif principal **(PN3)** est ainsi dominé par l'opposition entre **Pauline Marois** (chef du PQ et de l'opposition) et **Jean Charest** (chef du PLQ et du gouvernement). L'*objet de valeur* varie selon les différents articles et les différentes périodes. Généralement, il est possible de les résumer en deux grandes catégories : **discréditer l'adversaire** et **gagner les élections**. Ces deux *objets de valeur* sont poursuivis indistinctement par les deux acteurs.

Tableau 5.3 Programme narratif PN3

Sujet	*Objet de valeur*	*Antisujet*
P. Marois/J. Charest	**Discréditer l'adversaire/Gagner les élections**	**J. Charest/P. Marois**

Cet agglomérat révèle l'ampleur que la crise étudiante a prise au cours des semaines. Elle a occupé massivement l'espace médiatique mais aussi le débat politique. Le ton

des répliques échangées entre le chef du gouvernement et la chef de l'opposition est certainement un indicateur de l'importance politique de la grève étudiante :

> Charest et Marois s'échangent des questions sans réponses. Le dialogue de sourds n'est pas qu'une affaire entre le gouvernement et les leaders étudiants. S'y prêtent aussi avec un bel acharnement le chef libéral Jean Charest et son adversaire péquiste Pauline Marois. (K50SO0505Act0624)

En effet, le printemps érable deviendra, plus tard, un des principaux *enjeux électoraux*, comme il est possible de le déduire des extraits suivants :

> En entrevue au Soleil, elle a expliqué comment elle comptait s'y prendre, notamment en leur rappelant tout ce que sa formation politique a fait pour le patrimoine et l'économie de la région et, surtout, que le carré rouge n'est pas l'unique enjeu de la campagne électorale. (K52SO0802Act1380)

> La question des droits de scolarité « sera un enjeu majeur de la prochaine campagne, mais pas suffisant pour faire oublier le bilan libéral ». (K02PR0616Pol1295)

Le plus souvent, le ton des échanges entre les deux politiciens est *agressif* et *insultant*. Les communications sont souvent centrées sur la confrontation entre les deux principaux acteurs, lesquels poursuivent souvent le même objectif. Dans les extraits qui suivent, on souligne l'*objet* **discréditer l'adversaire politique**. Ce dernier, à différence des objets comme **gagner les élections** ou **gel des frais de scolarité**, lesquels sont des objets but, est un objet moyen et il couvre le rôle d'adjuvant pour des programmes narratif centré sur un objet but. C'est le type de relation qu'il a entre l'objet but **gagner les élections** et l'objet moyen **discréditer l'adversaire politique** :

> « Pauline Marois a plié l'échine », dit Charest. M. Charest n'a pas manqué d'attaquer la chef du Parti québécois Pauline Marois sur le conflit étudiant. (K09RC0615no_0659)

> Bilan de fin de session : Pauline Marois attaque Jean Charest dans le dossier étudiant. La session parlementaire qui s'achève est « l'une des pires que nous avons connues », a déclaré dans son bilan la chef du Parti québécois (PQ) et de l'opposition officielle, Pauline Marois. Mme Marois a appelé les Québécois - divisés par le conflit étudiant, selon elle - à changer « de gouvernement et de direction » (K09RC0615no_0663)
>
> Marois dénonce le « Gouvernement usé et corrompu ». (K02PR0802Act1521)
>
> Charest passe à l'attaque. (K02PR0821Act1627)
>
> Débat des chefs : deux heures solides d'échanges musclés. Le débat électoral qui s'est déroulé dimanche soir a prouvé qu'une joute à quatre peut s'avérer musclée, claire, et riche en échanges. (K09RC0819no_0866)
>
> Un tête-à-tête virulent. ÉLECTIONS PROVINCIALES 2012. L'animosité entre Jean Charest et Pauline Marois crevait l'écran, hier, lors de leur affrontement télévisé en face à face. (K52SO0821Act148)

Le thème des élections et de la campagne électorale est présent dans les articles de journaux à partir du début du mois de mai. À ce point, l'*objet* **gagner les élections** commence à dominer dans **PN3** :

> Élections cet été? « Un manque de respect », dit Marois. Déclencher la campagne électorale au début du mois d'août serait « un manque de respect profond » envers la population et une « manipulation » de la part du premier ministre Jean Charest, estime la chef du Parti québécois, Pauline Marois. [...] En vertu de la loi 78, les étudiants doivent retourner en classe à la mi-août. Si le scénario du déclenchement des élections le 1er août se concrétise, la reprise des cours aura lieu en pleine campagne. Mme Marois ne voit là pas une coïncidence. Selon elle, le premier ministre pourrait vouloir instrumentaliser une nouvelle fois les manifestations étudiantes et se poser comme le représentant de la loi et de l'ordre, « comme on l'a vu dans son PowerPoint ». (K09RC0711no_0731)
>
> Marois réclame des élections. Charest se moque de l'intérêt général selon la chef péquiste. Le refus du gouvernement Charest de négocier de bonne foi avec

> les étudiants, c'est la goutte qui fait déborder le vase aux yeux de Pauline Marois. La chef du Parti québécois réclame des élections. (K53JM0427Nou0594)

Dans ces derniers extraits, la crise étudiante tient un rôle ambigu, laissant entrevoir la possibilité qu'elle soit *la cause du déclenchement des élections*. En effet, la **crise étudiante** est un *adjuvant* ou un *opposant* des programmes narratifs des deux chefs politiques. Quelques fois, **Charest** utilise la **grève étudiante** comme *adjuvant* pour **discréditer Marois**; parfois c'est la chef péquiste qui adopte cette stratégie :

> Le premier ministre Jean Charest brandit la menace du chaos, sur fond de référendum sur la souveraineté, si le Parti québécois (PQ) prend le pouvoir aux prochaines élections. [...] Sur la vision politique du PQ, il l'a décrite comme étant « fondée sur le chaos et sur ce qui s'est passé [avec le mouvement étudiant] : le manque de respect pour les individus ». Pour appuyer ses dires, il a cité l'exemple de la candidature pour le PQ du leader étudiant Léo Bureau-Blouin, qui incarne selon ses dires le non-respect d'un droit fondamental : le libre accès des étudiants à leur salle de classe. (K52SO0726Act1340)

> Pauline Marois a prononcé un discours enflammé à haute saveur électorale dans lequel elle a pourfendu le bilan du gouvernement Charest, marqué à son avis par « la corruption, l'endettement, l'injustice et maintenant le conflit social ». (K53JM0506Nou0708)

> Même s'il laisse entendre qu'il n'y aura pas d'élections avant la Fête nationale, Jean Charest a sévèrement attaqué Pauline Marois, soutenant qu'elle s'est disqualifiée de la fonction de premier ministre durant la crise étudiante. « Elle a fait la démonstration qu'elle n'a pas ce qu'il faut pour être premier ministre en agissant comme elle a agi », a-t-il déclaré. (K53JM0507Nou0728)

> « Pauline Marois rembourserait les étudiants. En plus de promettre d'annuler la hausse des droits de scolarité décrétée par le gouvernement Charest, Pauline Marois s'engage à rembourser les étudiants si le PQ devait prendre le pouvoir à la prochaine élection. » (K53JM0503Nou0672)

> Ils reprochent aussi au gouvernement de vouloir profiter du conflit étudiant dans le but de réaliser des gains électoraux. « Jamais il n'a été question de

> chercher à exploiter une affaire comme ça, a nuancé le premier ministre Charest. (...) Mme Marois, c'est la rue », a-t-il ajouté. (K53JM0614Nou1218)

> « Il veut qu'on oublie son bilan; eh bien, moi, je vais en parler de son bilan », assure la chef du PQ. La prochaine campagne électorale ne portera pas que sur le « carré rouge » et sur la loi et l'ordre. Il sera aussi question d'éthique, de confiance et de corruption. (K53JM0616Nou1227)

> Contre Marois. La semaine dernière, la fuite d'un document du Parti libéral a permis de démontrer que M. Charest entend, lors de la prochaine campagne, opposer son équipe et son Plan Nord à une Pauline Marois avec son référendum et la rue. (K53JM0618Nou1243)

Au fur et à mesure que les semaines avancent, et surtout après le déclenchement des élections, les caractéristiques des « discours de campagne électorale » occupent la majorité des articles de journaux. La question étudiante devient alors marginale mais elle est toujours citée comme *un des enjeux les plus importants* de la *campagne électorale*. Dans ce contexte, d'autres types d'enjeux apparaissent, comme ceux énumérés par **Marois** dans les articles suivants :

> L'insistance de Jean Charest à pointer du doigt le bout de tissu rouge que Pauline Marois porte à son corsage ne vise qu'un seul objectif : « Il veut qu'on oublie son bilan; eh bien, moi, je vais en parler de son bilan », assure la chef du PQ. La prochaine campagne électorale ne portera pas que sur le « carré rouge » et sur la loi et l'ordre. Il sera aussi question d'éthique, de confiance et de corruption. À ce sujet, les travaux de la commission Charbonneau débutent à peine et déjà le gouvernement se trouve dans l'embarras avec l'ex-ministère des Transports, Sam Hamad, dans un très mauvais rôle. (K53JM0616Nou1227)

> Respect. « Nous devons décider dans quel type de société nous voulons vivre. Je vous propose une société qui avance dans le respect de chaque citoyen, dans le respect de nos institutions et le respect de la démocratie. Pauline Marois a fait le choix d'embrasser le mouvement de contestation, de porter ses symboles et même de recruter ses candidats. Pauline Marois propose de plier, de céder et de leur donner tout ce qu'ils demandent », a-t-il enchaîné. La presque totalité des attaques de Jean Charest ciblait son adversaire péquiste. (K53JM0802Nou1426)

> « Les stratégies des trois principaux chefs de parti sont déjà très claires. Jean Charest a donné un caractère carrément référendaire à l'élection du 4 septembre : pour ou contre la hausse des frais de scolarité; pour ou contre la préservation de la paix sociale?. Il a accolé hier Pauline Marois à la rue, aux manifestations, aux perturbations diverses, au point de porter elle-même le carré rouge et d'avoir même recruté comme candidat du Parti québécois l'un des leaders de cette crise. Le premier ministre place les Québécois devant un choix " entre la stabilité ou l'instabilité". » (K53JM0802Nou1428)

> « Pauline Marois et Jean Charest ont déployé hier tout leur arsenal d'arguments et de blâmes dans le premier de trois face-à-face électoraux à être diffusés sur le réseau TVA. Soixante minutes d'échanges musclés où les deux chefs ont défendu leur peau sans pour autant verser dans l'insulte gratuite. » (K53JM0821Nou1530)

Avant cette « dérive » électoraliste, d'autres *acteurs* dominent le discours. Par exemple, la **violence**, est un concept qui est souvent utilisé par le premier ministre pour attaquer les étudiants et les manifestants et pour discréditer la grève. Il est également utilisé pour discréditer Pauline Marois. En effet, cet élément est un *adjuvant* en support de **PN3** :

> Violence. Aux yeux du gouvernement, la CLASSE ne peut plus prendre part aux négociations, puisqu'elle « encourage la violence » et a contribué aux débordements des récentes manifestations qui ont eu lieu dans les rues de Montréal. [...] Jean Charest s'en est pris à sa vis-à-vis péquiste. « Pauline Marois ne croit pas ce qu'elle dit sur les frais de scolarité », a pesté le premier ministre. Cette idée que le gouvernement se sert de ce conflit à des fins électoralistes est « grotesque », a d'ailleurs signalé M. Charest. » (K50JQ0427Nou0486)

La **violence** se présente sous une forme particulière dans un cas de *stigmatisation d'une non-action* qui aurait dû être accomplie par Pauline Marois. En d'autres termes, la passivité de la chef péquiste est utilisée par Charest pour l'attaquer. Le premier ministre l'accuse ainsi de ne pas agir pour *contrer les actes de **violence***. La rhétorique utilisée par **Charest** contre **Marois** évoque la relation de complémentarité entre *non-*

opposant et *adjuvant*. Si face à la **violence** la chef péquiste est supposée tenir le rôle d'*opposant*, son immobilisme ne peut que l'étiqueter comme *non-opposant*, ce qui implique logiquement le rôle d'*adjuvant passif*. Cette dynamique est assez répandue dans cet agglomérat de clusters:

> « Une "faute impardonnable". Charest condamne " le silence " de Marois sur les dérapages du conflit étudiant Jean-Luc Lavallée. » (K53JM0414Nou0454)

> « Charest accuse Marois de mollesse. Le premier ministre Jean Charest accuse Pauline Marois de faire preuve de mollesse au moment où les négociations entre les associations étudiantes et la ministre Line Beauchamp butent sur un débat sémantique. « La chef de l'opposition a-t-elle du Jell-O dans la colonne vertébrale », a demandé le premier ministre pendant la période de questions, hier, à l'Assemblée nationale. M. Charest reprochait à Pauline Marois de dénoncer mollement certains gestes commis en marge du conflit étudiant. » (K50JQ0419Nou0407)

> Ils accusent la chef du Parti québécois, Pauline Marois, de ne pas condamner les actions de "perturbation économique". (K02PR0323Act0275)

> Après un discours aux accents nettement préélectoraux, il s'en est pris rudement à son adversaire Pauline Marois. Selon M. Charest, "par sa valse-hésitation" dans le conflit étudiant, Mme Marois "a fait la démonstration qu'elle n'a pas ce qu'il faut pour être premier ministre". (K02PR0507Pol0737)

> Sans surprise, il a répété que Pauline Marois veut organiser un nouveau référendum et qu'elle a refusé de condamner la "violence et l'intimidation", deux mots omniprésents depuis quelques semaines dans son discours. (K02PR0616Pol1304)

La chef péquiste devient un véritable *adjuvant actif* au support de la cause étudiante lors qu'elle commence à porter le **carré rouge**, qui lui aussi est un *adjuvant* :

> Selon lui [Charest], la chef péquiste s'est comportée de façon « indéfendable » en portant le carré rouge pendant la crise. « Dire que c'est dans la rue que ça se

> règle, et encourager les étudiants à boycotter leurs cours, ce n'est pas responsable de la part d'un parlementaire », a-t-il ajouté. (K50JQ0507Nou0605)
>
> Pauline Marois a fait le choix d'embrasser le mouvement de contestation, de porter ses symboles et même de recruter ses candidats. Pauline Marois propose de plier, de céder et de leur donner tout ce qu'ils demandent ", a-t-il enchaîné. La presque totalité des attaques de Jean Charest ciblait son adversaire péquiste (K53JM0802Nou1426)

Par ailleurs, **Marois** adopte d'autres stratégies pour discréditer le gouvernement et elle utilise largement la **crise étudiante** comme *adjuvant* :

> Marois veut renverser le gouvernement Charest. QUÉBEC - QUÉBEC -- Alors que les rumeurs d'élections battent leur plein, le PQ est prêt à enfiler les gants. Pauline Marois a déposé une motion de censure, hier, pour défaire le gouvernement Charest, lui reprochant sa gestion " catastrophique " de la crise étudiante. (K53JM0502Nou0654)
>
> Le premier ministre Jean Charest refuse de s'excuser pour les propos sarcastiques prononcés vendredi dernier à l'endroit des étudiants alors qu'une manifestation tournait à l'affrontement avec les forces policières dans les rues de Montréal. « Est-ce que le premier ministre est prêt à s'excuser pour les propos qu'il a tenus vendredi dernier lors du Salon sur le Plan Nord », a demandé, hier, la chef de l'opposition, Pauline Marois, à l'ouverture de la période de questions à l'Assemblée nationale. (K50JQ0425Nou0468)
>
> Monopolisée par la crise étudiante et la division qu'elle a suscitée, la session parlementaire qui se termine aura été « l'une des pires que le Québec ait connues », a lancé hier en point de presse Pauline Marois, chef de l'opposition officielle. Selon elle, Jean Charest se doit de déclencher des élections le plus rapidement possible. [...] « Il ne reste plus qu'une solution pour mettre fin à cette crise qui a déjà trop divisé les Québécois : seules des élections générales permettront de retrouver la paix sociale, et le plus vite sera le mieux », a soutenu Mme Marois. (K02PR0616Pol1295)
>
> Charest se moque de l'intérêt général selon la chef péquiste. Le refus du gouvernement Charest de négocier de bonne foi avec les étudiants, c'est la

> goutte qui fait déborder le vase aux yeux de Pauline Marois. (K50JQ0427Nou0486)
>
> Selon Marois, Charest « diabolise les étudiants ». La chef du Parti québécois, Pauline Marois, n'y est pas allée de main morte, hier, en attaquant de front l'attitude « lamentable » du gouvernement libéral de Jean Charest. (K53JM0415Nou0469)
>
> « Le conflit étudiant est un élément parmi d'autres, évidemment, il a pris beaucoup de place parce que M. Charest l'a laissé traîner », reconnaît d'emblée la chef péquiste. (K52SO0802Act1380)
>
> Bilan de fin de session : Pauline Marois attaque Jean Charest dans le dossier étudiant. La session parlementaire qui s'achève est « l'une des pires que nous avons connues », a déclaré dans son bilan la chef du Parti québécois (PQ) et de l'opposition officielle, Pauline Marois. (K09RC0615no_0663)

Dans cet agglomérat, le cluster du journal *Le Devoir* semble se distinguer en termes de langage utilisé, même s'il rapporte souvent les mêmes nouvelles au même moment que les autres journaux. Ce constat n'a pas pu, par contre, être exploré davantage dans le cadre de cette thèse.

> Crise sociale - Charest n'admet aucun tort. Le premier ministre affirme avoir fait tous les efforts nécessaires pour régler le conflit. Le premier ministre Jean Charest ne se reconnaît aucun tort dans la crise sociale dans laquelle le conflit étudiant a plongé le Québec pendant 12 semaines. (K12DE0507Pol0624)
>
> « Rien ne fonctionne plus, car la confiance de la population envers le premier ministre est rompue. Le premier ministre et le Parti libéral s'accrochent au pouvoir. Ce faisant, le premier ministre entraîne le Québec dans sa chute. Nous sommes dans un cul-de-sac. ». C'est en ces termes que la chef du Parti québécois, Pauline Marois, a amorcé hier le débat de cinq heures à l'Assemblée nationale sur la motion de censure contre le gouvernement. (K12DE0504Pol0593)

Il faut ajouter que, si le seuil pour filtrer les liens est réduit d'à peine 0,03, l'agglomérat s'élargit énormément, atteignant 23 clusters très fortement connectés. Ceci indique que *ce nœud narratif est prépondérant dans le corpus* et que d'autres études plus ponctuelles et détaillées pourraient suivre cette première exploration.

5.1.3 Loi spéciale n. 12 : L'objet magique

Un autre groupe de clusters a été identifié avec un seuil de la valeur cosinus fixé à 0,74. Cet agglomérat contient sept clusters connectés entre eux, chacun appartenant à un journal différent. Les journaux les plus connectés sont *Le Journal de Montréal* (JM_29) avec *Le Journal de Québec* (JQ_33)(figure 5.7) et ils restent relativement écartés du reste des clusters. Au contraire, le journal *Le Soleil* (SO_39) a plusieurs connexions importantes avec plusieurs journaux comme *Le Devoir* (DE_2), *La Presse* (PR_32) et *Métro* (ME_12). Le site web *Radio-Canada* (RC_7) est connecté surtout avec *La Presse* et *Le Soleil*. Le cluster *Le Devoir* est celui dont la taille est la plus importante. De plus, à cause de leur taille réduite, les deux clusters du groupe Québecor (JM_29 et JQ_33), ne font pas partie du tiers de clusters retenus dans la phase d'annotation. Ce qui souligne encore plus la distance entre ces deux journaux et les autres par rapport à la relation avec la macrostructure qui est traitée dans cet agglomérat.

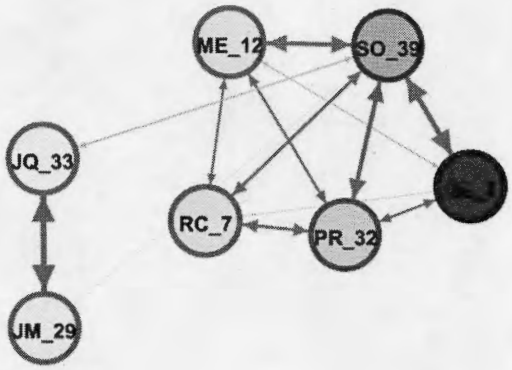


Figure 5.7 Représentation graphique de l'agglomérat traitant le projet de loi 12. Les clusters plus foncés désignent un cluster plus volumineux

Tableau 5.4 Taille des clusters de l'agglomérat sur la loi 12

Cluster	Taille
DE_2	59
SO_39	41
PR_32	30
RC_7	23
ME_12	22
JM_29	19
JQ_33	17

Tableau 5.5 Programme narratif PN4

Sujet	*Objet de valeur*	*Adjuvant*
Gouvernement	**Hausse des droits de scolarité**	**Loi spéciale**

Dans cet agglomérat, on remarque surtout la mise en discours d'une dynamique narrative entre trois éléments. L'un d'eux est un « objet magique », c'est-à-dire le projet de loi spécial n. 78, devenu la **loi n. 12** ou tout simplement, la **loi spéciale**. Les autres sont le *sujet* (**gouvernement**) et l'*objet de valeur* (**hausse des droits de scolarité**) du programme narratif **PN4** qui constituent la macrostructure de cet

agglomérat. Comme l'on peut observer dans les extraits d'article, le discours sur la loi spéciale met souvent de l'avant les « pouvoirs spéciaux » que cet élément détient dans son rôle d'*adjuvant* à **PN4** :

> Conflit étudiant - La carotte et la matraque. Dans un conflit entre l'État et ses employés, l'adoption d'une loi spéciale répressive sonne la fin de la récréation. Le gouvernement s'en sert comme d'une carte ultime qui met fin aux négociations et ordonne un retour au travail. (K02DE0526Pol0822)

> LA LOI SPÉCIALE DÉPOSÉE. La loi d'exception permettra à Québec de contrôler étroitement toutes les manifestations. (K32PR0518Act0914)

> Un consensus vieux de 50 ans vole en éclats. Mobilisation étudiante. En imposant une loi spéciale pour mettre fin au conflit, le gouvernement va à l'encontre d'un consensus social vieux de 50 ans au Québec selon lequel les étudiants ont le droit de faire la grève. (K39SO0517Act0783)

> Loi 78 - Abus de pouvoir. Le gouvernement Charest a choisi de dénouer la grève étudiante sur les droits de scolarité par la manière forte, suite logique de sa gestion d'une crise qu'il n'a jamais comprise ni maîtrisée. (K02DE0519Pol0754)

L'*antisujet* de ce programme narratif est couvert par les **étudiants,** les **manifestants** ou les **associations étudiantes**, qui sont en effet les trois principaux acteurs opposés au gouvernement. Ces acteurs sont soutenus par des politiciens comme M.me Marois ou M. Khadir et leurs partis politiques.

> Déclaration de guerre aux étudiants. Des amendes pouvant aller jusqu'à 125 000 $ par jour pour les associations étudiantes. Le projet de loi 78 visant à mettre fin au conflit étudiant a soulevé l'ire des associations et de l'opposition, hier. (K02DE0518Pol0734)

> Hier, les trois principales associations défendant les 150 000 jeunes en grève ont dénoncé la loi spéciale qui suspend jusqu'à l'automne les cours sur les campus frappés par le boycottage. (K39SO0517Act0784)

> Une « déclaration de guerre » disent les étudiants. Mobilisation étudiante. La loi spéciale de Jean Charest est une « déclaration de guerre » aux étudiants. Elle sera contestée parce qu'elle est « déraisonnable » et qu'elle vise à « tuer les associations étudiantes ». (K39SO0518Act0791)

L'élément **loi spéciale** couvre également un autre rôle actantiel, soit l'*objet de valeur*, ce qui donne lieu à un nouveau programme narratif **PN5**.

Tableau 5.6 Programme narratif PN5

Sujet	*Objet de valeur*	*Adjuvant*
Étudiants/Aassociations étudiantes	**Contre loi spéciale**	**Recours juridiques/Désobéissance civile**

Ce programme est essentiellement composé comme *sujet* par les **étudiants** ou les **associations étudiantes** qui cherchent à « faire suspendre » ou « faire retirer » la **loi spéciale**, en adoptant plusieurs stratégies et *adjuvants*, dont la majorité dans le cadre de **recours juridiques** :

> Loi spéciale - Une vague d'indignation. Du Barreau à Amnistie internationale en passant par un regroupement d'historiens, le projet de loi spéciale (qui a été adopté hier) a été accueilli hier par un déluge de protestations. (K02DE0519Pol0748)

> Les juristes prennent la rue pour dénoncer la loi spéciale. C'était hier soir au tour des avocats et des notaires d'ajouter leur voix au concert de protestations contre la loi spéciale limitant la liberté de manifester adoptée il y a plus d'une semaine par l'Assemblée nationale du Québec. (K02DE0529Soc0854)

Parmi les différentes configurations prises par le programme narratif **PN5** il en existe une qui se démarque de façon plus nette. Le *sujet* de cette configuration est une association étudiante particulière, la **CLASSE**, qui adopte une stratégie spécifique pour contrer la **loi spéciale**, basée sur l'*adjuvant* **désobéissance civile** :

> Désobéissance civile : La CLASSE pourrait désobéir. CONFLIT ÉTUDIANT. L'aile la plus militante du mouvement étudiant, la CLASSE, n'exclut pas la désobéissance civile et songe à défier la loi spéciale du gouvernement libéral visant à mettre un terme au conflit sur la hausse des droits de scolarité. (K39SO0519Act0808)

> Désobéissance, jure la CLASSE. CONFLIT ÉTUDIANT. Montréal - La CLASSE ne veut pas attendre qu'un tribunal tranche sur la constitutionnalité de la loi spéciale du gouvernement Charest : l'association étudiante jure de la défier. (K39SO0522Act0857)

Cette stratégie se répand rapidement au sein de toute la communauté manifestante, laquelle a commencé à grandir après l'approbation de la **loi spéciale**. En effet, la **loi spéciale** a constitué un élément qui a transformé la grève étudiante en une véritable **crise sociale**, les manifestants n'étant plus limités à la communauté étudiante. De plus en plus, les manifestants se sont multipliés, incluant des citoyens voulant manifester contre un gouvernement qui « accomplissait un abus de pouvoir » (K02DE0519Pol0754) :

> Exaspérés devant un conflit qui s'éternise, les citoyens se sont rangés derrière le gouvernement et non pas derrière sa loi spéciale, croient les porte-parole des fédérations étudiantes. Un sondage CROP-Le Soleil - La Presse publié hier matin indiquait entre autres que 66 % des répondants sont pour la loi spéciale annoncée par le gouvernement Charest. (K39SO0520Act0852)

> Conflit étudiant - La rue a répondu. Des dizaines de milliers de personnes ont bravé hier la loi 78, retournant au gouvernement par la force du nombre l'absurde de son code répressif. (K02DE0523Pol0786)

> Loi 78 : la rue choisit la désobéissance pacifique. Au centième jour d'un conflit qui s'enlise, la rue a répondu à la loi spéciale adoptée la semaine dernière par l'Assemblée nationale par un immense pied de nez : la quasi-totalité des manifestants - 250 000 selon les organisateurs - ont bifurqué du trajet soumis aux forces de l'ordre, bafouant ainsi des dispositions de la loi 78 sous l'oeil généralement tolérant des policiers. (K02DE0523Soc0778)

> Les présidents de la CSN, de la FTQ et de la CSQ ont également déploré le dépôt d'une loi spéciale, hier soir. (K12ME0517Act0398)

Les **associations étudiantes** ont testé, par de nombreuses **voies juridiques,** d'obtenir la suspension de la **loi spéciale** :

> Loi spéciale - Rendez-vous devant les tribunaux. Les étudiants s'adresseront à la cour pour faire invalider la loi spéciale. Les organisations étudiantes promettent de contester devant les tribunaux la loi spéciale adoptée hier par l'Assemblée nationale dans le but d'apaiser la crise. (K02DE0519Soc0747)

> Front commun pour contester la loi 78. CONFLIT ÉTUDIANT. Un front commun d'associations étudiantes, syndicales, communautaires et environnementales a déposé hier matin deux requêtes au palais de justice de Montréal pour faire annuler la loi spéciale du gouvernement Charest. (K39SO0526Act0919)

> Loi 78 : la requête en sursis rejetée. Montréal - La requête en sursis visant à suspendre l'application de certaines dispositions de la loi spéciale au coeur de l'impasse dans le conflit étudiant a été rejetée. Le juge François Rolland de la Cour supérieure, qui a étudié la demande, a conclu qu'un débat de fond à propos de la loi 12 - issue du projet de loi spéciale 78 - était nécessaire avant de pouvoir la suspendre. La requête en sursis, présentée par les associations étudiantes, visait à suspendre l'application de certaines dispositions de la loi. (K39SO0628Act1237)

> Loi 78 : "inapplicable en droit". La Commission des droits de la personne dit que la loi porte atteinte aux libertés fondamentales. 3. La loi 78 porte atteinte à des libertés fondamentales garanties par la Charte des droits et libertés de la personne, a tranché hier la Commission des droits de la personne et des droits de la jeunesse. (K39SO0720Act1320)

La controverse sur la **loi spéciale** a déclenché un débat mettant de l'avant un autre rôle actantiel significatif, soit le *destinataire*. En effet, les interventions des acteurs publiques tendent de plus en plus à mettre en évidence les conséquences et les implications qu'une loi de ce type comporte, soulignant surtout ses impacts face aux

droits fondamentaux de la personne. Dans ce cadre, les actions accomplies par le gouvernement dans l'application de **PN4** sont jugées par le *destinataire* qui est représenté par les **droits fondamentaux** de la personne :

> La loi spéciale inquiète le Barreau. CONFLIT ÉTUDIANT. Bien que le Barreau du Québec poursuive « les mêmes objectifs que le gouvernement et souhaite une sortie de crise », le bâtonnier Louis Masson a émis de « sérieuses inquiétudes » à l'égard de la loi 78. « Le Barreau du Québec est notamment préoccupé par les limitations apportées au droit d'association et au droit de manifestation. De plus, nous critiquons la judiciarisation des débats et le recours à la justice pénale prévus dans le projet de loi », exposait Me Masson dans un communiqué du Barreau publié hier avant l'adoption de la loi. (K39SO0519Act0814)

> La fin des assos étudiantes ? L'application de la loi brisera de facto le droit d'association, disent les juristes. De sévères limitations à la liberté d'association, doublées d'une grande restriction des moyens d'action des associations étudiantes, un rapport de force anéanti : la loi spéciale concoctée par le gouvernement Charest pourrait bien signifier la fin du mouvement étudiant tel que le Québec le connaît depuis plusieurs décennies, craignent les juristes et le Barreau du Québec. (K02DE0519Pol0750)

> La toute récente loi 78 est sans doute l'une des lois spéciales québécoises les plus (sinon la plus) lourdement attentatoires aux droits fondamentaux protégés par les chartes canadienne et québécoise. Que l'on analyse le texte sous l'angle de la liberté d'association et du droit d'agir collectivement, de la liberté de conscience, de la liberté d'expression ou du droit de manifester pacifiquement, presque tous les articles de cette nouvelle loi soulèvent, à leur face même, de sérieux doutes quant à leur compatibilité avec les chartes applicables au Québec. (K39SO0523Opi0870)

Dans ce contexte, le **gouvernement** a essayé de se défendre, mettant également de l'avant le *destinataire* de son programme narratif, principalement manifesté sous forme de **droit à l'éducation** :

> « Ce n'est pas une loi matraque ». Le ministre Fournier prétend qu'il s'agit plutôt de mettre en avant le droit à l'éducation. La très controversée loi 78,

> adoptée sous bâillon en fin d'après-midi hier après une nuit et une journée de débat, n'est pas la « loi matraque dont certains parlent », a insisté le ministre de la Justice, Jean-Marc Fournier, puisqu'il met en avant un droit, celui de l'accès à l'éducation. » (K02DE0519Pol0763)

5.1.4 Manifestations : Police versus Étudiants

Un autre agglomérat de clusters regroupe six clusters connectés à un seuil supérieur à 0,74. À un seuil de 0,70, l'agglomérat en regroupe 13. Les plus connectés (figure 5.8) sont ceux du *Le Journal de Montréal* (JM_8), *La Presse* (PR_7) et *Le Journal de Québec* (JQ_36), auxquels se connectent le quotidien *Le Devoir* (DE_27) et le site web de *Radio-Canada* (RC_40).

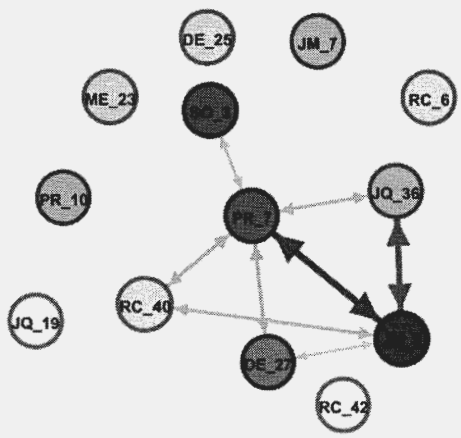


Figure 5.8 Représentation graphique de l'agglomérat traitant les manifestations et les affrontements entre police et manifestant. Les clusters plus foncés désignent un cluster plus volumineux

Tableau 5.7 Programme narratif PN6

Sujet	*Objet de valeur*	*Antisujet*	*Adjuvant*	*Opposant*
Manifestant	**Contrer la hausse**	**Police**	**Manifestations**	**Répression**

Il s'agit ici d'un groupe de clusters dont les articles présentent les événements liés aux manifestations. Plus spécifiquement, cet agglomérat met de l'avant les

affrontements entre la **police** et les **manifestants**, ce qui constitue la *modalité principale du traitement journalistique des manifestations*. En d'autres termes, nous avons de bons indices pour croire que, lors d'une manifestation, les journaux soulignent surtout les dynamiques d'affrontement. Certainement, les affrontements pendant les manifestations constituent la « valeur de l'information » (*newsworthiness*) principale. Généralement, entre deux événements, l'un polémique et l'autre irénique (contraire de polémique), les médias tendent à couvrir davantage le premier. Ceci dopent des questions purement stratégiques qui dépendent d'un phénomène bien connu, c'est-à-dire l'importance que les « S » (sang, sport, sexe, scandale, $) ont dans les intérêts des lecteurs des journaux. En effet, chaque cluster est peuplé surtout d'articles qui mettent en évidence les cas d'intervention policière durant les manifestations. Le programme narratif principal **PN6** est donc composé par le *sujet* **manifestant** et l'*antisujet* **police**, ou vice versa.

> Plus d'une centaine de manifestants se sont rassemblés devant le quartier général du Service de police de la Ville de Montréal (SPVM), hier soir, quelques heures après avoir manifesté dans les rues de la métropole pour protester contre la hausse des droits de scolarité. (K08JM0308Nou0119)

> C'est l'affrontement. GRÈVE ÉTUDIANTE. De 500 à 600 étudiants et élèves en grève ont bloqué hier après-midi l'entrée de l'immeuble de Loto-Québec, au centre-ville de Montréal, pour protester contre la hausse des droits de scolarité, avant d'être évincés par la police de Montréal qui a dû user de la force pour les disperser. Cinq manifestants ont été arrêtés et quatre personnes, dont un policier, ont été transportées à l'hôpital. (K07PR0308Act0131)

> LE CENTRE-VILLE PARALYSÉ. 1 800 manifestants crient leur indignation dans le Quartier international. Environ 1 800 personnes opposées à la hausse des droits de scolarité se sont réunies hier après-midi au centre-ville de Montréal pour participer à une marche qui s'est déroulée sous haute tension. (K08JM0314Nou0177)

> La marche contre la brutalité policière a donné lieu à des scènes violentes et désolantes. Vitrines saccagées, voitures de police renversées, magasins pillés :

> des milliers de manifestants ont pris d'assaut le centre-ville de Montréal hier. [...] Les policiers ont répliqué en faisant exploser des grenades assourdissantes. « C'est complètement débile. Les policiers foncent sur tout le monde comme des sauvages. C'est quoi leur problème? », s'indigne une manifestante, en larmes. (K08JM0316Nou0189)

Pour les manifestants, les manifestations sont un *adjuvant* pour atteindre l'objectif de leur programme narratif principal **PN6**, c'est-à-dire lutter pour faire annuler la **hausse des droits de scolarité**. Ceci est vrai surtout lors des manifestations étudiantes ayant caractérisé la première moitié de la période couverte dans cette étude.

> Une autre marche « jusqu'à la victoire ». GRÈVE ÉTUDIANTE. Des milliers d'étudiants en grève ont de nouveau manifesté dans les rues de Montréal, hier. « Une manifestation nocturne jusqu'à la victoire ». Voilà le leitmotiv des étudiants contre la hausse des droits de scolarité depuis quelques jours. (K07PR0430Act0629)

Les manifestations ont pris différentes formes, allant des défilés pacifiques aux contestations plus directes au regard des politiciens et ce, jusqu'aux actes de violences et vandalisme. Cette diversité est due au fait que les manifestants et les groupes organisés sont différents entre eux et ils ne partagent pas toujours la même vision ou le même objet but :

> La « Grande Mascarade ». Plusieurs centaines d'étudiants costumés et masqués ont envahi le Quartier des spectacles au centre-ville de Montréal hier après-midi, au terme d'une marche haute en couleur pour protester contre la hausse des frais de scolarité. (K08JM0330Nou0317)

> Visite-surprise chez Jean Charest. CRISE ÉTUDIANTE DU JAMAIS VU. Jean Charest a eu une visite-surprise, hier soir. Les manifestants qui participaient au neuvième rassemblement nocturne de suite ont marché jusqu'à sa résidence située à Westmount. (K07PR0503Act0679)

> La série de manifestations se poursuit. GRÈVE ÉTUDIANTE. Deux rassemblements distincts ont eu lieu dans les rues de Montréal, hier soir. En

> plus de la 15e marche nocturne qui s'est amorcée à la place Émilie-Gamelin, un autre groupe de protestataires « moins pacifiques » s'est réuni à la station de métro Acadie, dans le quartier Parc-Extension. (K07PR0509Act0764)
>
> Plus tôt, environ 1 500 personnes s'étaient réunies à la place Émilie-Gamelin, comme le veut la tradition des manifestations nocturnes des dernières semaines. Les gens se sont mis en marche vers 21 h, avec un groupe de « mères en colère » à la tête. (K08JM0518Nou0893)
>
> Plusieurs vitrines ont été fracassées, dont celles de la Banque CIBC, au 1010 rue Sherbrooke Ouest, des banques Nationale et TD Canada Trust au coin des rues Stanley et du poste de quartier 21. (K08JM0426Nou0579)
>
> LA MANIF DE LA COLÈRE. CRISE ÉTUDIANTE RIEN NE VA PLUS. Aussitôt les négociations rompues entre la ministre Line Beauchamp et les représentants des étudiants, ceux-ci sont redescendus dans la rue pour manifester. [...] les protestataires ont exprimé leur frustration en brisant des vitres de commerces, en lançant des balles de peinture sur des édifices et en vandalisant quelques voitures. La majorité des manifestants semblaient pacifiques, mais certains cachaient leur visage d'un masque, et quelques membres du Black Bloc étaient sur place. Les policiers, très nombreux, étaient visibles et ont prévenu les manifestants que si certains causaient des méfaits, le rassemblement serait déclaré illégal. (K07PR0426Act0577)
>
> Ça dégénère au centre-ville. CRISE ÉTUDIANTE. Vitrines fracassées et interventions au gaz poivre ponctuent la manifestation nocturne. L'annonce du dépôt par le gouvernement Charest d'une loi d'exception suspendant le trimestre des étudiants et des élèves en grève, certains depuis février, a donné lieu à une manifestation nocturne qui a semé le désordre dans son sillage, hier soir et cette nuit, à Montréal. Des vitrines ont été fracassées dans le centre-ville, les policiers ont essuyé des projectiles lancés par des manifestants et ils ont utilisé le gaz poivre pour disperser la foule. (K07PR0517Act0898)

Pour la plupart, les interventions des policiers sont justifiées par la mission ultime d'un corps policier soit *la défense de la loi et le maintien de l'ordre*. Leur but est d'éviter le plus possible le désordre, les actes de violence et de vandalisme, ainsi que de punir ceux ne respectent pas la loi.

> Étudiants et policiers entre les matraques et les roses. Policiers et étudiants doivent-ils collaborer pour empêcher les actes violents lors des manifestations? Pour éviter les débordements, le SPVM redouble d'efforts pour inciter les étudiants à dénoncer les casseurs sur les réseaux sociaux. (K27DE0430Soc0547)

> Une CLAC hargneuse, des étudiants festifs. Journée de manifs. Plus d'une centaine d'arrestations, des policiers blessés, des vitres fracassées, des voitures aspergées de peinture : la manifestation anticapitaliste organisée dans le cadre de la fête des Travailleurs a tournée au vinaigre en fin d'après-midi, hier, au centre-ville de Montréal. (K07PR0502Act0655)

Toutefois, un débat s'est ouvert sur l'usage massif et inapproprié de la force policière. Dans plusieurs articles, on illustre les dynamiques de telles interventions. Le doute sur la légitimité de ces interventions se manifeste surtout lors de l'analyse des *adjuvants* en soutien aux forces policières, comme les **arrestations** et l'utilisation de différents **outils de répression** souvent surutilisés :

> Manifestations en marge du débat sur la loi spéciale. Couverture en direct - Des manifestations sont en cours à Montréal, à Québec et à Sherbrooke, alors qu'à l'Assemblée nationale les élus débattent du projet de loi spéciale déposé par le gouvernement de Jean Charest. (K40RC0517no_0441)

> Manifestation monstre contre la brutalité policière - Pagaille au centre-ville. Le jeu du chat et de la souris entre policiers et manifestants se termine par environ 125 arrestations, mais peu de vandalisme. (K27DE0316Act0190)

> Grève étudiante. Exaspérée par les manifestations improvisées, la police de Québec a arrêté 14 étudiants, hier, à qui elle a remis des contraventions salées. Et elle compte sévir de nouveau s'il y a récidive. [...] À l'intérieur du véhicule, les poignets des manifestants ont été enserrés de menottes de plastique. Pour avoir entravé la circulation ou occupé la chaussée, les étudiants ont reçu des constats d'infraction qui leur coûteront au minimum 444 $, incluant les frais, en vertu du Code de la sécurité routière (K03SO0329Act0264)

> Après qu'une manifestation regroupant un peu plus d'une centaine d'étudiants se fut soldée par l'arrestation de 81 d'entre eux en après-midi, une autre en

> regroupant près de 400 s'est déroulée dans le calme sous escorte policière en soirée. (K03SO0428Act0535)
>
> C'était 100 % injustifié!". Mobilisation étudiante. Des manifestants contestent une opération policière et songent au recours collectif. Le vendredi 27 avril, la police de Québec mettait fin abruptement à une manifestation étudiante sur Grande Allée, à deux pas de l'Assemblée nationale. Bilan : 81 constats d'infraction, un sommet dans la capitale depuis le début du conflit étudiant. (K03SO0506Act0634)
>
> La police a procédé à une série d'arrestations vers 23h, sans préciser le nombre étant donné que l'opération de dispersion battait son plein. Des participants avaient pourtant tenté à plusieurs reprises d'opposer leur pacifisme à la casse et au zèle de certains individus. (K08JM0426Nou0579)
>
> Plus d'un millier de personnes se sont rassemblées place Émilie-Gamelin et ont marché pacifiquement dans les rues du centre-ville, hier soir. La marche s'est mise en branle vers 20 h 50 sous la pluie. Dix minutes plus tard, déjà les policiers annonçaient que le rassemblement était illégal. « Des projectiles ont été lancés en direction de policiers à vélo au tout début de la marche », a expliqué Daniel Fortier, porte-parole du Service de police de la Ville de Montréal (SPVM). (K07PR0427Act0586)
>
> Échaudés par les débordements de la veille, au cours desquels des vitrines de commerces ont volé en éclats et 122 personnes ont été arrêtées, les policiers du Service de police de la Ville de Montréal (SPVM) se sont faits très visibles dès le début du rassemblement, vers 21 h, à la place Émilie-Gamelin. (K07PR0518Act0904)
>
> Arrestations massives à Québec : la FECQ lance un appel à la raison. (K40RC0528no_0520)

Le gouvernement est ainsi accusé d'utiliser la police comme *adjuvant* pour atteindre son objectif principal, c'est-à-dire la hausse des droits de scolarité. Au-delà des responsabilités politiques, plus difficiles à prouver, des responsabilités au niveau du corps policier ont été soulignées. En particulier, des actes de brutalité policière ont été documentés par la *Commission spéciale d'examen des événements du printemps 2012*,

qui a publié un rapport en mars 2014 en faveur du gouvernement du Québec. Dans ce document, on souligne en particulier la nécessité à se doter d'un système plus adapté pour traiter les cas de violence policière :

> Même s'il n'entrait pas directement dans le mandat de la Commission d'examiner les décisions concernant les dénonciations d'abus de pouvoir et autres comportements dérogatoires de policiers qui ont fait l'objet de plaintes déposées devant les instances déontologiques ou judiciaires, le nombre de doléances entendues nous a poussés à nous pencher sur ce sujet. Au terme de notre réflexion, nous en sommes venus à la conclusion que le système, dans sa forme actuelle, n'est pas approprié pour traiter adéquatement des cas de violences policières commises lors de manifestations. (CSEP2012, 2014, p. 317)

Dès l'entrée en vigueur de la loi spéciale, les manifestations se multiplient. Elles deviennent ainsi des *adjuvants* pour les **manifestants** qui protestent contre la **loi spéciale**. Dans cet agglomérat de clusters, les affrontements avec les policiers sont toujours en évidence et ceci, de la même manière que pour les manifestations de la première moitié de la période couverte. Le seul élément qui s'ajoute est l'utilisation des **casseroles** par les manifestants. Dans l'agglomérat analysé précédemment (5.1.3) les articles ne traitent pas des manifestations contre la loi spéciale, mais du débat autour de la loi. Dans cet agglomérat, au contraire, la loi spéciale est le motif principal des manifestations.

