Université de Montréal

Diversité et recommandation : une investigation sur l'apport de la fouille d'opinions pour la distinction d'articles d'opinion dans une controverse médiatique

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

Thèse présentée en vue de l'obtention du grade de Doctorat en sciences de l'information

Août, 2019

Université de Montréal

Université de Montréal
École de bibliothéconomie et des sciences de l'information, Faculté des arts et des sciences

Cette thèse intitulée

Diversité et recommandation : une investigation sur l'apport de la fouille d'opinions pour la distinction d'articles d'opinion dans une controverse médiatique

Présentée par
Marcela Baiocchi

A été évaluée par un jury composé des personnes suivantes

Sabine Mas
Présidente-rapporteur

Dominic Forest
Directeur de recherche

Juliette de Maeyer
Membre du jury

Mathieu Valette
Examinateur externe

Résumé

Les plateformes de consultation d'articles de presse en format numérique comme *Google Actualités* et *Yahoo! Actualités* sont devenues de plus en plus populaires pour la recherche et la lecture de l'information journalistique en ligne. Dans le but d'aider les usagers à s'orienter parmi la multitude de sources d'information, ces plateformes intègrent à leurs moteurs de recherche des mécanismes de filtrage automatisés, connus comme systèmes de recommandation. Ceux-ci aident les usagers à retrouver des ressources informationnelles qui correspondent davantage à leurs intérêts et goûts personnels, en prenant comme base des comportements antérieurs, par exemple, l'historique de documents consultés. Cependant, ces systèmes peuvent nuire à la diversité d'idées et de perspectives politiques dans l'environnement informationnel qu'ils créent : la génération de résultats de recherche ou de recommandations excessivement spécialisées, surtout dans le contexte de la presse en ligne, pourrait cacher des idées qui sont importantes dans un débat. Quand l'environnement informationnel est insuffisamment divers, il y a un manque d'opportunité pour produire l'enquête ouverte, le dialogique et le désaccord constructif, ce qui peut résulter dans l'émergence d'opinions extrémistes et la dégradation générale du débat.

Les travaux du domaine de l'intelligence artificielle qui tentent de répondre au problème de la diversité dans les systèmes de recommandation d'articles de presse sont confrontés par plusieurs questions, dont la représentation de textes numériques dans le modèle vectoriel à partir d'un ensemble de mots statistiquement discriminants dans ces textes, ainsi que le développement d'une mesure statistique capable de maximiser la différence entre des articles similaires qui sont retournés lors d'un processus de recommandation à un usager. Un courant de recherche propose des systèmes de recommandation basés sur des techniques de fouille d'opinions afin de détecter de manière automatique la différence d'opinions entre des articles de presse qui traitent d'un même thème lors du processus de recommandation. Dans cette approche, la représentation des textes numériques se fait par un ensemble de mots qui peuvent être associés, dans les textes, à l'expression d'opinions, comme les adjectifs et les émotions. Néanmoins, ces techniques s'avèrent moins efficaces pour détecter les différences entre les opinions relatives à un débat public argumenté, puisque l'expression de l'opinion

dans les discussions politiques n'est pas nécessairement liée à l'expression de la subjectivité ou des émotions du journaliste.

Notre recherche doctorale s'inscrit dans l'objectif de (1) systématiser et de valider une méthodologie de fouille d'opinions permettant d'assister l'identification d'opinions divergentes dans le cadre d'une controverse et (2) d'explorer l'applicabilité de cette méthodologie pour un système de recommandation d'articles de presse. Nous assimilons la controverse à un type de débat d'opinions dans la presse, dont la particularité est la formation de camps explicitement opposés quant à la façon de voir et de comprendre une question d'importance pour la collectivité. Notre recherche apporte des questionnements sur la définition d'opinion dans ce contexte précis et discute la pertinence d'exploiter les théories discursives et énonciatives dans les recherches de fouille d'opinions. Le corpus expérimental est composé par 495 articles d'opinion publiés dans la presse au sujet de la mobilisation étudiante du Québec en 2012 contre la hausse de droits de scolarité annoncée par le gouvernement de Jean Charest. Ils ont été classés dans deux catégories, ETUD et GOUV, en fonction du type d'opinion qu'ils véhiculent. Soit ils sont favorables aux étudiants et à la continuité de la grève soit favorables au gouvernement et critiques envers le mouvement de grève.

Sur le plan méthodologique, notre recherche se base sur la démarche proposée par les travaux qui explorent des techniques du champ de la linguistique du corpus dans la fouille d'opinions, ainsi que les concepts de la sémantique interprétative de François Rastier. Elle systématise les étapes de cette démarche, en préconisant la description des textes du corpus, pour relever et interpréter les mots spécifiques qui contrastent les types d'opinions qui devront être classés. Ce travail permet de sélectionner des critères textuels interprétables et descriptifs des phénomènes énonciatifs étudiés dans le corpus qui serviront à représenter les textes numériques dans le format vectoriel. La démarche proposée par ces travaux a été validée avec l'utilisation du corpus de presse constitué pour l'expérimentation. Les résultats démontrent que la sélection de 447 critères textuels par une approche interprétative du corpus est plus performante pour la classification automatique des articles que le choix d'un ensemble de mots dont la sélection ne prend pas en compte de facteurs linguistiques liés au corpus. Notre recherche a également évalué la possibilité d'une application dans les systèmes de

recommandation d'articles de presse, en faisant une étude sur l'évolution chronologique du vocabulaire du corpus de l'expérimentation. Nous démontrons que la sélection de critères textuels effectuée au début de la controverse est efficace pour prédire l'opinion des articles qui sont publiés par la suite, suggérant que la démarche de sélection de critères interprétables peut être mise au profit d'un système de recommandation qui propose des articles d'opinion issus d'une controverse médiatique.

Mots clés : systèmes de recommandation de presse, fouille d'opinions, opinion publique, genre textuel, sémantique interprétative, linguistique de corpus.

Abstract

Web-based reading services such as *Google News* and *Yahoo! News* have become increasingly popular with the growth of online news consumption. To help users cope with information overload on these search engines, recommender systems and personalization techniques are utilized. These services help users find content that matches their personal interests and tastes, using their browser history and past behavior as a basis for recommendations. However, recommender systems can limit diversity of thought and the range of political perspectives that circulate within the informational environment. In consequence, relevant ideas and questions may not be seen, debatable assumptions may be taken as facts, and overspecialized recommendations may reinforce confirmation bias, special interests, tribalism, and extremist opinions. When the informational environment is insufficiently diverse, there is a loss of open inquiry, dialogue and constructive disagreement—and, as a result, an overall degradation of public discourse.

Studies within the artificial intelligence field that try to solve the diversity problem for news recommender systems are confronted by many questions, including the vector model representation of digital texts and the development of a statistical measure that maximizes the difference between similar articles that are proposed to the user by the recommendation process. Studies based on opinion mining techniques propose to tackle the diversity problem in a different manner, by automatically detecting the difference of perspectives between news articles that are related by content in the recommendation process. In this latter approach, the representation of digital texts in the vector model considers a set of words that are associated with opinion expressions, such as adjectives or emotions. However, those techniques are less effective in detecting differences of opinion in a publicly argued debate, because journalistic opinions are not necessarily linked with the journalist's subjectivity or emotions.

The aims of our research are (1) to systematize and validate an opinion mining method that can classify divergent opinions within a controversial debate in the press and (2) to explore the applicability of this method in a news recommender system. We equate controversy to an opinion debate in the press where at least two camps are explicitly opposed in their understanding of a consequential question in their community. Our research raises

questions about how to define opinion in this context and discusses the relevance of using discursive and enunciation theoretical approaches in opinion mining. The corpus of our experiment has 495 opinion articles about the 2012 student protest in Quebec against the raise of tuition fees announced by the Liberal Premier Minister Jean Charest. Articles were classified into two categories, ETUD and GOUV, representing the two types of opinions that dominated the debate: namely, those that favored the students and the continuation of the strike or those that favored the government and criticized the student movement.

Methodologically, our research is based on the approach of previous studies that explore techniques from the corpus linguistics field in the context of opinion mining, as well as theoretical concepts of François Rastier's Interpretative Semantics. Our research systematizes the steps of this approach, advocating for a contrastive and interpretative description of the corpus, with the aim of discovering linguistic features that better describe the types of opinion that are to be classified. This approach allows us to select textual features that are interpretable and compatible with the enunciative phenomena in the corpus that are then used to represent the digital texts in the vector model. The approach of previous works has been validated by our analysis of the corpus. The results show that the selection of 447 textual features by an interpretative approach of the corpus performs better for the automatic classification of the opinion articles than a selection process in which the set of words are not identified by linguistic factors. Our research also evaluated the possibility of applying this approach to the development of a news recommender system, by studying the chronological evolution of the vocabulary in the corpus. We show that the selection of features at the beginning of the controversy effectively predicts the opinion of the articles that are published later, suggesting that the selection of interpretable features can benefit the development of a news recommender system in a controversial debate.

Keywords : news recommender systems, opinion mining, public opinion, genre, interpretative semantics, corpus linguistics.

Table des matières

Liste des tableaux

Liste des figures

Gèn: Mountains standing close together: the image of Keeping Still.
Thus the superior man does not permit his thoughts to go beyond his situation.

« The wise do not hold opinions.
They are aware of the needs of others. »
(Lao Tsu, Tao Te Ching)

Remerciements

Ce travail était motivé par le désir de stabiliser ma réponse à la complexité à laquelle je suis quotidiennement exposée. Que devrais-je penser ? Si la promesse d'un monde colonisé par des machines a voulu garantir aux hommes la dominance et la maîtrise, force est de constater l'ampleur de la menace de ce monde qui tente de nous défigurer et de nous affaiblir. Ce travail se veut une contribution par des voix calmes et murmurantes au milieu du vacarme ininterrompu de la machine. Quiconque possède un cœur silencieux pourrait bien les entendre.

Les premières lignes de ces remerciements sont dédiées à mon directeur de thèse, Dominic Forest, et à son appui inconditionnel et toujours enthousiaste. Merci de m'avoir motivée à poursuivre vers ce qui pourrait « avoir un sens » pour moi. Ce n'est pas par hasard que la recherche de sens m'a fait naviguer dans des îles lointaines habitées par des théories sémantiques, jamais vues auparavant. La confiance que vous avez démontrée en mes capacités a été fondamentale à mon arrivée au bout de ce voyage.

Je suis également reconnaissante envers le comité de recherche, composé des professeures Audrey Laplante et Lyne Da Silva, qui ont posé les questions que je devais me poser moi-même. Grâce à votre aide et votre expertise, j'ai réussi à fournir quelques réponses. Je remercie également la professeure Michèle Hudon, qui m'a apporté en mains propres l'article scientifique qui fut le pivot de toute la réflexion de cette thèse doctorale. Au personnel

de l'EBSI qui m'a accompagnée pendant toutes ces années, je voudrais aussi exprimer ma reconnaissance pour le support toujours accordé.

Je tiens également à souligner la contribution de l'ancienne société Cedrom-SNI et de son ancien directeur, François Aird, qui m'ont gentiment accordé les droits des articles de presse qui font l'objet de mes expériences. Je remercie aussi le soutien de Thibault de La Grange, de Monique Perron, ainsi que le travail de mes collègues Mehdi Zmouli et Audrey Larivière pour les traitements effectués sur les données.

Pour tous les moments où je ne voyais pas la lumière au bout du tunnel, je remercie ceux qui m'ont montré ma propre étincelle : Pierre Girard et Marília Canavarros Girard, ma famille brésilienne-québécoise dans le pays de l'hiver, et Joseph Castel, *landsman* et *ghost writer* préféré.

À mes chers parents, Ivan Baiocchi Filho et Willene Carvalho Baiocchi, pour le soutien sans restriction, à tous les niveaux. Ce travail vous rend hommage.

À mon cher Mestre Gabriel, qui me guide dans le chemin de la science supérieure.

Et à Dieu, la grande force motrice, créatrice, qui m'a fait esprit et souffle vital, dotée de la capacité du verbe, que j'exerce ici avec amour et humilité.

Marcela Baiocchi

6 août 2019

Introduction

La lecture de l'actualité en ligne représente une des activités les plus importantes effectuées sur le Web. Selon les données publiées par *Statistique Canada* en 2016 sur l'utilisation des médias dans le pays (Statistique Canada, 2016), nous assistons depuis 2003 à un déclin de 18 % de la consommation de journaux imprimés. Corrélativement, ce même sondage a observé une croissance du nombre de personnes qui utilisent le Web pour suivre l'actualité, qui a augmenté de 30 % en 10 ans, soit entre 2003 et 2013. Plus récemment, la *Reuters Institut* a réalisé un autre sondage auprès de 70 000 personnes provenant de 36 pays, constatant que la disponibilité en ligne de contenus journalistiques gratuits a contribué à la croissance du nombre d'usagers qui utilisent le Web pour suivre l'actualité (Newman et coll., 2017).

Le déploiement de plateformes de recherche et de consultation d'articles de presse, communément appelés « agrégateurs de nouvelles »[1], atteste l'importance de ce nouveau lectorat de contenus numériques. L'agrégateur de nouvelles est une application qui génère un sommaire d'actualités à l'aide d'un agent automatique, regroupant des articles provenant de plusieurs sources d'information dans le monde. Sur le Web, les services gratuits *Google Actualités* et *Yahoo! Actualités*, créées en 2002 et 2011 respectivement, se trouvent parmi les agrégateurs les plus populaires, comportant des milliers de sources d'information journalistique. Il existe également les plateformes d'accès privées telles que les canadiennes *Eureka.cc*[2], avec environ dix mille sources d'information, et *Infomart*, totalisant plus de cinq mille sources.

---

[1] Le grand dictionnaire terminologique du Québec définit le terme « agrégateur de contenu » de la manière suivante : « Logiciel ou application Web qui permet à l'internaute de s'abonner à des fils de syndication, de recevoir automatiquement, regroupés dans une même fenêtre, le nouveau contenu des fils répertoriés, provenant de plusieurs sources, et de le lire dès qu'il est disponible » (Office de la langue française [OLF], 2013). Nous utilisons aussi le mot « application » pour nous référer aux agrégateurs de nouvelles.

[2] Informations obtenues à partir du site http://www.eureka.cc/.

Dans le but d'aider les usagers à s'orienter parmi la multitude de sources d'information, les agrégateurs de nouvelles intègrent des systèmes de filtrage personnalisés qui permettent de générer des recommandations de contenus susceptibles de les intéresser. Nous les désignons ici comme systèmes de recommandation d'articles de presse (SRAP). Ces systèmes filtrent les articles présents dans la base de données en fonction des préférences personnelles des usagers. Les préférences peuvent être explicitement exprimées par les usagers lorsqu'ils configurent des paramètres de personnalisation dans l'application, ou peuvent être inférés à partir des actions qu'ils effectuent : les clics, les évaluations (*likes),* l'historique d'articles consultés et le partage de contenus sont parmi les actions utilisées qui peuvent aider à inférer leurs préférences, lesquelles sont converties dans une représentation informatique qui sert de base à la constitution des profils d'usagers. Le but du SRAP est de prédire de nouveaux articles susceptibles d'intéresser un profil d'usager donné. Les recommandations sont généralement affichées sous la forme d'une liste d'articles non lus avec l'intitulé « Voir aussi » ou « Articles recommandés » dans l'interface des agrégateurs.

Une grande partie des SRAP reposent sur l'analyse du contenu de textes numériques et utilisent le modèle vectoriel de Salton et McGill (1983). Il s'agit d'un modèle algébrique destiné à donner une représentation des textes en fonction des mots extraits de ces derniers. Avec le modèle vectoriel, chaque article présent dans une collection est représenté par un vecteur $v = (m_{1j}, m_{2j}..., m_{nj})$ contenant n mots, où m est un mot et j indique la présence ou l'absence de ce dernier[3]. Certains traitements statistiques ou linguistiques sont appliqués pour filtrer le vocabulaire des articles et pour obtenir les mots les plus discriminants, que nous appellerons ici « traits discriminants ». Dans les SRAP, le modèle vectoriel peut être également employé pour fournir une représentation informatique du profil de l'usager, en fonction de l'ensemble des traits discriminants extraits des articles qu'il a consultés dans le passé. La recommandation s'effectue généralement au moyen de calculs qui détectent le niveau de similarité entre les vecteurs des articles de la base qui n'ont pas été encore consultés et le vecteur du profil de l'usager (Lei et coll., 2011).

[3] La valeur *j* peut représenter une mesure qui rend compte de la fréquence absolue du mot dans l'article en question, ou encore de sa fréquence pondérée (par exemple, la fréquence relative).

Les recommandations générées dans les SRAP s'avèrent pertinentes pour la consultation de nouvelles sur le Web permettant aux usagers de filtrer le flux d'information selon leurs intérêts personnels et de découvrir de nouveaux contenus. Cependant, plusieurs chercheurs manifestent leur inquiétude à l'égard de l'impact des SRAP sur la diversité d'idées et de perspectives politiques dans l'environnement informationnel. Ils affirment que la génération de recommandations excessivement personnalisées, surtout dans le contexte de la presse en ligne, pourrait engendrer l'homogénéisation de contenus dans ces environnements (Mutz et Young, 2011 ; Parisier, 2011 ; Van Alstyne et Brynjolfsson, 2005). Le filtrage aurait pour effet de maximiser l'importance de certains types de contenus, qui sont similaires à ceux déjà consultés dans le passé, et de confiner les usagers à des lectures qui sont plus en accord avec leurs points de vue politiques ou idéologiques.

> With the customized access and search capabilities of IT, individuals can focus their attention on career interests, music and entertainment that already match their defined profiles, and they can arrange to read only news and analysis that align with their preferences. Individuals empowered to screen out material that does not conform to their existing preferences may form virtual cliques, insulate themselves from opposing points of view, and reinforce their biases. (Van Alstyne et Brynjolfsson, 2005, p. 865-866)

> To the extent that a person goes to the same news website on an ongoing basis for information about the news of the day, that site may track the content he/she accesses and then use that information in choosing what is made most available on the page. However, unless different stories within a single website have different partisan angles, it is unlikely to prioritize one story over another on the basis of partisanship. Recommender agents are far better at inferring users' topics of interest than their partisanship. (Mutz et Young, 2011, p. 1032-1033)

La question de la diversité de perspectives dans les médias est intimement liée à la production d'un espace public démocratique et respectueux de la liberté d'expression des individus. Dans un contexte où les sociétés deviennent de plus en plus complexes en conséquence du développement technologique et scientifique, il est important que les médias puissent donner lieu à un débat qui amène à la confrontation de différentes perspectives, de manière à permettre aux nombreux intervenants de la société à se faire entendre. Mais il faut également que les technologies servant de support à l'accès à l'information soient capables

d'exposer ces perspectives de manière équitable. Puisque les agrégateurs et les SRAP sont de plus en plus présents dans l'environnement informationnel, la diversité dans la recommandation de contenus constitue un enjeu fondamental.

L'importance de ces questions amène à envisager un SRAP qui fonctionne comme un « modérateur » pour les individus lorsqu'ils consultent les agrégateurs de nouvelles pour s'informer des discussions publiques courantes. Au lieu de prédire les préférences des usagers et de recommander des contenus qui correspondent à ces préférences, ce système devrait les amener à confronter leurs penchants à certaines positions, en recommandant des contenus avec d'autres perspectives divergentes. Il peut sembler à première vue que ce type de recommandation rendrait plusieurs personnes inconfortables. Mais dans un contexte où le besoin informationnel est de comprendre les enjeux impliqués dans une discussion d'intérêt public, ce système pourrait permettre une lecture équitable et beaucoup plus informative.

Partant de la réflexion sur la nécessité de diversité dans le développement des SRAP, nous avons posé la question suivante : comment détecter automatiquement des articles qui sont comparables sur le plan du contenu, mais qui divergent quant aux opinions qu'ils véhiculent ?

D'une part, la question circonscrit un type de production textuelle commun dans la presse contemporaine : l'article d'opinion. À l'inverse de l'article du type nouvelle, l'article d'opinion s'inscrit dans le but communicatif explicite de présenter une opinion à un lecteur sur un thème précis, en offrant une vision approfondie de la question débattue.

D'autre part, la question nous amène à considérer les modèles de recommandation qui reposent sur l'analyse automatique de textes et sur le modèle vectoriel. Or, les textes qui véhiculent des opinions différentes sur un thème partagent une grande partie du vocabulaire se référant au thème traité, ce qui rend difficile le choix des termes discriminants qui peut capter la différence d'opinions. Cependant, il y a dans la production textuelle d'un article d'opinion des procédés linguistiques, parfois très évidents et parfois très subtils, qui permettent de dégager le projet persuasif de son auteur et qui peuvent être propres à l'opinion qu'il défend : la manière dont l'auteur se légitime, le vocabulaire qu'il utilise et les stratégies argumentatives

qu'il met en œuvre nous offrent des éléments qui peuvent être exploités pour donner une représentation vectorielle optimale et distinctive du type d'opinion véhiculée.

Le projet doctoral que nous proposons aborde, d'une part, la nécessité grandissante d'approfondir la diversité dans le développement des SRAP, et d'autre part, les méthodes d'identification automatique d'opinion. Il propose de systématiser une démarche de classification d'articles d'opinions à partir d'une analyse préalable du corpus constitué, qui explore les modes de structuration textuelle de l'opinion dans les textes. Il propose également d'étudier le bien-fondé de cette démarche méthodologique dans une perspective de recommandation d'articles d'opinion. Dans cette recherche, nous ne proposons pas de développer une méthode de recommandation, mais de créer un cadre expérimental permettant d'évaluer les observables linguistiques qui peuvent démontrer une meilleure efficacité à discriminer les articles d'opinion en fonction de l'opinion qu'ils véhiculent.

1. Justification empirique et théorique de la recherche

L'intérêt de parvenir à une plus grande diversité dans la recommandation d'articles de presse a fait l'objet d'un certain nombre de travaux proposant l'emploi de techniques de fouille de textes (Abbar et coll., 2013 ; Desarkar et Shinde, 2014 ; Li et coll., 2011 ; Shi et coll., 2012 ; Vargas et coll., 2014). Ces techniques, qui relèvent du domaine du traitement automatique des langues (TAL) de l'intelligence artificielle (IA), sont employées pour détecter le degré de différence (de l'anglais *dissimilarity*) entre l'ensemble d'articles qui sont recommandés à un usager, en utilisant comme base le modèle vectoriel. L'objectif des techniques de diversification est d'éviter de créer une liste de recommandation trop homogène, avec des articles qui sont trop similaires et répètent les mêmes informations. Les méthodes de diversification proposent ainsi de générer une liste d'articles à recommander qui, tout en étant pertinente à un usager donné (en regard de ses préférences personnelles), peut présenter une certaine différence relativement au contenu véhiculé.

D'autres études sur la recommandation d'articles de presse intéressés à la question de la diversité ont exploré les techniques de fouille d'opinions, lesquelles sont proposées pour distinguer les articles sur une même thématique en fonction du type d'opinion véhiculée (Kawai et coll., 2007 ; Zhang et coll., 2009). Dans ces systèmes, l'opinion d'un article est associée au vocabulaire subjectif employé dans les textes et les articles sont représentés dans le format vectoriel en fonction de ce vocabulaire. Les mots exprimant de la subjectivité, comme les sentiments et les émotions, sont utilisés pour créer les vecteurs des articles et des profils des usagers. Le processus de recommandation proposé par ces études vise à fournir des recommandations plus diversifiées en faisant intervenir une mesure de distance entre les vecteurs des articles et les vecteurs des profils des usagers, qui sont constitués par l'historique d'articles consultés.

Liu (2012) définit la fouille d'opinions comme un champ de recherche voué à l'analyse automatique d'opinions, de sentiments, d'évaluations, d'attitudes et d'émotions exprimés par les individus dans le langage écrit. Le terme « fouille d'opinions » est également utilisé comme synonyme d'analyse de sentiments (Liu, 2012 ; Piryani et coll., 2017). L'essor de la fouille

d'opinions date de quelques années et son intérêt est attesté par les réalisations d'ateliers et de campagnes d'évaluation — tels que DEFT 2007, DEFT 2009, DEFT 2015 et BlogTrack 2008[4] — ainsi que par de nombreuses publications dans des revues scientifiques (Piryani et coll., 2017 en fournit une ample revue). Les processus et les étapes impliquées dans la fouille de textes sont très similaires à ceux de la fouille d'opinions, ce qui nous amène à considérer cette dernière comme une spécialisation de la fouille de textes. Par contre, la fouille d'opinions se distingue par le type d'observable linguistique qui est privilégié dans l'analyse : tandis que la plupart des applications en fouille de textes s'intéressent au lexique présent dans les textes, dans une perspective d'extraction de concepts pour assister l'identification thématique, la fouille d'opinions privilégie les mots qui ont un apport subjectif pour détecter le type d'opinion véhiculée au sein d'un thème délimité.

Si la voie de recherche proposée par la fouille d'opinions pour le développement des SRAP semble intéressante d'un point de vue théorique en fonction des questionnements linguistiques valables qu'elle soulève, il faut néanmoins qu'elle interroge également la définition de l'opinion journalistique. Par exemple, la méthode de recommandation proposée par Kawai et coll. (2007) et Zhang et coll. (2009) préconise de représenter l'opinion des journalistes par les émotions exprimées dans les textes. Le processus de recommandation vise à suggérer des articles qui s'opposent aux patrons détectés dans le profil : si dans le profil de l'usager prédominent des articles qui véhiculent des mots subjectifs associées à la tristesse, le système permet de recommander des articles véhiculant des émotions associées à la joie. Le résultat obtenu par ces techniques sur la recommandation est limité, puisque la recommandation ne permet pas de distinguer les articles en fonction des opinions qu'ils véhiculent sur un thème donné, mais simplement de distinguer les bonnes des mauvaises nouvelles :

[4] DEFT - DÉFI Fouille de Textes (http://deft.limsi.fr); Text Retrieval Conference (http://trec.nist.gov/); TALN & RECITAL – 17e conférence sur le Traitement Automatique des Langues Naturelles (http://www.atala.org/TALN-2009-RECITAL-2009-17e).

> For example, a search using the keywords Iraq and terror with existing news portal sites will return articles about Iraq with such topics as suicide bombing in Iraq, which are likely to create a sad sentiment. A system that can recommend, for example, articles that create a happy sentiment, such as released hostage in Iraq, should thus be useful. (Kawai et coll., 2007, p. 613)

Les discussions et les débats dans la presse sont jalonnés de disputes et se déroulent plus généralement autour de la défense de valeurs qui ne sont pas nécessairement associées à l'expression émotionnelle des individus. Réduire l'opinion d'un texte journalistique en fonction des émotions qu'une nouvelle peut susciter chez le lecteur, c'est faire fi de la complexité du phénomène par lequel les opinions sont produites et diffusées dans les médias.

D'emblée, une distinction entre les différents genres journalistiques s'avère nécessaire, ce qui n'est généralement pas fait dans les travaux en fouille d'opinions (Eensoo et coll., 2011), ni dans les recherches qui utilisent la fouille d'opinions pour diversifier les recommandations des SRAP. Un éditorial ou une lettre de lecteur défendent une opinion plus explicitement qu'une nouvelle. Les nouvelles sont associées au rapport d'événements (Grosse, 2001) et les opinions qu'elles véhiculent sont souvent les paroles des tiers que le journaliste fait témoigner. De même, les critiques de produits ont un vocabulaire beaucoup plus subjectif et esthétisant qu'un éditorial qui argumente pour la défense d'un point de vue. Le premier privilégiera l'évaluation des qualités du produit en question, tandis que le second va privilégier le recours à des valeurs ou à des faits qui viennent justifier ou appuyer son opinion.

Représenter l'opinion d'un texte par ses mots subjectifs peut s'avérer une méthode efficace pour certains genres, comme les critiques, mais limitante pour d'autres, comme les articles d'opinion de types « tribune »[5], où l'exposition d'arguments n'est pas obligatoirement liée aux impressions subjectives de l'auteur. Ceci n'est pas pour autant considéré lorsque les études sur la détection automatique d'opinions proposent une représentation informatique de textes : l'utilisation de mots subjectifs pour représenter ces derniers dans le format vectoriel est largement proposée comme méthode à privilégier (Liu, 2012).

[5] « Rubrique consacrée à la publication d'articles n'engageant pas la responsabilité du journal, mais celle des seuls auteurs qui les rédigent ». (Office de la langue française [OLF], 2013).

À ce stade, il est prudent de remettre en cause la validité des traitements informatiques qui reposent uniquement sur l'analyse de mots subjectifs pour caractériser les opinions dans les textes. Cela implique de remettre en cause la conceptualisation adoptée par les recherches de fouille d'opinions de ce qui constitue l'expression linguistique d'une opinion. Il devient clair qu'une conceptualisation fortement axée sur les expressions subjectives et émotives est loin de couvrir tous les cas où l'opinion journalistique se fait présente.

De nouvelles voies de recherche envisagent la dimension communicationnelle et discursive de l'expression de l'opinion et proposent l'adoption d'un cadre théorique linguistique dans les recherches, en particulier les travaux réalisés par Valette (2004), Eensoo et Valette (2012, 2014a, 2014b, 2015) et aussi par Vernier et coll. (2009a ; 2009b). Ces travaux remettent en question la place accordée à l'analyse exclusive du vocabulaire subjectif et défendent la nécessité d'analyser les textes du point de vue de l'énonciation, laquelle est définie, dans la lignée de Benveniste (1967), comme la façon dont les locuteurs s'approprient le langage et organisent les énoncés et les productions textuelles (Eensoo et Valette, 2012). L'énonciation constitue dans ce sens une empreinte significative du locuteur sur le texte. Cette empreinte est repérable par des marques linguistiques qui rendent clair son positionnement à l'égard de l'interlocuteur, du monde qui l'entoure et de son propre propos.

En s'appuyant sur cette perspective énonciative et sur les concepts théoriques de la sémantique interprétative de François Rastier (Rastier, 1987, 2001 ; Rastier et coll., 1994), Eensoo et Valette (2012, 2014a, 2014b, 2015) proposent des méthodes d'exploitation du corpus pour l'élaboration de critères textuels[6] destinés à la représentation vectorielle des textes pour la fouille d'opinions. L'objectif est de sélectionner un ensemble de critères textuels qui soient capables d'expliciter les phénomènes linguistiques qui distinguent le positionnement énonciatif des différents types d'opinions dans le corpus. Ainsi, au lieu de choisir des traits discriminants liés au vocabulaire subjectif, ils proposent des techniques pour aller chercher

[6] Afin de maintenir la terminologie proposée par ces auteurs, nous proposons la distinction entre le terme « traits discriminants » et « critères textuels », réservant l'emploi de ce dernier seulement dans le contexte où les traits discriminants choisis pour représenter les textes dans le format vectoriel sont faits à partir d'un processus de sélection par la méthode proposée par ces auteurs.

dans les corpus des critères textuels qui ont un lien sémantique, et qui décrivent comment le matériel linguistique est utilisé par les locuteurs pour défendre une opinion dans un texte, à l'intérieur d'une situation de communication donnée. Les auteurs explorent en particulier le présupposé théorique de la sémantique interprétative selon lequel un texte présente une interaction particulière entre trois « composantes sémantiques » [7] qui peuvent renseigner sur ce positionnement énonciatif : 1) la thématique, qui a trait aux contenus abordés, 2) la dialectique, qui a trait à l'organisation temporelle et argumentative et 3) la dialogique, qui est liée à l'image que l'énonciateur se fait du texte et à sa prise d'attitude à l'égard du contenu de son énoncé. Le travail d'exploitation du corpus vise à valider l'hypothèse selon laquelle le choix de critères ayant un lien sémantique et interprétable à l'égard des composantes est plus efficace pour les tâches de classification automatique.

L'approche de fouille d'opinions fondée sur la sémantique interprétative propose un ensemble de méthodes et de calculs qui permettent de repérer dans les corpus des critères textuels compatibles avec les composantes sémantiques. Ces critères sont de natures variées : il peut s'agir de substantifs, mais aussi de pronoms, d'adjectifs, de déterminants, d'adverbes, de conjonctions, de préposition et même de ponctuation. La démarche de fouille d'opinions basée sur la sélection de critères démontre une très bonne performance, et ce, dans des corpus portant sur des thématiques variées, qu'il s'agisse de détection de textes racistes et antiracistes (Valette, 2004), d'analyse de sentiments de récits personnels portant sur la santé (Eensoo et Valette, 2014b) ou de détection d'articles de presse xénophobes à l'endroit d'une communauté ethnique (Eensoo et Valette, 2014a). Le succès de ces expérimentations et la pertinence des considérations théoriques à l'égard du problème de recherche posé motivent l'exploration de cette voie de recherche.

