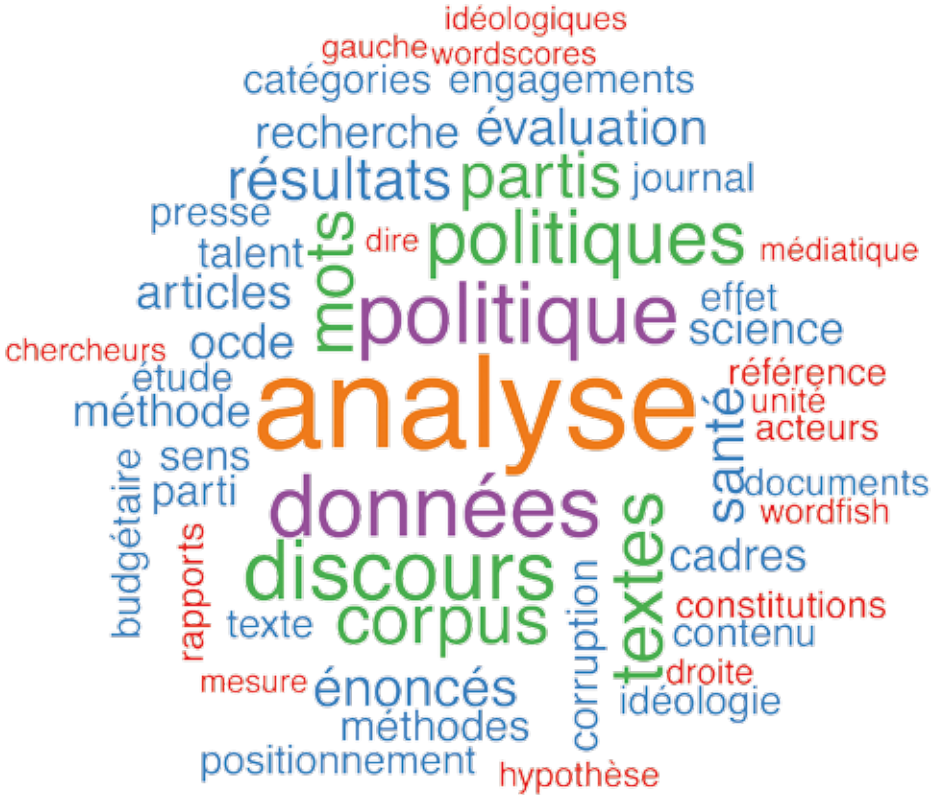
L'analyse textuelle

des idées, du discours et des pratiques politiques



Sous la direction de
Pierre-Marc Daigneault
François Pétry

CAPP Centre d'analyse des politiques publiques

PUL

L'analyse textuelle des idées, du discours et des pratiques politiques

MÉTHODES DE RECHERCHE EN SCIENCES HUMAINES

Collection dirigée par Louis M. Imbeau

Au cœur des sciences humaines, la question de la méthode alimente les débats, non seulement entre les « écoles » (modernisme/postmodernisme, qualitativisme/quantitativisme, monisme/pluralisme, individualisme/holisme, etc.), mais aussi entre les chercheurs à l'intérieur de chaque école.

La méthode est aussi au cœur de la formation des chercheurs. En plus de la maîtrise de plusieurs méthodes de recherche, devenir chercheur implique l'habileté à jeter un regard critique sur son propre travail et sur celui des autres.

Cette collection veut contribuer aux débats sur la méthode et à la formation méthodologique des chercheurs des sciences humaines. Dans cet esprit, on y accueillera aussi bien des essais critiques s'adressant aux spécialistes que des manuels à l'intention des chercheurs, qu'ils soient expérimentés ou en formation.

DANS LA MÊME COLLECTION

William Fox, *Statistiques sociales.* Traduction et adaptation de Louis M. Imbeau (avec la collaboration d'Augustin Simard et de Thierry Rodon), PUL et De Boeck, 1999 (14e tirage, 2012).

Gordon Mace et François Pétry, *Guide d'élaboration d'un projet de recherche*, 2e édition, PUL et De Boeck, 2000.

François Dépelteau, *La démarche d'une recherche en sciences humaines. De la question de départ à la communication des résultats*, 2e édition, PUL et De Boeck, 2000 (7e tirage, 2011).

Vincent Lemieux et Mathieu Ouimet, *L'analyse structurale des réseaux sociaux*, PUL et De Boeck, 2004.

André Sanfaçon, *La dissertation historique. Guide d'élaboration et de rédaction*, 2e édition, PUL, 2005.

Patrick Gonzalez et Jean Crête, *Jeux de société. Une initiation à la théorie des jeux en sciences sociales*, PUL, 2006.

François Pétry et François Gélineau, *Guide pratique d'introduction à la régression en sciences sociales*, 2e édition, PUL, 2009.

Louis M. Imbeau, *Statistiques sociales avec IBM SPSS*md*. Cahier d'exercices de la 19e version*, 2e tirage, 2012.

Christian Papinot, *La relation d'enquête comme relation sociale. Épistémologie de la démarche de recherche ethnographique*, 2014.

Jean-Herman Guay, *Statistiques en sciences humaines avec R*, 2e édition, 2014.

Jimmy Bourque et Salah-Eddine El Adlouni, *Manuel d'introduction à la statistique appliquée en sciences sociales*, 2016.

Gordon Mace et François Pétry, *Guide d'élaboration d'un projet de recherche*, 3e édition revue et augmentée, 2017.

L'analyse textuelle des idées, du discours et des pratiques politiques

SOUS LA DIRECTION DE
PIERRE-MARC DAIGNEAULT ET FRANÇOIS PÉTRY

Presses de
l'Université Laval

Les Presses de l'Université Laval reçoivent chaque année du Conseil des arts du Canada et de la Société de développement des entreprises culturelles du Québec une aide financière pour l'ensemble de leur programme de publication.

Financé par le gouvernement du Canada
Funded by the Government of Canada

Canada

Mise en pages : Diane Trottier
Maquette de couverture : Laurie Patry

Image de couverture : L'image intégrée au design de la couverture a été réalisée par Dominic Duval à l'aide du logiciel R et d'une version complète et quasi finale du manuscrit de cet ouvrage. Le nuage de mots-clés présente les mots non triviaux les plus courants de cet ouvrage et les représente en une police plus ou moins grande en fonction de leur importance.

ISBN 978-2-7637-3198-8
PDF 9782763731995

Dépôt légal 3e trimestre 2017

www.pulaval.com

Table des matières

Dans la même collection II

Remerciements XV

PRÉFACE
Les boîtes à outils de l'analyse de données textuelles (ADT) : des chantiers aussi prometteurs que périlleux XVII
Benoît Rihoux
Un regard amicalement critique sur l'ADT XVII
Un domaine fragmenté XVIII
De la science des mots vers la science des chiffres XIX
De multiples promesses XX
De sérieux périls XXII
Conseils d'un voisin critique et amical XXIII

INTRODUCTION
Quelques repères pour appréhender l'analyse des données textuelles dans toute sa diversité 1
Pierre-Marc Daigneault et François Pétry
La raison d'être de l'ouvrage 2
Appréhender un objet disparate et fragmenté 3
Les contributions 6

PREMIÈRE PARTIE
MÉTHODES D'ANALYSE DE DISCOURS

CHAPITRE 1
La parole politique comme performance multimodale et interactionnelle. Une proposition d'analyse 19
Olivier Turbide
1.1 La parole politique en action comme performance 20
1.1.1 Les scènes de la performance 21
1.1.2 L'incarnation d'un personnage 22
1.2 Présentation de l'extrait tiré d'une interview de *talk-show* 24

1.3 Deux principes méthodologiques de l'approche interactionnelle et multimodale 26
1.3.1 Co(n)texte et interprétation 26
1.3.2 Complexité de la parole et atomisation des actions 33
Conclusion 34

ANNEXE 1 : CONVENTION DE TRANSCRIPTION. 36

CHAPITRE 2
Théoriser à partir de données qualitatives secondaires : comparaison de deux méthodes d'analyse des données textuelles 41
Isabelle F.-Dufour et Marie-Claude Richard
2.1 La théorisation ancrée 43
2.2 La méthode générale inductive 44
2.3 Méthodologie 46
2.3.1 Les chercheuses 46
2.3.2 Le corpus. 46
2.3.3 Le dispositif expérimental 47
2.4 Analyses selon la théorisation ancrée (TA) 47
2.4.1 Phase 1 : codage ouvert 47
2.4.2 Phase 2 : codage axial 50
2.4.3 Phase 3 : codage sélectif. 51
2.4.4 Théorisation selon la TA 52
2.5 Analyses selon la méthode générale inductive (MGI) 54
2.5.1 Phase 1 : analyse selon la codification/catégorisation ouverte 54
2.5.2 Phase 2 : codification/catégorisation selon les concepts sensibilisateurs 55
2.5.3 Théorisation selon la MGI 56
2.6 Comparaison des résultats obtenus avec les deux méthodes 59
2.6.1 Avantages et inconvénients de chacune des méthodes 59
2.6.2 Enjeux liés à l'analyse de données qualitatives secondaires 61
2.6.3 Limites de l'étude et pistes pour les recherches ultérieures 62
Conclusion 63