> Dans la rue malgré la loi spéciale. Conflit étudiant. Les manifestations se sont poursuivies dans le calme hier soir, malgré l'imposition de la loi 78 qui restreint le droit de manifester. (K03SO0522Act0864)

> Une marée humaine contre la loi spéciale. Crise étudiante. Des dizaines de milliers de personnes ont envahi les rues du centre-ville, hier, pour manifester contre la hausse des droits de scolarité et contre la loi spéciale, en ce 100e jour de grève étudiante.[...] La CLASSE, instigatrice du rassemblement, estime que 250 000 personnes ont participé à la marche. Des sources policières ont plutôt avancé le chiffre de 100 000 manifestants. (K07PR0523Act0970)

> Marche pacifique et tintamarre. CRISE ÉTUDIANTE. Les policiers encerclaient environ 250 manifestants à l'angle des rues St-Denis et Sherbrooke, vers minuit hier soir, après avoir reçu des pierres lancées par les marcheurs. La marche nocturne d'hier, la 30e depuis le début de la crise étudiante, avait débuté dans le calme, à la place Émilie-Gamelin. (K07PR0524Act1003)

> Comme depuis plus d'une semaine, la manifestation contre la hausse des droits de scolarité et contre la loi 78 a été déclarée illégale avant son départ, car aucun itinéraire n'a été fourni aux policiers. (K08JM0527Nou1013)

> De la maNUfestation à l'arrestation. Les altercations ont été musclées en fin de soirée. Les manifestants « tout nus » ont rejoint l'habituelle marche nocturne hier dans les rues de Montréal. L'ambiance, au départ festive, est devenue électrique. La police comptait plus d'une trentaine d'arrestations en fin de soirée. (K08JM0608Vot1143)

> Les casseroles déferlent. Le tintamarre se répand aux quatre coins de Montréal et du Québec. Au lendemain d'arrestations massives à Montréal et à Québec, des milliers de mécontents ont une nouvelle fois bravé l'interdit en prenant la rue - serrant fermement casseroles et cuillers de bois - sans avoir fait connaître leur itinéraire aux forces de l'ordre à l'avance. [...] Plus de 518 personnes ont été interpellées ce soir-là, à Montréal, contre 176 à Québec. (K27DE0525Soc0803)

Une partie importante de cet agglomérat traite des manifestations plus pacifiques. Toutefois, *la macrostructure de l'affrontement*, c'est-à-dire la confrontation entre manifestants et policiers, demeure aussi latente dans ce cas :

> Après les heurts, le calme. MANIFESTATIONS ÉTUDIANTES. Quelques centaines de grévistes ont défilé pacifiquement au centre-ville hier. Au lendemain d'une journée extrêmement tendue, étudiants et policiers se sont retrouvés dans un calme relatif, hier, au centre-ville de Montréal. (K07PR0309Act0144)

> Une marche pacifique malgré la tension. CRISE ÉTUDIANTE L'IMPASSE PERSISTE. Pour la quatrième journée consécutive, policiers et manifestants se sont retrouvés dans les rues de Montréal pour une marche nocturne, qualifiée

> d'illégale par les autorités moins de deux heures après son départ. (K07PR0428Act0609)

> Le calme après la tempête. 28e manifestation nocturne à Montréal. La quatrième manifestation depuis l'adoption de la loi spéciale s'est déroulée pacifiquement, contrairement aux trois rassemblements précédents qui avaient été marqués par de la violence et de nombreuses arrestations. (K07PR0522Act0954)

> Les manifestants, surtout des étudiants, mais aussi des enseignants et des travailleurs de tous âges ont d'abord monté la rue Peel, puis se sont dirigés vers la rue Sherbrooke. Plusieurs personnalités. Plusieurs personnalités publiques, dont la chef du Parti québécois, Pauline Marois, et le député indépendant. Pierre Curzi, se sont jointes à la foule. La marche s'est déroulée pacifiquement malgré la présence de quelques manifestants cagoulés du Black Bloc. (K08JM0323Nou0252)

> Manifestation pacifique et tolérée. CONFLIT ÉTUDIANT. Montréal - Même si elle a rapidement été déclarée illégale peu après s'être mise en branle, la 28e manifestation étudiante nocturne d'affilée s'est déroulée dans une ambiance bon enfant dans les rues de Montréal, hier soir. (K03SO0522Act0855)

Plusieurs autres acteurs apparaissent en s'agençant à la macrostructure principale. L'un d'entre eux est un *adjuvant* du gouvernement, le maire de Montréal monsieur **Gérald Tremblay**, qui multiplie les efforts pour contrer les manifestants. En particulier, il approuve un règlement municipal réglementant davantage les manifestations. Ce dernier devient également un outil utilisé par la police pour contrecarrer les manifestations :

> Manifestations - Tremblay songe à interdire les masques. « Une personne qui veut manifester et qui a des revendications légitimes n'a pas à se cacher ». (K27DE0317Act0204)

> Vers l'interdiction du masque. Un pas de plus a été franchi pour interdire le port du masque dans les manifestations à Montréal. (K27DE0504Soc0591)

5.1.5 Élections : l'objet de valeur

Un autre agglomérat qui se démarque à un seuil de 0,77 est composé de cinq clusters constitués comme suit : *Le Devoir* (DE_44), *La Presse* (PR_19), *Le Soleil* (SO_45), *Le Journal de Montréal* (JM_33) et *Le Journal de Québec* (JQ_10) (figure 5.9). En diminuant le seuil à 0,76, d'autres clusters s'accrochent, appartenant tous aux journaux du groupe Québecor. Les clusters les plus liés sont ceux du *Journal de Québec* avec ceux du *Journal de Montréal*, suivis par celui de *La Presse* avec *Le Devoir*. Ce dernier est central relativement au réseau sélectionné, car il possède plus de connexions que les autres.

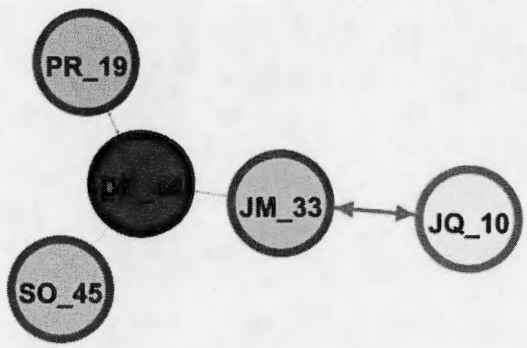


Figure 5.9 Représentation graphique de l'agglomérat traitant les élections avec le seuil du filtre des liens fixé à 0,77. Les clusters plus foncés désignent un cluster plus volumineux

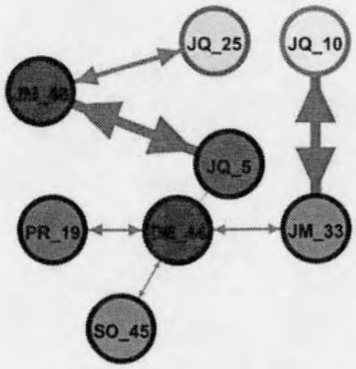


Figure 5.10 Représentation graphique de l'agglomérat traitant les élections avec le seuil du filtre des liens fixé à 0,76. Les clusters plus foncés désignent un cluster plus volumineux

Cet agglomérat est totalement concentré sur la campagne électorale et les stratégies électorales des partis politiques. Bien que les élections aient été déclenchées seulement au mois d'août, le ton électoral des interventions des politiciens dans la sphère publique est apparu bien avant cette date. En effet, il est possible d'observer dans cet agglomérat, que le « discours électoraliste » a traversé presque toute la grève étudiante, plus particulièrement du mois de juin au mois de septembre, avec un pic au mois d'août.

Cet agglomérat se distingue nettement de celui présenté au point 5.1.2, où la macrostructure est caractérisée par deux acteurs principaux, **Marois** et **Charest**, qui assument de manière alternée les rôles de *sujet* et d'*antisujet*. De plus, l'*objet de valeur* de l'agglomérat du point 5.1.2 est plus ou moins fixe et clair, car il alterne entre le concept de **discréditer l'adversaire politique** et l'objectif d'augmenter le taux d'appréciation par les électeurs. Dans l'agglomérat présenté ici, les acteurs du *sujet* sont plus variés et l'*objet de valeur* est, au contraire, stabilisé sur un objectif mieux précisé, soit **gagner les élections**. Dans ce contexte, la grève étudiante apparaît

toujours comme un des enjeux principaux de la campagne électorale, tel quel déjà observé au point 5.1.2.

Tableau 5.8 Programme narratif PN7

Sujet	*Objet de valeur*	*Adjuvant/Opposant*
Politiciens/Partis politiques	**Gagner les élections**	**Sondages/Grève étudiante**

Le rôle de la grève dans le programme narratif dont l'*objet* est l'objectif de **gagner les élections** varie selon le *sujet*. Le discours électoral est très souvent présent lors de la publication de **sondages**, qui montrent les tendances du vote des électeurs et apparaissent déjà au mois de mai. Les résultats de certains sondages décrivent la grève étudiante comme un *adjuvant* du **PLQ** :

> Le PLQ en tête à Québec. Alors que le PQ maintient sa suprématie en régions, les libéraux ont réussi à déclasser la CAQ à Québec -- une première -- et conservent aussi la pole position à Montréal par une majorité infime. Le Parti libéral se nourrit de nombreux nouveaux appuis dans la capitale nationale, vraisemblablement en raison du conflit étudiant. Il détient même une avance confortable avec 36 % des intentions de vote, devant la CAQ de François Legault, à 28 %, et le PQ qui obtient 25 %. [...] Le Parti libéral a gagné neuf points à Québec depuis janvier, alors que la CAQ en a perdu cinq. Les appuis au PQ n'ont pratiquement pas bougé. « La ville de Québec est celle qui appuie le plus le gouvernement dans le conflit étudiant. Ce sera un défi pour François Legault d'aller reconquérir la ville de Québec aux libéraux », analyse le sondeur Christian Bourque, vice-président à la recherche chez Léger Marketing. (K33JM0504Nou0682)

Dans ce contexte, la grève est un *opposant* pour le **PQ** :

> Le sondage est clair : le PLQ continue de profiter du conflit étudiant, tandis que le carré rouge demeure un handicap pour le PQ. (K44DE0616Pol1045)

Pour d'autres sondages, la grève est un *opposant* de la **CAQ** et, implicitement, du **PLQ** :

> Débat polarisé. La CAQ a été pénalisée par la très forte polarisation du débat politique ces derniers mois en raison du conflit étudiant. La position adoptée par le parti de François Legault d'appuyer le gouvernement Charest sur l'augmentation des frais de scolarité, tout en proposant de bonifier l'aide aux étudiants issus de la classe moyenne et de faciliter le remboursement des prêts, était responsable, mais, comme c'est souvent le cas pour des positions médianes, elle n'a que peu retenu l'attention. (K33JM0711Vot1365)

À la veille du déclenchement des élections, la grève semble être un *adjuvant* pour le **PQ** :

> C'est un pari risqué que prend Jean Charest en déclenchant des élections en plein été. Si le chef libéral a de l'élan, le Parti québécois de Pauline Marois est aux portes du pouvoir avec le tiers des intentions de vote. [...] « Aujourd'hui, s'il y avait élection, ce serait un gouvernement minoritaire péquiste » , [...] Au cours des derniers mois, le conflit étudiant a davantage profité aux libéraux, qui avaient vu leurs appuis gonflés. La crise n'étant plus à l'avant-scène depuis quelques semaines, le Parti libéral du Québec (PLQ) de Jean Charest glisse en seconde place, à 31 %. (K33JM0801Nou1423)

> Les répondants placent le PQ premier pour régler le conflit étudiant et choisissent les libéraux pour développer le potentiel énergétique du Québec. (K33JM0817Nou1502)

En général, les stratégies des partis font explicitement référence à la grève étudiante et ce surtout pour le **PLQ** et le **PQ** :

> En revanche, Jean Charest, qui a pris pour seule cible le PQ de Pauline Marois, table sur la polarisation autour du conflit étudiant pour s'imposer. Au désordre et à la turbulence que peut inspirer le PQ, le chef libéral oppose la stabilité, la loi et l'ordre. De son côté, la chef du Parti québécois, Pauline Marois, a diffusé un message, sans que ce soit une réplique à la publicité du chef libéral, qui misait sur l'unité, la fierté. » (K44DE0623Pol1090)

> Mais on a beau être en plein été, comme en 1994, il y a des enjeux « politisateurs » forts : crise étudiante, usure du gouvernement, dénonciation de cas de malversation, boum des ressources naturelles. « Ce n'est sûrement pas mauvais pour nous », dit un stratège péquiste. [...] Le principal talon d'Achille de Pauline Marois auquel les libéraux comptent s'attaquer sera ses positions lors de la crise étudiante. Son carré rouge, omniprésent jusqu'à la fin juin, « ça va lui nuire », insiste l'un. (K44DE0714Pol1182)

> Qui plus est, le sondage a été effectué après des semaines d'accalmie sur le front étudiant, mais le premier ministre Charest a fait en sorte que la campagne coïncide avec la rentrée dans les cégeps. Il ne s'est pas caché : la défense de la loi et de l'ordre au nom de la majorité silencieuse sera au cœur de sa campagne. Encore faudrait-il que les associations étudiantes entrent dans son jeu. (K44DE0802Pol1256)

> Dans de telles conditions, Jean Charest aura beaucoup de difficultés à ne pas être engouffré par les sables mouvants dans lesquels il s'enfonce présentement. La rentrée des étudiants ne lui a pas fourni, par ailleurs, la perche qu'il espérait lors du lancement de la campagne, le 1er août, et son Plan Nord soulève peu d'intérêt. Il s'agissait des deux piliers de sa campagne. Il ne lui reste plus qu'un coup de théâtre lors des débats de la semaine prochaine pour changer l'issue du dernier acte de la tragédie classique qu'aura été sa carrière politique. (K33JM0817Nou1505)

> Le printemps dernier, les étudiants promettaient une rentrée scolaire forcée par la loi spéciale, à compter du 15 août, qui serait tumultueuse. Le 1er août, le premier ministre a donné un caractère référendaire à l'élection qu'il lançait : les Québécois devraient choisir entre se laisser gouverner par la pression de la rue avec Pauline Marois ou la loi et l'ordre avec lui. (K33JM0901Nou1588)

> Jean Charest a fait campagne sur la stabilité, proposant le statu quo (business as usual, comme on dit en latin), mais les deux autres ont pris des engagements assez radicaux : reprise des hostilités avec Ottawa, politiques identitaires agressives, abolition de la loi 12 qui a mis fin aux protestations des étudiants pour le PQ. (K33JM0902Vot1595)

Enfin, d'autres sondages soulignent la volonté des citoyens d'aller en élections, puisque le gouvernement ne semble pas être en mesure de régler le conflit :

> Sondage Léger Marketing-Le Devoir - Le Québec veut des élections. C'est la seule façon de régler le conflit étudiant - Libéraux et péquistes sont à égalité. (K44DE0616Pol1040)

> Les résultats du dernier sondage Léger Marketing-Le Devoir le confirment : des élections générales déclenchées en août prochain -- ce que souhaitent une majorité de Québécois -- seraient une véritable boîte à surprise. (K44DE0616Pol1045)

Une configuration particulière de ce programme narratif PN7 voit le **PQ** utiliser de manière plus directe et explicite la **grève** comme *adjuvant*. Ceci se concrétise davantage par la candidature pour le **PQ** du président d'une des trois associations étudiantes, **Léo Bureau-Blouin** :

> Léo Bureau-Blouin rejoint les rangs du Parti québécois. Pauline Marois confirme la candidature de Bernard Généreux. L'ancien président de la Fédération étudiante collégiale... . L'ancien président de la Fédération étudiante collégiale du Québec (FECQ), Léo Bureau-Blouin, fait le saut en politique provinciale. Il sera le candidat du Parti québécois (PQ) dans la circonscription de Laval-des-Rapides. (K44DE0725Pol1225)

> Le recrutement de l'ex-président de la Fédération étudiante collégiale du Québec (FECQ), Léo Bureau-Blouin, comme candidat est condamnable à ses yeux et confirme que Pauline Marois « propose de céder » aux exigences des étudiants. (K44DE0802Pol1255)

5.1.6 Cégep : le retour en classe

En réduisant le seuil à une valeur de 0,7, un autre agglomérat de cinq clusters apparaît. Le cœur de cet agglomérat est constitué de clusters provenant de trois quotidiens, *Le Devoir* (DE_18), Le Journal de Montréal (JM_18 et JM_28) et *Le Journal de Québec* (JQ_7) (figure 5.11) et du site web *Radio-Canada* (RC_23). À ce même agglomérat s'ajoutent encore trois autres clusters si on diminue le seuil de 0,01, ce qui permet d'intégrer toutes les sources d'information sauf *Métro*. Trois clusters de *La Presse* (PR_52), du *Le Soleil* (SO_46) et de *Radio-Canada* (RC_28) (figure 5.12) s'ajoutent

avec ce dernier seuil. Les clusters appartenant à *Le Devoir*, *Le Journal de Montréal* et *Le Journal de Québec* sont centraux relativement au réseau sélectionné.

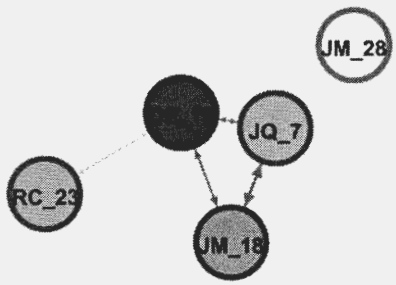


Figure 5.11 Représentation graphique de l'agglomérat traitant les élections avec le seuil du filtre des liens fixé à 0,7. Les clusters plus foncés désignent un cluster plus volumineux

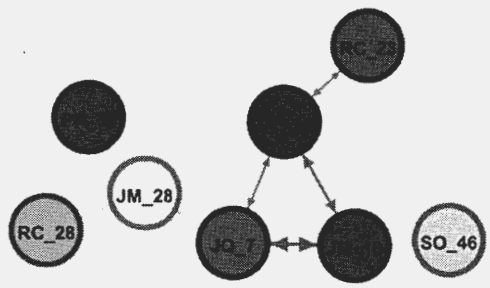


Figure 5.12 Représentation graphique de l'agglomérat traitant les élections avec le seuil du filtre des liens fixé à 0,69. Les clusters plus foncés désignent un cluster plus volumineux

Cet agglomérat met en évidence les évènements qui impliquent l'administration des **cégeps** et de leurs étudiants, avec une attention particulière aux votes de grève et aux débats sur le **retour en classe** (**PN8**). Plusieurs programmes narratifs s'entremêlent.

Cependant, une macrostructure émerge de manière dominante par l'appariement récurrent d'un actant avec un acteur, soit l'actant *objet de valeur* à l'acteur **retour en classe**.

Tableau 5.9 Programme narratif PN8

Sujet	*Objet de valeur*	*Adjuvant*	*Opposant*
Cégeps	**Retour en classe**	**Injonction/Gouvernement/ Professeurs/Carrés blancs**	**Professeurs/étudiants grévistes/**

Dans les différentes configurations prises par ce programme narratif, certaines sont plus fréquentes que d'autres. L'une d'entre elles est caractérisée par l'adjuvant **injonction** et par les conséquences de son utilisation :

> Une injonction force le retour en classe des étudiants du cégep d'Alma. (K18DE0331Act0294)

> Tension accrue entre les policiers et les manifestants, retour en classe forcé dans un cégep, procédures judiciaires et intimidation. (K18DE0412Act0372)

> Léo Bureau-Blouin, président de la Fédération étudiante collégiale du Québec (FECQ). Forcer le retour en classe des grévistes ne fera que créer des tensions potentiellement dangereuses entre les étudiants. (K18JM0412Nou0432)

Les **administrations des cégeps** sont un acteur qui intervient comme *sujet* dans le débat, mais ce n'est pas leur seul rôle actantiel. Souvent, les administrations trouvent aussi le rôle d'*adjuvant* du **gouvernement** ou, quand ils sont *sujets*, c'est le gouvernement qui devient leur *adjuvant* :

> Injonctions, actions musclées et menaces de mort. Droits de scolarité : la tension monte encore d'un cran. Manifestement loin d'avoir atteint son comble après 60 jours de débrayage, la tension est encore montée d'un cran hier, dans le conflit opposant les étudiants et le gouvernement sur la hausse des droits de

> scolarité. À bout de patience, les administrations des collèges et des universités et celle des étudiants opposés à la grève ont de plus en plus recours aux tribunaux pour mettre fin au conflit. (K18DE0413Act0374)

> Pressions pour un retour en classe. L'UdeM et le cégep Saint-Jean-sur-Richelieu vont reprendre les cours dès lundi. La voix de la ministre de l'Éducation, Line Beauchamp, qui demandait cette semaine aux établissements d'enseignement d'offrir les cours malgré le boycottage, semble avoir été entendue. Le cégep Saint-Jean-sur-Richelieu prend exemple sur le collège de Valleyfield, qui a imposé un retour en classe jeudi, un retour toutefois bloqué par les manifestants opposés à la hausse des droits de scolarité ces deux derniers jours. (K18DE0414Act0384)

> « La fin de la récréation. Bureau parlementaire - La ministre de l'Éducation, Line Beauchamp, demande aux cégeps et aux universités de donner leurs cours aux étudiants qui voudront bien y assister. Tant pis pour les absents. » (K07JQ0412Nou0354)

Au fur et à mesure que la grève se prolonge, la situation dans les cégeps devient de plus en plus complexe et l'**annulation de la session** devient un des scénarios les plus probables. La menace de l'annulation joue un rôle d'*adjuvant* dans le débat pour le **retour en classe** :

> Situation bientôt critique dans les cégeps. La situation est sur le point de devenir critique dans les cégeps, où les étudiants sont en grève depuis un mois ou plus. Certains cégeps en sont à élaborer des scénarios de reprise des cours avant les vacances d'été. Dans au moins une quinzaine de cégeps sur 48, les étudiants sont en grève depuis un mois ou plus. (K07JQ0403Nou0280)

> Les sessions seront-elles annulées? Au moment où les associations étudiantes s'avouent « dépassées » par les réactions de leurs membres, la Fédération des cégeps envisage l'option de l'annulation de la session. (K07JQ0509Nou0619)

> Le Conservatoire de musique de Montréal, où la grève dure depuis quatre semaines, a annoncé que la session des étudiants qui avaient cumulé deux absences et plus dans chaque cours était annulée. Les cours seront repris à la

> prochaine session, sans remboursement et avec la mention « incomplet temporaire ». (K18DE0421Act0440)

> « La session est foutue ». Plusieurs profs et cégépiens commencent à souhaiter l'annulation ou l'abandon de la session. Alors que l'espoir d'une sortie de crise est ténu, le chaos persiste dans les cégeps. Pour certains, vouloir sauver la session est devenu mission impossible. « Je considère que la session est foutue en raison de la nature même des injonctions et de l'absence de dialogue entre le gouvernement et les étudiants », a confié au Devoir Olivier Ménard, professeur de français au collège Montmorency. (K18DE0512Soc0672)

Le **gouvernement** et les **administrations des cégeps** élaborent conjointement un plan pour permettre le **retour en classe** et en arriver à une **sortie de la crise** :

> Rencontre pour une sortie de crise à Québec. Beauchamp et Charest rencontrent la direction des cégeps et des universités. Dans l'impasse depuis de nombreuses semaines, le conflit étudiant semble avoir atteint un tournant. (K18DE0504Soc0585)

> Incertitudes quant à la rentrée à l'automne. La session commencera à la mi-août dans la plupart des cégeps. De quoi auront l'air les sessions à la rentrée ? Dans un contexte de loi spéciale et de conflit dans une impasse profonde depuis la rupture des négociations, plusieurs se le demandent. (K18DE0601Soc0890)

> Bouchées doubles pour les cégeps. Pendant que les étudiants du collégial concernés par la grève printanière se prononçaient pour ou contre un retour en classe cette semaine, les 14 cégeps touchés se préparaient à une rentrée scolaire et à un automne sur un mode intensif jamais vécu auparavant. (K18DE0818Soc1372)

> Retour confirmé dans 3 cégeps. Dur coup hier pour le mouvement étudiant. Les étudiants de trois cégeps ont choisi de mettre fin au boycottage des cours et de rentrer en classe, abaissant à un peu plus de 100 000 le nombre de grévistes. Tour à tour hier, les étudiants du Cégep Édouard-Monpetit, du Cégep Marie-Victorin ainsi que du Collège de Maisonneuve ont choisi de mettre un terme à la grève étudiante. (K07JQ0814Nou1241)

Au mois d'août, les derniers **votes de grève** couvrent explicitement un rôle d'*opposant possible* face aux plans du gouvernement et des administrations des cégeps :

> Dur coup pour la grève générale illimitée. Trois cégeps ont voté hier pour le retour en classe la grève se poursuit au Cégep du Vieux-Montréal. Les mandats de grève générale illimitée tombent un à un dans les cégeps. Le Collège Édouard-Montpetit, le Cégep Marie-Victorin et le Collège de Maisonneuve ont voté le retour en classe aujourd'hui, s'ajoutant aux trois cégeps qui ont ratifié la fin de la grève la semaine dernière. (K18DE0814Act1343)

> La grève prend fin dans le réseau collégial. Les cégeps du Vieux-Montréal et de Saint-Laurent votent pour le retour en classe. Le vote pour un retour en classe au cégep du Vieux-Montréal (CVM) et au cégep de Saint-Laurent, deux bastions parmi les plus militants, a signé hier la fin de la grève générale illimitée dans tout le réseau collégial. (K18DE0818Soc1373)

Dans ce contexte, les **professeurs** des cégeps ont un rôle important et jouent souvent celui d'*opposant actif* et ceci, en raison de leur appui à la grève. Dans plusieurs cas, ils montrent un soutien actif à la grève, avec des prises de position nette (K07JQ0411Nou0338). D'autres fois leur soutien est indirect, mais il demeure actif (K18JM0411Nou0429, K18JM0519Nou0904).

> Les directions des universités et des cégeps ont complètement abdiqué leurs responsabilités depuis le début du boycott des cours par les étudiants et de nombreux professeurs. Leurs devoirs consistaient à garantir le libre accès aux salles de cours, à s'assurer que les professeurs offriraient leurs cours aux étudiants qui se présenteraient et qu'ils prendraient les présences. (K07JQ0411Nou0338)

> « On ne négociera pas de calendrier de reprise de cours tant que les étudiants ne seront pas de retour en classe », a indiqué le président de la Fédération nationale des enseignantes et enseignants du Québec (FNEEQ), Jean Trudelle. (K18JM0411Nou0429)

> Des notes « bonbons »? À Valleyfield, on s'inquiète du rattrapage automnal. Le Collège de Valleyfield se demande comment il fera pour rattraper 11 semaines de grève en un peu plus d'un mois. De son côté, le syndicat des professeurs lance un cri d'alarme : dans les conditions imposées Québec, on craint d'être obligé de lésiner sur la qualité de l'enseignement et de donner des « notes bonbons » aux étudiants. (K18JM0519Nou0904)

Leur rôle d'*opposant* se poursuit dans d'autres configurations actantielles, notamment en raison des conditions de travail qui leur sont imposées pour le rattrapage de la session :

> Qui va payer la facture de 20 millions pour la reprise de cours ? Grève -- Cégeps. Les négociations s'amorcent aujourd'hui entre les enseignants et les cégeps où les sessions d'études ont été interrompues en raison de la grève des étudiants cet hiver. (K18JM0607Nou1129)

> Aucune entente à l'horizon. ÉDUCATION -- SESSION À RATTRAPER Un écart de 30 M$ sépare Québec de ses profs de cégep. Rien ne va plus entre Québec et les professeurs de cégep, alors que ces derniers songent à exercer des moyens de pression lors de la reprise des cours en août. Malgré plusieurs jours de négociations, les professeurs n'arrivent pas à s'entendre avec la Fédération des cégeps et le ministère de l'Éducation quant au rattrapage de la session d'hiver. (K07JQ0627Nou1093)

> Entente pour la rentrée. Cégeps. À la veille de la rentrée scolaire, le ministère de l'Éducation a accepté de gonfler les effectifs professoraux dans les 14 institutions collégiales les plus touchées par le conflit étudiant, mais pas à n'importe quelles conditions. Si le conflit se prolonge, les nouveaux professeurs ne seront pas payés. Le ministère de l'Éducation a annoncé hier qu'il compenserait les cégeps pour l'embauche de l'équivalent de 180 postes de professeur à temps plein, à la suite d'une entente de principe entre la Fédération nationale des enseignants du Québec, affiliée à la CSN (FNEEQCSN), et le Comité patronal de négociation des collèges. (K07JQ0808Nou1220)

> Les professeurs seront au poste. Quelque 180 enseignants de plus devront être embauchés. Un des morceaux du casse-tête de la rentrée est tombé en place hier : les enseignants du collégial ont obtenu les ressources supplémentaires qu'ils

exigeaient pour la reprise de la session d'hiver et la session d'automne condensée. (K18DE0808Pol1296)

5.1.7 Quelle est votre opinion?

Un dernier agglomérat apparaît avec un seuil de 0,67. Il est composé de sept clusters appartenant à cinq quotidiens (figure 5.13), *Le Devoir* (DE_11 et DE_15), *La Presse* (PR_34 et PR_47), *Le Journal de Montréal* (JM_21), *Le Journal de Québec* (JQ_53) et *Le Soleil* (SO_9). Cet agglomérat rassemble les clusters les plus volumineux de *La Presse* et du journal *Le Devoir*, et les troisièmes en ordre de grandeur du *Journal de Montréal* et du journal *Le Soleil*. En d'autres termes, sauf pour le *Journal de Québec* (JQ_53 ne possède que 11 articles) et *Métro* (dont aucun cluster n'a été sélectionné), cet agglomérat regroupe les clusters les plus volumineux de chaque journal (tableau 5.10). Ces résultats correspondent également avec les résultats sur la fréquence du code « opinion_difficile_à_annoter » (cfr. 4.5.3) car, comme il est possible de visionner dans le tableau 5.10, ce sont les journaux *La Presse* et *Le Devoir* à avoir le plus de clusters dans l'agglomérat présent et les clusters les plus volumineux.

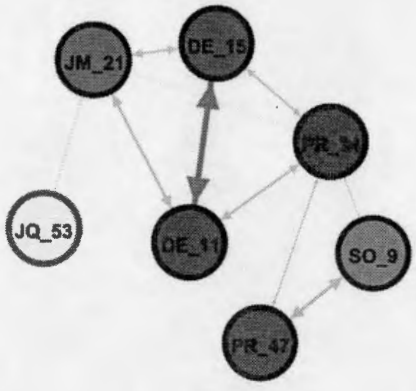


Figure 5.13 Représentation graphique de l'agglomérat traitant les articles d'opinion. Les clusters plus foncés désignent un cluster plus volumineux

Tableau 5.10 Taille de chaque cluster avec son rang dans la partition du journal correspondant

Cluster	Taille	Rang
PR_34	67	1
DE_11	61	1
PR_47	60	2
DE_15	60	2
JM_21	53	3
SO_9	46	3
JQ_53	23	11

Dans cet agglomérat, on observe un phénomène fréquent : les clusters regroupent des articles assez variés et pour lesquels *il n'émerge pas de macrostructure commune*. Ceci constitue le seul cas pour lequel *l'hypothèse de recherche ne correspond pas aux phénomènes observés*. Pour cette raison, il est donc difficile d'en faire un résumé exhaustif. Il s'agit d'une exception et elle ne remet pas en cause l'hypothèse, mais il souligne la complexité de l'analyse des artefacts sémiotiques. Cet élément sera amplement discuté dans le prochain chapitre (cfr. 6.1).

Néanmoins, il existe un élément qui constitue un commun dénominateur mais qui ne permet pas d'interpréter cet agglomérat par le biais d'une macrostructure tel que définie dans cette thèse. Ce dénominateur commun est *l'opinion*. En d'autres termes, la majorité des articles sont des articles d'opinion sur différents évènements liés à la grève. Ainsi, on observe la présence massive d'éditoriaux ou de lettres des lecteurs et plusieurs opinions sont exprimées également dans des chroniques ou articles qui n'appartiennent pas nécessairement à la classe des éditoriaux.

Le cluster JQ_53 est celui qui a la connexion la plus faible avec les autres clusters et dans lequel on observe l'expression d'opinions sur des sujets distincts des autres clusters. Le sujet principal est le parti politique **Québec Solidaire**. Le plus souvent,

les opinions présentés dans ces articles portent sur les interventions ou les événements dans lesquels le co-chef de ce parti politique, **Amir Khadir**, est un protagoniste :

> J'ai honte. Je trouve honteux et innommable de mettre les menottes à un homme ayant titre de député élu, Amir Khadir qui, certes, représente la gauche, mais ça en prend au moins un qui ait le courage de le faire dans un pays où Dieu est le dollar. (K53JQ0608Vot0964)

> Les pays scandinaves. Au lieu de « pelleter dans la cour des familles, de la classe moyenne et des étudiants les responsabilités fiscales qui incombent à l'ensemble de la société », le gouvernement Charest devrait plutôt suivre l'exemple des pays scandinaves qui prônent la gratuité scolaire, estime le député solidaire. Aux yeux d'Amir Khadir, l'entêtement de Jean Charest à vouloir dégeler les frais de scolarité mérite que le gouvernement libéral soit défait. « Nous sommes confiants qu'une majorité de la population québécoise est contre le doublement des frais de scolarité », a-t-il affirmé. (K53JQ0229Nou0069)

Ce sujet n'est toutefois pas exclusif à JQ_53. Des opinions sur ce même sujet sont exprimées dans d'autres clusters de cet agglomérat :

> Une nouvelle forme d'activisme politique. Les arrestations du député Amir Khadir et de sa fille illustrent avec force la nouvelle forme d'activisme politique plus contestataire mené par les mouvements sociaux, en lien avec Québec solidaire (QS), qui agit comme leur porte-parole dans les institutions de l'État. Monsieur Khadir n'est pas que le député du Plateau. Il est le lobbyiste en chef de la société civile organisée à l'Assemblée nationale. QS est une fédération de mouvements sociaux issus des syndicats, du tiers secteur, de la lutte contre la pauvreté, du féminisme et de l'environnement. (K15DE0608Idé0972)

Cet agglomérat des clusters où la phase d'annotation a été caractérisée par le même obstacle en l'absence de macrostructure commune. En effet, les articles font référence à beaucoup de programmes narratifs différents, le niveau discursif devient complexe et rend ainsi difficile l'interprétation globale des clusters et, par conséquent, de

l'agglomérat. Par exemple, dans l'extrait suivant, on observe plusieurs idées et opinions énoncées sur plusieurs sujets, ce qui est un phénomène fréquent dans les articles de cet agglomérat :

> Une ville au bord de la crise de nerfs. Je sais que vous êtes comme moi : plus capables, vraiment plus capables de supporter cette crise préfabriquée, et artificielle, qui est devenue un imbroglio syndicalo-socio-politique hors de contrôle et qui met en relief tout ce qui ne marche pas bien et n'est pas beau au Québec en ce moment. La température maussade qui refuse de nous laisser voir un peu de printemps n'aide pas, c'est certain. Mais je pense que le vacarme du gros hélicoptère de la SQ, qui survole le centre-ville à chaque soir est très démoralisant - d'une manière insidieuse. [...] UNE SOCIÉTÉ DYSFONCTIONNELLE. Des étudiants ont recours aux tribunaux pour pouvoir suivre leurs cours! Mais les ordres de la Cour sont défiés par d'autres étudiants - et des syndicats de profs payés à ne rien faire. Les directeurs d'école plient et annulent les cours. Les leaders étudiants, qui ont refusé les offres du gouvernement, s'improvisent grands mandarins, proposant toutes sortes de solutions administratives pour rencontrer leurs exigences. [...] UNE CAMPAGNE DE SABOTAGE. Et je ne vous dis pas ce qu'on lit sur les médias sociaux : une litanie d'injures exagérées, de désinformation, de démagogie hargneuse, de dénonciations malhonnêtes, très souvent écrites dans un français déboîté. Au début on applaudissait les étudiants, festifs, énergiques, qui défilaient - on sympathisait avec eux. On n'en est plus là. Maintenant, ce sont ces mêmes étudiants qui angoissent, devant une année perdue, ou une carrière gâchée et qui veulent reprendre leurs cours. [...] Le Québec n'a pas besoin d'ennemis. Il est parfaitement capable de s'affaiblir et de se détruire tout seul. (K21JM0504Nou0692)

Toutefois, il est possible de regrouper la majorité des articles dans une même catégorie, soit le débat entre ceux qui sont *pour la grève* et ceux qui sont *contre la grève :*

> Une minuscule minorité. Depuis le début du boycottage des cours, les leaders du mouvement et les médias nous répètent que « les » étudiants sont en colère, que « les » étudiants sont descendus dans la rue. La réalité, c'est que les jeunes participant aux manifestations constituent une petite minorité, parfois même une minuscule minorité. [...] En effet, ils oublient que leurs « actions d'éclat »

causent chaque jour plus d'ennuis aux citoyens, ceux-là mêmes qui paient le gros des coûts des études universitaires. (K34PR0309Déb0140)

Les étudiants sont en journées pédagogiques. Les étudiants sont en grève. C'est ce que rapportent tous les médias. Pourtant, il me semble que ce n'est pas le bon mot pour décrire la situation. Nous devrions peut-être retourner sur les bancs d'école, nous aussi, réviser nos expressions. Une grève est une action collective consistant en une cessation concertée du travail par les salariés d'une entreprise. Les étudiants ne sont pas des travailleurs, pas encore. (K34PR0310Act0147)

Quoi faire? Le dynamisme et la ferveur des jeunes Québécois ont été confisqués par un petit groupe d'idéologues et d'extrémistes invoquant leurs supposés droits pour abuser de ceux des autres. (K53JQ0610Vot0978)

Je n'ai jamais été favorable à une intervention hâtive des premiers ministres dans les conflits comme celui des étudiants. Ils doivent éviter de jouer aux pompiers chaque fois que s'annonce un problème. Mais il arrive un point où seule l'intervention du premier ministre peut changer le cours des choses. Et je pense honnêtement qu'on est rendu là. Dans ce genre de crise, c'est au chef du gouvernement qu'il incombe de jouer le rôle d'un « bon père de famille » et de réconcilier tout le monde. (K09SO0524Act0890)

Jean Charest a raison de se défendre âprement lorsqu'il est suggéré que son gouvernement a orchestré la crise avec les étudiants dans l'objectif de s'en servir pour atténuer l'insatisfaction à l'endroit des libéraux. Ça serait pousser le machiavélisme un peu loin. Mais ça ne veut pas dire que sa gestion du dossier a été exemplaire pour autant, ni exempte de toute considération politique. (K09SO0507Édi0649)

Dans les endroits malheureux où une vraie révolution s'impose, la dernière chose dont les citoyens ont besoin est de se faire expliquer pourquoi ils devraient faire la révolution. Ils le savent trop bien. Mais ici, au Québec? Ici au Québec, c'est différent. Nous sommes pleins. Gras dur. Subventionnés, syndiqués. « Paddés » mur à mur. Protégés par la convention, permanents, indexés... Et nous avons des richesses naturelles... pour ceux que cela intéresse. Alors, la question est : pour qui est-ce que le discours de gauche nous dépeint-il toujours comme des perdants, des victimes, des gens à qui on doit « donner une chance »? (K21JM0508Vot0738)

Dans ce contexte, quelques caractéristiques spécifiques peuvent être soulignées. L'élément remarqué le plus fréquemment est la quasi omniprésence dans les articles d'un acteur qui recouvre explicitement le rôle de *destinataire*. Ceci est probablement dû au fait que, lorsqu'on exprime une opinion, on a tendance à évaluer un acte de performance, comme par exemple, évaluer une action accomplie par un héros. Un acte de performance implique la présence d'un destinataire : par exemple, le roi évalue positivement le fait que le héros ait sauvé la princesse, et leur permette ainsi de se marier. Dans plusieurs articles, les opinions reflètent des jugements sur des interventions publiques ou des actions accomplies par les acteurs impliqués dans le débat. Un des critères de cette évaluation est l'impact que les actions ont dans la **société**, qui constitue un des *destinataires* plus fréquemment observés :

> Je ne pense aucun bien de ceux qui jouent à la révolution et sont tentés par la violence. Aucun. Mais réduire la grève étudiante à quelques casseurs et apprentis Che Guevara, c'est ne rien comprendre à ce qui se passe au Québec. Car la querelle sur les frais de scolarité a changé de proportion depuis un bon moment. La grève étudiante est le révélateur d'un malaise social profond. (K21JM0419Nou0503)

> La jeunesse qui pousse le Québec à la maturité. Face à l'échec du modèle d'autorité néolibéral, les « Y » avancent une vision humaniste à long terme. (K11DE0428Soc0531)

> Une fière jeunesse, M. Reid ! Monsieur Reid, je ne sais pas ce que vous pensez, dans votre for intérieur, de ces disciplines que vous jugez peut-être trop improductives, mais je puis vous dire que ces étudiants qui, depuis le 20 février, sont dans les rues tous les jours, qui font preuve d'une inventivité, d'une patience, d'une créativité inouïes, sont parmi les plus brillants, allumés et impliqués dans leurs études que je connaisse. (K11DE0510Idé0655)

> Le poing sur la table. Il n'appartient pas aux étudiants, mais au gouvernement du Québec, de faire les difficiles choix de société. (K34PR0419Déb0501)

> Point chaud - Face-à-face de générations. « Les jeunes se réveillent, ne les écrasez pas! », demande Jean-Marc Léger. Une fois un trait tiré à la grève étudiante, « les Québécois ne verront plus les jeunes de la même manière », selon le président et fondateur de Léger Marketing, Jean-Marc Léger. (K11DE0522Soc0769)

> Obstruction systématique. Pour protéger le portefeuille des étudiants, leurs leaders se drapent dans des principes de « justice sociale » et de « choix de société ». Depuis plusieurs jours, je vois de nombreux chroniqueurs et commentateurs donner une image idéalisée des étudiants qui militent dans la rue. Changer le monde, défendre des idéaux, se battre pour la justice, lancer le printemps québécois... autant de raisons pour lesquelles les jeunes descendraient dans la rue avec une fougue qui fait envie. Je suis désolée de briser cette image romantique pour révéler des côtés moins flatteurs de cette génération dont je fais partie. Dans mes fonctions au sein de la CDJ-ADQ, j'ai côtoyé plusieurs de ces leaders étudiants. Plusieurs d'entre eux sont proches soit de partis politiques, soit de groupes communistes visant à faire de l'obstruction systématique à tout fonctionnement de la société telle qu'on la connaît. (K34PR0426Déb0575)

Un autre exemple de *destinataire* observé est l'**éducation** :

> Ras-le-bol des idées néolibérales. « Résumer le conflit à un choc des générations serait une façon commode d'en évacuer l'aspect idéologique ». Le gouvernement n'arrivera jamais à rien de bon, dans le conflit étudiant, tant qu'il ne comprendra pas mieux à qui et à quoi il a affaire, c'est-à-dire à une génération différente des autres, dont les intérêts débordent les questions d'éducation et qui n'a pas fini de prendre la rue pour se faire entendre. (K11DE0526Soc0814)

> Grève étudiante : au-delà des sous. Le mouvement de grève étudiante qui a démarré sur un refus de la hausse des droits de scolarité autour du slogan « bloquons la hausse » a pris une coloration différente de celle qu'il avait au départ du fait de l'intransigeance du gouvernement et des effets politisants de l'action politique elle-même. [...] Plus encore, en défendant le droit à l'éducation, les étudiants et ceux qui les appuient fraient la voie à une autre conception de l'éducation et de la société que celle qui prévaut actuellement, un peu plus près de celle que défendait Condorcet lors de la Révolution française. [...] Dans ces conditions, l'intransigeance du gouvernement a permis au mouvement de se déployer et de se radicaliser. Quand un gouvernement n'a que

> la police à offrir à sa jeunesse en colère, il y a lieu de s'inquiéter. Pas tant pour la jeunesse que pour le gouvernement. (K11DE0404Idé0327)

D'autres articles identifient la valeur de la **démocratie** comme critère d'évaluation :

> Les étudiants sont de plus en plus nombreux à se joindre au mouvement de grève afin de contrer la hausse des frais de scolarité universitaires. Pourtant, il y a moins de manifestants dans les rues. La question est donc de savoir si le vote est vraiment représentatif de l'opposition à l'augmentation des coûts ou s'il ne s'agirait pas plutôt d'un militantisme actif qui prendrait le pas sur une majorité silencieuse qui préférerait ne pas se prononcer. (K21JM0308Nou0122)

> [...] selon Sébastien Ricard, qui n'hésite pas à relier l'effervescence actuelle à la naissance d'un éventuel « printemps québécois ». « Le timing est parfait, observe-t-il. On est dans une période au Québec ou la disponibilité d'esprit et de temps est là pour s'attarder à ces questions et pour vraiment mettre en marche les idées et les débats. [...] L'effervescence et les débats qui se sont dégagés de ces événements nous ont inspirés et nous ont encouragés à se pencher nous aussi sur la notion de la démocratie, indique Ricard. » Évidemment, on n'avait pas du tout prévu que la grève étudiante serait déclenchée entre-temps et que Dominic Champagne (qui participe également à notre événement) organiserait sa mobilisation générale le 22 avril (à l'occasion du Jour de la Terre). (K21JM0407Wee0389)

> Élections - Le panneau tendu aux étudiants. Les représentants et porte-parole des associations étudiantes... . Les représentants et porte-parole des associations étudiantes québécoises comprenaient déjà, et ils le disent maintenant, que la question des droits de scolarité soulève, compare, voire oppose différentes conceptions de la nature et du fonctionnement de la société et de la démocratie. (K15DE0714Idé11)

Enfin, d'autres articles portent plus d'attention à l'**accessibilité aux études**, qui, au contraire d'éducation et démocratie, n'est pas une *valeur idéologique*. Dans ces cas, la manière de respecter ces valeurs peut changer, mais ils demeurent des valeurs amplement partagées par la société. Dans le cas de l'accessibilité, le niveau d'abstraction est différent et il pourrait ne pas être accepté comme destinataire. Ce qui n'est pas possible pour les autres deux exemples.