Les travaux de Valette (2004) et Eensoo et Valette (2012, 2014a, 2014b, 2015) proposent des calculs particuliers relevant de la textométrie pour une étude préalable du corpus visant à repérer les critères textuels pour la classification automatique de textes d'opinion. La première étape de cette méthode consiste à constituer des sous-corpus contenant les textes qui représentent chaque type d'opinion à classer. Les calculs textométriques mettent

[7] Ce concept est proposé dans le cadre de la sémantique interprétative et sera expliqué en détail dans le chapitre 3.

en contraste les sous-corpus et font émerger des textes des observables linguistiques variés qui sont spécifiques à chacun. Ces observables sont par la suite analysés et qualifiés comme critères textuels, en fonction d'un travail interprétatif entamé par un analyste. Ce dernier considère certains paramètres statistiques de sélection et regarde les contextes dans lesquelles les mots apparaissent. Cette analyse permet de repérer les critères compatibles avec les composantes sémantiques et d'expliciter les critères textuels qui relèvent les différences entre le positionnement des locuteurs.

La textométrie fait partie du champ général de la linguistique de corpus et englobe un ensemble de méthodes pour assister l'interprétation des corpus numériques. Les affinités entre la sémantique interprétative et la textométrie ont été soulignées par Pincemin (2012a) dans un article où l'auteure fait état de plusieurs recherches articulant les apports de la théorie de Rastier et les calculs fondateurs de la textométrie, dont les détails seront expliqués dans cette recherche. Cependant, il ne faut pas confondre la linguistique de corpus avec la textométrie. Dans son article, Pincemin explique que ces deux disciplines ont en commun le fait de porter les investigations sur des corpus numériques, mais que la linguistique de corpus s'intéresse à la description de la langue, alors que la textométrie s'intéresse davantage à l'élaboration de données pour le développement d'applications.

> L'analyse statistique des données textuelles (ADT), ou Textométrie est un ensemble particulier de pratiques relevant du champ général de la linguistique de corpus. Elle comprend des traitements statistiques (analyse factorielle des correspondances, spécificités fondées sur le modèle hypergéométrique, etc.) et des outils de visualisation des corpus (nuages de mots, histogrammes, etc.) et documentaires (concordanciers) destinés à l'aide à l'interprétation des textes. (Eensoo et Valette, 2015, p.3).

Ancrée dans l'étude des choix linguistiques opérés par des locuteurs réels, la textométrie ne prétend pas formuler un ensemble de règles directement applicables à l'analyse automatique de textes. Elle propose des méthodes permettant de découvrir les observables linguistiques qui peuvent être qualifiées de pertinentes pour une application donnée, puisqu'elles permettent de révéler les caractéristiques de l'élaboration textuelle et le positionnement des locuteurs dans les corpus. Néanmoins, cette dépendance du corpus et des démarches interprétatives pour le développement d'applications informatiques soulève des questions importantes quant à la généralisation et la reproductibilité des observations de ces

méthodes (Jacques, 2005). Cela ne permet pas de conclure qu'une démarche de sélection de critères est pertinente pour des contextes d'application qui exigent un traitement plus rapide des textes numériques, comme les SRAP. Notre recherche propose de contribuer à l'avancement des études sur cet aspect en s'appuyant sur la démarche méthodologique proposée par Eensoo et Valette (2012, 2014a, 2014b, 2015) dans le domaine de la fouille d'opinions.

2. Objectifs de recherche

L'objectif général de notre recherche est de *systématiser et de valider une démarche méthodologique de fouille d'opinions basée sur la textométrie pour l'identification de textes véhiculant des opinions divergentes dans une controverse et d'explorer, par une analyse de la progression chronologique du vocabulaire du corpus, l'applicabilité des résultats observés dans le développement des SRAP*. Il s'agit d'un travail expérimental et quantitatif, inspiré d'autres expérimentations dans le champ de la fouille d'opinions réalisées par Valette (2004) et Eensoo et Valette (2012, 2014a, 2014b, 2015) et qui portera sur un ensemble d'articles d'opinion sur la grève étudiante de 2012 au Québec contre la hausse des frais de scolarité.

Les objectifs spécifiques de la recherche sont les suivants :

1. À partir d'une analyse textométrique de deux sous-corpus d'articles d'opinion dont les opinions sont opposées, sélectionner les critères textuels qui différencient les deux parties en conflit sur le plan thématique, dialectique et dialogique.

2. Évaluer les différentes catégories de critères textuels issus d'une analyse textométrique dans le cadre d'une application de classification automatique, afin de valider la démarche de sélection de critères pour distinguer un article selon la position qu'il défend à l'intérieur d'une controverse.

3. Effectuer des analyses textométriques sur différentes périodes du corpus et explorer la progression chronologique de critères textuels qui démontrent une meilleure performance pour la tâche de classification automatique.

Le premier objectif cherche à analyser les textes, organisés dans des sous-corpus, par le biais de calculs textométriques et à sélectionner des critères interprétables qui les différencient. Il cherche aussi à élucider les stratégies d'élaboration textuelle qui sont caractéristiques de l'opinion défendue pour les auteurs situés dans un camp et dans l'autre de la controverse et à catégoriser les critères textuels selon qu'ils se rattachent au plan thématique, dialectique ou dialogique.

Le second objectif consiste à créer une approche de classification automatique paramétrée en fonction des différentes catégories de critères textuels issus de l'analyse textométrique. Nous allons utiliser des mesures d'évaluation classiques (rappel et précision) pour valider la performance de notre démarche pour l'identification de textes d'opinion, et ce pour chaque catégorie de critères textuels. L'objectif est de tester l'efficacité des critères sélectionnés par la méthode textométrique pour distinguer les articles d'opinion organisés dans le corpus.

Les deux premiers objectifs réfèrent au déploiement de la démarche méthodologique de fouille d'opinions qui sera systématisée dans le cadre de notre recherche.

Avec le troisième objectif, nous voulons vérifier s'il est possible d'identifier les critères textuels les plus performants pour la tâche de classification au début de la controverse. Cela nous permettra de nous prononcer sur l'applicabilité de la démarche de sélection de critères textuels dans le développement des SRAP.

3. Questions de recherche

Notre recherche propose de répondre aux questions suivantes :

1. Dans un corpus d'articles sur la Grève étudiante au Québec en 2012 qui a été classé préalablement par un humain dans des classes opposées, quels sont les critères textuels permettant de caractériser et de distinguer les opinions exprimées dans les textes sur les plans thématique, dialectique et dialogique ?

2. Quels types de critères textuels issus de l'analyse textométrique entamée à question 1 sont les plus performants pour prédire la classe des articles d'opinion lors d'un processus de classification automatique ?

3. À quel moment dans la controverse apparaissent les critères textuels les plus performants pour la tâche de classification ?

Chapitre 1. Revue de la littérature

1. Introduction

L'objectif premier de cette revue est de faire l'état des connaissances sur notre problématique de recherche. Dans la première partie, nous présentons un exposé sur l'avènement et l'historique des systèmes de recommandation. Nous situons les systèmes de recommandation dans le développement historique des systèmes de repérage de l'information, en explicitant les affinités entre ces deux genres de systèmes. Nous caractérisons les systèmes de recommandations d'articles de presse (SRAP) comme un type de système de recommandation et nous décrivons leurs particularités. Nous allons également expliquer les techniques de recommandation utilisées dans les SRAP afin de contextualiser notre problématique de recherche. Par la suite, nous examinons un ensemble de travaux qui ont proposé des solutions techniques pour diversifier les opinions et les perspectives véhiculées par les articles qui sont suggérés par les SRAP.

L'objectif secondaire de cette revue consiste à mettre au clair les définitions et les concepts clés de notre recherche, afin d'en assurer une utilisation cohérente et précise. En nous appuyant sur une recension d'écrits comprenant des sources variées (des articles scientifiques, des ouvrages, des encyclopédies, des dictionnaires) nous donnons des précisions sur les concepts d'opinion et de controverse, utilisés dans le cadre de cette recherche. Ces concepts sont par la suite décrits à l'intérieur d'un cadre théorique qui considère l'aspect communicationnel des échanges linguistiques.

2. Les systèmes de recommandation : historique, définitions et méthodes

2.1 Introduction

Les systèmes de recommandation ont leur origine dans les recherches qui ont mené à l'élaboration du filtrage collaboratif, méthode développée dans les années 1990 par le centre de recherche *Xerox Palo Alto* (PARC) et destinée à contrer la surcharge informationnelle provoquée par l'envoi massif de courriels non sollicités (Martin et coll., 2011 ; Parisier, 2011). Le projet de filtrage collaboratif développé par le PARC consistait à enregistrer les données sur les réactions des usagers aux courriels reçus (messages lus, exclus ou archivés) et les réutiliser pour différencier les messages plus importants des pourriels[8]. En 1994, le créateur du magasin en ligne *Amazon.com*, Jeff Bezos, reprend la technique de filtrage collaboratif et crée un système de recommandation de livres basé sur l'historique d'achats effectués par les clients. En appariant les profils des clients qui avaient des préférences de consommation similaires (ceux qui avaient acheté les mêmes produits), le système recommandait les livres susceptibles de les intéresser.

Le succès du site *Amazon* a contribué considérablement à l'intérêt pour les systèmes de recommandation dans les deux dernières décennies (Martin et coll., 2011). Ils sont devenus des incontournables dans la plupart des sites Web et des catalogues en ligne grâce à leur potentiel d'offre de recommandations personnalisées pour la découverte de nouveaux produits comme des livres, des films, des albums de musique, etc. (Poirer et coll., 2010). Nous pouvons citer comme exemple le site *Netflix*, qui recommande des films, ainsi que le site *Yahoo!*

[8] Selon la définition proposée par l'Office québécois de la langue française, il s'agit d'un « Message électronique importun et souvent sans intérêt, constitué essentiellement de publicité, qui est envoyé à un grand nombre d'internautes sans leur consentement, et que l'on destine habituellement à la poubelle. » (Office québécois de la langue française [OLF], 2013).

Actualités, qui suggèrent aux usagers des articles de presse correspondants à leurs préférences personnelles.

Un des premiers travaux sur les systèmes de recommandation a défini ces derniers comme des systèmes dans lesquels « les personnes fournissent en entrée des recommandations que le système agrège puis les font suivre à des destinataires appropriés » (Resnick et Varian, 1997, p.56, traduction libre). Cette définition explicite le fait que les recommandations des systèmes sont générées en fonction de la recommandation des usagers à des contenus existants, fournies comme données d'entrée sur la forme de clics, d'attribution de notes et d'autres mécanismes d'évaluation. D'autres écrits définissent les systèmes de recommandation comme un mécanisme de filtrage personnalisé intégré à un système d'information, dont la fonction est de suggérer des ressources informationnelles qui correspondent aux préférences des usagers (Lynch, 2001 ; Nageswara Rao et Talwar, 2008 ; Park et coll., 2012). Le but de ce mécanisme de filtrage est de créer un rapprochement entre le profil des usagers, qui sont construit en fonction de leur goûts, et les contenus présents dans la base d'informations, facilitant la découverte de ces derniers. (Furner, 2002 ; Lynch, 2001 ; Nageswara Rao et Talwar, 2008 ; Park et coll., 2012 ; Poirer et coll., 2010).

Bon nombre de chercheurs (Burke, 2002 ; Lei et coll., 2011 ; Lynch, 2001 ; Parisier, 2011) voient les moteurs de recherche à l'instar de *Google* comme un type de système de recommandation, car ils intègrent des mécanismes de filtrage dans le repérage de l'information. Dans ces derniers, les données provenant d'actions que les usagers effectuent dans le système (par exemple, consulter, apprécier ou noter un document[9]), ainsi que des données relatives au contexte de l'usager (comme la localisation géographique) sont utilisées pour prédire leurs préférences pour de nouveaux documents lors du processus de recherche. Ainsi, le résultat d'une requête apparaît déjà trié dans une liste ordonnant les documents les

[9] Nous avons utilisé le terme « article » pour référer à des types de documents qui sont traités par les SRAP. Nous utilisons le terme « document » dans le but d'englober d'autres types de documents qui peuvent être traités par les systèmes de recommandation (par exemple, de documents contenant des métadonnées sur les films, les livres, etc.).

plus pertinents, en fonction non seulement des mots clés présents dans la requête, mais aussi des préférences personnelles inférées.

2.2 Recommandation ou repérage ?

L'inclusion des systèmes de repérage personnalisés dans la catégorie des systèmes de recommandation installe une certaine confusion par rapport à la définition de ces derniers. Les systèmes de recommandation sont-ils un genre de système de repérage ou plutôt un mécanisme de filtrage ? Dans quel sens l'action de prédire les documents qui sont pertinents pour un usager diffère-t-elle dans un contexte de recommandation et de repérage de l'information ?

L'étude des systèmes de recommandation s'insère dans une tradition de recherches sur le filtrage de l'information, qui se sont différenciées historiquement des études sur le repérage de l'information (Furner, 2002). Belkin et Croft (1992), remettent en question cette différenciation en comparant les mécanismes des systèmes de filtrage et de repérage. Ils constatent que leurs différences ne sont pas théoriquement significatives. D'une part, les auteurs défendent que d'un point de vue global et théorique, les objectifs de deux types de systèmes s'équivalent, puisqu'ils tentent de répondre de façon adéquate à un besoin informationnel. En outre, les enjeux techniques qui touchent le processus de réponse à un besoin informationnel sont essentiellement les mêmes pour les deux types de systèmes.

Belkin et Croft (1992) avancent que le processus de filtrage de l'information contient des entités et des opérations très similaires au processus de repérage de l'information (voir les figures 1 et 2 pour une comparaison). Les différences existantes ne touchent pas l'essentiel du processus, mais plutôt des aspects spécifiques de ce dernier. Elles se trouvent sur le plan du type de collection de documents, de la régularité de la mise à jour de cette dernière dans le système, dans la manière dont les usagers expriment leur besoin informationnel et dans la façon dont les documents sont présentés aux usagers.

Ainsi, dans les systèmes basés sur le modèle de repérage, les collections sont relativement statiques dans le sens où leur mise à jour n'est pas nécessairement régulière. Il s'agit de collections dans lesquelles les documents ont une longue durée de vie et dont la récence n'est pas un critère documentaire très important pour l'usage du système. Dans le

repérage, cet usage du système est « unique », car une personne avec un besoin d'information ponctuel soumet une requête qui traduit ce besoin afin d'interroger le système. Dans le filtrage, la collection est plutôt dynamique, car les documents sont intégrés de façon plus régulière dans la base de données et l'usage du système ne demande pas d'action directe de l'usager, puisque son besoin informationnel est exprimé par son profil qui réunit un ensemble d'évidences hétérogènes sur son intérêt (documents consultés, profils similaires, etc.). Les auteurs soulignent aussi que dans les systèmes d'information basés sur le modèle de filtrage, l'usage du système est répété, car l'usager est informé régulièrement de l'arrivée de nouveaux documents correspondant à son profil (Belkin et Croft, 1992).

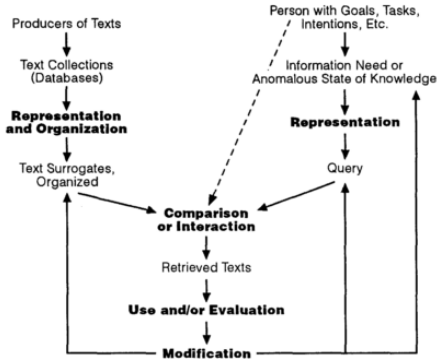


Figure 1. Modèle générique de repérage de l'information (Belkin et Croft, 1992, p. 31)

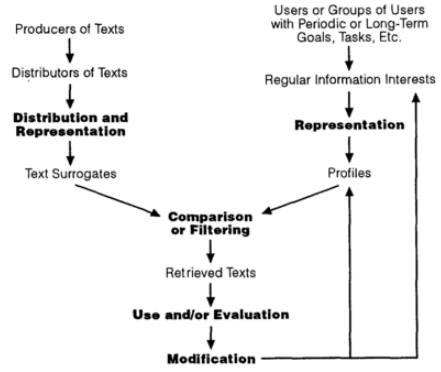


Figure 2. Modèle générique de filtrage de l'information (Belkin et Croft, 1992, p. 31)

D'ailleurs, Belkin et Croft (1992) expliquent que le domaine du filtrage de l'information a beaucoup bénéficié d'études réalisées sur le repérage de l'information. Les auteurs donnent l'exemple de la représentation de textes et de requêtes avec le modèle vectoriel ainsi que les modèles de comparaison entre les textes et les requêtes, comme les mesures de similarité. Ces solutions techniques se trouvent alors employées de manière subtilement différente dans le modèle de filtrage, puisqu'au lieu de comparer les représentations des textes à une requête, le modèle de filtrage implique la comparaison de textes à des profils, eux aussi constitués de mots extraits des textes.

L'idée selon laquelle le repérage de l'information et le filtrage de l'information sont comme les deux faces de la même médaille (Belkin et Croft, 1992) est développée par Furner (2002), pour qui les systèmes de repérage sont aussi des systèmes de recommandation.

D'abord, Furner (2002) explique que l'objectif des systèmes de recommandation est identique à celui des systèmes de repérage, c'est-à-dire que tout système qui implémente un mécanisme dans lequel les objets de la collection sont triés selon l'ordre prévu d'utilité à un usager est un système de repérage.

Ensuite, Furner (2002) généralise la notion de recommandation. Il considère que toute action qui vise à attribuer un jugement de valeur et de pertinence à un document est fondamentalement une action de recommandation. Ainsi, si une personne ou un engin automatique indexe un document à l'aide de mots clés, il attribue un jugement sur la pertinence de ces mots pour décrire la valeur du document en question. Alors cet agent « recommande » un ensemble d'attributs qui décrivent ce document en anticipant leur utilité pour l'usager qui veut le chercher.

En s'appuyant sur le modèle d'information décisionnelle ERIn (*évaluation-recommandation-information*), Furner (2002) définit la recommandation comme étant l'action d'exprimer une proposition observable sur la valeur perçue d'un document. L'expression de cette proposition est effectuée par un agent situé dans le système lui-même ou en dehors de celui-ci. Cela peut être un usager, l'auteur du document, un indexeur ou un agent automatique qui indexe ce document dans le format vectoriel, par exemple. L'agent qui recommande un document doit indiquer l'ordre de préférence dans lequel les options disponibles dans la

collection doivent être considérées. De cette manière, indexer, citer un document ou formuler une requête sont des actions de recommandation analogues qui fournissent les indices que les systèmes utilisent pour attribuer une valeur de pertinence, d'utilité ou d'approbation aux documents.

Par exemple, si l'auteur d'un document cite un ouvrage dans son texte, le système peut utiliser cette recommandation dans le cadre d'un processus de repérage. C'est le cas notamment des bases de données bibliographiques qui reposent sur un index de citations comme la *Science Citation Index* (SCI). Pareillement, si la consultation d'un document par un usager dans un système est comptabilisée comme vote pour ledit document, ce jugement de pertinence est utilisé comme source par le système pour prédire le besoin informationnel de l'usager en question. Selon cette vision de Furner (2002), les systèmes ne font que synthétiser ces différents jugements afin de prédire un rang de documents qui se rapprochent le plus possible des préférences exprimées par les usagers.

La perspective que Furner (2002) et Belkin (2000) proposent et que nous envisageons dans le cadre de ce travail, c'est de voir les systèmes de recommandation comme synonyme des systèmes de repérage, car les deux agissent sur la prédiction de l'utilité des objets d'une collection à un usager particulier :

> It is a system, operating on behalf of information seekers, that evaluates documents (whether through direct observation or through analysis of testimonials), arrives at preference orderings, and presents those orderings, so that searchers can come to their own relevance judgments and make their own reading decisions on the basis of the evidence supplied by the system (Furner, 2002, p. 755)

Par rapport au modèle générique de filtrage de l'information de la figure 2, notre problème de recherche se rapporte à l'étape « Distribution et Représentation » (Belkin et Croft, 1992, traduction libre) dans laquelle les textes obtiennent une représentation vectorielle pour des fins d'analyse statistique de leur contenu. Notre recherche ne traite pas de la comparaison entre ces représentations et les profils d'usagers.

2.3 Les méthodes de recommandation et les particularités des SRAP

Les méthodes de recommandation sont classées en trois grandes catégories : 1) le filtrage basé sur le contenu, qui analyse le contenu des articles consultés par les usagers pour générer des recommandations similaires ; 2) le filtrage collaboratif, qui procède par analyse du voisinage entre les usagers qui ont des préférences similaires ; et 3) le filtrage hybride, qui combine les méthodes de filtrage basé sur le contenu et le filtrage collaboratif pour optimiser les résultats de la recommandation (Furner, 2002 ; Karimi et coll., 2018 ; Jannach et coll., 2011 ; Nageswara Rao et Talwar, 2008 ; Park et coll., 2012 ; Poirer et coll., 2010). Burke (2002) ne considère pas la méthode hybride comme une méthode en soit, car elle consiste à combiner le filtrage basé sur le contenu et le filtrage collaboratif pour contrer les problèmes associés à chaque type. Plus récemment, une quatrième méthode basée sur la représentation de connaissances (*Knowledge Representation*) est proposée pour un type particulier de recommandation de produits à valeur élevée comme les maisons et les voitures (Jannach et coll., 2011). Dans ce genre de recommandation, les objets à recommander sont représentés par un ensemble de concepts relatifs à un domaine d'expertise (par exemple, les caractéristiques d'une voiture). Ces systèmes aident les usagers à trouver des produits de consommation dont les caractéristiques correspondent le plus précisément possible à leurs besoins personnels.

Le filtrage basé sur le contenu est fondé sur l'analyse informatique du contenu des articles à recommander. Dans ce type de filtrage, les articles sont représentés par des vecteurs contenant un ensemble de mots qui ont été extraits de ces derniers. Le système essaye de prédire de nouveaux articles qui pourraient intéresser un usager en fonction de son profil, constitué par un autre vecteur contenant les mots extraits des articles lus dans le passé. Ce profil est continuellement mis à jour par le système. Les algorithmes de recommandation tentent d'apparier de nouveaux articles aux profils existants, en calculant la similarité entre les vecteurs constitués (Jannach et coll., 2011 ; Karimi et coll., 2018 ; Lei et coll., 2011 ; Pazzani et Billsus, 2007).

Les calculs de similarité prennent en compte la contribution de chaque mot présent dans les vecteurs pour rapprocher les profils des usagers aux articles à recommander. Des mesures statistiques sont appliquées pour sélectionner dans les articles l'ensemble des mots retenus comme traits discriminants pour constituer les vecteurs. La mesure TF-IDF (*Term*

Frequency-Inverse Document Frequency) proposée par Salton et McGill (1983) est parmi une des plus populaires (Ahn et coll., 2007 ; IJntema et coll., 2010 ; Rao et coll. 2013). Cette mesure statistique permet d'obtenir les traits les plus discriminants des articles en pondérant leurs fréquences en fonction de toutes les fréquences observées dans l'ensemble des articles de la collection. D'autres techniques sont également proposées, comme le modèle génératif probabiliste connu comme « Allocation de Dirichlet Latente » (Blei et coll., 2003 ; Desarkar et Shinde, 2014 ; Li et coll., 2011), qui pondère statistiquement les fréquences des mots en fonction de leur co-présence dans les articles.

Le calcul de similarité peut être fait avec des mesures statistiques qui estiment la distance entre les vecteurs dans l'espace vectoriel, comme le calcul des k-plus-proches-voisins (Lei et coll., 2011), le coefficient de Jaccard (Abbar et coll., 2013 ; Li et coll., 2011 ; Poirer et coll., 2010) ou la similarité cosinus (Ahn et coll., 2007 ; IJntema et coll., 2010 ; Rao et coll. 2013).

Selon Lei et coll. (2011) la recommandation dans les SRAP peut être également modélisée à partir de la méthode de classification supervisée avec l'utilisation de techniques d'apprentissage automatique. Dans cette méthode, des exemples d'articles rassemblés dans une classe prédéterminée sont fournis aux machines pour qu'elles puissent « apprendre » les patrons statistiques de la classe en question (Forest et coll., 2009). Cet apprentissage donne la possibilité aux machines de générer un modèle de classification (dorénavant, classifieurs) qui est appliqué pour prédire de nouveaux articles de la classe étudiée. Dans le cas des SRAP, les articles lus par un usager peuvent constituer des exemples pour entraîner les classifieurs du système. Ces classifieurs peuvent par la suite prédire si un nouvel article publié et pas encore lu pourrait intéresser l'usager.

Dans le filtrage collaboratif, le contenu des articles n'est pas analysé dans le processus de recommandation. Les systèmes basés sur cette méthode engagent les usagers dans un mécanisme de vote de type explicite ou implicite. Dans le premier type, les usagers sont incités à porter une appréciation et à exprimer une évaluation positive ou négative sur un article consulté. Dans le second type, les actions comme les clics ou la lecture d'un article sont comptées comme un vote pour ce dernier (Resnick et Varian, 1997). Ces votes, traduits en valeurs numériques, sont par la suite utilisés pour comparer les usagers entre eux. Ceux qui

partagent le même profil de vote sont considérés comme ayant les mêmes intérêts et, par conséquent, sont regroupés. Ainsi, le système prédit l'intérêt d'un usager pour un article donné en cherchant d'autres usagers dont les préférences lui sont apparentées et qui ont déjà porté une appréciation positive à l'article en question. Concrètement, si l'usager A a porté une appréciation positive pour l'article *a* et que les usagers A et B ont des goûts similaires, l'article *a* sera aussi recommandé à l'usager B, peu importe son contenu. Les profils des usagers, constitués de vecteurs qui contiennent les données identifiant les articles et les évaluations correspondants, sont aussi rapprochés en utilisant des mesures de corrélation ou de similarité (Burke, 2002).

Le filtrage basé sur le contenu et le filtrage collaboratif présentent quelques inconvénients. Le principal inconvénient du filtrage basé sur le contenu, est qu'il est subordonné totalement à l'analyse du contenu pour représenter, d'une part, les nouveaux articles qui sont ajoutés dans le système et, d'autre part, les profils des usagers. Par exemple, si un usager consulte beaucoup d'articles sur un même thème, son profil accumulera au fil du temps un nombre important de mots provenant de ces articles, de sorte que les recommandations générées par le système risquent d'être assez homogènes pour cet usager (Burke, 2002). Inversement, si les articles consultés sont assez divers, les vecteurs qui composent les profils des usagers peuvent aussi présenter une grande diversité et diminuer l'efficacité du système à prédire les préférences aux nouveaux articles ajoutés (Lei et coll., 2011). Les systèmes à filtrage collaboratif doivent par contre trouver des solutions pour contrer un problème connu comme démarrage à froid, qui est causé lorsqu'aucune information sur les usagers n'a été obtenue ou lorsqu'un article ajouté à la base n'a pas été encore évalué par les usagers et risque de rester inconnu (Burke, 2002).

La présence des contraintes associées aux méthodes de recommandation existantes a collaboré aux progrès des méthodes de filtrage hybrides. Celles-ci ont la caractéristique principale d'alterner entre différentes techniques pour contrer le manque de disponibilité d'informations dans le système (Burke, 2002 ; Martin et coll., 2011). Par exemple, lorsque les informations à propos des usagers et de leurs préférences sont inconnues dans un système basé sur le filtrage collaboratif, il est possible de faire intervenir un filtrage basé sur le contenu par une technique de commutation de méthodes (Burke, 2002 ; Poirer et coll., 2010). D'autres

techniques d'hybridation incluent la fusion des méthodes, permettant par exemple d'utiliser les informations sur le voisinage entre les usagers comme une information additionnelle dans la représentation vectorielle du contenu des articles, ou encore l'hybridation échelonnée, dans laquelle un type de filtrage sert à raffiner la recommandation faite par un autre (Burke, 2002).

L'inférence d'informations relatives au contexte des usagers peut également aider à contrer les problèmes relatifs au démarrage à froid éprouvé par le filtrage collaboratif. Par exemple, les informations relatives à la localisation géographique (Chu et Park, 2009 ; Montes-Garcia et coll., 2013), à la date de l'article, à la crédibilité de la source (Chu et Park, 2009) ou encore à la popularité de l'article dans le système (Chu et Park, 2009 ; Ge et coll., 2010 ; Li et coll., 2011) peuvent être utilisées pour prédire l'intérêt des usagers à certains types de contenu.

La lecture de nouvelles constitue un type spécifique d'accès à l'information qui pose des difficultés différentes pour la modélisation des systèmes de recommandation (Ahn et coll., 2007 ; Chu et Park, 2009 ; Li et coll., 2011). Le caractère périssable des nouvelles est souvent mentionné par les études sur les SRAP (Chu et Park, 2009 ; Lei et coll., 2011 ; Li et coll., 2011 ; Özgöbe et coll., 2014). Le fait que d'autres événements peuvent tout d'un coup devenir très populaires dans les médias entraîne une perte d'intérêt rapide pour les contenus qui sont présents dans le système. Ce type de difficulté est moins important pour d'autres types de contenu comme les livres ou les films, dont la durée de vie est beaucoup plus longue et dont l'intérêt ne tende pas à se périmer complètement au fil du temps (Ahn et coll., 2007 ; Özgöbe et coll., 2014). En plus, dans le contexte de la recommandation d'articles de presse, la récence va toujours avoir un impact sur la pertinence de la recommandation, à moins que l'usager veuille consulter des événements passés.

Karimi et coll. (2018) distingue trois types de techniques qui incorporent la récence comme facteur de pertinence dans la modélisation des SRAP, parmi d'autres facteurs : le pré-filtrage, qui permet d'enlever les articles anciens avant de proposer une nouvelle recommandation ; la modélisation avec récence, qui prend en compte la date de publication de l'article lors de conception du processus de recommandation (Li et coll., 2011) ; le post-filtrage, qui permet de diminuer l'importance des articles plus anciens dans le rang final d'articles proposés. Wen et coll. (2012) ont proposé d'utiliser la date de publication des

articles dans le processus de recommandation comme un coefficient qui est combiné au calcul de pertinence permettant d'apparier les articles aux profils d'usagers constitués.

En analysant 112 articles scientifiques sur les SRAP, Karimi et coll. (2018) ont observé que plus de la moitié des études dans le domaine utilisent le filtrage basé sur le contenu (59 articles). Parmi la totalité d'articles analysés par les auteurs, seulement 19 ont reporté l'utilisation du filtrage collaboratif dans le contexte des SRAP. Cependant, 45 des études proposent le filtrage hybride. La prééminence du filtrage basé sur le contenu dans les SRAP s'explique par la richesse des données provenant des textes qui peuvent être traitées avec des techniques qui sont assez répandues dans le domaine du repérage de l'information (Karimi et coll., 2018). Burke (2002) remarque aussi que le filtrage collaboratif est plus performant pour les types de systèmes qui ne font pas une mise à jour très fréquente en termes d'intégration de nouveaux articles. Ainsi, lorsque la collection est relativement petite et plus au moins statique, la densité des évaluations permet plus facilement de rapprocher les usagers qui partagent des préférences.

Le filtrage basé sur la représentation de connaissances n'a pas été trouvé dans les articles qui composent notre revue de la littérature au sujet des SRAP.

2.4 Conclusion

Dans cette section, nous avons analysé l'avènement des systèmes de recommandation et nous avons constaté que leur étude provient des recherches sur le filtrage collaboratif. Ils s'inscrivent dans une tradition de recherches relatives principalement au filtrage de l'information. Nous avons également constaté que l'évolution récente des systèmes de recherche sur le Web, surtout au niveau des techniques de personnalisation, a peu à peu effacé les frontières entre les notions de filtrage et de repérage. Cela nous amène à considérer que ces deux types de systèmes partagent globalement un objectif commun, celui de répondre adéquatement à l'expression d'un besoin informationnel. Dans les systèmes de recherche, ce besoin est exprimé par la requête et dans les systèmes de recommandation, par les actions réalisées par les usagers. En ce sens, nous avons retenu la définition de Furner (2002) sur les systèmes de recommandation. Il s'agit d'un système qui récupère les traces transmis par les usagers pour générer un ordre de préférence sur la base de laquelle les documents de la

collection doivent être rangés, de façon à laisser aux usagers concernés le choix de leurs propres lectures. Il s'agit d'une définition non restrictive et qui englobe les différentes méthodes de recommandation existantes. La création des méthodes de recommandation dépend de la façon dont les concepteurs de ces systèmes considèrent le besoin informationnel, décision qui doit prendre en compte aussi le type de contenu à recommander. Diversifier les contenus lors de la recommandation de nouvelles, par exemple, peut aider à satisfaire le besoin d'un usager désireux de s'informer à propos des différentes opinions véhiculées dans un débat public.

Nous avons également souligné que notre recherche se penche sur la composante « Distribution et Représentation » du modèle de filtrage de l'information proposé par Belkin et Croft (1992). Cette composante traite de la représentation informatique des textes numériques dans le modèle vectoriel. Notre recherche vise à créer un cadre expérimental nous permettant de connaître les observables linguistiques les plus performants pour distinguer les articles d'opinion à l'intérieur d'une discussion polarisée, dans laquelle les points de dissension sont bien évidents. Il ne s'agit pas ici de développer une méthode de recommandation qui fera le rapprochement entre les profils et les articles, mais de déployer un cadre expérimental nous permettant d'évaluer la pertinence d'un ensemble d'observables linguistiques pour classer les articles d'opinion selon le type d'opinion véhiculée.