CHAPITRE 3
La domination d'une idéologie. Histoire des usages politiques du concept de talent (1945-2015) 67
Adrien Thibault
3.1 Une présence diffuse 70
3.1.1 Un vocable journalistique usuel. 70
3.1.2 Un vocable de plus en plus courant. 73
3.1.3 Un vocable de la grandeur 75

3.2 Une présence générale ... 77
3.2.1 Des champs entiers et multiples ... 77
3.2.2 Du champ artistique au champ économique ... 79
3.2.3 Variations mélodiques sur un même champ ... 84
3.3 Un concept politique ... 85
3.3.1 Un concept qui sert à qualifier la politique ... 86
3.3.2 Un concept qui sert aux politiciens ... 87
3.3.3 Un concept qui sert des politiques ... 90
Conclusion ... 92

DEUXIÈME PARTIE
MÉTHODES DE CLASSIFICATION MANUELLES ET AUTOMATISÉES AVEC CATÉGORIES INCONNUES

CHAPITRE 4
Les mots de la campagne : la fouille de textes appliquée à l'étude de la communication électorale ... 97
Dominic Forest, Frédérick Bastien, Ariane Legault-Venne, Olivier Lacombe, Hélène Brousseau
4.1 Le discours politique à l'ère numérique ... 99
4.2 La fouille de textes : principes et méthode ... 100
4.2.1 La constitution du corpus ... 103
4.2.2 Le filtrage ... 103
4.2.3 La transformation vectorielle ... 104
4.2.4 Application des algorithmes de fouille ... 105
4.2.5 Évaluation, interprétation et intégration ... 107
4.3 Une application de la fouille de textes : les mots de la campagne ... 108
4.3.1 Constitution du corpus ... 110
4.3.2 Résultats ... 111
4.3.2.1 Le cadrage de la crise des réfugiés syriens à travers les mots des partis ... 111
4.3.2.2 Le contrôle de l'ordre du jour à travers la cohésion dans le discours des partis ... 114
Conclusion ... 119

CHAPITRE 5
Les rapports de l'OCDE consacrés à la santé et leur réception médiatique en France. Objectiver « l'influence » médiatique par l'analyse textuelle informatisée ... 123
Constantin Brissaud
5.1 Analyser l'influence de l'OCDE à l'aide d'une méthode lexicométrique ... 125
5.2 Objectifs ... 127

5.3 Présentation du logiciel Iramuteq et de la méthode utilisée 128
5.4 Mode de constitution des corpus de textes 130
5.4.1 Le corpus « Rapports ». 130
5.4.2 Constitution du corpus « Presse ». 131
5.5 Résultats et discussion 132
5.5.1 Thèmes contenus dans les corpus 132
5.5.2 Segments de textes caractéristiques. 135
Conclusion 137

CHAPITRE 6
Représenter la corruption : définition d'un problème public à travers la production médiatique française 145
Sofia Wickberg
6.1 Cadre théorique. 147
6.1.1 Discours médiatique et cadrage 147
6.1.2 Définition du problème par le cadrage médiatique 148
6.2 Données et méthodes 150
6.2.1 Collecte des données 150
6.2.2 Présentation des cadres d'interprétation 152
6.3 Résultats. 154
6.3.1 Attribution de responsabilité et risque de dépolitisation. 154
6.3.2 Rhétorique du drame, du scandale et de la morale 156
Conclusion 159

TROISIÈME PARTIE
MÉTHODES DE CLASSIFICATION MANUELLES ET AUTOMATISÉES AVEC CATÉGORIES CONNUES

CHAPITRE 7
L'analyse textuelle automatisée et l'analyse des sentiments pour comprendre les cadrages médiatiques sur l'euthanasie : potentiel et défis. 167
Lisa Birch et Sandra P. Escalera
7.1 Considérations théoriques 169
7.2 Méthode. 171
7.2.1 Étape 1 : constitution d'un corpus de textes. 172
7.2.2 Étape 2 : création de dictionnaires généraux 174
7.2.3 Étape 3 : identification des cadres concurrentiels et codage automatisé 176
7.2.4 Étape 4 : estimation du ton 177
7.2.5 Étape 5 : production de graphiques 178

7.3 Résultats. 178
7.3.1 Les cadres de l'enjeu des politiques de fin de vie 178
7.3.2 L'évolution de la couverture médiatique des cadres concurrentiels par cas 179
Conclusion 185

CHAPITRE 8
Les dieux, les monarques et la constitution : une analyse de contenu 191
Jean Crête
8.1 De la présence des dieux dans les constitutions 192
8.1.1 Coopération 196
8.1.2 Soumission 197
8.1.3 Incertitude 199
8.2 Méthode. 200
8.2.1 Variable Sécularité-Transcendance 200
8.2.2 Variable régime politique 203
8.2.3 Variables coopération, soumission et incertitude 203
8.3 Résultats et discussion 204
8.3.1 Référentiel et régime politique 205
Conclusion 212

CHAPITRE 9
L'optimisme des unités d'évaluation de technologies et modes d'intervention en santé : analyses statistiques fondées sur l'analyse du contenu de leurs rapports. 217
Mathieu Ouimet, Pascal Lalancette, Alexandre Racine
9.1 Mise en contexte 218
9.2 Ancrage théorique et hypothèses 220
9.3 Méthodologie 223
9.4 Résultats. 228
Conclusions et discussion 234

QUATRIÈME PARTIE
MÉTHODES DE POSITIONNEMENT MANUELLES ET AUTOMATISÉES

CHAPITRE 10
Les partis promettent-ils des politiques qui correspondent à leur idéologie ? 241
François Pétry, Dominic Duval, Lisa Birch, Jean Crête
10.1 Approches empiriques pour mesurer la fiabilité des partis 242
10.2 Méthode 244
10.3 Résultats. 248
Conclusion 259

CHAPITRE 11
L'analyse de contenu automatisée et les entretiens font-ils bon ménage ? Caractériser l'idéologie d'une réforme de politique sociale à l'aide de Wordscores 267
Pierre-Marc Daigneault, Dominic Duval, Louis M. Imbeau
11.1 La méthode Wordscores 269
11.1.1 Principes généraux 269
11.1.2 Procédures de calcul des scores 270
11.1.3 Bilan de la méthode 272
11.2 Les données d'entretiens et Wordscores 273
11.3 Objectifs, devis de recherche et cadre conceptuel 276
11.3.1 Objectifs 276
11.3.2 Devis de recherche 276
11.3.3 Cadre conceptuel 278
11.4 Méthodes 279
11.4.1 Corpus 279
11.4.2 Procédures 280
11.5 Résultats 281
Discussion et conclusion 286

CHAPITRE 12
Idéologie partisane et déficit budgétaire : le conservatisme budgétaire des premiers ministres provinciaux au Canada, 1971-2015 293
Louis M. Imbeau et Mickael Temporão
12.1 D'une dichotomie gauche – droite à une dichotomie vision totale – vision partielle du budget 296
12.2 La mesure du conservatisme budgétaire 299
12.3 Résultats 301
Conclusion 305

CHAPITRE 13
Mesurer les préférences budgétaires des maires et mairesses québécois à l'aide de Wordfish 309
Jérôme Couture
13.1 Wordfish 311
13.2 Méthodologie 315
13.3 Hypothèses de recherche 318
13.4 Résultats 321
Conclusion 322

COLLABORATEURS ET COLLABORATRICES 327

Liste des figures

Figure 1.1 Extrait de l'interview P. Lagacé (IR) – G. Barrette (IÉ), *Deux hommes en or*, Télé-Québec, 7 mars 2014 (manifestations verbales seulement) 25

Figure 1.2 Extrait de l'interview P. Lagacé (IR) – G. Barrette (IÉ), *Deux hommes en or*, Télé-Québec, 7 mars 2014 (mouvements gestuels A-B-C) 29

Figure 1.3 Extrait de l'interview P. Lagacé (IR) – G. Barrette (IÉ), *Deux hommes en or*, Télé-Québec, 7 mars 2014 (mouvement gestuel D)..... 33

Figure 2.1 La recomposition familiale réussie selon la MGI..... 56

Figure 3.1 Croissance du vocable en valeur relative 74

Figure 3.2 Croissance du vocable en valeur comparée 75

Figure 3.3 Analyse factorielle des correspondances – *Le Monde*, 1945-2015... 80

Figure 3.4 Dendrogramme de la CHD à 6 classes – *Le Monde* (1944-1964) 82

Figure 3.5 Dendrogramme de la CHD à 6 classes – *Le Monde* (2005-2015) 83

Figure 3.6 Fréquence du lemme *talent* dans les discours des présidents français, par mandat..... 88

Figure 3.7 Mots associés à *talent* dans les discours de Georges Pompidou (1969-1974) 89