> À mes camarades au carré rouge... Je porte le carré rouge et suis contre la hausse des droits de scolarité. Selon moi, afin de permettre la meilleure accessibilité possible aux études supérieures et d'enrichir ainsi cette belle société qui est la nôtre, la gratuité scolaire serait envisageable dans un avenir rapproché, bien qu'elle ne soit possible qu'à la suite de plusieurs modifications au système actuel. . L'accessibilité à l'université est un choix de société. (K15DE0514Idé0693)

> UNE GRÈVE LÉGITIME. Grève des étudiants justifiée? [...] En augmentant les droits, on empêchera certaines de ces familles d'envoyer leurs enfants à l'université. D'ailleurs, toutes les études sérieuses démontrent que lorsque les droits augmentent, moins d'étudiants s'inscrivent. (K34PR0216Déb0018)

> En classe! Il est temps que la ministre Line Beauchamp lance un ultimatum aux manifestants. Ce n'est pas une grève, car les étudiants ne sont pas syndiqués. Tout le monde est perdant dans ce conflit. À ce que je sache, c'est le gouvernement qui doit suggérer des solutions. Ce ne sont pas les étudiants qui doivent faire des propositions. Les étudiants devraient apprendre à budgéter. (K09SO0509Opi0683)

Dans d'autres cas, les énonciateurs des articles ont recours à plusieurs destinataires en même temps :

> Dix grandes leçons. L'affrontement entre les étudiants et le gouvernement au sujet des droits de scolarité nous révèle quelques enseignements sur la société québécoise. 1 Les « grévistes » veulent être écoutés, mais n'entendent rien.[...] 2 Les « grévistes » ne veulent pas négocier, mais gagner. [...] 3 La démocratie étudiante est gravement malade. [...] 4 Ce mouvement est corporatiste et même réactionnaire. [...] 5 Le terrorisme est de retour au Québec, sous l'appellation de désobéissance civile. [...] 6 Les syndicats et les groupes populaires ont fait dévier le débat. [...] 7 Les enseignants qui soutiennent ou même encouragent les « grévistes » nuisent à la profession. [...] 8 Le gouvernement Charest ne peut pas céder à ces revendications. [...] 9 Le recours aux tribunaux est entré dans les moeurs. [...] 10 La démocratie est à la fois vulnérable et solide. Les Che et autres Trotsky en herbe jouent une pantomime de Mai 68 et s'imaginent faire la révolution. (K34PR0516Déb0872)

Pour un certain nombre de cas, il est toutefois moins simple d'expliciter l'acteur correspondant au rôle de destinataire. Dans certains autres cas, il est par contre possible d'identifier un concept autour duquel gravitent les considérations énoncées, comme le cas du concept de **participation politique** :

> Point chaud - L'enjeu philosophique mondial du conflit étudiant. « Ayons toujours un oeil sur le Québec », conseille Alain Badiou. Le Québec et sa crise actuelle pourraient-ils servir à mieux penser le monde? Oui, assure Alain Badiou, ancien leader de Mai 68 qui ne renie pas son passé maoïste. Rencontre avec un des philosophes français les plus connus et controversés de l'heure. Que pensez-vous du conflit étudiant au Québec? Ce qui m'intéresse d'abord, c'est l'amplitude et la détermination du phénomène. Au fond, ce qui se passe chez vous, c'est une résistance brutale et étendue à un phénomène mondial, qui veut que le modèle de l'entreprise s'applique à toutes les activités humaines, quelles qu'elles soient. Comme l'entreprise, l'université devrait s'autofinancer, alors qu'historiquement, elle s'est édifiée selon des règles toutes différentes. Évidemment, le conflit prend la forme particulière et très localisée d'un combat contre le programme d'augmentation des droits de scolarité universitaire, qui s'est ensuite étendu à une opposition contre la gestion gouvernementale de la crise. Mais on sent bien au coeur de ce soulèvement une subjectivité révoltée contre l'idée que le paradigme de toute chose est l'entreprise. Et ce point de résistance mobilise, pour l'instant, un débat de grande ampleur, qui nous concerne tous, et dont la fin n'est pas prédictible. Feriez-vous un rapprochement avec la révolte étudiante de Mai 68, alors que, dirigeant maoïste, vous appeliez à la révolution? Oui, par ses manières de faire, ses allures, son inventivité. (K11DE0611Soc0998)

> Être jeune et faire de la politique autrement. Le mouvement étudiant actuel pourrait-il avoir un impact sur la vie politique québécoise? Tous les observateurs de la jeunesse québécoise et internationale des dernières années l'ont remarqué : les jeunes participent peu aux pratiques politiques traditionnelles : scrutins, adhésion à un parti, etc. Devant un tel constat, une question se pose : le mouvement étudiant actuel, qui déferle dans les rues de Montréal contre la hausse des droits de scolarité et contre la loi 78, pourrait-il avoir un impact sur la vie politique québécoise? (K15DE0602Soc0898)

Beaucoup de documents analysés dans cet agglomérat sont des lettres écrites par des lecteurs et ce surtout pour le journal *Le Devoir :*

> Grève étudiante - Individualisme contre sens de la communauté. Je suis étudiante en communication et « j'investis » dans notre futur. Je veux réussir dans la vie, j'ai un iPhone et je n'ai pas de dreadlocks. [...] Un vote de grève symbolique (de 10 jours!?) qui ne passe pas. Et pourquoi? Parce qu'en grosse majorité, on pense à son examen en mars, à son cours de thaï box et à son stage chez Cossette. Allô la communauté! Et il y a de quoi s'inquiéter. Rare. Voilà dévoilé un symptôme plus profond ici... Un symptôme inquiétant, relié directement à la hausse des droits de scolarité... Si l'éducation est en train de changer, ce n'est pas juste à cause des programmes « rentables » favorisés, de la facture étudiante qui s'alourdit, de la gestion et du financement des universités qui se privatisent; c'est aussi... à cause des étudiants. (K11DE0218Idé0018)

La présentation et la synthèse des annotations de ces clusters ne sont pas exhaustives et un approfondissement de cet agglomérat serait vraisemblablement pertinent. Comme ceci ne coïncide pas avec les objectifs de notre recherche cette analyse demeurera incomplète et approximative.

5.2 Groupes de macrostructures distinctives

Dans la section précédente, nous avons mis en évidence les macrostructures les plus récurrentes et les plus communes entre les différents journaux. Pour ce faire, nous avons projeté les centroïdes des clusters sur un espace commun afin d'obtenir une visualisation de leurs relations à l'aide de la mesure de similarité cosinus. En configurant un seuil pour filtrer le poids des liens entre les nœuds, nous avons obtenu différentes représentations des relations inter-clusters sous forme de graphiques, à partir desquels il a été possible d'identifier des agglomérats de clusters similaires. Ces agglomérats constituent les « grandes » macrostructures communes entre les journaux.

Dans cette section, l'opération inverse est exécutée. Ainsi, à partir de l'espace des centroïdes, on construit un protocole pour identifier les clusters qui se différencient le plus et ceci, en tenant compte de l'appartenance au journal. Or, l'identification des différences entre clusters et entre journaux constitue une opération plus difficile à

réaliser compte tenu de la nature des hypothèses de cette recherche et de la méthode construite. En effet, cette dernière, n'est pas optimisée pour ce genre de tâche, puisqu'elle est basée sur la notion de similarité. Toutefois, un protocole a été identifié pour effectuer, avec quelques limites, ce genre d'analyse très pertinente dans une phase d'exploration d'un corpus de type journalistique.

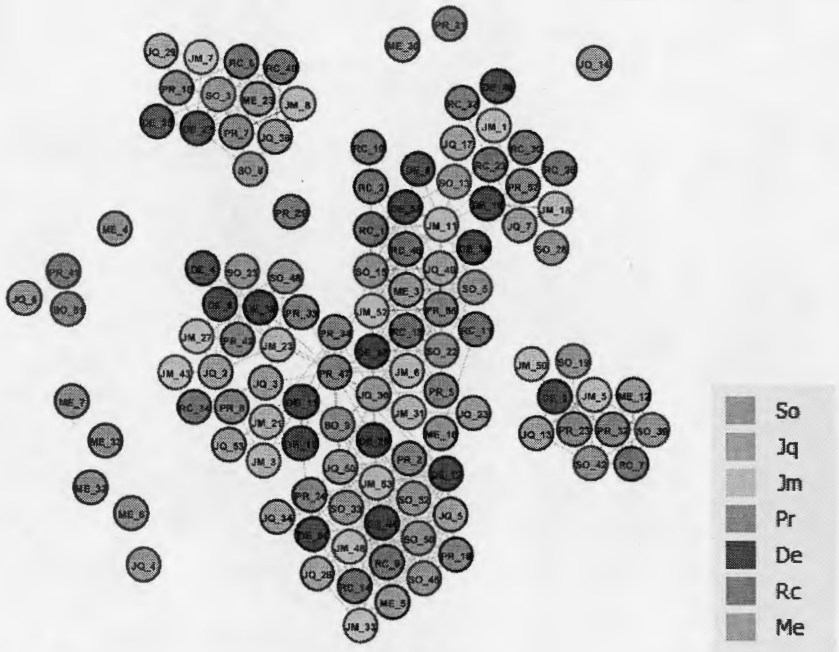


Figure 5.14 Représentation graphique des 117 clusters qui ont été annotés avec le seuil du filtre des liens fixé à 0.6. Chaque couleur correspond à une source d'information.

Le protocole a été construit de manière spéculaire par rapport à celui utilisé pour identifier les agglomérats de macrostructures. Le principe de base est d'*observer les clusters les plus isolés du graphique*. En effet, ceux qui possèdent le moins de connexions sont peu similaires au reste des clusters. Pour ce faire, deux paramètres ont été modifiés dans la phase de construction du graphique. Le premier est le nombre de nœuds retenus. Ainsi, seuls les 117 clusters passés à l'étape d'annotation en raison de leur taille ont été retenus. En d'autres termes, seulement le tiers plus volumineux

des clusters de chaque journal a été retenu pour la construction du graphique. Ceci implique que seulement 117 nœuds ont été projetés. Le deuxième paramètre est le seuil pour filtrer le poids des liens entre les nœuds. Cette valeur a été fixée à 0,6 afin d'obtenir la figure 5.14. On y remarque certains nœuds peu connectés qui se situent aux marges et d'autres plus connectés qui demeurent au centre du graphique. La couleur du nœud dépend du journal. À l'occasion, d'autres seuils plus faibles seront utilisés.

5.2.1 Métro

Dans le graphique ainsi construit (figure 5.15), on observe plusieurs clusters appartenant au journal *Métro* qui se situent aux marges de l'espace. Quatre clusters sont susceptibles d'être analysés : ME_4, Me_6, ME_7 et ME_31. Pour des raisons de simplicité, nous ne rapporterons que l'analyse de ME_6 et ME_7. Ces clusters montrent deux profils différents. Le premier est caractérisé par de très forts liens avec un autre agglomérat et ne présente pas des caractéristiques spécifiques et distinctives. L'autre, au contraire, montre une typologie de cluster qui regroupe des caractéristiques spécifiques au journal.

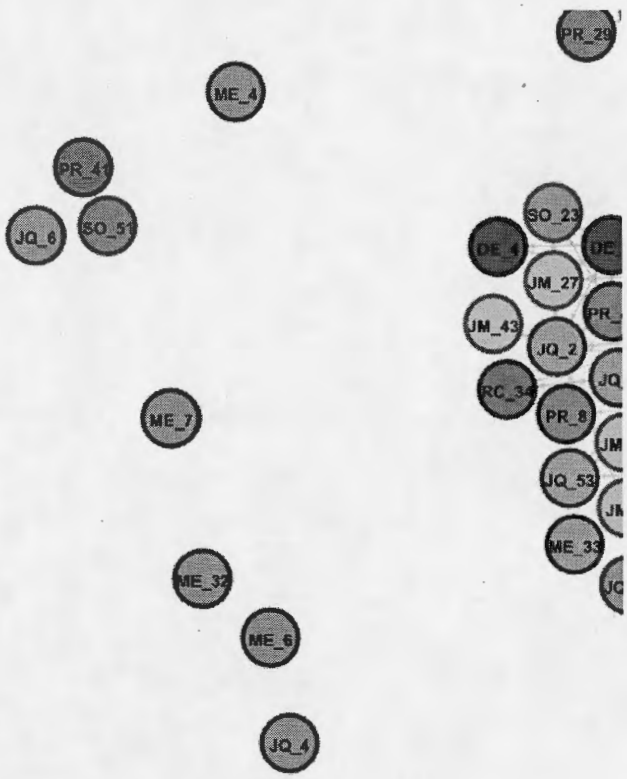


Figure 5.15 Détail de la figure 5.14

5.2.1.1 Les opinions exprimées dans *Métro*

Dans le cas de ME_6, nous avons d'abord exploré les liens que ce cluster possède. Ainsi, avec un seuil fixé à 0,5, dix clusters s'y retrouvent reliés. Ces liens nous fournissent des indices pour étudier le cluster ciblé. En effet, les liens du cluster ME_6 (figure 5.16) indiquent clairement que ME_6 est relié à l'agglomérat présenté au point 5.1.7, c'est-à-dire au groupe des articles d'opinion. Il est possible d'extraire des exemples à partir de clusters comme DE_11, DE_15 (*Le Devoir*), PR_34, PR_47 (*La Presse*), JM_21 (*Le Journal de Montréal)* et SO_9 (*Le Soleil*). Enfin, une analyse plus détaillée du cluster ME_6 montre que ses composants sont des articles d'opinion et qu'ils sont aussi caractérisés par les mêmes schémas logico-sémantiques qui ont été observés pour les clusters appartenant à l'agglomérat présenté au point 5.1.7.

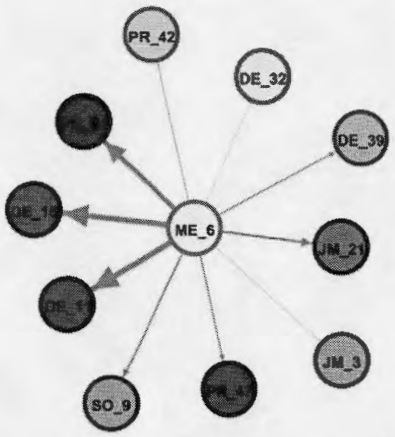


Figure 5.16 Représentation graphique des liens du cluster ME_6

De la même manière, les articles de ce cluster mettent surtout en valeur le *destinataire*. Les plus répandus sont les **droits fondamentaux** et la **démocratie** :

> Lettre d'une étudiante. N'importe quelle société est censée respecter les droits humains fondamentaux. Un de ces droits est celui à la santé, c'est pourquoi nous avons accès à un système de santé universel au Québec. Autre droit de l'humain : celui à la justice. Voilà d'où nous viennent les procès qui servent à juger des cas d'injustice. Ces temps-ci, le gouvernement Charest se plaît à jouer avec un des droits humains fondamentaux : celui à la liberté. (K06ME0227Act0035)

> Nos étudiants sont ainsi devenus nos maîtres et nous font maintenant une leçon de démocratie! (K06ME0307Car0064)

> Il semble que nos étudiants comprennent mieux que nos élus l'importance du processus de consultation pour assurer le bon fonctionnement d'une société démocratique. (K06ME0425Car0287)

Le cas de ME_6 démontre que les agglomérats identifiés dans la première partie de ce chapitre peuvent être plus grands et inclure de façon pertinente des clusters qui sont légèrement moins similaires du point de vue lexico-sémantique. Ceci n'invalide pas la manière de trouver les agglomérats les plus importants mais constitue un indice pour mieux approfondir cette analyse. Après une étape d'exploration, un

approfondissement de ce cas pourrait être pertinent. Toutefois, ceci n'entre pas dans le cadre de cette thèse puisque notre méthode vise à assister la phase d'exploration d'un corpus.

5.2.1.2 Le jugement des contribuables

Dans le cas de ME_7 (figure 5.15), une macrostructure spécifique a été identifiée. En effet, ce cluster met en évidence un élément qui apparaît rarement de manière explicite dans les 117 clusters que nous avons annotés, soit *les impacts que les actions du gouvernement auraient sur les contribuables*. Dans ce contexte, plusieurs acteurs sont évoqués, dont la majorité occupe le rôle de *destinataire* des programmes narratifs du gouvernement. L'un d'entre eux est la **classe moyenne**, à laquelle la majorité de contribuables appartient :

> La classe moyenne durement touchée. Alors que le gouvernement martèle aux étudiants de faire « leur juste part » en ce qui concerne les droits de scolarité, les étudiants affirment que ce sont les familles de la classe moyenne qui seront les grandes perdantes. « La hausse des droits de scolarité est une nouvelle taxe pour la classe moyenne », a indiqué Martine Desjardins, présidente de la Fédération étudiante universitaire du Québec. Selon le Code civil du Québec, les parents ont l'obligation de contribuer au financement des études de leurs enfants, et ce, même après que ceux-ci ont atteint leur majorité. (K07ME0322Act0138)

Un autre acteur est la **famille**, puisque les décisions budgétaires prises par le gouvernement ont des impacts sur les finances familiales :

> Des scénarios écartés. Que notre système de prêts et bourses soit le plus généreux en Amérique du Nord, comme le souligne Jean Charest pour discréditer les revendications étudiantes, ne change rien. Dès que les revenus des parents sont pris en compte pour déterminer le montant dont l'étudiant a besoin pour vivre, on écarte une foule de scénarios. Qu'en est-il si mon père est riche, mais refuse de payer pour mes études en travail social, lui qui voudrait me voir étudier en médecine? Qu'en est-il si ma mère est riche, mais refuse de me parler depuis des mois pour une raison ou une autre? Le régime de prêts et

> bourses ne veut et ne peut pas prendre en charge la réalité, se réduisant à considérer les étudiants comme une pile de chiffres. (K07ME0402Act0178)

> Le véritable courage. Les deux lettres suivantes répondent à une autre publiée dans cette page mercredi. Je dirais que, oui, certes, défier les masses est courageux. Mais moi, le courage, j'aimerais le voir chez un ministre des Finances, ou tout autre ministre, vérifiant chacune de NOS cents dépensés à gauche et à droite, sans vraiment faire attention. J'ai parfois (même souvent) l'impression que ces personnes qui sont payées pour faire un travail supposément irréprochable, ferment les yeux sur des dépenses abusives ou non nécessaires, comme le ferait un employé d'usine qui voit un de ses collègues poser un geste stupide, qui coûte de l'argent à la compagnie, mais qui ne dit rien par peur de représailles ou pour faire partie de la gang! D'un père de famille qui va payer ces fameux frais de scolarité, et qui, comme tout le monde, est exaspéré et va payer ses taxes et ses impôts. (K07ME0413Act0231)

Enfin, les travailleurs constituent un autre acteur important. Ils soulignent à plusieurs reprises leur contribution centrale aux finances du gouvernement :

> Injustice sociale. Les étudiants demandent au gouvernement du Québec un gel des frais de scolarité. Et nous les travailleurs... eh bien, on aimerait ne pas payer plus de taxes. Le gouvernement s'entête, il ne veut pas nous ajouter, à nous les travailleurs, les dépenses qu'engendrerait un gel des frais de scolarité. Si le gouvernement arrêtait de subventionner les écoles privées et qu'il prenait cet argent pour les frais de scolarité universitaires, on aurait l'impression, nous les travailleurs, que nos taxes et nos impôts profitent vraiment aux Québécois. Qui va dans les écoles privées? Sûrement les enfants de familles qui en ont les moyens. (K07ME0416Act0237)

Il est important de mentionner que *Métro* n'est pas le seul journal à mettre de l'avant cette macrostructure. Par exemple, le journal *La Presse* contient des articles qui l'évoquent dans au moins deux clusters PR_33 et PR_44 :

> Des parents inquiets. Pendant que les étudiants manifestent, leurs parents s'inquiètent, car ils savent que ce sont eux qui devront absorber une large part de la hausse des droits de scolarité. (K33PR0316Act0205)

> La justice sociale. Pour la hausse. J'ai trois enfants. S'ils vont à l'université, je paierai leurs droits de scolarité. J'en ai les moyens, c'est vrai. Mais pour tout vous dire, je ne vois pas vraiment ce que j'ai de tellement plus important à faire sur cette Terre avec mon argent que de donner l'occasion à mes enfants d'étudier aussi longtemps qu'ils le veulent. Quoi? Vous me dites qu'au nom de la justice sociale, on devrait geler les droits? Ou mieux encore, instaurer la gratuité universitaire? Ce sera « le gouvernement » qui paiera pour mes enfants? Cool, le gouvernement! Je vais économiser un paquet d'argent avec ça. Mais n'appelez pas ça « justice sociale ». (K42PR0322Act0257)

Les autres journaux ont des articles similaires à ceux de ME_7 mais, par contre, sont éparpillés dans plusieurs clusters, comme PR_42, PR_33, JM_3, SO_23, DE_39 et JQ 2. Enfin, les autres journaux n'ont pas de cluster possédant les mêmes caractéristiques que ME_7 en termes de taille, de langage et de macrostructure. Par conséquent, si un cluster regroupant les articles qui contiennent cette macrostructure s'est formé pour *Métro*, pour les autres journaux en revanche, la macrostructure est évoquée dans des articles qui n'appartiennent pas à un seul cluster. Enfin, pour cette macrostructure, le journal *Métro* utilise un langage plus homogène et distinctif, ce qui permet probablement à notre méthode d'identifier un cluster complètement dédié à cette macrostructure.

5.2.2 Journal de Montréal

Les deux journaux de Québecor, *Le Journal de Montréal* et *Le Journal de Québec*, sont très similaires. En effet, nous avons remarqué à plusieurs reprises que les clusters des deux journaux se ressemblent. Ceci est surtout dû au fait que les deux journaux publient souvent les mêmes articles. Un exemple est en constitué par les clusters JQ_7 et JM_18 dont la majorité des articles sont les mêmes. Trois exemples sont présentés dans le tableau 5.11, où on peut observer que, par leurs titres, ces articles sont presque identiques.

Tableau 5.11 Exemples de titres similaires pour les articles pareils entre le *Journal de Québec* et le *Journal de Montréal*

Journal	Code article	Titre article
Journal de Québec	K7JQ0403Nou0280	Situation bientôt critique dans les cégeps
Journal de Montréal	K18JM0403Nou0349	Situation critique dans les cégeps
Journal de Québec	K7JQ0809Nou1225	Échec pour les grévistes
Journal de Montréal	K18JM0809Nou1461	Échec pour les grévistes
Journal de Québec	K7JQ0627Nou1093	Aucune entente à l'horizon
Journal de Montréal	K18JM0627Nou1305	Pas d'entente à l'horizon

Les habitudes de publication du groupe Québecor constituent ainsi un obstacle à l'identification des caractéristiques distinctives de ses journaux.

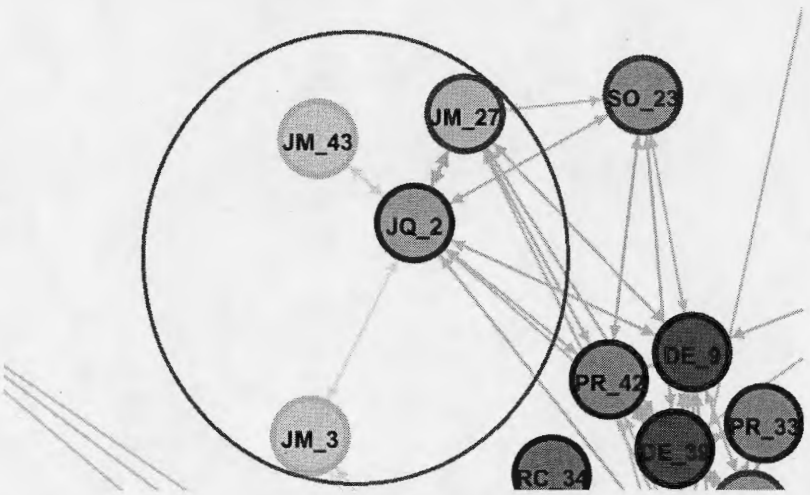


Figure 5.17 Détail de la figure 5.14

5.2.2.1 Une question d'argent

Tous les clusters du *Journal de Montréal* ont au moins une connexion lorsque le seuil est fixé à 0,6. Le cluster avec le moins de connexion est JM_43 (figure 5.17), qui en possède une à 0,67 avec le *Journal de Québec* (JQ_2). Comme on peut le constater dans la figure 5.17, les deux journaux de Québecor sont connectés par plusieurs liens.

Le langage spécifique de JM_43 met en valeur de termes qui apparaissent rarement dans les autres clusters et, tout particulièrement, le mot « milliard », suivi par les mots « budget », « million », « dépense », et autres (figure 5.18).

Figure 5.18 Nouage de mots les plus fréquents du cluster JM_43

Ce cluster présente surtout des articles qui traitent des questions financières (figure 5.18), où le gouvernement prend le rôle de *sujet*. Certains articles soulignent le *manque de ressources financières des universités*, d'autres *les propositions d'amélioration de la gestion des finances publiques.* Dans ce dernier cas, les propositions sont présentées comme des solutions pour éviter la hausse des droits de scolarité. En ce sens, l'acteur **amélioration de la gestion des finances publiques** est un *adjuvant* des programmes narratifs qui ont comme *objet* **l'annulation de la hausse**.

> Allons! Vous allez me faire croire qu'il n'y a pas 3 % de dépenses douteuses dans ce ministère? Ou si vous voulez, 3 % de gains d'efficacité possibles? Qu'on pense seulement aux commissions scolaires, qui gèrent près de 7 milliards $ par année. Et dont les parties de golf et les frasques dépensières se poursuivent malgré les remontrances des politiciens. À lui seul, le service de la dette de ces commissions scolaires coûte 650 millions $ par année! (K43JM0316Vot0195)

> Je ne vise pas l'éducation en particulier. Au contraire! Si on gérait avec un peu plus de rigueur, on pourrait probablement faire atterrir plus d'argent dans nos écoles, là où sont les besoins. Mais regardez là, la pieuvre. Vous avez le choix. Le réseau de la santé, avec ses cadres qui poussent comme des champignons, coûte presque 30 milliards $! (K43JM0316Vot0195)

> En fait, si on améliorait l'efficacité de l'État de 0,8 % seulement, on aurait le 500 millions $ pour les universités. (K43JM0316Vot0195)

> Heureusement, certains sont plus allumés. Des leaders étudiants d'associations universitaire (FEUQ) et collégiale (FECQ) proposent enfin de couper dans le gras! Des compressions de 300 millions $ sur cinq ans dans la gestion des universités, pour réinvestir l'argent dans la recherche et l'enseignement. Notamment en gelant des fonds liés à l'informatique, aux communications ou aux investissements immobiliers. En sabrant les budgets de gestion, et en diminuant les salaires des recteurs universitaires. (Personnellement, je me serais aussi attaqué à leurs pensions parfois scandaleuses, mais bon.). Le but : limiter la hausse des droits de scolarité, tout en évitant d'accroître le fardeau fiscal des Québécois. (K43JM0413Vot0444)

Dans ce contexte, différents acteurs ayant le rôle d'*antisujet* apparaissent et ce, en relation avec les actions du *sujet* **gouvernement**. En général, la variété de ces acteurs dépend du fait que le discours sur la **hausse** s'inscrit dans un débat plus large sur l'**amélioration de la gestion des finances publiques** et qui prend en compte différentes typologies de hausse, comme la **hausse du prix de l'essence**. Dès lors, les *antisujets* ne sont pas seulement les étudiants mais aussi les **contribuables**, les **familles**, etc.

> Pour permettre au gouvernement d'atteindre son objectif de retour à l'équilibre budgétaire comme prévu, en 2013-2014, les contribuables québécois devront continuer à piger dans leurs poches. Un couple avec deux enfants ayant un revenu familial de 75 000 $ devra débourser 900 $ de plus, cette année, et 1 100 $ l'an prochain. Cet effort comprend la contribution santé du gouvernement Charest, qui atteint 200 $, cette année, de même que la hausse de la taxe sur l'essence et l'augmentation de la TVQ, qui est passée à 9,5 % le 1er janvier. Ce calcul fait par le Conseil du trésor exclut bien sûr la hausse des droits de scolarité. (K43JM0320Nou0227)

> La présidente de la Ligue des contribuables du Québec, Claire Joly, affirme catégoriquement que le ministre des Finances, Raymond Bachand, a raté ses propres cibles en matière de réduction des dépenses. Le gouvernement a dépensé 3,6 milliards de plus que prévu. Le ministre s'était engagé solennellement, on s'en souviendra, à ce que le gouvernement supporte 62 % de l'effort pour revenir au déficit zéro. La Ligue des contribuables a fait ressortir à la veille du dépôt du budget 2012-2013 qu'une famille dont le revenu est de 60 000 $ versera à l'État cette année 1 000 $ de plus qu'en 2009, seulement à travers la hausse de la TVQ, celle de la taxe sur l'essence, ainsi que l'impôt santé. (K43JM0320Vot0224)

> Les contribuables n'en peuvent plus. Ils veulent que le gouvernement contrôle mieux ses dépenses, réduise la bureaucratie, améliore la gestion dans l'appareil gouvernemental. La croissance des dépenses a été de 5 % par année en moyenne depuis l'arrivée au pouvoir de ce gouvernement et la dette a plus que doublé. (K43JM0320Vot0224)

> C'est la faute à Bachand. Le gouvernement a un gros problème. Et pas seulement avec les étudiants. Personne ne veut payer davantage. C'est vrai que, pressés comme des citrons, les contribuables québécois donnent au fisc chaque année beaucoup plus que tous leurs voisins, qu'ils soient en Ontario, au Nouveau-Brunswick ou au Vermont... Selon le plus récent rapport annuel de l'Agence du revenu du Québec, là où les plus « saines habitudes de vie » sont gracieusement offertes aux fonctionnaires, les contribuables ont été soulagés de 80 milliards, l'an dernier, le quart de cette somme provenant uniquement de l'impôt sur le revenu des particuliers. Collecter plus de 20 000 millions dans une société où 40 % des gens ne gagnent pas 20 000 $ par année, voilà une performance remarquable. Ce qui fait presque léviter le ministre des Finances, Raymond Bachand. « L'Agence de revenu du Québec est l'une des plus agressives », a-t-il fièrement signalé à l'Assemblée nationale, il y a quelques jours. Seuls les fous l'applaudiront... Taxage. Impôts, taxes de vente ou sur l'essence, taxes municipales et scolaires, permis, droits, franchise, frais, cotisations, prélèvements, primes, surprimes... Même le chien, comme le chat, est taxé. Mais tout ça ne suffit pas. Les demandes affluent, les lobbys affamés se multiplient et les défenseurs de nos droits « sociaux » en veulent toujours plus... Eh bien, la dette augmente... Loin des regards (K43JM0602Nou1061)

On remarque également que cette macrostructure s'insère dans le programme narratif du **gouvernement** qui a comme *objet* le **déficit zéro** puisque l'acteur **amélioration**

de la gestion des finances publiques est un *adjuvant*. On remarque aussi que les termes « milliard » et « million » deviennent plus fréquents :

> Sans aucune marge de manœuvre, le gouvernement Charest a livré un budget prudent et sans artifice, hier, saupoudrant quelques millions ici et là, afin de revenir au déficit zéro l'an prochain. Le Québec est toujours dans le rouge. Acculé au pied du mur, le ministre des Finances a déposé un budget de consolidation. (K43JM0321Nou0237)

> Après avoir réalisé un déficit de 3,3 milliards de dollars (500 millions de moins que prévu), on anticipe un déficit de 1,5 milliard en 2012-2013. Les dépenses continuent d'augmenter bien que leur croissance ait été réduite à 2 %, un taux « historiquement bas ». Elles atteindront 70,9 milliards alors que l'État engrangera des revenus de 69,4 milliards en 2012-2013. La dette brute a gonflé de 10 milliards, atteignant aussi un niveau historique (183,7 milliards), soit 55 % du produit intérieur brut (PIB). Elle atteindra 191,7 milliards en 2013. Électoralisme. Dans un tel contexte, Raymond Bachand s'est défendu de présenter un budget électoraliste, soulignant qu'il ne contenait « que 200 millions de nouvelles mesures ». Vérification faite, l'impact total des mesures annoncées atteint seulement 211,2 millions pour la prochaine année. En fait, ce budget aura bien peu d'effet sur le contribuable moyen à court terme. Sans surprise, M. Bachand n'a annoncé aucune « nouvelle » hausse de taxes ni d'impôts (K43JM0321Nou0237)

Le lien entre ce cluster et JQ_2 illustre un phénomène particulier. Ce cluster est similaire du point de vue sémantique et lexical à JQ_2 puisque leur valeur cosinus est de 0,67. Toutefois, JQ_2 souligne d'autres éléments et les macrostructures ne coïncident pas. Par exemple, le cluster du *Journal de Québec* présente *la valeur économique du diplôme* ou le concept d'*investissement scolaire pour justifier la hausse*, ce qui implique des macrostructures différentes de celles apparaissant dans JM_43. Cependant, il existe également des clusters qui sont proches de JM_43 du point de vue sémantique et qui structurent aussi leur contenu d'une manière similaire, comme JQ_49, SO_6 et PR_17. Au contraire du cluster du *Journal de Montréal*, ces clusters contiennent peu d'articles et n'ont pas été retenus pour la phase d'annotation. Ceci indique que le *Journal de Montréal* dédie probablement plus d'espace que les

autres journaux à la macrostructure présentée dans ce paragraphe, ce qui constitue certainement une des pistes pour étudier les spécificités de ce quotidien.

5.2.3 Journal de Québec

En général, les considérations énoncées pour le *Journal de Montréal* valent aussi pour le *Journal de Québec*. Toutefois, le *Journal de Québec* présente des clusters isolés dans l'espace des 117 clusters (figure 5.19). Ces clusters sont au nombre de trois et ils seront analysés dans les prochains paragraphes.

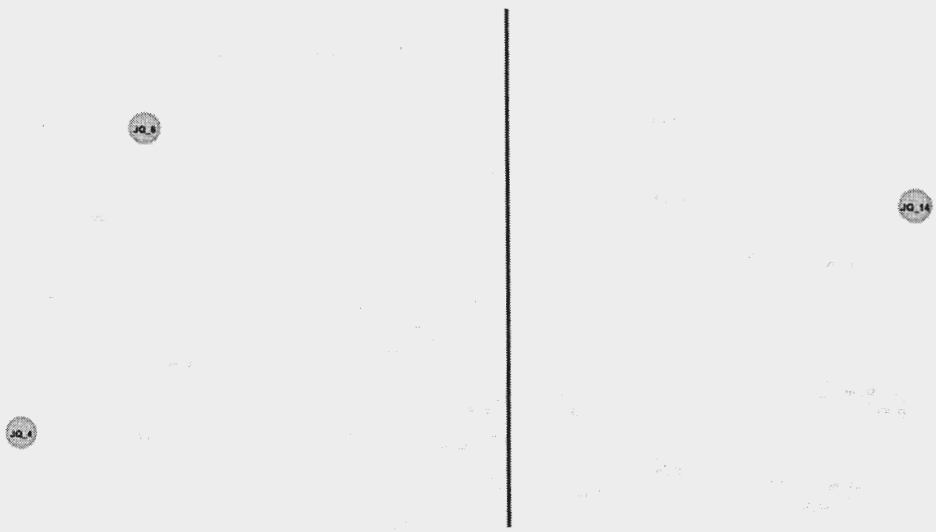


Figure 5.19 Détail de la figure 5.14. Les clusters JQ_4 et JQ_6 sont représentés à gauche, le cluster JQ_14 à droite. La figure montre leur isolement

5.2.3.1 Quelle légitimité pour les syndicats?

Le premier cluster isolé qui sera analysé est JQ_4. Il est caractérisé par l'apparition fréquente d'un acteur, le **syndicat**. En effet, le schéma lexical spécifique de ce cluster se distingue par la présence du mot « syndicat » et d'autres termes faisant partie du même champ sémantique comme « syndical », « centrale » (Centrale des syndicats du Québec), « CSQ », « CSN », « FTQ », etc. On observe les mots les plus fréquents à l'aide de la technique de nuage de mots (figure 5.20.).

L'annotation du cluster a identifié l'existence d'une macrostructure commune entre les articles explicitant davantage le schéma lexical qui est à peine aperçu par la distribution des mots les plus fréquents du cluster (figure 5.20). Le lecteur pourra également trouver un certain nombre de stéréotypes qui reprennent les présupposés culturels et politiques qui, historiquement, mettent en opposition la gauche et la droite. Cette dernière partie politique, par exemple, considère les étudiants et les syndicats de travailleurs comme des « profiteurs » du système économique. Ceci est plus évident lorsqu'une crise sociale, comme celle du printemps érable, se déclenche. La gauche, au contraire, défend le rôle des syndicats puisqu'ils sont une partie intégrante de la vision « plus globale de la société » et non strictement économique.

Figure 5.20 Nouage de mots plus fréquents du cluster JQ_4

Cette macrostructure est caractérisée essentiellement par les acteurs : la **centrale syndicale CSQ** et son président **Rejean Parent**, le syndicat **CSN** et son président **Louis Roy** et le syndicat **FTQ** et son président **Michel Arsenault**. Ces acteurs assument généralement le rôle d'*adjuvants* dans la phase de dialogue et de négociation entre les **étudiants** et le **gouvernement**.

> Réjean Parent de la CSQ, Louis Roy de la CSN et Michel Arsenault de la FTQ étaient tous trois présents, lors des négociations entre le gouvernement et les associations étudiantes. (K04JQ0507Nou0601)

Les articles présentent fréquemment une critique envers les syndicats et mettent en doute la légitimité de leur intervention dans le conflit étudiant. En d'autres termes, certains articles contestent la légitimité d'une telle intervention dans les **négociations** et donc leur rôle d'*adjuvant.* Cette configuration de la macrostructure est la plus fréquente.

> Pouvez-vous m'expliquer pourquoi les représentants de ces trois centrales syndicales ont été invités à siéger sur une table de négociation destinée à dénouer l'impasse de la crise étudiante? Pourquoi pas le Conseil du patronat? L'Union des producteurs agricoles? Les filles d'Isabelle? Les Schriners? Qu'ont en commun les syndicats de travailleurs et les associations étudiantes, sinon la même affection pour la couleur rouge? Les étudiants NE SONT PAS des travailleurs, leur boycott N'EST PAS une grève et leurs piquets NE SONT PAS légaux. Combien de temps faudra-t-il le répéter, bon Dieu de bon sang? (K50JM0507Nou0730)

Les raisons d'une telle opinion sont souvent basées sur le raisonnement que les étudiants ne sont pas des travailleurs.

> La CLASSE buissonnière. Les médias devraient cesser d'utiliser ce mot « grève » dans le cas du conflit avec les étudiants. De plus, la CLASSE ne peut contenir le terme « syndicat » dans son nom, car il ne s'agit que d'une association étudiante, donc non reconnue comme « syndicat ». Elle ne peut donc pas faire de grève mais un « boycott ». (K04JQ0420Vot0409)

> « DES INTRUS? Qu'est-ce que les présidents de la FTQ, de la CSN et de la CSQ (Michel Arsenault, Louis Roy et Réjean Parent) faisaient autour de la table de négo vendredi et samedi? Quand on sait que 78 % des dépenses dans nos universités vont à payer les salaires de leurs membres, voulaient-ils s'assurer de ne pas faire leur part dans l'entente entre étudiants et gouvernement? Au lieu de défendre l'intérêt de ceux qui étudient, les leaders étudiants semblent avoir laissé le renard syndical s'occuper de leur poulailler.

> Le plus long conflit étudiant de l'histoire du Québec n'aura, dans ce sens, pas été totalement inutile. Il aura permis aux tribunaux de démontrer que chaque étudiant à le droit d'accéder à ses cours, peu importe ce que souhaitent les syndicalistes - étudiants. » (K04JQ0507Vot0598)

> Transparence souhaitée. [...] Au moment où les appuis en argent sonnant des grandes centrales syndicales aux mouvements étudiants soulèvent la controverse, un député conservateur envisage de forcer les organisations à divulguer l'utilisation qu'ils font des cotisations de leurs membres. (K04JQ0517Vot0729)

> La crise expliquée aux enfants. [...] « Peux-tu m'expliquer? » Bien sûr, fiston. LE CLUB DE MICKEY. Commençons par le début. Il y a quelques mois, la ministre du Travail a confronté les syndicats et tenté de les mettre à leur place... « C'est quoi, un syndicat? ». C'est une association. Avant, cette association défendait les droits des travailleurs qui étaient exploités par des patrons sans scrupule. Mais avec le temps, les syndicats ont grossi et sont devenus des corporations qui n'ont qu'une obsession : étendre leur influence et augmenter leur part de marché en vendant le plus de cartes possible. Tu comprends? « Comme le club de Mickey? ». Exactement. Plus il y a de gens qui joignent le club, plus le club est fort et a du pouvoir. (K04JQ0521Nou0785)

> « Les banquiers. Il y a quelques semaines, j'écrivais que les syndicats se servaient de la « cause étudiante » pour régler leurs comptes avec le gouvernement. Ils n'ont pas digéré le fait que le gouvernement veuille abolir le placement syndical dans le milieu de la construction, disais-je, et ils ont décidé d'envoyer les jeunes au front pour mener leur bataille et déstabiliser leur adversaire. (K04JQ0607Nou0954)

> Le conflit étudiant au coeur des priorités. CHARGEMENT DE GARDE À LA CSQ. La Centrale des syndicats du Québec (CSQ) fera tout pour se faire entendre aux prochaines élections et ne nie pas qu'elle pourrait appuyer les étudiants dans leurs actions pour faire tomber le gouvernement. Questionnée au sujet de l'appui que la CSQ offrira au mouvement étudiant dans ses démarches pour déloger le gouvernement, la présidente fraîchement élue, Louise Chabot, a confirmé ses intentions de soutenir les étudiants. (K4JQ0630Nou1103)

La plupart des articles qui ont été annotés possèdent un correspondant dans le *Journal de Montréal*. Plusieurs articles ont été publiés sous des titres différents, comme pour

K04JQ0630Nou1103 (« Le conflit étudiant au cœur des priorités ») et K06JM0630Nou1315 (« La CSQ compte bien appuyer les étudiants »), mais ils partagent souvent plus du 80% du contenu. La seule différence significative entre les deux journaux du groupe Québecor est qu'aucun des clusters regroupant ces articles s'est formé pour *Le Journal de Montréal*. Ces articles sont éparpillés dans différents clusters, comme dans JM_6, JM_50, JM_3, et restent loin du centroïde qui définit le cluster. L'absence de ce cluster constitue une piste de recherche pour identifier les différences entre les deux journaux du groupe Québecor.

5.2.3.2 Le combat de Proulx et Morasse

À partir des 117 clusters retenus dans la phase d'annotation, un deuxième cluster du *Journal de Québec* se démarque, soit JQ_14 (figure 5.21). Ce cluster présente l'histoire de deux étudiants de l'Université Laval, **Jean-François Morasse** et **Laurent Proulx**, qui ont mené un combat **juridique** pour obtenir le **retour en classe** dans leurs cours respectifs. Dans le contexte du *boycottage de cours* pratiqué par les étudiants, ces deux acteurs se sont battus pour forcer les tribunaux à obliger les professeurs et les étudiants à retourner en classe.