Nous avons également présenté les principales méthodes de recommandation. La richesse de données à explorer dans les textes explique la prépondérance du filtrage basé sur le contenu dans le développement SRAP. Nous avons vu que dans le filtrage basé sur le contenu, la représentation des articles et du profil de l'usager est faite sur la base du modèle vectoriel, en fonction des mots extraits des articles. La prédiction de la pertinence d'un article à un usager donné est un calcul de similarité entre les vecteurs. La sélection des traits discriminants pour la représentation vectorielle est faite par des mesures qui calculent leur représentativité statistique des mots dans le corpus de textes, comme le TF-IDF.

Le caractère périssable des articles de presse est un facteur à considérer dans le développement des SRAP. Cela est le cas spécialement pour les nouvelles, qui ont une durée de vie relativement courte. La publication de nouvelles dans les journaux est assez rapide et d'autres événements peuvent gagner rapidement de la popularité. Nous pouvons nous

demander à quel point cela s'avère aussi pour d'autres genres journalistiques. Les articles d'opinion publiés dans la presse, tels que les éditoriaux, les chroniques politiques ou les articles de tribune libre, peuvent avoir un intérêt de longue durée, car ce sont des genres qui offrent une interprétation et une analyse plus approfondie des événements courants.

3. Diversification de contenus dans les SRAP

3.1 Introduction

L'objectif de cette section est de démontrer la pertinence de notre recherche pour contrer les problèmes relatifs à la diversification de contenus dans les SRAP. Dans la première partie, nous expliquerons les risques associés au manque de diversité dans les environnements voués à la diffusion d'informations journalistiques. Dans la deuxième partie, nous présentons un ensemble de travaux qui se sont penchés sur le problème de la diversité des SRAP, afin d'offrir un panorama des recherches dans le domaine. Nous abordons également des travaux qui ont essayé de répondre au besoin de la diversité de contenus dans les SRAP en proposant l'utilisation de techniques en fouille d'opinions. Ces derniers sont plus liés à nos questions de recherche dans la mesure où ils proposent la recommandation d'articles sur un même thème, mais qui se distinguent par le type d'opinion véhiculée.

La dernière partie présente une discussion épistémologique au sujet de la fouille d'opinions. Nous voulons démontrer qu'un regard différent sur certains présupposés théoriques qui soutiennent les recherches en fouille d'opinions dans les SRAP peut répondre plus adéquatement au problème de recherche que nous avons posé. Nous allons clore cette section avec une discussion sur l'ensemble des écrits exposés.

3.2 Le phénomène de l'exposition sélective

Certains travaux critiquent la tendance des SRAP, surtout ceux basés sur l'approche collaborative, à créer un environnement informationnel excessivement sélectif, où les parcours d'accès à l'information sont de plus en plus limités en fonction des choix et des comportements individuels des usagers (Mutz et Young, 2011 ; Parisier, 2011 ; Sunstein, 2007 ; Van Alstyne et Brynjolfsson, 2005). Ces auteurs soulignent d'ailleurs un risque important de ces systèmes qui pourraient, dès lors, confiner les usagers dans une « bulle de filtrage » (Parisier, 2011). Cela aurait pour effet de les exposer aux contenus qui correspondent davantage à leurs idées et de contribuer à renforcer leurs convictions politiques.

En analysant l'évolution des recherches publiées dans la revue *Public Opinion Quarterly* au cours des trois dernières décennies, Mutz et Young (2011) constatent que malgré

l'impact des innovations technologiques de communication (incluant la télévision et plus récemment l'Internet), certains thèmes restent toujours actuels, comme la question des effets persuasifs des médias. Cette question si populaire connaît toutefois une évolution dans son traitement : le type d'effet devenu problématique avec l'avancement des nouvelles technologies ne se réfère plus au pouvoir des médias d'influencer ou de changer l'opinion publique, mais de renforcer les opinions individuelles et de donner naissance à un extrémisme nocif pour la démocratie (Mutz et Young, 2011).

L'évolution historique des recherches sur la diversité des contenus politiques dans les médias peut aider à comprendre cette préoccupation. Il y a 25 ans, la télévision, considérée comme le principal média de la vie politique, était souvent critiquée pour le manque de diversité des chaînes de programmation. Les programmes de nouvelles, obéissant aux normes d'objectivité journalistique, étaient accusés de faire disparaître le type de journalisme partisan qui les précédait, où différentes perspectives politiques étaient présentées clairement au public et encourageaient d'ailleurs le débat au sein de la société (Mutz et Young, 2011).

Avec l'avènement de la télévision par câble et plus récemment avec l'Internet, nous assistons à une augmentation fulgurante de nouveaux contenus, autant de genres (cinéma, téléroman, divertissement, etc.) que de nouveaux espaces de diffusion de contenus comme les blogues, les laboratoires d'idées, les sites d'organisations civiles, etc. Ce qui pourrait constituer un indice de diversification et de pluralité dans l'espace communicationnel s'est révélé à l'inverse être une nouvelle source de problèmes. En effet, puisque la quantité de sources d'informations est énorme et devient ingérable, les individus sont plus enclins à choisir les contenus qui correspondent le mieux à leurs visions politiques et à éviter ceux qui contredisent ces mêmes visions (Mutz et Young, 2011). Ce phénomène, connu comme « exposition sélective », est à la base des préoccupations comme celles de Sunstein (2007), qui défend que l'exposition exclusive des individus à un ensemble d'opinions avec lesquelles ils sont d'accord fragmente et polarise l'espace de communication politique, conduisant à l'appauvrissement de ce dernier et de la société dans son ensemble.

Les SRAP ont été conçus dans le but de faciliter la tâche de filtrage parmi la multitude de sources de presse présentes sur le Web (Mutz et Young, 2011). Cependant, puisqu'ils sont basés sur des techniques de filtrage qui regroupent des usagers ayant des préférences et des

habitudes de lecture communes (ou qui recommande des contenus similaires à ceux déjà consultés), ils finissent aussi par renforcer le phénomène de l'exposition sélective. Le risque évident de ces techniques est de rendre disponibles exclusivement les contenus correspondant aux préférences des usagers, collaborant à produire une certaine homogénéisation à l'égard des contenus qu'ils leur sont offerts.

Bien que les causes du phénomène d'exposition sélective ne soient pas uniquement provoquées par la façon dont les systèmes sont conçus (par exemple, Bakshy et coll. (2015) souligne que les choix individuels jouent également un rôle important dans le phénomène), le potentiel qu'elle représente d'empêcher la formation d'un environnement informationnel diversifié est évoqué par certaines études sur la recommandation d'articles de presse (Abbar et coll., 2013 ; Gemmis et coll., 2015). Il est également évoqué le problème de la personnalisation excessive des recommandations et la nécessité de développer des techniques capables de fournir plus de diversification dans les SRAP (Karimi et coll., 2018 ; Park et coll., 2012). Dans la section prochaine, nous allons exposer des travaux qui ont traité de la question de la diversité dans les SRAP, afin de comprendre comment cette dernière est définie et si elle répond ou pas au problème de l'exposition sélective.

3.3 Diversification dans les SRAP : un facteur de qualité

Dans les évaluations des systèmes de recommandation, la précision mesure la capacité du système à prédire les articles qui seraient lus par un usager donné. Ces évaluations sont souvent menées dans un cadre expérimental hors-ligne (Karimi et coll., 2018), en utilisant des données enregistrées dans les bases de données à propos des activités réelles effectuées par les usagers dans le système, comme les articles consultés ou les recherches réalisées. Les démarches d'évaluation des SRAP comparent les résultats obtenus par le processus de recommandation expérimental avec les activités effectives. Par exemple, si les articles suggérés par le processus de recommandation correspondent aux articles qui ont été effectivement lus par l'usager, la précision de la prédiction pour ce système est jugée élevée. Vargas et coll. (2014) ont particulièrement critiqué le fait que les évaluations des SRAP, trop axées sur le calcul de la précision relative à la prédiction des articles pertinents, ne sont pas capables de capturer toutes les autres facettes qui peuvent être révélatrices de l'intérêt des

usagers. Ainsi, l'évaluation sur la diversité ou à la nouveauté qui peut être apportée par le processus de recommandation n'est pas réalisée par les recherches.

Depuis 2011, les études sur les SRAP démontrent un intérêt plus prononcé par la question de la diversité dans le processus de recommandation (Karimi et coll., 2018). La diversité est vue par certains auteurs comme un critère lié à la qualité de la recommandation (Desarkar et Shinde ; 2014 ; Ekstrand et coll., 2014 ; Herlocker et coll., 2004 ; McNee et coll., 2006 ; Vargas et coll., 2014). Certaines recherches ont d'ailleurs constaté que la diversification de contenus dans la recommandation est perçue positivement par les usagers (Ekstrand et coll., 2014 ; Pu et coll., 2011 ; Ziegler et coll., 2005) et que la précision du système en matière de prédiction n'entraîne pas nécessairement de la satisfaction (Ekstrand et coll., 2014).

Dans le contexte des recherches sur la diversification dans les SRAP, la diversité est généralement associée à la variété interne de la liste d'articles recommandés (Vargas et coll., 2014). Vargas et coll. (2014) définissent la diversité comme une qualité des systèmes de recommandation lorsque ceux-ci prennent en compte la variété de goûts personnels et de besoins que les usagers sont susceptibles d'avoir, par exemple la variété de thèmes ou de genres. Karimi et coll. (2018) soulignent également l'importance de considérer la diversité comme un critère de qualité. Pour ces auteurs, il faut pondérer les conclusions sur les comportements et les habitudes de lecture que les usagers démontrent, car si un usager lit beaucoup sur le thème des élections présidentielles de son pays, cela ne signifie pas qu'il veut encore recevoir comme recommandation d'autres articles portant sur le même thème. La diversité dans la recommandation permet d'éviter l'ennui que les usagers peuvent éprouver lorsqu'ils voient des recommandations trop similaires correspondant aux préférences que le système a inférées à partir de leur historique de lecture (Desarkar et Shinde, 2014).

La diversité est aussi définie par opposition à la sérendipité (Karimi et coll., 2018 ; Pu et coll., 2011), qui est perçue comme la proposition de suggestions inattendues dans le processus de recommandation (Gemmis et coll., 2015 ; Musto et coll., 2017 ; Shani et Gunawardana, 2011), ou encore par opposition à la nouveauté (Ge et coll., 2010 ; Shani et Gunawardana, 2011), définie comme la capacité du système à présenter des documents avec des sujets nouveaux et inconnus des usagers (Shani et Gunawardana, 2011). À l'inverse de la sérendipité ou de la nouveauté, qui doit être perçue par rapport à la recommandation des

documents pris individuellement, la diversité doit être perçue à l'intérieur d'un sous-ensemble de documents qui est recommandé par le processus (Pu et coll., 2011). Par exemple, McNee et coll. (2006) préconisent une diversification interne à la liste de recommandation, en expliquant que l'efficacité du processus de recommandation qui prend en compte la diversité doit être jugée par l'ensemble des résultats retournés et non par les résultats individuels de chaque document qui compose la liste.

Dans ce contexte, le principal défi pour le développement des SRAP est de trouver un équilibre entre la précision de la prédiction (l'efficacité du système à prédire un article qu'un usager cible est susceptible de consulter) et la diversité, qui est une mesure de différence interne entre les résultats retournés par le processus (McNee et coll., 2006). Cependant, l'équilibre entre la précision et la diversité est difficile à atteindre, puisqu'une liste trop diverse peut aussi devenir non pertinente pour l'usager (Karimi et coll., 2018).

Desarkar et Shinde (2014) définissent la diversité comme une différence maximale entre des objets similaires qui recommandés dans une liste. Ils proposent un algorithme de recommandation qui considère d'une part le critère de la pertinence et d'autre part un critère de diversité. Préoccupés par des questions sur la protection de données privées, Desarkar et Shinde (2014) proposent de définir la pertinence par un critère de popularité, qui est calculée en fonction du nombre de fois que l'article a été lu par d'autres usagers du système. Le processus de recommandation proposé est échelonné sur deux étapes distinctes. Dans la première étape, l'algorithme sélectionne un nombre d'articles candidats en fonction de la pertinence pour le profil de l'usager. Dans la seconde étape, un sous-ensemble de cette première sélection est formé à partir d'un calcul qui maximise la différence entre chaque article en fonction d'une mesure de distance statistique entre les vecteurs. La mesure de diversité prend comme traits discriminants certains mots et les entités nommées qui sont mentionnées dans les articles. Selon les auteurs, l'avantage d'intégrer la diversité dans un système de recommandation est pouvoir couvrir un ensemble plus large d'articles sur d'autres thèmes que le calcul de pertinence seul ne saurait pas considérer.

Li et coll. (2011) ont proposé la plateforme SCENE, un SRAP qui inclut dans le processus de recommandation le critère de pertinence et de diversité. Comme dans Desarkar et Shinde (2014), le modèle de recommandation proposé se compose de deux étapes. Dans la

première, les articles qui ont été récemment publiés et qui sont également les plus populaires sont regroupés en fonction de leur similarité, puis appariés aux profils des usagers. Dans la seconde étape, cet appariement est raffiné pour suggérer les articles correspondants plus aux patrons de lecture des usagers, en fonction de leur profil. Dans chaque étape, les auteurs ont proposé des techniques pour diversifier les recommandations, basées sur des mesures de distance entre les vecteurs représentant les articles. La diversification avait pour objectif d'augmenter la couverture des thèmes tout en minimisant la similarité entre les articles dans la liste de recommandation finale.

Dans une autre approche, Li et Li (2013) ont proposé un modèle plus complexe, basé sur la constitution d'un hypergraphe qui modélise les relations entre les usagers, les articles et les différents types de traits discriminants provenant des articles comme les entités nommées et les mots. Dans cette approche, ils ont démontré une amélioration en précision et en diversité par rapport à celle obtenue dans l'étude antérieure (Li et coll., 2011). Le processus de recommandation se fait d'abord par le partitionnement de l'hypergraphe dans des sous-graphes qui expriment des relations plus cohésives entre les éléments de l'hypergraphe. Ensuite, un algorithme permet d'établir un rang entre les sous-graphes et le profil de l'usager cible, optimisant la prédiction des articles à recommander. Les profils des usagers ont été créés sur la base des articles consultés dans le passé.

Souhaitant apporter une solution au problème de l'exposition sélective, Gemmis et coll. (2015) ont proposé d'augmenter la sérendipité des SRAP, développant des techniques qui donnent au système la capacité de trouver des articles qui sont à la fois nouveaux et inattendus et que l'usager aurait du mal à découvrir par lui-même (Herlocker et coll., 2004). Ils appliquent une approche désignée « infusion de connaissances » (*Knowledge Infusion*) pour augmenter la sérendipité dans le système. Dans cette approche, le système enregistre en mémoire des relations entre les concepts présents dans les articles préférés des usagers avec d'autres concepts qui sont extraits de ressources ontologiques (ressources linguistiques et encyclopédiques qui décrivent des relations entre des concepts) comme *Wikipédia*. Les relations entre les concepts sont utilisées pour faire la prédiction de nouveaux articles pertinents. Les évaluations effectuées par les chercheurs démontrent que le niveau de précision et de sérendipité obtenu par leur approche est meilleur lorsque comparé avec des approches

basées sur le contenu. Une étude avec des usagers réels a permis de constater que 69 % des recommandations ont été jugées pertinentes et que 46 % ont été jugées comme ayant un bon degré de sérendipité.

Les recherches exposées ci-haut considèrent la diversité du point de vue de la variété thématique et de la variété de genres (Vargas et coll., 2014), mais ne tiennent pas compte de la question de l'exposition sélective en tant que telle, à l'exception de Gemmis et coll. (2015) qui mentionne explicitement le problème. Les techniques pour augmenter la diversité proposent des mesures pour maximiser la distance statistique entre un ensemble d'articles qui sont pertinents pour un usager donné, en choisissant certains traits discriminants comme les mots et les entités nommées. L'estimation statistique de la diversité ne mesure pas forcément la différence de perspectives entre deux articles sur le même thème. Les recherches qui ont abordé la question de la diversification en utilisant des techniques en fouille d'opinions sont surtout préoccupées par la question de la diversité d'opinions véhiculées en lien avec un thème spécifique. Nous allons les présenter dans la section suivante.

3.4 La fouille d'opinions appliquée à la diversification de contenus dans les SRAP

Nous avons identifié quelques travaux proposant l'utilisation de la fouille d'opinions pour diversifier les opinions et les perspectives des articles qui sont suggérés par les SRAP (Abbar et coll., 2013 ; Kawai et coll., 2007 ; Zhang et coll., 2009). Par rapport aux techniques de diversification exposées dans ce chapitre, ces travaux s'intéressent plus spécifiquement au développement d'un processus de recommandation pour varier les perspectives véhiculées à propos d'un thème qui intéresse l'usager. Avant de présenter ces études, nous voulons d'abord expliquer les différentes méthodes utilisées dans le domaine de la fouille d'opinions.

3.4.1 Fouille d'opinions : définitions et méthodes

La fouille d'opinions est définie comme un champ de recherche vouée à l'analyse d'opinions et de sentiments exprimés dans les textes (Liu, 2012 ; Pang et Lee, 2008). Elle concerne l'emploi d'un ensemble de techniques issues du traitement automatique du langage (TAL) et du champ de l'intelligence artificielle qui permettent l'analyse informatique de la

subjectivité dans les textes numériques. Ces techniques permettent de détecter les opinions véhiculées dans un texte ou de classer les textes selon les types de sentiments ou d'opinions véhiculées. Le domaine a un intérêt scientifique et économique non négligeable en fonction de la grande disponibilité d'informations qui sont partagées spontanément par les usagers dans plusieurs types de plateformes de contenu dans le Web, tels les forums de discussion, les blogues et les réseaux sociaux. Les recherches se sont démontrées particulièrement prolifiques pour l'étude des opinions de consommateurs visant la gestion de la réputation de marques et les processus décisionnels relatifs à l'achat de produits (Dave et coll., 2002 ; Kaiser et coll., 2011 ; Vernier et coll., 2009a ; Piryani et coll., 2017). Les articles de presse sont parmi les types de contenus les plus exploités dans les études du domaine (Piryani et coll., 2017).

La fouille d'opinions s'applique à un certain nombre de tâches, dont la classification automatique de textes en fonction de leur tonalité. Dans cette dernière, l'analyse informatique des textes numériques permet d'attribuer une classe positive, négative ou neutre à des textes d'après la présence de certains mots subjectifs. D'autres travaux de fouille d'opinions proposent une classification en fonction de sentiments ou d'émotions prédominantes dans les textes ou encore la détection de passages objectifs et subjectifs (Liu, 2012 ; Pang et Lee, 2008). Au sujet de la description des tâches de la fouille d'opinions, Grouin et coll. (2009) ont souligné l'importance de séparer dans les textes les segments véhiculant des informations factuelles de ceux véhiculant des valeurs, afin de pouvoir caractériser l'opinion générale ou la tonalité exprimée dans un texte :

> Une analyse d'opinion commence par la détection du caractère plus ou moins subjectif d'un texte ou d'un passage, c'est-à-dire, par déterminer s'il est porteur d'un « sentiment », d'un jugement, d'une opinion, ou au contraire de données essentiellement factuelles. Les parties de texte qui contiennent une opinion sont ensuite analysées pour donner une valeur à l'opinion exprimée, soit suivant une polarité positive/négative, soit suivant une échelle de valeurs. Enfin, le jugement exprimé sur un sujet particulier peut être influencé par, ou laisser transparaître, des opinions d'un type plus général par exemple une opinion politique. (Grouin et coll., 2009, p.35)

Le raffinement des techniques de fouille d'opinions dans les dernières années a permis la réalisation d'analyses plus complexes. Ainsi, en plus de la classification par tonalité,

certains travaux proposent des méthodes pour identifier les entités qui énoncent les opinions (Balahur et coll., 2013 ; Bethard et coll., 2004 ; Liu, 2012 ; Pang et Lee, 2008 ; Vechtomova, 2010 ; Vernier et coll., 2009b), pour créer des résumés automatiques d'opinions (Hu et Liu, 2004 ; Lun-Wei et coll., 2006 ; Soo-Min et Hovy, 2005), pour analyser les phrases évaluatives dans les textes (Jindal et Bing, 2006) ou encore pour repérer des opinions dans les moteurs de recherche (Vechtomova, 2010). Liu (2012) a proposé un classement de ces différentes directions de recherche en fonction de trois niveaux de complexité qui correspondent aux niveaux textuels sur lesquels les analyses portent :

1. Au niveau du document : le but est d'attribuer une classe (positive, négative ou neutre) à un document entier.

2. Au niveau de la phrase : le but est d'attribuer une classe (positive, négative ou neutre) à une phrase ou encore de distinguer les phrases subjectives des phrases objectives.

3. Au niveau d'une entité nommée ou d'un aspect d'une ou de plusieurs entités nommées : le but est d'attribuer une classe (positive, négative ou neutre) à un énoncé d'opinion sur une entité dans le texte ou à un aspect de cette entité (par exemple, une caractéristique d'un produit). Cette technique est souvent utilisée lorsqu'on veut repérer les opinions sur une entité quelconque.

Les expérimentations en fouille d'opinions s'inscrivent dans deux grandes familles de méthodes informatiques, soit la méthode non supervisée et la méthode supervisée. La méthode non supervisée (Chaovalit et Zhou, 2005 ; Maurel et coll., 2007) est faite par l'assignation d'une valeur sémantique à des mots susceptibles de véhiculer une opinion positive, négative ou neutre. Lorsque le traitement obtient l'ensemble des valeurs assignées aux mots, le score de tout l'article peut être calculé en appliquant des mesures d'agrégation comme la somme ou la moyenne (Chaovalit et Zhou, 2005). Dans cette dernière méthode, la partie la plus importante est la définition des valeurs qui vont être assignées aux mots, ce qui est fait par l'utilisation d'une ressource lexicale connue comme dictionnaire de sentiments, qui contient une liste de mots avec des valeurs sémantiques définies (par exemple, mots positifs ou mots négatifs).

La méthode supervisée se fait par apprentissage automatique, opération permettant d'extraire des modèles statistiques des articles représentés dans le format vectoriel et classés préalablement par un humain. Le processus d'apprentissage et de création des classifieurs est effectué par des algorithmes développés dans le domaine de l'intelligence artificielle, par exemple le classifieur bayésien naïf (NB) ou machine à vecteurs de support (SVM)[10]. Ces derniers mémorisent les patrons statistiques des classes en fonction des traits utilisés pour représenter les articles et ensuite prédisent la classe de nouveaux articles (Forest et coll., 2009). La méthode supervisée peut également utiliser un dictionnaire de sentiments pour extraire des articles des traits discriminants correspondant à des mots subjectifs. Lorsque l'utilisation de dictionnaires de sentiments est proposée, peu importe la méthode choisie, elle est considérée comme une méthode symbolique (Maurel et coll., 2007).

Turney (2002) a présenté une méthode non supervisée basée sur l'analyse de phrases subjectives pour détecter les critiques positives et négatives octroyées à des produits. L'auteur a employé un Étiqueteur de la Partie du Discours (EPD) pour attribuer aux mots contenus dans les phrases une étiquette correspondant à sa catégorie linguistique (syntaxique et morphologique). La méthode proposée a d'abord fait la distinction entre les passages subjectifs et objectifs dans le texte, en fonction des patrons de combinaisons morphosyntaxiques révélés par l'EPD. Par la suite, l'orientation sémantique de chaque phrase a été estimée à l'aide d'une ressource lexicale contenant des valeurs sémantiques positives et négatives pré-assignées à un ensemble de mots subjectifs. L'orientation sémantique moyenne de toutes les phrases a été finalement calculée à l'aide d'un algorithme qui a attribué une tendance de tonalité au document (positive ou négative).

Du côté des méthodes supervisées, un des travaux les plus connus est celui de Pang et coll. (2002) sur un corpus constitué de critiques de films. Un classifieur a été constitué à l'aide d'exemples de données dans un corpus d'entraînement composé par des documents préalablement classés comme positifs ou négatifs. Les expérimentations ont démontré une très bonne performance avec l'utilisation de classifieurs bayésien naïf ou machine à vecteurs de support pour l'attribution des classes positives et négatives, avec une précision de 81 % et

[10] Nous allons expliquer le fonctionnement de ces algorithmes dans le chapitre 4.

82,9 % respectivement (Pang et coll., 2002). Dans une étude ultérieure, Pang et Lee (2004) ont amélioré la performance de la classification en présentant une méthode pour exclure de l'analyse des « phrases objectives », celles qui ne présentaient pas une structure syntaxique susceptible de porter une évaluation sur un objet. Les résultats obtenus ont été meilleurs que l'expérimentation proposée en 2002.

Piryani et coll. (2017) ont conduit une étude scientométrique au sujet de la fouille d'opinions et de l'analyse des sentiments sur 488 articles repérés dans la plateforme *Web of Science*. Cette étude comprend des techniques d'analyse automatique de données provenant de la plateforme *Web of Science,* pour comprendre la trajectoire des recherches effectuées dans le domaine. La recherche scientométrique comprenait des articles scientifiques, des comptes rendus, des actes de conférences, des éditoriaux et des chapitres de livres, recueillis entre 2000 et 2015 et en langue anglaise. Les auteurs ont découvert que 92,44 % des approches basées sur l'apprentissage automatique (méthode supervisée) portent sur l'analyse au niveau du document, selon la nomenclature proposée par Liu (2012). La classification de documents est aussi assez dominante dans les méthodes symboliques, couvrant 65,22 % des articles.

Le recours à des ressources linguistiques comme les dictionnaires de sentiments est particulièrement fréquent dans le domaine de la fouille d'opinions (Kawai et coll., 2007 ; Maurel et coll., 2007 ; Torres-Moreno et coll., 2007 ; Petz et coll., 2014 ; et et coll., 2007, 2009a, 2009b ; Zhang et coll., 2009) et leur emploi est envisagé pour répondre au besoin de caractériser de façon précise les segments jugés pertinents pour l'analyse automatique de l'opinion. L'utilisation de ces ressources est majoritairement présente dans les méthodes non supervisées, mais certains auteurs soulignent que les prétraitements des textes des documents permettent d'atteindre une meilleure performance des algorithmes pour la méthode supervisée (Maurel et coll., 2007 ; Petz et coll., 2014). Dans la méthode non supervisée, des ressources comme *SENTIWORDNET* (Esuli et Sebastiani, 2006) sont également proposées dans les recherches (Balahur et Steinberger, 2012 ; Lee et coll., 2008 ; Ohana et Tierney, 2009). Ces ressources contiennent un ensemble de mots subjectifs avec des valeurs pré-assignées et sont employés pour annoter les corpus et pour attribuer la valeur sémantique positive, négative ou neutre aux mots subjectifs présents.

La construction de dictionnaires de sentiments à partir des corpus par des techniques automatiques ou semi-automatiques (Dave et coll., 2002 ; Vernier et coll., 2009b) est également fréquente. Par exemple, Dave et coll. (2002) ont proposé un algorithme pour attribuer une note à chaque mot subjectif d'un échantillon de documents, en tenant compte du nombre de fois que ces mots apparaissent dans les documents associés à une classe positive ou négative. La valeur du mot était un coefficient allant de -1 à 1 qui indiquait sa proximité avec chaque classe. Ils ont ensuite utilisé cette ressource lexicale construite de façon automatique pour détecter les opinions dans les critiques de produits de consommation.

Les travaux en fouille d'opinions proposent également d'autres types de prétraitements des textes pour optimiser leur représentation pour les algorithmes, comme la reconnaissance d'entités nommées, l'analyse lexicale et l'analyse morphologique (Acosta et Bittar, 2007). Aussi, sont proposées des procédures de filtrage du corpus visant à sélectionner les segments textuels plus pertinents à l'analyse des opinions dans les textes (Abbar et coll., 2013 ; Acosta et Bittar, 2007 ; Charton et Acuna-Agost, 2007 ; Maurel et coll., 2007)

L'amélioration attestée par les approches symboliques motive des directions de recherches qui revendiquent explicitement un cadre théorique linguistique pour orienter le traitement des textes numériques qui véhiculent des opinions (Eensoo et Valette, 2014a). À l'inverse des autres approches mentionnées, celles-ci ont pour caractéristique d'analyser les textes du point de vue du positionnement des énonciateurs dans les discours et par la recherche de marques d'énonciation, d'argumentation ou axiologiques[11] qui précisent ce positionnement (Eensoo et Valette, 2014a, 2014b, 2015 ; Vernier et coll., 2007, 2009b). Par exemple, Vernier et coll. (2009b) proposent une méthode pour repérer et catégoriser automatiquement des structures évaluatives dans les billets de blogues afin de découvrir et catégoriser les opinions émises dans les textes à propos d'un concept cible, comme une personne ou un produit. Ces

[11] Vernier et coll. (2009b) associent l'axiologie à la notion de polarité positive/négative : « [elle] comporte les évaluations référant aux champs d'expérience humaine : esthétique (beau/laid), pragmatique (utile/inutile, important/dérisoire, efficace/inefficace), cognitif ou intellectuel (intéressant/inintéressant), éthique ou morale (bien/mal, bon/mauvais), hédonique-affectif (agréable/désagréable, plaisir/souffrance) » (Vernier et coll., 2009b, p.53)

auteurs s'appuient sur des théories d'études sur l'énonciation et sur l'axiologie pour modéliser les principales caractéristiques des phénomènes évaluatifs. En se basant sur les études sur l'axiologie et l'ouvrage de Charaudeau (1992), ils ont annoté manuellement un ensemble de marqueurs d'un corpus d'entraînement composé de billets de blogues, afin de regrouper dans un dictionnaire un ensemble de mots et d'expressions pertinentes pour l'analyse des textes. À partir de ce dictionnaire, ils ont extrait les structures évaluatives d'un corpus de test et ils ont ensuite catégorisé automatiquement les phrases de ce corpus de test en fonction de leur configuration énonciative (implicite ou explicite), de la modalité de l'opinion qu'ils présentaient (opinion, appréciation, accord/désaccord) et de leur aspect axiologique (favorable, défavorable ou ambiguë). Les résultats démontrent un bon taux de précision pour la détection, mais un taux de rappel relativement bas. Ce dernier a été causé, selon les auteurs, d'une part par les caractéristiques du corpus sélectionné (qui présentait des opinions surtout argumentatives) et d'autre part parce que les évaluations détectées ne concernaient pas toujours les concepts cibles choisis.

Dans la même lignée, les travaux de Eensoo et Valette (2014a, 2014b, 2015) et Valette (2004) revendiquent un cadre théorique précis, basé sur les études en sémantique de François Rastier (Rastier, 1987, 2001 ; Rastier et coll., 1994). La principale différence de leur approche comparée à celle de Vernier et coll. (2009b) c'est qu'ils n'utilisent pas de dictionnaires pour extraire dans les textes les données textuelles correspondantes à des marques linguistiques prédéfinies. Ils préconisent plutôt d'utiliser celles qui sont attestées dans le corpus et qui font état du mode particulier par lesquels des locuteurs réels emploient le langage dans le corpus en question. Ainsi, la recherche de traits discriminants pour les tâches de fouille d'opinions, qu'ils appellent « critères textuels », est faite de manière exploratoire en utilisant des calculs textométriques. Ces calculs, utilisés dans le domaine de la linguistique de corpus pour assister l'interprétation de textes, sont employés dans les recherches de fouille d'opinions pour comparer des sous-corpus constitués de documents qui véhiculent des opinions différentes sur un même thème. Ils permettent ainsi d'identifier les différences entre ces sous-corpus, différences qui sont attestées par la spécificité des critères textuels présents dans chacun.

3.4.2 Fouille d'opinions et diversification

Nous avons constaté une prépondérance d'approches symboliques dans l'ensemble des travaux qui appliquent la fouille d'opinions pour diversifier la recommandation d'articles dans les SRAP. Ceux-ci procèdent à une distinction des articles sur un même thème en fonction des mots subjectifs présents. Dans ces travaux, la notion d'opinion est plutôt associée à la « perspective journalistique » ou au « point de vue » du journaliste. Le choix d'observables linguistiques pour représenter cette perspective journalistique pour les algorithmes est proposé différemment dans les recherches. Le recours à des dictionnaires de sentiments pour attribuer des valeurs sémantiques au vocabulaire subjectif des textes ou pour extraire des textes les mots subjectifs qui serviront à les représenter dans le modèle vectoriel est très présent.

Abbar et coll. (2013) proposent une méthode non supervisée pour le développement d'un SRAP dans lequel les articles à recommander devraient présenter une mesure de diversité maximale par rapport à un article source thématiquement similaire consulté par l'usager. Leur méthode consiste à maximiser la diversité d'articles thématiquement analogues dans le processus de recommandation, en analysant leurs différences par rapport aux sentiments et aux entités (personnes ou organisations) mentionnés dans les commentaires des usagers à propos de l'article. Les auteurs défendent l'hypothèse selon laquelle les commentaires laissés par les usagers sont plus révélateurs des différences entre deux articles similaires. Ils soutiennent que les usagers ont tendance à amplifier ces différences, en rendant visible par exemple le biais du journaliste par rapport à l'événement reporté.