Figure 3.8 Mots associés à *talent* dans les discours de François Hollande (2012-2015) 89

Figure 4.1 La démarche générique de fouille de textes..... 103

Figure 4.2 Exemple de matrice *documents x mots* dans laquelle chaque document (colonne) est traduit numériquement par la fréquence d'un ensemble de mots (ligne)..... 105

Figure 4.3 Présence du thème des réfugiés dans les documents des cinq principaux partis, pondérée par 10 000 mots 112

Figure 4.4 Portrait des mots utilisés par les partis en lien avec la thématique des réfugiés syriens 113

Figure 5.1 Dendrogramme des thèmes du corpus « Rapports » 133

Figure 5.2 Dendrogramme des thèmes du corpus « Presse »..... 133

Figure 6.1 Fréquence des cadres d'action 154

Figure 7.1 Couverture des cadres (% dans le total des articles 1995-2015) 180

Figure 7.2 Couverture des cadres, 1995 à 2015 (% du total des cadres couverts par les journaux)..... 182

Figure 7.3 Ton moyen des segments codés 184

Figure 8.1 Les catégories du référentiel 203
Figure 8.2 Coopération selon le référentiel 206
Figure 8.3 Soumission selon le référentiel 207
Figure 8.4 Incertitude selon le référentiel 208
Figure 8.5 Coopération selon le référentiel et le régime 209
Figure 8.6 Soumission selon le référentiel et le régime 210
Figure 8.7 Incertitude selon le référentiel et le régime 211
Figure 9.1 Probabilité prédite (en points de %) que l'usage d'une TMIS soit recommandé favorablement en fonction de son statut d'utilisation au moment de l'évaluation 231
Figure 9.2 Probabilité prédite (en points de %) que l'usage d'une TMIS soit recommandé favorablement en fonction de différents scénarios . . . 232
Figure 10.1 Positionnement idéologique de tous les partis et pour tous les types d'énoncés 249
Figure 10.2 Positionnement de l'idéologie et des engagements spécifiques du Parti progressiste-conservateur et du Parti conservateur 255
Figure 10.3 Positionnement de l'idéologie et des engagements spécifiques du Parti libéral 255
Figure 10.4 Positionnement de l'idéologie et des engagements spécifiques du Nouveau Parti démocratique 255
Figure 10.5 Positionnement de l'idéologie et des engagements spécifiques du Bloc québécois 256
Figure 10.6 Positionnement de l'idéologie et des engagements spécifiques réalisés par le PLC et le PCC au gouvernement 259
Figure 11.1 Stratégie de validation convergente 277
Figure 11.2 Positionnement de la réforme et de l'activation (scores transformés) 282
Figure 11.3 Positionnement des répondants (scores transformés). 284
Figure 11.4 Positionnement des répondants classés selon leur idéologie (scores transformés) 285
Figure 12.1 Zones idéologiques relatives aux dépenses et aux recettes. 297
Figure 12.2 Zones idéologiques relatives aux dépenses, aux recettes et au solde 298
Figure 12.3 Indice de conservatisme budgétaire (ICB) par province, 1971-2015 . 302
Figure 12.4 Conservatisme budgétaire des premiers ministres de sept provinces canadiennes, 1971-2015 (courbes de tendance « Lowess »). 304
Figure 13.1 Dimension de l'analyse Wordfish 317
Figure 13.2 Distribution des scores Wordfish. 317

Liste des tableaux et encadrés

Tableau 2.1 Exemples de codes et de catégories : codage ouvert 49
Tableau 2.2 Catégories et sous-catégories définitives au terme du codage ouvert 50
Tableau 2.3 Composantes de la catégorie centrale d'analyse 52
Tableau 2.4 Exemples de codes, rubriques et catégories : codification ouverte des données 55
Tableau 2.5 Comparaison des deux méthodes d'analyse 61
Tableau 3.1 Prégnance du vocable en valeur absolue 70
Tableau 3.2 Prégnance du vocable en valeur relative 73
Tableau 3.3 Champs lexicaux et champs sociaux 77
Tableau 3.4 Caractérisation des classes issues de la CHD selon la variable *date* .. 81
Tableau 4.1 Statistiques du corpus 111
Tableau 4.2 Résultats obtenus sur le corpus d'apprentissage 116
Tableau 4.3 Résultats obtenus sur le corpus de test 117
Tableau 6.1 Mots-clés utilisés pour la recherche 151
Tableau 6.2 Distribution des articles de l'échantillon par journal et par année 152
Tableau 6.3 Cadres d'action 153
Tableau 6.4 Cadres rhétoriques 153
Tableau 6.5 Fréquence des cadres d'action par aire géographique 155
Tableau 6.6 Fréquence des cadres rhétoriques 156
Tableau 7.1 Corpus d'articles de 1995-2015 173
Tableau 7.2 Exemples des mots-clés et unités de sens de chaque cadre (thème) 177
Tableau 7.3 Cadres spécifiques à l'euthanasie et aux politiques de fin de vie 179
Tableau 8.1 Régime politique et référentiel 205
Tableau 9.1 Définition opérationnelle de la variable dépendante 226
Tableau 9.2 Description des variables et statistiques descriptives 229
Tableau 10.1 Fréquence de mentions des catégories de l'échelle gauche-droite dans les énoncés idéologiques (ID) et d'engagements (EN) de chaque parti 1997-2011 252
Tableau 11.1 Positionnement de la réforme et de l'activation, scores originaux et transformés (MV) 281

Tableau 11.2 Positionnement des répondants, scores originaux et transformés (MV) 283
Tableau 12.1 Indice de conservatisme budgétaire (ICB) : statistiques descriptives 303
Tableau 13.1 Exemple fictif d'une matrice de mots analysée par Wordfish 312
Tableau 13.2 Résumé des composantes de l'analyse Wordfish 318
Tableau 13.3 Analyse descriptive des variables 321
Tableau 13.4 Analyse de régression (moindres carrés ordinaires) 322

Encadré 8.1 Exemples d'énoncés dans la catégorie séculier déclaré 201
Encadré 8.2 Exemples d'énoncés dans la catégorie transcendant 202
Encadré 8.3 Exemples d'énoncés dans la catégorie séculier-transcendant 202

Remerciements

Un livre de cette ampleur est nécessairement le fruit de l'appui et du travail – souvent plus ou moins visible – d'une myriade de personnes et d'organisations. Nous tenons d'abord à remercier Steve Jacob, directeur du Centre d'analyse des politiques publiques (CAPP), et le Fonds de recherche du Québec – Société et culture (FRQSC), sans qui, faute de fonds, cet ouvrage n'aurait jamais vu le jour sous cette forme. Plus généralement, nous tenons à remercier les membres de l'équipe de recherche « Évaluer la performance publique et la mesurer par l'analyse textuelle », financée par le programme de Soutien aux équipes de recherche du FRQSC, pour leur intérêt et leur enthousiasme pour ce projet. Nous sommes également reconnaissants à la Société québécoise de science politique (SQSP) de nous avoir offert la possibilité de tenir l'atelier « Les idées, le discours et les pratiques politiques au prisme de l'analyse des données textuelles » à l'occasion de son Congrès annuel 2016, qui s'est tenu du 19 au 21 mai, à l'Université Laval. Cet ouvrage est en effet tributaire de la diversité et de la qualité des communications qui y ont été présentées, ainsi que des échanges stimulants qui se sont tenus tout au long de l'atelier. Nous sommes reconnaissants aux participants de l'atelier qui ont accepté de commenter les présentations, en particulier Daniel Béland, ainsi que Benoît Rihoux, qui a bien voulu écrire la préface de l'ouvrage. Nous désirons enfin exprimer notre gratitude à toutes les personnes qui ont contribué à concrétiser ce projet : André Baril (travail d'édition), Marie-Hélène L'Heureux (travail d'édition, de mise en forme et de révision linguistique lors de la présoumission), Laurie Patry (graphisme) et Dominic Duval (création du nuage de mots clés utilisé en couverture de l'ouvrage). Nous demeurons bien entendu les seuls responsables des erreurs et des lacunes qui pourraient subsister dans cet ouvrage.

Pierre-Marc Daigneault et François Pétry

PRÉFACE

Les boîtes à outils de l'analyse de données textuelles (ADT) : des chantiers aussi prometteurs que périlleux

Benoît Rihoux

UN REGARD AMICALEMENT CRITIQUE SUR L'ADT

Contrairement à la plupart des contributeurs et contributrices au présent ouvrage, je ne suis ni un spécialiste ni un usager régulier de l'ADT. Je me considère surtout comme un « voisin critique et amical » de celle-ci. Mon regard est aussi celui du gestionnaire de programmes de formation pour doctorants en science politique et disciplines connexes, en particulier en tant que responsable académique de l'école de méthodes du Consortium européen de recherche politique (ECPR). C'est dans ce cadre précisément que j'ai été amené à découvrir la richesse dans ce domaine, ce qui m'a amené, avec mes collègues de l'ECPR, à faire le choix délibéré d'offrir une diversité de formations en la matière : analyse de discours, analyse de contenu, analyse de données qualitatives et analyse textuelle quantitative.