Figure 5.21 Nouage de mots plus fréquents du cluster JQ_14

Laurent Proulx gagne une manche. INJONCTION ACCORDÉE. Après avoir remporté une bataille en Cour supérieure, l'étudiant Laurent Proulx a eu droit à un débat enflammé de deux heures trente, où la majorité des participants n'étaient pas de son avis, hier, lors de son cours d'anthropologie à l'Université Laval. Suite à sa victoire personnelle contre sept avocats, l'étudiant de 24 ans s'est présenté une quinzaine de minutes avant l'heure prévue. (K14JQ0404Nou0291)

Inspiré par Laurent Proulx. GRÈVE ÉTUDIANTE -- INJONCTION. Un autre étudiant de l'Université Laval, exaspéré par la grève, demandera une injonction pour obtenir le libre accès à ses cours, ce matin, au palais de justice. Jean-François Morasse, 25 ans, plaidera sa cause seul, en même temps que celle de Laurent Proulx, qui souhaite continuer d'assister à son cours sans obstacle. Le jeune homme veut terminer un certificat en arts plastiques pour débuter un baccalauréat de design graphique en septembre prochain. Sa requête est identique à celle de son confrère. (K14JQ0412Nou0352)

Accès aux cours pour Proulx et Morasse. INJONCTIONS. Jean-François Morasse pourra aller à ses cours d'arts plastiques à l'Université Laval, après avoir obtenu une injonction du tribunal, tout comme Laurent Proulx qui a vu la sienne reconduite. C'est dans une salle bondée d'étudiants, au palais de justice de Québec, que Jean-François Morasse, 25 ans, s'est présenté seul devant le tribunal afin de faire valoir ses droits. (K14JQ0413Nou0364)

> Laurent Proulx règle à l'amiable. INJONCTION CONTRE L'UL. Laurent Proulx, cet étudiant de l'Université Laval qui avait réussi à obtenir une injonction provisoire à deux reprises pour lui permettre d'accéder à son cours d'anthropologie, a réglé à l'amiable hier son litige. Alors qu'il devait revenir devant les tribunaux lundi, Laurent Proulx s'est finalement désisté de la demande d'injonction contre l'Université Laval, la Confédération des associations d'étudiants et d'étudiantes (CADEUL) et les associations des étudiants en anthropologie et en sciences sociales. (K14JQ0421Nou0425)

> Injonction reconduite. UNIVERSITÉ LAVAL. Après un débat d'une journée, l'injonction provisoire permettant à l'étudiant en arts plastiques Jean-François Morasse d'aller à ses cours à l'Université Laval, a été reconduite jusqu'au 4 mai. Le juge Jean-François Émond, de la Cour supérieure, a reconduit la demande d'injonction provisoire et se prononcera vendredi prochain sur une ordonnance de sauvegarde afin d'éviter aux différentes parties d'avoir à revenir devant le tribunal tous les dix jours. Jean-François Morasse a l'intention de se faire entendre sur le fond de cette affaire afin que les étudiants puissent avoir libre accès à leurs cours si un autre conflit devait à nouveau perturber le déroulement d'une session. « Ça soulève beaucoup de questions quant aux droits des individus et des regroupements et je trouve que c'est important d'aller au fond des choses », a-t-il mentionné. Nouvelle association. Quant à Laurent Proulx, cet autre étudiant de l'Université Laval qui a obtenu une injonction de la Cour supérieure cet hiver, il a l'intention de lancer une nouvelle association étudiante. « Le mandat qu'on se donnerait est le libre accès aux cours et que les étudiants puissent recevoir les services pour lesquels ils ont défrayé les coûts. » (K14JQ0427Nou0479)

L'histoire racontée par ce cluster se termine avec la phase d'évaluation des actions de ces étudiants. **Laurent Proulx** abandonne son cours alors que **François Morasse** réussit à terminer sa scolarité. Ces deux étudiants sont des *symboles*. Ils représentent les étudiants qui sont lésés par les événements.

> Laurent Proulx décroche. AGENCE QMI -- Après avoir obtenu une injonction et avoir réclamé haut et fort la liberté de pouvoir étudier malgré la grève étudiante, Laurent Proulx a finalement abandonné le cours d'anthropologie qu'il suivait à l'Université Laval. [...] Jean-François Morasse, qui a lui aussi eu recours aux tribunaux, pour empêcher les grévistes de bloquer l'accès à ses cours, a pour sa part terminé sa session. « J'ai reçu mes notes. Je passe dans tous mes cours », a-t-il confié. Malgré la tourmente des derniers mois, l'étudiant a

> réussi à terminer les six derniers cours qui lui manquait pour obtenir un certificat en arts plastiques. (K14JQ0609Nou0971)

Ce cluster n'est pas le seul à avoir traité des événements qui impliquent ces étudiants. En effet, l'autre journal situé dans la ville de Québec, *Le Soleil*, présente un cluster similaire. Ces deux journaux ont discuté très souvent de cette question, avec SO_26 qui contient 21 articles et JQ_14 qui en contient 18. Le *Soleil* toutefois présente deux clusters séparés, un pour l'histoire de Morasse (SO_27) et l'autre pour Proulx (SO_26). Les clusters du *Soleil* ne font pas parti du tiers retenu dans la phase d'annotation.

5.2.3.3 La saison touristique de Montréal en péril

Le dernier cluster qui caractérise le *Journal de Québec* est JQ_6 (figure 5.22). Ce cluster compte 25 articles traitant des actions accomplies par les étudiants en lien avec les grands événements de l'été montréalais, soit le **grand prix de F1** et les **festivals**. En effet, les associations étudiantes et d'autres organisations de manifestants ont utilisé ces événements comme *vitrine privilégiée* pour protester **contre la hausse des frais de scolarité**. La fréquence des mots (figure 5.22) montre que le champ lexical le plus fréquent est lié aux spectacles et aux événements qui caractérisent l'été montréalais, comme le *Festival Juste pour Rire*, le *Grand Prix de Formule 1* et d'autres.

Figure 5.22 Nouage de mots plus fréquents du cluster JQ_6

Un élément qui est mis en valeur par ce cluster est le concept de **perturbation**, qui est utilisé par les étudiants comme *adjuvant*. En effet, la stratégie utilisée pour profiter de la vitrine médiatique se base sur la *perturbation des spectacles*. Dans ce contexte, l'ordre socio-économique est perçu comme une valeur supérieure à la justice sociale.

> Des étudiants songent à perturber. COURSE AUTOMOBILE -- GRAND PRIX DE MONTRÉAL. MONTRÉAL - MONTRÉAL -- (Agence QMI) Des étudiants envisageraient de perturber les activités du Grand Prix de formule 1 de Montréal. (K06JQ0511Spo0654)

Un autre élément très présent dans ce cluster est le programme narratif des **présidents des festivals** ou du **ministre du Tourisme**. Ces acteurs interviennent fréquemment dans l'espace public pour limiter les dommages potentiels de perturbation des festivals sur la **saison touristique** de Montréal :

> Nicole Ménard préoccupée. CONFLIT ÉTUDIANT -- IMPACT SUR LE TOURISME et Jean-Luc Lavallée. La ministre du Tourisme, Nicole Ménard, s'inquiète du possible impact négatif des manifestations étudiantes sur la

> prochaine saison touristique à Montréal. « C'est très préoccupant, je l'avoue », a indiqué, hier, la ministre, questionnée sur les effets possibles de la diffusion d'images de confrontations entre policiers et manifestants étudiants partout dans le monde. Craignant que les touristes hésitent à se rendre dans la métropole en raison de cette image de désordre, Mme Ménard soutient qu'elle est en contact constant avec Tourisme Montréal. (K06JQ0517Nou0722)

> Plein son casque. Ce matin, j'ai une pensée pour tous les commerçants de Montréal. Plusieurs hôtels du centre-ville, habituellement en demande quand se pointe le cirque de la formule 1, ont encore beaucoup de chambres libres. Ça peut être dû à une foule de facteurs. Pas seulement aux scènes d'affrontements entre policiers et manifestants qu'on voit dans les médias, ici et ailleurs. Mais en ce début d'été, les organisateurs de festivals commencent à s'inquiéter. (K06JQ0530Vot0873)

Dans ce cluster, les événements qui touchent le **grand prix de F1** et le **festival Juste pour Rire** sont les plus nombreux. **Gilbert Rozon**, le président du festival Juste pour Rire, devient un des acteurs principaux de cette macrostructure. Afin de minimiser les perturbations sur le festival, M. Rozon s'implique en personne pour négocier avec les étudiants :

> Dumontier se fait rassurant. GRAND PRIX DU CANADA. MONTRÉAL - MONTRÉAL -- (Agence QMI) L'organisation du Grand Prix de Formule 1 de Montréal se veut rassurante malgré les rumeurs de possibles perturbations liées au conflit étudiant. « Le Grand Prix du Canada dispose d'un plan complet visant à assurer en tout temps le confort et la sécurité des participants et des spectateurs. Quelles que soient les circonstances, notre objectif demeure toujours le même : offrir à chacun de ces groupes un événement plaisant et de grande qualité », a indiqué hier par communiqué François Dumontier. (K06JQ0602Spo0899)

> Rozon parlera aux étudiants. Le Festival Juste pour Rire pâtit des manifestations. Le grand manitou du Festival Juste pour Rire Gilbert Rozon rencontrera les leaders des quatre associations étudiantes en début de semaine afin de les inciter à ne pas perturber les nombreux festivals cet été à Montréal. (K06JQ0603Nou0910)

L'*antisujet* le plus récurrent de **Rozon** est l'association la **CLASSE** laquelle, comme pour les négociations avec le gouvernement, est exclue des rencontres entre Rozon et les autres associations étudiantes. La réaction de cette association rendra plus difficile la poursuite des discussions :

> La Coalition n'a pas reçu d'invitation de Rozon. Si l'homme d'affaires Gilbert Rozon rencontre les leaders étudiants aujourd'hui ou demain pour « faire appel à leur gros bon sens » et rétablir une « paix sociale » cet été au Québec, ce sera sans la CLASSE. Réunie en congrès à Valleyfield, hier, la Coalition large de l'Association pour une solidarité syndicale étudiante (CLASSE) dit ne pas avoir été approchée par l'entourage de M. Rozon. (K06JQ0604Nou0925)

> « Face à face cordial. Gilbert Rozon rencontre les étudiants. MONTRÉAL - MONTRÉAL -- Les leaders étudiants ont tenu à rassurer Gilbert Rozon, les dirigeants de grands événements et la population lors d'une rencontre « cordiale » qui a eu lieu hier après-midi. « Ç'a été une rencontre très positive et respectueuse », a déclaré le grand patron de Juste pour rire, qui est resté discret quant aux détails de cette réunion qui s'est déroulée en présence de Paul-Émile Auger (TACEQ), Martine Desjardins (FEUQ) et Éliane Laberge (FECQ). « L'idée, c'était de mettre cartes sur table. Nous avons voulu exprimer notre mécontentement par rapport aux différentes sorties publiques qui ont été faites, mais nous voulions aussi examiner ses demandes, a déclaré Martine Desjardins. Nous l'avions déjà fait dans les médias, mais de le faire en étant face à face, c'était beaucoup plus intéressant. Ça aurait dû être fait avant. ». (K06JQ0605Spe0928)

> « Pas de spectacle-bénéfice pour les étudiants. JUSTE POUR RIRE. Le Festival Juste pour rire ne présentera aucun spectacle-bénéfice au profit des associations étudiantes. Au lendemain de son rendez-vous avec les leaders étudiants, Gilbert Rozon tient à rectifier certains faits publiés dans les pages du quotidien montréalais La Presse. « Mes sœurs ont décidé, avec d'autres producteurs, qu'elles allaient présenter un spectacle. Par contre, ce spectacle ne sera pas offert dans le cadre de Juste pour rire, a déclaré Gilbert Rozon lorsque joint par le Journal, hier avant-midi. Juste pour rire n'organise pas de spectacle-bénéfice pour l'une ou l'autre des parties. ». Luce et Lucie Rozon, qui évoluent toutes les deux au sein de la division Juste pour rire scène, se sont associées à Daniel Thibault, qui a notamment coécrit les deux saisons de la série Mirador, afin de mettre sur pied cet événement humoristique présenté par la Coalition des humoristes indignés. » (K06JQ0606Spe0942)

Dans les autres journaux, il existe des clusters qui ressemblent à JQ_6 (figure 5.22), par exemple JM_30, JM_41 (*Le Journal de Montréal*), PR_20, PR_22 (*La Presse*), So_29 (*Le Soleil*) et ME_4 (*Métro*). Toutefois, la majorité d'entre eux n'ont pas été retenus dans la phase d'annotation à cause de leur taille.

5.2.4 La Presse

Dans l'espace des 117 clusters annotés, le journal *La Presse* ne présente aucun cluster isolé. Les clusters « les plus isolés » ont une seule connexion. Parmi ceux-ci (figure 5.23), le cluster PR_41 se démarque des autres puisqu'il fait partie d'une paire isolée de clusters, laquelle, de plus, est composée de journaux du même groupe. En effet, le cluster qui l'accompagne SO_51, appartient au journal *Le Soleil*. Les autres clusters, PR_21, PR_23 et PR_29, sont tous connectés à de plus grands groupes de clusters.

Comme dans le cas du groupe Québecor, les deux journaux du groupe Power Corporation, *La Presse* et *Le Soleil*, sont souvent connectés entre eux, car ces clusters sont souvent très similaires. Pour des raisons de simplicité et de lisibilité, nous nous limitons à analyser la paire de clusters composée de PR_41 et SO_51.

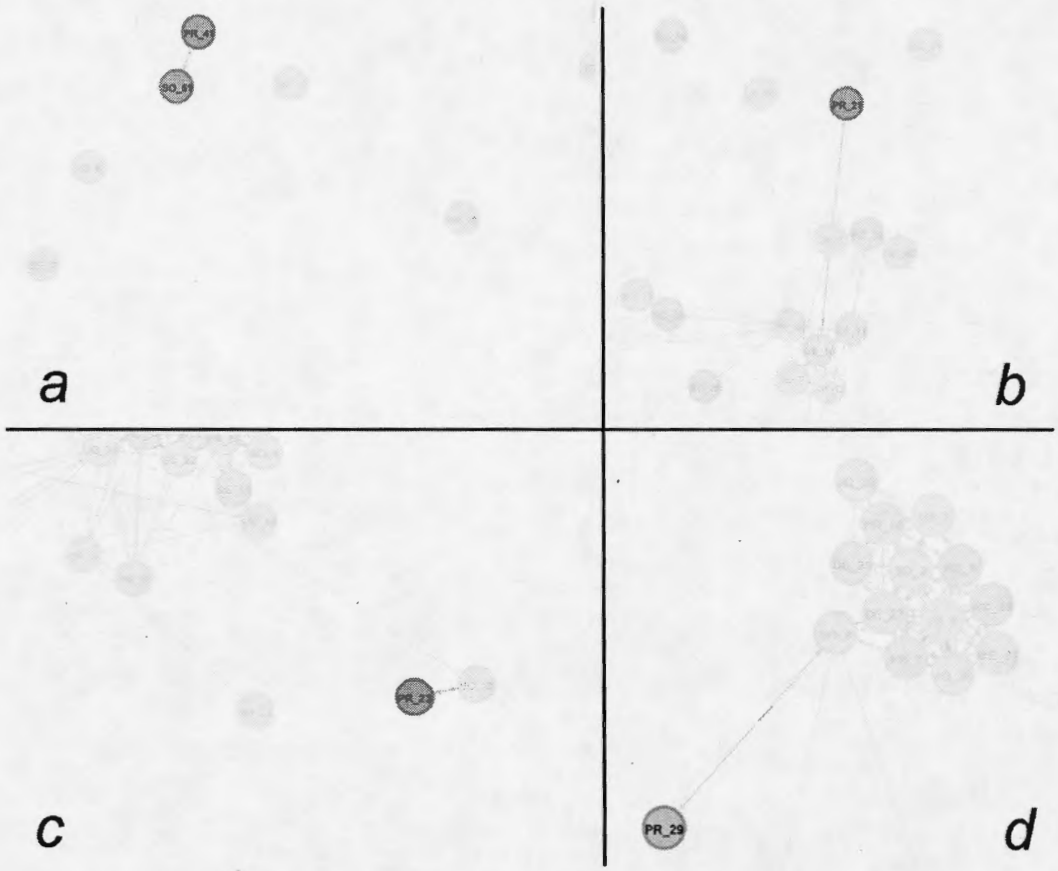


Figure 5.23 Détails de la figure 5.14. Les clusters PR_41 (a), PR_21 (b), PR_23 (c) et PR_29 (d) sont représentés

5.2.4.1 Khadir, sa fille et les arrestations

Le groupe de clusters composé par PR_41 et SO_51 (figure 5.23) fait état des événements dans lesquels le député du parti Québec Solidaire, **Amir Khadir,** et sa fille **Yalda Machouf-Khadir** ont été impliqués. Si on tient compte seulement de l'espace des 117 clusters annotés, ces deux clusters sont relativement uniques. Un élément supplémentaire pouvant contribuer à l'interprétation de ce cluster émerge lorsqu'on élargit l'analyse des similarités. En effet, on peut observer que le langage utilisé dans ces clusters est similaire à celui des clusters traitant des manifestations étudiantes, des affrontements entre les étudiants et la police et des différentes

manifestations violentes caractérisées par des arrestations de masse, par exemple DE_27, PR_10 et SO_8.

Dans ces clusters, les acteurs principaux tiennent la première place mais dans des structures narratives différentes. Aussi, la fille du député, **Yalda Machouf-Khadir** apparaît souvent comme *sujet passif*, qui subit l'action d'être **arrêtée** ou **privée de sa liberté**. De plus, dans la majorité des articles, la « valeur de l'information » (*newsworthiness*) des évènements qui la concerne dépend surtout du fait qu'elle est la fille d'un député.

> Des conditions sévères pour les accusés. Blocage du pont Jacques-Cartier. Les manifestants arrêtés mardi pour avoir bloqué le pont Jacques-Cartier ont dû s'engager à respecter de sévères conditions avant d'être mis en liberté hier. [...] Certains accusés ont demandé des exceptions pour éviter de se trouver en violation de conditions lorsqu'ils retourneront à leurs cours à l'UQAM et au cégep du Vieux Montréal, ou lorsqu'ils iront à la Grande Bibliothèque, près du parc.[...] Quelques accusés ont obtenu des exceptions pour pouvoir fréquenter un colocataire ou un ami de coeur. C'est le cas de Yalda Machouf-Khadir, fille du député de Québec solidaire, Amir Khadir. Les jeunes adultes de 18 à 23 ans font face à des accusations de complot, d'entrave au travail des policiers et de méfait de plus de 5000 $ pour avoir empêché l'usage du pont Jacques-Cartier lors d'une manifestation matinale. (K41PR0517Act0892)

> « DES EMPLOYÉS DU GRAND PRIX VISÉS. LA CRISE ÉTUDIANTE - À L'HEURE DU GRAND PRIX. À l'aube des festivités de la Formule 1, le Service de police de la Ville de Montréal (SPVM) a procédé à une rafle chez les militants étudiants, hier. Des agents ont perquisitionné chez le député Amir Khadir, dont la fille a été arrêtée. Malgré les critiques qui s'élèvent, les autorités tenteront sous peu d'arrêter d'autres contestataires après avoir découvert que certains d'entre eux occupaient des emplois temporaires au Grand Prix de Montréal, a appris La Presse. » (K41PR0608Act1208)

> Le week-end derrière les barreaux. CRISE ÉTUDIANTE À L'HEURE DU GRAND PRIX. Les militants étudiants arrêtés jeudi, dont Yalda Machouf-Khadir, font face à une pléthore d'accusations. Quatre des personnes arrêtées lors de la rafle de jeudi dans les milieux militants passeront le week-end

> derrière les barreaux. C'est le cas de la fille du député Amir Khadir, Yalda Machouf-Khadir, qui fait face à une multitude de nouvelles accusations liées à la crise étudiante. Selon le cas, les suspects arrêtés dans l'opération policière ont comparu hier en lien avec le saccage du bureau de l'ex-ministre Line Beauchamp le 12 avril, le grabuge à l'Université de Montréal le 13 avril ou un épisode de voies de fait contre une photographe du Journal de Montréal le 22 mai. La seule personne qui a été accusée pour tous ces événements à la fois est Yalda Machouf-Khadir, 19 ans. Elle fait face à 11 accusations, dont introduction par effraction, complot, méfait de moins de 5000 $, déguisement dans le but de commettre un crime, voies de fait contre un agent de la paix, voies de fait contre une photographe, vol de moins de 5000 $ et méfait de plus de 5000 $. (K41PR0609Act1229)

> La fille du député de Québec solidaire, Amir Khadir, accusée en lien avec des manifestations survenues lors des dernières semaines à Montréal, passera la fin de semaine derrière les barreaux. Yalda Machouf-Khadir fait face à 11 chefs d'accusation reliés à divers événements. La Couronne s'est opposée à sa libération et à celle de trois des quatre autres personnes également arrêtées pour les mêmes motifs. Les audiences de libération sous caution doivent avoir lieu lundi. (K51SO0609Act1124)

De son côté, le député **Khadir** s'exprime sur les événements qui concernent sa fille. Dans ce cadre, il assume souvent le rôle d'*adjuvant* ou d'*adjuvant passif*. En effet, l'*antisujet* du programme narratif dans lequel sa fille est impliquée comme *sujet*, c'est-à-dire la **police**, est nommé à plusieurs reprises par Kadir lui-même pendant ses interventions publiques, et d'une manière qui renforce son rôle d'*adjuvant* :

> « Ma fille assumera ses responsabilités ». LA CRISE ÉTUDIANTE - À L'HEURE DU GRAND PRIX. Amir Khadir a réagi à l'arrestation de sa fille Yalda hier matin, blâmant les policiers au passage. QUÉBEC - Le député de Québec solidaire, Amir Khadir, se dit « choqué » par les « brusqueries » des policiers du Service de police de la Ville de Montréal (SPVM) qui ont arrêté sa fille Yalda au domicile familial, hier matin. (K41PR0608Act1216)

> « Ma fille assumera ses responsabilités », a déclaré, hier, le cochef de Québec solidaire, quelques heures après l'arrestation de Yalda Machouf-Khadir, 19 ans, par la police de Montréal. Elle est suspectée d'avoir participé à des méfaits à l'occasion du conflit étudiant, devenu conflit social au fil des semaines. Sans se

> prononcer sur ce qui est reproché à sa fille, dont il ignore les détails, Amir Khadir se dit convaincu qu'elle a agi pour le « bien commun ». Il rappelle qu'elle est présumée innocente. »(K51SO0608Act1102)
>
> Khadir déplore la réaction des policiers. Amir Khadir dénonce l'utilisation de gaz lacrymogènes contre les manifestants, jeudi, et s'interroge sur une « commande politique » à l'intervention policière. Il verse dans la « théorie du complot », selon la ministre Line Beauchamp. (K51SO0303Act0110)
>
> Une militante « radicale, masquée et impliquée ». Yalda Machouf Kadhir est une leader dans le mouvement radical étudiant. Vêtue de noir, le visage masqué, elle a participé aux manifestations qui ont tourné en saccages à l'Université de Montréal, au cégep du Vieux Montréal, ainsi qu'au bureau de l'ex-ministre de l'Éducation Line Beauchamp, plus tôt au printemps. Ses empreintes ont été relevées sur une porte qui a été forcée dans le bureau de Mme Beauchamp. C'est notamment ce qui ressort de l'enquête sous cautionnement de Mme Khadir, qui s'est tenue hier après-midi à Montréal. Après cinq jours de détention, la jeune femme de 19 ans est entrée dans le box avec le sourire. Son père, le député Amir Khadir, sa mère, l'épidémiologiste Nima Machouf, ainsi que plusieurs supporteurs arborant le carré rouge ont assisté à l'audience, qui se tenait devant la juge Hélène Morin. Celle-ci annoncera ce matin si la jeune Kadhir, de même qu'un autre manifestant actif, Zachary Daoust, recouvreront la liberté. Le jeune homme, sans emploi depuis juin 2011, n'avait pas les 1000 $ nécessaires pour assurer son cautionnement. (K41PR0612Act1258)

Le journal *Le Soleil* (SO_51) présente quelques différences par rapport au cluster PR_41 (figure 5.23). En effet, dans SO_51, la majorité des articles met en valeur l'acteur **Amir Khadir,** situé dans le même contexte que celui de sa fille, c'est-à-dire l'**arrestation**. En effet, le député subit également l'action de **se faire arrêter** par la **police** pour des raisons liées aux **manifestations étudiantes**. Dans certains articles, le ton devient très accusateur à son égard, ce que nous n'avons pas observé dans *La Presse* :

> Gandhi, King puis... Mandela. Khadir arrêté. Régis Labeaume s'est moqué des propos tenus par le député Amir Khadir, hier. « J'ai été bien impressionné de voir Amir Khadir se comparer à Gandhi et à Martin Luther King. Je me disais

> que s'il avait fait cinq minutes de prison hier, probablement qu'il aurait ajouté qu'il se prenait pour Nelson Mandela », a-t-il lancé depuis New York [...] L'arrestation du député de Québec solidaire (QS), lors d'une manifestation jugée illégale, mardi soir, n'émeut pas le maire de Québec. [...] « Gouvernement corrompu ». Le député de QS a commenté son arrestation hier matin. « Je regrette quand même qu'il y ait des ordres, des commandes en haut de ces policiers qui se laissent instrumentaliser par un gouvernement corrompu », a-t-il dénoncé. (K51SO0607Act1090)

> Excès de zèle. Le député Amir Khadir a participé à une manifestation qu'il savait « illégale » cette semaine à Québec. Ce qui lui a valu les menottes. Et les remontrances de ses adversaires politiques. Vrai que M. Khadir aurait dû, comme élu siégeant à l'Assemblée nationale, s'abstenir de prendre part à une manifestation illégale. (K51SO0609Édi1120)

Tous ces événements sont traversés par les dichotomies *légal/illégal* et *légitime/illégitime*. Le débat entre les actions des étudiants potentiellement illégales mais considérées légitimes par les acteurs et leurs adjuvants constitue une dimension latente de ce cluster. En raisons du rôle politique (député pour QS) de l'acteur principale de ce cluster, soit Amir Khadir, les évènements qui concernent sa famille sont le *symbole* de cette situation de tension entre visions et valeurs opposées.

5.2.5 Le Soleil

Dans l'espace des 117 clusters, le journal *Le Soleil* ne présente aucun cluster isolé. Toutefois, on observe un seul cluster (SO_51) avec une connexion et deux clusters avec deux connexions (SO_19 et SO_28) (figure 5.24). Le premier a déjà été traité au le paragraphe 5.2.4.1. Dans le prochain paragraphe, nous analyserons un des deux clusters avec deux connexions et plus particulièrement, celui qui ne possède que de connexions avec le journal *La Presse*, qui fait partie du même groupe, soit Power Corporation.

Le cluster SO_19 possède deux connexions et toutes les deux avec le journal *La Presse* (PR_47 et PR_23). L'autre cluster avec deux connexions, c'est-à-dire SO_29, est relié à des journaux n'appartenant pas au groupe Power Corporation (DE_18 et JM_18) (figure 5.24).

Figure 5.24 Détail de la figure 5.14. Les clusters SO_19 et SO_28 sont représentés avec leurs connexions.

5.2.5.1 La population et les sondages

Les clusters SO_19 et PR_23 traitent d'un des moyens utilisés par la presse pour sonder l'opinion publique, le **sondage**. Ces articles illustrent ainsi la vision des électeurs québécois face à la crise étudiante. En général, les articles soulignent la présence d'un acteur spécifique, soit l'acteur **population**, lequel regroupe le plus souvent plusieurs autres sous-catégories d'acteurs. De plus, les opinions et les actions que ces acteurs accomplissent en relation avec la grève étudiante sont un sujet des articles.

Dans certains cas, la population est présentée sous la forme de **famille** et assume le rôle d'*adjuvant* du programme narratif principal des **étudiants** qui luttent **contre la hausse** :

> Les familles en appui. Des centaines de citoyens dans les rues de Québec pour supporter les étudiants. 3. Des familles et des gens de toutes les générations sont descendus dans les rues de Québec hier pour appuyer les étudiants et dénoncer les hausses des droits de scolarité du gouvernement de Jean Charest. (K19SO0319Act0189)

D'autres articles soulignent, au contraire, le rôle d'*opposant* de l'acteur **population** :

> La population lasse de la crise. Mobilisation étudiante. Au lendemain de la grande manifestation contre la hausse des droits de scolarité, la population presse les étudiants et le gouvernement d'en finir avec la crise. Un sondage CROP commandé par la Fédération étudiante collégiale du Québec (FECQ) révèle que 78 % des répondants se disent favorables à ce que les étudiants et le gouvernement négocient pour mettre fin aux manifestations étudiantes. (K19SO0323Act0224)

> Québec et les étudiants priés de s'entendre. Sondage. Au lendemain de la grande manifestation contre la hausse des droits de scolarité, la population presse les étudiants et le gouvernement d'en finir avec la crise. Un sondage CROP commandé par la Fédération étudiante collégiale du Québec (FECQ) révèle que les trois quarts des répondants se disent favorables à ce que les étudiants et le gouvernement négocient pour mettre fin aux manifestations étudiantes. (K23PR0323Act0276)

> Les Québécois appuient la hausse. Sondage CROP-Le Soleil-La Presse. Malgré la manifestation monstre, les coups d'éclat et la profusion de carrés rouges, les étudiants sont en train de perdre la bataille de l'opinion publique sur la hausse des droits de scolarité. Les Québécois souhaitent toutefois que le gouvernement négocie avec eux. (K19 SO0331Act0290)

> L'appui aux étudiants s'effrite. Exclusif SONDAGE CROP / LA PRESSE. Davantage de Québécois épousent la position gouvernementale. Après 12 semaines de conflit avec les étudiants, le gouvernement gagne des points auprès

> des Québécois sur la question de la hausse des droits de scolarité, révèle un sondage CROP réalisé à la demande de La Presse. (K23PR0504Act0692)

Toutefois, les **résultats des sondages** constituent le véritable centre d'agrégation qui forme cette macrostructure. Ils assument le rôle d'*adjuvant* ou d'*opposant* dans les programmes narratifs des deux principaux acteurs du corpus, c'est-à-dire le **gouvernement** et les **étudiants**. Dans certains cas, les **sondages** révèlent les intentions de vote des électeurs. Souvent, elles sont perçues par les spécialistes comme un produit des actions d'opposition au **mouvement étudiant** de la part du **gouvernement** :

> Sondage CROP-Le Soleil-La Presse. La crise étudiante profite à Jean Charest. Le chef libéral devance un peu plus la péquiste Pauline Marois et le caquiste François Legault. Son gouvernement est légèrement moins impopulaire. Une « fenêtre » électorale s'entrouvre. Le coup de sonde de la firme CROP, conduit à la veille du conseil général que tiendra à Victoriaville le Parti libéral du Québec, révèle qu'il y a du mouvement dans l'opinion publique, favorable à la formation du premier ministre. Si un scrutin avait eu lieu les 2 et 3 mai, le Parti libéral du Québec (PLQ) aurait obtenu 31 % des voix. (K19SO0504Act0605)

> Cent jours de grève sans départager les partis. SONDAGE CROP - LE SOLEIL - LA PRESSE. La grève étudiante de plus de 100 jours n'a pas départagé les belligérants de la prochaine campagne électorale. Les libéraux de Jean Charest sont très légèrement en avance, en chiffres absolus, sur les péquistes de Pauline Marois. La Coalition avenir Québec (CAQ) de François Legault n'est pas décomptée. (K19SO0526Act0930)

D'autres sondages se concentrent plutôt sur l'appui ou le non-appui des électeurs à la grève étudiante :

> La majorité rangée derrière Charest. SONDAGE CROP - LE SOLEIL - LA PRESSE. Les carrés verts, favorables à la hausse des droits de scolarité, ont gagné la bataille de l'opinion publique selon un sondage CROP-Le Soleil-La Presse dont 68 % des répondants se sont rangés derrière la position du gouvernement et seulement 32 % derrière celle des associations étudiantes qui

> réclament un gel des droits de scolarité. Le sondage révèle également que 66 % des Québécois sont pour la loi spéciale annoncée par le gouvernement Charest. (K19SO0519Act0821)

> Mot d'ordre : négociez! SONDAGE CROP - LE SOLEIL - LA PRESSE. Les Québécois appuient les dispositions de la loi spéciale pour casser le mouvement de grève étudiante. Mais ils jugent que la législation ne règle rien, qu'elle est contraire aux droits et libertés. Négociez avec les jeunes, ordonnent-ils au gouvernement. Un coup de sonde, conduit par CROP du 22 au 25 mai auprès de 1500 personnes, révèle d'apparentes contradictions dans l'humeur des citoyens.

Un élément qui apparaît plus fréquemment, est le rôle d'*adjuvant* de l'opinion des électeurs (communiquée au moyen des sondages) pour le programme narratif ayant comme *objet* le thème des **négociations**. Aussi, les sondages révèlent la volonté de la population de résoudre la crise par le moyen de la négociation :

> Pour la hausse... et la négociation. SONDAGE CROP-LA PRESSE GRÈVE ÉTUDIANTE. Les Québécois ont des opinions polarisées sur la grève étudiante, mais ils refusent d'accorder leur appui total à un camp ou à l'autre. Si la majorité est favorable à une hausse des droits de scolarité, ils sont tout aussi nombreux à demander au gouvernement de mettre de l'eau dans son vin par le truchement de la négociation. (K23PR0331Act0336)

Un autre élément récurrent est, selon les sondages du groupe Power Corporation, l'appui des électeurs au **projet de loi 78**, c'est-à-dire à la loi spéciale, ce qui se transforme souvent en un appui au travail de la **police** :

> Le travail des policiers salué. LOI SPÉCIALE SONDAGE CROP-LA PRESSE. Bouleversés par ce qu'ils voient aux bulletins d'information depuis des semaines, les Québécois approuvent largement les interventions policières et jugent disproportionnées les manifestations à répétition. Pour la population, les excès et le vandalisme sont bien plus le fait de « casseurs » que des étudiants. Dans sa dernière enquête, la maison CROP observe que pas moins de 85 % des gens avouent éprouver un malaise certain (27 %), voire profond (58 %), devant les manifestations violentes des dernières semaines. Seulement 7 % des gens y sont indifférents. (K23PR0519Act0923)

> Appui au travail de la police. CRISE ÉTUDIANTE SONDAGE CROP-LA PRESSE. Québec - En dépit des centaines d'arrestations en une seule soirée, des plaintes en déontologie policière et des images embarrassantes sur les médias sociaux, les Québécois n'ont pas changé d'avis : la police fait correctement son travail. Le dernier sondage CROP réalisé cette semaine constate que selon 77 % des répondants - et 76 % des gens de la grande région de Montréal -, les forces de l'ordre ont fait du « bon » travail dans le contrôle des manifestations étudiantes; 23 % sont d'avis contraire. La police obtient un taux important d'appui à Québec, à 81 %, et chez les aînés, à 84 %. Mais 70 % des moins de 35 ans sont aussi satisfaits du travail des policiers. Une claire majorité de Québécois estime que la police devrait appliquer « très ou assez » rigoureusement les dispositions de la loi d'exception; 58 % des gens sont de cet avis, et 71 % des répondants à Québec. (K23PR0526Act1034)

Toutefois, les résultats des sondages sont très variables sur les différents sujets traités et encore plus particulièrement sur l'adoption de la **loi spéciale** :

> Les Québécois divisés. CRISE ÉTUDIANTE SONDAGE CROP-LA PRESSE. Québec - Les Québécois sont profondément partagés à l'égard de la loi d'exception adoptée la semaine dernière. Ils croient que cette loi ne réglera rien, mais ils en approuvent pourtant le contenu. Une mince majorité de Québécois, 51 %, appuie l'adoption de la loi spéciale, mais une forte majorité de répondants se range derrière la quasi-totalité de ses dispositions. (K23PR0526Act1038)

5.2.6 Le Devoir

Le journal *Le Devoir* présente des caractéristiques différentes par rapport aux journaux analysés jusqu'à maintenant. Ce journal n'est pas associé de manière stricte à un autre journal, comme cela se produit pour ceux des groupes Québecor et Power Corporation. Toutefois, aucun des 117 clusters n'est isolé. Les clusters avec le moins de connexions sont DE_4 et DE_46 (figure 5.25), lesquels ont deux connexions à un seuil de 0,6. Le troisième cluster le moins connecté est DE_6 qui comporte des liens avec plusieurs journaux. En général, les clusters de ce journal sont des nœuds très importants dans le réseau.

Figure 5.25 Détail de la figure 5.14. Les clusters DE_4, DE_6 et DE_46 sont représentés.

Le cluster DE_4 est connecté seulement à des clusters du même journal, soit DE_9 et DE_39 (figure 5.26). Il sera le seul à être utilisé pour l'exploration de spécificités du journal *Le Devoir*.

Figure 5.26 Détail de la figure 5.14. Le cluster DE_4 et ses connexions

5.2.6.1 L'éducation avant tout!

Le groupe d'articles appartenant au cluster DE_4 présente des *opinions sur le rôle et sur la mission de l'institution universitaire*. Ce thème est caractérisé par l'importance majeure que deux actants spécifiques prennent dans la macrostructure, soient le *destinataire* et le *destinateur*. Le *destinataire* est de toute première importance dans un contexte d'évaluation des actions de l'université, alors que le *destinateur* tient un rôle central dans un contexte d'analyse de la légitimité et de la mission des universités. Ces deux rôles actantiels constituent le cœur de la macrostructure qui domine le cluster et ils sont souvent actorialisés par l'**éducation**. La majorité des articles se concentre sur le fait que l'**éducation** est à la fois la mission principale des universités et le principal critère pour l'évaluation de leur enseignement. Enfin, chaque idéologie, entendue comme système de valeurs, place une ou quelque valeurs au sommet d'une hiérarchie. Le Journal *Le Devoir* semble y placer l'éducation. Pour des journaux de droite, la valeur la plus importante est généralement l'économie.

La macrostructure dominante est le plus souvent complétée par un *opposant* qui se concrétise dans les **intérêts privés** ou **marchands**. Ainsi, *Le Devoir*, de manière

assez unique à l'intérieur du corpus, propose un nombre important d'articles qui argumentent contre la « dérive marchande » ou la « marchandisation du savoir » et de l'enseignement supérieur et ceci, en mettant en valeur l'**éducation** comme principe ultime de la mission universitaire :

> Droits de scolarité - Étudier pour... étudier. Que le recteur Guy Breton et sa suite prennent la peine de préciser, après un tel préambule sur l'importance d'étudier pour travailler, que « cela ne s'applique pas seulement aux entreprises privées, mais aussi aux institutions, aux ministères, [...], etc. », ne change pas grand-chose à la vision purement instrumentaliste du savoir qu'il nous présente. Il va de soi que l'université permet, par exemple, d'acquérir des compétences qui seront ensuite reconnues par un ordre professionnel. Il va de soi que la recherche et l'enseignement au sein des universités contribuent au développement économique d'une société. En revanche, il nous apparaît extrêmement réducteur de limiter les objectifs d'une telle institution à ces seules préoccupations. [...] En premier lieu, un grand nombre d'étudiants fréquentent l'université d'abord et avant tout pour acquérir une culture, un savoir, un bagage intellectuel. (K04DE0223Idé0049)

> Libre opinion - Gare au modèle de l'« Université McDonald ». La grève des étudiants contre l'augmentation radicale et rapide des droits de scolarité apparaît comme le symptôme d'un malaise profond par rapport au détournement de la mission de certains services publics et, en particulier, de la mission essentielle de l'enseignement supérieur. (K04DE0712Lib1170)

> Libre opinion - Les universités mises au service des entreprises. La ministre Beauchamp justifie la hausse des droits de scolarité en prétextant que l'augmentation de la contribution étudiante servira à améliorer la « qualité » de l'éducation. Or, c'est bien plutôt l'inverse : la hausse ne profitera pas aux étudiants, mais à des entreprises qui souhaitent brancher directement l'université sur les besoins de l'économie. [...] Quant aux étudiants, c'est la hausse des droits de scolarité et l'endettement qui permettront de les intégrer au nouveau modèle de l'université-entreprise. Selon la vision commerciale de l'université mise de l'avant par le gouvernement libéral, les étudiants sont du capital humain ou, comme le disait le recteur Guy Breton, « des cerveaux » qui doivent être moulés « au service des entreprises ». (K04DE0309Édi0137)

Malade, l'université? En moins de 20 ans, les universités québécoises ont connu une transformation profonde de leur mission fondamentale, passant du modèle à la Humboldt consacré à la vie de l'esprit au modèle du désengagement de l'État, faisant la part belle aux capitaux privés. Pour le meilleur ou pour le pire? Et comment comprendre la hausse dans ce contexte? À chaque époque sa menace à l'intégrité de l'institution universitaire. (K04DE0310Act0142)

La liberté du chercheur en jeu. La tendance mondiale au désengagement de l'État dans les universités au profit du privé sème des inquiétudes. La perte de l'indépendance des chercheurs est-elle un mythe? Les bailleurs de fonds contrôlent-ils la recherche? Un étudiant qui ne répond pas à une question d'examen parce qu'il ignore la réponse, c'est fréquent. Mais un étudiant qui connaît la réponse, mais refuse de la fournir, c'est autrement plus étonnant. L'exception s'est produite il y a quelques années au Massachusetts Institute of Technology (MIT). Le cachottier se disait lié au secret professionnel du projet de recherche sur lequel il travaillait, financé par un bailleur privé. Cette anecdote souvent racontée par l'éminent linguiste Noam Chomsky cristallise les craintes de plusieurs : la recherche libre est-elle encore possible? Devant la croissance du financement privé, le doute persiste. (K04DE0310Act0147)

Hausse des droits de scolarité - John Rawls contre la conception entrepreneuriale de l'université. Le débat actuel va bien au-delà de l'accessibilité et du caractère. Deux fois par mois, Le Devoir lance à des passionnés de philosophie, d'histoire et d'histoire des idées le défi de décrypter une question d'actualité à partir des thèses d'un penseur marquant. . Selon le philosophe américain John Rawls (1921-2002), la justice distributive requiert que les avantages soient attachés à des positions sociales auxquelles tous peuvent parvenir s'ils ont les talents requis. Chaque personne qui a un talent et qui veut le développer doit être en mesure de le faire. C'est le principe de la juste égalité des chances (equality of fair opportunity). Il ne s'agit pas seulement d'assurer une égalité de droit, garantie par la loi, mais de parvenir à institutionnaliser une égalité de fait : le système scolaire doit effectivement permettre à un enfant issu d'une classe défavorisée d'accéder à une carrière adaptée à son talent. (K04DE0526Soc0830)

Compétitivité et autres contrats de performance ont transformé l'université. « Des états généraux s'imposent, à condition qu'on ne limite pas cet exercice aux seuls droits de scolarité ». Doit-on revoir le mode opératoire du réseau universitaire du Québec? Faut-il revoir ses principes directeurs? Est-ce que ce réseau répond aux aspirations et aux défis sociétaux d'aujourd'hui? Peut-on faire

> autrement? Doit-on faire autrement? Pourquoi et comment? Et de quoi sera fait demain? Voilà autant de questions que Le Devoir a soumises à l'examen de deux professeurs réputés, soit Jean Bernatchez, professeur-chercheur en administration et politiques scolaires à l'Université du Québec à Rimouski, et Yves Gingras, professeur d'histoire à l'UQAM et codirecteur de l'ouvrage. Les transformations des universités du 13e au 21e siècle (PUQ, 2006). Chose certaine, pour ces deux professeurs, des états généraux portant sur le présent et le devenir de nos universités s'imposent. (K04DE0818Soc1371)

Le discours contre la marchandisation du savoir est intégré au contexte de la grève étudiante. Le plus souvent, ces arguments sont présentés en *soutien aux idées contre la hausse des frais de scolarité*. Ce soutien est véhiculé également par un autre *destinataire*, l'**accessibilité à l'éducation**.

> Libre opinion - Des oubliées : les études supérieures et la recherche. Dans la foulée de l'augmentation des droits de scolarité décrétée par le gouvernent du Québec -- qui passeront de 2168 $ en 2011-2012 à 3793 $ en 2016-2017 -- et de la grève de quelque 200 000 étudiants collégiaux et universitaires, la plupart des observateurs ont discuté de l'effet négatif de telles hausses sur l'accès aux études de premier cycle. Dans Le Devoir du 23 mars, Pierre Doray et Amélie Groleau nous rappelaient qu'à la suite du dernier dégel des frais de scolarité, les universités francophones avaient subi une baisse de plus de 26 000 inscriptions entre 1992 et 1997. Peu d'observateurs ont toutefois noté l'effet délétère d'une telle hausse sur l'accès aux études de cycle supérieur et, par conséquent, sur la capacité de recherche des universités québécoises. Les données de l'Association canadienne pour les études supérieures nous montrent que les inscriptions dans les programmes de doctorat furent également touchées, avec un certain décalage compte tenu du temps nécessaire au passage des études de premier cycle aux études doctorales. Pour les trois principales universités de recherche du Québec (Laval, McGill et Montréal), qui comptent pour plus des trois quarts des doctorants en 1995, les inscriptions au doctorat sont passées de 6792 à 5880 entre 1995 et 2001, soit une baisse de plus de 13 %. Certaines universités furent également plus touchées que d'autres : les programmes de doctorat de l'Université de Montréal, par exemple, ont subi une baisse de plus de 22 % (2865 en 1995 contre 2229 en 2001). (K04DE0410Édi0356)

Les arguments contre la hausse de droits de scolarité sont également accompagnés de considérations sur la *gestion des finances publiques et des universités*. Un *adjuvant*,

actorialisé par l'**amélioration de la gestion des universités**, apparaît en soutien aux idées contre la hausse des droits de scolarité.