La mesure de diversité développée par Abbar et coll. (2013) est basée sur la distance euclidienne entre les articles représentés dans un espace vectoriel. Ils ont formulé le problème de la façon suivante : chercher des articles candidats qui sont à une distance de pertinence *r* de l'article que l'usager consulte (article-source) et, parmi ceux-ci, identifier le nombre *k* articles les plus divers selon une mesure de diversité qui capte les différences entre les entités nommées et les sentiments mentionnés dans les commentaires. Ainsi, les auteurs considèrent deux distances entre les articles : la distance de pertinence, qui détermine un ensemble d'articles similaires à l'article-source de l'usager et la distance de diversité, qui mesure la « différence » entre les commentaires de l'article-source et les autres articles pertinents.

L'utilisation d'un dictionnaire de sentiments a permis d'identifier les sentiments positifs et négatifs exprimés dans les commentaires.

En comparant les scores de diversité obtenus à l'aide de l'algorithme développé pour l'expérimentation, Abbar et coll. (2013) démontrent que les sentiments et les entités nommées présentent des performances comparables. Une évaluation conduite avec des usagers réels révèle par contre la que génération de la liste d'articles recommandés en fonction des sentiments exprimés, ne capte pas très bien la diversité des articles politiques. Ils constatent que pour le corpus d'articles politiques, la diversité est plus évidente si les commentaires sont représentés en fonction des entités nommées mentionnées dans les commentaires.

D'autres auteurs ont exploré les émotions exprimées dans les articles de presse pour diversifier les recommandations dans le SRAP. Kawai et coll. (2007) et Zhang et coll. (2009) ont proposé une méthode non supervisée pour calculer la divergence de perspective entre deux articles de presse portant sur un même thème sur la base des émotions exprimées dans les textes. Leur méthode consistait à extraire les mots subjectifs du corpus d'articles de presse à l'aide d'un dictionnaire de tonalité contenant le poids de chaque mot subjectif relativement au sentiment exprimé par quatre dimensions d'émotions prédéfinies : Joie/Tristesse, Acceptation/Dégoût, Anticipation/Surprise et Peur/Colère[12]. Les algorithmes créaient un vecteur de sentiments pour chaque article qui traduisait l'inclination de l'ensemble des mots subjectifs vers un des deux sentiments de chacune des dimensions. La méthode permettait de comparer les articles selon le type d'émotion qu'ils véhiculaient (information fournie par la dimension) ainsi que la polarité de l'émotion comprise dans chaque dimension.

Kawai et coll. (2007) ont proposé une méthode pour la détection d'émotions dans le cadre du système *Fair News Reader* afin de recommander des articles contenant des sentiments discordants aux patrons de consultation des usagers du système. Pour cela, en plus de la génération de vecteurs pour chaque article, le système constituait des vecteurs de

[12] Dans les recherches de Kawai et coll. (2007) et de Zhang et coll. (2009), les valeurs attribuées aux mots du corpus correspondent à des dimensions d'émotion prédéfinies tels que Joie/Tristesse, Acceptation/Dégoût, Anticipation/Surprise et Peur/Colère. Ces catégories s'inspirent de la classification des huit éléments fondamentaux des émotions humaines du psychologue Plutchik (1991).

sentiments pour les profils des usagers, selon leur historique de consultation. La recommandation se faisait ensuite en fonction d'une mesure d'écart type entre ces deux vecteurs. Par exemple, un article avec un vecteur de sentiments exprimant la « Joie » était recommandé à un usager dont l'historique de consultation indiquait la prédominance d'articles exprimant la « Tristesse ». Zhang et coll. (2009) ont exploré la détection d'émotions dans le cadre du système *Sentiment Map,* qui permettait de visualiser les articles divergents dans une carte géographique interactive et d'avoir un aperçu complet des tendances d'émotions pour une couverture de presse sur un thème spécifique.

La méthode proposée par Kawai et coll. (2007) et Zhang et coll. (2009) est contestable dans ce qu'elle entend comme définition de « perspective journalistique ». Celle-ci est réduite à l'impact émotionnel qu'une nouvelle peut susciter chez le lecteur. Or, chercher l'impact émotionnel des articles sur un même thème permet de discriminer les « bonnes » nouvelles des « mauvaises » nouvelles, mais ne permet pas nécessairement de distinguer les articles en fonction de la perspective journalistique véhiculée. Cet extrait de Kawai et coll. (2007) révèle comment les auteurs conçoivent l'utilité de leur technique de diversification :

> For example, a search using the keywords Iraq and terror with existing news portal sites will return articles about Iraq with such topics as suicide bombing in Iraq, which are likely to create a sad sentiment. A system that can recommend, for example, articles that create a happy sentiment, such as released hostage in Iraq, should thus be useful. (Kawai et coll., 2007, p. 613)

L'exemple de certains articles d'opinion publiés dans la presse sur le conflit israélo-palestinien démontre qu'opposer les articles sur la base des émotions n'est pas la façon la plus appropriée de représenter la divergence d'opinions. Il serait bien plus facile de trouver des articles qui défendent une des causes, palestinienne ou israélienne, et qui emploient un ton de « colère » ou de « peur » semblable. Wei-Hao et Hauptmann (2006) constatent également ce problème dans leur étude :

> The high but almost equivalent number of subjective sentences in two perspectives suggests that perspective is largely expressed in subjective language but subjectivity ratio is not enough to tell if two document collections are written from the same

(Palestinian v.s. Palestinian) or different perspectives (Palestinian v.s. Israeli). (Wei-Hao et Hauptmann, 2006, p. 1058)

Wei-Hao et Hauptmann (2006) ont d'ailleurs démontré que des méthodes statistiques qui ne considèrent pas les mots subjectifs sont aussi efficaces pour détecter les articles véhiculant des opinions divergentes. Les chercheurs ont présenté un test pour déterminer si deux collections d'articles étaient écrites selon des perspectives opposées, en se basant sur un calcul de divergence statistique et en considérant des substantifs. Pour réaliser le test, ils ont constitué trois corpus : le premier contenant un ensemble d'articles sur le conflit israélo-palestinien du site *bitterlemons.org*, chacun classé selon la perspective défendue (palestinienne ou israélienne) ; le deuxième contenant des transcriptions des débats des candidats Bush et Kerry lors des élections présidentielles américaines de 2004, classés par auteur ; et le troisième contenant un ensemble d'articles provenant de l'agence *Reuters*, chacun classé dans sept thématiques distinctes. Les chercheurs voulaient observer si la divergence statistique observée entre les paires formées à l'intérieur des deux premiers corpus, qui reflétait des perspectives divergentes sur un même thème, était importante par rapport à celle observée entre les sept thématiques du troisième corpus, en fonction des substantifs présents dans le texte (sans discriminer les mots subjectifs). En utilisant la mesure Kullback-Liebler[13], ils ont constaté que l'écart statistique entre les paires d'articles ayant des perspectives opposées était comparable à l'écart observé entre les différentes thématiques du troisième corpus, suggérant que des informations lexicales diverses peuvent être efficaces pour distinguer les opinions dans les textes.

Certains travaux dans le champ de l'apprentissage automatique soulignent l'importance de considérer des facteurs autres que les mots subjectifs pour déterminer si deux articles sont différents en termes de perspectives véhiculées. Fortuna et coll. (2009) par exemple, se sont intéressés à la détection du biais éditorial dans les articles publiés par les agences de presse. En utilisant des techniques d'apprentissage automatique, ces auteurs ont démontré qu'il est

[13] La divergence Kullback-Liebler est une mesure statistique pour calculer la dissimilarité entre deux distributions de probabilités.

possible de prédire l'appartenance d'un article à une agence en se basant sur les substantifs employés dans l'article. Dans leurs expérimentations, les auteurs utilisent un corpus d'articles provenant des sites de CNN et d'Al Jazeera traitant des enjeux politiques du Moyen-Orient. Une analyse statistique du lexique employé permet de constater la différence et le biais connu de chacune de ces sources quant au traitement éditorial de cet enjeu. Par exemple, CNN privilégie les mots *insurgency*, *militants* et *terrorists* tandis qu'Al Jazeera préfère parler de *resistance* et de *rebels*.

Il devient clair que la perspective d'un journaliste est aussi le résultat d'un choix lexical servant à cadrer l'enjeu en question pour promouvoir certaines croyances et valeurs. Les partisans politiques connaissent bien le pouvoir que représente l'utilisation sélective du langage pour encourager des attitudes qui sont plus en accord avec leurs idées. Par exemple, Schuldt et coll. (2011) démontrent que l'emploi de « réchauffement global » et « changement climatique » dans les discours publics révèle en grande mesure dans quelle perspective ce problème environnemental est abordé. C'est ainsi que « changement climatique » est plutôt employé dans un contexte où la responsabilité humaine est identifiée comme étant la cause du problème tandis que « réchauffement global » est utilisé davantage pour associer ce dernier à des causes naturelles. Les recherches de fouille d'opinions doivent prendre acte de cette particularité des textes argumentatifs s'ils veulent proposer des méthodes de détection efficaces du type d'opinion véhiculée par ces derniers.

Sur la base de cette revue de la littérature, nous pouvons constater la nécessité de comprendre plus précisément comment les journalistes et les chroniqueurs s'expriment lorsqu'ils essayent d'argumenter ou de défendre une opinion et plus généralement, comment les recherches en fouille d'opinions définissent leur objet d'étude. Certains genres journalistiques ont une visée argumentative plus explicite, mais pas nécessairement empreinte d'affects ou d'émotions. D'autres genres journalistiques, comme les nouvelles, n'ont pas de visée argumentative explicite et pourtant, une personne qui l'interprète peut repérer aisément plusieurs stratégies persuasives.

L'identification des prémisses et des présuppositions sur lesquels se base la définition des concepts et des objets scientifiques est importante pour comprendre les implications sur la conduction de la recherche scientifique, ainsi que les limites et les possibilités offertes à

l'investigation (Dick, 1999). Dans un article publié en 2011, Eensoo et Valette (2011) entament une réflexion sur les bases épistémologiques de la fouille d'opinions et sur comment celle-ci se présente comme une extension des critiques sociologiques qui ont été formulées par rapport aux sondages des opinions dans les années 30. Nous abordons ce sujet dans la section suivante.

3.5 Mesurer l'opinion : l'évolution du sondage jusqu'à la fouille d'opinions

L'intérêt pour le phénomène de l'opinion est apparu de façon accentuée dans les études sur l'opinion publique, avec l'essor des instituts et techniques de sondage d'opinion dans les années 1930. Le sondage est devenu depuis ce temps un instrument privilégié pour connaître « ce que les gens pensent ». Il est particulièrement utilisé dans le domaine du marketing pour l'analyse des préférences de consommation et, plus tard, dans le domaine politique, pour la recherche et la diffusion des intentions de vote lors des élections. Les techniques de sondage d'opinion voulaient proposer une « science de l'opinion publique » par le biais d'un ensemble de méthodes quantitatives qui contribueraient à la connaissance des phénomènes sociaux (Eensoo et coll., 2011).

Dans son ouvrage « La fabrique de l'opinion », Blondiaux (1998) analyse le processus historique de l'assimilation de l'opinion publique aux résultats des sondages et démontre comment les discussions menées dans les années 1920 et 1930 sur l'élaboration d'instruments statistiques de sondage ont oblitéré la définition même de ce qu'il tentait de mesurer. L'opinion a été très rapidement assimilée à une expression verbale d'une attitude, que l'auteur définit comme étant des dispositions individuelles « acquises et durables qui jouent comme matrices de comportements envers certains objets ou certaines situations » (Blondiaux, 1998, p. 131). Dépourvue de toute complexité à l'égard des déterminations sociales qui engendre le phénomène de l'opinion publique, l'opinion a été assimilée à une prédisposition mentale envers un objet, de nature quantifiable et assujettie à des opérations mathématiques (Blondiaux, 1998) :

> Avant même l'apparition des sondages, l'urgence n'est déjà plus de penser l'opinion publique en tant que phénomène social, de spéculer à l'infini sur la solidité de ce pilier de la théorie démocratique, mais de compter. Elle est de collecter et

> d'additionner les opinions individuelles. Un tel événement est le signe qu'une révolution est en train de s'opérer dans la réflexion de l'objet opinion publique. L'objet est en train d'émigrer et de se déplacer, de quitter les sphères de la théorie politique pour descendre dans les salles de classe, là où les psychologues sociaux mettent en œuvre différentes procédures de mesure empirique des attitudes individuelles, calquées sur les tests d'intelligence. (Blondiaux, 1998, p. 129)

Plusieurs sociologues ont mis en cause cette notion d'opinion publique et son équivalence à une somme d'opinions individuelles (Blondiaux, 1998 ; Bourdieu, 1973 ; Hermet et coll., 2000 ; Rieffel, 2010). Rieffel (2010) par exemple, propose que l'opinion publique n'est pas définie comme l'agrégation d'opinions individuelles, ni la manifestation de simples croyances, mais la résultante d'une « élaboration concertée de points de vue, d'une confrontation négociée et sans cesse renouvelée » (p.36). Pour sa part, Hermet et coll. (2000) postulent qu'en termes réalistes, l'opinion publique qui conçoit les sondages n'existe pas puisqu'il est impossible d'agréger des jugements qui varient de nature et d'intensité : en effet, les réponses à un sondeur peuvent provenir de convictions fermes, d'autres de simples impressions.

Bon nombre d'objectifs et procédures associés aux sondages d'opinion se voient aujourd'hui transposés dans la fouille d'opinions, en faisant de cette dernière une solution technologique actuelle pour aller chercher dans le Web ce que les gens pensent (Eensoo et coll., 2011). Les techniques de fouille d'opinions sont majoritairement employées dans les applications marketing de surveillance de marques destinées à dégager et à analyser les avis des consommateurs à propos de produits. Le développement du domaine s'oriente de plus en plus vers la réalisation de ce genre d'application. Cette orientation épistémologique rappelle par plusieurs aspects la discussion sur la légitimité des outils de mesure de l'opinion utilisés dans les sondages et la nécessité d'une réflexion plus approfondie sur ce qui constitue l'opinion et sur les instruments les mieux adaptés à son analyse.

Un parallèle peut être établi entre la fouille d'opinions et les sondages au niveau de l'assimilation de l'opinion « générale » comme la somme d'opinions individuelles (Eensoo et Valette, 2011). Dans la fouille d'opinions, cette assimilation concerne particulièrement la conception de l'expression textuelle de l'opinion comme la somme d'énoncés véhiculant une opinion. En effet, la plupart des travaux en fouille d'opinions définissent l'opinion comme

étant une phrase qui pourrait répondre à la question « Qu'est-ce que X pense de Y ? » (Bethard et coll., 2004 ; Pang et Lee, 2008). Cette notion n'est pas sans similitudes avec celle que Blondiaux (1998) avait identifiée dans les travaux de psychologie sociale sur les sondages, c'est-à-dire une prédisposition mentale envers un objet ou une situation (Blondiaux, 1998). Si dans les sondages, l'opinion publique est l'agrégation des réponses individuelles sur une question, dans le contexte de la fouille d'opinions, cette même conception se voit transposée dans le texte. Ainsi, l'opinion serait l'ensemble d'énoncés où un sujet X émet une opinion positive ou négative sur l'entité Y.

En analysant un ensemble de recherches dans le domaine de la fouille d'opinions, Eensoo et coll. (2011) constatent que la définition de l'opinion en tant qu'objet linguistique est épistémologiquement très influencée par une vision logiciste du langage. Ils prolongent ce débat en faisant une critique de la prépondérance du positivisme logique comme base épistémologique des principales applications informatiques qui se penchent sur le fonctionnement de la langue. Le positivisme logique est un courant épistémologique qui a tenté d'élaborer un langage idéal pour représenter la science, basé sur le principe de l'économie. Les philosophes du Cercle de Vienne croyaient que la connaissance est tout phénomène sensoriel qui peut être exprimé dans le langage par une proposition logiquement vraie. En ce sens, les concepts ne pouvant pas être traduits dans des énoncés logiques (les concepts métaphysiques et éthiques notamment) devraient être rejetés par le langage scientifique, puisqu'ils présentent un obstacle à l'expression de la pensée pure.

Cette perspective logique du langage a été notamment explorée par des domaines comme l'ingénierie des connaissances, discipline qui cherche à donner une représentation formelle de la langue qui puisse être exploitable par les machines (Bachimont, 2000). L'essor de cette discipline a marqué notamment l'émergence dans le domaine informatique des ontologies : des schémas de concepts interconnectés par des propriétés sémantiques et syntaxiques qui sont utilisés en intelligence artificielle pour perfectionner les mécanismes de raisonnement déductif, de classification automatique, d'interopérabilité entre les systèmes informatiques, de recherche d'information, entre autres. L'utilisation de ressources lexicales ontologiques comme *SENTIWORDNET* (Balahur et coll., 2013 ; Lee et coll., 2008 ; Ohana et

Tierney, 2009), pour attribuer au vocabulaire subjectif des valeurs sémantiques, atteste de l'influence du positivisme logique dans le domaine.

Le positivisme logique, selon Eensoo et coll. (2011), procède à une nette séparation entre les « faits », ce qui est observable ou représentable logiquement dans les énoncés, et les « valeurs », tout concept qui ne peut pas prendre la forme d'un rapport verbal objectif, ancré sur l'expérience concrète. Selon Eensoo et coll. (2011), cette influence est perceptible dans l'opposition fait/valeur et elle a été transposée dans les tâches courantes de la fouille d'opinions. Les recherches proposent de fragmenter les textes dans des segments linguistiques « objectifs » ou « subjectifs » et de porter l'analyse de l'opinion sur les derniers seulement, parce qu'ils sont plus susceptibles de véhiculer des évaluations et des jugements. Les auteurs critiquent la faiblesse de cet apriorisme à travers des exemples tirés des recherches de fouille d'opinions, qui montrent comment certaines valorisations sont véhiculées par des segments dits « factuels » :

> (…) le syntagme *long battery life*, considéré a priori comme factuel, assigne en fait une valeur positive à l'ordinateur, et cette valeur est contenue dans la dénomination puisque la durée d'autonomie de la batterie figure parmi les critères positifs importants pour cet outil technique. (Eensoo et coll., 2011, p. 18)

La vision logiciste de la langue a généré deux effets sur la notion d'opinion telle que comprise par les recherches de fouille d'opinions. D'une part, l'hypothèse selon laquelle les mots utilisés pour exprimer une évaluation ou un jugement sont invariables d'un point de vue sémantique. Dans le cadre de l'analyse de la tonalité par exemple, où les techniques cherchent à classer des textes véhiculant l'opinion comme positive ou négative, la valeur sémantique des adjectifs est présupposée stable et indépendante de tout contexte d'énonciation. Liu (2001) note l'inconvénient de ce présupposé sur les applications de fouille d'opinions quand il explique le changement de connotation du mot « imprévisible » par rapport à deux contextes étudiés, soit les critiques de films et de voitures. Dans le premier, l'expression « ce film est imprévisible » a une connotation positive tandis que dans le domaine de la voiture, « cette voiture est imprévisible » offre une appréciation négative. Cela démontre que la connotation des mots est dépendante de son contexte d'utilisation et qu'il est difficile de généraliser les

valeurs sémantiques qui sont pourtant présupposées stables par les dictionnaires de sentiments qui sont utilisés dans les traitements informatiques du domaine.

L'autre effet serait la supposition d'une compositionnalité du sens mis en avant par les méthodes de fouille d'opinions. Ce principe, comme l'explique Polguère (2008), « veut qu'un énoncé linguistique soit directement calculable — dans sa composition lexicale et sa structure syntaxique — à partir de la combinaison du sens de chacun de ses constituants » (Polguère, 2008, p. 57). Ainsi, dans les tâches qui visent la classification automatique de textes, l'opinion véhiculée dans un texte est vue comme la somme des différentes parties du texte (plus généralement les phrases) pour lesquelles on attribue une valeur positive ou négative. Ce postulat est également une source de problèmes pour les applications de la fouille d'opinions, comme le soulignent Eensoo et coll. (2011) :

> Ainsi, Turney (2002) fait remarquer que la stratégie consistant à repérer les morceaux de texte supposés positifs ou négatifs et à les additionner pour déterminer la valeur finale du signe ne donne pas les mêmes résultats selon le domaine (artefacts ou objets culturels). Par exemple, lorsque l'on traite les opinions relatives à des voitures par une approche compositionnelle, la valeur de l'entier est proche de la somme des parties, mais ce n'est pas le cas pour les films. Dans ce dernier cas, le jugement porté a un caractère holistique (le jugement global est plus complexe que la somme des parties). (Eensoo et coll., 2011, p. 19)

Eensoo et coll. (2011) prônent la nécessité de rétablir la stabilité épistémologique du domaine de la fouille d'opinions, distinguant parmi les différentes choses que l'on peut dénommer opinion ce qui relève d'une appréciation personnelle (bon/mauvais), d'une évaluation (recommandé/non recommandé) et d'un exercice de raisonnement (arguments, pour/contre). C'est d'ailleurs dans ce sens que ces auteurs préconisent l'importance de considérer les différents genres textuels dans le domaine de la fouille d'opinions, ainsi que le rôle de l'énonciateur dans l'émission du message. Tout acte communicatif est encadré par l'objectif précis d'un sujet qui parle et qui choisit son mode d'expression en fonction d'un genre textuel et de la pratique sociale où il s'insère. Ce retour à l'énonciateur implique d'assigner à la fouille d'opinions un rôle d'identification et de restitution de l'émetteur du message, en opérant un rapprochement à des théories énonciatives où l'opinion est vue comme

une construction collective « dont on retrouve des traces dans les matériaux textuels émis par et pour des membres des groupes sociaux visés » (Eensoo et coll., 2011, p. 29-30).

Ce changement de paradigme considère l'expression écrite de l'opinion à l'intérieur d'un cadre communicationnel et s'oppose d'emblée à l'hypothèse de compositionnalité du sens préconisé par la linguistique formelle. Linguistique formelle qui impose, elle, dans le cadre de la fouille d'opinions, l'analyse individuelle de phrases pour inférer la valeur de l'opinion globale d'un texte. Eensoo et Valette (2015a) expliquent que jusqu'aux années 2000, les applications qui concernaient la thématique, la terminologie et le lexique, comme la désambiguïsation lexicale et l'EPD, relevaient d'une sémantique de la phrase et que les modèles formels de l'analyse syntaxique se sont avérés efficaces pour l'extraction de l'information précise, par exemple dans les applications de recherche d'information. Elles étaient, par contre, moins efficaces dans l'analyse de grands corpus, notamment pour la classification de textes. Pourtant, dans les tâches de classification automatique d'opinions, les applications sont davantage confrontées à l'unité texte, exigeant un regard plus global sur cet objet.

> L'essor des applications en fouille de textes subjectifs dans le courant des années 2000 (fouille d'opinions, analyse des sentiments, détection des émotions, etc.) implique également une évolution des tâches : alors que le TAL privilégiait les unités référentielles et souvent lexicales (entités nommées, concepts, termes, thèmes), il est aujourd'hui confronté à des valeurs. Certes, les méthodes d'extraction et de classification n'ont guère évolué : dans beaucoup d'applications, les adjectifs sont aux textes subjectifs ce que les substantifs sont aux concepts (…) et on a tendance à appliquer aux premières les méthodes qui ont fait leur preuve sur les secondes. Dépasser le « lexicalisme » du TAL est un des enjeux de la linguistique, car l'inventaire des objets de la linguistique susceptibles d'être appréhendés par le TAL est, en effet, loin d'être clos. Il est par exemple probable que les contraintes de genres, de discours, que la structure actancielle des textes et que le schéma de la communication soient utiles à l'interprétation des émotions, sentiments ou des opinions (Eensoo et Valette, 2015a, p. 2).

Eensoo et Valette (2015a) préconisent que les questions posées par le TAL évoluent d'un paradigme logico-formel vers une problématique herméneutique et interprétative, permettant d'envisager le texte comme un objet complexe et déterminé par le projet de communication et par la pratique sociale où il s'insère :

> En somme, tout se passe comme si les questions qui se posent au TAL évoluaient d'une problématique logico-formelle dominée par le primat référentiel et le choix historique de la phrase (et son avatar : l'énoncé) comme unité d'analyse, vers une problématique herméneutique et interprétative dont l'objet est la réception et l'interprétation des textes considérés comme des unités de sens complexes déterminées par un projet de communication. (...) Ce moment de flottement paradigmatique est l'occasion d'esquisser des méthodes fondées non pas sur les présupposés théoriques du paradigme logico-grammatical, mais sur un paradigme herméneutique et interprétatif peu exploré encore en TAL. (Eensoo et Valette, 2015a, p. 3)

3.6 Conclusion

Dans cette section, nous avons évoqué un certain nombre de problèmes générés par la personnalisation dans les SRAP relatifs à l'homogénéisation de contenus. La personnalisation de ces systèmes est vue comme un facteur qui contribue au phénomène de l'exposition sélective, puisqu'elle renforce la tendance des individus à faire des choix partisans en termes de sélection d'informations et de sources journalistiques (Mutz et Young, 2011).

Les études qui proposent la diversification dans les SRAP tendent à définir la diverstié comme une distance statistique entre les vecteurs des articles qui sont recommandés pour un usager cible (Desarkar et Shinde, 2014 ; Li et Li, 2013 ; Li et coll., 2011). Ainsi, la diversité dans la recommandation est associée à une différence interne de la liste de recommandations dans laquelle la distance entre les articles pertinents recommandés est maximisée. Dans ces études, la diversité n'est pas nécessairement associée à la notion de diversité d'opinions sur un thème qui intéresse l'usager. Les recherches en fouille d'opinions explorent de façon plus poussée cette dernière voie, en proposant des techniques pour distinguer les articles sur un même thème en fonction d'indices linguistiques qui expriment la subjectivité, comme les émotions (Abbar et coll., 2013 ; Kawai et coll., 2007 ; Zhang et coll., 2009). Dans ces travaux, les méthodes proposées préconisent la représentation des articles de presse (principalement les nouvelles) par un ensemble de mots subjectifs ou d'émotions associés à des mots subjectifs. Cette représentation produit une performance inférieure pour la recommandation d'articles de domaines comme la politique (Abbar et coll., 2013) et les méthodes ne permettent pas de distinguer les articles en fonction des différences d'opinions, comme elles le prétendent, mais en fonction des types d'émotions que les articles peuvent susciter chez le lecteur (Kawai et

coll., 2007 ; Zhang et coll., 2009). Une autre chose remarquable dans ces études sur la fouille d'opinions appliquée dans les SRAP est que le débat à propos des différences entre les genres journalistiques n'est jamais présenté. Nous pensons qu'il est important de connaître de quelle manière l'opinion est exprimée dans les différents genres journalistiques, afin de pouvoir proposer des techniques pour les analyser.

Nous avons aussi vu qu'un positionnement épistémologique est important pour envisager le concept d'opinion et son expression écrite. Les techniques actuelles en fouille d'opinions, basées sur l'analyse du vocabulaire subjectif des textes, sont plus adaptées pour certains types d'opinions, comme les critiques, mais présentent des résultats moins satisfaisants pour l'analyse de textes argumentatifs. Ce problème est l'effet d'une conceptualisation limitée de l'opinion et de la façon d'envisager son expression linguistique : l'opinion est définie comme une évaluation d'un sujet à propos d'un objet comportant une appréciation positive ou négative.

En fonction de ces constats, nous avons examiné la notion d'opinion telle que la conçoivent les recherches en fouille d'opinions. Elle s'apparente à la notion d'opinion préconisée par le domaine du sondage de l'opinion. Nous avons démontré que la fouille d'opinions peut être présentée comme un prolongement de ce domaine de recherche, dans la mesure où elle propose de concevoir l'opinion d'un point de vue quantitatif. En ce qui concerne la démarche méthodologique de fouille d'opinions, définir l'opinion comme l'appréciation ou le jugement d'un sujet sur un objet a la conséquence de considérer le texte qui véhicule une opinion comme une somme de phrases subjectives, ce qui laisse échapper son irréductible complexité. Cependant, il existe d'autres procédés qui ne sont pas nécessairement repérés dans les structures phrastiques, mais qui servent de support au lecteur pour comprendre l'opinion véhiculée par un texte. La façon dont les émetteurs encadrent les questions, le vocabulaire utilisé, ou même de références culturelles et sociohistoriques qui sont évoquées peuvent constituer des indices du type d'opinion défendue.

La nécessité de développer un regard plus global sur l'unité texte est importante dans notre recherche, pourvu que l'objectif poursuivi soit de détecter la divergence d'opinions entre deux articles portant sur le même enjeu thématique. Comme vu dans cette section, les travaux de fouille d'opinions tendent à supposer que l'expression d'opinion est réduite à un

vocabulaire subjectif et à faire de la phrase l'unité minimale d'inférence de cette expression. Les recherches de fouille d'opinions ne remettent pas en question ces choix épistémologiques et laissent inexplorées les études linguistiques qui ont un regard plus global sur le texte comme unité de sens, ainsi que les conditions de production qui déterminent en grande partie les choix linguistiques des émetteurs. Dans notre recherche, nous voulons explorer cette voie théorique.

Au prochain point, nous allons analyser de plus près la notion d'opinion dans le contexte du journalisme, afin de comprendre plus précisément de quelle façon cette notion peut être abordée, ainsi que son aspect concrètement textuel.

4. L'opinion : expression individuelle et engagement social

> *Il faut une grande maturité pour comprendre que l'opinion que nous défendons n'est que notre hypothèse préférée, nécessairement imparfaite, probablement transitoire, que seuls les très bornés peuvent faire passer pour une certitude ou une vérité.*
>
> (Milan Kundera dans *Une rencontre*)

4.1 Introduction

L'opinion en tant qu'expression d'un sujet à l'autre est souvent envisagée dans sa réalité concrète et plus tangible, comme un énoncé qui exprime un point de vue subjectif à propos d'un objet. Cette dimension linguistique, ancrée sur l'énoncé, constitue la base pour le développement d'instruments statistiques destinés à l'analyser et la mesurer. Les travaux en fouille d'opinions font constamment allusion à l'existence d'un « langage subjectif » qui constituerait la configuration linguistique ou textuelle de l'opinion : un ensemble d'expressions émotionnelles et subjectives que les individus utilisent pour argumenter ou pour porter un jugement à propos de quelque chose. Pourtant, l'action d'argumenter, de juger la qualité d'un objet ou d'exprimer un état émotionnel à propos d'une chose correspond à des types d'opinions variés. Comme remarquent Somasundaran et coll. (2007), les questions « Êtes-vous préoccupé par le changement climatique ? » et « Quel sera l'effet de présenter un rapport sur l'Iran au Conseil de sécurité ? » dispose les sujets à fournir des réponses qui se diffèrent par leur nature, la première faisant appel à une réaction plutôt émotionnelle et subjective et la seconde, à la défense argumentative d'un point de vue (Eensoo et coll., 2011).

Cela illustre la dualité intrinsèque à la notion d'opinion. Quel est le type d'opinion que nous devons considérer en tant qu'analystes ? S'agit-il du contenu privé et fondamentalement individuel d'un sujet cognitif interpellé par une question ou plutôt du contenu produit par un sujet dilué dans la conscience d'un groupe social, d'une culture et d'une société ? Une deuxième question peut être également formulée dans une perspective linguistique : parlons-nous de l'opinion comme d'un énoncé contenant certaines caractéristiques syntaxiques ? Ou

plutôt comme d'un acte de langage inséré dans une situation communicative et qui porte les traces de cette situation ?

Dans cette partie de la revue de la littérature, nous entendons approfondir ces questions afin de contextualiser notre objet d'étude : l'opinion publiée dans la presse, qui connaît sa naissance dans l'évolution historique de la presse. Nous voulons également préciser le concept de controverse adopté dans le cadre de cette recherche et la relation de celle-ci avec les caractéristiques du débat dans la presse contemporaine. Les questions qui touchent l'expression écrite de l'opinion seront traitées dans les sections subséquentes.

4.2 L'émergence et l'évolution de l'opinion dans la presse : de l'origine jusqu'aux controverses médiatiques contemporaines

Nous allons exposer dans cette section l'évolution historique de la notion de l'opinion dans le contexte de la presse. Cette démarche nous permettra de contextualiser et de préciser l'utilisation du concept d'opinion dans le cadre de notre recherche, laquelle est associée à la pratique journalistique de diffusion d'opinions dans les médias.

4.2.1 L'émergence de l'opinion dans la presse

Le sens du mot opinion a connu une évolution depuis l'antiquité jusqu'à nos jours. Nous pouvons remonter au mot grec *doxa,* de tradition philosophique platonicienne, qui peut être traduit comme « croyance douteuse ». L'acception de la *doxa* a d'abord été revêtue d'un statut inférieur à celui du savoir et opposée à la notion « d'idée » ou *logos*. L'idée renvoyait à la raison, à l'universel et à la science, alors que la *doxa* faisait référence au subjectif, au mondain et au particulier.

Dans « La République », Platon nuance cette opposition en rapprochant la *doxa* à la sophistique : une opinion peut être bien fondée ou mal fondée, en fonction des critères de vérité qui sont réunis dans son énonciation. Cette idée est développée davantage dans Ménon, où Platon formule qu'une opinion énoncée sur la forme d'un jugement vrai peut se convertir en science (Paveau, 2003). Ainsi, l'opinion peut être une expression reliée au sens commun, aux jugements de valeur, et dépourvue de savoir, mais peut être également un jugement

raisonné, lequel provient de l'observation réfléchie de phénomènes et de l'articulation logique de propositions, vérifiables et démontrables empiriquement.