Lorsque j'examine une « boîte à outils » méthodologique, quelle qu'elle soit, j'adopte toujours une double perspective : ouverte et critique à la fois. D'abord ouverte, car toute méthode présente des forces spécifiques. Plus précisément : toute méthode, potentiellement, peut être précieuse pour *certains* types de données et pour *certains* types de questions de recherche. L'erreur – ou l'illusion – la plus fréquente en matière de méthodes, me semble-t-il, est d'adopter la croyance selon laquelle une méthode

donnée constitue la méthode miracle pour répondre de manière décisive à toute question de recherche. C'est à mon sens une tendance qu'on observe aujourd'hui, dans certains pans de la science politique, avec les méthodes expérimentales (Rihoux, 2013).

Il est vrai que le domaine de l'ADT est en croissance rapide, à la fois dans ses versants qualitatif et quantitatif (*cf.* infra). Elle est en réalité plus universellement présente qu'on puisse le suspecter, bien au-delà de la boîte à outils des chercheurs et chercheuses en sciences sociales, car tous les moteurs de recherche Web (et le super moteur de recherche qu'est Google) sont, eux aussi, construits autour d'algorithmes d'ADT.

UN DOMAINE FRAGMENTÉ

En première approximation, l'ADT se définit surtout par son objet, c'est-à-dire les « données textuelles » – l'étiquette ADT couvrirait donc, simplement, toutes les méthodes qui envisagent les textes comme des données d'analyse. Toutefois, au-delà de cette définition un peu tautologique, il faut bien convenir que le champ de l'ADT est pluriel. Il est d'ailleurs tellement pluriel et fragmenté qu'il est difficile d'en définir les contours précis, et il y a fort à parier que l'exercice même de la définition des contours ne recueillerait pas un plein consensus parmi les auteurs des chapitres du présent ouvrage (voir à sujet l'introduction de Daigneault et Pétry, dans cet ouvrage). Il est d'ailleurs symptomatique de relever que les différents manuels en la matière ne s'accordent pas sur une définition commune. La raison en est que l'ADT est certes un champ très dynamique, mais aussi un champ très divisé, ou plus précisément segmenté.

Pour simplifier, il semble que ce vaste domaine comporte deux grands segments, assez largement distincts, et qui coïncident assez bien (hélas, pourrait-on ajouter) avec le grand clivage entre approches « qualitative » et « quantitative » en sciences sociales.

D'une part, au sein des approches qualitatives, on retrouve principalement la grande variété – l'éparpillement pourrait-on presque dire – des écoles de l'analyse de discours. On peut définir cette dernière, de manière très large, comme une approche socio-sémantique qui s'intéresse au moins autant au contexte de relation de pouvoir dans lequel un texte est produit et à l'effet de ce texte

sur le récepteur et sur le corps social qu'au texte lui-même dans ses caractéristiques sémantiques et dans son contenu intrinsèque. Dans ce vaste champ assez divisé, on retrouve entre autres l'analyse structurale, l'analyse sémiologique, l'analyse sémiotique, ou encore l'analyse critique de discours, elle-même issue de la théorie critique de l'École de Francfort. Cette dernière se caractérise par un rejet plus ou moins marqué d'approches plus quantitatives.

D'autre part, on trouve le vaste domaine de ce que certains appellent l'analyse de contenu ou l'analyse lexicométrique ou textométrique, et que d'autres nomment l'analyse textuelle quantitative. Ce domaine partage une préoccupation commune (qui constitue aussi un postulat commun) selon laquelle les données textuelles constituent des données organisées, donc susceptibles d'analyses par traitements informatiques plus ou moins automatisés. Une évolution marquante dans ce segment de l'ADT est le passage du codage manuel au codage assisté par ordinateur (à l'aide de lexiques et dictionnaires générés par ordinateur) et, plus récemment, à des modes d'extraction plus puissants et automatisés, à l'instar des procédures du type Wordscores ou Wordfish permettant d'extraire automatiquement les positions de contenu politique ou autre dans un texte donné ou dans un corpus de textes. Une procédure voisine qui a connu un essor récent est l'analyse des sentiments *(sentiment analysis* aussi appelée *opinion mining*) qui permet de faire ressortir automatiquement dans un texte donné ou dans un corpus de textes, les principales émotions et attitudes structurantes. D'une certaine manière, en termes d'objectif final, ces deux dernières techniques viennent concurrencer des approches plus qualitatives et manuelles telles des analyses sémantiques ou sémiotiques.

DE LA SCIENCE DES MOTS VERS LA SCIENCE DES CHIFFRES

Bien sûr, comme discuté plus haut, l'analyse des données textuelles ne se résume pas aux approches quantitatives. Cependant, si l'on prend le recul de la longue durée, on peut considérer que l'ADT, au moins dans sa large composante plus quantitative et informatisée, correspond à un profond et remarquable changement de paradigme. Historiquement, depuis l'antiquité et au fur et à mesure du développement des sciences humaines et sociales, l'approche la plus fréquente des textes était une approche spéculative, basée avant tout sur le raisonnement,

la démonstration, l'exégèse. C'est encore cette approche qui nourrit aujourd'hui des disciplines comme la philosophie et la philologie.

Le point fondamental est que ce que nous appelons aujourd'hui l'approche « empirique » (par opposition à l'approche « normative » ou « spéculative »), que l'on peut faire remontrer à Aristote, et plus près de nous à Mill et à Durkheim, s'est développée pendant très longtemps en-dehors de l'analyse textuelle : le matériau de la recherche n'était pas le texte ou le discours, mais les *phénomènes* sociopolitiques, à l'instar des sciences « exactes » qui analysent, mesurent et manipulent des phénomènes physiques, mécaniques, chimiques ou biologiques (De Meur et Rihoux, 2002, p. 19-20). La science politique empirique, en particulier la science politique quantitative, qui s'est développée depuis les années 1950 avec le tournant comportementaliste (*behavioral turn*) et qui est aujourd'hui devenue dominante s'est, pour l'essentiel, détournée des textes et a porté son attention vers l'observation des comportements politiques individuels, collectifs et institutionnels.

Or, avec le développement de l'ADT, en particulier dans son versant quantitatif, la science formalisée – la « science des chiffres » – a opéré un tournant et a rejoint la « science des mots », en s'intéressant enfin à un corpus – les discours sociopolitiques sous leurs multiples formes – qui a toujours existé, mais qui était resté largement marginal lorsqu'il s'agissait d'analyser plutôt les faits sociopolitiques observables. Certes, l'analyse d'éléments textuels n'était pas entièrement absente même dans les années 1950 à 1980. Par exemple, dans les grandes enquêtes électorales avec des échantillons de grande taille, des questions ouvertes étaient parfois insérées. Néanmoins, le traitement de ces bribes de données textuelles restait fort limité : soit une simple interprétation (assez libre et donc peu rigoureuse), soit un codage manuel en catégories permettant des traitements statistiques simples (par exemple, une classification de ce qui motive les électeurs à voter pour tel ou tel parti).

DE MULTIPLES PROMESSES

La grande richesse – au moins potentielle – de l'ADT réside d'abord dans la richesse des données pour lesquelles elle est conçue. Mais que constituent en réalité les « données textuelles » ?

Il est utile de considérer les données textuelles comme des données sociales faisant l'objet d'une communication sous forme de mots et de phrases d'un producteur vers un récepteur. Mais les données textuelles constituent un type particulier de « données communicationnelles » : outre le texte, il y a aussi l'image (par exemple, les affiches électorales, la gestuelle d'un candidat politique) et le son (par exemple, un discours ou un entretien enregistré, l'intonation lors de ce discours ou entretien) (Bauer, Gaskell et Allum, 2000, p. 5-6). Et, plus important : ce qui distingue le texte de l'image et du son, c'est le type de moyen de communication utilisé, c'est-à-dire le fait que le discours soit reproduit sur un document imprimé, manuscrit ou, de plus en plus fréquemment, électronique et donc dématérialisé.

Une caractéristique centrale de ce type de données textuelles est leur diversité et leur richesse en matière de supports et de forme. Souvent, on se retrouve face à des « données désordonnées » (*messy data*), dont la plupart n'ont pas été générées par le chercheur lui-même, et qui requièrent donc toute une série d'opérations de compilation, de sélection, de préparation, de formatage, de nettoyage, etc. avant de procéder à l'analyse proprement dite. Parmi ces données si diverses, on peut en distinguer deux grandes catégories : les données textuelles « informelles » ou « formelles » (Bauer, Gaskell et Allum, 2000, p. 5-6). Les premières sont produites plus « à chaud » par le locuteur, par exemple dans une expression spontanée ou dans la réponse à une question ouverte dans un questionnaire d'enquête ou pendant un entretien.