> Grève étudiante - Des idées pour mettre fin au « gaspillage ». Les fédérations étudiantes proposent de couper dans les budgets universitaires. Les fédérations étudiantes proposent de geler certains budgets des universités afin de pouvoir maintenir le gel des droits de scolarité. (K04DE0412Act0371)

> Universités - Un sous-financement bien réel. Lorsque le gouvernement du Québec a annoncé un éventail de mesures visant à assurer un meilleur financement aux universités, incluant une hausse graduelle des droits de scolarité, il répondait à un besoin bien réel. Le sous-financement des universités québécoises est documenté et reconnu de tous. Il se chiffre à plus de 600 millions de dollars par année. Or, aujourd'hui, l'objectif fondamental est occulté par la recherche d'économies qui pourraient être réalisées par les universités. Comme administrateurs universitaires, nous nous inquiétons de cette dérive et des risques qu'elle comporte. (K04DE0512Idé0680)

Enfin, le journal lui-même prend position en assumant, à la première personne, le rôle de destinateur et ceci, par un débrayage énonciatif qui est unique à l'intérieur du corpus. Le débrayage est :

> l'opération par laquelle l'instance d'énonciation se disjoint et projette hors d'elle, lors de l'acte de langage et en vue de la manifestation, certains termes liés à sa structure de base pour constituer ainsi les éléments fondateurs de l'énoncé-discours (Greimas et Courtés, 1979, p. 79)

Généralement, le débrayage implique trois catégories d'entité, soit la personne, le temps et l'espace. Dans le cas qui nous présentons ici, il s'agit surtout d'un débrayage qui implique la personne, et plus particulièrement l'auteur ultime de l'article.

Ceci se produit de manière très spécifique. En effet, le débrayage énonciatif constitue rarement une dynamique principale des articles et ceci l'est encore moins lorsque le destinateur est le journal lui-même. Quand le *je* apparaît dans un espace et dans un

temps donnés, le débrayage devient un élément dominant de l'énoncé. Dans le cas des éditoriaux ou des lettres, le débrayage est plus marqué, étant donné qu'une opinion est exprimée. Mais dans tous ces cas, le *je* qui constitue le débrayage implique le journaliste ou l'auteur de l'article ou de la lettre. Au contraire, dans le cas présenté ici, le débrayage concerne un *je* qui personnifie le journal au complet. Dans l'article qui suit, *Le Devoir* prend la parole en première personne et exprime une opinion contre le *recteur de l'Université de Montréal*, accusé d'avoir permis des actes d'intimidations envers les étudiants à l'intérieur de l'institution qu'il dirige.

> Libre opinion - Nous demandons la démission du recteur. Ces derniers jours ont été désastreux à l'Université de Montréal. Les mots ne suffisent pas pour les décrire. Nous pouvons et nous devons néanmoins condamner ces gestes de violence et d'intimidation. Parce qu'il s'agit de gestes de violence et d'intimidation. [...] Cette semaine, l'administration est encore allée plus loin. Beaucoup trop loin. Agents de sécurité, policiers, anti-émeutes se sont croisés et sont intervenus dans les couloirs de l'université comme s'il s'agissait d'un lieu de passage habituel, comme si l'université n'avait jamais été ce sanctuaire qu'elle a jadis pu être, comme si l'université n'était plus ce lieu de protection mais bien de répression. [...] Bref, seule la direction a le pouvoir de bafouer les principes fondateurs de l'université. Que reste-t-il, par exemple, du principe de collégialité quand on place les professeurs dans une situation impossible? Quand on ne leur donne pas la possibilité de respecter des votes de grève pris par leurs étudiants et qu'on les oblige, sous peine d'amendes salées, à donner leurs cours et ce, dans une ambiance des plus pesantes et traumatisantes? Valeur de l'université. Par ailleurs, une université est-elle encore une université quand en ses lieux se déroulent arrestations, gardes à vue, souricières et charges de l'anti émeute? Quand des étudiants sont considérés comme de dangereux criminels alors qu'ils tentent simplement de faire respecter des décisions prises démocratiquement et collectivement? (K04DE0831Lib1440)

Le cluster présenté ici est très distinctif et caractérise le journal *Le Devoir* de manière unique. En particulier, le dernier article n'a de similarité avec aucun des milliers d'articles traités dans cette recherche. Toutefois, cette considération vaut également pour le cluster au complet. En effet, même en tenant compte des clusters de plus

petite taille, ce cluster représente un groupe d'articles très différents du reste des articles contenus dans le corpus.

5.2.7 Radio-Canada

Pour le site web *Radio-Canada*, aucun cluster n'est isolé et deux ont une seule connexion, soit RC_10 et RC_28 (figure 5.27). Comme *Le Devoir*, cette source d'information n'a pas de connexions fréquentes avec un journal en particulier. Sa caractéristique principale est celle d'être un site web et, au contraire des autres, ses articles sont souvent publiés à l'intérieur de la journée même où les événements se déroulent. Dans certains cas, comme pendant la couverture des manifestations, les articles décrivent les événements en direct, ce qui est unique dans le corpus.

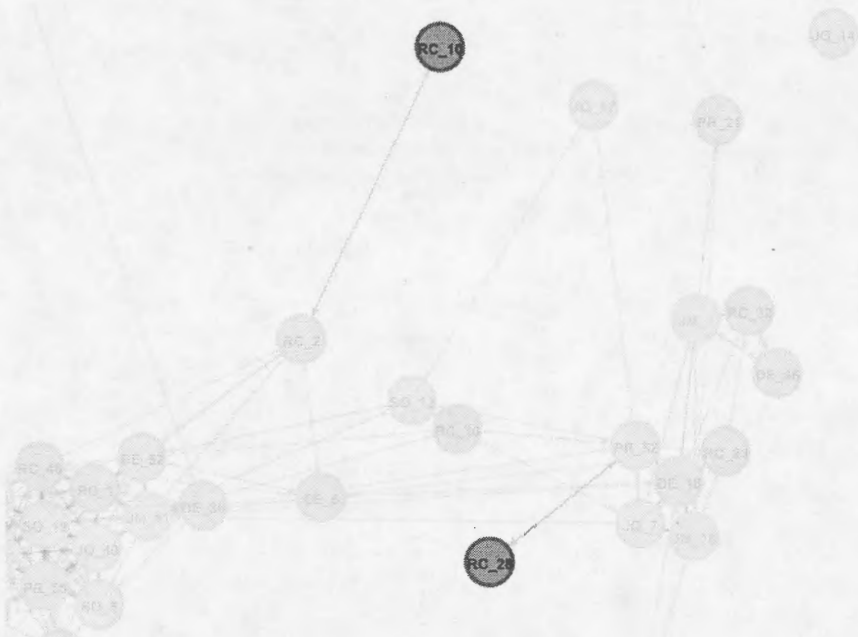


Figure 5.27 Détail de la figure 5.14. Les clusters RC_16 et RC_28 sont représentés

Comme dans le cas des derniers journaux analysés, seul le cluster connecté avec la même source d'information sera présenté. Pour *Radio-Canada*, le cluster à être analysé est RC_10 qui est connecté à RC_2.

5.2.7.1 Chronique des manifestations

Les deux clusters connectés de *Radio-Canada* (RC_10 et RC_2) regroupent des articles sur les *manifestations étudiantes*. Les extraits qui suivent représentent des articles prototypes de RC_10. Ce sont généralement des chroniques qui racontent les évènements ou qui fournissent des informations générales sur les manifestations. En particulier, RC_10 regroupe surtout les chroniques des manifestations qui se sont déroulées devant **les édifices du Parlement**, à Québec:

> Droits de scolarité : rassemblement étudiant sur la colline Parlementaire. Des étudiants se sont donné rendez-vous à l'occasion de la rentrée parlementaire pour manifester devant l'Assemblée nationale contre la hausse des droits de scolarité. (K10RC0214no_0004)

> Près de 100 000 étudiants en grève aujourd'hui. Des milliers d'étudiants sont attendus jeudi après-midi devant l'Assemblée nationale, à Québec, pour dénoncer de nouveau la hausse des droits de scolarité annoncés dans le dernier budget du gouvernement de Jean Charest. (K10RC0301no_0036)

> Droits de scolarité : des milliers d'étudiants devant l'Assemblée nationale. De 4000 à 5000 étudiants de partout au Québec sont massés devant l'Assemblée nationale jeudi après-midi pour protester contre la hausse des droits de scolarité. Quelques étudiants ont franchi les barrières de sécurité qui avaient été érigées pour l'occasion. Des policiers de l'escouade anti-émeute de la Sûreté du Québec sont intervenus en utilisant des gaz lacrymogènes pour disperser les étudiants. (K10RC0301no_0038)

> Les étudiants manifestent à Québec en ce jour de budget. Quelques dizaines d'entre eux se sont rassemblés devant le Parlement de Québec mardi en attendant le dépôt du budget du gouvernement Charest en après-midi. (K10RC0320no_0097)

> Journée de manifestations étudiantes à Québec. Les étudiants manifestent à Québec en ce jour de budget. Quelques dizaines d'entre eux effectuent une opération de visibilité, mardi matin, à l'intersection de la rue de la Couronne et du boulevard Charest. (K10RC0320no_0100)

> Carrés rouges devant l'Assemblée nationale pour le 22 juillet. Des manifestants se donnent rendez-vous devant l'Assemblée nationale à Québec en ce 22 juillet pour dénoncer la hausse des droits de scolarité, comme c'est le cas depuis le 22 mars. Des rassemblements sont organisés dans plusieurs autres villes du Québec. (K10RC0722no_0764)

Le cluster RC_2 regroupe des articles sur des évènements moins spécifiques. Ainsi, il contient des articles sur plusieurs manifestations étudiantes. La spécificité de ces deux clusters réside dans la manière de couvrir les défilés étudiants. Une étude détaillée sur ces deux clusters pourrait mettre en évidence un style journalistique sans doute spécifique à cette source d'information.

CHAPITRE VI

DISCUSSION

Dans ce chapitre, nous discuterons des résultats obtenus et de certains aspects méthodologiques. Les réflexions ici regroupées constitueront l'évaluation de la méthode, ce qui implique deux considérations préliminaires. D'abord, la méthode, qui a été présentée comme une preuve de concept, ne peut pas être évaluée par une comparaison avec un travail de référence puisque, à notre connaissance, il n'existe pas de recherches qui ont analysé le printemps érable en utilisant les mêmes cadres théoriques que cette thèse. Les travaux que nous avons cités au point (1.7.3) ne peuvent pas être utilisés comme référence, puisque le cadre théorique auquel ils réfèrent diffère de celui de ce travail et aussi parce que les corpus analysés sont différents en termes de taille et de typologie. En général, ces travaux exécutent des analyses sur des corpus de petite taille ou sur des unités syntagmatiques très différentes des nôtres (cfr 1.7.3). Certaines études, par exemple, n'analysent que les premières pages. Deuxièmement, il n'est pas possible de bâtir une évaluation quantitative pour l'application des outils non supervisés. Cette problématique a été introduite au point 3.6.2. En effet, ce type de méthode est le plus souvent évalué par une phase d'annotation, tel que proposé dans cette thèse. Certains lecteurs pourraient objecter sur ce point[37]. En effet, il est possible de créer des heuristiques, mais le dégrée de fiabilité ne sera pas élevé. Pour l'augmenter, il sera nécessaire d'améliorer

[37] L'impossibilité de construire une mesure quantitative pour l'évaluation de méthodes non supervisées appliquées aux données textuelles

les ressources économiques, humaines, technologiques, etc. du projet de recherche. Par exemple, il aurait été possible de calculer une valeur sur la base du pourcentage des clusters qui contiennent effectivement une macrostructure et ceux qui ne la contiennent pas. Toutefois, cette mesure devrait prendre en compte toute une série de facteurs, dont le niveau de confiance de l'annotateur et l'accord interjuge, ce qui implique la mise en place d'un questionnaire de qualité et d'une équipe d'annotateurs. De plus, étant donné la spécificité de la méthode utilisée et du métalangage greimassien, les annotateurs auraient dû être des sémioticiens. Les ressources pour bâtir une telle infrastructure n'étaient pas disponibles. Dans un contexte de démonstration de faisabilité d'une méthode hybride entre sémiotique et intelligence artificielle, nous nous sommes donc limités à une évaluation « qualitative » des résultats obtenus et à une discussion sur la méthode.

Les questions de recherche de cette thèse correspondent à deux *objectifs de recherche* : A) explorer la couverture médiatique du cas d'étude en se basant sur le concept de macrostructure et B) exécuter cette analyse au moyen d'une méthode d'analyse de texte assistée par ordinateur qui associe un outil computationnel spécifique à l'analyse narrative du texte journalistique. Le point A) sera discuté dans la première section de ce chapitre et le point B) dans la deuxième.

6.1 Réflexions sur les résultats

Les résultats présentés au chapitre V montrent que la méthode permet d'explorer un corpus de type journalistique et d'en souligner les macrostructures les plus fréquentes. En effet, il a été possible d'identifier *sept grands agglomérats de clusters* dont le niveau discursif des articles est généré par des macrostructures communes. Pour les six premiers agglomérats, la détection d'un ou de deux axes du vouloir (sujet et objet de valeur) a permis l'analyse des articles à partir d'une macrostructure commune,

constituée par cet axe et par le programme narratif correspondant. Pour certains agglomérats, des axes du pouvoir (adjuvant et opposant) ou, plus rarement, de la transmission (destinateur et destinataire), ont également pu être dégagés. Ceci constitue certainement un élément positif dans l'évaluation de notre méthode Puisque, il a été possible d'analyser, à l'aide de l'ordinateur, un grand corpus de textes de type journalistique en ciblant ses macrostructures. Cette analyse a relevé que les articles similaires possèdent des macrostructures similaires et que ces dernières possèdent des caractéristiques narratives.

Toutefois, pour le dernier agglomérat, il n'a pas été possible d'identifier une macrostructure comme nous en avions fait l'hypothèse précédemment. Toutefois, un élément s'est dégagé clairement pendant la phase d'annotation de ces articles et c'est la mise en valeur du *destinataire*, ce qui permet certainement de faire une interprétation pertinente des articles contenus dans l'agglomérat. L'analyse de cet agglomérat a montré la présence d'une catégorie particulière d'articles, soit les *articles d'opinion.* En effet, les articles de cet agglomérat appartiennent principalement à la catégorie des éditoriaux, des lettres des lecteurs et autres. L'analyse a aussi montré que certains articles, qui ont été classifiés par les journaux sous la catégorie de chroniques, étaient en réalité structurés sous la forme d'articles d'opinion. En effet, plusieurs articles classifiés comme « nouvelles » ou « actualités », qui sont généralement dédiés à des chroniques, constituent en réalité de véritables expressions d'opinion. Dans ce contexte, la mise en valeur du destinataire est justifiée et il constitue un outil au service de l'interprétation. En effet, lorsque dans un article on argumente *pour* ou *contre* quelque chose, on se prête inévitablement à une dynamique sémiotique qui « pointe » vers la dernière phase du schéma narratif canonique, c'est-à-dire la phase d'évaluation, ce qui implique l'émergence du destinataire. En d'autres termes, exprimer une opinion comporte généralement la structuration d'une argumentation qui se prête, par définition, à l'évaluation de sa

légitimité. En effet, raconter des événements répond à des dynamiques différentes de celles en fonction lors de l'expression d'une opinion.

Généralement, l'argumentation d'une opinion est fortement marquée par une *idéologie* ou une *logique des idées*, entendue comme un système d'idées et de principes logiquement cohérents. Prenons une application classique du modèle actantiel, fournie par Greimas lui-même (Greimas, 1966, p. 181), soit la distribution actantielle de l'idéologie marxiste (tableau 6.1). Dans ce cas, les actions du militant marxiste *sont évaluées* par le *destinataire*, l'*humanité*. C'est au nom de l'*humanité* que l'*homme* exécute des actions pour atteindre la *société sans classe*.

Tableau 6.1 Schéma actantiel de l'idéologie marxiste selon Greimas

Actant	**Acteur**
Sujet	Homme
Objet	Société sans classe
Destinateur	Histoire
Destinataire	Humanité
Opposant	Classe bourgeoise
Adjuvant	Classe ouvrière

La même dynamique est présente dans la majorité des articles d'opinion analysés dans le septième agglomérat. Le plus souvent, les articles fournissent un élément qui présente une argumentation selon la formule « c'est au nom de *x* que cette opinion est légitimée », ce qui implique la mise en valeur du *destinataire*. Une analyse plus approfondie de cet agglomérat fournirait certainement des pistes pour la définition de dynamiques spécifiques au genre «article d'opinion».

L'analyse du septième agglomérat n'est pas le seul élément qui montre les limites de la méthode. Les plus grandes difficultés ont été rencontrées lorsque nous avons essayé de dégager des macrostructures spécifiques par journal. Il a été possible de définir une macrostructure pour les articles des clusters ciblés, par exemple ceux

décrits au point 5.2.3.2. Toutefois, il a été plus difficile d'identifier des macrostructures qui définissent des caractéristiques spécifiques par journal. Ceci peut dépendre de plusieurs facteurs, et en premier lieu, des caractéristiques intrinsèques à la méthode. En effet, la méthode est basée sur la *notion de similarité*, ce qui implique une mise en relief des macrostructures similaires plutôt que de celles qui sont distinctes. D'autres outils pourraient être explorés pour réaliser cette tâche. Deuxièmement, l'approche par macrostructure sémantique appliquée à un corpus de type journalistique et ciblant les aspects sémantiques des textes peut constituer un « biais méthodologique » pour la réalisation de ce genre de tâche. En effet, lorsqu'on exécute une analyse du contenu sémantique pour définir les structures globales du sens et que le corpus analysé est composé d'articles de journaux et que ces articles traitent des mêmes événements, l'identification des spécificités par journal est difficilement réalisable. Dans notre cas, les journaux traitent des mêmes manifestations, des mêmes sessions parlementaires, des mêmes déclarations publiques, etc. Dans un tel cadre, l'analyse du niveau discursif et de surface serait plus pertinente pour l'identification des éléments distinctifs par source d'information. Le niveau de surface pourrait souligner le style ou la manière à travers lesquels les évènements sont traités, ce qui est plus susceptible de constituer les spécificités par journal.

L'accomplissent le plus important que la méthode permet de réaliser est certainement la *découverte* d'un certain nombre d'*hypothèses de recherche*. Lorsqu'on explore un corpus, un des objectifs est d'identifier des pistes de recherche pouvant mener à des études approfondies du corpus. Dans notre cas, chaque agglomérat ou cluster décrit au chapitre 5 constitue certainement une hypothèse de recherche qui pourrait être approfondie davantage. En d'autres termes, à partir de l'identification d'une série de macrostructures, l'analyse de contenu peut être approfondie ou précisée davantage. Ceci constitue un des avantages importants d'une sémiotique computationnelle, c'est-

à-dire la découverte de « nouveaux observables » par des méthodes assistées par ordinateur. Comme Rastier le spécifie à plusieurs reprises (Rastier, 2011), une analyse à grande échelle effectuée à l'aide de l'ordinateur pourrait permettre d'identifier des éléments qui, sans aucun doute, n'auraient pas pu être identifiés autrement. Par exemple, dans notre cas, nous avons identifié un certain nombre de macrostructures plus importantes que d'autres. Est-ce qu'une analyse effectuée par une méthode traditionnelle et sans assistance par ordinateur aurait permis d'identifier les mêmes macrostructures? Si oui, en combien de temps et en utilisant quelles ressources? Il n'est pas possible de répondre à cette question dans le cadre de ce travail, mais elle pourrait certainement inspirer de travaux futurs.

6.2 La méthode et ses implications sémiotiques

L'exploration d'un corpus de type journalistique par l'analyse de ses macrostructures a été réalisée avec l'assistance d'un ordinateur, ce qui réalise le deuxième objectif de la thèse (B), soit associer un outil computationnel à l'analyse narrative du texte journalistique. Ceci est accompagné d'une autre contribution majeure, soit le transfert de connaissances de l'informatique à la sémiotique, car la méthode constitue le principal élément de ce travail. Cette thèse s'est développée sous la forme d'une démonstration de faisabilité, dans le but ultime d'évaluer une méthode d'assistance computationnelle à l'analyse sémiotique du texte journalistique. Si d'une part, l'obtention de résultats cohérents avec les attentes habituelles d'une recherche en sémiotique est déterminante pour l'évaluation de la méthode (voir point 6.1), de l'autre, une réflexion sur la méthode l'est autant. En effet, l'utilisation d'une telle méthode hybride doit inévitablement passer par une compréhension détaillée des aspects techniques des outils utilisés et par une prise de conscience des caractéristiques de ces outils lors de leur application à des produits sémiotiques. En d'autres termes, pour le développement des méthodes de la sémiotique

computationnelle, il est nécessaire de bien comprendre quelles sont *les opérations cognitives et sémiotiques que les outils automatiques exécutent à la place de l'être humain*. Il est donc impératif de comprendre comment ces opérations sont exécutées afin de déterminer la façon dont ces outils peuvent être utilisés pour l'analyse sémiotique. Cette prise de conscience permet d'éviter des interprétations erronées des résultats ou une utilisation non appropriée des outils, laquelle pouvant mener à de fausses conclusions. Cette compréhension et cette prise de conscience sont les conditions nécessaires pour la construction d'une sémiotique computationnelle. Enfin, le modèle de la « boîte noire » (*black box*), c'est-à-dire l'utilisation de programmes informatiques (peu importe s'ils sont sous la forme d'algorithmes ou de logiciels) sans connaître les opérations qu'ils exécutent, *ne doit pas* et *ne peut pas* constituer une base solide pour la sémiotique computationnelle.

Dans les paragraphes qui suivent, un certain nombre d'étapes de la méthode sont analysées du point de vue technique ou sémiotique. Chacune de ces étapes a une importance majeure pour une utilisation correcte des outils dans le cadre d'une recherche sémiotique. Certaines d'entre elles sont des opérations simples, comme la tokenisation. D'autres présentent des problèmes techniques plus pointus et dont il est important d'en connaître la nature afin d'éviter une utilisation biaisée des outils, par exemple le problème de la « malédiction dimensionnelle ». Cette discussion méthodologique ne prétend pas être exhaustive, mais elle est certainement représentative des problèmes principaux que la recherche en sémiotique doit affronter dans un contexte computationnel.

6.2.1 Tokenisation

L'opération de *tokenisation* est une procédure de *reconnaissance des signes graphiques* qui sont susceptibles de constituer un *token*, lequel a été désigné comme étant un *sinsigne* (cfr. 2.1.1). Cette opération est plus complexe à réaliser pour une

machine que pour un être humain, car il est difficile de formaliser et de transférer l'intelligence et l'expérience nécessaires à l'exécution de cette tâche à un ordinateur. La complexité de l'opération dépend surtout de la langue et du système d'écriture auquel le chercheur est confronté. Prétraiter un texte en arabe nécessite des outils différents de ceux utilisés pour prétraiter l'anglais. En général, la tokenisation est une procédure qui requiert un certain nombre d'outils de reconnaissance des formes et qui est mise en place en fonction de la langue à traiter.

La réussite de l'opération de tokenisation dépend principalement de la définition des *délimiteurs* des unités lexicales. Par exemple, on peut considérer comme unité lexicale chaque séquence de caractères précédée et suivie par un espace. Cette définition peut fonctionner pour une langue néo-latine, comme l'italien ou le français, mais ne fonctionnera pas aussi bien pour l'arabe ou l'allemand. Conséquemment, des stratégies différentes seront définies en fonction de la langue et du système d'écriture traités. Identifier les délimiteurs des mots requiert la connaissance des conventions orthographiques du système d'écriture auquel on est confronté. En amharique par exemple, les séparateurs des mots et des phrases sont explicitement indiqués, ce qui n'est pas le cas pour le thaï. Chaque système d'écriture possède donc sa *convention orthographique* pour l'identification d'un mot et il est généralement possible de distinguer les c*onventions basées sur les espaces* et les *conventions qui n'utilisent pas de délimiteurs de mots standards*. Dans le premier cas, qui correspond à la majorité des langues européennes, les mots sont généralement séparés par un espace et il suffit d'identifier les suites de caractères délimitées par les espaces. Dans le deuxième cas, comme pour le chinois, le thaï ou, en partie l'allemand, les mots ne sont pas séparés par un espace. Les méthodes de tokenisation sont très différentes selon la langue à laquelle elles sont apliquées et il n'est pas opportun d'en faire la synthèse ici.

Le français, qui est la langue traitée dans cette thèse, appartient au premier cas. Bien que le processus de tokenisation puisse apparaitre simple, cette typologie de langue présente d'autres enjeux, comme l'*identification des suites spéciales de caractères*, dont les règles varient d'une langue à une autre et d'un domaine du savoir à un autre. Par exemple, certaines conventions peuvent être spécifiques à des textes de mathématique ou de philosophie. Certains algorithmes peuvent donc être très spécifiques au corpus analysé, selon la langue utilisée et le domaine en cause.

Dans le contexte de la thèse, deux types de suites spéciales sont présentées : les suites *prévisibles* et dépendantes de conventions inscrites dans le système d'écriture ou du domaine du savoir et les suites *imprévisibles*, qui dépendent d'erreurs de toutes sortes. Pour la première catégorie par exemple, il est possible de prévoir les abréviations des auxiliaires être ou avoir, c'est le cas de l'expression anglaise « I'm », où il est possible de déterminer comment l'apostrophe doit être traitée dans différentes situations ou, comme le génitif en anglais. On retrouve dans cette première catégorie toutes les procédures de désambiguïsation des occurrences de certains signes graphiques dont le comportement ou le rôle sont prévisibles. L'apostrophe (') en est un exemple, mais celui du point (.) est plus répandu, comme délimiteur de phrases, d'acronymes ou d'abréviations. Ce premier enjeu a été pris en charge par la méthode Treetagger, qui a été utilisée pour l'annotation morphosyntaxique et la lemmatisation. Toutefois, plusieurs erreurs ont été trouvées et un processus de correction manuelle s'est imposé.

À cette même catégorie, s'ajoutent les mots composés ou des cas particuliers d'utilisation du trait d'union. En effet, dans certains cas en français, une suite de caractères qui contient un trait d'union peut être considérée comme une seule unité lexicale, par exemple les mots composés, mais dans d'autres cas, le trait d'union doit plutôt être considéré comme un séparateur de mots, comme dans les phrases

interrogatives (ex. « A-t-il mangé? »). L'identification de ces suites spéciales de caractères est exécutée par des règles construites *ad hoc* qui sont généralement mises en œuvre par des *expressions régulières* communément appelées *regex*. Il est possible de définir les regex comme un langage spécifique pour la manipulation de caractères. Reprenons l'exemple du point, qui peut être un délimiteur de phrases, d'abréviations, d'acronymes et aussi un délimiteur pour les décimales. Pour chacun de ces cas, le *lexicographe computationnel* doit adopter une stratégie permettant de les identifier et les désambiguïser, ce qui implique un regex capable de distinguer le point délimiteur d'abréviation de celui qui délimite les phrases.

En ce qui concerne les suites imprévisibles, les stratégies d'identification ne peuvent s'appuyer sur les connaissances que le chercheur a de la langue, du système d'écriture ou du domaine du savoir dans lequel le corpus s'inscrit. Ces suites spéciales de caractères constituent le plus souvent des erreurs de transcription ou des erreurs de reconnaissance optique des caractères. Dans la pratique, il n'est pas rare de devoir composer avec des erreurs comme des espaces ou des caractères non reconnus et il est difficile d'identifier chaque suite de caractères qui dérive des erreurs. Pour cette raison, *chaque opération de tokenisation comporte une marge d'erreur inévitable* et qui ne peut être estimée que dans un très petit nombre de cas. Dans le cas du présent travail, un grand nombre de ces erreurs ont été corrigées par l'utilisation de règles construites *ad hoc* mises en œuvre par des *expressions régulières,* les *regex*.

En d'autres termes, la tokenisation est une opération complexe, en raison des conventions orthographiques des langues, des conventions du domaine du savoir et des différentes problématiques comme les suites de caractères prévisibles et imprévisibles. La qualité du texte traité, qui peut contenir des erreurs de transcription ou d'autres types ajoute à cette complexité. L'identification des *lexies* (Rastier, 2012, 2009), qui sont les véritables unités lexicales d'une langue, est un problème délicat. Il

peut exister des lexies constituées d'un mot, d'autres d'un mot composé ou de suites de mots, comme « pommes de terre » ou « au fur et à mesure ». En somme, la tokenisation est une étape difficile à réaliser et des stratégies spécifiques doivent être considérées. Une marge d'erreur doit être tolérée. Un travail important de nettoyage des textes par des règles *regex* construites *ad hoc doit être effectuées afin de réduire l'effet néfaste de ce type de bruit sur les résultats*.

6.2.2 Lemmatisation

Certaines opérations de la phase d'annotation automatique ont des implications sémiotiques plus importantes que d'autres. C'est le cas de la lemmatisation, dont il est opportun d'en connaitre les caractéristiques sémiotiques et les limites. La lemmatisation est l'opération qui identifie le *morphème porteur du signifié* pour les unités lexicales des langues flexionnelles. Dans un cadre d'analyse de texte assistée par ordinateur, cette opération est accomplie pour *normaliser* les formes des mots, c'est-à-dire pour regrouper les tokens en *types*. Certains auteurs (Eco, 1978) ont associé cette dichotomie au modèle du signe de Hjelmslev (1943) qui suit :

(matière) *substance* *forme*	CONTENU
forme *substance* (matière)	EXPRESSION

Figure 6.1 Expression et contenu selon Hjelmslev

Ce modèle (figure 6.1) implique une structure du signe à deux plans, celui du plan du contenu (signifié) et celui de l'expression (signifiant). Le modèle de Hjemslev complète la théorie du signe proposée par Saussure en ajoutant à chaque plan trois niveaux de transformation. Nous citons un passage du texte de Eco à partir duquel la réflexion qui suit a débuté :

> Selon ce modèle, on définit comme matière de l'expression tout continu amorphe auquel un système sémiotique déterminé donne forme en en découpant des éléments pertinents et structurés et en les produisant ensuite comme substance; et l'on définit comme matière du contenu l'univers en tant que champ de l'expérience auquel une culture déterminée donne forme en en découpant des éléments pertinents et structurés et en les communiquant ensuite comme substance. La différence entre un élément de la forme et un élément de la substance est celle qui intervient entre un *type* et une occurrence concrète (*token*). (Eco, 1978, p. 141)

En d'autres termes, la matière de l'expression et la matière du contenu sont un « tout continu amorphe » alors que la substance est la matière structurée par la forme. Ces deux plans ne sont pas indépendants, mais étroitement reliés l'un à l'autre. En effet, sur le plan du contenu, la matière amorphe est découpée en unités qui constituent la substance du contenu ce qui est permis par la forme. Par exemple, le signifié général de « mourir » peut avoir différentes formes : « décéder » ou « partir », etc. Chacune de ces formes est porteuse d'un signifié sur le plan du contenu qui se manifeste dans une forme particulière sur le plan de l'expression. Chacun des mots suivants « décéder », « mourir » ou « partir » est une forme de l'expression qui correspond à des formes du contenu. En effet, ces mots sont utilisés dans des situations différentes et ont un signifié différent ou, à tout le moins, véhiculent des intentions différentes. Cependant, *la substance demeure la même, soit la notion générale de mourir*. Quel est donc le lien entre forme et contenu d'un côté, et type et token de l'autre? La dichotomie entre type et token peut être appréhendée grâce au modèle de Hjelmslev afin de considérer « la différence entre un élément de la forme et un élément de la substance [comme celle] qui intervient entre un type et une occurrence concrète (token) (Eco, 1978, p. 141).

Selon Eco, type et token correspondent respectivement à forme et substance. Dans ce sens, le type « mourir » peut avoir différents tokens, comme « partir », « décéder », etc.

Cependant, dans un cadre d'analyse linguistique et dans le contexte de l'opération de lemmatisation, ce parallélisme n'est pas applicable tel quel, ce qui constitue certainement une limite importante de la méthode. L'opération de lemmatisation permet de « normaliser » les différentes occurrences et formes qu'un mot peut prendre et donc de trouver les types mais seulement dans un cadre linguistique du genre « partir » (type) versus « partais », « part » ou « partirais ». Elle ne permet pas de normaliser toutefois les formes et occurrences d'un mot comme «mourir» (type) par rapport à «décéder», «mourir» ou «partir». Dans certains contextes, le mot « partir » est aussi un token de « mourir » et la réduction sémantique devrait donc suivre celle linguistique.

En d'autres termes, cette vision sémantique de la relation entre type et token est très difficile à concrétiser dans un cadre computationnel, car elle nécessite un algorithme de lemmatisation qui, en analysant le contexte d'occurrence d'un mot, peut trouver son type linguistique mais aussi son type sémantique. Par exemple, l'algorithme devrait transformer la phrase « il est parti serein » selon la liste suivante de lemmes, « il », « être », « mourir », « serein », ce qui est actuellement très difficile à exécuter par des outils informatiques. La parfaite mise en relation entre la substance du contenu et la substance de l'expression devra donc être accomplie par un outil de lemmatisation plus sophistiqué, fondé sur une théorie sémiotique du type, afin de dépasser les limites de l'approche lexicographique. Le modèle de Hjelmslev offre un cadre sémiotique satisfaisant pour interpréter la différence entre type et token et, en particulier, pour comprendre le rôle de la lemmatisation dans un cadre d'analyse de texte assistée par ordinateur.

Malgré les limites actuelles, l'opération de normalisation ou de réduction du token à son lemme n'est pas inutile et elle est accomplie essentiellement pour mettre en valeur les éléments sémantiques des textes et pour réduire les impacts potentiels des

aspects syntaxiques. En effet, regrouper les différents tokens autour d'un type qui correspond au lemme exige de pouvoir comparer les mots du point de vue sémantique et non selon la position syntaxique qu'ils occupent dans la phrase. Par exemple, si dans un texte il existe 100 occurrences du mot « mourir », distribuées en dix ou 20 formes différentes, l'analyse ne sera pas efficace, car la significativité du mot « mourir » est diminuée par les formes qu'elle prend dans le texte, chacune reflétant une portion de son occurrence totale. Ainsi, réduire chaque occurrence à un seul type augmente la significativité du verbe « mourir ».

Dans une application d'analyse de texte, il est préférable d'orienter les opérations de prétraitement vers les aspects sémantiques plutôt que syntaxiques. Il est donc moins important de déterminer si un verbe a été utilisé plus fréquemment à la troisième personne qu'à la première, que de connaître le nombre de fois où il a été utilisé et avec quel autre lemme. Évidemment, cette hypothèse peut varier en fonction de l'application et de l'objectif de la recherche. Le présent projet est orienté sur les aspects sémantiques du texte et c'est pour cette raison que l'étape de lemmatisation est accomplie pour regrouper les tokens sous le même type correspondant au morphème porteur du signifié d'un mot, c'est-à-dire le lemme. Toutefois, une vision sémantique de l'opération de lemmatisation, tel que décrite dans le modèle de Hjemslev, serait certainement une valeur ajoutée à la méthode.

6.2.3 Segmentation

L'*unité syntagmatique élémentaire* est une autre opération qui implique des choix de type sémiotique. Cette unité est l'observable sur lequel les méthodes computationnelles exécutent leurs manipulations. Lorsque le paragraphe, l'article au complet ou la phrase sont choisis comme unité élémentaire, les résultats doivent être interprétés de manière différente. Chacun de ces choix comporte aussi des problèmes techniques de type computable ou linguistique.

Par exemple, l'opération de segmentation en phrases peut comporter différents niveaux de complexité et cela, selon la notion d'unité syntagmatique élémentaire que le chercheur est prêt à adopter. De manière simple, on peut dire qu'il est possible de distinguer une séquence de syntagmes délimitée par une ponctuation et une séquence de syntagmes indépendante du point de vue grammatical. Ainsi, on peut donc distinguer la *phrase* de la *proposition,* cette dernière devant être comprise dans sa conception linguistique classique. Une phrase est généralement délimitée par un signe de ponctuation, alors qu'une proposition est une suite syntagmatique complète qui ne nécessite pas de la ponctuation pour être identifiée. Par conséquent, si une phrase contient deux propositions coordonnées, il est possible de les séparer en deux unités syntagmatiques distinctes. Par exemple, la phrase « il disait d'aller faire une visite à des amis, mais il allait jouer au casino en réalité » peut être considérée comme une seule unité syntagmatique élémentaire, ou encore comme deux unités. Cette typologie de segmentation peut devenir très complexe à cause de la propriété récursive du langage. Les phrases contenant des propositions emboîtées ou interrompues et reprises ensuite ne constituent pas des exceptions rares pour des langues comme le français ou l'anglais. La segmentation en propositions peut être résolue avec des outils avancés de reconnaissance de la structure syntaxique de la phrase.

Dans un contexte d'analyse de texte, l'unité syntagmatique élémentaire peut aussi être un segment composé de plus d'une phrase ou du texte au complet. Dans certains contextes de recherche, la segmentation par mot pôle peut être également pertinente. Cette segmentation consiste à trouver la liste des segments contenant le mot pôle choisi. Cette liste s'appelle *concordance* ou, en anglais, *KWIC* (*KeyWord In Context*). Ce type de segmentation peut être effectué avant ou après la segmentation par phrases, car il est possible d'obtenir une concordance par *fenêtre de mots* ou par *fenêtre de phrases*. Par exemple, il est possible d'établir que l'unité syntagmatique élémentaire de notre concordance est la suite de mots ayant dix mots à gauche et dix mots à droite

du mot pôle; ou encore, il est possible de construire une concordance dans laquelle les unités syntagmatiques élémentaires sont composées de la phrase qui contient le mot pôle, en plus d'une phrase à gauche et une phrase à droite. D'autres types de segmentation existent également, comme la segmentation en paragraphes ou en sections, mais elles ne peuvent être réalisées que si le chercheur dispose préalablement des informations nécessaires à leur identification.

Enfin, la segmentation détermine les unités syntagmatiques élémentaires que le chercheur entend utiliser et qui sont le plus adaptées à son contexte de recherche. Chaque décision prise à ce niveau est donc très importante pour les résultats et pour la configuration des algorithmes à utiliser dans les phases successives de la chaîne de traitement, par exemple dans la construction du modèle vectoriel. Dans notre cas, cette étape a été relativement simple, car nous avons choisi l'article au complet, mais la méthode permet également d'utiliser d'autres types de segmentation. L'important est de demeurer conscient du fait que l'unité syntagmatique élémentaire est l'observation utilisée par les modèles computationnels, ce qui affecte directement les résultats.

6.2.4 Le paradigme distributionnel

Le cœur de notre approche méthodologique réside dans l'utilisation du paradigme vectoriel pour l'analyse de texte. Toutefois, la représentation du texte dans un espace vectoriel n'est qu'un des modèles qui peuvent être utilisés dans l'analyse de texte assistée par ordinateur. Notre choix a été motivé pour plusieurs raisons : la simplicité du modèle; sa diffusion dans la communauté d'analystes du domaine de la fouille de textes (*text mining*) et le fait que le modèle vectoriel partage les mêmes racines linguistiques et épistemologiques que les méthodes et le cadre sémiotique choisis pour ce travail. En effet, le paradigme distributionnel est assurément apparenté au paradigme structuraliste (Rieger, 1997; Sahlgren, 2006, 2008), ce qui permet

d'approfondir les liens qu'ils entretiennent et les implications sémiotiques de l'utilisation du paradigme vectoriel dans notre thèse.

Le lien principal entre ces deux univers apparemment éloignés est le concept de *valeur* (Sahlgren, 2006, 2008). Pour la sémiotique saussurienne, la valeur se définit à l'intérieur d'un *système discret*, dans lequel les éléments qui le constituent s'opposent et, par cette opposition, constituent la valeur de chaque signe. Ce dernier est donc porteur d'une « fraction de sens » qui se construit en opposition à d'autres signes. Cette conception est complémentaire à l'hypothèse distributionnelle. En particulier, le transfert d'un texte dans un modèle vectoriel correspond à la construction d'un *système discret*, exactement comme prévu par Saussure. Un mot prend sa *valeur numérique* (qui est la valeur de la fonction de pondération) en relation avec le système où il est situé. Le chiffre qui correspond à sa valeur numérique change en fonction des mots ou des textes qui sont insérés dans le système. Si un mot ou un texte est ajouté, la valeur numérique de chaque mot change, car dans ce système discret chacune de ses composantes prend une position en opposition aux autres. En d'autres termes, les caractéristiques discrètes de la *valeur linguistique* sont traduites dans une valeur numérique qui est déterminée par ses relations avec les autres éléments présents dans le système.

Cette ressemblance entre la valeur linguistique et la valeur mathématique du modèle vectoriel a été observée par d'autres auteurs, parmi lesquels Sahlgren est probablement le premier (Sahlgren, 2006, 2008). D'autres auteurs ont utilisé la notion de valeur pour souligner le chevauchement entre l'analyse sémiotique et un domaine connexe à la sémantique vectorielle, l'analyse de données (*data mining*) :

> La deuxième ressemblance fondamentale entre l'analyse sémiotique et l'analyse de données réside dans le fait que les deux ne travaillent que sur des différences: tout traitement algorithmique est fondé sur la manipulation

> d'éléments qui n'ont, pour la machine, qu'une valeur différentielle. [...] Finalement, chaque « individu » est un point défini par sa position dans un système de représentation, par des coordonnées sur un espace multidimensionnel où chaque dimension est une variable. À la fin du processus, la substance de départ (la personne, le mot) n'est connue que par sa forme, par les écarts qui permettent de différencier une observation d'une autre. (Compagno, 2017, p. 51)

Ces réflexions font partie d'un débat qui n'est qu'à son début, c'est-à-dire le *débat pour la construction d'une sémiotique computationnelle*. Notre avis sur le chevauchement entre la valeur linguistique et la valeur numérique du modèle vectoriel sera explorée davantage dans le prochain paragraphe, lorsque nous discuterons des caractéristiques spécifiques à la fonction de pondération.

6.2.4.1 Le rôle de la pondération

La *pondération* est la fonction qui assigne une *valeur numérique* à chaque mot pour chaque unité syntagmatique élémentaire. Elle est donc une des fonctions computables les plus importantes de notre méthode. Elle comporte également des implications sémiotiques. Pour explorer les caractéristiques de cette fonction, nous commençons par décrire le contexte de sa mise en œuvre.

La réflexion sur les différentes fonctions de pondération a été surtout développée dans le domaine de la recherche d'informations, pendant les années 1960 et 1970 grâce à l'équipe de Salton. Dans ce cadre, l'introduction de la fonction de pondération Tf-Idf a mené à une augmentation des performances des moteurs de recherche. Le but de la recherche d'informations est de créer une *requête* et de sélectionner les documents qui lui sont les plus similaires. La qualité des documents sélectionnés en fonction d'une requête est généralement évaluée par les indices de *précision* et de *rappel*. La précision mesure la proportion d'information pertinente qui a été sélectionnée par le moteur de recherche. Le rappel mesure la proportion

d'information pertinente qui a été sélectionnée, mais en fonction de la quantité d'information pertinente totale qui est disponible dans la base de données pour la requête. En d'autres termes, la précision se concentre sur la quantité d'information pertinente qui a été *sélectionnée*, alors que le rappel évalue la quantité d'information pertinente qui a été *oubliée*. Un bas taux de précision signifie que, parmi les documents sélectionnés par le moteur de recherche, peu sont pertinents en fonction de la requête. Un bas taux de rappel signifie que le moteur de recherche a « oublié » des documents pertinents et ne les a donc pas sélectionnés.

La pondération joue un rôle important dans les performances d'un moteur de recherche. D'une part, il a été démontré que la fréquence des caractéristiques est déterminante pour améliorer le rappel. Ceci est dû au fait que plus une caractéristique présente dans la requête se retrouve dans les mêmes documents, plus ces documents sont susceptibles d'être sélectionnés. Cependant, la fréquence de caractéristiques ne suffit pas. En effet, si une caractéristique particulière présente dans la requête n'est pas propre à un certain nombre de documents, mais qu'elle est également distribuée dans le corpus, alors le moteur de recherche tendra à sélectionner la majorité des documents du corpus, ce qui réduit énormément le taux de précision. Pour dépasser ces limites, l'équipe de Salton (1988) a introduit la notion de *fréquence documentaire inverse*, afin d'augmenter la possibilité de faire ressortir les *termes discriminants* c'est-à-dire les caractéristiques des documents qui sont les plus pertinentes pour discriminer et distinguer les documents entre eux. Ceci implique que la Tf-Idf est une fonction de pondération *qui met en valeur les différences entre les documents*, car les termes ayant une valeur élevée seront ceux présentant une fréquence élevée et une distribution limitée dans le corpus.

La pondération est aussi très importante dans un contexte d'analyse de texte assistée par ordinateur. Une bonne partie de la puissance du modèle vectoriel dépend de cette

fonction et de plus, elle constitue le cœur de la représentation vectorielle du texte. La pondération tient aussi un rôle de soutien des hypothèses et des postulats liés à la représentation mathématique du « sens ». Dans l'exemple des documents « Doc 1 » et « Doc 2 » de la matrice *U* (cfr. 3.3.2.3), la représentation vectorielle du « sens » de ces phrases est assumée par la fréquence des lemmes qu'ils contiennent. Affirmer que la fréquence des lemmes dans un document est suffisante pour décrire le « sens » du document constitue une simplification qui ne ferait sûrement pas l'unanimité au sein de la communauté des sémiologues ou des linguistes. Toute représentation mathématique du texte mène inévitablement à des *compromis* entre la simplicité du modèle mathématique et la complexité de toute sémiosphère.