C'est précisément la notion d'opinion au sens d'un jugement vrai qui a gagné de la force avec l'avènement de la presse écrite au XVIIIe siècle et la formation d'un « public », constitué de l'élite intellectuelle et économique bourgeoise. Habermas (1991) décrit l'émergence de ce public dans les espaces de discussion politique dans l'ère moderne comme une entité clairement séparée de l'État. Désignée de sphère publique, elle se configurait comme une réunion d'individus dans des espaces physiques (comme les salons et les cafés de l'époque) pour débattre des sujets à l'ordre du jour, pour contester ou légitimer le régime politique, bref, pour exercer la critique et délibérer sur les enjeux d'intérêt commun. Habermas explique que l'assimilation du terme opinion par celui d'opinion publique a découlé d'un processus graduel dans lequel un groupe de personnes s'est reconnu intellectuellement compétent et capable de former ses propres jugements, par le biais de la réflexion critique et par l'emploi individuel de la raison. La reconnaissance de cette conscience individuelle est tributaire de diverses transformations historiques, comme la privatisation du religieux et de la propriété, ainsi que de l'émancipation de la société civile en opposition à l'ordre absolutiste. Elle a permis à ce groupe, à ce public de s'organiser collectivement comme une instance médiatrice, servant d'intermédiaire entre la vie privée (la famille, le marché, l'individu) et la vie publique (le gouvernement, le pouvoir établi).

Habermas (1991) explique que l'émergence d'un espace de discussion où les personnes privées font usage de leur raison a retrouvé son expression institutionnelle dans le journalisme et a consacré la notion de « publicité » comme principe fondateur de la délibération rationnelle, essentielle à l'exercice de la démocratie dans les sociétés modernes. Selon ce principe, la diffusion du meilleur argument dans les médias est plus importante que le statut social des interlocuteurs (Jacobs et Townsley, 2011). En relayant les opinions et les arguments de ces individus, les journaux ont contribué à intégrer la notion d'opinion à une nouvelle culture politique, fondée sur la force de l'argumentation. L'opinion publique serait en ce sens la résultante d'un processus de communication et de délibération menée par les individus dotés de la capacité de bien raisonner, et de former une opinion collective capable d'influencer directement les décisions des instances politiques qui les gouvernent.

Une des principales contributions de Habermas (1991) selon Susen (2011) a été la proposition d'un cadre théorique et conceptuel pour analyser historiquement la dichotomie entre le privé et le public et plus généralement, la relation entre l'individu et la société. Habermas a démontré qu'il existe une interdépendance entre le privé et le public, et que l'autonomie de l'individu ne peut être définie qu'en rapport avec la société à laquelle il appartient. La sphère publique serait, dans ce sens, un concept permettant de comprendre comment l'autonomie des individus est mutuellement définie et exprimée socialement dans l'organisation de la société et de ses institutions, ou encore, comment l'expression individuelle est le résultat d'une intense négociation symbolique qui est élaborée et ancrée dans la vie en société.

> Since humans actors cannot escape the various socialization processes imposed upon them by their environment, the purest form of privacy cannot eliminate individuals dependence upon society. Individuals can assert their privacy only in relation to, rather than in isolation from, the existence of other individuals. In this sense, the public sphere is nothing but the socialized expression of individuals' reciprocally constitued autonomy (Susen, 2011, p. 43).

Pour Habermas (1991), le développement du capitalisme mercantile du XVIe siècle et le renversement des régimes absolutistes ont créé les conditions d'émergence d'une sphère publique sans précédent dans l'histoire, puisqu'elle était d'une part potentiellement ouverte à tout individu linguistiquement compétent et d'autre part, parce qu'elle avait la capacité de passer au scrutin le pouvoir politique constitué. Née au sein des régimes démocratiques, la sphère publique bourgeoise avait un potentiel émancipateur, grâce à sa capacité de promouvoir l'engagement civique à travers un processus communicatif de formation d'opinion et de volonté délibérative capable de faire progresser la vie des individus et par conséquent, de la société (Susen, 2011).

La conceptualisation de la sphère publique par Habermas (1991) a eu un impact considérable sur les études politiques et sociologiques. Malgré cette influence, plusieurs auteurs qui ont étudié l'œuvre habermasiènne sont d'accord pour considérer la sphère publique comme un concept idéalisé et de vocation normative (Fraser, 1990 ; Kellner, 2014 ; Susen, 2011). Les principales critiques formulées à l'égard du concept habermasien de la sphère

publique contestent son potentiel émancipateur, puisqu'elle n'incluait pas la participation des minorités, comme les femmes et la classe ouvrière (Fraser, 1990). En outre, si nous pouvons admettre que le pouvoir émancipateur de la sphère publique se fondait sur une discursivité d'opposition à un pouvoir arbitraire, il faut également considérer que les intérêts privés de ses participants, issus de l'élite économique capitaliste qui a le contrôle sur les grands conglomérats médiatiques, ont déprécié peu à peu le type de débat rationnel visant le bien de tous (Susen, 2011).

Kelnner (2014) explique que les critiques formulées à l'égard du concept de la sphère publique oublient que la vraie intention de Habermas (1991) est de souligner les mutations subies par cette dernière, plus précisément sa trajectoire historique, de son émergence jusqu'à son déclin. Dans son livre, Habermas (1991) souligne le processus de « reféodalisation » de la sphère publique dans la fin du XX^{e} siècle et démontre comment la croissance du pouvoir des grandes corporations et leur pénétration dans le monde politique ont provoqué le déclin de la sphère publique, ainsi que la progressive transformation des citoyens participants à des discussions politiques en des consommateurs de contenus fabriqués par les médias (Kellner, 2014). Avec la colonisation des moyens de communication de masse, l'opinion devient administrée, surveillée, gérée et manufacturée par les intérêts privés des compagnies et des annonceurs, ainsi que par les intérêts politiques qui régulent dans la société le marché des médias.

> While in the bourgeois public sphere, public opinion, on Habermas's analysis, was formed by political debate and consensus, in the debased public sphere of welfare state capitalism, public opinion is administered by political, economic, and media elites which manage public opinion as part of systems management and social control. (Kellner, 2014, p.23)

D'autres critiques contestent la forme singulière et monolithique que Habermas (1991) a conférée à la théorisation de la sphère publique dans la société (Fraser, 1990 ; Susen, 2011). Pour ces auteurs, la réflexion de Habermas manque de réalisme face à la composition de la société, puisque concrètement, celle-ci se constitue de plusieurs sphères publiques de discours et d'action. La formation de ces sphères est tributaire de la réaction de groupes minoritaires contre les intérêts économiques et politiques dominants. Fraser (1990) mentionne l'émergence

d'une multiplicité de sphères publiques, crées par les minorités ethniques, sociales et sexuelles, pour fonder des espaces alternatifs de discursivité et d'influence, avec des institutions et des canaux de communication propres. Dans le développement plus récent de sa réflexion, Habermas (1996) s'affilie à cette vision, admettant qu'une des caractéristiques les plus marquantes de la politique moderne est l'intense dispute entre différents groupes sociaux pour faire avancer leurs intérêts et pour influencer les décisions politiques. L'avènement des nouvelles technologies a fortement contribué à l'émergence de ces « arénas de discursivité » (Fraser, 1990) permettant la formulation et circulation d'interprétations différentes sur l'identité, les intérêts et les besoins des différents groupes sociaux qui ont été historiquement marginalisés.

L'étude du journalisme en tant que champ disciplinaire est fortement influencée par la notion de la sphère publique, en ce qui a trait à la réflexion sur le rôle des médias dans les processus de délibération et de formation de l'opinion publique. Le concept de sphère publique supporte normativement la pratique journalistique, en attribuant à cette dernière la mission de surveiller la politique, d'expliciter les questions d'intérêt commun et de permettre aux citoyens de prendre des décisions sur les enjeux qui les concernent. Il est communément admis que les médias peuvent participer à la création d'une démocratie vivante et capable de représenter les aspirations des citoyens, pourvu qu'elles remplissent certaines conditions pour que la délibération ait lieu dans un espace de communication médiée. Ces conditions sont, selon Jacobs et Townsley (2011), l'inclusion, le raisonnement et la publicité. L'inclusion se réfère à la participation des personnes concernées et à l'exposition équitable des différents points de vue. Le raisonnement implique l'articulation et la considération d'arguments divers, pour que la discussion arrive à un terrain commun qui soit profitable à tous. La publicité quant à elle, implique la transparence du processus, la mise en scène du débat et la libre circulation des arguments exposés.

Les médias, en fonction de leur puissance technologique, peuvent facilement remplir l'exigence de publicité, mais pas complètement celle de l'inclusion et du raisonnement, étant donné l'organisation et l'amplitude des sociétés contemporaines (Jacobs et Townsley, 2011). Les différentes réflexions à propos du support offert par les médias aux processus démocratiques ont considéré l'importance de représenter dans l'espace médiatique les

différentes perspectives des acteurs sociaux. Les modèles de communication influencés par la vision libérale de la démocratie par exemple, défendent que les médias supportent en effet la délibération rationnelle dans la mesure où ils offrent un espace ouvert aux différents points de vue de la société, en invitant par exemple les personnes concernées par les problématiques discutées ou en facilitant la mise en scène de positions provenant de sources d'information variées. Des versions plus élitistes de ce libéralisme préconisent que les médias doivent représenter les différentes perspectives avec un journalisme essentiellement professionnel, mené par le travail d'experts et d'intellectuels qui ont la compétence de bien interpréter les événements de la société. Ceux-ci doivent orienter le public quant à la meilleure décision à prendre tout en lui apportant plus de connaissances sur les enjeux complexes de la société et les implications de leurs choix. D'autres courants liés à des études poststructuralistes sont plus enclins à défendre la maximisation de la production de récits identitaires, dans le but de faire circuler la vision de différents groupes de la société, spécialement sur les sujets qui concernent les minorités, dans un espace médiatique diversifié et plus ouvert. Cela constituerait une façon de faciliter l'articulation de leurs opinions et des questions importantes qui les concernent avec une autonomie plus grande que celle offerte par les canaux officiels disponibles (Jacobs et Townsley, 2011).

Jacobs et Townsley (2011) expliquent que le récent développement des technologies de communication, surtout avec l'avènement de l'Internet, a favorisé l'émergence d'un espace de discussion publique plus diversifié et plus ouvert à des acteurs qui ne proviennent pas des médias traditionnels. Ils situent l'émergence de cet espace dans la création des sections d'opinions et commentaires des grands journaux comme le *The New York Times*, dans les années 1970, destinées à pondérer l'influence de sources officielles, en invitant des experts et d'autres représentants de la société à commenter les événements de l'actualité. Les auteurs démontrent comment le commentaire et l'opinion ont gagné en importance dans l'espace médiatique contemporain, suscitant des transformations importantes sur la circulation d'informations et sur la formation de l'opinion publique. L'espace d'opinion, comme les auteurs le désignent, offre la possibilité d'étudier le rapport entre les médias et les processus délibératifs des sociétés contemporaines dans une perspective qui reflète de façon plus réaliste

la configuration de l'espace médiatique actuel, en considérant notamment l'existence et l'influence de plusieurs sphères publiques formées par différents groupes sociaux.

> (…) we have a theory of media and deliberation that places more attention on the spaces of opinion and commentary, as those are the spaces where it is possible to stage a dialogue between official and informal publics – either by allowing representatives of civil society to speak for themselves about the issues of the day or by allowing others (e.g., columnists, intellectuals, and other expert commentators) to challenge official stakeholders on behalf of civil society (Jacobs et Townsley, 2011, p. 67)

Jacobs et Townsley (2011) soutiennent que les réflexions plus récentes de Habermas (1996) sur la possibilité de participation de différents groupes sociaux dans la sphère publique officielle — celle institutionnalisée et composée par les grands conglomérats médiatiques — confèrent plus d'importance à l'espace de l'opinion. Habermas essaie d'élucider les mécanismes de formation d'opinion dans les sociétés contemporaines et de comprendre comment les discussions informelles, menées par les petites associations de citoyens, font leur chemin vers la sphère publique officielle, exerçant une influence sur les questions qui sont portées à la discussion par cette dernière. Habermas constate que malgré la difficulté des publics informels (c'est-à-dire les groupes sociaux organisés) à se faire reconnaître dans une organisation médiatique hautement centralisée, les mobilisations qu'ils organisent auprès de la société civile peuvent éventuellement instaurer un climat d'instabilité politique et avoir un impact sur les discussions véhiculées par les médias. Ce potentiel transformateur des publics informels est selon Habermas un moteur essentiel pour contrer l'influence et la concentration de pouvoir des médias.

Jacobs et Townsley (2011) soulignent que le modèle de Habermas (1996) sur le rapport entre les publics informels et le public officiel manque de démonstrations empiriques capables de rendre compte des dynamiques réelles qui permettent de susciter le débat dans la société. L'investigation de l'espace d'opinion par les auteurs est une tentative de rendre compte de ces phénomènes, en analysant empiriquement ces dynamiques.

Dans le prochain point, nous allons aborder plus en détail les caractéristiques de l'espace de l'opinion conceptualisé par Jacobs et Townsley (2011).

4.2.2 L'espace de l'opinion

Dans l'ouvrage « The Space of Opinion — Media Intellectuals and the Public Sphere », Jacobs et Townsley (2011) font une étude systématique des espaces médiatiques marqués par la diffusion de l'opinion et analysent le rôle du commentaire spécialisé dans les sociétés civiles contemporaines. Jacobs et Townsley (2011) caractérisent l'émergence d'un espace destiné à contrebalancer la dépendance des journaux aux sources d'information officielles, souvent gouvernementales (Jacobs et Townsley, 2011). Ils le désignent comme « l'espace de l'opinion ». Son essor remonte à la création de tribunes libres et d'opinions dans les années 1970 par le journal *The New York Times*, où des journalistes et d'autres représentants de la société civile étaient invités à commenter les actualités.

Les résultats de la recherche de Jacobs et Townsley (2011) attestent que l'espace de l'opinion dans les médias américains, en plus d'être marqué par la présence de chroniqueurs issus d'institutions journalistiques officielles, est également ouvert à une diversité de voix provenant d'autres institutions : les universités, les groupes de pression, les laboratoires d'idées (*think tanks*), les cabinets d'avocats, les maisons d'édition, les organisations politiques et les instituts de sondages d'opinion. L'émergence de la blogosphère et d'autres cannaux de diffusion en-ligne atteste de cette tendance, car elle a permis l'apparition de nouveaux chroniqueurs par une voie différente des médias traditionnels. D'ailleurs ces nouveaux intervenants jouent un rôle décisif dans la formation d'opinion des individus dans la mesure où, en alimentant le débat public avec leurs opinions, ils finissent par créer un lectorat captif et intellectuellement identifié à leurs idées.

Selon Jacobs et Townsley (2011), l'avènement de l'espace de l'opinion a créé les conditions pour un débat plus indépendant et ouvert. Dans ce nouvel espace de l'opinion, l'idée que les discussions publiques reposent sur le principe de délibération rationnelle dont parlait Habermas (1991) est toujours présente, mais les auteurs soulèvent l'importance de bien comprendre la diversité de styles rhétoriques qui sont mobilisés par les chroniqueurs lorsqu'ils élaborent des stratégies argumentatives pour capter l'attention de leur lectorat.

Jacobs et Townsley (2011) ont étudié les aspects sociologiques, culturels, institutionnels et historiques de l'espace de l'opinion ainsi que les caractéristiques des

chroniqueurs : comment écrivent-ils, quels types d'autorités et d'expertise ces derniers mobilisent-ils dans leurs commentaires et quels types de délibération suscitent-ils dans la société ? Pour entreprendre ce travail de description, les chercheurs ont sélectionné un échantillon d'articles d'opinion des journaux *The New York Times* et *USA Today* et des transcriptions des émissions *News Hour*, *Face the Nation*, *Crossfire* et *Hannity & Colmes*, dans les années 1993 à 1994 puis de 2001 à 2002. Les journaux sélectionnés sont parmi les plus lus aux États-Unis et les émissions représentent les principales infovariétés où se déroule le débat politique dans le pays.

Jacobs et Townsley (2011) analysent en particulier les styles rhétoriques employés par les chroniqueurs dans l'espace de l'opinion des médias américains. Le but de cette analyse selon eux est de savoir comment les chroniqueurs élaborent leurs arguments et défendent leurs propos publiquement. Ils décrivent l'existence d'une dimension esthétique et de dramatisation des conflits sociaux dans laquelle les chroniqueurs présentent les protagonistes et les antagonistes dans des récits impliquant de l'intrigue et de l'honneur, aussi bien que des récits qui évoquent les valeurs morales essentielles, telles que l'opposition entre le bien et le mal ou entre l'admirable et l'exécrable. À travers l'examen des articles et transcriptions des sources mentionnées, les chercheurs ont détecté et classé 5 catégories de styles d'opinion qui sont prédominantes : l'argumentation typique, l'argumentation morale, la formulation de questions, le recadrage du débat et la présentation d'informations.

L'« argumentation typique » et l'« argumentation morale » sont les styles les plus employés par les chroniqueurs (Jacobs et Townsley, 2011). Ils sont souvent utilisés dans les débats sur une problématique particulière, présentée dans une structure antagoniste du type pour-contre. Dans l'argumentation typique, les chroniqueurs fournissent un ensemble d'informations de manière stratégique pour contextualiser la problématique et pour supporter un côté du débat — le pour ou le contre. Dans l'argumentation morale, ce sont les valeurs morales qui prennent place dans l'argument, car l'enjeu est débattu sous l'angle de ce qui est « bon » et ce qui est « mauvais » selon les valeurs défendues au sein de la société. L'analyse de données montre que l'argumentation morale est un peu plus marquée à la télévision que dans les journaux américains.

Le style « formulation de questions » est exclusif à la télévision et se manifeste par la présence d'interlocuteurs qui posent des questions pour cerner le sujet et orienter le débat. Le style recadrage est similaire à celui de la formulation de questions, cependant, mais il n'est pas exclusif aux émissions télévisées et est plutôt associé aux opinions des chroniqueurs de la communauté scientifique, comme les avocats ou les écrivains, qui tentent de réorienter le débat en fournissant un type d'opinion plus spécialisée, plus marquée par l'exposition de savoirs scientifiques. En ce qui concerne le dernier style, « présentation d'informations », la caractéristique fondamentale est l'approvisionnement de faits et d'évidences dans un débat en cours.

Jacobs et Townsley (2011) ont aussi analysé l'occurrence, dans les discours des chroniqueurs, des références qu'ils font dans les articles à des autorités et à des experts dans des domaines de connaissances diverses. Ces références, aussi connues comme appels d'autorité, visent à légitimer l'argument et renforcer l'opinion défendue par le chroniqueur. Ils ont identifié des appels d'autorité directe, qui relèvent de l'expérience ou de l'expertise d'une personne, ainsi que des appels indirects qui relèvent des découvertes scientifiques, des faits historiques, des références à des textes populaires ou à d'autres genres de produits culturels (télévisions, films, musique, etc.).

L'analyse entreprise par Jacobs et Townsley (2011) permet de constater que le type de débat qui se déroule sur les médias ne peut pas être analysé dans une perspective normative propre à la conception de communication raisonnée que préconisait Habermas (1991). L'échange argumentatif est traditionnellement défini comme la recherche d'un consensus et d'un accord raisonnable, mais la réalité démontre que les discussions font plutôt appel à la défense de valeurs. Les acteurs en opposition essayent ainsi de se distinguer par les valeurs morales qu'ils défendent, leur discussion amenant moins à une résolution d'un désaccord qu'à la fermeté d'une vision du monde qu'ils défendent. Il s'agit de valeurs correspondant à des codes culturels comme l'identité, l'idéologie ou la religion ou encore correspondant à des visions sur le bien et le mal, le fiable et le suspect, le sacré et le profane.

La composition fragmentée et partisane de l'espace d'opinion (Jacobs et Townsley, 2011) fait susciter un intérêt croissant sur le caractère des discussions menées, marquées par un antagonisme de plus en plus présent. Les débats de type pour-contre dont parle Jacobs et

Townsley (2011) rappellent la notion de controverse utilisée dans certains programmes de recherche dans les sciences sociales, comme dans les *sciences studies* de Bruno Latour (Latour, 2006) et plus récemment par les chercheurs en analyse du discours, comme Charaudeau (2017). La controverse, telle qu'elle est définie par ces derniers, réunit des particularités propres aux débats qui sont menés dans l'espace d'opinion. Elle est également antagonique, se configure comme une dissension entre des groupes opposants qui discutent d'une question donnée et se déploie sur des arènes publiques plus ou moins institutionnalisées, comme les assemblées parlementaires, les médias (journaux, réseaux sociaux, télévision) et certaines organisations (associations, groupements militants et associatifs) (Charaudeau, 2017). Les débats acharnés sur le changement climatique ou sur les organismes génétiquement modifiés (OGM) constituent des exemples de controverses dont l'ampleur dépasse les limites des communautés scientifiques dans lesquelles ils ont éclos et ils mobilisent une pluralité d'organisations, d'experts, de groupes de pression et d'individus.

> C'est par la controverse que commence la possibilité d'une participation lucide et responsable à la vie sociale et politique, c'est par la controverse que peut se développer et s'entretenir une culture du dissensus, fondement même du dialogue social (Charaudeau, 2017, p. 107).

Dans la prochaine section, nous voulons présenter une revue de la littérature au sujet des controverses et les différentes acceptions de ce mot, afin de préciser davantage son rapport à l'espace d'opinion et son utilisation dans le cadre de notre recherche.

4.2.3 La controverse : un champ de recherche de l'espace l'opinion contemporaine

L'étude de la controverse en tant que phénomène social s'insère dans une tradition sociologique initiée par les *science studies*, un champ multidisciplinaire apparu dans les années 1970 qui a insisté sur l'analyse historique des sciences d'un point de vue social et politique. Les *sciences studies* prennent comme objet d'étude les processus de dispute et essayent de comprendre comment les actions collectives engendrent les transformations sur les rapports de forces présents dans les sociétés (Lemieux, 2007). Ce champ de recherche préconise l'étude de controverses scientifiques, spécialement celles qui dépassent le cercle fermé de la discussion savante, comme moyen de comprendre les facteurs qui induisent les

transformations sociales. Elle se positionne comme un programme de recherche descriptive et exploratoire qui, contrairement à d'autres traditions d'études en sciences sociales, ne cherche pas à théoriser les structures sociales de manière à voir dans celles-ci les facteurs préexistants qui expliquent les raisons des conflits sociaux (Latour, 2006 ; Lemieux, 2007).

L'étude des controverses a été développée avec la théorie de l'Acteur-Réseau (Latour, 2006). Cette théorie défend une approche sociologique permettant d'envisager les faits sociaux comme résultants d'une multiplicité de relations, impliquant des acteurs humains et non humains, ainsi que des discours. Dans les années 1980, Bruno Latour développe la discipline de la cartographie des controverses, une version didactique de la théorie Acteur-Réseau. Il présente la cartographie des controverses comme un exercice pédagogique qui vise à créer des dispositifs informatiques permettant d'observer, de décrire et de représenter visuellement les controverses sociotechniques (Venturini, 2009). Celles-ci sont décrites comme des discussions publiques sur des enjeux qui dépassent la sphère restreinte de l'expertise scientifique et rejoignent d'autres domaines du monde social comme l'économie, la politique et l'éthique. Dans la perspective de la cartographie des controverses, les controverses ne se restreignent pas aux questions scientifiques, elles peuvent englober toutes sortes de débats dans lesquels les individus discutent de questions liées à la société, au pouvoir, aux lois et à la politique (Venturini, 2009, 2012). Venturini (2009) illustre ce caractère en utilisant l'exemple des discussions sur le réchauffement climatique :

> Consider, for instance, the controversy on global warming. It all started as a specialized dispute among climatologists and in a few decades, it grew to involve a huge number of scientific disciplines, industrial lobbies, international institutions, social movements, ecosystems, natural species, biological networks, geophysical and atmospheric phenomena. A few years ago, no one would have seen the connection between cars and glaciers. Today we know that they may be opposed on the climatic chessboard, as well as air conditioning and polar bears, sea levels and economical growth, airplanes and crops. (Venturini, 2009, p. 262)

De façon générale, la controverse est assimilée à une situation interlocutive marquée par le conflit, dans laquelle s'engagent au moins deux groupes défendant des positions contraires (Charaudeau, 2017 ; Goodnight, 1991 ; Govier, 1999). Pour Goodnight (1991), la controverse implique une opposition, un échange de perspectives conflictuelles sur un enjeu

d'importance. Pour sa part, Govier (1999) présente trois caractéristiques des controverses : 1) les personnes qui prennent position à l'égard des enjeux débattus doivent être en désaccord ; 2) il doit exister au minimum deux points de vue face à chaque enjeu du débat ; et 3) les personnes doivent faire plus qu'exprimer des points de vue divergents, ils doivent argumenter en vue de délibérer sur la question.

Charaudeau (2017) définit la controverse comme une discussion argumentée et l'inscrit dans une typologie d'échanges de parole. Pour cet auteur, les échanges linguistiques entre des interlocuteurs sont départagés en fonction de leur objectif, qui peut être celui de la coopération ou de la confrontation. La controverse correspondrait à un type d'échange de confrontation en fonction du rapport antagonique établi entre les interlocuteurs. Par contre, la controverse serait différente d'autres types d'échanges antagoniques comme la discussion et le débat. Pour Charaudeau, la discussion est une notion générique englobant les échanges de confrontation, mais elle marque une situation interlocutive dans laquelle les partenaires échangent leurs opinions sur des thématiques variables et dans le but d'arriver à une résolution (par exemple, la discussion entre des collègues dans une entreprise pour prendre une décision d'affaires). Pour sa part, le débat serait lié à une mise en scène d'une discussion, à un moment de confrontation entre les interlocuteurs devant un public, et répondrait à l'objectif de persuasion de ce dernier. La controverse pour Charaudeau peut avoir également un caractère public, mais se distinguerait du débat et de la discussion en ce qu'elle porte sur une thématique spécifique. Dans une controverse, les points de vue sont opposés et argumentés autour d'une même question, avec l'espoir de faire accepter, par un public tiers, une « vérité » contre une autre.

Charaudeau (2017) propose également de distinguer les controverses en montrant qu'elles reposent sur des régimes de vérité différents. Il oppose ainsi la controverse scientifique à la controverse sociale. La controverse scientifique selon lui impliquerait une situation d'interlocution entre des experts, et qui met en jeu des connaissances scientifiques. Ce type de controverse pourrait être tranché par une preuve empirique capable de démontrer la vérité d'un argument. À l'autre extrémité se trouveraient les controverses sociales, lesquelles se déroulent dans l'espace d'opinion (Jacobs et Townsley, 2011) : elles sont publiques par définition, et ont lieu dans des espaces ouverts et plus ou moins institutionnalisés (les assemblées, les associations, les médias, etc.). Par contre, les controverses sociales ne sont pas

caractérisées par la quête de vérité comme les controverses scientifiques, puisqu'elles impliquent l'exposition de la subjectivité des personnes en relation d'interlocution. Les discussions sont marquées par des thèmes sociétaux et par l'exposition de valeurs fondamentales qui ont leur origine dans le « regard que le sujet porte sur l'empirie du monde dans laquelle il est plongé, et dont il tire un savoir d'expérience » (Charaudeau, 2017 p. 49). Il s'agit donc de « savoirs d'opinion » qui sont des « jugements de conviction » (p.51) relevant d'une vérité subjective et partagée par les membres d'un groupe social. Les savoirs d'opinion construisent des explications sur le monde et « Contrairement aux savoirs de connaissance (…) ne se soumettent pas à une vérité extérieure, mais sont porteurs d'un point de vue sur le bien-fondé des choses du monde » (p. 50).

Les controverses sociales se manifestent à la suite d'événements ponctuels dans les sociétés. Charaudeau (2017) en fournit quelques exemples : le mariage homosexuel, à la suite d'une loi parlementaire ; l'immigration, à la suite d'une guerre qui provoque l'arrivée d'immigrants dans une région ; le réchauffement climatique, à la suite de l'observation des changements de température, etc. Charaudeau rejoint la vision de controverses des *sciences studies* (Latour, 2006) en soulignant qu'elles peuvent être abordées et discutées en fonction de différents angles et faisant appel à une multiplicité de connaissances techniques, scientifiques, éthiques ou identitaires. Ceux qui participent à une controverse parlent au nom du groupe qu'ils représentent (religieux, politique, syndical, associatif, citoyen) et leur positionnement se fait au nom de la « *défense de valeurs* afin de faire entendre un point de vue qui se voudrait témoin d'un engagement moral » (Charaudeau, 2017, p.55). Par exemple, la question sur l'avortement mobilise d'une part les féministes, qui défendent la liberté associée à leur corps, et d'autre part les religieux, qui défendent la sacralité de la vie.

Charaudeau (2017) souligne également la difficulté de trancher la controverse par le consensus et explique que celle-ci demeure le plus souvent indécidable pour les intervenants. Il compète au public, qui se place comme juge devant la controverse, de s'identifier aux valeurs que les interlocuteurs en antagonisme mettent en relief.

> Partant d'un thème qui concerne l'ensemble de la société, et qui d'une façon ou d'une autre heurte la conscience morale ou le sentiment d'injustice, voire provoque

> de l'indignation empreinte d'émotion, elle provoque l'intervention des représentants de divers secteurs sociaux voulant défendre une certaine position. De ce fait, la controverse sociale ne permet pas l'établissement d'un cadre commun d'évaluation à propos duquel il serait possible de trancher par consensus. Le lieu discursif n'est pas de *décidabilité,* mais de proposition d'une opinion. (Charaudeau, 2017, p. 51)

Lemieux (2007) offre une perspective englobante sur la notion de controverse, qui tout en étant compatible avec les définitions que nous avons exposées ici (Charaudeau, 2017 ; Goodnight, 1991 ; Govier, 1999), permet de rendre compte des critères de distinction entre les types d'échanges conflictuels relevés par Charaudeau (2017) ainsi que les types de controverses que cet auteur distingue. Lemieux (2007) soutient que la controverse n'a pas une forme pure et propose une définition analytique, dans laquelle celle-ci est assimilée à un conflit présentant une structure triadique. Ainsi, dans la controverse, « un différend entre deux parties est mis en scène devant un public, tiers placé dès lors en position de juge » (Lemieux, 2007, p. 195). Pour Lemieux, ce public peut se constituer seulement d'experts, c'est-à-dire de personnes ayant des compétences distinctives pour juger l'enjeu, ou peut encore dépasser le cercle privé et atteindre le public de profanes. En ce sens, Lemieux propose de voir les controverses à l'intérieur d'un axe borné par deux figures opposées : d'une part, la controverse est discutée par un groupe privé et d'autre part, la controverse est discutée par le grand public, hors du contrôle de la communauté d'experts. Selon Lemieux (2007) cette définition permet d'améliorer la comparabilité des nombreux cas que nous appelons controverse et voir à quel degré le phénomène à l'étude est effectivement triadique (à quel point la controverse implique l'opposition entre deux parties et un public) et à quel degré le public que la controverse réunit se compose de pairs plutôt que de profanes.

4.3 Conclusion

Dans cet exposé, nous avons caractérisé le type d'opinion qui fait l'objet de notre recherche. L'opinion se rapproche de la conception de Habermas (1991) de l'opinion publique qui a pris naissance avec l'essor de la presse écrite, c'est-à-dire un dispositif d'engagement politique émanant de personnes privées organisées comme un public et utilisé pour contester l'état des forces. Ce dispositif implique également une activité de « publicisation » et se situe dans un cadre institutionnel précis, qui est la presse. Dans ce sens, nous pouvons définir

l'opinion comme l'expression linguistique d'un acte communicationnel intégré à une pratique institutionnelle (débats et discussions dans la presse) et qui se pose comme une action critique envers le pouvoir. Cette expression peut être individuelle, mais l'autonomie de cet individu est seulement définie en fonction du milieu social et en fonction des pratiques dans lesquelles cet individu agit.

Il est important de souligner que le caractère actuel du débat contemporain dans la presse, dans lequel les opinions circulent, est différent de celui qui a été décrit et même idéalisé par Habermas (1991) lorsqu'il a proposé le concept de la sphère publique. Deux aspects importants doivent être retenus. Premièrement, il faut prendre en compte l'idée reçue que les personnes qui débattent dans l'espace public le font de façon raisonnée dans le but d'atteindre un consensus. Cette hypothèse contredit la réalité des débats publics contemporains. Nous considérons que Jacobs et Townsley (2011) soulèvent un point important lorsqu'ils caractérisent l'espace de l'opinion comme un espace de dramatisation de conflits et de disputes et dans lequel les valeurs sont discutées et négociées, mais surtout renforcées par les groupes qui discutent et qui s'opposent.