Toujours dans le domaine des données informelles, on pourrait ajouter une très grande diversité de données dites « organiques », c'est-à-dire riches, mais générées sous une forme qui n'est pas très structurée, ou de structure variable (Groves, 2011), par exemple l'énorme masse de données textuelles dématérialisées produites par les divers réseaux sociaux de type Facebook ou Twitter, données qui sont en effet d'une très grande richesse et qui constituent aussi une partie de ce que certains nomment les données massives (*big data*). Ces données, ainsi que la masse de données très diverses, y compris textuelles, archivées sur le Web, peuvent à présent être extraites de plus en plus aisément via des outils spécifiques d'extraction de données Web (*web scraping*) et de fouille de texte (*text mining*) (Munzert et coll., 2014).

À l'opposé, les données « formelles » sont produites par des agents qui disposent d'un certain savoir spécialisé – par exemple : un article de la presse quotidienne, une retranscription d'une émission radiophonique, un discours prononcé (et préparé) par un homme politique, une retranscription d'un débat parlementaire, un texte de manifeste électoral, un accord gouvernemental, un texte de loi, etc. Nombre de ces données sont de plus en plus accessibles par voie électronique, et il devient également possible d'accéder à beaucoup d'entre elles via des outils d'extraction de données Web et de fouille de texte.

Bien évidemment, l'ADT est également très prometteuse en raison de la diversité et la puissance croissante des logiciels de traitement de données conçus à cette fin. En la matière, on trouve bien sûr la grande diversité de logiciels d'analyse de données qualitatives assistée par ordinateur (*CAQDAS*) avec près d'une trentaine d'options disponibles à présent, parmi lesquelles on peut mentionner NVivo, MAXQDA, QDA Miner, Atlas.ti et Alceste. Tous ces logiciels permettent une grande diversité de traitements quantitatifs, et également des formes de visualisation riches et originales, comme plusieurs chapitres du présent ouvrage peuvent témoigner. Une richesse complémentaire est qu'il est possible, avec ces logiciels spécifiques, de gérer, de relier et d'analyser différents types de données « qualitatives », parmi lesquelles les données textuelles, mais aussi d'autres données sonores, visuelles, voire multimédias. Il faut aussi mentionner des solutions plus classiques comme l'exportation de fichiers déjà codés vers des logiciels statistiques (SPSS, Stata et bien sûr R) et des solutions de plus en plus souples et puissantes via l'environnement de programmation R (en particulier via RQDA). Mentionnons enfin les nouveaux types de traitement de données textuelles, comme la « *quantitative narrative analysis* » (Franzosi, 2010) basée sur des modèles relationnels et en réseaux, et dépassant donc des approches statistiques basées sur des postulats de linéarité et d'additivité.

DE SÉRIEUX PÉRILS

La richesse et la diversité des données qui se prêtent à l'ADT recèlent aussi bien des périls. Il semble que le recours à l'ADT présente au moins trois difficultés fondamentales, trois écueils auxquels tout chercheur fait potentiellement face.

Le premier écueil est le contexte de la production des « données ». Il nous a toujours semblé que l'usage du terme « données » (au sens de « *data* » en anglais) était malheureux en sciences sociales – car, en effet, les « données » ne sont pas données ; elles sont toujours *produites*, avec tous les biais, filtres et manipulations possibles en amont du traitement d'ADT proprement dit. C'est précisément ici que réside un des intérêts des approches ADT du type « analyse de discours » (*cf.* supra), qui par définition accordent une grande importance aux conditions de production du discours qui va générer les « données textuelles ».

Le deuxième écueil est que tout recours à l'ADT, en particulier dans ses approches plus quantitatives, requiert l'une ou l'autre forme de catégorisation, tout particulièrement l'élaboration de lexiques. Or le recours à des catégorisations automatisées (*cf.* supra) peut très vite générer des résultats qui font sens statistiquement, mais qui font nettement moins sens substantivement. Une bonne pratique pour éviter cet écueil est que le chercheur détermine lui-même certains paramètres de la catégorisation, c'est-à-dire qu'il borne en quelque sorte les opérations automatisées générées par le logiciel d'ADT.

Le troisième écueil, qui n'est pas spécifique à l'ADT, mais qui est un risque couru par toute analyse quantitative sur des données en grand nombre est celui de se lancer dans des opérations de « *data crunching* » très puissantes sur le plan statistique ou en termes d'analyse formalisée, et brassant des corpus de très grande taille, mais en perdant au passage le contact précieux avec la théorie. Ce risque devient de plus en plus fréquent à mesure que les logiciels disponibles gagnent en puissance. Dès lors, la question se pose de savoir quels résultats empiriques ont vraiment un sens théoriquement, et quels autres résultats ne constituent que des artefacts de l'analyse statistique.

CONSEILS D'UN VOISIN CRITIQUE ET AMICAL

Au final, l'ADT apparaît dans l'ensemble séduisante, comme le démontre assez brillamment Roberto Franzosi dans son ouvrage *From Words to Numbers* (2004). Elle est séduisante non seulement par la richesse de ses données et de ses outils (informatisés, en particulier), mais aussi parce qu'elle s'est jusqu'à présent développée dans un espace interdisciplinaire fertile et dépassant

les sciences sociales au sens strict du terme : ce champ d'innovation couvre aussi des domaines aussi divers que les statistiques fondamentales, la linguistique, l'informatique, la philosophie ou encore la théorie des réseaux. Voilà pourquoi l'ADT recèle encore certainement un énorme potentiel, entre autres dans les différentes sciences sociales et sciences humaines, dont la science politique.

Néanmoins, il est très facile de mal exploiter le potentiel de l'ADT. Voilà pourquoi, en considérant bien les potentiels et écueils de l'ADT, et en considérant également les « bonnes pratiques » méthodologiques en sciences sociales, je formulerais trois principaux conseils (version positive) ou trois principales mises en garde (version négative) aux chercheurs et chercheuses s'engageant dans l'exploitation de l'ADT, en particulier ceux et celles s'orientant vers les outils plus quantitatifs.

Mon premier conseil est de ne pas escompter que l'ADT apporte les réponses à l'ensemble des questions posées par une recherche donnée ; ceci est très peu probable ; la situation la plus fréquente est que l'ADT permettra de répondre à un type bien précis de question de recherche, et il faudra donc sans doute recourir à d'autres méthodes pour d'autres volets de la recherche. Ensuite, mon second conseil est de ne pas utiliser les logiciels d'ADT de manière rapide ou en suivant une logique de « pousse-boutons » ; ceci génèrera certes des résultats, mais ils seront nettement moins fertiles et pertinents que si l'usage du logiciel est précédé de moult précautions. Enfin, mon troisième conseil est de ne pas privilégier systématiquement les procédures automatisées, en particulier dans la construction des lexiques ; très souvent, une procédure au moins semi-manuelle permettra de nourrir les analyses de connaissances contextuelles que seul le chercheur peut posséder.

Somme toute, ces mises en garde sont très semblables à celles que je formulerais envers ceux et celles qui souhaitent exploiter les méthodes quantitatives, mais elles sont plus vigoureuses encore dans le domaine de l'ADT, au vu de la puissance des logiciels et de la nature particulièrement riche des données. Ceci ne doit cependant décourager personne d'utiliser ces outils extrêmement fertiles de l'ADT ; j'en appelle plutôt à un usage de ces méthodes qui soit à la fois bien raisonné et vigilant – comme cela devrait être le cas pour toute méthode empirique.

RÉFÉRENCES

Bauer, Martin W., George Gaskell et Nicholas C. Allum (2000), « Quality, Quantity and Knowledge Interests : Avoiding Confusions », dans Martin W. Bauer et George Gaskell (dir.), *Qualitative Researching with Text, Image and Sound. A Practical Handbook for Social Research*, London, Thousand Oaks et New Delhi, Sage.

De Meur, Gisèle et Benoît Rihoux (2002), *L'analyse quali-quantitative comparée (AQQC-QCA) : approche, techniques et applications en sciences humaines*, trad. Sakura Yamasaki, Louvain-la-Neuve, Academia-Bruylant.

Franzosi, Roberto (2004), *From Words to Numbers. Narrative, Data and Social Science,* Cambridge, Cambridge University Press.

Franzosi, Roberto (2010), *Quantitative Narrative Analysis,* Quantitative Applications in the Social Sciences, Thousand Oaks, Sage 175.

Groves, Robert M. (2011), « Three Eras of Survey Research », *Public Opinion Quarterly,* vol. 75, n° 5, p. 861-871.

Munzert, Simon, Christian Rubba, Peter Meiβner et Dominic Nyhuis (2014), *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*, New-York, Wiley.

Rihoux, Benoît (2013), « Recension de : Druckman, James N., Donald P. Green, James H. Kuklinski et Arthur Lupia (dir.) (2011), *Cambridge Handbook of Experimental Political Science*, Cambridge, Cambridge University Press », *Revue française de science politique,* vol. 63, n° 5, p. 948-949.