Cependant, les hypothèses et les postulats linguistiques en cause lors de la transformation d'un corpus en une matrice selon une fonction de pondération sont complémentaires de la *conception structuraliste du sens*. Plus particulièrement, les hypothèses des fonctions de pondération sont complémentaires de la définition du concept de *valeur* et ceci, surtout si on considère la notion de poids mise en place par la fonction de pondération Tf-Idf. Dans le *Cours de linguistique générale*, Saussure définit le concept de valeur en le distinguant de celui de *signification*, même si les deux termes demeurent complémentaires. La signification est le mouvement du signifiant vers le signifié, donc de l'*image acoustique* vers le *concept*. Par exemple, en anglais le mot « sheep » a la même signification que le mot français *mouton*. Toutefois, le mot « sheep » ne peut pas être utilisé dans des contextes similaires à ceux dans lesquels le mot « mouton » est employé. En effet, pour indiquer une pièce de viande de mouton servie pendant un repas, en anglais on utilise le mot « mutton », alors qu'en français on garde le même terme. Dans ce cas, « sheep » n'a donc pas la même *valeur* que « mouton » car il est remplacé par le mot « mutton » dans certains contextes. La valeur est une notion qui intègre plus directement la langue et le concept de système au contexte syntagmatique de l'usage des mots. Comme on l'a vu,

le mot « mutton » diffère du mot « sheep » et cette différence existe dans le système « langue », ce qui confère une *valeur* différente aux deux mots. Autrement dit, la valeur indique la position spécifique de chaque mot dans le système discret de la langue. Chaque mot organise la pensée d'une communauté linguistique qui « n'est qu'une masse amorphe et indistincte » (Saussure, 1916, p. 120) :

> Philosophes et linguistes se sont toujours accordés à reconnaître que, sans le secours des signes, nous serions incapables de distinguer deux idées d'une façon claire et constante. Prise en elle-même, la pensée est comme une nébuleuse où rien n'est nécessairement délimité. Il n'y a pas d'idées préétablies, et rien n'est distinct avant l'apparition de la langue. (Saussure, 1916, p. 120)

Pour Saussure, la substance de la pensée est un continu amorphe qui peut assumer des *formes* définies seulement par le biais de la langue; ce concept a été repris et complété par Hjelmlev. Pour Saussure, la langue est un système discret de signes qui donne forme à la pensée amorphe. Dans ce cadre, le signe linguistique se met « en condition de signifier » seulement à l'intérieur d'un *système discret* et non analogique, composé d'éléments qui s'opposent entre eux. Saussure propose donc ce type de schéma (figure 6.2) :

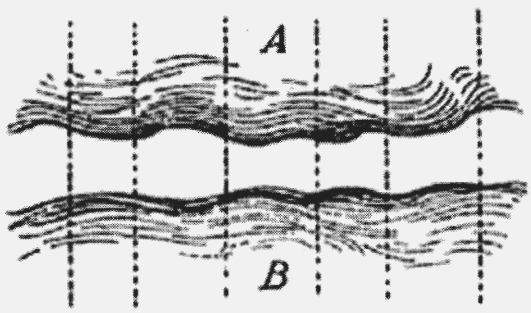


Figure 6.2 L'exemple de la feuille de papier (Saussure, 1916, p. 120)

où (A) illustre les « idées confuses » et (B) représente le flou continu de tous les sons possibles, alors que les lignes verticales sont les mots qui donnent forme à la pensée.

L'image qui permet le mieux de comprendre ce schéma est celle de la *feuille de papier*. La langue est comparée à une feuille où la pensée est le recto et le signe perceptible est le verso. Lorsqu'on coupe le papier, il est impossible de couper le verso sans couper le recto. De la même manière, la langue coupe la pensée amorphe et construit un système discret.

Par sa proposition « dans la langue, il n'y a que des différences », Saussure accorde de l'importance à la position que le signe prend dans le système à l'intérieur du processus de la signification. *Ce sont les différences entre les mots qui donnent la valeur à chaque forme de pensée constituée et perceptible par la langue*. Ainsi, la signification se définit verticalement à travers la relation arbitraire entre signifiant et signifié, alors que le plan horizontal est dominé par les relations discrètes et intrasystémiques entre les signes. Ceci est décrit par un schéma proposé par Saussure lui-même :

Figure 6.3 Illustration de la dépendance intrasystémique de la valeur linguistique (Saussure, 1916, p. 122)

Dans cette illustration, Saussure montre comment la signification s'organise sur deux plans, soit celui du signifiant et du signifié et celui des relations intrasystémiques. La valeur est l'élément qui donne du poids aux relations intrasystémiques et permet l'intégration de ces dernières dans les processus de la signification. En d'autres termes, le processus de signification d'un mot n'est pas limité à la relation arbitraire entre signifiant et signifié, car *sa valeur est établie en évaluant les relations que le*

mot entretient avec les autres. Enfin, pour Saussure, *la valeur d'un mot est le résultat d'une fonction qui permet de le distinguer des autres mots.* En effet, il affirme :

> En outre, l'idée de valeur, ainsi déterminée, nous montre que c'est une grande illusion de considérer un terme simplement comme l'union d'un certain son avec un certain concept. Le définir ainsi, ce serait l'isoler du système dont il fait partie; ce serait croire qu'on peut commencer par les termes et construire le système en en faisant la somme, alors qu'au contraire c'est du tout solidaire qu'il faut partir pour obtenir par analyse les éléments qu'il renferme [...]. Puisque la langue est un système dont tous les termes sont solidaires et où la valeur de l'un ne résulte que de la présence simultanée des autres [...]. (Saussure, 1916, p. 123)

À notre avis, la notion de valeur constitue une base théorique et méthodologique de nature sémiotique pour comprendre la notion de « poids » présente dans la fonction Tf-Idf. En effet, de manière similaire à la notion de valeur chez Saussure, l'objectif de la fonction Tf-Idf est d'intégrer la globalité du système dans lequel un mot (ou lemme) s'insère afin d'évaluer son « poids » dans un document. En assignant un poids à chaque occurrence d'un mot (lemme), la fonction Tf-Idf évalue toujours sa valeur, c'est-à-dire la position qu'il occupe à l'intérieur de l'« univers clos » qui constitue le corpus. Si la valeur linguistique d'un mot se définit à l'intérieur du système-langue, le poids Tf-Idf d'un mot (lemme) est défini à l'intérieur du système-corpus. Le système dans lequel un mot se positionne, qu'il soit général comme la langue ou relatif comme un corpus, ne peut être construit en commençant « par les termes et [...] en en faisant la somme », mais en débutant par le « tout solidaire », c'est-à-dire par les relations entre les éléments qui le constituent.

Il nous apparaît aussi que Rastier va dans le même sens quand il affirme que :

> Ainsi, la valeur n'est pas un signe, mais une relation entre signifiés. Elle exclut une définition atomiste du signe, qui le pourvoirait a priori d'une signification – car une signification est un résultat, non une donnée. Elle interdit la définition

> compositionnelle du sens, puisqu'en tant que principe structural elle établit la détermination du local par le global. Il faut alors admettre que le contenu du signe n'est pas un concept universel, mais un signifié relatif à une langue, voire à un texte et à un corpus (Rastier, 2011, p. 30)

Bien que Rastier ne fasse aucun lien avec la notion de Tf-Idf, sa compréhension du concept de valeur est complémentaire à la nôtre. Plus particulièrement, Rastier souligne que la valeur établit « la détermination du local par le global », ce qui est l'objectif exact de la fonction Tf-Idf. Rastier continue en affirmant que :

> [...] sur l'axe syntagmatique, les classes sont définies par observation des cooccurrences : or, la création de listes de cooccurrences est une des opérations les plus simples et les plus éprouvées que permet la linguistique de corpus. Ainsi le concept de valeur étend-il également son efficacité aux groupements syntagmatiques que définissent les contextes : il importe donc à présent de redéfinir la valeur comme un rapport du texte à ses unités constituantes d'une part, à son corpus d'autre part (Rastier, 2011, p. 30)

Ce dernier extrait met en évidence le rôle du concept de valeur dans un cadre d'analyse propre à la linguistique de corpus. La valeur a un rôle très important dans la détermination du poids que les unités du texte ont en fonction du corpus ce qui implique, selon nous, d'identifier la valeur relative d'un mot à travers le jeu de cooccurrence auquel il participe dans les textes et dans le corpus.

En résumé, nous estimons que la fonction Tf-Idf est une hypothèse « involontaire » pour *approximer mathématiquement* le concept de *valeur* chez Saussure. Comme pour le linguiste, la fonction Tf-Idf évite d'isoler le mot du corpus dont il fait partie et pour ce faire, la fréquence des mots est combinée à la fréquence documentaire inverse. De cette façon et comme la valeur linguistique, le Tf-Idf met en évidence les oppositions et les différences entre les termes avec pour objectif de trouver les seuils discriminants entre les mots. Enfin, comme le dirait Saussure, *dans le corpus il n'y a que différences.*

6.2.4.2 Les spécificités techniques du modèle sémantique vectoriel

Le modèle sémantique rend possible la représentation vectorielle d'un texte. Cette représentation possède des caractéristiques particulières directement dépendantes du phénomène dans lequel ses objets, les textes, s'inscrivent. Ce phénomène est le langage. Les caractéristiques principales du modèle sémantique vectoriel mènent à deux problèmes : le *problème des matrices creuses* et la *malédiction dimensionnelle*.

Le problème des matrices creuses, appelé *sparsity, sparseness* ou *data sparseness problem* en anglais, est un sujet bien connu en fouille de données et il caractérise plusieurs typologies de jeux de données. Dans le contexte de la représentation vectorielle du texte, ce phénomène est dû à deux facteurs intrinsèques au langage. Le premier est la *richesse lexicale*. Il existe un nombre limité d'unités phonologiques élémentaires dans les langues alphabétiques et syllabiques, mais grâce au principe de la compositionnalité (Simone, 2005), elles peuvent s'agencer pour former des centaines de milliers de mots différents. Cette variété lexicale des langues et des textes constitue un véritable défi pour la représentation vectorielle. Tel que présenté dans les paragraphes précédents, *chaque mot qui apparaît dans un document est considéré comme une variable nominale et est transformé en un format computable*. La conversion d'une variable nominale en une variable computable est une notion élémentaire en statistique (Bertrand R., 1986) qui, dans un contexte vectoriel, mène à la définition d'une matrice (de type documents-mots) contenant une colonne pour chaque mot. Ceci implique que l'entièreté du vocabulaire d'un document, par exemple, est utilisée pour former les colonnes de la matrice qui le représente mathématiquement. Si le texte fait partie d'un corpus, la matrice s'agrandira du point de vue des colonnes, car plusieurs autres mots différents seront contenus dans d'autres textes du corpus et du point de vue des lignes, car chaque nouveau document ajouté constituera une nouvelle ligne. Enfin, le vocabulaire entier d'un corpus sera utilisé pour décrire chaque document, ce qui implique que seule une mince partie des

cellules de la matrice sera remplie, la grande majorité étant composée de zéro. Par exemple, chaque phrase, n'utilisant que quelques dizaines de mots parmi les milliers de mots du corpus, n'apporte que peu d'information en relation avec les milliers d'informations possibles. Cette « pénurie » d'information du document conduit à des matrices creuses, c'est-à-dire des matrices contenant surtout des zéros, généralement entre 90 % et 99 %. Par exemple, la phrase « La linguistique est enseignée à l'université », après prétraitement, est représentée ainsi à la ligne « Doc 1 » du tableau 6.2 :

Tableau 6.2 Exemple d'une matrice binaire pour la phrase « La linguistique est enseignée à l'université », suivi d'une deuxième phrase.

Doc/ Mots	**aller**	**apprendre**	**classe**	**enseigner**	**étudiant**	**linguistique**	**philosophie**	**professeur**	**salle**	**sémiotique**	**université**	***n lemme***
Doc 1	0	0	0	1	0	1	0	0	0	0	1	...
Doc 2	1	1	1	0	1	0	1	0	0	0	0	...
n doc	...	...	...	...	...	...	...	...	...	...	...	...

Dans cet exemple, 12 lemmes forment les 12 colonnes de la matrice. Pour la ligne « Doc 1 », trois cellules sont remplies et neuf sont vides. Dans des cas réels, le vocabulaire est plus grand et les matrices atteignent des dizaines de milliers de colonnes qui, pour la majorité, contiennent des zéros. La proportion d'informations pertinentes pour effectuer des analyses est donc très mince.

Le deuxième facteur intrinsèque au langage qui mène aux matrices creuses est la *loi de Zipf*, qui caractérise la distribution des fréquences lexicales d'une langue et qui est liée à un autre principe qui caractérise le langage, soit la *loi du moindre effort* (Simone, 2005). La loi de Zipf est formalisée par la formule mathématique suivante :

$f * r = C$ (cfr. 2.1.2). Cette loi montre que la fréquence d'un mot est inversement proportionnelle à son rang. Par exemple, si, chaque mot a une fréquence *f* et que le mot le plus fréquent possède une valeur de fréquence x, alors les mots successifs auront une valeur de fréquence plus petite, qui se réduit de manière proportionnelle à leur rang. Par exemple, si le mot le plus fréquent a une valeur *f* de 8 000, alors le dixième mot le plus fréquent a une valeur *f* de 800, le centième mot de 80, etc. La loi de Zipf montre que *seulement une partie réduite du vocabulaire d'une langue est très fréquente*, la plus grande partie du vocabulaire étant très peu présente. Ce phénomène conduit également à des matrices creuses et à des variables nominales très rares. Par exemple, le verbe « jouxter » est très peu employé, même dans un corpus littéraire, ce qui implique que la colonne correspondante sera composée à plus de 99.9 % de zéros. La loi de Zipf augmente donc la « pénurie d'information » avec laquelle une représentation vectorielle du texte doit composer.

Le phénomène de la malédiction dimensionnelle est directement lié au problème des matrices creuses et il a des impacts importants dans la construction de l'espace vectoriel. Il s'applique à toute matrice ayant des centaines ou des milliers de dimensions, ce qui est de plus en plus fréquent dans les jeux de données disponibles de nos jours. En effet, les données à grande dimensionnalité sont très fréquentes dans plusieurs disciplines (Giraud, 2015; Masulli *et al.*, 2015), comme en biotechnologie par exemple avec les données sur les séquences d'ADN, en traitement des images, où chacune d'entre elles peut avoir des centaines de milliers de caractéristiques selon sa résolution et le nombre de pixels et en marketing, où des milliers de variables décrivent le comportement des individus, etc. Le texte fait partie de cette liste de domaines, car une collection de textes est représentée par des milliers de dimensions. Tous ces jeux de données ont en commun des caractéristiques liées au phénomène de la malédiction dimensionnelle (Giraud, 2015; Steinbach *et al.*, 2004). Sans entrer dans les aspects mathématiques du phénomène, il est possible d'en synthétiser les

impacts de la manière suivante : dans un espace à grande dimensionnalité, les espaces sont vastes et les points sont « isolés dans cette immensité », ce qui implique que plus des dimensions sont ajoutées, plus le calcul des similarités ou des corrélations entre les entrées du jeu de données analysées devient difficile. Ainsi, chaque caractéristique, variable ou dimension qui est ajoutée mène à une réduction du « pouvoir prédictif ».

Pour décrire le phénomène du point « perdu dans un espace immense », nous nous inspirons de l'exemple décrit dans Bishop (2006). Imaginons que nous disposions d'un jeu de donnée à deux dimensions. Chaque individu peut donc être représenté dans un graphique à deux dimensions comme dans la figure suivante (figure 6.4).

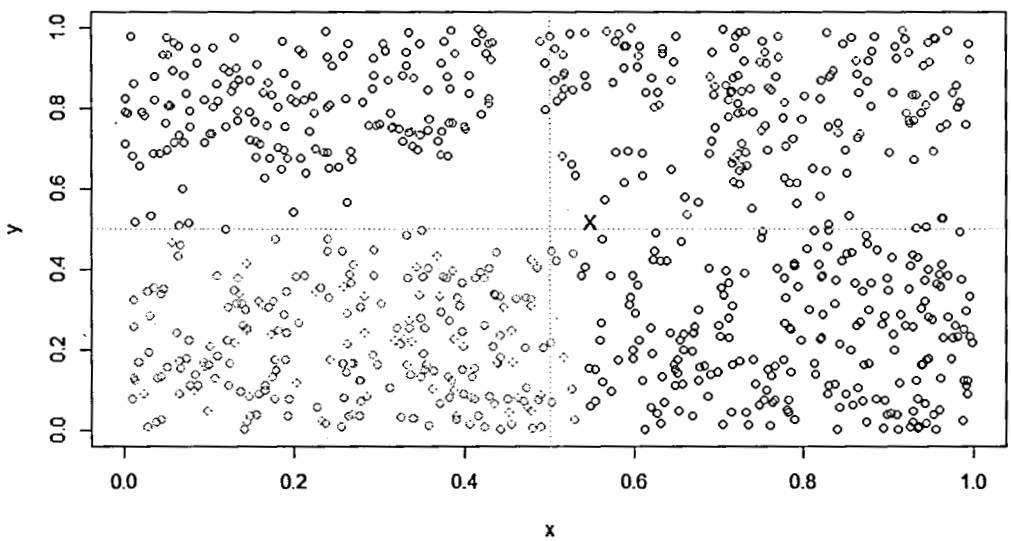


Figure 6.4 Jeu de données à deux dimensions représentées dans un espace plan

Dans cet exemple, il y a environ 800 individus qui sont situés dans l'espace à deux dimensions. Les caractéristiques x et y qui décrivent les individus peuvent avoir une valeur entre 0 et 1. En examinant la figure 6.4, quatre classes d'individus peuvent être identifiées, les individus en bleu ayant la caractéristique y supérieure à 0.5 et la

caractéristique x inférieure à 0.5, les individus rouges, possédant la caractéristique x supérieure à 0.5 et la caractéristique y inférieure à 0.5; et ainsi de suite pour la classe verte et la classe noire. Imaginons que nous voulions utiliser cette classification pour identifier la classe d'un nouvel individu, représenté par le « X » (figure 6.4). La question est donc la suivante : à quelle classe le nouveau point appartient-il? Le point pourrait faire partie de la classe rouge parce qu'il entre dans le rectangle associé à la classe rouge, mais on peut aussi croire qu'il appartient à la classe noire, car le point le plus proche appartient à cette dernière classe. Quelle sera notre stratégie de classification?

Ce genre de questionnement devient encore plus complexe dans un espace à trois dimensions (figure 6.5), car les points sont encore plus isolés. L'œil humain n'étant pas capable de visualiser plus de trois dimensions, chaque espace perçu par l'être humain ne peut donc comporter que trois dimensions.

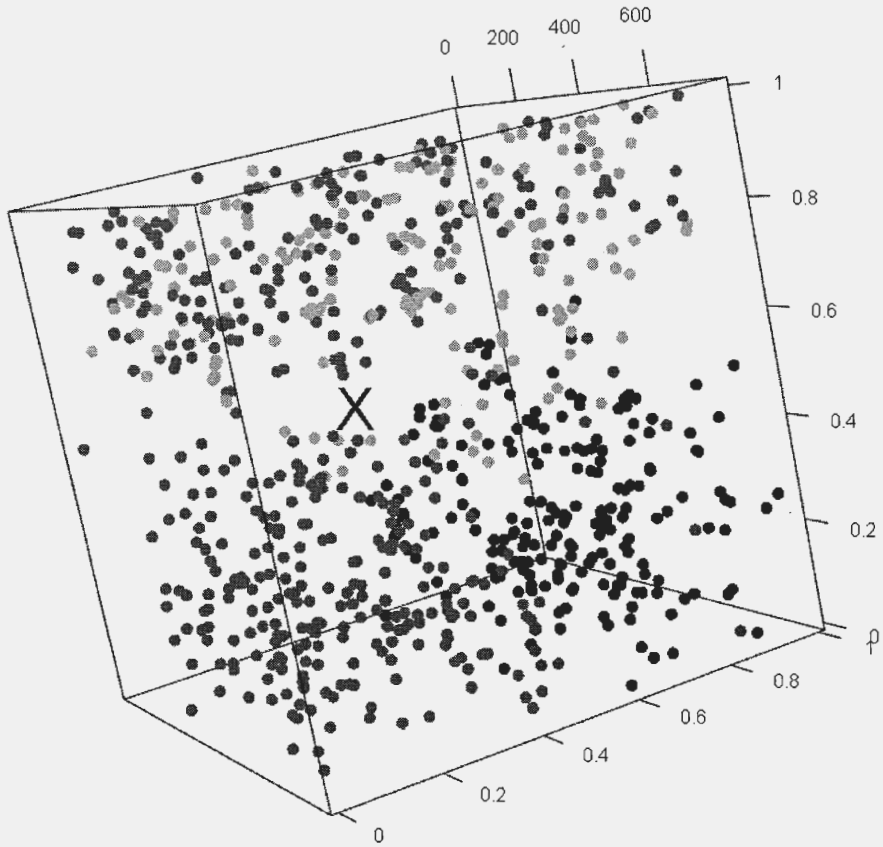


Figure 6.5 Représentation d'un jeu de données en trois dimensions

Or, la représentation vectorielle d'un texte peut comporter des milliers de dimensions, ce qui correspond à des milliers de plans ou à des milliers d'axes à l'intérieur desquels les points se situent. Dans un espace en milliers de dimensions, même les concepts les plus simples de la géométrie euclidienne, comme la distance, deviennent flous et ambigus. La distance entre deux points est effectivement moins précise dans un espace comportant des milliers de dimensions (Aggarwal *et al.*, 2001; Hinneburg *et al.*, 2000). Ceci entraîne donc que la discrimination entre les points les plus proches ou les points les plus loin devient très difficile à évaluer. En augmentant les dimensions d'une matrice, on observe les phénomènes suivants :

1- La distance minimale entre deux points augmente

2- La majorité des points ont une distance similaire et donc la notion de point le plus proche est difficilement utilisable

Pour décrire ces phénomènes, reprenons l'exemple de Giraud (2015), qui comporte quatre histogrammes représentant la distance entre toutes les paires de points (figure 6.6).

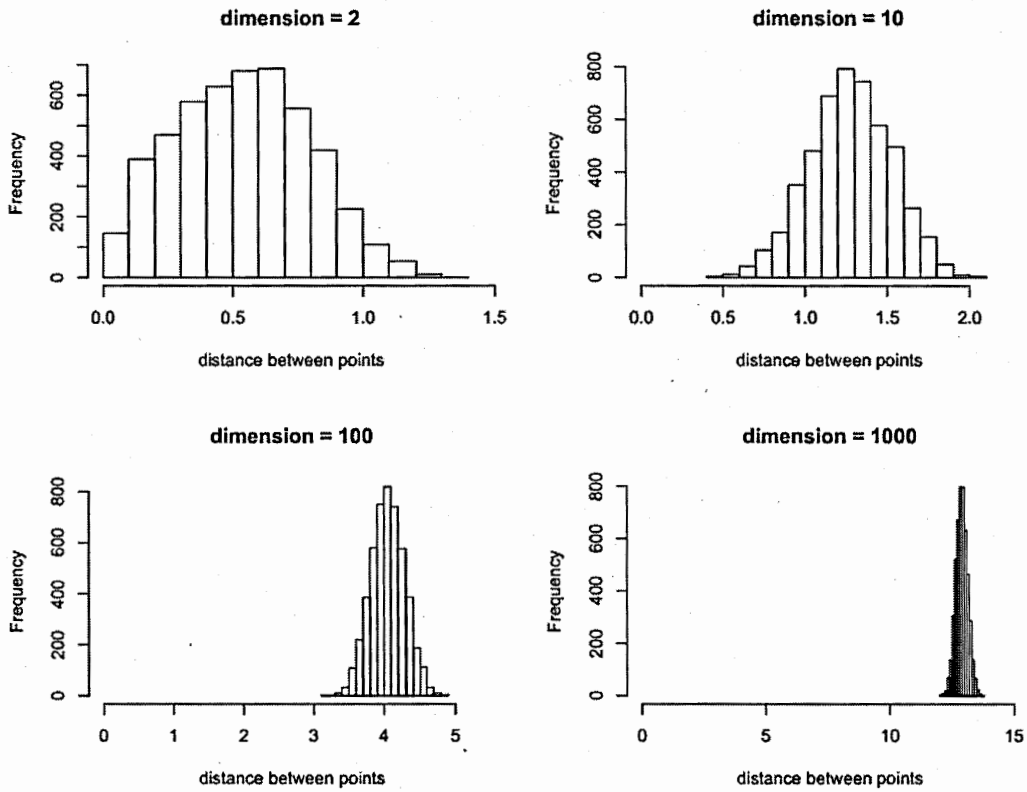


Figure 6.6 Histogrammes des distances de toutes les paires possibles parmi 100 points échantillonnés de manière uniforme. La différence entre les quatre histogrammes est le nombre de dimensions que les points ont dans leur représentation vectorielle (dimension = 2, 10, 100 et 1000) (Giraud, 2015, p. 4)

Dans un espace à deux dimensions, on remarque que beaucoup de paires ont une distance supérieure à 0.5 et qu'un certain nombre de paires sont très proches, la distance étant inférieure à 0.3. En augmentant les dimensions de deux à 10, 100 ou

1000, la distance minimale observée augmente et la majorité des distances observées entre les paires de points sont situées dans une échelle réduite (par exemple, entre 12 et 14 pour un espace à 1 000 dimensions).

Dans un contexte de type textuel, ces considérations peuvent être résumées comme suit :

> the unifying thread that binds together many short context applications and methods is the fact that similarity decisions must be made between contexts that share few (if any) words in common. (Pedersen, 2008, p. 1)

Ce passage souligne donc le fait que les caractéristiques significatives identifiant deux documents similaires sont minimes et que la majorité des documents se situe à une distance difficile à interpréter.

Pour ces raisons, il est préférable de réduire ces impacts néfastes au moyen de techniques de réduction des dimensions, pour lesquelles il existe plusieurs types d'approche. Ce sujet n'a pas été traité dans cette thèse, même s'il aurait mérité un approfondissement. Certaines de ces méthodes sont utilisées également pour améliorer les performances du clustering. En ce sens, ceci représente une limite de notre méthode. Plus particulièrement, il serait utile d'approfondir la connaissance des méthodes de réduction par *lissage* ou basées sur des *classes construites a priori*, mais surtout les méthodes de réduction des dimensions par des techniques comme l'*analyse en composantes principales* (en anglais, *Principal Component Analysis*) et le *positionnement multidimensionnel* (en anglais, *Multidimensional Scaling*). L'utilisation de ces dernières est, par exemple, très répandue pour l'amélioration des résultats de clustering (Aggarwal et Zhai, 2012). Une brève description des concepts liés à ces techniques est esquissée au paragraphe suivant.

6.2.5 Le filtrage des caractéristiques

Le filtrage des caractéristiques est une opération qui peut être exécutée grâce à une variété de méthodes. Elle peut être abordée comme un problème de *sélection des caractéristiques les* plus importantes (en anglais, *features selection*) ou comme un problème de *réduction de la dimensionnalité*. En réalité, ces deux approches s'entremêlent entre elles, car l'opération de sélection des caractéristiques correspond à une réduction des dimensions. Toutefois, elles demeurent très différentes l'une de l'autre. La réduction de la dimensionnalité est une opération accomplie sur des jeux de données ne comportant pas de catégories préalables (*unlabeled data*) et elle peut être exécutée avec une famille spécifique de méthodes, dont font partie la *décomposition en valeurs singulières* (en anglais, *Singular Value Decomposition*), l'*analyse en composantes principales* (en anglais, *Principal Component Analysis*) et le *positionnement multidimensionnel* (en anglais, *Multi-Dimensional Scaling*). Chacune de ces méthodes réduit les dimensions par des calculs matriciels, ce qui consiste à « regrouper » les dimensions en un plus petit nombre par une sorte d'opération de *compactage*. En d'autres termes, ces méthodes « résument » l'information à plusieurs dimensions en un nombre réduit de composantes. Elles peuvent être utilisées pour plusieurs tâches qui vont de la transformation des matrices pour les préparer aux analyses à la visualisation des données. Lors de la réduction dimensionnelle, les composantes de la matrice ne sont plus interprétables comme des caractéristiques, car elles sont transformées en objets de nature différente.

Par contre, le filtrage par sélection des caractéristiques permet de maintenir la nature de chaque dimension telle quelle, ce qui en fait une approche très utilisée lorsque l'interprétation des dimensions restantes est importante. Plusieurs méthodes sont utilisées pour établir quelles sont les caractéristiques les plus signifiantes d'un jeu de données. Certaines sont utilisées avec un jeu de données possédant des catégories définies préalablement (*labeled data*) comme le *khi-deux* ou l'information mutuelle,

ce qui correspond à identifier les spécificités des classes composant un échantillon d'apprentissage. Dans le cas des données non catégorisées, la sélection des caractéristiques est plus complexe et plusieurs typologies d'heuristique et d'algorithmes sont employées.

Dans le cas de cette thèse nous nous sommes limités à sélectionner les caractéristiques par des catégories obtenues grâce à l'annotation morphosyntaxique, tel que décrit dans le chapitre de la méthode (cfr. 3.4.1). Nous avons choisi de ne pas exécuter la réduction dimensionnelle afin de garder des variables signifiantes. Toutefois, d'autres chaînes de traitement, caractérisées par une combinaison des méthodes, pourraient également être envisagées, surtout dans le but d'améliorer les performances du clustering.

6.2.6 Le clustering

Une des questions de recherche de la thèse est la suivante : comment peut-on assister *computationnellement* une analyse sémiotique du texte journalistique avec une approche théorique basée sur la notion de narrativité? Cette question mènerait à un recensement de la panoplie d'outils et d'algorithmes existants, afin d'en identifier les plus pertinents. Toutefois, il serait un projet de recherche trop large. Plus simplement nous avons choisi d'en tester un. Ainsi nous avons identifié dans le *clustering* la possibilité d'assister *computationnellement* l'analyse narrative du texte journalistique.

Cette considération dérive essentiellement de la comparaison entre deux des méthodes non supervisées les plus utilisées pour l'analyse de texte, soit la comparaison entre les *topic modeling* et le *clustering*. Les deux outils exécutent une tâche en apparence similaire, mais qui est en réalité différente. Les topic modeling ont comme but la détection des *distributions de mots* qui caractérisent les documents analysés. Ces distributions sont interprétées comme étant des thèmes. Le clustering,

que décrit dans cette thèse, regroupe des documents qui sont similaires entre eux par rapposrt à leur distribution lexicale. Malgré leur similitude, les deux outils présentent de grandes différences et ce, surtout lorsqu'on utilise un topic modeling de type probabiliste. Dans ce cas, même les modèles mathématiques à la base des outils sont différents.

Toutefois, une certaine confusion demeure dans l'utilisation fréquente du clustering pour trouver de thèmes. En effet, pendant des décennies, les résultats du clustering ont été interprétés avec un modèle sémiotique qui est plus adapté aux outils de topic modeling qu'à ceux de clustering. Ce modèle est essentiellement basé sur des concepts comme ceux d'*isotopie* et de *thème*. Dans ce contexte, le cœur de l'analyse est constitué par la visualisation du vocabulaire le plus fréquent contenu dans les documents que chaque cluster regroupe. Le vocabulaire le plus fréquent est ensuite interprété comme une distribution de mots spécifiques au cluster et enfin, ces distributions sont associées à des thèmes. Le topic modeling exécute des opérations qui sont plus appropriées pour la détection des distributions de mots interprétables comme des thèmes. Ainsi, quand les outils de topic modeling se sont diffusés à partir du début des années 2000 avec l'arrivée des modèles probabilistes de détection automatique des thèmes (par exemple, le *probabilisticLSA* ou la *Latent Dirichlet Allocation*), le clustering est rapidement devenu « une méthode dépassée ».

Toutefois, le modèle sémiotique utilisé pour interpréter les résultats du clustering doit différer de celui adapté à l'analyse du topic modeling. À tout le moins, le modèle sémiotique du thème n'est qu'un des modèles qui peuvent être utilisés pour analyser les résultats du clustering. Toutefois, le clustering exécute des opérations différentes de celles du topic modeling et, au contraire de ce dernier, ses résultats peuvent être interprétés avec des modèles sémiotiques différents.

En effet, le concept de similarité lexicale entre les documents est la base même du clustering et cette question de la similarité lexicale est liée à la similarité sémantique. La découverte du ou des thèmes défini la similarité constituant alors une bonne façon de l'interpréter. Toutefois, la *similarité sémantique peut être interprétée avec d'autres modèles sémiotiques*. Dans cette thèse, nous avons proposé le concept de macrostructure et sa configuration narrative. En d'autres termes, lorsque le clustering regroupe des articles similaires, selon notre proposition, il le fait parce que le lexique et les cooccurrences lexicales reflètent une macrostructure sémantique de type narratif. Les relations entre les mots ne sont donc pas interprétées comme une *distribution associée à un thème*, mais comme un *système complexe de relations narratives*. Il s'agit d'un modèle interprétatif différent de celui du thème pour analyser les résultats du clustering. En se basant sur le concept de similarité sémantique, il est probable que d'autres modèles interprétatifs peuvent être construits ou repris.

Enfin, cette thèse propose un modèle théorique différent pour analyser et interpréter les résultats du clustering. Lorsqu'on veut soutenir l'analyse narrative du texte journalistique par ces outils computationnels, il est nécessaire d'en choisir un qui permet d'obtenir des résultats cohérents avec les présupposés qu'un modèle sémiotique narratif implique. Le topic modeling ne constitue certainement pas un des outils aptes à ce genre d'analyse, puisque le modèle sémiotique le plus adapté est celui de l'isotopie et du thème. Au contraire, le clustering constitue une solution adéquate, puisque le calcul de la similarité sémantique met en relief des relations entre mots qui peuvent être interprétées par une méta-sémiotique de type narratif. Le clustering constitue donc une méthode efficace pour soutenir l'analyse narrative du texte journalistique.

6.2.7 Annotation

L'étape d'annotation est plus que les autres celle qui peut être améliorée et, surtout, *ouvre de nouvelles pistes de recherche pour la sémiotique computationnelle*. Dans le contexte de notre méthode, la phase d'annotation est l'étape que nous permet d'ajouter de l'information de type sémiotique aux résultats obtenus par le processus automatique. Lorsque les articles ont été regroupés par des critères de similarité sémantique, des échantillons de chaque groupe peuvent être sélectionnés et annotés. L'annotation permet ainsi d'évaluer la qualité des groupes d'articles obtenus automatiquement. Lorsque les textes originaux sont repris pour la phase d'annotation, la méthode implique la mise en place d'une théorie et d'une méthode pour effectuer cette tâche, qui peuvent changer selon les hypothèses et des objectifs de la recherche. Le rôle que la sémiotique joue dans ce contexte est très important : *fournir de méthodes pour analyser et interpréter les résultats*. Ceci représente sûrement un des rôles que la sémiotique peut jouer dans un contexte computationnel, c'est-à-dire de la considérer comme une « théorie de l'annotation », tel que déjà proposé :

> Semiotics theory thus has the potential to serve as a general framework for annotation design. This could be achieved through examination of the annotations applied to data consisting of human-created content. [...] To sum up, semiotics could serve as a foundational theory for annotation, and this would reveal the generality of signs underlying human content. (Tanaka-Ishii, 2015, p. 999)

Dans le cadre de cette thèse, le cadre structuraliste et, plus particulièrement, le métalangage d'inspiration greimassienne a été choisi (cfr. 2.2.1). Cette théorie et ses méthodes ont été critiquées à plusieurs reprises au cours des derniers 50 ans (Ablali, 2016). Pour combler certaines de ses lacunes, nous avons ajouté au modèle classique des sous-catégories actantielles (Hébert, 2016). La puissance de ces sous-catégories ressort principalement lors de la détection de l'axe du pouvoir. Dans cet axe, les combinaisons entre ces sous-catégories garantissent l'identification des rôles

actantiels spécifiques qui enrichissent davantage le spectre des relations possibles entre les actants. Comme l'illustre le tableau 2.2, un individu « qui laisse un de ses amis se noyer » (exemple A) est, pour le schéma original, un simple opposant. En réalité, son rôle actantiel est bien plus complexe. Un opposant actif, par exemple, accomplit des actions concrètes afin que la personne ne soit pas sauvée, alors qu'un adjuvant passif est un adjuvant qui n'accomplit aucune action.

Toutefois, l'utilisation de ce métalangage dans un cadre d'analyse du texte journalistique et d'un corpus de grande taille met en relief d'autres limites. En effet, pour certaines typologies de texte, comme les éditoriaux ou d'autres types d'articles d'opinion, plusieurs difficultés apparaissent. Le *saut* entre le plan discursif observé et la complexité de la couche d'annotation qui est impliquée par les structures sémio-narratives rend alors le modèle actantiel *inutilisable*. En d'autres termes, les structures sémio-narratives rendent le plan discursif des articles d'opinion trop complexe, ce qui pose des problèmes d'application du modèle. Lorsque dans un article d'opinion, le journaliste évoque plusieurs programmes narratifs différents, l'annotation de ses structures narratives devient difficile. Une des solutions possibles pour simplifier l'annotation est de mettre en valeur l'acte d'énonciation et ses simulacres, mais ainsi l'analyse tend alors à se limiter sur les intentions communicatives du journaliste ou de l'éditeur. Tel que présenté au point 5.1.7, le seul élément qui nous a permis d'identifier des structures sémio-narratives communes dans les articles d'opinion est le *destinataire*. Mais cet élément est insuffisant pour identifier une macrostructure. Pour la plupart, les journalistes évoquent des acteurs et des plans narratifs différents, lesquels peuvent difficilement être résumés dans un schéma commun. La modélisation qui implique la différenciation entre les programmes narratifs de base et les programmes narratifs d'usage ne simplifie pas la tâche d'annotation, mais, au contraire, la complexifie ce qui est peu pratique lors de l'annotation de plusieurs articles. Ceci représente donc une limite importante dans une tâche d'analyse de

contenu. Enfin, pour l'analyse et l'interprétation d'une grande quantité d'articles d'opinion, l'adoption d'un cadre théorique plus adapté est préférable.

La sémiotique comme « théorie de l'annotation » n'est pas la seule et unique piste de développement de la sémiotique computationnelle. Dans cette thèse, la sémiotique a joué un rôle important à plusieurs reprises, comme dans la constitution des hypothèses de recherche et dans un certain nombre de choix méthodologiques. Par exemple, la notion de *régularités sémiotiques*, est requise pour identifier les hypothèses de recherche, les postulats et les outils computationnels à utiliser. Mais cette thèse n'a pas approfondi la piste de la *modélisation algorithmique et computationnelle sur la base des théories sémiotiques*. Elle pourrait, par exemple, se concrétiser avec *l'intégration de la phase d'annotation dans la panoplie d'outils informatiques utilisés*. En d'autres termes, même lorsqu'elle est partielle, l'*automatisation, de la phase d'annotation* est une des pistes de recherche les plus intéressantes à développer pour la sémiotique computationnelle.

À notre connaissance, il existe au moins trois axes de recherche pour l'exploitation méthodologique et théorique de cet élément. Le premier est de transformer les méta-sémiotiques en systèmes formels et logiques afin d'explorer les possibilités de leur représentation dans des modèles computables. Une voie a déjà été tracée par Galofaro (2013), qui a proposé une formalisation partielle des structures narratives greimassiennes. Dans ce contexte, le modèle formel est construit pour mettre en évidence la propriété récursive du métalangage greimassien (figure 6.7) et les différences avec le modèle de Propp, ce dernier étant défini comme un *automate fini* (*finite-state machine*) (figure 6.8) (Galofaro, 2013).

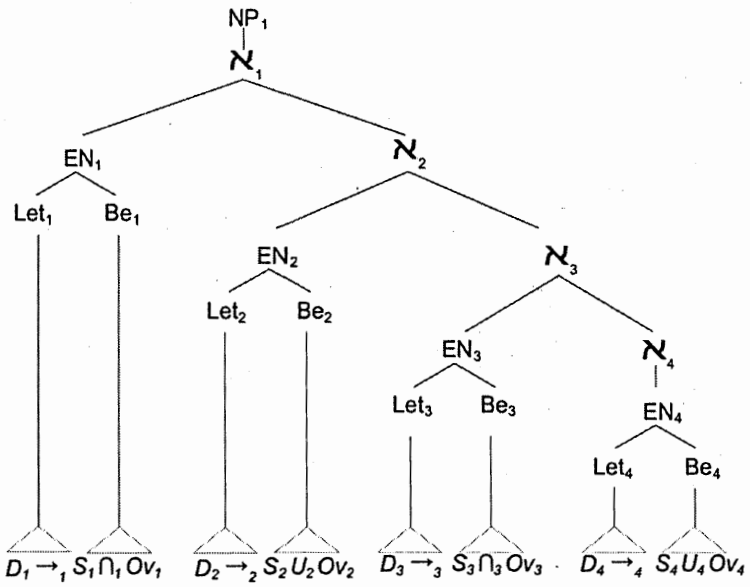
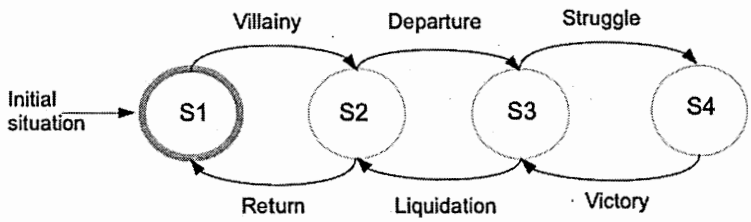


Figure 6.7 Genèse narrative comme une structure récursive (Galofaro, 2013, p. 241)



Figute 6.8 Génération d'une trame narrative simple par un automate fini (Galofaro, 2013, p. 233)

Le travail de Galofaro est un des exemples qui se rapproche le plus du présent travail. Mais, dans un certain sens, il ne constitue que la pointe d'un iceberg. Depuis quelques décennies, des dizaines de travaux en informatique intégrant les concepts de la sémiotique narrative dans des modèles formels (Lakoff et Narayanan, 2010; Zarri, 1997, 2010, 2015) ont été réalisés. Ainsi, la conférence *Computational Model of Narrative* (Mani, 2012), permet d'accéder à de nombreux articles portant sur la représentation formelle des différentes théories de la narration. Ces articles contribuent depuis plus d'une décennie aux progrès théoriques et méthodologiques d'un champ d'études propre à l'informatique, c'est-à-dire les ontologies, mais qui

commence à produire des retombés au sein des humanités numériques (Ciotti, 2016). Par exemple, Lieto et Damiano (2014) utilisent le modèle actantiel pour une représentation des rôles narratifs. Gervás (2013b) se sert de la formalisation de Propp comme un automate fini, avec une implémentation informatique qui permet de produire automatiquement des trames. Dans un de ses travaux les plus anciens, il propose un système de raisonnement par cas (*case-based reasoning*) pour la génération de trames qui se basent sur un ensemble de contes de fées préalablement annotés (Gervás *et al.*, 2005). Ces deux exemples ne constituent qu'une faible part du nombre de travaux existants qui sont pertinents dans le cadre de la sémiotique computationnelle.

Une autre avenue pour l'exploration théorique et méthodologique de l'« automatisation » de la phase d'annotation est constituée par un champ de recherche moins riche que le précédent en termes de travaux existants, mais tout aussi autant important. Il est possible de le résumer par les travaux de Franzosi, lequel a développé des méthodes pour l'analyse quantitative du contenu textuel qui intègrent l'analyse des structures narratives des textes (Franzosi, 2010). Son approche est similaire à celle proposée dans ce travail. Dans ce cadre, plusieurs pistes de recherches peuvent être proposées, en utilisant le traitement automatique du langage naturel pour l'annotation automatique des phrases ou des paragraphes. Par exemple, il existe des ressources comme *FrameNet* qui peuvent être utilisées pour l'annotation automatique des phrases. Cette ressource se base sur la théorie de la sémantique du cadre (*frame semantics*) de Fillmore (Fillmore, 1976) et constitue certainement une des possibilités les plus intéressantes.

La troisième piste est relativement similaire à la précédente, mais elle est basée sur une typologie de travaux différents. Ces travaux font un usage massif des modèles non supervisés de l'apprentissage automatique. Certains d'entre eux ont mené au

développement de méthodes pour inférer automatiquement des chaînes narratives simples. Se basant également sur des outils classiques en traitement automatique du langage naturel, ces travaux exploitent surtout les modèles de Markov et les modèles probabilistes (Chambers et Jurafsky, 2008; Cheung *et al.*, 2013).

CONCLUSION

Le *but de la recherche* est de mettre en place une chaîne de traitement automatisé pour l'analyse sémiotique d'un corpus journalistique et d'en démontrer la faisabilité. En particulier, le premier objectif (cfr.1.6) est d'explorer la technique du clustering pour l'assistance à l'analyse narrative du texte journalistique et le deuxième est de faire une analyse du traitement journalistique du printemps érable. Les résultats de notre recherche répondent aux attentes, puisque ces deux objectifs ont été atteints. D'abord, la thèse a montré comment explorer un corpus journalistique de grande taille par la détection de ses macrostructures les plus récurrentes (question de recherche A) et ceci, à l'aide d'un modèle formel et computable (question de recherche B). Cette exploration a permis d'exécuter une véritable analyse du phénomène à l'étude (objectif A). La phase d'annotation a également permis d'identifier la nature narrative de ces macrostructures. Enfin, puisque plusieurs résultats intéressants sur le traitement journalistique du printemps érable ont été dégagés, la recherche prouve la faisabilité d'une méthode d'analyse narrative assistée par ordinateur basée sur le clustering (objectif B) (cfr. 1.5).