En second lieu, nous sommes d'accord qu'il faut considérer la presse d'opinion comme un espace plus indépendant et ouvert que celui institutionnalisé par le journalisme et qui a été au fil du temps recolonisée par l'intérêt économique des grandes corporations comme soutien Habermas (1991). Le développement des technologies et l'émergence des réseaux sociaux contribuent à la diversité et une certaine indépendance de cet espace, dans la mesure où il s'ouvre à l'émergence de nouveaux chroniqueurs, commentateurs ou intellectuels par une voie non traditionnelle, c'est-à-dire moins centralisée et relativement indépendante des grands conglomérats médiatiques. La diversification de l'espace de l'opinion permet à un nombre grandissant de groupes organisés de la société civile de prescrire les enjeux à débattre dans les médias institutionnalisés, par la voie de la mobilisation.

Un des phénomènes particuliers dans lesquels l'expression d'opinion occupe une place d'importance est celui des controverses. Plusieurs enjeux discutés aujourd'hui sont marqués par un fort antagonisme entre les acteurs impliqués. Dans le cadre de cette recherche, nous reprenons la définition analytique que propose Lemieux (2007) pour caractériser la controverse : « un différend entre deux parties est mis en scène devant un public, tiers placé

dès lors en position de juge » (p. 195). Cette définition nous permettra d'orienter le choix de la controverse pour notre expérimentation et de la situer par rapport au public qu'elle mobilise. Les controverses sont des situations interlocutives qui alimentent le débat public et constituent un dispositif important pour la démocratie, pour la diffusion d'opinions par le biais de laquelle les positionnements de la société circulent (Charaudeau, 2017).

Puisque la controverse est impliquée dans un enjeu de pouvoir, elle peut être caractérisée par l'exacerbation de positions qui ne cherchent pas nécessairement à établir un consensus, mais à réaffirmer des valeurs, communément associées à des visions de monde et partagées par des groupes organisés dans la société civile.

En définissant l'opinion publiée dans la presse comme objet de notre recherche, nous tenons à considérer les articles d'opinion dans les journaux comme notre matériel d'analyse. D'un point de vue empirique, ces textes ont une fonction prioritairement argumentative et présentent une certaine variété de styles rhétoriques et de formes de composition, comme les études menées par Jacobs et Townsley (2011) nous l'ont montré. Ils sont également rattachés à certains genres textuels, que nous dénommons génériquement comme genre de l'opinion et qui seront détaillés dans les prochaines sections.

Nous voulons plus spécifiquement connaître les genres qui sont utilisés dans l'espace d'opinion et aussi examiner la question suivante : qu'est-ce qui permet de dire qu'un texte véhiculant une opinion est argumentatif ou qu'il s'insère dans un enjeu argumentatif ?

5. L'opinion : expression linguistique et genres textuels

5.1 Introduction

Dans la section précédente, nous avons défini l'opinion comme l'expression linguistique d'un acte communicationnel qui se produit dans le contexte de la pratique journalistique. Cela nous a permis de cibler l'objet de notre recherche : il s'agit de productions textuelles élaborées par des sujets dans une situation communicationnelle spécifique et utilisées comme des instruments d'engagement politique et social qui sont fondamentaux à l'exercice de la démocratie. Ces productions sont associées à des genres textuels, communément rencontrés dans l'espace d'opinion (Jacobs et Townsley, 2011) : les commentaires, les éditoriaux, les chroniques et les lettres de lecteurs sont parmi les genres caractéristiques de cet espace.

Comme vu aussi dans la section précédente, les débats qui se produisent dans l'espace d'opinion présentent une variété de styles rhétoriques. Jacobs et Townsley (2011) soulignent le caractère dramatisant des conflits dans l'espace d'opinion, qui prennent la forme de récits opposant les protagonistes de la société en fonction de valeurs morales profondément ancrées sur la culture ou sur les notions de bien et de mal. La quête de rationalité, au sens de la démonstration par des preuves de la justesse d'un argument, existe dans le débat. Mais il ne faut pas négliger la variété de recours que les chroniqueurs utilisent pour chercher l'adhésion de son auditoire.

Nous avons également défini et caractérisé une situation typique des débats publics, qu'est la controverse. La dispute dans laquelle s'opposent les groupes dans ce genre de débat peut mener à l'exacerbation des positions défendues, jusqu'à même l'agression verbale. Le type particulier d'échanges dans les controverses a été désigné par quelques auteurs de discours polémique (Amossy, 2014 ; Charaudeau, 2017 ; Garand, 1998 ; Kerbrat-Orecchioni, 1980). Ici, le mot « discours » réfère à la fois à un type de situation de communication impliquant des émetteurs et de récepteurs et à un type de configuration linguistique qui reflète cette situation dans les productions textuelles. Le discours polémique est propre aux situations

opposant des locuteurs qui se trouvent en désaccord et dans lesquelles il est question d'attaque explicite des positions défendues par un « autre », ciblé comme adversaire.

Cette section s'intéresse aux aspects linguistiques de l'expression de l'opinion, tout en considérant le cadre communicationnel dans lequel nous avons inscrit la notion d'opinion abordée dans cette recherche. Notre objectif ici est de caractériser l'opinion journalistique sur deux axes : d'une part, sur le plan des genres textuels qui sont propres à la pratique journalistique de diffusion d'opinions et d'autre part sur les caractéristiques de ces genres, en fonction de la situation de communication dans laquelle ils sont produits.

Il convient d'abord de situer historiquement l'apparition des genres journalistiques et d'expliquer leurs différences, afin de comprendre les caractéristiques distinctives des genres dans l'espace d'opinion. Par la suite, les études sur l'argumentation visant à approfondir la caractérisation de ces genres seront exposées, en particulier l'étude de l'argumentation dans le cadre communicationnel qui est le nôtre. En plus, puisque les débats qui se déroulent dans la presse ne sont pas nécessairement liés à la quête d'un accord raisonné comme le prétend la vision normative du journalisme, nous allons explorer une littérature portant sur le discours polémique. Il importe de connaître spécifiquement les types de stratégies mises en œuvre dans les textes et qui se produisent dans des situations antagoniques comme les controverses.

5.2 Évolution historique des genres journalistiques

Cette section a pour objectif de présenter des études qui expliquent l'évolution des genres journalistiques et de contextualiser l'apparition du genre de l'opinion. Avant de commencer cet exposé, nous voulons introduire brièvement la notion de genre adoptée dans le cadre de notre recherche. Celle-ci est abordée dans la section suivante.

5.2.1 Qu'est-ce qu'un genre ?

Dans les espaces de commercialisation de biens culturels, le genre est le principe organisateur des ressources : dans une librairie par exemple, les ouvrages littéraires sont généralement organisés dans des sections comme « biographie », « poésie » ou « théâtre ». En fonction de ce principe organisateur, le genre est conventionnellement défini comme une

catégorie d'œuvres. L'ensemble de ces catégories permettent non seulement d'organiser, mais aussi de hiérarchiser la production culturelle et intellectuelle disponible.

Selon Chandler (1997), la plupart des études sur le genre ont tendance à associer ce dernier à des types de textes. Dans cette perspective, les textes appartenant à un genre doivent détenir certains traits communs, autant au niveau de la forme qu'au niveau du contenu. Un genre est alors défini comme un système de conventions structurelles, stylistiques et thématiques. Cependant, Chandler explique que les conventions établies à un moment donné pour caractériser les productions textuelles existantes ne suffisent pas à déterminer un genre. D'une part parce que la réalité démontre que les textes qui sont produits par une culture exhibent plus de caractéristiques que celles normalisées par un genre à un moment donné dans l'histoire. D'autre part parce que les caractéristiques considérées particulières à un genre ne sont pas uniques à ce dernier. Cette diversité, selon l'auteur, rend difficile l'assimilation du genre à un type de texte. Mais elle est intrinsèque à la pratique même de production d'œuvres culturelles, où de nouveaux genres émergent à l'intérieur d'un processus de différenciation interne et de combinaison de genres existants. Par exemple, une œuvre qui s'inscrit dans le genre roman peut présenter de nouveaux éléments stylistiques provoquant un écart entre l'œuvre en question et les caractéristiques conventionnées par ce genre particulier.

Rastier (2001b) souligne les mêmes problèmes abordés par Chandler (1997) en distinguant deux conceptions du genre : 1) la conception classificatoire dans laquelle le genre est déterminé par un ensemble de critères linguistiques visant à hiérarchiser les productions textuelles, et 2) la conception typologique, où le genre constitue un modèle hypothétique servant à caractériser les occurrences du type. Selon Rastier (2001b), la première conception doit faire face aux problèmes généraux de toutes les taxinomies, qui consistent à déterminer la variabilité des critères qui déterminent les classes et les sous-classes à l'intérieur d'une organisation logique. La seconde conception doit régler le rapport entre le type et les occurrences, car le type n'a de sens que lorsqu'il permet d'identifier une occurrence. Or, comme nous l'avons mentionné, la valeur stylistique d'une œuvre culturelle se réalise par une certaine divergence par rapport à ce qui est conventionné par un type, et il est difficile de se prononcer sur la stabilité d'un type qui pourrait servir de modèle (Rastier, 2001a, 2001b).

L'idée que le genre n'est pas un type de texte est imprégnée dans les travaux sur l'étude des genres en sciences de l'information. Andersen (2008) par exemple explique que pour les sciences de l'information, le genre s'inscrit dans plusieurs activités humaines de production et d'utilisation de textes et contribue à caractériser ces activités d'un point de vue de production et d'accès à la connaissance. Cette tradition d'étude socio-humaniste du genre provient de la *North American School* et corrèle le genre à une pratique rhétorique. Dans cette conception, étudier le genre signifie comprendre les manières dont les textes deviennent des actes typifiés à l'intérieur du contexte social, et liés à un but communicatif propre, plutôt que savoir les qualités formelles associées aux textes. Dans cette vision, les textes sont produits et utilisés comme des instruments stratégiques de communication qui façonnent et affectent les modes par lesquels les individus interagissent.

Inspirés par l'*Activity Theory,* certains auteurs soulignent le pouvoir du genre à instaurer une situation communicative particulière dans laquelle le texte est utilisé pour accomplir un objectif ou une tâche (Russell, 1997 ; Winsor, 1999). Par exemple, Bakhtin (1986) affirme que les genres sont des types relativement stables d'énonciations (*utterances*) qui se développent dans le monde social pour coordonner l'activité humaine : c'est le moyen formel par lequel les individus s'impliquent dans une situation discursive, pour réaffirmer, réfuter ou complémenter les énoncés d'autres individus.

Plusieurs auteurs ont mis en évidence la capacité du genre textuel à établir un contrat interprétatif permettant la communication entre les producteurs et les récepteurs. Cette approche écarte la définition du genre comme type textuel (Chandler, 1997). Par exemple, Wilson et Robinson (1990) définissent les genres comme formes de littératures qui créent une attente de la part des récepteurs et qui guident également le travail des producteurs. Le genre se perçoit par certaines régularités linguistiques, mais il n'est pas un objet abstrait de la réalité pratique dans lequel il s'inscrit. La vision socio-humaniste du genre que nous avons exposée prend acte d'éléments typiques d'une situation de communication : l'émetteur, le public visé et le type de message qui détermine la manière dont les textes sont produits et interprétés.

Par le terme « genre de l'opinion », nous voulons pour l'instant exprimer les modèles d'écriture qui sont utilisés par les journalistes, chroniqueurs et autres commentateurs lorsqu'ils tentent de donner explicitement leur opinion à propos d'un sujet d'intérêt public dans un

espace médiatique qui comprend la présence d'un public et d'un lectorat. Ils sont insérés dans un type de pratique professionnelle et créent des attentes par rapport à l'instance de réception. Cela inclut notamment les éditoriaux, les lettres ouvertes, les commentaires, les articles de tribune libre et bien d'autres. Une définition en rapport avec notre méthodologie, mais compatible avec les concepts présentés ici, sera présentée en détail dans le chapitre 2 (p.122).

5.2.2 Développement historique des genres journalistiques

L'émergence de la presse s'est accompagnée de l'apparition progressive de genres textuels. C'est dans ce contexte aussi que nous pouvons retracer l'apparition des premiers genres journalistiques, dont le genre de l'opinion. Grosse (2001) situe l'origine du genre de l'opinion au début du XIX^{e} siècle, avec l'apparition des premiers commentaires et lettres des lecteurs. L'apparition des débats d'opinions dans la presse européenne coïncide historiquement avec la naissance de l'État national italien, et ces genres reflètent l'ambiance combative et patriotique de l'époque. Selon Grosse, ce même modèle d'évolution est observé dans d'autres États, notamment en Allemagne et en France.

Grosse (2001) propose un modèle de classement pour les genres journalistiques en fonction de leur évolution historique (figure 3). Les genres sont également regroupés en fonction de catégories textuelles qui correspondent à l'intention dominante et manifeste (ou intention affichée) des textes : Information, Opinion[14], Conseil, Fiction, Divertissement, Publicité et Hyperstructures. Une lecture du haut vers le bas du modèle permet de voir la séquence temporelle et l'évolution historique des genres. Les « champs transitoires » indiquent une étape de transition des genres liés à une catégorie vers une autre catégorie textuelle postérieure. Ils regroupent aussi des genres « transitoires », qui combinent les caractéristiques de genres précédentes puis postérieures. Dans ce sens, les genres qui se situent dans un regroupement lié à une catégorie textuelle sont considérés comme plus stables quant à leur intention dominante.

[14] Dans l'article publié dans la revue *Semen* en 2001, le modèle de Grosse affiche le nom « Information » au lieu de « Opinion ». Par contre, il mentionne le trosième groupe comme Opinion dans le texte. Nous avons corrigé la figure 3 en fonction.

Nous pouvons ainsi voir dans les trois premiers regroupements du modèle de Grosse (2001) la transformation d'une presse éminemment informative, dont le travail se résumait à rapporter les faits courants et les affaires du souverain, vers une presse plus engagée et plus critique. Grosse explique que les premiers genres informatifs sont issus des brèves, des petits textes destinés à rapporter les nouvelles de la cour, et sont apparus au début du XVII^{e} siècle. Le genre de l'opinion (groupe « Opinion ») est apparu deux siècles plus tard. Cette époque de transformations sociales a favorisé l'émergence d'une posture journalistique plus combative vis-à-vis du pouvoir.

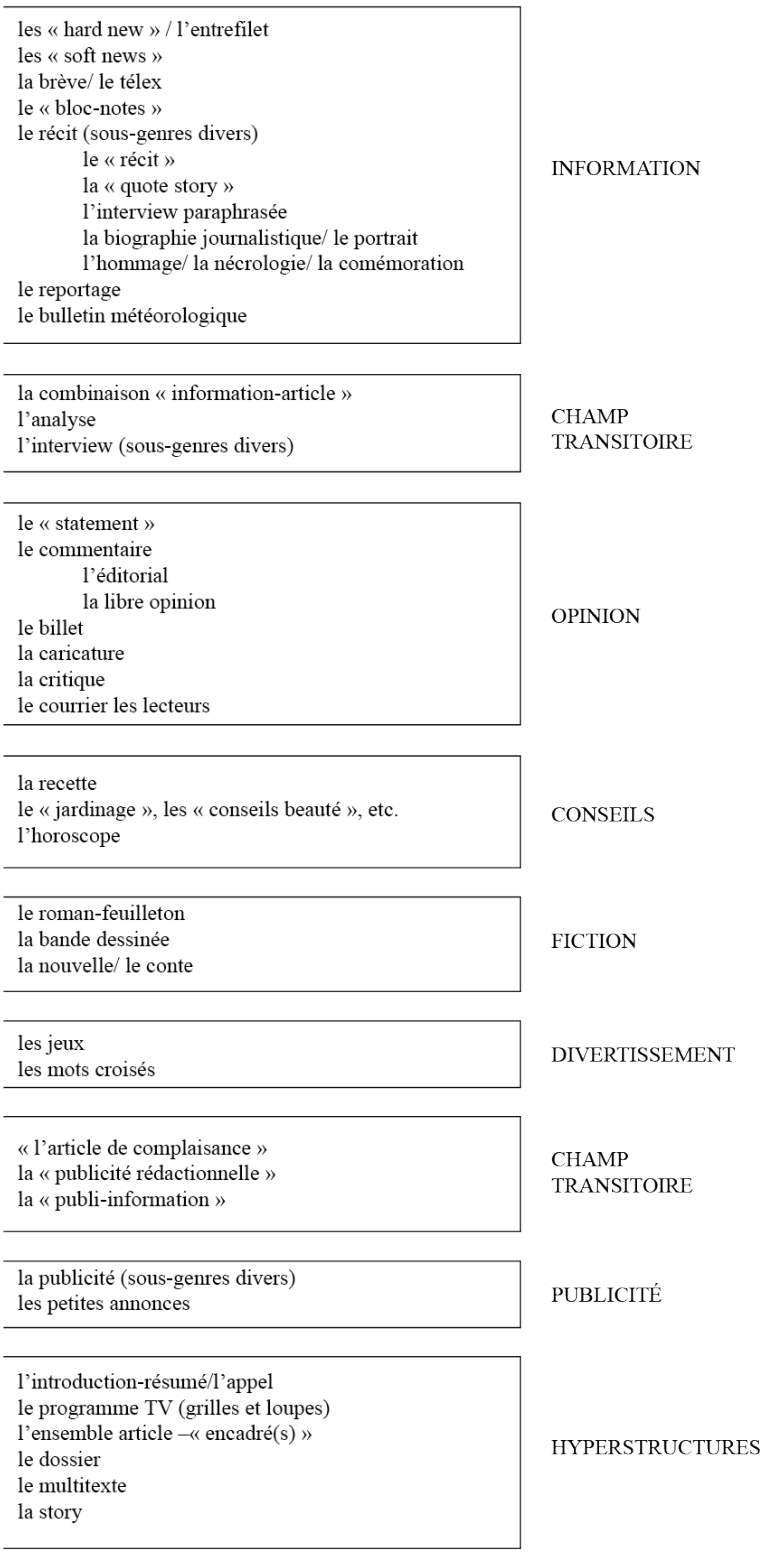


Figure 3. Modèle de l'évolution des genres journalistiques (Grosse, 2001, p.10)

Le premier regroupement « Information » affiche les genres qui ont fait leur apparition en premier lieu dans la presse, dans laquelle le compte rendu de faits est prédominant. Ce sont les nouvelles, les *hard news*, le reportage, entre autres. La caractéristique principale de ces genres est le souci d'objectivité et d'impartialité, ordinairement forgées par la position de distance du journaliste, qui se livre seulement au rapport des faits et non de son opinion personnelle (même s'il peut toutefois demander des avis et des conseils à des tiers, que ce soit des personnes publiques, des experts ou quelqu'un qui a été un témoin oculaire du fait qu'il rapporte). Le champ transitoire entre « Information » et « Opinion » signale le glissement progressif vers les genres de l'opinion, pour lesquels il est accordé une place plus importante à l'analyse, à la critique et à l'avis personnel. L'entretien, qui implique un sujet interviewé qui parle et qui s'exprime à la première personne en réponse à l'intervention du journaliste, est un sous-genre situé dans le champ transitoire et indique le passage progressif d'une catégorie à l'autre. Dans la catégorie « Opinion » se trouvent les genres dans lesquels l'opinion du journaliste ou du chroniqueur devient plus explicite, intégré au but communicatif du texte.

Le modèle affiche par la suite l'ordre d'apparition d'autres genres journalistiques, phénomène influencé d'une part par la demande d'un lectorat de consommateurs désireux d'obtenir des informations sur les biens et services (« Conseils », « Divertissement » et « Publicité »), et d'autre part par la convergence des genres. Dans « Hyperstructures », Grosse (2001) classe les genres créés à partir du regroupement d'autres genres précédents et qui se sont configurés comme des « ensembles rédactionnels ». Le dossier est le cas le plus typique de ce phénomène, se constituant comme un recueil détaillé d'articles journalistiques à propos d'un événement. Un autre exemple, le genre « appels » dans la catégorie « Hyperstructures » fait référence aux petits résumés affichés dans la première page d'un journal ou d'un site Web et dirigeant le lecteur vers une version plus détaillée de chaque article. Selon Grosse, les appels remontent aux anciennes brèves qui rapportaient en résumé les nouvelles de la cour.

Grosse (2001) explique que le commentaire, est un des premiers genres de l'opinion à naître, adoptait le modèle d'écriture fourni par la rhétorique classique enseignée dans les écoles et les universités. La composition des commentaires dans la presse émergente était caractérisée par l'emploi des figures rhétoriques et des techniques d'argumentation issues du

discours oratoire classique (*exordium, narratio, argumentatio, conclusio*). Le commentaire s'est diversifié postérieurement, donnant naissance aux éditoriaux et aux billets.

Une discussion à propos de la rhétorique et de sa relation avec l'argumentation nous permettra d'éclairer la spécificité du genre de l'opinion. Nous avons consulté des articles et des ouvrages qui étudient la rhétorique et l'argumentation dans une perspective linguistique et discursive, afin de vérifier l'apport de ces études pour caractériser notre objet d'étude. La prochaine section présente la recension et l'analyse de ces écrits.

5.3 Rhétorique et argumentation

Dans le domaine des sciences du langage, la rhétorique et l'argumentation sont souvent considérées comme deux domaines d'étude différents (Amossy et Koren, 2009). La rhétorique peut être vue comme une « discipline particulière qui englobe l'argumentation, mais le terme rhétorique peut aussi désigner une branche d'étude qui s'oppose à l'argumentation » (Amossy et Koren, 2009, p. 7). En France, l'étude de l'argumentation a été fortement influencée par les travaux d'Anscombre et Ducrot (1983), qui ont particulièrement insisté sur l'aspect linguistique de l'argumentation. Dans le livre « Argumentation dans la langue », Ascombre et Ducrot distinguent « l'argumentation linguistique », liée à construction du sens logique de l'énoncé, et « l'argumentation rhétorique » qui relève de l'art de raisonner et de convaincre par la parole. Cette opposition entre argumentation et rhétorique, qui se fonde d'une part sur la construction logique de l'énoncé et d'autre part sur l'intentionnalité communicative, est le résultat de plusieurs modifications que la rhétorique classique a subies pendant les vingt siècles suivant sa naissance dans le monde occidental (Amossy, 2013 ; Amossy et Koren, 2009 ; Ducrot et Todorov, 1972).

Afin de clarifier ces différents aspects, nous allons aborder l'évolution de ces deux courants d'études dans les sciences du langage. Nous allons ensuite présenter comment certains auteurs affiliés à l'analyse du discours définissent la rhétorique et l'argumentation dans la perspective discursive et de quelle façon nous pouvons reconnaître la présence de l'argumentation dans le genre de l'opinion.

5.3.1 Études de la rhétorique et de l'argumentation

La rhétorique en tant que discipline remonte à l'Antiquité. Elle a été conceptualisée et formalisée par Aristote, qui l'avait définie comme « la faculté de considérer, pour chaque question, ce qui peut être propre à persuader » (Aristote, 1991, p. 82). Son origine est liée au régime démocratique de la *polis* grecque et à la pratique de la discussion publique. Elle a été souvent définie comme l'art de la parole (Amossy, 2013 ; Amossy et Koren, 2009 ; Ducrot et Todorov, 1972) : faire un discours rhétorique signifie utiliser un ensemble de techniques et de stratégies verbales destinées à influencer et à convaincre un auditoire pour le faire adhérer à une vision ou à une idée. Dans ce sens, la rhétorique avait chez les anciens une visée pragmatique (Ducrot et Todorov, 1972), puisqu'elle impliquait l'intention de susciter une action de la part de l'auditeur concernant la vie collective, en lui proposant un ensemble de preuves destinées à le persuader.

Les preuves sont des démonstrations logiques discernables par des opérations d'induction et de déduction sur les énoncés, mais aussi déductibles en fonction du caractère moral de l'orateur. La rhétorique aristotélicienne a souligné l'importance de l'*éthos* dans la persuasion, une dimension liée au caractère moral de l'orateur et à l'image qu'il imprime dans son discours. Aristote a aussi consacré un livre entier dédié au *pathos* dans la rhétorique, qui est l'effet persuasif du discours sur les émotions de l'auditoire (Amossy et Koren, 2009). Selon Amossy (2013), il ne faut pas sous-estimer l'importance de l'*éthos* et du *pathos* dans la persuasion. Il s'agit de dimensions de la rhétorique qui sont aussi importantes que les démonstrations logiques pour entraîner la conviction. Le pouvoir de convaincre quelqu'un par la parole ne dépend pas seulement de la démonstration, puisque l'accord que le locuteur cherche à établir n'est pas nécessairement lié à une vérité absolue, mais à ce qui peut être raisonnable ou plausible de faire ou de croire, selon les valeurs partagées par la société. La persuasion dépend de la crédibilité, de l'autorité et de la légitimité de celui qui parle, aussi bien que de la force des émotions que ce dernier peut susciter à travers son discours.

La rhétorique classique était fondée sur 5 parties, d'abord de l'ordre de l'organisation du discours et ensuite, de la mise en scène :

> 1) *inventio* : sujets, arguments, lieux, techniques de persuasion et d'amplification ; 2) *dispositio* : arrangement des grandes parties du discours (exorde, narration, discussion, péroraison) ; 3) *elocutio* : choix et disposition des mots dans la phrase, organisation dans le détail ; 4) *pronuntiatio :* énonciation du discours ; 5) *memoria* : mémorisation. (Ducrot et Todorov 1972, p. 99)

Amossy (2013) souligne que la rhétorique avait dès ses débuts un caractère communicationnel, puisqu'en tant que discipline, elle reposait sur l'existence d'un public capable de raisonner et de tomber d'accord. Le *pronuntiatio* et la *memoria* représentent cet aspect, dans la mesure où ils renvoient à une certaine « dramatisation » dont l'orateur devait faire preuve sur la scène publique. Ce caractère communicationnel a été cependant progressivement effacé dans l'étude postérieure de la rhétorique. Au Moyen Âge, la rupture entre l'*inventio* et la *dispositio* d'une part et de l'*elocutio* de l'autre a cantonné la rhétorique à l'étude du style et de l'ornement dans l'écriture, cantonnant aussi les questions entourant le raisonnement à des disciplines comme la logique. Sur cette rupture, Amossy et Koren (2009) expliquent :

> Dans l'histoire de la rhétorique telle qu'on la trouve dans les manuels, il est courant de marquer la rupture qui s'est opérée entre l'art de raisonner et de persuader, d'une part, et l'art de bien-dire, d'autre part (…). À partir du moment ou *l'inventio*, ou recherche de matériaux à utiliser dans le discours, et la *dispositio* ou organisation de ces mêmes matériaux, ont été renversées au compte du raisonnement dialectique, à savoir de la philosophie, l'*elocutio* relative au style est devenue l'essentiel de la rhétorique, réduisant dès lors celle-ci à la question des figures et des tropes » (Amossy et Koren, 2009, paragr. 3).

Devenue un art de l'ornement dans la composition des discours, la rhétorique a été associée plus spécifiquement aux études littéraires et son étude est devenue essentiellement axée sur l'énumération des figures de style, fondant le programme disciplinaire de la stylistique (Ducrot et Todorov, 1972). Ducrot et Todorov distinguent deux volets de cette discipline : un volet de caractère plus normatif qui a fondé les pratiques du bien-écrire, en se basant sur des œuvres de la rhétorique classique et un autre volet qui soutient une conception plus psychologisante, voyant dans les figures employées par un auteur les traces de sa subjectivité.

Dans le monde anglo-saxon, la dialectique, héritière de l'*inventio* et de la *dispositio,* s'est développée sous l'égide de la linguistique formelle. Celle-ci a restreint la rhétorique à l'étude de la communication efficace, en soulignant les critères et les normes du raisonnement valide, s'intéressant surtout aux discours de la vie ordinaire, comme le discours politique et juridique (Amossy, 2013 ; Amossy et Koren, 2009). La linguistique formelle a étudié plus spécifiquement les procédures argumentatives qui permettent de soutenir les bons arguments et de justifier les propositions. Dans cette lignée normative, les travaux de Woods et Walton (2004 : 1982) se sont concentrés à l'étude des paralogismes (ou arguments fallacieux), des arguments qui ont une apparence de validité, mais qui sont logiquement invalides. Un autre courant centré sur l'argumentation est celui de la pragma-dialectique. Aussi de caractère normatif, cette discipline a fondé un modèle de discussion critique pour permettre l'accord sur l'acceptabilité des positions débattues, en définissant les règles de la bonne discussion et de la construction d'arguments (Amossy, 2013 ; Amossy et Koren, 2009). Affiliés à des aspects strictement linguistiques et logiques de la construction d'arguments, ces études ont évacué toute question relative à l'*éthos* et au *pathos*, ou à la question de l'intentionnalité communicative qui faisait partie de la rhétorique classique (Amossy, 2013).

Du côté français, l'étude de l'argumentation par Anscombre et Ducrot (1983) a fondé une école centrée sur l'étude des connecteurs de l'argumentation. Pour Anscombre et Ducrot (1983), l'argumentation doit être étudiée du point de vue linguistique, et chercher à expliquer la construction de l'énoncé sans avoir recours à la situation dans laquelle la communication a lieu ni aux intentions des locuteurs. Dans la vision de ces auteurs, l'argumentation est un enchaînement d'énoncés menant à une certaine conclusion. Ils nommaient rhétorique une composante à l'intérieur de l'énoncé (ou d'un l'ensemble d'énoncés) qui permettait à ce dernier à s'orienter vers une conclusion :

> Un locuteur fait une argumentation lorsqu'il présente un énoncé E1 (ou un ensemble d'énoncés) comme destiné à en faire admettre une autre (ou ensemble d'autres) E2 (Anscombre et Ducrot, 1983, p.8).

Malgré la diversité d'approches qui ont pour origine de la rupture entre l'*inventio* et la *dispositio* d'un côté et de l'*elocutio* de l'autre, plusieurs auteurs ont essayé de restaurer les

bases de la rhétorique classique, en proposant des voies d'études unissant d'une part les procédés linguistiques liés à la construction logique d'arguments et d'autre part l'aspect de l'*elocutio*, lié aux choix individuels des locuteurs pour persuader. Bally (1951, cité par Ducrot et Tudorov, 1972) a été le premier à proposer la rhétorique comme une discipline descriptive et à soulever les facteurs qui imprègnent le discours des sentiments et des expressions individuelles des locuteurs. En introduisant la notion d'énonciation, il a fait remarquer que tout énoncé linguistique contient des caractéristiques particulières qui ne sont pas uniquement explicables par le système de la langue, puisqu'ils relèvent d'un choix judicieux opéré par l'énonciateur sur les moyens de s'exprimer, du lexique jusqu'aux catégories grammaticales employées. Cette perspective a ouvert la voie aux études discursives de la rhétorique et de l'argumentation, que nous allons présenter dans le prochain point.

5.3.2 La rhétorique comme cadre communicatif entourant les échanges à visée persuasive

Dans le champ de la linguistique, l'étude de l'usage de la langue dans des situations impliquant un processus de communication entre deux partenaires a marqué l'émergence de théories du discours dans les années soixante. En dépit de la diversité d'approches et de notions existantes, les recherches en analyse du discours (AD) convergent vers un intérêt prononcé pour l'étude du texte, qui est souvent défini comme une production discursive : il serait le résultat tangible du discours, autrement dit, de la mise en œuvre de la langue lors d'un échange de communication entre deux partenaires.

Trois courants se font remarquer dans l'approche discursive : les théories énonciatives, l'analyse du discours et la sémantique interprétative (Paveau et Safarti, 2003). Ces théories présupposent l'existence d'une dimension linguistique et d'une dimension extralinguistique qui renvoient à la traditionnelle opposition entre langue et parole posée par Ferdinand Saussure (Saussure, 2016). Pour ce dernier, l'objet de la linguistique serait la langue en tant que « système de signes », lequel est employé par une communauté d'individus. La parole, définie comme un acte individuel, serait quant à elle la mise en œuvre de ce système dans une situation communicative. Dans la vision de l'AD, le discours serait la manifestation individuelle de la langue dans une situation de communication. Ainsi, l'étude des productions

langagières, qu'elles soient orales ou textuelles, devrait se rapporter aux conditions spécifiques dans lesquelles elles sont produites.

La situation de communication peut se définir comme un « phénomène d'échange entre deux partenaires (que ceux-ci soient présents l'un à l'autre, ou non) qui doivent se reconnaître semblables et différents » (Charaudeau, 1995, p. 99). Cette définition présuppose une interaction communicationnelle orale ou écrite entre deux pôles, l'émetteur et le récepteur. Ils sont différents parce qu'ils reconnaissent la spécificité de leur rôle à l'intérieur de l'échange et ils sont semblables par le rapport qui rend leur communication possible, c'est-à-dire l'univers de connaissances partagées sur le monde et sur la situation d'interlocution dans laquelle ils se retrouvent. La situation de communication comprend d'une part les intentions des partenaires impliqués dans l'échange et d'autre part, les aspects psychologiques, sociaux, culturels et historiques qui sont à l'origine de la production discursive et qui instaurent différents types de rapports déterminant le registre du langage, la façon dont l'acte langagier est produit et aussi interprété.