INTRODUCTION

Quelques repères pour appréhender l'analyse des données textuelles dans toute sa diversité

Pierre-Marc Daigneault et François Pétry

Le langage, qu'il soit sous forme écrite ou orale, est le principal médium de la politique (Grimmer et Stewart, 2013 ; Trimble et Treiberg, 2015 ; Coman et coll., 2016). Les décideurs politiques, les gestionnaires publics, les acteurs de la société civile et les citoyens ont en effet recours au langage pour faire entendre leur voix, c'est-à-dire pour se positionner face à leurs adversaires, pour défendre leurs intérêts et pour tenter d'influencer les orientations de l'État. Le langage est en outre essentiel aux journalistes et aux universitaires pour rendre compte de la politique. Celui-ci n'est toutefois jamais neutre : il véhicule de l'information, certes, mais aussi certaines valeurs et une manière d'interpréter notre monde. Par de savants effets de cadrage, les experts en marketing politique ne manquent d'ailleurs pas d'exploiter la charge normative et affective rattachée au langage pour tenter de convaincre les publics visés.

L'importance fondamentale du langage dans la vie politique signifie que les chercheurs ont accès à des données textuelles – écrites ou parlées – abondantes, riches et variées. Ainsi, les politologues mobilisent dans leurs travaux plusieurs sources de données textuelles : plateformes électorales, débats parlementaires, documents législatifs, énoncés de politique, rapports de recherche, communiqués de presse, articles de presse, entretiens et micromessages (« gazouillis » ou *tweets*). En facilitant la numérisation, le

partage, le traitement et l'analyse des textes, la révolution numérique a par ailleurs contribué, dans des proportions jusqu'ici inégalées, à accroître la disponibilité des données textuelles (Small et coll., 2014). L'abondance de ce type de données peut toutefois représenter un problème, en ce sens qu'il y a tout simplement trop de textes à analyser étant donné les ressources limitées dont disposent les chercheurs (Grimmer et Stewart, 2013). Si les méthodes automatisées peuvent être utiles dans ce contexte, encore faut-il connaître leur existence, être en mesure d'en évaluer les forces et faiblesses et savoir les utiliser de manière appropriée, ce qui n'est pas toujours le cas.

LA RAISON D'ÊTRE DE L'OUVRAGE

Face à notre constat de méconnaissance des principales méthodes de l'analyse textuelle en sciences sociales en général et en science politique en particulier, nous avons organisé l'atelier « Les idées, le discours et les pratiques politiques au prisme de l'analyse des données textuelles » à l'occasion du Congrès annuel de la Société québécoise de science politique (SQSP) qui s'est tenu du 19 au 21 mai 2016 à l'Université Laval à Québec. Le nombre et surtout la qualité des présentations de l'atelier ont renforcé notre volonté de nous lancer dans la réalisation de cet ouvrage collectif.

La visée de cet ouvrage est d'abord et avant tout pédagogique. Plus précisément, nous tentons d'atteindre trois objectifs : 1) présenter les principes d'un échantillon choisi d'approches, de méthodes et d'outils d'analyse de données textuelles et ce, tant du côté qualitatif que quantitatif ; 2) illustrer concrètement le fonctionnement de ces méthodes en les appliquant à des objets d'étude de la science politique et d'autres disciplines en sciences sociales ; 3) apprécier le potentiel et les limites de ces méthodes et, de manière générale, explorer les nouvelles avenues offertes par l'analyse textuelle.

Cet ouvrage a une fonction pédagogique, certes, mais il diffère du manuel à deux égards. D'une part, l'ouvrage n'aborde pas toutes les méthodes d'analyse de contenu disponibles à l'heure actuelle comme le ferait un manuel. D'autre part, il s'agit d'un ouvrage collectif dont les chapitres présentent des études complètes et originales qui vont bien au-delà des simples

illustrations méthodologiques qui caractérisent un manuel. Cet ouvrage s'adresse d'abord aux étudiants gradués et aux chercheurs en science politique et en sciences sociales intéressés par les méthodes d'analyse des données textuelles. Il s'adresse également aux chercheurs qui s'intéressent aux sujets et cas traités dans les différentes contributions (le cadrage des politiques de fin de vie, la corruption, l'assistance sociale, etc.).

APPRÉHENDER UN OBJET DISPARATE ET FRAGMENTÉ

Les méthodes d'analyse de données textuelles constituent un objet d'étude très fragmenté pour lequel un effort de cartographie s'impose. Nous mobilisons quatre dimensions pour appréhender cet objet.

1) Analyse de contenu ou de discours. La première dimension concerne la mise en opposition de deux grandes approches d'analyse textuelle, soit l'analyse de contenu et l'analyse de discours (p. ex., Tonkiss, 2004 ; Trimble et Treiberg, 2015 ; Coman et coll., 2016 ; Mace et Pétry, 2017 ; voir aussi la préface de Rihoux, dans cet ouvrage). L'analyse de contenu recoupe les méthodes qui visent à analyser de manière « scientifique » (systématique, rigoureuse et objective) un corpus donné (Franzosi, 2004). L'accent y est généralement mis sur le contenu manifeste d'un corpus, soit les messages dominants, explicites et directement accessibles aux chercheurs qui y sont véhiculés sous forme de mots, thèmes et arguments. Cette approche est fondée sur le postulat selon lequel « le discours reflète de manière plus ou moins neutre une réalité qui existe en soi, hors du langage » (Mace et Pétry, 2017, p. 84). Alors que, dans les années 1940 et 1950, les pionniers de cette approche (Abraham Kaplan et Bernard Berelson) insistaient sur le traitement quantitatif des données textuelles, la quantification n'est généralement plus considérée comme un attribut essentiel de l'analyse de contenu (Franzosi, 2004 ; voir aussi Leray et Bourgeois, 2016). De son côté, l'analyse de discours désigne beaucoup plus qu'une simple méthode d'analyse des données pouvant être utilisée dans tous les contextes. L'analyse de discours est en effet une approche fondée sur des postulats épistémologiques et théoriques constructivistes selon lesquels le contenu d'un texte n'est jamais neutre ni « donné » au chercheur (Fairclough, 2004 ; Mace et Pétry, 2017, p. 84). Par conséquent, l'analyse de discours, qui a des ambitions plus larges

que la simple analyse de contenu, est centrée sur l'interprétation du sens du contenu sous-jacent et implicite d'un texte :

> Dans une perspective compréhensive ou critique [...], elle s'attache à mettre en lumière la *dimension latente* du discours, c'est-à-dire les messages et valeurs implicitement véhiculés par un discours à travers son agencement (syntaxe), ses composantes linguistiques, métaphores, pronoms, figures de style comme la métonymie, l'hyperbole, les connotations, etc.), l'inscription dans son contexte sociohistorique, ainsi que les références à d'autres discours (intertextualité). (Coman et coll., 2016, p. 135-136)

Généralement de nature qualitative, l'analyse de discours donne parfois lieu à des applications quantitatives (Coman et coll., 2016). Dans sa variante critique, l'analyse de discours vise à mettre au jour les rapports de pouvoir politique ou de domination sociale qui sont véhiculés et reproduits par le langage (Fairclough, 2004).

2) Analyse manuelle ou automatisée. La seconde dimension de notre grille concerne la distinction entre les méthodes manuelles et automatisées d'analyse textuelle. Comme leur nom l'indique, les méthodes manuelles sont caractérisées par le fait que ce sont les chercheurs qui codent et analysent « à la main » les données textuelles. Quant aux méthodes automatisées, elles désignent les méthodes où les chercheurs confient une grande partie du travail d'analyse à un logiciel ou à un algorithme informatique. L'automatisation n'élimine toutefois pas totalement le rôle du chercheur. Par ailleurs, l'utilisation de l'analyse automatisée comporte à la fois des avantages et des inconvénients. Un avantage réside dans l'économie en temps et en argent rattachée à son utilisation. Il devient ainsi possible d'analyser à coût raisonnable des objets qui nécessitent l'analyse de très grands corpus de données textuelles. Un deuxième avantage est que cette démarche donne une lecture parfaitement fidèle : les résultats de plusieurs codages indépendants, effectués avec la même méthode, sur le même corpus par des opérateurs différents, concordent parfaitement. Autrement dit, il y a parfaite « reproductibilité » des résultats pour chacun des textes, peu importe qui, quand, où, et pourquoi les analyses sont effectuées. Un troisième avantage de l'analyse automatisée est l'absence de biais caractéristiques du codage manuel.

Le principal inconvénient de l'analyse automatisée est qu'elle repose sur une conception erronée du langage (Grimmer et Stewart, 2013, p. 2) qui ne prend pas ou très rarement en compte le contexte, l'expérience personnelle de la source, les figures de style, l'utilisation des symboles, etc. Il s'agit d'une analyse sommaire moins apte à capter la complexité que ne le serait l'analyse manuelle par des analystes formés à cet effet. Néanmoins, l'utilisation de l'une ou l'autre peut dépendre de la question de recherche. Les deux méthodes ciblent, d'une certaine façon, différents niveaux d'analyse. Pour nous en convaincre, Hart (2001) compare par analogie l'analyse par codage manuel à la perspective d'un policier travaillant dans les rues d'un quartier et l'analyse automatisée à un pilote d'hélicoptère survolant toute la ville. Les deux ont leur utilité particulière et amènent une perspective que l'autre n'a pas. Ainsi, pour paraphraser Grimmer et Stewart (2013, p. 2), les méthodes automatisées sont pertinentes dans la mesure où « elles augmentent et amplifient une lecture attentive et réfléchie des textes ».