De manière plus précise (cfr. 1.5.1), la méthode a permis d'explorer les hypothèses sémiotiques qui formalisent la nécessité d'identifier des *macrostructures par le biais de schémas lexicaux*. En général, le clustering identifie des groupes d'observations similaires grâce à des régularités détectées dans la distribution des variables qui les décrivent. Dans le cas des données textuelles, le clustering détecte de groupes d'articles similaires grâce aux *régularités lexicales* détectées. Ainsi, l'identification des macrostuctures récurrentes dans les articles de chaque cluster démontre la

pertinence de l'hypothèse sémiotique (hypothèse A). Pour le dire autrement, les clusters regroupent des articles similaires sur la base de leur comportement lexical et l'annotation a permis de valider la similarité sémantique entre les articles. Ceci nous porte à affirmer que, dans un cadre d'analyse de corpus de grande taille, la *régularité lexicale correspond à la régularité sémantique* et *constitue la base pour la similarité sémantique entre documents*.

La sous-hypothèse sémiotique A.1 (cfr. 1.5.1.1) a été explorée dans la phase d'annotation. Pour plusieurs clusters, il a été possible de faire correspondre les schémas récurrents identifiés à des structures narratives. Ceci implique que la régularité lexicale et sémantique détectée peut être interprétée comme une similarité structurale entre documents et que, à son tour, la structure sémantique sous-jacente aux textes comporte une nature narrative. L'identification des schémas actantiels les plus récurrents dans les clusters a permis de confirmer cette sous-hypothèse. Cependant, elle n'a été confirmée que partiellement, puisque les clusters appartenant à la catégorie des articles d'opinion ne respectent pas cette règle de fonctionnement ou du moins ne facilitent pas l'analyse narrative des textes qu'ils contiennent.

L'hypothèse computationnelle est dépendante de celle sémiotique et a ainsi été confirmée partiellement. En effet, il est sans doute vrai que le clustering permet de trouver des macrostructures. L'utilisation de cette technique pour l'exploration d'un corps journalistique est certainement pertinente et devrait être explorée davantage par les sémioticiens. Il est aussi vrai que cette technique, mieux que d'autres (ex. le topic modeling) peut assister l'analyse narrative du texte journalistique (cfr. 6.2.5). Toutefois, lorsqu'elle est appliquée à des textes qui se prêtent difficilement à l'analyse narrative, les résultats ne sont pas interprétables de la même manière et les clusters doivent plutôt être analysés avec un filtre interprétatif différent. Cependant, il est possible d'affirmer que le clustering permet d'identifier les macrostructures

récurrentes qui caractérisent un corpus, quelle qu'en soit la nature (c'est-à-dire, narrative, argumentative, etc.). Plus spécifiquement, il est possible de regrouper des documents en fonction de leur similarité sémantique et de leurs macrostructures sous-jacentes.

La majeure partie de la recherche a été au service du développement de la chaine de traitement et de la deuxième question de recherche, soit comment assister computationnellement une analyse sémiotique de texte et, plus particulièrement une analyse narrative du texte journalistique. Ceci est dû au fait que *le présent travail est de nature méthodologique*. En effet, la thèse illustre une approche pour l'avancement méthodologique de la sémiotique computationnelle. En particulier, elle souligne l'importance de décomposer les méthodes sémiotiques « traditionnelles » en unités séparées, afin qu'on puisse évaluer la possibilité de les convertir en modèles formels et computables (cfr. 2.1.2). Cette phase de décomposition est nécessaire tout au moins pour les projets de la sémiotique computationnelle qui veulent utiliser les outils issus de l'intelligence artificielle. La manière la plus simple pour intégrer ces outils à l'analyse sémiotique est d'identifier des tâches précises et spécifiques que les outils peuvent exécuter. En d'autres termes, les méthodes computationnelles ne peuvent pas être appliquées à la résolution de problèmes complexes et généraux. La sémiotique computationnelle peut fonctionner si l'approche de recherche est basée sur l'identification de problèmes simples et spécifiques.

En ce sens, la sémiotique computationnelle contribue au parcours ouvert par les humanités numériques, qui ont de plus en plus porté l'attention de la communauté scientifique sur le croisement entre les disciplines des sciences dites pures et celles des sciences dites molles. La typologie de travaux méthodologiques de la sémiotique computationnelle est certainement un terrain fertile pour répondre aux questions types des humanités numériques, comme par exemple: « Comment une technologie peut-

elle s'insérer dans une compréhension interprétative de ces objets signifiants? Et à l'inverse : comment des humanités peuvent-elles être numériques? » (Meunier, 2019b). Selon nous, la voie de la sémiotique computationnelle doit passer par une recherche méthodologique appliquée visant au transfert des connaissances de l'intelligence artificielle vers les sciences humaines.

La contribution principale de la thèse est le transfert de connaissances de l'informatique et de l'analyse de données à la sémiotique. Dans ce travail méthodologique, la démonstration de faisabilité d'une méthode d'assistance à l'analyse narrative du texte journalistique fournit des éléments pour évaluer l'intégration d'outils computationnels à l'analyse sémiotique. Les réflexions théoriques et épistémologiques qui constituent l'approche théorique de la thèse (cfr. 2.1), contribuent également à l'avancement de la sémiotique computationnelle, qui constitue à part entière une nouvelle piste de développement pour la sémiotique. Lorsqu'on considère la vocation empirique de la sémiotique (cfr. 1.1.4) et l'avancement technologique du XXIème siècle, l'ouverture à de nouvelles méthodologies ne peut que constituer le déroulement naturel de la recherche sémiotique. Cette thèse contribue ainsi à la compréhension de ce champ qui est la sémiotique computationnelle et, plus particulièrement, de son troisième volet (cfr. 1.1.3), soit celui qui considère l'ordinateur comme un outil au service de l'analyse sémiotique.

La thèse ouvre aussi la porte à plusieurs pistes de recherche. Certaines d'entre elles ont été présentées lors de la discussion. Par exemple, au point 6.2.2, les limites de la phase de lemmatisation ont été soulignées et une théorie sémiotique pour l'amélioration de cette étape a été proposée. Au point 6.2.6, nous avons souligné la différence entre le clustering et d'autres méthodes non supervisées utilisées pour l'analyse de texte, comme le topic model. Les différences structurelles ont aussi été

soulignées en mettant de l'avant le rôle de la sémiotique dans l'interprétation des résultats de ces différents algorithmes. En d'autres termes, une des pistes de recherche possibles pour la sémiotique computationnelle est l'exploration de nouvelles théories ou filtres sémiotiques pour l'interprétation des résultats du clustering ou d'autres méthodes computationnelles. Enfin, au point 6.2.7, plusieurs autres pistes de recherche ont été évoquées en lien avec l'utilisation de l'assistance computationnelle à l'analyse narrative du texte. Dans ce contexte, il a été présenté la perspective qui voit la sémiotique comme théorie de l'annotation et celle qui voit la sémiotique comme un modèle formel cohérent avec l'assistance computationnelle. Ensuite, plusieurs possibilités méthodologiques pour accroître et améliorer l'assistance computationnelle ont été présentées, comme la détection non supervisée des chaînes narratives, l'exploration de ressources comme *FrameNet* ou l'exploration d'algorithmes d'annotation automatique basés sur des ontologies narratives (voir *computational models of narrative*) préalablement construites. L'exploration de ce champ de recherche est sans doute un des plus intéressants pour la sémiotique computationnelle.

Ces premières réflexions conduisent à deux approches générales qui peuvent guider la recherche de la sémiotique computationnelle. L'une possède une nature inférentielle et se base sur le *raisonnement inductif*. Le but de ces recherches serait d'identifier des règles générales de fonctionnement sémiotique à partir des observations tirées de l'analyse d'artefacts sémiotiques assistée par ordinateur. Dans ce contexte, certains outils comme le clustering ou d'autres pourrait être étudiés pour identifier les processus ou systèmes de la *sémiosis*. La deuxième approche générale, au contraire, se baserait sur le raisonnent déductif et utiliserait des méthodes pour cibler des observations répondant à des règles générales de fonctionnement sémiotique. C'est d'ailleurs le procédé logique qui a été adopté dans ce travail. Dans les deux cas, les résultats des recherches pourraient conduire à la confirmation ou à l'infirmation des

postulats ou présupposés sémiotiques, permettant ainsi à la sémiotique de se renouveler.

La nature interdisciplinaire de la sémiotique fait en sorte que cette thèse ouvre de pistes de recherche dans plusieurs autres domaines connexes, parmi lesquels les plus importants sont certainement ceux de la communication, des études des médias, du journalisme et de la science politique. En effet, notre cas d'étude, le printemps érable, est lié à la littérature sur le « protest event analysis », c'est-à-dire l'analyse des événements protestataires. Il existe une grande littérature à ce sujet (Koopmans et Rucht, 2002; Oliver *et al.*, 2003) et plusieurs auteurs se sont aussi penchés sur l'utilisation de méthodes de l'analyse de texte assistée par ordinateur (Wüest *et al.*, 2013). En particulier, cette thèse propose une méthode qui peut soutenir un des volets principaux du champ de recherche du « protest event analysis », soit l'analyse du texte journalistique (Danzger, 1975; Davenport, 2010; Oliver et Myers, 1999). La méthode proposée fournit un outil pour les chercheurs intéressés aux études des mouvements sociaux et à leur couverture journalistique. Évidemment, cette méthode est aussi pertinente dans des champs d'étude plus généraux de la communication et des études de médias, puisqu'elle peut être utilisée avec d'autres typologies de corpus journalistique.

Enfin, la principale piste de recherche qui est renforcée par les résultats de cette thèse est celle de la sémiotique computationnelle et du croisement entre sémiotique et intelligence artificielle. Tel qu'au premier chapitre (cfr. 1.1), il existe déjà plusieurs travaux à ce sujet. Plusieurs auteurs déjà cités au cours de cette thèse (ex. Meunier, Tanaka-Ishii, Rieger, Andersen, De Souza, Figge, Gudwin, Stamper, Zemanek, etc.) sont sans doute de précurseurs de la sémiotique computationnelle. Il y en a plusieurs d'autres qui, même s'ils ne s'inscrivent pas explicitement sous l'étiquette de « sémiotique computationnelle », contribuent également à son avancement (ex.

Rastier). Nous tenons à souligner l'importance de ce croisement entre sémiotique et intelligence artificielle, qui à partir de la fin des années 80, constitue une voie obligée pour la sémiotique. Dans un article publié en 1989 et intitulé « Sémiotique et intelligence artificielle » (Thérien, 1989), l'auteur Gilles Thérien[38], insiste sur la *convergence entre sémiotique et intelligence artificielle* et, plus particulièrement sur le rôle que la sémiotique peut avoir dans le développement de l'intelligence artificielle, car « la sémiotique a inventé des modèles théoriques qui peuvent servir à l'intelligence artificielle ». Un des arguments en soutien à cette convergence provient des relations disciplinaires qui constituent les sciences cognitives, représentées par le populaire hexagone des sciences cognitives. C'est d'ailleurs dans ce contexte que la première connexion entre sémiotique et intelligence artificielle s'est produite puisque « l'hexagone [...] commandité par la fondation Sloan pourrait être facilement modifié en hexagone sémiotique ». Un autre argument en faveur de cette convergence réside dans le concept de représentation. En effet, les deux disciplines partagent un but similaire, soit celui de « décrire le fonctionnement du système, d'en établir les règles et d'interpréter ses effets ». En sémiotique le *signe* est le principal porteur de cette problématique. En intelligence artificielle, *on manipule les signes*. En réalité, ce sont les mêmes problématiques, car la manipulation du signe présuppose le concept de signe et une définition de ce qu'il *re*-présente (Derrida, 1967). Par ailleurs, la sémiotique qui théorise sur le signe, même si elle ne s'interroge pas explicitement sur sa manipulation, en effleure tout de même les dynamiques, car manipulation et signe sont intimement liés dans ce contexte. Enfin, Thérien met en évidence un certain nombre de théories qui sont plus adaptées à cette convergence. Ce sont des théories, concepts ou outils qui ont atteint « un haut degré de formalisation et d'efficacité, [soit] la théorie greimassienne, la grammaire narrative de Van Dijk et Kintsch et les travaux

[38] Gilles Thérien a été professeur titulaire au département d'études littéraires de l'Université du Québec à Montréal (UQAM). Il a participé à la fondation du programme de doctorat en sémiologie à l'UQAM. Il a été élu membre de l'Académie des lettres et des sciences humaines de la Société royale du Canada.

de Gardin en archéologie ». En d'autres termes, il s'agit là de théories qui, en raison de leur niveau de formalisme, s'adaptent mieux au contexte computationnel.

Toutefois, certains des propos et conclusions de Thérien sont à notre avis, discutables. Par exemple, nous croyons que l'auteur se trompe en soutenant « qu'une véritable convergence pourrait s'opérer seulement autour du traitement de l'image ». Le développement du traitement automatique du langage naturel, de la fouille de texte et de l'analyse de texte assistée par ordinateur, ainsi que le travail de cette thèse, montrent exactement l'inverse, c'est-à-dire que c'est par le texte que la sémiotique computationnelle doit commencer son exploration. Ce qui reproduit, à bien y penser, les étapes de la naissance de cette discipline, dérivée de la linguistique. Ceci est surtout vrai en raison des avancements de l'intelligence artificielle dans le domaine et du traitement d'images lequel présentement, ne fournit pas des outils suffisamment performants et efficaces pour que le transfert de connaissances vers les sciences humaines soit justifié. Au-delà des conclusions de Thérien l'importance de la convergence entre sémiotique et intelligence artificielle semble claire. Trente ans après de cette publication, la convergence est devenue un véritable champ de recherche, sous le nom de sémiotique computationnelle.

Comme d'autres disciplines, la sémiotique vit le « défi numérique » et elle doit y faire face. Avec l'avancement technologique, les paradigmes des sciences humaines « sont appelés à se modifier et à se renouveler » (Diminescu et Wieviorka, 2015). La révolution du « big data » et l'approche « data-driven » qu'elle comporte ne peuvent pas demeurer inexplorées pour une discipline comme la sémiotique, qui a une forte vocation empirique. En devenant computationnelle, la sémiotique doit savoir conserver sa *pertinence* et son rôle. *Les sémioticiens ne doivent pas devenir des informaticiens*. De part sa nature interdisciplinaire, la sémiotique a tendance à perdre sa pertinence au profit de problématiques, méthodologies et théories des disciplines

dans lesquelles elle opère. Certains auteurs, par exemple, ont souligné cette tendance dans des recherches en psychanalyse (Beividas, 2016) Dans le domaine de l'analyse de texte assistée par ordinateur, la sémiotique doit maintenir sa pertinence et jouer sur ses spécificités, celles-ci mêmes qui lui ont permis de s'affranchir de la linguistique et de poursuivre, comme le dirait Barthes (Barthes, 1985), cette belle aventure qu'est la sémiotique.

ANNEXE A

DÉTAILS SUR LE CADRE PROBABILISTE À LA BASE DE L'ÉTIQUETAGE MORPHOSYNTAXIQUE

Les *processus de Markov* ont été diffusément utilisés dès les années 80 pour l'annotation automatique des parties du discours dans un contexte supervisé (Church, 1989; Derouault et Jelinek, 1984).

Un processus de Markov est un processus stochastique qui se base sur l'hypothèse que les événements du futur sont exclusivement déterminés par les événements du présent et non par les événements du passé. Cette hypothèse est appelée *propriété de Markov*. Généralement, ces processus sont utilisés dans des séquences temporelles, pour lesquelles on veut connaître la probabilité qu'un événement futur se produise, sachant quel est l'état du présent. Les prévisions météorologiques en sont un exemple. Pour déterminer quelle sera la météo de demain, on prend pour acquis qu'il suffit de connaître les conditions météorologiques d'aujourd'hui. Dans ce cas, on construit un modèle probabiliste permettant de définir la probabilité d'avoir une belle journée ou une mauvaise journée au temps $t+1$, sachant quelles sont les conditions météorologiques du temps t. Sur la base des conditions actuelles, ce modèle établit une probabilité pour chaque état et détermine ainsi l'état le plus probable au temps $t+1$.

Dans le cas de l'étiquetage morphosyntaxique, le processus de Markov est appliqué à des séquences de mots. Dans ce cadre, le but est d'établir la séquence la plus probable d'étiquettes morphosyntaxiques $T = \{t_1, t_2, t_3, \dots, t_n\}$ pour chaque séquence de mots $W = \{w_1, w_2, w_3, \dots, w_n\}$. Par exemple, imaginons de vouloir trouver les étiquettes morphosyntaxiques T de la phrase W : « La diminution paraît, toutefois, moins nette en France et en Italie. » D'un point de vue mathématique, il s'agit de trouver la séquence la plus probable d'étiquettes pour cette phrase, qui peut être représentée avec la formule suivante : $argmax_t P(T|W)$. La première partie de la formule ($argmax_t$) indique la fonction qui détermine la séquence avec la probabilité la plus grande alors que la seconde partie de la formule correspond au théorème de Bayes : $P(T|W) = P(T)\frac{P(W|T)}{P(W)}$. Cette formule peut être simplifiée de nouveau car $P(W)$, qui est la probabilité de la séquence des mots, est constant. On peut donc réécrire la formule comme suit : $argmax_t P(T) \cdot P(T|W)$, c'est-à-dire la probabilité de l'étiquette multipliée par la probabilité de la séquence d'étiquettes connaissant les mots. Décrivons cette séquence encore plus en détail : $P(T) = P(t_1, t_2, t_3, \dots, t_n) = P(t_1) \cdot P(t_2|t_1) \cdot P(t_3|t_1, t_2) \cdot \dots \cdot P(t_n|t_1, t_2, t_3, \dots, t_{n-1})$

Toutefois, un processus de Markov assume la propriété de Markov, ce qui permet d'obtenir des séquences de probabilité moins complexes à calculer. Plus particulièrement, un processus de Markov assume qu'il suffit d'une petite séquence d'étiquettes pour prédire la prochaine étiquette. Cette assomption peut être étendue aux modèles n-gramme. Le *modèle bigramme* assume que, pour prédire la probabilité de la prochaine étiquette, il est suffisant de connaître seulement l'étiquette présente. On assume donc que la probabilité de la prochaine étiquette dépend exclusivement de l'étiquette présente. Dans un modèle de type bigramme, l'équation précédente est transformée comme suit : $P(T) = P(t_1, t_2, t_3, \dots, t_n) = P(t_1) \cdot P(t_2|t_1) \cdot P(t_3|t_2) \cdot \dots \cdot P(t_n|t_{n-1})$. Ainsi, dans le cas de la phrase W, la probabilité que le verbe qui suit

le mot « diminution » soit à la troisième personne est plus grande que celle qu'il soit à la première personne.

D'autres modèles peuvent être élaborés en élargissant l'assomption de Markov à *n-gramme*. Un des modèles le plus utilisé, et qui a donné les meilleurs résultats, est le modèle trigramme. Les modèles n-gramme sont donc construits en calculant la probabilité $P(t_i|t_j)$ sur le corpus annoté. Par exemple, la probabilité d'obtenir l'étiquette t_i (VERBE) en connaissant l'étiquette t_j (NOM) qui le précède est calculée en utilisant les ressources annotées, comme le *French Treebank*. La probabilité $P(VERBE|NOM)$ est établie en calculant les occurrences de l'étiquette VERBE suivie de l'étiquette NOM et en divisant par les totaux des occurrences de l'étiquette VERBE.

Une fois ce calcul obtenu, on peut aussi considérer la probabilité qu'un mot corresponde à une étiquette particulière, ce qui est le deuxième terme de la formule du début, $argmax_t P(T) \cdot P(T|W)$. Pour simplifier, on peut assumer que la correspondance du mot ainsi que l'étiquette soient indépendantes du contexte, et que pour chaque mot on puisse obtenir des probabilités différentes correspondant à chaque étiquette possible. Par exemple, pour le mot « paraît », il existe une probabilité d'obtenir l'étiquette VERBE, une autre l'étiquette NOM, etc. Ces probabilités sont calculées sur les ressources annotées, en faisant la somme de toutes les occurrences d'un mot, par exemple « paraît » correspondant à l'étiquette VERBE, puis en divisant par les occurrences du mot « paraît » dans le corpus. Pour la séquence des mots de la phrase W, on peut écrire l'équation suivante : $P(W|T) = P(w_1|t_1) \cdot P(w_2|t_2) \cdot \ldots \cdot P(t_n|t_n)|t_n)$. Pour les premiers trois mots de la phrase W, il est possible d'établir la séquence d'étiquettes suivante : La|DETdiminution|NOMparaît|VERBE, ce qui correspond à la probabilité $P(DET\ NOM\ VERBE|La\ diminution\ paraît)$. L'estimation de ces probabilités est

calculée ainsi : $P(DET|DÉBUT) \cdot P(NOM|DET) \cdot P(VERBE|NOM) \cdot P(La|DET) \cdot P(diminution|NOM) \cdot P(paraît|VERBE)$.

Dans ce contexte, on peut dire que, pour un modèle bigramme, chaque état correspond à une étiquette morphosyntaxique, et que les *probabilités de transition* d'un état à l'autre sont les probabilités $P(t_i|t_j)$. On peut ajouter que les *probabilités d'émission* d'un symbole sont $P(w_i|t_j)$, ce qui est la probabilité d'un mot d'être émis à partir d'une étiquette morphosyntaxique particulière. Enfin, quelle est donc la probabilité que le mot « paraît » soit étiqueté comme verbe à partir de la séquence d'étiquettes VERBE|NOM? Pour répondre à cette question, il faut évaluer toutes les séquences possibles pour tous les mots. Pour résoudre ce type de tâches et trouver la meilleure séquence d'étiquettes T pour la séquence des mots W, on utilise des algorithmes comme le *Viterbi*.

ANNEXE B

LISTE DES ANNOTATIONS MORPHOSYNTAXIQUES UTILISÉES PAR TREETAGGER

Code annotation	Catégorie morphosyntaxique
ABR	abréviation
ADJ	adjectif
ADV	adverbe
DET:ART	article
DET:POS	adjectif possessif (ma, ta, ...)
INT	interjection
KON	conjonction
NAM	nom propre
NOM	nom commun
NUM	nombre
PRO	pronom
PRO:DEM	pronom démonstratif
PRO:IND	pronom indéfini
PRO:PER	pronom personnel
PRO:POS	pronom possessif (mien, tien, ...)
PRO:REL	pronom relatif
PRP	préposition
PRP:det	préposition plus article (au, du, aux, des)
PUN	ponctuation
PUN:cit	ponctuation de citation
SENT	marqueur de fin de phrase
SYM	symbole
VER:cond	verbe conditionnel
VER:futu	verbe futur
VER:impe	verbe impératif
VER:impf	verbe imparfait
VER:infi	verbe infinitif
VER:pper	verbe participe passé
VER:ppre	verbe participe présent
VER:pres	verbe présent
VER:simp	verbe passé simple
VER:subi	verbe subjonctif imparfait

VER:subp	verbe subjonctif présent

ANNEXE C

DÉTAILS DE L'ALGORITHME DBSCAN

DBscan (*Density-Based Spatial Clustering of Applications with Noise*) est un algorithme de clustering axé sur la densité des points et proposé par Ester *et al.* (1996). La particularité de l'algorithme DBscan est de produire un partitionnement de données en considérant les *zones les plus denses de l'espace*, c'est-à-dire des groupes de données très similaires entre eux et très distincts des autres. Pour ce faire, l'algorithme évalue la densité sur la base de deux paramètres : le premier est ε, soit la distance maximale entre le centroïde et le point le plus loin, et le deuxième est $MinPts$, qui correspond au nombre minimal d'individus qui peuvent être contenus dans un cluster. Pour déterminer ces groupes selon les paramètres de densité, DBscan utilise l*a méthode des k plus proches voisins*. Par exemple, si ε est égal à 0,3 et *MinPts* à 5, la méthode des k plus proches voisins doit trouver au moins cinq individus situés dans un rayon de 0,3 pour former un cluster. Concrètement, ces paramètres constituent une estimation *a priori* de la densité minimale d'un cluster. Une autre caractéristique de l'algorithme est la production d'un *partitionnement partiel*, ce qui implique que les individus qui ne répondent pas aux paramètres de densité sont exclus de la partition et sont réunis dans un même cluster (le cluster « -1 »), qui est considéré comme étant du bruit.

ANNEXE D

DESCRIPTION DE L'IDENTIFIANT UNIQUE DES ARTICLES DU CORPUS

L'identifiant unique de chaque article est composé de 13 caractères alphanumériques. Les premiers deux caractères correspondent à la source d'information; les quatre suivants correspondent à la date (deux pour le mois et deux pour le jour); les trois qui suivent correspondent à la catégorie de l'article; les quatre derniers sont le numéro de série. Dans le chapitre V, les codes sont précédés par le cluster d'appartenance. Le cluster est indiqué avec la lettre « K », suivie par le numéro du cluster.

Voici un exemple, sans le préfixe indiquant le numéro du cluster :

Identifiant unique	**Explication des composantes de l'identifiant**
DE0613Act0124	DE = devoir
	06 = juin
	13 = jour 13
	Act = Actualités (catégorie de l'article)
	0124 = numéro de série

Le même article aura le code K05DE0613Act0124, s'il appartient au cluster numéro 5, ou K13DE0613Act0124, s'il appartient au 13, et ainsi de suite.

Liste des codes utilisés pour identifier les sources d'information :

Code	**Source d'information**

RC	Radio-Canada (site web)
DE	Le Devoir
JQ	Le Journal de Québec
JM	Le Journal de Montréal
ME	Métro
PR	La Presse
SO	Le Soleil

Ensuite, nous présentons la liste exhaustive de toutes les catégories d'articles trouvées pour toutes les sources d'information, qui ont servi à créer l'identifiant unique. Le présent travail n'a utilisé que superficiellement cette métadonnée. Son nettoyage n'était donc pas nécessaire pour la poursuite de la recherche.

Carrières	Enjeux	Politique
Actualités	Environnement	Société
Affaires	Festivals	Spectacles
Arts	Idées	Sports
Arts & Spectacles	La Presse Affaires	Vers rio 2012
Arts Et Spectacles	Le féminisme renouvelé	Vie de stars
Arts HUMOUR	Le samedi	Vie et Société
Arts LECTURES	Les moments «politisateurs»	Vivre
Arts MÉDIAS	Lettres	Votre Argent
Arts THÉÂTRE	Libre opinion	Votre Opinion
Arts TÉLÉVISION	Livres	Votre Vie
Arts VISUELS	Montréal Plus	Week-end
Arts et spectacles	News	Weekend
Arts magazine	Nos régions	no_type
Cahier Spécial	Nouvelles	Économie
Carrière et formation	Opinion	Éditorial
Carrières	Opinions	Éducation
Convergence	Perspectives	Éthique et religions
Culture	Philosophie	
Dimanche	Plan Nord	
Débats	Plus-value	

RÉFÉRENCES

Ablali, D. (2016). La « sémantique de corpus », le programme inachevé de Sémantique structurale. *Semiotica*, *2017*(214), 159–172.

Abonyi, J. et Feil, B. (2007). *Cluster analysis for data mining and system identification*. (s. l.) : Springer Science & Business Media.

Adam, J.-M. (1997). Unités rédactionnelles et genres discursifs: cadre général pour une approche de la presse écrite. *Pratiques*, *94*, 3‑18.

Adam, J.-M. (2011). *La linguistique textuelle : introduction à l'analyse textuelle des discours* (3. éd.). Paris : Colin.

Adam, J.-M. (2016). Pratiques, la linguistique textuelle et l'analyse de discours, dans le contexte des années 70. *Pratiques. Linguistique, littérature, didactique*, (169‑170).

Adam, J.-M. et Heidmann, U. (dir.). (2005). *Sciences du texte et analyse de discours: enjeux d'une interdisciplinité*. (270)Etudes de lettres. Genève : Slaktine Erudition.

Aggarwal, C. C. (2015). *Data mining: the textbook*. New York, NY : Springer.

Aggarwal, C. C. (2016). *Outlier analysis* (2nd edition). New York, NY : Springer Science+Business Media.

Aggarwal, C. C. (2018). *Machine learning for text*. Cham : Springer.

Aggarwal, C. C., Hinneburg, A. et Keim, D. A. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Space. Dans *Database Theory — ICDT 2001* (p. 420-434). Springer, Berlin, Heidelberg.

Aggarwal, C. C. K. et Chandan, R. (2014). *Data Clustering: Algorithms and Applications* (0 éd.)Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Chapman and Hall/CRC.

Aggarwal, C. C. et Zhai, C. (dir.). (2012). *Mining text data*. New York, NY : Springer.

Ahuja, R. K., Magnanti, T. L. et Orlin, J. B. (1993). *Network flows: theory, algorithms, and applications*. Upper Saddle River, N.J. : Prentice-Hall.

Akimoto, T. et Ogata, T. (2012). Macro structure and basic methods in the integrated narrative generation system by introducing narratological knowledge. Dans *2012 IEEE 11th International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC)* (p. 253-262).

Allen, J. (1995). *Natural language understanding* (2nd ed). Redwood City, Calif : Benjamin/Cummings Pub. Co.

Aluísio, S. M., Specia, L., Pardo, T. A., Maziero, E. G. et Fortes, R. P. (2008). Towards brazilian portuguese automatic text simplification systems. Dans *Proceedings of the eighth ACM symposium on Document engineering* (p. 240 -248). ACM.

Amini, M.-R. et Bach, F. (2015). *Apprentissage machine: de la théorie à la pratique*. Paris : Eyrolles.

Andersen, P. B. (1997). *A Theory of Computer Semiotics: Semiotic Approaches to Construction and Assessment of Computer Systems*. Cambridge : Cambridge University Press.

Anderson, S. J., Dewhirst, T. et Ling, P. M. (2006). Every document and picture tells a story: using internal corporate document reviews, semiotics, and content analysis to assess tobacco advertising. *Tobacco Control*, *15*(3), 254-261.

Andsager, J. L. et Powers, A. (1999). Social or economic concerns: How news and women's magazines framed breast cancer in the 1990s. *Journalism & Mass Communication Quarterly*, *76*(3), 531-550.

Arquembourg, J. (2016). L'antibiorésistance dans les médias français, un problème insaisissable (Le Monde, 1948–2014). *Journal des Anti-infectieux*, *18*(1), 5-7.

Arthur, D. et Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. Dans *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (p. 1027 - 1035). Society for Industrial and Applied Mathematics.

Atkins, S., Clear, J. et Ostler, N. (1992). Corpus design criteria. *Literary and linguistic computing*, *7*(1), 1-16.

Atkinson, M., Belayeva, J., Zavarella, V., Piskorski, J., Huttunen, S., Vihavainen, A. et Yangarber, R. (2010). News mining for border security intelligence. Dans *Intelligence and Security Informatics (ISI), 2010 IEEE International Conference on* (p. 173-173). IEEE.

Attewell, P. A. et Monaghan, D. B. (2015). *Data mining for the social sciences: an introduction* (First edition). Oakland, California : University of California Press. Récupéré de Library of Congress

Baeza-Yates, R. et Ribeiro-Neto, B. (1999). *Modern information retrieval* (vol. 463). (s. l.) : ACM press New York.

Baeza-Yates, R. et Ribeiro-Neto, B. (2011). *Modern Information Retrieval: The Concepts and Technology Behind Search* (2nd éd.). USA : Addison-Wesley Publishing Company. Récupéré de ACM Digital Library.

Baker, P. (2006). *Using Corpora in Discourse Analysis*. (s. l.) : A&C Black.

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., van der Goot, E., Halkia, M., ... Belyaeva, J. (2013). Sentiment Analysis in the News. *arXiv:1309.6202 [cs]*. Récupéré de arXiv.org : http://arxiv.org/abs/1309.6202

Barcus, F. E. (1961). A content analysis of trends in Sunday comics, 1900-1959. *Journalism & Mass Communication Quarterly*, *38*(2), 171-180.

Bardin, L. (1977). *L'analyse de contenu.* Paris : PUF.

Barry, A. O. (2002). Les bases théoriques en analyse du discours. Texte de Méthodologie de la Chaire de Recherche du Canada en Mondialisation, Citoyenneté et Démocratie.

Barthes, R. (1966). Introduction à l'analyse structurale des récits. *Communications*, *8*(1), 1-27.

Barthes, R. (1985). *L'aventure sémiologique*. Paris : Ed. du Seuil.

Bastian, M., Heymann, S. et Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. Dans *Third international AAAI conference on weblogs and social media.*

Battaglino, C., Damiano, R. et Lombardo, V. (2014). Moral Values in Narrative Characters: An Experiment in the Generation of Moral Emotions. [Moral Values in Narrative Characters]. Lecture Notes in Computer Science. Dans A. Mitchell, C. Fernández-Vara et D. Thue (dir.), *Interactive Storytelling* (p. 212 -215). Springer International Publishing.

Beividas, W. (2016). La sémioception et le pulsionnel en sémiotique. Pour l'homogénéisation de l'univers thymique. *AS - Actes Sémiotiques*. Récupéré de epublications.unilim.fr : http://epublications.unilim.fr/revues/as/5613

Bell, P. et Milic, M. (2002). Goffman's Gender Advertisements revisited: combining content analysis with semiotic analysis. *Visual Communication*, *1*(2), 203-222.

Benzécri, J. P. (1973). *L'analyse des données*. Paris : Dunod.

Benzécri, J.-P. (1966a). Linguistique et Mathématique. *Revue Philosophique de la France Et de l'Etranger*, *156*, 309–374.

Benzécri, J.-P. (1966b). Linguistique et Mathématique. *Revue Philosophique de la France Et de l'Etranger*, *156*, 309–374.

Benzécri, J.-P. (1981). *Pratique de l'analyse des données, Linguistique et lexicologie*. Paris : Dunod.

Benzécri, J.-P. (1982). *Histoire et préhistoire de l'analyse des données*. Paris : Dunod.

Benzécri, J.-P. et Benzécri, F. (1980). *Analyse des Correspondances: exposé élémentaire*. Paris : Dunod.

Berelson, B. (1952). *Content analysis in communication research*. New York, NY, US : Free Press. [07203].

Berendt, B. (2016). Text Mining for News and Blogs Analysis. Dans C. Sammut et G. I. Webb (dir.), *Encyclopedia of Machine Learning and Data Mining* (p. 1-9). Springer US.

Bernard, M. et Bohet, B. (2017). *Littérométrie: outils numériques pour l'analyse des textes littéraires*. Paris : Presses Sorbonne nouvelle.

Berry, M. W., Dumais, S. T. et O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM review*, *37*(4), 573-595.

Bertrand, D. (2000). Éléments de narrativité. Dans *Précis de sémiotique littéraire*. Nathan.

Bertrand, R. (1986). *Pratique de l'analyse statistique des données*. Montréal : PUQ.

Bharat, K., Curtis, M. et Schmitt, M. (2016). *Method and apparatus for clustering news online content based on content freshness and quality of content source. US9361369 B1*. Récupéré de http://www.google.com/patents/US9361369

Biber, D. (1993). Representativeness in corpus design. *Literary and linguistic computing*, *8*(4), 243-257.

Bilger, M. (2000). *Corpus: méthodologie et applications linguistiques* (vol. 3). Paris : Honoré Champion.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Singapore : Springer.

Bloom, R. L., Obler, L. K., De Santi, S. et Ehrlich, J. S. (1994). *Discourse analysis and applications: Studies in adult clinical populations*. (s. l.) : Lawrence Erlbaum Associates.

Bloomfield, L. (1933). *Language*. New York : H. Holt. Récupéré de Open WorldCat.

Bolasco, S. (2005). Statistica testuale e text mining: alcuni paradigmi applicativi. *Quaderni di Statistica, 7*.

Bommier-Pincemin, B. (1999). *Diffusion ciblée automatique d'informations: conception et mise en oeuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents* (Phd dissertation). Paris : Paris 4.

Bonville, J. de. (2006). *L'analyse de contenu des médias*. De Boeck Supérieur.

Bouillon, P. et Vandooren, F. (1998). *Traitement automatique des langues naturelles*. Louvain-La-Neuve : Aupelf-Uref- Editions Duculot.

Bouras, C. et Tsogkas, V. (2010). W-kmeans: clustering news articles using wordnet. Dans *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (p. 379-388). Springer.

Bouroche, J.-M. et Saporta, G. (1992). *L'analyse des données*. (1854) (5e édition)Que sais-je? Paris : Presses universitaires de France.

Boyd, D. et Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society, 15*(5), 662-679.

Bramer, M. A. (2007). *Principles of data mining*Undergraduate topics in computer science. London : Springer.

Bremond, C. (1966). La logique des possibles narratifs. *Communications*, *8*(1), 60-76.

Brill, E. et Mooney, R. J. (1997). An overview of empirical natural language processing. *AI Magazine*, *18*, 10.

Brown, A. L. et Day, J. D. (1983). Macrorules for summarizing texts: The development of expertise. *Journal of verbal learning and verbal behavior*, *22*(1), 1-14.

Bruner, J. S. (1986). *Actual minds, possible worlds*. Cambridge, Mass. : Harvard University Press.

Bruner, J. S. (1991). Narrative Construction of Reality. *Critical Inquiry*, *18*(1), 1-21.

Bruner, Jerome. (2004). Life as Narrative. *Social Research*, *71*(3), 691-710.

Bruner, Jérôme. (2006). La culture, l'esprit, les récits. *Enfance*, *58*(2), 118-125.

Bullinaria, J. A. et Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, *39*(3), 510-526.

Calabrese, S. (2014). *Neuronarratologia: Il futuro dell'analisi del racconto*. Bologna : ArchetipoLibri.

Campion, B. (2015). Évaluer le récit comme acte cognitif. Quel cadre pour les approches expérimentales ? *Cahiers de Narratologie. Analyse et théorie narratives*, (28).

Cangelosi, A. et Parisi, D. (2002). *Simulating the evolution of language*. London; New York : Springer.

Carley, K. (1990). *Content analysis*. The encyclopedia of language and linguistics. Edinburgh: Pergamon Press.

Carley, K. (1993). Coding Choices for Textual Analysis: A Comparison of Content Analysis and Map Analysis. *Sociological Methodology*, *23*, 75-126.

Chabrol, C. (1973). *Sémiotique narrative et textuelle: Alexandrescu, Barthes, Bremond, Greimas, Maranda, Schmidt, Van Dijk.* Paris : Larousse.

Chambers, N. et Jurafsky, D. (2008). Unsupervised Learning of Narrative Event Chains. Dans *Proceedings of ACL-08, Association for Computational Linguistics*.

Charaudeau, P. (2013). *Les médias et l'information: L'impossible transparence du discours* (Rev. ed). Bruxelles : De Boeck Université.

Charaudeau, P., Maingueneau, D. et Adam, J.-M. (2002). *Dictionnaire d'analyse du discours*. Paris : Seuil.

Chartier, L. (2003). *Mesurer l'insaisissable: méthode d'analyse du discours de presse*. Montréal : Presse de l'université du Québec. [00056].

Chartron, G. (1988). *Analyse des corpus de données textuelles, sondage de flux d'informations* (Thèse de doctorat). France : Université Paris Diderot - Paris 7.

Cheung, J. C. K., Poon, H. et Vanderwende, L. (2013). Probabilistic Frame Induction. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 837-846.

Chomsky, N. (1967). Recent Contributions to the Theory of Innate Ideas. *Synthese*, *17*, 2–11.

Chowdhury, S. G., Routh, S. et Chakrabarti, S. (2014). News analytics and sentiment analysis to predict stock price trends. *Int. J. Comput. Sci. Inform. Technol*, *5*(3), 3595-3604.

Church, K. W. (1989). A stochastic parts program and noun phrase parser for unrestricted text. Dans *International Conference on Acoustics, Speech, and Signal Processing,* (p. 695-698 vol.2).

Cibois, P. (1933). Principe de l'analyse factorielle. *Educational Psychology*, *24*, 417–441.

Ciotti, F. (2016). Toward a Formal Ontology for Narrative. *MATLIT: Materialidades da Literatura*, *4*(1), 29-44.

Clark, A., Fox, C. et Lappin, S. (2010). *The handbook of computational linguistics and natural language processing*. Malden, Ma. : Wiley-Blackwell.

Clear, J. (1992). Corpus sampling. Dans G. Leitner (dir.), *New Directions in English Language Corpora* (p. 21-32). Berlin ; New York : Mouton de Gruyter.

Cléroux, S. (2015). *Le traitement journalistique du « printemps érable » : Comprendre les logiques agissant sur le processus de fabrication de la nouvelle* (Thèse de doctorat). Université d'Ottawa. Récupéré de www.ruor.uottawa.ca : http://www.ruor.uottawa.ca/handle/10393/32105

Compagno, D. (2017). Signifiant et significatif. Réflexions épistémologiques sur la sémiotique et l'analyse des données. *Questions de communication*, (31), 49-70.

Compagno, D. (dir.). (2018). *Quantitative Semiotic Analysis*Lecture Notes in Morphogenesis. Springer.

Cornuéjols, A. et Miclet, L. (2003). *Apprentissage artificiel: concepts et algorithmes* (Première édition). Paris : Eyrolles.

Cotte, D. (2001). De la Une à l'écran, avatars du texte journalistique. *Communication et langages*, *129*(1), 64-78.

CSEP2012. (2014, mars). *Rapport printemps érable* [Rapport, Gouvernement du Québec]. (Commission spéciale d'examen des événements du printemps 2012).

Dacos, M. et Mounier, P. (2015, mars). *Humanités numériques* [Rapport de recherche]. Institut français. Récupéré de HAL Archives Ouvertes : https://hal.archives-ouvertes.fr/hal-01228945

Dale, R., Moisl, H. et Somers, H. L. (2000). *Handbook of natural language processing*. New York : M. Dekker.

Damiano, R., Lombardo, V. et Pizzo, A. (2005). Formal Encoding of Drama Ontology. Lecture Notes in Computer Science. Dans G. Subsol (dir.), *Virtual Storytelling. Using Virtual Reality Technologies for Storytelling* (p. 95-104). Springer Berlin Heidelberg.

Danzger, M. H. (1975). Validating conflict data. *American Sociological Review*, 570-584.

Davenport, C. (2010). *Media Bias, Perspective, and State Repression: The Black Panther Party*. Cambridge : Cambridge University Press. [

De Souza, C. S. (2005). *The semiotic engineering of human-computer interaction*. Cambridge, MA : MIT press.

Delforce, B. (1985). L'objectivité de la presse : les critères du jugement d'objectivité chez le lecteur et les représentations relatives à l'expression. *Études de communication. langages, information, médiations*, (5), 36-69.

Delfosse, P. (1979). *Réformisme et presse ouvrière: histoire et sémiotique*. (1)Dossiers Media. Bruxelles : Bruxelles Labor.

Dennett, D. C. (1991). *Consciousness explained*. Harmondsworth : Penguin.

Derouault, A.-M. et Jelinek, F. (1984). Modèle probabiliste d'un langage en reconnaissance de la parole. *Annales des Télécommunications*, *39*(3), 143-151.

Derrida, J. (1967). *De la grammatologie*. Paris : Éditions de Minuit.

Desrochers, A. (2016). *La neutralité journalistique dans la couverture du conflit étudiant de 2012 : le cas du Montréal Campus* (Mémoire de maîtrise). Montréal : Université du Québec à Montréal.

Diminescu, D. et Wieviorka, M. (2015). Le défi numérique pour les sciences sociales. *Socio. La nouvelle revue des sciences sociales*, (4), 9-17.

Dobre, D. (2013). *Analyse du discours de presse: projet sémiotique*. Bucarest : Editura Universităţii din Bucureşti.

Dreyfus, G. (2008). *Apprentissage statistique*. Paris : Eyrolles.

Drisko, J. et Maschi, T. (2016). *Content analysis*Pocket Guides to Social Work Research methods. Oxford : Oxford University Press.

Drucker, H., Wu, D. et Vapnik, V. N. (1999). Support vector machines for spam categorization. *Neural Networks, IEEE Transactions on*, *10*(5), 1048-1054.

Dubois, J. (dir.). (2002). *Dictionnaire de linguistique*. Paris : Larousse.

Dupuy, J.-P. (2009). *On the Origins of Cognitive Science*. Cambridge : A Bradford Book.

Dussault-Brodeur, M. (2015). *Le caractère politique de la violence contestataire : analyse de la grève étudiante de 2012 au Québec* (Mémoire de maîtrise). Montréal : Université du Québec à Montréal.

EC-AISS. (2018). *Nuove pratiche digitali. La ricerca semiotica alla prova.* (Anno XII-n. 23) Serie speciale della rivista on-line dell'Associazione Italiana di Studi Semiotici. Palermo : EC-AISS.

Eco, U. (1962). *Opera aperta*. Milano : Bompiani.

Eco, U. (1978). Pour une reformulation du concept de signe iconique. *Communications*, *29*(1), 141-191.

Eco, U. (1979). *Lector in fabula: la cooperazione interpretativa nei testi narrativi*. Milano : Bompiani.

Eco, U. (1997). Sulla stampa. *Cinque scritti morali*, 49-79.

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of communication*, *43*(4), 51-58.

Ertel, W. (2011). *Introduction to Artificial Intelligence*Undergraduate Topics in Computer Science. London : Springer London.

Escofier-Cordier, B. (1969). L'analyse factorielle des correspondances. *Cahiers du Bureau universitaire de recherche opérationnelle*, *13*, 25-59.

Ester, M., Kriegel, H.-P., Sander, J. et Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Dans *Kdd-96 Proceedings* (vol. 96, p. 226-231).