Perelman et Olbrechts-Tyteca (1988) ont défendu un remembrement de la rhétorique et de l'argumentation, en intégrant la notion de la situation de communication et l'importance de considérer, pour l'étude de l'argumentation, la présence d'interlocuteurs qui cherchent à obtenir l'adhésion de l'autre à un point de vue. En suivant la tradition aristotélicienne, les auteurs considéraient que l'argumentation ne devrait pas se concentrer sur la recherche de la validité des arguments ou sur leur véracité, mais devrait chercher à comprendre comment un accord raisonnable pourrait se faire dans un cadre communicationnel.

En partant du principe qu'il n'existe pas de vérité absolue, ils ont défendu l'argumentation comme un moyen de chercher l'accord entre les esprits sur ce qui peut paraître raisonnable et acceptable. Dans cette perspective, les preuves n'étaient pas uniquement un résultat de la validité des énoncés, elles faisaient appel à ce qu'Aristote a défini comme *topiques* ou *topoï,* soit les lieux communs, les opinions dominantes qui sont référencés par la culture des interlocuteurs (Amossy, 2013).

La nouvelle rhétorique de Perelman et Olbrechts-Tyteca (1988) s'est définie comme étant « les techniques discursives permettant de provoquer ou d'accroître l'adhésion des esprits

aux thèses qu'on présente à leur assentiment » (p. 6). Cette nouvelle rhétorique ne s'intéressait pas à l'étude de figures séparée de la situation de communication, mais à la compréhension des buts spécifiques que ces figures remplissent dans l'argumentation (Amossy, 2013).

Même si la nouvelle rhétorique de Perelman et Olbrechts-Tyteca (1988) offrait une possibilité d'étude de l'argumentation hors du cadre normatif de la linguistique formelle ou de la pragma-dialectique, elle ne proposait pas pour autant une manière d'envisager les textes argumentatifs d'un point de vue linguistique. Ces auteurs préconisaient l'analyse de l'argumentation en fonction des types de liaison entre les pensées, afin d'expliciter les aspects cognitifs permettant de construire des raisonnements argumentatifs. L'aspect proprement linguistique est expliqué de façon sommaire comme étant lié aux choix linguistiques individuels des locuteurs, dans le but de capter l'attention de l'interlocuteur ou de l'audience. Pour les auteurs, il n'existe pas de choix neutre. Même le langage ordinaire inscrit déjà l'argumentation dans un accord implicite :

> On repère généralement l'intention argumentative par l'indice que présente l'usage d'un terme s'écartant du langage habituel. Il va sans dire que le choix du terme habituel peut également avoir valeur d'argument ; d'autre part, il y aurait lieu de préciser où et quand l'usage d'un terme déterminé peut être considéré comme habituel ; *grosso modo,* nous pourrons considérer comme habituel le terme qui passe inaperçu. Il n'existe pas de choix neutre — mais il y a un choix qui paraît neutre et c'est à partir de celui-là que peuvent s'étudier les modifications argumentatives (Perelman et Olbrechts-Tyteca, 1988, p. 201).

Amossy (2013) adhère à la perspective de la nouvelle rhétorique, soulignant que l'importance de l'argumentation réside moins dans les procédures de validité logique, que dans le projet rhétorique dans son ensemble, en l'occurrence, le cadre d'échange verbal opéré dans une dimension sociale, institutionnelle ou culturelle. Ainsi, elle défend que les théories d'argumentation doivent être subordonnées à la rhétorique, qui constitue, comme proposent d'ailleurs Perelman et Olbrechts-Tyteca (1988), le cadre communicatif entourant les échanges. Amossy (2013) préconise une approche d'analyse dans laquelle les procédures de raisonnement, dans leur aspect linguistique, sont réinsérées dans le cadre communicationnel, impliquant un échange verbal à visée persuasive. L'argumentation serait dans ce sens un type de configuration linguistique qui se présente de façon implicite ou explicite dans les textes

produits dans un cadre communicationnel spécifique (Amossy, 2013). La démarche méthodologique qu'elle propose se nourrit de théories qui ont traité de l'argumentation dans la langue, à l'exemple de l'étude de connecteurs chez Anscombre et Ducrot (1983), mais aussi de la pragmatique et de la stylistique, qui cherchent à comprendre comment les intentions des émetteurs se manifestent dans les procédés linguistiques qu'ils mettent en œuvre dans un cadre communicationnel donné.

Sur le plan du langage, Amossy (2013) privilégie l'analyse de l'argumentation dans le discours en fonction des stratégies de persuasion mises en œuvre dans les textes par la description des procédés linguistiques qui attestent des intentions des locuteurs à imposer des croyances et des schémas interprétatifs sur la réalité. Amossy relève principalement l'étude des *topoï* et des connecteurs. Les premiers sont, dans le sens aristotélicien, les lieux communs, les récits partagés par une culture, ou encore le système de valeurs qui servent de toile de fond à l'entreprise persuasive des locuteurs. Le second concerne plus spécifiquement l'enchaînement logique des énoncés dans le texte, lié à des stratégies argumentatives qui créent entre les phrases les rapports de causalité, de conséquence, d'association, de restriction, etc. D'autres facteurs importants à considérer sont l'étude de la situation de communication, la prise en compte du caractère conversationnel des échanges dans ces derniers, l'étude des genres textuels et l'étude des figures de style (Amossy, 2013, p. 31-32).

En définissant l'argumentation comme une configuration discursive, Amossy (2013) soutient que cette configuration est présente dans plusieurs genres textuels, où il est question d'une situation de communication dans laquelle prédomine la persuasion.

Le genre de l'opinion serait un de ces genres où l'argumentation se fait perceptible. Il est produit dans une situation de communication marquée par la persuasion et présente une configuration discursive argumentative. L'utilisation de *topoï* correspond à ce que Jacobs et Townsley (2011) appellent des récits et qui ont pour objectif, comme l'a expliqué Amossy (2013), d'établir un système de croyances et de valeurs visant à chercher l'adhésion de l'auditoire. Aussi, le genre de l'opinion est marqué par l'utilisation de procédés linguistiques permettant l'enchaînement logique entre les énoncés, souvent attestés par la présence de connecteurs qui établissent des relations entre les différentes idées qui sont présentées. Il

cherche à convaincre aussi par la force de la preuve, induisant les conclusions auxquelles l'auditoire doit arriver.

Dans la prochaine section, nous allons traiter d'un type d'argumentation plus présente dans les controverses, que certains auteurs ont nommée discours polémique, analysant dans quelle mesure ce discours peut aussi nous aider à caractériser le genre de l'opinion.

5.4 L'argumentation dans les controverses : quelle place pour le discours polémique ?

Pour Kerbrat-Orecchioni (1980), le discours polémique se caractérise par une situation où le locuteur veut falsifier et disqualifier un discours adverse. Dans ce sens, il relève d'une situation dialogique marquée par l'opposition à un autre discours déjà produit (par exemple, une production textuelle, une œuvre, un article de presse, un discours d'un politicien, etc.) et qui est évoqué explicitement par l'interlocuteur. Le discours polémique se perçoit dans certains genres textuels (pamphlet, libelle, propagande, manifeste, satire) comportant également une dimension argumentative, puisqu'il met au service de la défense d'un argument tout l'arsenal de stratégies argumentatives et rhétoriques pour réfuter et pour discréditer la parole d'autrui, en même temps qu'il renforce et réaffirme son propre positionnement comme le plus acceptable. Kerbrat-Orecchioni soutient aussi que le discours polémique s'inscrit dans un contexte plutôt agressif et passionnel et dans ce sens, il est distinct d'autres discours marqués par l'antagonisme et la dissension. Il se perçoit plus généralement par l'expression d'émotions négatives comme la colère, l'indignation et l'injure.

Charaudeau (2017) ne caractérise pas le discours polémique comme une situation discursive, à l'exemple de Kerbrat-Orecchioni (1980). Pour cet auteur, il faut distinguer le conflit argumenté de l'affrontement, qui est la tentative de détourner une discussion sur un problème sérieux vers un combat personnel le plus souvent improductif. Selon Charaudeau (2017), l'attitude d'affrontement en tant que jeu stratégique de destruction verbale de l'autre est propre au discours polémique et ne peut pas se placer au même rang qu'une situation interlocutive d'échange d'arguments. L'affrontement cherche à éliminer l'adversaire discursif en le disqualifiant : « Il n'y a pas de contrat de parole qui dise, *a priori* — sauf dans des cas de

mise en spectacle — que les protagonistes doivent se mettre réciproquement en cause jusqu'à vouloir se dénier l'un l'autre » (Charaudeau, 2017, p. 81).

Pour Charaudeau (2017), le discours polémique en tant que stratégie peut se manifester dans différentes situations d'interlocution comme la conversation, la discussion, le débat et la controverse. Charaudeau définit le discours polémique comme l'« ensemble de procédés d'attaque et de défense qui ne peuvent déboucher sur aucun consensus et qui font que la discussion est sans issue possible » (p.79). Le discours polémique concernerait ainsi l'attaque frontale, l'action de détruire l'argument adverse en utilisant des procédés verbaux plus agressifs comme l'ironie, la dérision ou même les injures, jusqu'à rendre la situation interlocutive contreproductive. En tant que jeu stratégique, le discours polémique pour Charaudeau peut avoir un degré plus au moins acceptable, étant donné que « le processus de réfutation et disqualification est caractéristique de tout débat contradictoire » (p. 85). Cependant, l'exacerbation des attaques, la radicalisation de l'antagonisme, le refus du dialogue raisonnable ou la violence verbale supprimeraient l'argumentation de l'interlocution.

> La polémique, elle, relève d'une stratégie qui peut traverser différents genres et qui, s'immisçant dans une controverse, la bloque par attaque des positions et rejet des propos adverses, le tout en absence d'argumentation (Charaudeau, 2017, p.85)

Garand (1998) conteste la vision de Kerbrat-Orecchioni (1980) et de Charaudeau (2017) selon laquelle le discours polémique doit se caractériser par la violence ou par l'agressivité des propos qui sont échangés par les interlocuteurs. L'objection principale de Garand (1998) est que toute caractérisation des discours en fonction des stratégies argumentatives ou rhétoriques qui sont mises en œuvre aboutit invariablement à une liste de procédés linguistiques qui sont aussi caractéristiques d'autres productions linguistiques où il est question de confrontation. Selon cet auteur, il n'y a pas de signes distinctifs de « polémicité » permettant de distinguer les différentes situations discursives antagonistes. Garand dit qu'il n'est pas non plus très productif de qualifier le discours polémique en fonction de l'intention du locuteur, puisque la volonté de déformer ou de disqualifier la position de l'autre est aussi présente dans d'autres types de discours où il est question de l'argumentation.

Ces réflexions amènent Garand (1998) à situer le discours polémique dans la notion de conflictualité, propre au débat démocratique. Selon lui, le discours polémique engage certaines fonctions du langage permettant aux sujets de « se défendre, de se distinguer, de contester, de convaincre, de provoquer, de menacer, de dénoncer, de dominer, de manipuler, de disqualifier, etc. » (p. 216). Dans cette perspective conflictuelle, le discours polémique nécessite la rencontre de « deux sujets, deux forces, deux volontés autour de problèmes qui concernent la vie collective, mais aussi les conditions de viabilité des discours » (p. 217).

La définition du discours polémique comme le terrain du conflit s'oppose aussi à l'idée normative de délibération rationnelle à laquelle Charaudeau (2017) s'affilie lorsqu'il associe le discours polémique à l'absence même de l'argumentation. Sur ce point, Garand (1998) offre une perspective intéressante, en rappelant que les échanges impliquant l'antagonisme qui présuppose certaines conditions de réalisation. Il explique que l'exacerbation d'une position polémique, comme dans le cas extrême de l'échange d'injures, entraînerait la rupture même de la discussion. Aussi, Garand (1998) rappelle que dans une situation de conflit et de divergence profonde, même si les procédés visent à délégitimer la position contraire en attaquant l'adversaire, cette stratégie ne fonctionne pas toute seule. Il faut aussi défendre son point de vue et essayer de rallier le lecteur ou le public à sa cause :

> Attaquer une Cible ne suffit pas, encore faut-il rallier le tiers à sa cause : le but d'un polémiste n'est pas que le lecteur en conclue simplement qu'il n'aime pas ou ne partage pas les idées de tel adversaire, il doit l'amener à partager son opinion afin que la Cible se retrouve en position d'isolement. (Garand, 1998, p. 231)

Amossy (2014) rejoint la perspective théorique de Garand (1998) sur l'importance du discours polémique. Cette auteure soutient que le conflit doit être vu comme une forme de socialisation et que même l'exacerbation de différends dans un discours polémique peut être une contribution à la construction d'un espace public démocratique. Pour Amossy (2014), dans les démocraties pluralistes, les divisions sont inévitables et il ne faut pas juger le discours polémique ou le dévaloriser comme un discours moins argumentatif. L'auteure définit le discours polémique comme une modalité argumentative permettant aux différents acteurs sociaux d'influencer, de protester et d'inciter à l'action.

Pour Amossy (2014), le discours polémique se distingue par un type de structure actancielle qui se compose de deux groupes antagonistes, un proposant et un opposant, qui échangent devant un tiers. Par structure actancielle, elle réfère à la représentation de ces acteurs réels dans le récit polémique et au caractère dialogique de la situation réelle que ce dernier reflète textuellement : le discours polémique répond à des discours antérieurs et anticipe également les réactions des discours opposants. Dans ce sens, le discours polémique se configure dans un type de débat dichotomique autour d'une question d'intérêt public et qui est ancré dans le conflit et dans la polarisation. Le format polémique est même associé à la circulation de discours dans la société, lequel « émerge et se consolide de la diffusion dans l'espace public, d'un foisonnement de discours et d'interactions polémiques » (p. 210).

Il est intéressant de noter que la définition de discours polémique est liée à un type de situation (Amossy, 2014 ; Charaudeau, 2017 ; Garand, 1998 ; Kerbrat-Orecchioni, 1980) et que la manière dont elle est définie rassemble en plusieurs aspects les définitions de controverses qui sont abordées dans ce chapitre (p. 68). Le discours polémique serait dans ce sens une situation dans laquelle l'argumentation est instanciée en tant que configuration linguistique, mais qui présente des caractéristiques particulières en fonction de la structure actancielle (Amossy, 2013 ; Garand, 1998) que la situation de dispute configure sur ce genre de discussion. Cette structure est composée d'un proposant et un opposant (Amossy, 2014), par l'identification d'une cible qu'il faut attaquer, mais aussi par la défense d'une position (Amossy, 2014 ; Charaudeau, 2017 ; Garand, 1998 ; Kerbrat-Orecchioni, 1980).

La définition du discours polémique ressemble à la caractérisation d'un des styles rhétoriques relevés par Jacobs et Townsley (2011) dans lequel les chroniqueurs s'engagent dans un débat du type pour-contre. Puisque les controverses font partie des débats contemporains typiques de l'espace d'opinion, nous pouvons conclure que le genre de l'opinion peut aussi comporter les caractéristiques du discours polémique, dans la mesure où certaines situations impliquent un type de dialogue entre adversaires défendant des positions contraires. Ainsi, en plus de présenter une configuration textuelle argumentative par la présence de *topoï* et de constructions logiques entre des énoncés, le genre de l'opinion peut également révéler une structure actancielle illustrant la dispute qui se produit dans une

situation de communication et qui se manifeste par l'adoption de stratégies argumentatives particulières visant la destruction discursive de l'adversaire.

5.5 Conclusion

Le genre de l'opinion est un modèle d'écriture dont la visée argumentative est explicite, et diffèrent d'autres genres journalistiques en fonction de cela (Grosse, 2001). Ces genres sont largement présents dans l'espace d'opinion (Jacobs et Townsley, 2011). Ils sont utilisés par les chroniqueurs des journaux, par les spécialistes intervenant dans les médias et les laboratoires d'idées, par les lettres de lecteurs adressés aux journaux et par tous les acteurs de la société civile qui s'organisent institutionnellement dans le but de faire valoir leurs idées dans les divers canaux de communication contemporains.

La catégorisation des genres opérés par Grosse (2001) a l'avantage de percevoir que ces formes de littérature que sont les genres (Wilson et Robinson, 1990) sont le résultat d'évolutions historiques à l'intérieur d'un champ professionnel spécifique. Les transformations technologiques, économiques et politiques par lesquelles ce champ professionnel passe ont une influence sur l'émergence de nouvelles formes d'écritures, qui peuvent d'ailleurs garder en mémoire les pratiques anciennes, en les adaptant aux demandes plus courantes du champ professionnel. Grosse (2001) catégorise les genres journalistiques en fonction d'une intentionnalité, d'un but communicatif. Le genre de l'opinion se situe dans un projet de persuasion et d'argumentation, avec un modèle d'écriture qui est hérité de la rhétorique classique. Cette origine historique nous amène à explorer dans la littérature les études sur la rhétorique et l'argumentation, afin de nous éclairer sur la manière dont ces deux courants d'études caractérisent plus précisément le genre de l'opinion.

Nous avons vu que le clivage de la rhétorique classique au Moyen Âge a eu pour effet la création de deux champs d'études de vocation manifestement normative et descriptive, restreignant d'une part l'étude de l'argumentation aux règles logiques du raisonnement et d'autre part, la rhétorique à la description des figures de style. Le caractère normatif de ces théories offre des instruments pour décrire les critères de validité des arguments, ainsi que les procédés d'expression impliquant la volonté stylistique de l'énonciateur. La vision de l'analyse du discours proposé par Amossy (2013) préconise de son côté la nécessité de renouer

avec la visée pragmatique de la rhétorique classique. Cette perspective voit dans l'intention du locuteur une composante primordiale pour comprendre la relation entre l'argumentation et la langue, et plus spécifiquement comment les ressources linguistiques peuvent être mobilisées pour persuader. Comme le remarquent Perelman et Olbrechts-Tyteca (1988), il n'y a pas de choix neutre : les procédés linguistiques utilisés par les locuteurs témoignent de leurs intentions.

Nous avons inclus dans cette revue des écrits à propos du discours polémique, qui se trouve à être un mode d'argumentation marqué par le conflit (Garand, 1998) et plus présent dans les débats politiques et les controverses. Notre analyse permet de dresser le caractère dialogique du discours polémique, dans la mesure où celui qui élabore son texte cherche non seulement à défendre un point de vue, mais à désavouer son adversaire discursif en mobilisant tous les moyens argumentatifs et rhétoriques pour y parvenir. Dans ce sens, les discours polémiques sont aussi argumentatifs, mais plus spécifiques d'une situation de communication marquée par le conflit où la divergence entre deux points de vue est évidente.

Cependant, la vision du discours polémique comme un type de discours argumentatif n'est pas partagée par tous. Charaudeau (2017) ne reconnaît pas la situation communicative du discours polémique comme un enjeu de persuasion, mais comme une tentative de destruction du discours adversaire. L'auteur va même avancer que le discours polémique est marqué par une absence d'argumentation, puisque selon lui, il vise principalement la délégitimation des points de vue de l'opposant. Cette perspective est nuancée dans les écrits d'Amossy (2014), qui écarte une vision idéaliste de délibération rationnelle et de la polémique comme d'une simple violence verbale. Elle défend la nécessité de développer une vision non normative des échanges marqués par le conflit et de concevoir la polémique comme modalité argumentative propre aux débats contemporains. Garand (1998) à son tour, soutient que la caractérisation du discours polémique par les relations de conflit et par le type de structure actancielle que cette situation produit dans le discours est plus productive que la tentative de répertorier les traces de polémicité, car ces derniers peuvent être aussi présents dans d'autres types de discours.

Pour nous, le discours polémique n'est pas l'absence de l'argumentation comme le propose Charaudeau (2017). Nous sommes d'accord avec la proposition de Garand (1998) selon laquelle l'exacerbation du conflit dans les discussions qui se produit dans ce type de

situation de communication peut entraîner la rupture de la communication. Dans le discours polémique, des stratégies pour invalider le discours contraire sont employées, mais il doit également valider ses propres arguments.

L'analyse de ces écrits nous a permis de conclure que le genre de l'opinion, dont l'origine historique remonte le passage d'une presse informative à une presse critique, est propre à une situation de communication où sont en jeu la persuasion, la volonté de convaincre quelqu'un d'un point de vue sur une question. Ils sont ainsi des modèles d'écriture qui manifestent ce but communicatif. Ils présentent une configuration linguistique argumentative, que les auteurs cités nomment discours argumentatif, pour signaler l'usage de la langue avec une visée persuasive. L'argumentation se reconnaît d'une part par cette intention persuasive manifeste et d'autre part par le choix de certains procédés linguistiques, comme l'utilisation des *topoï* et de connecteurs permettant l'imposition de schémas interprétatifs et l'agencement logique des énoncés, induisant une conclusion à l'auditoire. Le genre de l'opinion peut également présenter des aspects du discours polémique lorsqu'il est produit dans une situation d'antagonisme clair entre un proposant et un opposant, comme l'a souligné Amossy (2014). Dans ce cas, les procédés qui sont aussi argumentatifs visent non seulement la défense d'une opinion, mais aussi la réfutation de l'opinion contraire.

Nous voulons dans la section suivante présenter quelques travaux en analyse de discours qui préconisent un cadre méthodologique pour l'étude de l'argumentation et du discours polémique et analyser la pertinence de ce cadre dans le contexte de notre recherche.

6. Analyse du discours et argumentation

6.1 Introduction

Après avoir circonscrit l'opinion dans la presse comme notre objet d'étude et l'avoir définie comme l'expression linguistique d'un acte communicationnel lié explicitement à un enjeu argumentatif et de persuasion, une question s'impose : comment pouvons-nous explorer le matériel linguistique organisé dans les textes pour distinguer ces derniers en fonction des opinions qu'ils défendent ?

Cette section explore des réponses à cette question en présentant des travaux qui exposent le programme méthodologique de l'analyse du discours argumentatif et du discours polémique. Nous voulons également explorer les contributions possibles de l'analyse du discours dans notre recherche en faisant état des principaux instruments d'analyse proposés par les auteurs étudiés.

6.2 Étude du discours et perspective communicationnelle

Les approches énonciatives ont été parmi les premières à signaler les formes d'objectivation de la subjectivité des émetteurs dans l'expression linguistique de leur échange, ainsi que du contexte extralinguistique entourant cette dernière. Le principal exposant de ce courant, Émile Benveniste, postulait en 1967 que le sens se construit à partir d'une sphère subjective : un « je » projeté dans le discours qui fonde son point de vue par rapport à un « tu » extérieur et par rapport à un mode de référence objectif. En assimilant la phrase comme l'instrument élémentaire de communication et l'unité de base du discours, Benveniste (1967) a préconisé l'étude d'indicateurs linguistiques qui reflètent les points de vue du sujet impliqué dans l'énonciation : « Ce sont les indicateurs de la deixis, démonstratifs, adverbes, adjectifs, qui organisent les relations spatiales et temporelles autour du sujet pris comme repère » (Benveniste 1967, p.262).

La deixis est un mot grec qui veut dire « fait de montrer » (Paveau et Safarti, 2003). Par ce concept Benveniste (1967) voulait signaler l'existence, dans les énoncés, de marques qui attestent le fonctionnement de la langue dans les situations de communication. Par exemple, le « je » renvoie à l'instance réelle de l'énonciation, mais aussi à l'identité de cette instance

projetée dans le discours. Pareillement, le « tu » indique la personne (réelle ou représentée) à qui ce premier s'adresse. D'autres marques linguistiques signalent le cadre spatio-temporel où la communication se déroule, comme les mots « ici », « maintenant », etc. Benveniste (1967) a également distingué les marques de modalité, qui révèlent le rapport que l'énonciateur a avec son propos, c'est-à-dire la manière dont le contenu de l'énoncé est envisagé par l'énonciateur. Par exemple, les modalités appréciatives sont attestées par la présence d'adjectifs (« beau », « bon », « mauvais ») ou d'adverbes quantitatifs (« plusieurs », « beaucoup », « moyennement ») associés à des objets (substantifs) présents dans les énoncés.

Du côté de l'analyse du discours, le modèle communicationnel de Charaudeau (1992, 1994, 2006a ; 2006b) accorde une attention particulière au phénomène de l'énonciation. Dans ce modèle, Charaudeau (2006b) dit que l'origine du processus de communication se trouve d'abord dans l'intentionnalité du sujet communiquant de signifier un monde à un autre (l'interlocuteur). Le discours serait la performance de cet acte en contexte, dans laquelle l'énonciateur, en tant qu'image discursive de l'interlocuteur réel, représente des objets du monde, sa propre identité et sa finalité communicative à travers l'organisation du texte. Le discours serait également associé à une configuration linguistique qui porte l'empreinte d'une intention (par exemple, l'intention de convaincre) et dans ce cadre conceptuel, le texte est perçu comme la résultante matérielle et linguistique de l'acte communicationnel.

Dans le but d'approfondir cette réflexion et connaître comment l'AD peut contribuer à l'étude du matériau linguistique exprimé dans les textes dans une situation de communication donnée, nous présentons le modèle de communication de Charaudeau (2006a, 2006b).

Le travail de Patrick Charaudeau nous intéresse particulièrement puisqu'il considère le rôle de la situation de communication dans l'étude du sens. Cet auteur a également écrit beaucoup sur ce qu'il désigne comme le « discours médiatique », particulièrement les textes journalistiques et le phénomène de l'argumentation. Le programme méthodologique de Charaudeau (1992) pour l'analyse du discours a été aussi abordé dans le cadre de certaines expérimentations de fouille d'opinions qui revendiquent un cadre linguistique d'analyse (Vernier et coll., 2009a, 2009b).

6.3 Le modèle socio-communicationnel de Charaudeau (1994)

Charaudeau (1994) définit le processus de communication comme une transaction de sens entre deux partenaires, qu'il nomme sujet communiquant et sujet interprétant. Ces derniers sont liés par la même finalité, qui est de comprendre et de se faire comprendre et par conséquent, ils s'engagent dans une même action de construction du sens. De ce fait, le processus de communication repose sur l'intention du sujet communicant de signifier un monde à son interlocuteur. Le sujet communiquant configure ce monde à travers un acte de prise de parole dans une situation de dialogue ou dans une situation de « mise en texte » (Charaudeau, 1994, Les opérations des sujets de la communication, paragr. 14). De son côté, le sujet interprétant est chargé de reconnaître le monde qui a été signifié par l'autre. Il discerne le sens des mots qui sont énoncés en fonction de la situation de communication qui l'unit à son interlocuteur.

À l'intérieur de ce processus, le texte est conçu comme la résultante tangible de la communication et devient interprétable en fonction du contrat de communication établi entre les deux partenaires de l'échange. Par « contrat de communication » Charaudeau (1994) veut exprimer que les échanges communicationnels sont toujours situés dans un cadre socio-institutionnel qui impose certaines conventions sur ce qui doit être dit et comment cela doit être dit. Les organisations de presse, les journaux et les groupes médiatiques constituent un exemple de ce cadre institutionnel qui donne d'avance les instructions sur la production et l'interprétation des énoncés.

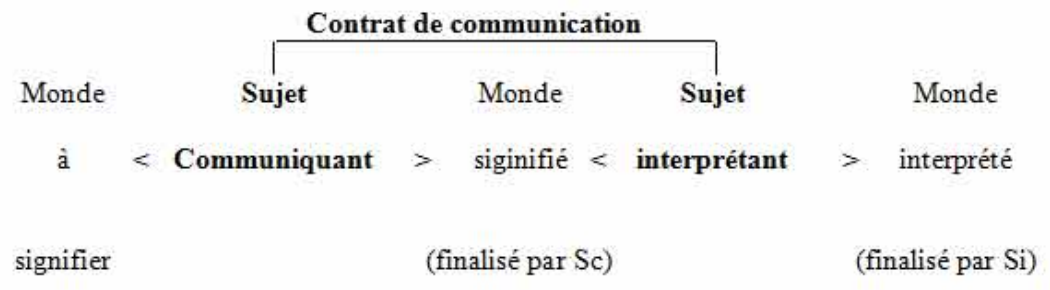


Figure 4. Modèle socio-communicationnel de Charaudeau (1994, Les opérations des sujets de la communication, paragr. 15)

Puisque le contrat de communication souligne cet aspect situé de l'échange communicationnel, Charaudeau (1994) le désigne également comme situation de communication. La situation attribue les rôles et les identités aux partenaires de l'échange, de sorte que la légitimité du sujet communicant de prendre la parole dépend de son statut social. Le sujet interprétant, pour sa part, reconnaît son rôle et son identité en fonction de l'enjeu communicatif, car il détient les compétences interprétatives exigées par le contrat (Charaudeau, 2006a). Charaudeau (1992) distingue, d'une part, l'identité psychologique et sociale qui est externe au langage et qui renvoie à la personne réelle (le locuteur) et, d'autre part, l'identité langagière qui se représente dans l'acte en tant que tel (l'énonciateur). Les identités du locuteur et de l'énonciateur ne sont pas complètement dissociées l'une de l'autre, comme le laisse entrevoir cet extrait :

> Par exemple, l'effet discursif ne sera pas le même selon que derrière l'énonciateur qui donne un *ordre* à l'adresse d'un *jeune*, il y a un locuteur *adulte* ou *jeune*, un *père*, un *professeur*, un *chef de service*, etc., car l'énonciation d'un ordre dépend de la position d'autorité du locuteur-émetteur. (Charaudeau, 1992, p. 133)

La situation de communication détermine également les propos échangés, c'est-à-dire ce qui est thématisé dans la communication (Charaudeau, 2006a, 2006b). En somme, elle fournit les « données » qu'instruisent le sujet communiquant et le sujet interprétant lors l'échange :

> Ces données fournissent (imposent) au sujet parlant des instructions discursives sur la façon de se comporter en tant qu'énonciateur, à propos de l'identité qu'il doit attribuer à son partenaire en tant que sujet destinataire, à propos de la façon d'organiser son discours (de manière descriptive, narrative et/ou argumentative), sur les topiques sémantiques qu'il doit convoquer (Charaudeau, 2006a, Introduction, paragr. 5).

Dans le modèle socio-communicationnel, le sujet communiquant est chargé de l'organisation du sens, de la « mise en texte » d'un sens qui appartient d'abord à lui-même et qu'il veut partager à un autre en fonction d'une finalité spécifique, qui peut être par exemple, dans le cas des textes argumentatifs, d'influencer ou convaincre quelqu'un de son point de

vue. Pour organiser le sens en fonction de cette finalité, le sujet communiquant se livre à deux types d'opérations langagières selon Charaudeau (1994) : la thématisation et la relation. L'organisation par la thématisation comprend les opérations qui donnent l'existence aux êtres du monde dans les discours (les personnes, les objets, les concepts) et qui rend compte de leurs propriétés, leurs changements d'état, leurs raisons d'être et d'agir. Il s'agit d'opérations d'identification, de qualification, d'explication et de description. En réalisant ces opérations, le sujet communiquant mobilise le sens des mots de son système linguistique. Il construit un réseau de relations entre les mots en fonction de leurs traits distinctifs, en les liant en fonction d'un critère de cohérence, mais aussi en fonction de son intention de communiquer quelque chose à l'intérieur d'une situation spécifique. Cette mise en relation du sens des mots lui permet également de produire des effets de « glissement de sens » qui rendent compte du sens implicite de son propos.

Le sujet interprétant, pour sa part, agit dans la co-construction du sens textuel, dans la mesure où il se situe aussi comme partenaire dans le contrat de communication et où il perçoit les marques formelles dans le texte. Sa compréhension du sens requiert l'identification des opérations qui ont présidé la construction du sens de langue par le sujet communiquant. Il réalise des inférences pour reconstruire le sens indirect et implicite sur la base de la mise en relation qui peut être effectuée entre le texte et les données externes, incluant la situation de communication, ses propres connaissances à propos de la société, ou ses connaissances à propos d'autres énoncés (œuvres, livres, articles, etc.). Charaudeau (1994) distingue d'une part la compréhension et d'autre part l'interprétation : la première est une opération de reconnaissance du sens de la langue (à l'aide entre autres des connaissances grammaticales et encyclopédiques) et la seconde est le processus de mise en relation du sens linguistique avec les conditions psychosociales qui président l'acte de communication.

La perspective théorique de Charaudeau (1994) impose une démarcation très évidente entre les dimensions linguistique et extralinguistique impliquées dans l'acte de communication. En effet, dans son modèle communicationnel, il distingue l'acte du discours (l'acte d'énoncer) et tout ce qui entoure cet acte qui n'est pas verbalisé ou repérable dans l'énoncé, mais qui est nécessaire à l'interprétation de ce dernier. Il s'agit en effet d'un « hors langage » qui renvoie aux conditions psycho-sociales entourant les productions linguistiques.

Les conditions internes, que sont les textes, les genres textuels, ainsi que les types de structuration que la langue rend possible, sont conçues comme des instruments disponibles aux locuteurs. De cette conception instrumentale des ressources de la langue découle l'idée que le texte ne peut pas être interprété en lui-même, mais seulement en fonction de la situation de communication :

> Autrement dit, les mots et les énoncés produits ne signifient pas en eux-mêmes, ils ne sont interprétables qu'en relation avec un « ailleurs » plus ou moins surdéterminant, un lieu de conditionnement que doivent partager les partenaires de l'échange : tout acte de langage est produit et interprété en fonction des conditions qui président à sa production et à son interprétation (Charaudeau, 2006b, L'articulation de l'acte de langage avec son environnement, paragr. 8).