3) Classification ou positionnement. La troisième dimension de notre grille porte sur les deux principales fonctions des méthodes d'analyse de contenu, soit la classification et le positionnement (Grimmer et Stewart, 2013 ; Mace et Pétry, 2017). Les méthodes automatisées de classification organisent les textes sur la base de catégories connues (logique déductive) ou inconnues (logique inductive). Les méthodes automatisées de classification avec des catégories connues nécessitent la création d'un dictionnaire d'analyse ou encore l'utilisation d'une méthode d'apprentissage automatique supervisé (*supervised machine learning : SVM*), ce qui implique au départ le codage manuel d'un corpus d'apprentissage. Par la suite, les analyses de réplication peuvent se faire de manière automatisée. L'approche par dictionnaire représente d'ailleurs l'approche la plus courante de classification pour des catégories définies pouvant être appliquée sans codage manuel. Celle-ci nécessite cependant la création du dit dictionnaire, ce qui peut s'avérer une tâche difficile (voir Grimmer et Stewart, 2013 pour une discussion *in extenso* des différentes formes d'analyse automatisée de contenu). Les méthodes automatisées de classification avec catégories inconnues consistent essentiellement à effectuer des regroupements (*clustering)* à l'aide d'algorithmes supervisés ou non, qui modélisent statistiquement et identifient

les catégories présentes dans les textes (p. ex., modèle thématique ou *topic modeling; Latent Dirichlet Allocation*; Blei et coll., 2003). Quant aux méthodes de positionnement, elles situent les textes, et par extension les acteurs politiques qui en sont à l'origine, à l'intérieur d'un espace de politique (*policy space*) ou une dimension d'intérêt pour le chercheur. Les méthodes automatisées de positionnement peuvent être supervisées (logique déductive), c'est-à-dire calibrées à l'aide de points de référence (p. ex., la méthode Wordscores de Laver, Benoit et Garry, 2003), ou non supervisées (logique inductive), c'est-à-dire qu'elles dépendent de la distribution statistique de l'utilisation des mots (p. ex., la méthode Wordfish de Slapin et Proksch, 2008).

4) Logique déductive ou inductive. La quatrième et dernière dimension de notre grille concerne le type de raisonnement, déductif ou inductif, qui sous-tend l'analyse textuelle (Crête et Imbeau, 1994, p. 34-36). Un raisonnement inductif commence par l'observation sur le terrain, vierge de tout *a priori* théorique, et tente de déceler des régularités dans ce qui est observé. Ces régularités seront ensuite confrontées aux travaux de recherche existants, afin de voir si les observations confirment ou remettent en cause les résultats des recherches précédentes. À l'inverse, le raisonnement déductif part d'une théorie à partir de laquelle on formule des attentes ou hypothèses. L'observation sur le terrain intervient après que les hypothèses aient été formulées afin de vérifier si ces hypothèses et la théorie dont elles sont issues sont conformes à la réalité des faits.

LES CONTRIBUTIONS

La présentation des contributions à cet ouvrage est structurée à partir des dimensions précédentes. L'ouvrage compte treize chapitres regroupés en quatre parties. La première partie de l'ouvrage est consacrée aux méthodes d'analyse de discours. Les trois chapitres qui la constituent s'inscrivent explicitement dans la mouvance de l'analyse de discours ou, à tout le moins, font preuve d'une ambition interprétative plus affirmée que les contributions des trois parties suivantes qui relèvent plutôt de l'analyse de contenu. Chacun de ces chapitres aborde l'analyse textuelle sous un angle discursif original et pertinent au vu de notre objectif de permettre aux chercheurs en sciences sociales de se familiariser avec une diversité de méthodes d'analyse textuelle. Il est intéressant

de noter qu'alors que les deux premières contributions reposent sur une méthode manuelle (qualitative), la troisième mobilise une approche automatisée (quantitative).

Olivier Turbide (chapitre 1) mobilise explicitement l'analyse de discours pour analyser le contenu latent de la communication politique. Il soutient qu'on assiste à une transformation des pratiques politiques médiatiques, soit le passage d'une parole politique monologale, impersonnelle et statique où le politicien constitue uniquement le support corporel au texte à livrer, à une parole coconstruite, mettant au jour des dynamiques d'adaptation et d'ajustement des comportements discursifs, vocaux et mimogestuels des politiciens en situation. Il vise à montrer que ces changements rendent nécessaire le recours à une approche qui tient compte du caractère coconstruit, interactionnel et multimodal du discours. Cette proposition méthodologique est illustrée à l'aide de la transcription détaillée du flux verbal et coverbal d'interactions médiatiques d'un extrait tiré d'une interview politique de *talk-show* diffusée à l'émission *Deux hommes en or*.

La contribution d'Isabelle F.-Dufour et de Marie-Claude Richard (chapitre 2) n'aborde pas un objet politique, mais elle est néanmoins du plus grand intérêt pour les chercheurs qualitatifs de toutes disciplines. Les auteures cherchent à évaluer dans quelle mesure deux méthodes « théorisantes » – la théorisation ancrée (TA) et la méthode générale inductive (MGI) – produisent des résultats équivalents. Elles évaluent en outre les limites et avantages respectifs de ces méthodes et dégagent les enjeux théoriques, éthiques et épistémologiques associés à l'analyse de données secondaires dans cette étude. Leur corpus est constitué de transcriptions d'entretiens semi-dirigés provenant d'une étude réalisée par d'autres chercheurs sur les facteurs facilitant ou nuisant à la stabilité des unions recomposées. Si les deux méthodes conduisent à des conclusions similaires concernant l'importance des rôles dans la recomposition familiale, la TA semble légèrement avantageuse par rapport à la MGI, car elle permet de distinguer des différences subtiles dans les propos des répondantes à partir desquelles une classification est développée. Cette différence ne peut toutefois être attribuée uniquement à la méthode employée : les résultats ne permettent donc pas de répondre de manière définitive à la question de la comparabilité des méthodes d'analyse. Les auteures suggèrent donc de sélectionner l'une ou l'autre de

ces méthodes sur la base de leurs avantages et inconvénients relatifs.

Adrien Thibault (chapitre 3) s'intéresse au concept de talent et défend la thèse selon laquelle ce concept est aujourd'hui l'un des schèmes de « l'idéologie dominante » visant à légitimer les élites françaises (Bourdieu et Boltanski, 1976). Dans une perspective historique, il procède à l'analyse lexicométrique du discours journalistique (articles de presse de trois quotidiens français) et politique (discours des présidents français). Il conclut à une augmentation dans l'utilisation du concept de talent depuis la fin de la Seconde Guerre mondiale et à une évolution sémantique du concept[1]. Alors que le talent était à l'origine associé à la politique comme pratique collective, il est désormais associé à des personnalités politiques et à une conception de la politique comme pratique individuelle. À l'aide de l'analyse factorielle des correspondances, Adrien Thibault démontre par ailleurs la diffusion du concept de talent dans différents champs sociaux (artistique, sportif, économique, etc.).

La deuxième partie de l'ouvrage porte sur les méthodes de classification manuelle ou automatisée des contenus à partir de catégories inconnues (logique inductive). Par contraste avec les trois chapitres précédents qui sont plutôt consacrés à l'analyse de discours, les trois chapitres de cette deuxième partie (tout comme ceux de la troisième et de la quatrième partie) s'inscrivent dans une perspective d'analyse de contenu. Les chapitres de la deuxième partie de l'ouvrage ont en commun d'analyser le contenu textuel par classification sur la base de catégories inconnues, en effectuant des regroupements permettant d'identifier des catégories dérivées des textes de manière inductive.

Dominic Forest, Frédérick Bastien, Ariane Legault-Venne, Olivier Lacombe et Hélène Brousseau (chapitre 4) expliquent la méthode de fouille de textes, de la constitution d'un corpus à l'interprétation des résultats, et mobilisent cette méthode pour analyser la communication des partis politiques (communiqués,

1. Bien que ce chapitre eût été à sa place en deuxième partie en raison de sa méthode automatisée de classification avec catégories inconnues, nous avons surtout retenu l'utilisation que fait l'auteur de cette méthode pour mettre au jour les structures de domination sous-jacentes à l'usage du concept de « talent » par les élites françaises.

billets de blogues et plateformes officielles) lors de l'élection fédérale canadienne de 2015. Dans un premier temps, ils utilisent un algorithme non supervisé afin d'expliciter le cadrage privilégié par chaque parti politique relativement à la thématique des réfugiés syriens qui était au centre de l'actualité à l'automne 2015. À l'aide d'une analyse factorielle des correspondances, ils démontrent que le Parti conservateur, au pouvoir au moment de lancer les élections, et les partis d'opposition ont interprété cet enjeu dans une perspective sécuritaire et humanitaire, respectivement. Dans un second temps, ils appliquent un algorithme supervisé à leur corpus afin d'évaluer dans quelle mesure les partis ont réussi à contrôler leur ordre du jour et à demeurer « *on message* » durant la campagne. Ils concluent que le Parti conservateur et le Parti libéral, le gagnant des élections de 2015, ont eu davantage de succès que les autres partis pour livrer un message cohérent aux électeurs.