Etxeberria, A. et Ibáñez, J. (1999). Semiotics of the artificial: The 'self' of self-reproducing systems in cellular automata. *Semiotica*, *127*(1-4), 295–320.

Everitt, B. S.; Landau, S., Leese, M. et Stahl, D. (2011). *Cluster Analysis*. John Wiley & Sons, Ltd. [04572].

Fabbri, P. (2008). *Le tournant sémiotique* (Y. Jeanneret, trad.). Paris : Hermès Science publications-Lavoisier.

Fabre, C. et Lenci, A. (2015). Sémantique distributionnelle. *Traitement automatique des langues*, *56*(2), 7-23.

Fan, D. P. (1988). *Predictions of Public Opinion from the Mass Media: Computer Content Analysis and Mathematical Modeling*. (s. l.) : Greenwood Publishing Group.

Fararo, T. J. (1993). Generating narrative forms. *The Journal of Mathematical Sociology*, *18*(2-3), 153-181.

Ferraro, G. (2012). Attanti: una teoria in evoluzione. Dans A. M. Lorusso, C. Paolucci et P. Violi (dir.), *Narratività: problemi, analisi, prospettive* (p. 43-60). Bologna : Bononia University Press.

Fetzer, J. H. (2011). Minds and Machines: Limits to Simulations of Thought and Action. *International Journal of Signs and Semiotic Systems*, *1*(1), 39-48.

Fillmore, C. J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, *280*(1), 20-32.

Firth, J. R. (1935). The technique of semantics. *Transactions of the philological society*, *34*(1), 36-73.

Firth, J. R. (1957). *Papers in Linguistics 1934–1951*. London : Oxford University Press.

Flament, C. et Rouquette, M.-L. (2003). *Anatomie des idées ordinaires: comment étudier les représentations sociales*. Paris : A. Colin.

Fontanille, J. (1998). *Sémiotique du discours*. Limoges : Presses Univ. Limoges.

Fontanille, J. (2005). Signes, textes, objets, situations et formes de vie: les niveaux de pertinence sémiotique. *Les objets du quotidien*, 192-203.

Fontanille, J. (2006). Pratiques sémiotiques : immanence et pertinence, efficience et optimisation. *Actes sémoitiques*, (104-106).

Fontanille, J. (2007). Affichages : de la sémiotique des objets à la sémiotique des situations. *AS - Actes Sémiotiques*. Récupéré de epublications.unilim.fr : http://epublications.unilim.fr/revues/as/1113

Fontanille, J. (2008). *Pratiques sémiotiques* (1re éd) Formes sémiotiques. Paris : Presses universitaires de France.

Forest, D. (2006). *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 : Université du Québec à Montréal.

Franzosi, R. (2008). *Content analysis*. 2éd. Thousand Oaks, California : SAGE Publications

Franzosi, R. (2010). *Quantitative narrative analysis* Series: Quantitative Applications in the Social Sciences. Thousand Oaks, California : SAGE Publications.

Franzosi, R. (2011). *Content analysis*. 3éd. London; Thousand Oaks, Calif. : SAGE.

Freeman, J. B. (1991). *Dialectics and the macrostructure of arguments: A theory of argument structure* (vol. 10). Berlin : Walter de Gruyter.

Freeman, J. B. (2011). An Approach to Argument Macrostructure. Argumentation Library. Dans *Argument Structure:* (p. 1-38). Springer Netherlands.

Fuchs, C., Lacheret-Dujour, A. et Victorri, B. (1993). *Linguistique et traitement automatique des langues*. Paris : Hachette.

Fung, G. P. C., Yu, J. X. et Lam, W. (2003). Stock prediction: Integrating text mining approach using real-time news. Dans *Computational Intelligence for Financial Engineering, 2003. Proceedings. 2003 IEEE International Conference on* (p. 395-402). IEEE.

Galofaro, F. (2013). Formalizing Narrative Structures: Glossematics, Generativity, and Transformational Rules. *Signata. Annales des sémiotiques / Annals of Semiotics*, (4), 227-246.

Gamson, W. A. (1989). News as framing: Comments on Graber. *American behavioral scientist*, *33*(2), 157-161.

Gazoni, R. M. (2018). A Semiotic Analysis of Programming Languages. *Journal of Computer and Communications*, *06*(03), 91-101.

Gervás, P. (2013a). Propp's Morphology of the Folk Tale as a Grammar for Generation. Dans *4th Workshop on Computational Models of Narrative (CMN '13)*. M. A. Finlayson, J. C. Meister, & E. G. Bruneau.

Gervás, P. (2013b). Propp's Morphology of the Folk Tale as a Grammar for Generation. Dans *4th Workshop on Computational Models of Narrative* (p. 106–122).

Gervás, P., Díaz-Agudo, B., Peinado, F. et Hervás, R. (2005). Story plot generation based on CBR. *Knowledge-Based Systems*, *18*(4–5), 235‑242.

Giraud, C. (2015). *Introduction to high-dimensional statistics*. (139)Monographs on statistics and applied probability. Boca Raton : CRC Press, Taylor & Francis Group.

Giroux, D. et Charlton, S. (2014). *Les médias et la crise étudiante* [Rapport d'analyse], p. 72. Québec : Centre d'études sur les médias - Université Laval.

Glasgow University Media Group. (1980). *More bad news*. London : Routledge and Kegan Paul. [00000 Great Britain. Television programmes: News programmes (BNB/PRECIS)Includes index.].

Goepfert, E.-M. (2010a). *Médias, politique et vie privée : analyse du phénomène de peopolisation dans la presse écrite française*. Lyon 2.

Goepfert, E.-M. (2010b). Traduction et construction du phénomène de peopolisation. *Signes, Discours et Sociétés: Revue semestrielle en sciences humaines et sociales dédiée à l'analyse des Discours*, (4), 10.

Greimas, A. J. (1966). *Sémantique structurelle: recherche et méthode*. Paris : Larousse.

Greimas, A. J. (1970). *Du sens essais semiotiques*. Paris : Éditions du Seuil.

Greimas, A. J. (1973a). Les actants, les acteurs et les figures. *Sémiotique narrative et textuelle*, 161‑176.

Greimas, A. J. (1973b). Un problème de sémiotique narrative : les objets de valeur. *Langages*, *8*(31), 13-35.

Greimas, A. J. (1979). *Introduction à l'analyse du discours en sciences sociales*. Paris : Hachette.

Greimas, A. J. (1983). *Du sens II: essais sémiotiques*. Paris : Seuil.

Greimas, A. J. et Courtés, J. (1979). *Sémiotique : dictionnaire raisonné de la théorie du langage*. Paris : Hachette.

Grelley, P. (2012). Contrepoint — La méthode expérimentale. *Informations sociales, n° 174*(6), 23-23.

Groupe d'Entrevernes, (1984). *Analyse sémiotique des textes: introduction, théorie, pratique*. Lyon : Presses universitaires de Lyon.

Gudwin, R. R. et Gomide, F. A. C. (1997). Computational semiotics: An approach for the study of intelligent systems-part I: Foundations. *Tenchical Report RT-DCA*.

Hammer, R. et Amey, P. (2009). La sensibilisation au don d'organes dans la presse: récits et expériences vécues. Dans *Les médias et le politique. Actes du colloque «Le français parlé dans les médias», Lausanne* (p. 1-4).

Han, J., Kamber, M. et Pei, J. (2011). *Data mining: concepts and techniques*. (s. l.) : Elsevier

Hariharan, G. (2012). *News Mining Agent for Automated Stock Trading*. (s. l.) : Citeseer.

Harris, Z. (1954). Distributional structure. *Word*, *10*(23), 146–162.

Harris, Z. S. (1952). Discourse Analysis. *Language*, *28*(1), 1 - 30. doi: 10.2307/409987

Harris, Z. S. et Dubois-Charlier, F. (1969). Analyse du discours. *Langages*, 8-45.

Hartigan, J. A. et Wong, M. A. (1979). A K-means clustering algorithm. *Applied Statistics*, *28*, 100–108.

Hastie, T., Tibshirani, R. et Friedman, J. (2013). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* (2e éd.) Springer Series in Statistics. (s. l.) : Springer..

Hébert, È.-L. (2017). *Les effets du traitement médiatique de la grève étudiante québécoise de 2012* (Mémoire de maîtrise). Montréal : Université du Québec à Montréal.

Hébert, L. (2016). Dictionnaire de sémiotique générale. Dans *Signo*. Récupéré de http://www.signosemio.com/documents/dictionnaire-semiotique-generale.pdf

Hennig, C., Meila, M., Murtagh, F. et Rocci, R. (dir.). (2016). *Handbook of Cluster Analysis*. (s. l.) : Chapman & Hall/CRC.

Herman, D. (2000). Narratology as a cognitive science. *Image [&] Narrative*, *1*(1).

Herman, D. (2002). *Story logic: problems and possibilities of narrative*. Lincoln, Neb : University of Nebraska Press

Herman, D. (2003a). How Stories Make Us Smarter. Narrative Theory and Cognitive Semiotics. *Recherches en Communication*, *19*(19), 133-154.

Herman, D. (2003b). *Narrative theory and the cognitive sciences*. (s. l.) : Center for the Study of Language and Inf.

Herman, D. (2009a). Beyond voice and vision: Cognitive grammar and focalization theory. *Point of view, perspective, and focalization: modeling mediation in narrative*, *17*, 119.

Herman, D. (2009b). Narrative ways of worldmaking. *Narratology in the age of cross-disciplinary narrative research*, *20*, 71.

Herman, D. (2011). *Basic Elements of Narrative*. (s. l.) : John Wiley & Sons.

Herman, D. (2013). *Storytelling and the Sciences of Mind.* (s. l.) : MIT press.

Heylen, K. et Bertels, A. (2016). Sémantique distributionnelle en linguistique de corpus. *Langages*, (201), 51-64.

Hinneburg, A., Aggarwal, C. C. et Keim, D. A. (2000). What is the nearest neighbor in high dimensional spaces? Dans *26th Internat. Conference on Very Large Databases* (p. 506-515).

Hou, L., Li, J., Wang, Z., Tang, J., Zhang, P., Yang, R. et Zheng, Q. (2015). NewsMiner: Multifaceted news analysis for event search. *Knowledge-Based Systems*, *76*, 17-29. Hsu, L.-F. (2010). Mining on Terms Extraction from Web News. Dans *ICCCI 2010: Computational Collective Intelligence. Technologies and Applications* (p. 188-194). Springer Berlin Heidelberg.

Huang, A., Milne, D., Frank, E. et Witten, I. H. (2008). Clustering Documents with Active Learning Using Wikipedia. Dans *Eighth IEEE International Conference on Data Mining, 2008. ICDM '08* (p. 839-844).

Huang, R. (2018). *Package « RQDA ».* Récupéré de http://cran.r-project.org/web/packages/RQDA/RQDA.pdf

Hubé, N. (2007). Qu'est-ce que l'actualité «politique»? Pour une analyse de la hiérarchisation de l'information. Regards croisés sur les «Unes» de la presse quotidienne française et allemande. Thèse de doctorat en science politique préparée en co-tutelle sous la direction de Jean-Baptiste Legavre et Nils Diederich, université Robert Schuman Strasbourg/université libre de Berlin, soutenue le 25 novembre 2005. *Trajectoires. Travaux des jeunes chercheurs du CIERA*, (1).

Iglesias, J. A., Tiemblo, A., Ledezma, A. et Sanchis, A. (2015). Web news mining in an evolving framework. *Information Fusion*, *28*, 90-98.

Imbert, G. (1988). *Le discours du journal à propos de" El País": pour une approche socio-sémiotique du discours de la presse* (vol. 35). (s. l.) : Centre national de la recherche scientifique.

Influence Communication. (2012a). *État de la nouvelle. Bilan 2012.* Récupéré de http://www.influencecommunication.com/sites/default/files/bilan-2012-qc.pdf

Influence Communication. (2012b, juillet). *Conflit étudiant – Analyse des premières pages (unes) des quotidiens La Presse, Le Journal de Montréal, Le Devoir et The Gazette 15 février et le 9 juin 2012.* Récupéré de http://www.influencecommunication.com/sites/default/files/Rapport_UNES_%C3%89tudiants_JUILLET_2012.pdf

Influence Communication. (2018). Qui est Influence. Dans *Influence*. Récupéré de http://www.influencecommunication.com/entreprise/qui-est-influence

Ingvaldsen, J. E., Gulla, J. A., Laegreid, T. et Sandal, P. C. (2006). Financial News Mining: Monitoring Continuous Streams of Text. Dans *Web Intelligence* (p. 321-324).

Jacomy, M., Venturini, T., Heymann, S. et Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE*, *9*(6), e98679.

Jacquemin, C. (2001). *Traitement automatique des langues pour la recherche d'information.* Paris : Hermès Science publ.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, *31*(8), 651–666.

James, G., Witten, D., Hastie, T. et Tibshirani, R. (2013). *An introduction to statistical learning.* (s. l.) : Springer.

Jamet, C. et Jannet, A.-M. (1999). *La mise en scène de l'information.* Paris : Editions L'Harmattan.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Harvard : Harvard University Press.

Juanzi, L., Jun, L. et Jie, T. (2009). A flexible topic-driven framework for news exploration. Dans *Proceeding of Conference of Knowledge Discover and Data Ming*.

Kannan, R., Woo, H., Aggarwal, C. C. et Park, H. (2017). Outlier Detection for Text Data : An Extended Version. *arXiv:1701.01325 [cs, stat]*. Récupéré de arXiv.org : http://arxiv.org/abs/1701.01325

Kantardzic, M. (2011). *Data mining: concepts, models, methods, and algorithms* (2nd ed). Hoboken, N.J : John Wiley : IEEE Press.

Kao, A. et Poteet, S. R. (2007). *Natural Language Processing and Text Mining*. (s. l.) : Springer Science & Business Media.

Kassambara, M. A. (2017). *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning* (1 edition). Erscheinungsort nicht ermittelbar : CreateSpace Independent Publishing Platform.

Kaufman, L. et Rousseeuw, P. J. (2005). *Finding groups in data: an introduction to cluster analysis* (vol. 344). (s. l.) : John Wiley & Sons.

Kervella, A. (2008). *Les discours de la presse française sur le terrorisme perpétré dans le cadre du conflit israélo-palestinien et du conflit tchétchène, face à la « guerre contre le terrorisme »*. Lyon 3.

Ketner, K. L. (1988). Peirce and Turing: Comparisons and conjectures. *Semiotica*, *68*(1-2), 33–62.

Kintsch, W. (2002). On the notions of theme and topic in psychological process models of text comprehension. Dans M. Louwerse et W. van Peer (dir.), *Thematics : Interdisciplinary Studies* (p. 157-170). Amsterdam/Philadelphia : John Benjamins Publishing Company.

Kintsch, W. et Van Dijk, T. A. (1975). Comment on se rappelle et on résume des histoires. *Langages*, (40), 98-116.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 2053951714528481.

Klinkenberg, J.-M. (2000). *Précis de sémiotique générale*. Paris : Seuil.

Koopmans, R. et Rucht, D. (2002). Protest event analysis. *Methods of social movement research*, *16*, 231-59.

Kordjazi, Z. (2012). Images matter: A semiological content analysis of gender positioning in contemporary English-learning software applications. *Novitas-ROYAL (Research on Youth and Language)*, *6*(1), 59-80.

Kraft, D. H. et Colvin, E. (2017). *Fuzzy Information Retrieval*. Morgan & Claypool.

Krieg, A. (2001). Emergence et emplois de la formule «purification ethnique» dans la presse française (1980-1994): Une analyse de discours. *L'Information Grammaticale*, *91*(1), 41-43.

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, California : SAGE Publications.

Lakoff, G. (2008). *The political mind: why you can't understand 21st-century politics with an 18th-century brain*. (s. l.) : Penguin.

Lakoff, G. et Narayanan, S. (2010). Toward a Computational Model of Narrative. Dans *2010 AAAI Fall Symposium Series*. Récupéré de www.aaai.org : https://www.aaai.org/ocs/index.php/FSS/FSS10/paper/view/2323

Lambert, F. (1986). *Mythographies: la photo de presse et ses légendes* (vol. 12). (s. l.) : Edilig.

Lan, M., Tan, C. L., Su, J. et Lu, Y. (2009). Supervised and traditional term weighting methods for automatic text categorization. *IEEE transactions on pattern analysis and machine intelligence*, *31*(4), 721-735.

Lapalut, S. (1995). Text clustering to support knowledge acquisition from documents. RR- 2639, INRIA.

Larose, D. T. (2006). *Data mining methods and models*. Hoboken, NJ : Wiley-Interscience.

Le, E. (2000). Pour une analyse critique du discours dans l'étude des relations internationales. Exemple d'application à des éditoriaux américains sur la guerre en Tchétchénie. *Études internationales*, *31*(3), 489-515.

Le Priol, F., Desclès, J.-P. et in Le Priol et Desclès. (2009). *Annotations automatiques et recherche d'information*. Paris : Hermès science publ. : Lavoisier.

Lebart, L., Piron, M. et Steiner, J.-F. (2003). *La sémiométrie: essai de statistique structurale*. Paris : Dunod.

Lebart, L. et Salem, A. (1994a). L'analyse des correspondances des tableaux lexicaux. Dans *Statistique textuelle* (p. 79-97). Paris : Dunod.

Lebart, L. et Salem, A. (1994b). *Statistique textuelle*. Paris : Dunod.

Lebreton, C. (2008). *Analyse sociologique de la presse québécoise pour adolescentes (2005/2006): entre hypersexualisation et consommation* (Mémoire). Montréal : Université du Québec à Montréal.

Lehmann, A. et Martin-Berthet, F. (2014). *Lexicologie: sémantique, morphologie, lexicographie* (4.éd)Collection Cursus Lettres. Paris : Colin.

Leiss, W., Kline, S. et Jhally, S. (1990). *Social communication in advertising: persons, products & images of well-being*. Scarborough, Ont.; New York : Nelson Canada ; Routledge. [01589].

Lemaire, B., Mandin, S., Dessus, P. et Denhière, G. (2005). Computational cognitive models of summarization assessment skills. Dans *Proceedings of the 27th Annual Meeting of the Cognitive Science Society (CogSci'2005)* (p. 1266-1271).

Léon, J. (2008). Théorie de l'information, information et linguistes français dans les années 1960. Un exemple de transfert entre mathématiques et sciences du

langage. Dans *Congrès Mondial de Linguistique Française* (p. 097). EDP Sciences.

Lieto, A. et Damiano, R. (2014). A Hybrid Representational Proposal for Narrative Concepts: A Case Study on Character Roles. Dans *5th Workshop on Computational Models of Narrative* (p. 106). M. A. Finlayson, J. C. Meister, & E. G. Bruneau.

Linoff, G. S. et Berry, M. J. A. (2011). *Data mining techniques: for marketing, sales, and customer relationship management* (3rd ed). Indianapolis, IN : Wiley Pub. [02353].

Liu, K. (2000). *Semiotics in information systems engineering*. Cambridge, UK : Cambridge University Press.

Liu, Y., Li, Z., Xiong, H., Gao, X. et Wu, J. (2010). Understanding of internal clustering validation measures. Dans *Data Mining (ICDM), 2010 IEEE 10th International Conference on* (p. 911-916). IEEE.

Lombardo, V. et Pizzo, A. (2014). Ontology–based visualization of characters' intentions. Dans *International Conference on Interactive Digital Storytelling* (p. 176-187). Springer.

Lorusso, A. M. et Violi, P. (2004). *Semiotica del testo giornalistico* (vol. 368). (s. l.) : Laterza.

Loula, A., Gudwin, R. et Queiroz, J. (dir.). (2007). *Artificial cognition systems*. Hershey, PA : Idea Group Pub.

Ma, C., Gan, G. et Wu, J. (2007). *Data Clustering: Theory, Algorithms, and Applications*ASA-SIAM Series on Statistics and Applied Probability. SIAM, Society for Industrial and Applied Mathematics. Récupéré de http://gen.lib.rus.ec/book/index.php?md5=A3F903E72B3FE79925019A6E747E2F31

Maarek, P. J. (2014). *Présidentielle 2012: une communication politique bien singulière*. Paris : Editions L'Harmattan.

Magli, P. et Pozzato, M. P. (1985). La grammatica narrativa di Greimas. Dans A. J. Greimas, *Del senso 2: narrativa, modalità, passioni*. Milano : Bompiani.

Mahajan, A., Dey, L. et Haque, S. M. (2008). Mining financial news for major events and their impacts on the market. Dans *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on* (vol. 1, p. 423-426). IEEE.

Maingueneau, D. (1979). L'analyse du discours. *Repères pour la rénovation de l'enseignement du français à l'école élémentaire*, *51*(1), 3-27.

Maingueneau, D. (1997). *L'analyse du discours*. (s. l.) : Hachette Supérieur.

Malhotra, S. et Dixit, A. (2013). An effective approach for news article summarization. *International Journal of Computer Applications*, *76*(16).

Mandler, J. M. (1984). *Scripts, stories and scenes: Aspects of schema theory* (Lawrence Erlbaum Associate). Hillsdale (NJ) : Psychology Press.

Mani, I. (2012). *Computational Modeling of Narrative*. San Rafael : Morgan & Claypool Publishers.

Manning, C. D., Raghavan, P. et Schütze, H. (2009). *Introduction to information retrieval* (Online edition). Cambridge, UK : Cambridge University Press.

Manning, C. D. et Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA : MIT press

Manning, P. K. (2004). Semiotics and Data Analysis. Dans M. Hardy et A. Bryman, *Handbook of Data Analysis* (p. 566-587). London, UK : SAGE Publications, Ltd.

Marcus, M. P., Marcinkiewicz, M. A. et Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, *19*(2), 313-330.

Markie, P. (2017). Rationalism vs. Empiricism. Dans E. N. Zalta (dir.), *The Stanford Encyclopedia of Philosophy* (Fall 2017). Metaphysics Research Lab, Stanford University. Récupéré de Stanford Encyclopedia of Philosophy : https://plato.stanford.edu/archives/fall2017/entries/rationalism-empiricism/

Marrone, G. (2001). *Corpi sociali: processi comunicativi e semiotica del testo*. Torino : Einaudi.

Martin, O. (2009). *L'analyse de données quantitatives*. Paris : A. Colin.

Martinet, J. (1975). *Clefs Pour La Semiologie*. Paris : Seghers.

Masulli, F., Petrosino, A. et Rovetta, S. (dir.). (2015). *Clustering High--Dimensional Data: First International Workshop, CHDD 2012*. Naples, IT : Springer.

Mayaffre, D. (2002). Les corpus réflexifs : entre architextualité et hypertextualité. *Corpus*, (1).

McAdams, D. P. et McLean, K. C. (2013). Narrative Identity. *Current Directions in Psychological Science*, Sage CA: Los Angeles, CA.

McKeown, K. et Radev, D. R. (1995). Generating summaries of multiple news articles. Dans *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval* (p. 74-82). ACM.

McQuarrie, E. F. et Mick, D. G. (1992). On Resonance: A Critical Pluralistic Inquiry into Advertising Rhetoric. *Journal of Consumer Research*, *19*(2), 180-197.

Meunier, Jean-Guy. (1989). Artificial intelligence and sign theory. *Semiotica*, *77*(1-3), 43-64.

Meunier, Jean-Guy. (2009). CARAT–Computer-Assisted Reading and Analysis of Texts: The Appropriation of a Technology. *Digital Studies / Le champ numérique*, *1*(3).

Meunier, Jean-Guy. (2014). Humanités numériques ou computationnelles: Enjeux herméneutiques. *Sens-Public*.

Meunier, Jean-Guy. (2017a). Humanités numériques et modélisation scientifique. *Questions de communication*, *31*(à paraître).

Meunier, Jean-Guy. (2017b). Sémiotique et computation. *Applied Semiotics / semiotique appliquée*.

Meunier, Jean-Guy. (2019a). La rencontre du sémiotique et du « numérique » : Le rôle d'une modélisation conceptuelle. *Semiotica*, (à paraitre).

Meunier, Jean-Guy. (2019b). Le paradoxe des humanités numériques. *Quaderni*, *n° 98*(1), 19-31.

Meunier, Jean-Guy et Forest, D. (2009). Lecture et analyse conceptuelle assistée par ordinateur: premières expériences. Dans J.-P. Desclés et F. Le Priol (dir.), *Annotation automatique et recherche d'informations*. Paris : Hermes - Lavoisier.

Meunier, J.-G. (2018). Vers une sémiotique computationelle? *Applied Semiotics / Semiotique appliquée*, (26), 75-107.

Meyer, D., Hornik, K. et Feinerer, I. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, *25*(5), 1-54.

Minsky, M. (1988). *Society of mind*. (s. l.) : Simon and Schuster.

Mitchell, T. M. (1997). *Machine Learning*. New York : McGraw-Hill.

Mittermayer, M.-A. et Knolmayer, G. (2006). *Text mining systems for market response to news: A survey*. Bern : Institut für Wirtschaftsinformatik der Universität Bern.

Moisl, H. (2015). *Cluster analysis for corpus linguistics*. (66) Quantitative linguistics. (s. l.) : De Gruyter Mouton..

Moretti, F. (2013). *Distant reading*. New York : Verso Books.

Mounin, G. (2004). *Dictionnaire de la linguistique* (4. éd)Quadrige. Paris : PUF.

Nadin, M. (1977). Sign and fuzzy automata. *Semiosis*, *1*(5), 19-26.

Nadin, M. (2011). Information and Semiotic Processes The Semiotics of Computation. *Cybernetics & Human Knowing*, *18*(1-2), 153-175.

Nazarenko, A., Habert, B. et Salem, A. (1997). *Les linguistiques de corpus*. Armand Colin. Récupéré de HAL Archives Ouvertes : http://hal.archives-ouvertes.fr/hal-00619268

Née, É. (dir.). (2017). *Méthodes et outils informatiques pour l'analyse des discours*Didact Méthode. Rennes : Presses Universitaires de Rennes.

Neuendorf, K. A. (2002). *The Content Analysis Guidebook*. Thousand Oaks, California : SAGE Publications.

Neveu, F. (2004). *Dictionnaire des sciences du langage*. Paris : Armand Colin.

Nöth, W. (1995). *Handbook of Semiotics*. Bloomington : Indiana University Press.

Oliver, P. E., Cadena-Roa, J. et Strawn, K. D. (2003). Emerging trends in the study of protest and social movements. *Research in political sociology*, *12*(1), 213-244.

Oliver, P. E. et Myers, D. J. (1999). How events enter the public sphere: Conflict, location, and sponsorship in local newspaper coverage of public events. *American journal of sociology*, *105*(1), 38-87.

Patry, R. et Nespoulous, J.-L. (1990). Discourse Analysis in Linguistics: Historical and Theoretical Background. [Discourse Analysis in Linguistics]. Springer Series in Neuropsychology. Dans Y. Joanette et H. H. Brownell (dir.), *Discourse Ability and Brain Damage* (p. 3-27). Springer New York.

Pêcheux, M. (1969). *Analyse automatique du discours*. Paris : Dunod.

Pedersen, T. (2008). Computational Approaches to Measuring the Similarity of Short Contexts : A Review of Applications and Methods. *arXiv:0806.3787 [cs]*. Récupéré de arXiv.org : http://arxiv.org/abs/0806.3787

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peirce, C. S. (1994). *The Collected Papers of Charles Sanders Peirce* (Electronic Edition). Virginia, U.S.A. : InteLex Corp. Charlottesville.

Peldszus, A. et Stede, M. (2013). From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, *7*(1), 1-31.

Pessoa de Barros, D. L. (1983). L'isotopie discursive. Papiere zur Textlinguistik, Papers in textlinguistics. Dans F. Neubauer (dir.), *Coherence in Natural-Language Texts* (p. 115-133). Hamburg : H. Buske.

Piskorski, J. et Atkinson, M. (2011). Frontex real-time news event extraction framework. Dans *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (p. 749-752). ACM.

Polguère, A. (2016). *Lexicologie et sémantique lexicale: notions fondamentales* (Troisième édition)Paramètres. Montréal, Québec : Les Presses de l'Université de Montréal.

Porteous, J. et Cavazza, M. (2009). Controlling Narrative Generation with Planning Trajectories: The Role of Constraints. [Controlling Narrative Generation with Planning Trajectories]. Lecture Notes in Computer Science. Dans I. A. Iurgel, N. Zagalo et P. Petta (dir.), *Interactive Storytelling* (p. 234-245). Springer Berlin Heidelberg.

Porteous, J., Cavazza, M. et Charles, F. (2010). Narrative Generation Through Characters' Point of View. Dans *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 - Volume 1* (p. 1297–1304). Richland, SC : International Foundation for Autonomous Agents and Multiagent Systems.

Porter, M. F. (2001). *Snowball: A language for stemming algorithms*. Récupéré de http://snowball.tartarus.org/texts/introduction.html

Poudat, C. et Landragin, F. (2017). *Explorer un corpus textuel: méthodes, pratiques, outils*. Paris : De Boeck supérieur.

Pozzato, M. P. (2005). *Leader, oracoli, assassini: analisi semiotica dell'informazione*. Roma : Carocci.

Que peut le métalangage ? (2013). *Signata. Annales des sémiotiques / Annals of Semiotics*, (4), 7-9.

Queiroz, J. et Merrell, F. (2008). On Peirce's Pragmatic Notion of Semiosis—A Contribution for the Design of Meaning Machines. *Minds and Machines*, *19*(1), 129-143.

Raber, D. et Budd, J. M. (2003). Information as sign: semiotics and information science. *Journal of Documentation*, *59*(5), 507-522.

Rapaport, W. J. (2012). Semiotic Systems, Computers, and the Mind: How Cognition Could Be Computing. *Int. J. Signs Semiot. Syst.*, *2*(1), 32–71.

Rastier, François. (1972). La systématique des isotopies. Dans A. J Greimas (dir.), *Essais de sémiotique poétique*. Paris : Larousse.

Rastier, François. (1995). La sémantique des thèmes - ou le voyage sentimental. Dans François Rastier (dir.), *L'analyse thématique des données textuelles* (p. 223-249). Paris : Didier. L

Rastier, François. (1996). La sémantique des textes: concepts et applications. *Hermès*, *16*, 15-37.

Rastier, François. (1997). Les fondations de la sémiotique et le problème du texte. Questions sur les Prolégomènes. Semiotic and cognitive studies. Dans A. Zinna (dir.), *Hjelmslev aujourd'hui* (p. 141-164). Turnhout : Brepols.

Rastier, François. (2001a). *Arts et sciences du texte* (1re éd)Formes sémiotiques. Paris : Presses universitaires de France.

Rastier, François. (2001b). Sémiotique et sciences de la culture. *Linx. Revue des linguistes de l'université Paris X Nanterre*, (44), 149-168.

Rastier, François. (2005a). Enjeux épistémologiques de la linguistique de corpus. *La linguistique de corpus. Presses Universitaires de Grenoble*, 31-46.

Rastier, François. (2005b). La Microsémantique. *Texto!*, *10*(2). Récupéré de http://www.revue-texto.net/1996-2007/Inedits/Rastier/Rastier_Microsemantique.html

Rastier, François. (2009). *Sémantique interprétative* (3e éd) Formes sémiotiques. Paris : Presses universitaires de France.

Rastier, François. (2010). Objets culturels et performances sémiotiques. L'objectivation critique dans les sciences de la culture. Dans L. Hébert et L. Guillemette, *Performances et objets culturels nouvelles perspectives* (p. 15-58). Sainte-Foy : Presses de l'Université Laval. Récupéré de Open WorldCat.

Rastier, François. (2011). *La mesure et le grain: sémantique de corpus*. Paris : H. Champion.

Rastier, François. (2012). *Langage et pensée: dualité sémiotique ou dualisme cognitif?* (s. l.) : Texto.

Rastier, François. (2018). Computer-Assisted Interpretation of Semiotic Corpora. Lecture Notes in Morphogenesis. Dans *Quantitative Semiotic Analysis* (p. 123-139). Springer, Cham.

Rastier, François, Cavazza, M. et Abeillé, A. (1994). *Sémantique pour l'analyse: de la linguistique à l'informatique*. (s. l.) : Masson.

Ravoux Rallo, E. (1996). Les mécanismes du récit. *Sciences humaines*, (60), 16-19.

Reese, S. D., Professor, O. H. G. J., Jr, O. H. G., Grant, A. E. et Grant, J. R. M. P. of J. A. E. (2001). *Framing Public Life: Perspectives on Media and Our Understanding of the Social World*. (s. l.) : Routledge. [00928 Google-Books-ID: LhaQAgAAQBAJ].

Reinert, M. (1990). Alceste une méthodologie d'analyse des données textuelles et une application: Aurelia De Gerard De Nerval. *Bulletin de méthodologie sociologique*, *26*(1), 24-54.

Rendón, E., Abundez, I., Arizmendi, A. et Quiroz, E. M. (2011). *Internal versus External cluster validation indexes*, *5*(1), 8.

Revaz, F. (1997). Le récit dans la presse écrite. *Pratiques*, *94*, 19-33.

Rey-Debove, J. (1979). *Sémiotique* (1. éd)Lexique. Paris : Puf. Récupéré de Library of Congress ISBN. (P99 .R48)

Rich, E., Knight, K. et Nair, S. B. (2009). *Artificial intelligence*. Neh Delhi : Tata McGraw-Hill. [00149].

Rieger, B. B. (1997). Computational Semiotics and Fuzzy Linguistics. Dans *Proc. of the 1997 International Conference on Intelligent Systems and Semiotics* (p. 541–551).

Ringoot, R. (2014). *Analyser le discours de presse*. Paris : Armand Colin.

Rizkallah, É. (2013). L'analyse textuelle des discours assistée par ordinateur et les logiciels textométriques : réflexions critiques et prospectives à partir d'une modélisation des procédés analytiques fondamentaux. *Cahiers de recherche sociologique*, (54), 141-160.

Rizkallah, É. et Della Faille, D. (2014). *Regards croisés sur l'analyse du discours*. (54)Cahiers de recherche sociologique. Montréal : Athéna édition.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53-65.

Rumelhart, D. E. (1975). Notes on a schema for stories. *Representation and understanding: Studies in cognitive science*, *211*(236), 45.

Ruwet, N. (2009). Analyse structurale d'un poème français: un sonnet de Louise Labé. *Linguistics*, *2*(3), 62–83.

Ryan, M.-L. (2015). Narratologie et sciences cognitives : une relation problématique*. *Cahiers de Narratologie. Analyse et théorie narratives*, (28).

Sahlgren, M. (2006). *The Word-space model* (Thèse de doctorat). Stockholm : Stockholm University.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, *20*(1), 33-54.

Salem, A. (1982). Analyse factorielle et lexicométrie : synthèse de quelques expériences. *Mots*, *4*(1), 147-168.

Salton, G. et Lesk, M. E. (1968). Computer Evaluation of Indexing and Text Processing. *J. ACM*, *15*(1), 8–36.

Salton, Gerard. (1971). *The SMART retrieval system: Experiments in automatic document processing*. NJ : Prentice-Hall, Upper Saddle River.

Salton, Gerard. (1988). *Automatic Text Processing : The Transformation, Analysis, and Retrieval of Information by Computer*. Boston, MA : Addison-Wesley Publishing Company.

Salton, Gerard. (1991). The Smart Document Retrieval Project. Dans *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (p. 356–358). New York, NY, USA : ACM.

Salton, Gerard et Buckley, C. (1988). Term-weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, *24*(5), 513–523.

Salton, Gerard et McGill, M. J. (1983). *Introduction to modern information retrieval.* New York : McGraw-Hill.

Salton, Gerard, Wong, A. et Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613-620.

Saporta, G. (2006). *Probabilités, analyse des données et statistique*. Paris : Technip.

Saussure, F. de. (1916). *Cours de linguistique générale*. Lausanne : Payot.

Savoy, J. (1999). A stemming procedure and stopword list for general French corpora. *JASIS, 50*(10), 944-952.

Sayyadi, H., Sahraei, A. et Abolhassani, H. (2008). Event Detection from News Articles. Communications in Computer and Information Science. Dans H. Sarbazi-Azad, B. Parhami, S.-G. Miremadi et S. Hessabi (dir.), *Advances in Computer Science and Engineering* (p. 981-984). Springer Berlin Heidelberg.

Sayyadi, H., Salehi, S. et AbolHassani, H. (2007). Survey on News Mining Tasks. Dans T. Sobh (dir.), *Innovations and Advanced Techniques in Computer and Information Sciences and Engineering* (p. 219-224). Springer Netherlands.

Schaeffer, J.-M. (2010). Le traitement cognitif de la narration. Dans *Narratologies contemporaines: approches nouvelles pour la théorie et l'analyse du récit* (p. 215-32). Archives contemporaines.

Schank, R. C. et Abelson, R. P. (1977). *Scripts, plans, goals und understanding: inquiry into human knowledge structures*. Hillsdale : Lawrence Erlbaum.

Scheufele, D. A. (1999). Framing as a theory of media effects. *Journal of communication, 49*(1), 103-122.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. Dans *Proceedings of the International Conference on New Methods in Language Processing*. Manchester; UK.

Schmid, H. (1995). Improvements In Part-of-Speech Tagging With an Application To German. Dans *In Proceedings of the ACL SIGDAT-Workshop* (p. 47–50).

Schöch, C. (2013). Big? Smart? Clean? Messy? Data in the Humanities. *Journal of Digital Humanities*, *2*(3), 2-13.

Schütze, H. (1993). Word space. Dans *Advances in Neural Information Processing Systems 5*. Citeseer.

Schütze, H. et Pedersen, J. (1993). A vector model for syntagmatic and paradigmatic relatedness. Dans *Proceedings of the 9th Annual Conference of the UW Centre for the New OED and Text Research* (p. 104-113). Citeseer.

Shah, N. A. et ElBahesh, E. M. (2004). Topic-based clustering of news articles. Dans *Proceedings of the 42nd annual Southeast regional conference* (p. 412-413). ACM.

Simone, R. (2005). *Fondamenti di linguistica*. Roma : Editori Laterza.

Sinclair, J. (1991). *Corpus, concordance, collocation*. (s. l.) : Oxford University Press.

Sinclair, J. (1996). *Preliminary recommendations on Corpus Typology* [Technical report]. EAGLE (Expert Advisory Group on Language Engineering Standards). Récupéré de http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html

Skvoretz, J. (1993). Generating narratives from simple action structures. *The Journal of Mathematical Sociology*, *18*(2-3), 135-140.

Speed, J. G. (1893). Do newspapers now give the news. Dans *Forum* (vol. 15, p. 705-711).

Stamper, R. K. (1973). *Information in Business and Administrative Systems*. New York, NY : John Wiley & Sons.

Steinbach, M., Ertöz, L. et Kumar, V. (2004). The challenges of clustering high dimensional data. Dans *New directions in statistical physics* (p. 273-309). Springer.

Sudhahar, S., Franzosi, R. et Cristianini, N. (2011). Automating Quantitative Narrative Analysis of News Data. Dans *JMLR: Workshop and Conference Proceedings* (vol. 17, p. 63-71).

Sueur, A. (1994). *Urss et mythologie avant la perestroika*. Thèse de doctorat. Paris 1.

Tahar, A. (2006). *Étude sémiotique du discours journalistique algérien d'expression française le titre du fait divers comme micro-objet sémiotique* (Thèse doctorat). Biskra : Université Mohamed khider Biskra.

Tan, P.-N. et Steinbach, M. (2019). Ch. 7: Cluster Analysis: Basic Concepts and Algorithms. Dans *Introduction to Data Mining* (Second edition). New York, NY : Pearson.

Tanaka-Ishii, K. (2010). *Semiotics of programming*. New York, NY : Cambridge University Press.

Tanaka-Ishii, K. (2015). Semiotics of Computing: Filling the Gap Between Humanity and Mechanical Inhumanity. Dans *International Handbook of Semiotics* (p. 981-1002). New York, NY : Springer.

Tang, X., Yang, C. et Zhou, J. (2009). Stock Price Forecasting by Combining News Mining and Time Series Analysis. Dans *IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT '09* (vol. 1, p. 279-282).

Thérien, G. (1989). Sémiotique et intelligence artificielle. *Études littéraires*, *21*(3), 67.

Traini, S. (2013). *Le basi della semiotica*. Milano : Bompiani.

Tremblay, P.-A., Roche, M. et Tremblay, S. (dir.). (2015). *Le printemps québécois* (1re éd.). Presses de l'Université du Québec.

Tuchman, G. (1980). *Making news: a study in the construction of reality*. New York : Collier Macmillan.

Tufféry, S. (2017). *Data mining et statistique décisionnelle: la science des données*. Paris : Technip.

Tukey, J. W. (1977). *Exploratory Data Analysis* (1 edition). Reading, Mass : Pearson.

Valette, M. (2018). Elements of a Corpus Semantics for Humanities. Application to the Classification of Subjective Texts. Lecture Notes in Morphogenesis. Dans *Quantitative Semiotic Analysis* (p. 141-150). Springer, Cham.

Van Dijk, Teun A. (1972). Foundations for typologies of texts. *Semiotica*, *6*(4), 297-323.

Van Dijk, Teun A. (1973). Grammaires textuelles et structures narratives. Dans C. D. Chabrol Alexandresku, Sorin, Barthes, Roland, Bremond, Claude, Greimas, Algirdas Julien, Maranda, Pierre, Schmidt, Siegfried J. ,. Van Dijk, Teun a (dir.), *Sémiotique narrative et textuelle*. Larousse.

Van Dijk, Teun A. (1980). *Macrostructures*. Hillsdale, NJ : Erlbaum.

Van Dijk, Teun A. (1983). Discourse Analysis: Its Development and Application to the Structure of News. *Journal of Communication*, *33*(2), 20-43.

Van Dijk, Teun A. (1985a). *Discourse and Communication: New Approaches to the Analysis of Mass Media Discourse and Communication*. (s. l.) : Walter de Gruyter.

Van Dijk, Teun A. (1985b). Structures of news in the press. *Discourse and communication*, 69-93.

Van Dijk, Teun A. (1988a). *News analysis: Case studies of international and national news in the press*. Hillsdale, NJ : Lawrence Erlbaum Associates.

Van Dijk, Teun A. (1988b). *News as discourse*. Hillsdale, NJ : Lawrence Erlbaum Associates.

Van Dijk, Teun A. (2008a). *Discourse and context*. Cambridge : Cambridge University Press. [02523].

Van Dijk, Teun A. (2008b). *Discourse and power*. Basingstoke : Palgrave Macmillan.

Van Dijk, Teun Adrianus et Kintsch, W. (1983). *Strategies of discourse comprehension*. (s. l.) : Citeseer.

Van Rijsbergen, C. J. (2004). *The geometry of information retrieval*. Cambridge, England ; New York : Cambridge University Press.

Vijayarani, S., Ilamathi, M. J. et Nithya, M. (2015). Preprocessing Techniques for Text Mining-An Overview. *International Journal of Computer Science and Communication Networks*, *5*(1), 7-16.

Volli, U. (2008). *Manuale di semiotica* (7e éd.). Roma : GLF editori Laterza.

Volli, U. (2014). L'analisi semiotica come ricerca empirica sul testo. *COSMO | Comparative Studies in Modernism*, *0*(4). Récupéré de www.ojs.unito.it : http://www.ojs.unito.it/index.php/COSMO/article/view/635

Weber, R. P. (1990). *Basic content analysis*. Thousand Oaks, California : Sage Publications.

Wierzchon, S. et Klopotek, M. (2018). *Modern Algorithms of Cluster Analysis* (vol. 34)Studies in Big Data. Cham : Springer.

Wüest, B., Rothenhäusler, K. et Hutter, S. (2013). Using Computational Linguistics to Enhance Protest Event Analysis. *SSRN Electronic Journal*.

Wynne, M. (2005). *Developing linguistic corpora: A guide to good practice* (vol. 92). Oxbow Books Oxford.

Yan, Y., Okazaki, N., Matsuo, Y., Yang, Z. et Ishizuka, M. (2009). Unsupervised Relation Extraction by Mining Wikipedia Texts Using Information from the Web. Dans *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language*

Processing of the AFNLP: Volume 2 - Volume 2 (p. 1021–1029). Stroudsburg, PA, USA : Association for Computational Linguistics.

Yao, L., Riedel, S. et McCallum, A. (2012). Unsupervised Relation Discovery with Sense Disambiguation. Dans *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1* (p. 712–720). Stroudsburg, PA, USA : Association for Computational Linguistics.

Young, K. et Saver, J. L. (2001). The Neurology of Narrative. *SubStance*, *30*(1), 72-84.

Yu, C. H., Jannasch-Pennell, A. et DiGangi, S. (2011). Compatibility between Text Mining and Qualitative Research in the Perspectives of Grounded Theory, Content Analysis, and Reliability. *The Qualitative Report*, *16*(3), 730-744.

Zarri, G. P. (1997). NKRL, a knowledge representation tool for encoding the of complex narrative texts. *Natural Language Engineering*, *3*(02), 231–253.

Zarri, G. P. (2010). Representing and Managing Narratives in a Computer-Suitable Form. Dans *2010 AAAI Fall Symposium Series*.

Zarri, G. P. (2015). Semantic/Conceptual Annotation Techniques Making Use of the Narrative Knowledge Representation Language (NKRL). Dans *FLAIRS Conference* (p. 131-136).

Zemanek, H. (1966). Semiotics and Programming Languages. *Commun. ACM*, *9*(3), 139–143.

Zhai, C. (2009). *Statistical language models for information retrieval*. San Rafael (Calif.) : Morgan & Claypool.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, Mass : Addison-Wesley.