Constantin Brissaud (chapitre 5) compare deux corpus, soit les rapports de l'Organisation de coopération et de développement économiques (OCDE) en matière de politiques de santé publiés entre 1992 et 2015 et des articles de presse des trois principaux quotidiens français. Grâce à une analyse lexicométrique, il vise à rendre compte du traitement médiatique des rapports de l'OCDE. Il s'agit en fait de la première étape d'une étude beaucoup plus ambitieuse visant à déterminer si et dans quelle mesure l'OCDE exerce une influence sur la réforme des systèmes de santé en France. La classification hiérarchique descendante et l'analyse par segments de textes caractéristiques révèlent que l'influence de l'OCDE sur le débat médiatique semble extrêmement limitée. En effet, la presse traite peu des propositions ou des enjeux de réforme de l'OCDE et, lorsqu'elle le fait, c'est uniquement en reproduisant certains indicateurs tels que la part totale du produit intérieur brut (PIB) consacrée aux dépenses de santé.

Sofia Wickberg (chapitre 6) s'intéresse à un sujet qui est hélas hautement d'actualité – la corruption – à partir d'une perspective constructiviste axée sur le discours[2]. Wickberg vise d'une part à

2. Bien que l'auteure se réclame explicitement d'une perspective d'analyse de discours, nous croyons que ce chapitre a davantage sa place ici, en deuxième partie. En effet, nous retenons surtout de ce chapitre la méthode manuelle

déterminer le cadrage de la corruption dans le discours médiatique français et d'autre part à identifier les principaux cadres interprétatifs utilisés par les journalistes couvrant cet enjeu. À partir d'une analyse de 155 articles de la presse française, elle identifie trois cadres d'action et quatre cadres rhétoriques. Ses résultats suggèrent deux conclusions principales. Premièrement, les journalistes ont tendance à représenter la corruption comme un problème d'éthique individuel, exagérant le rôle des acteurs par rapport aux potentielles racines structurelles de la corruption, ce qui contribue à réduire les solutions à une question de répression et de sanction. Deuxièmement, le traitement médiatique de la corruption focalise notre attention sur les conséquences de la corruption et tend à dramatiser la narration des événements pour rendre les récits plus captivants, ce qui présente le risque de susciter plus de cynisme que d'envie d'agir chez les lecteurs.

La troisième partie présente trois chapitres qui mobilisent la classification manuelle ou automatisée des contenus à partir de catégories connues, autrement dit, des catégories fixées à l'avance sur la base de théories générales. Les chapitres de la troisième partie adoptent donc tous un raisonnement déductif visant à tester des hypothèses explicites.

Lisa Birch et Sandra P. Escalera (chapitre 7) ont recours à l'analyse textuelle automatisée pour mieux comprendre le cadrage médiatique sur l'euthanasie et les politiques en fin de vie en Belgique, en France et au Québec. Le concept de cadrage médiatique au centre de ce chapitre est le même que celui qu'on retrouve dans le chapitre précédent par Sofia Wickberg. Toutefois, le raisonnement n'est plus vraiment le même ; d'inductif avec des catégories de classification inconnues au départ chez Wickberg, il devient déductif avec des catégories de classification connues au départ chez Birch et Escalera. L'hypothèse proposée est que les différences dans le cadrage médiatique de l'euthanasie ont contribué à orienter la nature des débats et à influencer certaines décisions législatives. Une autre différence majeure est que Sofia Wickberg procède à une analyse de données manuelles tandis que l'analyse est automatisée chez Lisa Birch et Sandra P. Escalera

d'analyse avec catégories inconnues utilisée par l'auteure pour identifier les cadres rhétoriques relatifs à la corruption.

qui font appel en particulier à la version française (Duval et Pétry, 2016) du *Lexicoder Sentiment Dictionary* de Young et Soroka (2012) pour faire l'analyse automatisée du ton positif ou négatif de la couverture médiatique de l'euthanasie. Il s'agit ainsi de la seule contribution de l'ouvrage qui s'attaque de manière explicite et systématique à la question de l'orientation positive ou négative du contenu à l'aide de l'analyse automatisée (voir cependant le chapitre de Thibault qui conclut, grâce à un examen des cooccurrences, que le concept de talent est utilisé pour glorifier ceux à qui il s'applique).

Jean Crête (chapitre 8) procède à une analyse exploratoire fort originale du contenu des constitutions de 193 États pour établir la relation entre divinité, monarchie et absence de conflit. Il propose plusieurs hypothèses de recherche issues de la théorie associant le concept de divinité à la présence de conflits, en particulier l'hypothèse selon laquelle les textes constitutionnels qui s'en remettent au divin sont moins associés à la notion de coopération que ne le sont les textes constitutionnels séculiers. Pour tester cette hypothèse, le contenu des constitutions est classé selon la fréquence de mention de mots préalablement définis dans des dictionnaires constitués par d'autres chercheurs. Un de ces dictionnaires contient les termes servant à identifier les références au divin, un autre contient les termes liés aux concepts de coopération. L'auteur conclut d'abord que les dieux sont fréquemment présents dans les constitutions – une majorité (62 %) d'entre elles présentent des caractéristiques de transcendance – et cette proportion augmente significativement pour les régimes monarchiques. Il arrive par ailleurs à des résultats positifs pour huit de ses neuf hypothèses, dont six passent le test de la signification statistique.

Mathieu Ouimet, Pascal Lalancette et Alexandre Racine (chapitre 9) analysent de manière systématique et rigoureuse le contenu des rapports des unités d'évaluation de technologies et modes d'intervention en santé (UETMIS) au Québec. Ils testent deux hypothèses issues de la théorie de l'agence qui cherchent à expliquer la variation du contenu de ces rapports. Alors que la première avance que la demande d'évaluation sera plus élevée pour les technologies déjà en usage que pour les technologies qui ne sont pas en usage, la seconde énonce que les recommandations seront plus favorables pour les technologies en usage. Les auteurs réalisent une analyse de contenu manuelle de 108 rapports

d'évaluation qui offre un faible soutien à la première hypothèse (demande accrue). Les résultats d'analyse sont en revanche concluants pour la seconde hypothèse (recommandation favorable) qui ne peut être rejetée par les auteurs.

La quatrième partie porte sur les méthodes manuelles et automatisées de positionnement. Trois des quatre chapitres qui la constituent ont la propriété d'être des analyses de contenu déductives tout comme ceux de la troisième partie, sauf qu'ils s'intéressent au positionnement des textes dans un espace de politique prédéfini plutôt qu'à leur classification.

François Pétry, Dominic Duval, Lisa Birch et Jean Crête (chapitre 10) utilisent la méthode du *Comparative Manifesto Project* (Volkens et coll., 2013), une méthode manuelle, afin de positionner le contenu idéologique des programmes électoraux des partis politiques canadiens sur un axe gauche-droite en fonction de catégories préétablies. Ils utilisent ensuite la même méthode pour positionner le contenu gauche-droite des promesses que les partis s'engagent à réaliser s'ils sont élus. La comparaison des deux corpus permet de tester l'hypothèse tirée de la théorie du mandat (Klingemann et coll., 1996) selon laquelle les partis sont fiables, c'est-à-dire qu'il y a cohérence entre leur idéologie, leurs engagements électoraux et leurs réalisations en la matière. Deux résultats clairs émergent de leurs analyses. Premièrement, hormis quelques variations mineures, les partis canadiens sont généralement fiables. Deuxièmement, un biais de mesure clair est identifié dans l'échelle gauche-droite du *Comparative Manifesto Project*, ce qui amène les auteurs à proposer diverses solutions pour le corriger.

Pierre-Marc Daigneault, Dominic Duval et Louis M. Imbeau (chapitre 11) font une analyse automatisée du contenu d'entretiens semi-dirigés afin de caractériser l'idéologie d'une réforme de politique sociale adoptée dans les années 1990 en Saskatchewan. Une étude qualitative antérieure de la même réforme les amène à poser l'hypothèse selon laquelle la réforme est positionnée au centre ou légèrement à droite en termes idéologiques. Cette hypothèse est testée à l'aide de Wordscores (Laver, Benoit et Garry, 2003), une méthode de positionnement fondée sur le recours à des points de référence établis a priori. Les textes de référence utilisés sont des textes universitaires caractéristiques de deux paradigmes opposés d'assistance sociale, soit le droit