L'interprétation constitue ainsi une opération d'inférence qui met en jeu plusieurs facteurs : doté de la connaissance de la langue, le sujet interprétant opère d'abord dans un processus de compréhension et de reconnaissance des signes linguistiques, qui se trouvent combinés dans les constructions syntaxiques. Par la suite, il reconstitue les intentions des locuteurs par une inférence qui prend en compte non seulement l'identification du sens des mots réalisés dans un contexte linguistique donné[15], mais aussi les effets qu'ils provoquent à l'égard de la situation de communication considérée, en fonction des identités et des positions sociales des locuteurs réels. En observant le discours médiatique, Charaudeau (2006b) tient compte des positions sociales des sujets et les considère essentielles à l'analyse des différents types de discours.

> Si on s'intéresse à l'analyse de différents types de discours, il sera nécessaire de s'interroger sur l'identité sociale du sujet qui est à l'origine de chacun d'eux. Par exemple, s'agissant de tel discours de presse, quels sont les traits d'identité sociale pertinents en rapport avec le texte produit : est-ce celui d'un journaliste du journal, celui d'un correspondant ou d'un envoyé spécial, ou est-ce celui d'un chroniqueur

[15] Charaudeau distingue le contexte linguistique de la situation de communication, extérieur à l'acte de langage. Le contexte linguistique est désigné comme « l'environnement verbal d'un mot considéré, quelle qu'en soit en dimension » (Charaudeau, 1992, p. 637).

> occasionnel, d'une personnalité extérieure au journal, etc. ? (Charaudeau, 2006b, Le lieu de la production, paragr. 9)

La prise en compte de phénomènes extérieurs à la langue ainsi que les différences posées entre la compréhension et l'interprétation du sens d'un énoncé ou d'un texte déterminent en grande partie l'appareil instrumental de l'analyse du discours préconisé par les travaux de Patrick Charaudeau. Charaudeau (1992) propose un ensemble de catégories d'analyse qui sont fondées sur les intentions et qui sont localisables dans les structures propositionnelles des textes. Ces structures se trouvent dans la composition des textes et sont définies comme des segments de rang supérieur à la phrase qui subsument une intention communicative. L'étude du discours quant à lui se caractérise comme un repérage de marques permettant des retracer les intentions des sujets. Nous allons aborder plus en détail ce sujet dans la section suivante.

6.4 Programme méthodologique de l'AD pour l'analyse de discours argumentatifs et polémiques

Dans l'ouvrage « Grammaire du sens et de l'expression », Charaudeau (1992) organise son appareil d'analyse du discours. L'auteur définit un ensemble de catégories de la langue qui décrivent cette dernière du point de vue des intentions de communication. Au total, sept catégories de la langue sont définies : 1) nommer ; 2) déterminer (fournir un mode d'identification aux êtres) ; 3) qualifier ; 4) décrire (les actions et les processus) ; 5) situer (dans le temps et l'espace) ; 6) expliquer (les relations logiques et les raisons d'être) ; 7) dire le point de vue (le positionnement face à ce qu'il est dit). À chacune des catégories de la langue, Charaudeau associe une fonction et une catégorie grammaticale dominante. Par exemple, les noms propres correspondent à l'opération de nommer, tandis que les verbes, à l'opération de décrire les actions. Par extension, le discours, qui est la mise en texte de l'acte de communication comprend des modes d'organisation spécifiques qui peuvent être dominants dans un texte : 1) Énonciatif, 2) Descriptif, 3) Narratif et 4) Argumentatif.

Chacun des modes d'organisation du discours élabore différemment la relation entre les objets du discours, c'est-à-dire les choses et les êtres. Dans le cas du discours argumentatif, le sujet communiquant emploie les catégories grammaticales qui établissent des rapports de

causalité entre les objets du discours et qui rendent compte de la force du lien causal (de possibilité, de probabilité, de nécessité ou d'inéluctabilité) (Charaudeau, 2007). Par la thématisation, il peut choisir une façon particulière de nommer les objets et d'ainsi cadrer le propos de manière à aborder les thèmes qui lui sont plus favorables. Il peut également employer des stratégies de captation pour faire adhérer son lecteur, en attribuant une axiologisation[16] aux objets. Ces stratégies peuvent se faire à partir de l'emploi d'un vocabulaire dramatisant ou émotif ou encore par l'utilisation de procédés qui donnent un style polémique, persuasif ou dramatisant. Par conséquent, les catégories grammaticales sont des indices qui peuvent faire l'objet de l'analyse du discours, dans la mesure où elles permettent de retracer l'intention du sujet communiquant.

> Le locuteur, plus au moins conscient des contraintes et de la marge de manœuvre que lui propose la situation de communication, utilise certaines des catégories de la langue qu'il ordonne dans les modes d'organisation du discours pour produire du sens, à travers la mise en forme d'un Texte. Pour le locuteur, parler est donc affaire de stratégie, tout comme s'il demandait : « Comment vais-je/dois-je parler (ou écrire), étant donné ce que je perçois de l'interlocuteur, ce que j'imagine qu'il perçoit, et attend de moi, du savoir que lui et moi avons en commun, et des rôles que lui et moi devons jouer ? » (Charaudeau, 1992, p. 642)

Dans la « Grammaire du sens et de l'expression », Charaudeau (1992) définit l'argumentation comme une catégorie discursive qui peut être présente dans un texte comme un mode « d'organisation dominante » et manifestée par une configuration textuelle spécifique (p. 645). Plus généralement, cette catégorie du discours comprend la performance d'autres

[16] L'axiologie est une branche de la philosophie qui s'intéresse à l'étude des valeurs morales, particulièrement liée au champ de l'éthique et de l'esthétique (du grec : *axia*, valeur, qualité). Dans le champ de la linguistique, l'axiologie renvoie à un ensemble d'opérations et de marques de « l'ordre de l'affectif » (Charaudeau, 1992, p. 604) que l'énonciateur utilise pour exprimer un jugement de valeur de type bon/mauvais ou souhaitable/regrettable. Elle diffère de la sémantique en ce sens qu'elle ne traite pas du sens en général, mais de l'analyse du sens dans une langue particulière, ou « l'étude des éléments de sens qu'une langue donnée retient pour élaborer le signifié des unités significatives » (Walter, 2001, p. 59). Nous employons le terme axiologique ici pour parler de ces marques, qui reposent sur un jugement de l'ordre de l'affectif, comme le propose Charaudeau (1992).

sous-tâches discursives, qui sont 1) la problématisation ; 2) le positionnement ; et 3) la preuve. Problématiser, selon Charaudeau (1992) c'est poser un thème, ce dont il est question dans le discours. Dans le positionnement, le locuteur donne son point de vue sur le thème posé et pondère les assertions proposées, en démontrant les avantages et les inconvénients de chaque position. Prouver, pour sa part, est relatif à la tâche d'assurer la validité de la prise de position, en établissant des rapports de causalité entre les assertions du discours et la force du lien entre elles. Charaudeau (1992) renoue également avec la rhétorique classique en proposant l'argumentation comme une double quête, la quête d'une part d'un idéal de rationalité et d'autre part d'un idéal d'influence (p. 784-785). Si la quête de rationalité repose plus spécifiquement sur les activités discursives liées à la composition logique des arguments, l'enjeu d'influence repose sur des stratégies, aussi linguistiques, qui visent à légitimer le locuteur en suscitant, chez l'interlocuteur, une empathie, une identification ou une admiration.

C'est au niveau de l'organisation énonciative du discours que Charaudeau (1992) identifie une des principales fonctions du langage. L'auteur définit l'énonciation comme un phénomène d'appropriation de la langue par le sujet parlant, dans lequel ce dernier est « amené à se situer par rapport à son interlocuteur, par rapport au monde qui l'entoure, et par rapport à ce qu'il dit » (p.572). D'un point de vue concret, l'énonciation se manifeste par des marques linguistiques qui renvoient à la position du sujet parlant : il s'agit de systèmes formels composés par les catégories grammaticales comme les pronoms personnels, les démonstratifs, le temps et les modes verbaux (indicatif, subjonctif, impératif, etc.). Mais l'énonciation peut être présente aussi dans d'autres marques linguistiques comme les adjectifs et les adverbes, qu'il faut chercher dans l'organisation du discours. Dans la « Grammaire du sens et de l'expression », Charaudeau analyse l'utilisation de ces marques linguistiques dans différents types de discours, en les associant à des effets de sens produits ou aux identités et rapports entre les interlocuteurs. L'utilisation du « vous » par exemple, peut conférer à la parole de quelqu'un un effet de politesse, au même titre que « tu » pourrais signifier une proximité réelle entre deux interlocuteurs.

D'ailleurs, l'expression de l'opinion est pour Charaudeau (1992) une modalité énonciative qui présuppose « un fait (ou une information) à propos duquel le locuteur se positionne en fonction de son « univers de croyance » (p. 601). Selon l'auteur, la variété des

modes d'expression des croyances caractérise les divers types d'opinions et atteste le degré de certitude du locuteur face à son propos. Par conséquent, les marques comme « je pense que », « je suis persuadé que », « je suis certaine » s'opposent sur le plan du degré de certitude aux constructions comme « j'ai le pressentiment que » ou « je suppose que ». Charaudeau distingue aussi la modalité énonciative de l'appréciation, dans laquelle le locuteur n'évalue pas la vérité du propos, mais sa valeur. L'expression prend la forme d'un jugement qui se fait en fonction des quatre « domaines de valeurs » auxquels il associe des exemples de marques qui l'explicitent : la valeur Éthique (« je trouve bien/mal que… »), la valeur Esthétique (« je trouve beau/laid que… »), la valeur Hédonique (« je trouve heureux/malheureux ») et la valeur Pragmatique (« trouve utile/inutile que… »).

En parlant de la modalisation, Charaudeau (1992) explique que les marques linguistiques permettent de mettre en évidence la construction identitaire de l'énonciateur/locuteur, sa prise en charge de l'énoncé ainsi que ses intentions. Cependant, il faut songer que ces marques ne sont pas monosémiques et que parfois, la modalisation est inscrite dans l'implicite du discours, lequel doit être analysé en fonction de la situation de communication :

> Une même marque peut recouvrir différents sens, selon les particularités du contexte dans lequel elle se trouve (polysémie). Par exemple, le verbe vouloir peut exprimer : - un « désir » dans : « Je veux partir » — un « ordre » dans : « Je veux que tu partes ! » ou dans : Veux-tu te tenir tranquille ! » — un « souhait » dans : « Je voudrais tellement partir » — une « demande » dans : « Veux-tu venir avec moi ? » (…) Par exemple, l'énoncé : « Je reviendrai demain », qui ne comporte pas de marque explicite de modalisation, pourra exprimer, selon la situation de communication : une « promesse », une « menace » (ou un « avertissement »), une « assertion d'évidence », ou simple acceptation (Charaudeau, 1992, p. 573).

La citation ci-dessus explique que les marques linguistiques ne sont pas monosémiques et que par conséquent, une même marque est utile pour expliquer une structure d'organisation du discours. Pareillement, les intentions du sujet communiquant ne sont pas repérables uniquement à travers les marques, mais également par un contexte extralinguistique, soit la situation de communication dans laquelle l'acte d'énonciation se déroule. Puisque les marques ne sont pas monosémiques et que l'absence d'une marque spécifique n'est pas indicative de

l'absence d'une intention communicative, il revient à l'analyste de se rapporter à la situation de communication pour effectuer l'interprétation des procédés qui relèvent de l'implicite du discours, comme les effets de « glissement de sens » dont parle Charaudeau (1992). La situation en tant qu'élément externe et non linguistique se présente comme un lieu de contraintes permettant l'élaboration de certaines hypothèses à l'égard de la position sociale des sujets et de leurs rapports à l'intérieur de la situation de communication. Mais elle est également un lieu de références culturelles propres à une communauté, ce qui permet de comprendre les effets de glissement de sens associés à des procédés métaphoriques, notamment.

En définissant l'argumentation comme une catégorie du discours transversal à plusieurs textes, l'AD préconise que l'étude de l'argumentation dans le discours doive considérer les stratégies qui relèvent de l'explicite et de l'implicite. Les éditoriaux, les billets de blogues ou les commentaires publiés dans les sections de « libre opinion » des journaux constituent des exemples de genres textuels où cette visée argumentative est plus explicite. Par contre, d'autres genres journalistiques, comme les nouvelles, peuvent avoir une dimension argumentative qui n'est pas complètement assumée par le locuteur (Emediato, 2011), mais qui se laisse entrevoir par les choix linguistiques qu'il effectue.

En analysant les titres de nouvelles au sujet de la politique dans des magazines brésiliens, Emediato (2011) a relevé un ensemble de stratégies implicites qui sont utilisées pour influer l'interprétation. Il relève ainsi des stratégies de nature argumentative effectuée sur le plan référentiel (thématisation, nomination, désignation), de la prédication (qualification) et de l'énonciation (modalisation) :

- Stratégie référentielle : concerne de façon générale le choix des thèmes abordés, mais aussi l'activation dans la mémoire de l'interlocuteur de valeurs symboliques associées à certains substantifs utilisés pour désigner les êtres (par exemple, désigner un agent agresseur par son identité ethnique).
- Stratégie de prédication : dans les nouvelles, la prédication consiste à se référer aux dires d'autrui par des verbes d'attitude qui représentent l'opinion du locuteur ou qui peuvent produire une interprétation favorable ou défavorable. Par exemple, le verbe « attaquer » suscite une interprétation plutôt défavorable

dans le contexte de la phrase « X attaque Y » tandis que le verbe « critiquer » est interprété de façon plutôt favorable.

- Stratégie d'énonciation : concerne la représentation de l'objet du discours par des mots indiquant l'hypothèse, la probabilité ou la certitude et qui impose certains points de vue sur ce qui est proposé et sur ce qu'il faut croire.

L'exemple fournit par Emediato (2011) sur les types des biais qu'un simple titre d'une nouvelle peut véhiculer, renforce l'idée que les stratégies argumentatives ne sont pas uniquement élaborées à partir d'un schéma ordonné et composé d'épisodes spécifiques comme suggère Charaudeau (1992) à propos de l'argumentation (thématisation, problématisation et preuve). Le genre de l'opinion peut avoir une élaboration argumentative plus affichée, mais rien n'empêche que ce genre de stratégie relevant de l'implicite ne soit employée. Le type de proposition théorique pour caractériser le discours que fait Charaudeau (1992) est critiqué par certains. Par exemple, Garand (1998) s'oppose à caractériser les discours en fonction de la présence de marques explicites, puisque plusieurs types de discours peuvent présenter des configurations linguistiques semblables, ce qui rend difficile l'entreprise de caractérisation de leurs différences.

Dans la section suivante, nous allons poursuivre la réflexion au sujet de l'expression linguistique de l'opinion en abordant un type de discours qui est très caractéristique des controverses, le discours polémique. Différemment de Charaudeau (1992), Garand (1998) s'intéresse à décrire le discours en fonction de l'établissement d'une structure actantielle lors d'une situation de communication où il y a présence d'une divergence marquée, et pas par un ensemble de marques linguistiques. Nous expliquons la perspective de Garand (1998) dans la section suivante.

6.5 Stratégies argumentatives du discours polémique

Comme nous l'avons vu, Garand (1998) dépeint le discours polémique à partir d'une structure actancielle qui fait état d'un différend. Pour Garand (1998), le discours polémique émane d'un sentiment de malaise provoqué par un tort commis par une autre personne (l'adversaire) et qu'il est nécessaire de réparer. Ce tort ne peut être jugé par aucune loi : il met

en jeu des positions idéologiques, politiques ou personnelles qui menacent l'existence même des groupes concernés.

> Le tort crée un différend duquel il n'est pas possible de dégager de loi commune, puisque le différend est précisément la confrontation de deux règles dont la supériorité de l'une sur l'autre ne peut être décidée par aucun Tiers. Le différend traduit l'impossibilité d'un métadiscours qui ferait loi (Garand, 1998, p. 241).

La perception du tort entraîne une résistance, établissant dans le récit des coupables et des victimes, qui sont selon Garand (1998) les « pivots de l'opération polémique » (p. 223).

> Au contraire, on assiste à des résistances idéologiques, établies en fonction des intérêts des groupes concernés, qui tendent spontanément à résister devant tout ce qui les obligerait à se redéfinir. L'inconnu est souvent synonyme de mort pour les individus, à fortiori pour les groupes sociaux pour qui la pérennité représente le seul possible envisageable (Garand, 1998, p.222).

Garand (1998) défend que l'analyse du discours polémique doive reconstruire les sources de la situation conflictuelle de manière à mettre en évidence les rapports entre le sujet et l'anti-sujet, c'est-à-dire l'identité discursive et personnelle que l'énonciateur se crée ainsi que celle qu'il forge pour son adversaire. L'analyse du discours polémique doit également éclairer les rapports entre ces deux protagonistes ainsi que : 1) le tort à réparer, 2) le tiers placé comme public, et 3) la référence, qui est le système de valeurs fondamentales — associées à une utopie ou à une idée de ce qui devrait exister dans le meilleur monde possible – qui est censé fonctionner comme arbitre de la dispute.

Garand (1998) décrit les stratégies argumentatives qui permettent de reconnaître le discours polémique. Ces stratégies peuvent se départager selon deux axes principaux :

1) Le rapport entre le sujet et l'anti-sujet, dans lequel se trouvent les stratégies argumentatives relatives à des catégories axiologiques (le bien et le mal, l'admirable et l'exécrable) et qui font appel au *pathos* du public, à l'expérience émotive des individus.

2) Le rapport entre les protagonistes du conflit et le tort constituant le pivot du discours polémique dans lequel se trouvent les types d'arguments qui visent, d'une part, à contester le découpage référentiel opéré par l'adversaire et, d'autre part, à dénoncer les vices de raisonnement de ce dernier, en signalant les contradictions, les incohérences ou les incompatibilités dans son discours.

Nous présentons dans les points suivants les principales stratégies identifiées par Garand (1998).

6.5.1 Le rapport entre le sujet et l'anti sujet

Dans cette catégorie se trouvent les stratégies de démonisation visant à construire un *ethos* négatif de l'adversaire. Par exemple, « associer adroitement certaines caractéristiques de l'adversaire à des traits de figures emblématiques qui ont subi l'opprobre de la collectivité » (Garand, 1998, p. 236) ; l'argument *ad hominem* qui « consiste à confronter le discours de l'adversaire à ses actes pour en faire ressortir l'inconséquence ou la contradiction » (p. 236) ; ou encore, disqualifier les partisans de l'opposant : « Si l'autorité de l'adversaire est difficile à déboulonner, on la met en doute en lui associant certains de ses partisans qui paraîtraient suspects » (p. 237). D'autres stratégies liées à l'inscription de l'anti-sujet dans le discours pour l'identifier de façon négative sont le refus de nommer ce dernier par son nom réel, ou utiliser des mots qui expriment le dégoût ou l'ironie envers lui. Par exemple, l'utilisation de « Monsieur » peut constituer selon Garand une expression ironique visant à ridiculiser l'adversaire dans le discours.

Garand (1998) souligne l'élaboration de ces procédés d'ironie, de distance, de rejet et d'humour à l'aide de marques graphiques (guillemets, italiques parenthèses, majuscule, points d'exclamation, d'interrogation et de suspension) pour exprimer la désolidarisation envers les mots ou les expressions utilisés par l'autre. Parmi les attaques personnelles élaborées dans le discours polémique, les injures constitueraient un cas limite de disqualification et présenteraient le risque, comme nous l'avons vu, de briser le contrat de parole entre les interlocuteurs.

Le sujet doit aussi se construire un *ethos* positif et une image valorisante, pour le légitimer en tant qu'autorité dans son discours. Les arguments qui fondent cette autorité sont

relatifs à sa crédibilité, à son expérience (ce qu'il a vu ou ce dont il a été témoin), sa nationalité, sa scolarité, sa neutralité partisane, etc. Il peut faire appel à d'autres sources d'autorité externes comme des experts qui sont en faveur de sa cause, des connaissances scientifiques ou historiques, des statistiques, etc. Ces mêmes éléments peuvent être utilisés pour établir une image *a contrario* de l'adversaire, en lui imputant par exemple de l'inexpérience, de l'attitude partisane, ou encore en dénonçant les sources d'autorité douteuses éventuellement citées par lui (Garand, 1998)

Les modalisateurs de discours sont également des recours linguistiques qui relèvent les stratégies du rapport du sujet et de l'anti-sujet. Certains procédés permettent d'atténuer la portée de ses propres énoncés (« peut-être », « semble-t-il ») ou au contraire, les renforcer (« tout porte à croire », « c'est évident »), ainsi que de relativiser ceux de l'anti-sujet (« croit-il », « prétend-il »). Le registre de langage permet également de projeter une certaine image du sujet (soit intellectuelle, une image de mépris, d'indignation) et de comprendre quel type de lectorat ce dernier cherche à convaincre. Outre ces aspects, Garand (1998) relève aussi l'utilisation de verbes d'action dans le discours indirect, qui servent à donner une orientation au point de vue de l'anti-sujet, par exemple, « prétendre », « affirmer », « avouer », « déclarer » « insinuer ».

6.5.2 Le rapport entre les protagonistes du conflit et le tort

Parmi les stratégies qui contestent le découpage référentiel se retrouvent les procédés de négation de réfutation du discours de l'adversaire, comme l'élaboration d'exemples a contrario, qui sont des cas invalidant les arguments ; le déplacement du problème (« la n'est pas la question ») ; ou encore la hiérarchisation des problèmes en vue de cadrer ce qu'il faut discuter (« ce problème est secondaire », « le vrai problème est »). Quant aux stratégies visant dénoncer les vices de raisonnement de l'anti-sujet, se retrouvent les procédés de distinction métalinguistique, qui essayent de corriger les propos de l'adversaire interlocutif (« ce n'est pas la même chose », « ne confondez pas la fierté avec l'orgueil ») ; le renversement du point de vue, visant à valoriser des choses qui se sont présentées de façon négative dans le discours opposant ; et la reformulation, qui essaie de rendre explicite ce qui était implicite dans le

discours adverse et mettre en lumière le non-dit de l'adversaire, ce qu'il cache (« ce que l'on ne dit pas cependant ») (Garand, 1998).

6.6 Conclusion

Au début de cette section, nous avons posé la question suivante : comment pouvons-nous explorer le matériel linguistique organisé dans les textes pour les distinguer en fonction des opinions qu'ils défendent ? Le programme méthodologique de l'analyse du discours préconise que l'analyse du sens dans le discours argumentatif se fasse par le repérage de marques explicites et implicites et par l'interprétation de ces marques en fonction d'un contexte extralinguistique fourni par le cadre socioculturel et institutionnel dans lequel se déroule la communication entre les interlocuteurs.

La perspective de l'analyse du discours est axée sur l'intention des sujets et l'argumentation, définie comme configuration discursive qui peut être plus ou moins prédominante dans les échanges linguistiques, comme les textes (Charaudeau, 1992 ; Amossy, 2013). Cette configuration linguistique se fonde sur une intention spécifique dans le cadre d'échanges où l'interlocuteur veut persuader un autre du point de vue qu'il défend. Les caractéristiques plus proprement linguistiques de textes argumentatifs sont à chercher dans les modes de structuration et dans les stratégies argumentatives employées. Les écrits analysés convergent pour dire que l'empreinte de l'intention du locuteur est plus au moins présente dans la composition textuelle et plus accessible dans l'étude des thèmes, de l'énonciation et de la modalisation en langue. Charaudeau (1992, 2007) indique certains procédés discursifs propres à l'argumentation : la proposition d'un enjeu, la manifestation explicite d'un positionnement, et la présentation d'une preuve ou une justification qui appuie son positionnement.

Cependant, l'argumentation peut également contenir des procédés persuasifs qui ne sont pas repérables par les marques et qui sont seulement compris en se rapportant à la situation de communication. Dans ce contexte, la situation de communication est un élément important pour la compréhension de l'enjeu communicatif et des stratégies argumentatives employées. Comprendre l'enjeu de la communication et la position sociale des sujets dans l'interlocution permet de connaître les référentiels sociaux et culturels partagés et d'inférer le

sens des procédés persuasifs : l'emploi d'expressions à visée manipulatrice, ou encore les glissements de sens qui s'opèrent grâce à certains choix linguistiques font partie des procédés inférés à partir de la situation.

Nous avons abordé aussi dans notre revue de la littérature le discours polémique, lequel est associé à des situations de communication qui sont marquées par le conflit et la tentative d'invalider l'adversaire discursif par le moyen de stratégies persuasives.

Garand (1998), différemment de Charaudeau (1991), ne propose pas d'étudier le discours en fonction de marques explicites, puisque comme il l'a expliqué, plusieurs marques indiquant la présence du discours polémique peuvent être présentes dans d'autres types de discours, compliquant la détection de la spécificité du discours polémique. Il propose ainsi d'étudier le discours polémique par la reconnaissance d'une structure actantielle composée de deux protagonistes, le sujet et l'anti-sujet, en plus de la présence d'un tort qu'il faut réparer. Nous avons répertorié les stratégies argumentatives qui permettent de reconnaître cette structure actantielle dans le discours polémique.

7. Conclusion du chapitre : discussion sur la revue de la littérature

Cette section résume les principales conclusions tirées de la revue de la littérature, rappelant les définitions retenues dans le cadre de cette recherche. Elle analyse également l'application du cadre méthodologique de l'AD dans le contexte de la fouille d'opinions, en présentant les possibilités et les limites.

7.1 Systèmes de recommandation et diversification

Le premier objectif de notre revue de littérature est de présenter l'état de l'art sur le problème de recherche. Nous avons premièrement décrit les méthodes et les scénarios de recommandation afin de comprendre le fonctionnement de ces systèmes. Notre revue de la littérature constate que le choix de la méthode de recommandation (collaborative ou basée sur le contenu) dépend des types de documents à recommander et aussi des caractéristiques propres à la production et à la consultation de ces documents. Dans les SRAP, la prépondérance de la méthode basée sur le contenu s'explique par la disponibilité de données textuelles provenant des articles de presse eux-mêmes, ce qui permet de rendre le processus de recommandation indépendant des évaluations des usagers. Nous avons également circonscrit notre objet de recherche sur la question de la représentation vectorielle des articles lors du processus de recommandation. Notre recherche ne vise pas à créer une méthode de recommandation, mais à proposer un cadre expérimental dans lequel les articles d'opinion thématiquement analogues sont représentés dans un format optimal, afin que les classifieurs puissent détecter le type d'opinion qu'ils véhiculent.

Nous pensons que les articles d'opinion ont une durée de vie plus grande que les nouvelles et que certains problèmes liés à la recommandation de nouvelles ne sont pas pertinents dans notre contexte de recherche, qui envisage la recommandation d'articles d'opinions exclusivement. Ainsi, la question de la récence est moins importante dans le contexte de l'opinion journalistique. Les controverses ont aussi la caractéristique d'avoir une durée de vie plus grande que les événements qui circulent quotidiennement dans la presse.

Comme souligne Charaudeau (2017), les controverses détiennent une popularité variable et leur intérêt peut être rallumé chaque fois qu'un nouvel événement suscite le retour au débat.

Dans notre revue de la littérature, nous avons également exposé l'état de l'art des recherches qui s'intéressent à augmenter la diversité de contenu dans les SRAP. Nous avons vu que dans la plupart des études qui concernent ce problème, la notion de diversité n'est pas nécessairement associée à la diversité d'opinions par rapport à un thème d'intérêt public. Elle est définie comme une mesure de distance entre chaque article d'un ensemble recommandé dans une liste. Cette mesure qui est calculée en fonction de certains traits discriminants présents dans les articles (comme les entités nommées ou les mots). Par contre, les recherches en fouille d'opinions qui explorent la diversité dans les SRAP s'attardent surtout sur la différence entre les perspectives véhiculées dans les articles à propos d'un thème spécifique. Le problème avec cette dernière approche est que la perspective est envisagée comme une attitude émotionnelle face à un événement. Dans les systèmes de recommandation proposés par ces recherches, les articles sont distingués par le type d'émotion que la lecture de l'article suscite chez le lecteur.

Nous avons vu dans la revue de la littérature que les travaux de fouille d'opinions tendent à associer l'expression de l'opinion dans les textes à l'emploi d'un vocabulaire subjectif. Les travaux posent aussi la phrase comme unité minimale d'inférence de cette expression. L'analyse des enjeux épistémologiques touchant la conception de l'opinion pour les sondages d'abord, et pour les sciences sociales ensuite, nous a permis de tracer une différence entre deux conceptions de l'opinion : une conception instrumentale, assimilée à un énoncé appréciatif d'une source vers une cible et une autre, qui prend en compte l'aspect socialement institué des échanges communicationnels et qui reconnaît l'influence de la situation de communication et du cadre énonciatif institué sur les productions textuelles. Afin de combler cette faille théorique, nous avons dans notre recherche exploré la voie des théories discursives qui s'insèrent autant dans une problématique du texte que dans l'empreinte significative de l'énonciateur dans ces productions. Cette voie a été déjà explorée par certains travaux dans le domaine de la fouille d'opinions (Eensoo et Valette, 2012, 2014a, 2014b, 2015 ; Valette, 2004 ; Vernier et coll., 2009a ; 2009b).

7.2 L'opinion dans une perspective discursive

La recension d'écrits au sujet de l'opinion a permis notamment de rendre compte de l'évolution historique de l'opinion dans la presse et des études sociologiques qui se sont intéressées au phénomène de l'opinion publique. Nous avons circonscrit la notion d'opinion dans notre recherche comme l'expression linguistique d'acte communicationnel socialement institué par la pratique journalistique et qui sert d'instrument de contestation à l'intérieur du débat public. Ce choix nous a permis de cibler certains genres textuels qui sont propres à la pratique journalistique de production et de diffusion de l'opinion et de comprendre que l'expression de l'opinion s'intègre à un enjeu argumentatif de visée persuasive. Un panorama des théories discursives au sujet de la rhétorique et de l'argumentation a dévoilé la spécificité de l'élaboration textuelle argumentative, mais a aussi précisé le caractère polémique de cette dernière dans les controverses qui se produisent dans l'espace d'opinion.

Les auteurs analysés définissent l'argumentation comme une catégorie discursive, un type d'organisation textuelle qui peut être plus ou moins dominant dans un texte et dont les procédés peuvent être implicites ou explicites. Le programme méthodologique de l'analyse du discours préconise que l'étude du sens dans le discours argumentatif doive se faire par le repérage de stratégies de persuasion, perçues par les choix linguistiques effectués au niveau de l'énonciation et de la modalisation, principalement. Les procédés implicites peuvent être inférés en fonction de l'identité psychosociale des interlocuteurs ou en fonction de certains choix lexicaux qui encadrent le discours dans un ensemble de valeurs culturelles et morales partagées par la culture. Dans le cas des controverses, où il y a une situation de conflit déclaré, les articles d'opinion ont une visée argumentative plus avouée et présentent des stratégies polémiques, avec des procédés qui visent à délégitimer un groupe adversaire ou à invalider une opinion contraire, soit les attaques personnelles et les arguments d'ordre moral, qui en constituent de solides exemples.

La linguistique énonciative a préconisé l'étude des deixis, qui sont des marques grammaticales portant les traces de la situation de communication et de la subjectivité du sujet à l'origine de l'acte du langage. Charaudeau (1992) a reformulé cette hypothèse en distinguant les modes d'organisation du discours et les stratégies qui permettent au sujet de signifier son monde, tels que la thématisation, l'organisation énonciative et la modalisation. Il propose

d'outiller l'analyse du discours d'un ensemble de catégories discursives associées à des marques grammaticales. Celles-ci permettraient de déceler les stratégies discursives mises en place. Cependant, le fait que certains énoncés puissent être porteurs d'un type d'opération discursive même s'ils ne présentent pas de marques explicites laisse croire que d'autres observables linguistiques sont susceptibles de participer à une certaine configuration discursive dans les textes.

7.3 Application du cadre de l'AD dans les applications de fouille d'opinions : perspectives et limites

Dans une perspective quantitative, sur laquelle se basent les traitements informatiques, la conception théorique de Charaudeau (1992) invite à repérer les marques linguistiques qui signalent les objets du discours pour envisager les corpus textuels et pour en étudier le sens. Il s'agit de voir le discours comme une macro-proposition qui pourrait être décrite par une syntaxe discursive. Par exemple, pour Charaudeau (1992), l'expression de l'opinion dans les discours est marquée par l'emploi de certains types de modalités énonciatives qui permettent aux locuteurs d'extérioriser les croyances, les attitudes et les appréciations. Suivant cette perspective, l'analyse de l'opinion dans un texte doit identifier les marques explicites qui font état d'un type spécifique d'opération de mise en discours (par exemple, l'opération de qualification) et caractériser les différentes intentions associées à ces procédés.

Par intérêt pour modéliser le langage en fonction des perspectives linguistiques associées aux discours, les recherches de Vernier et coll. (2009a ; 2009b) se basent sur le cadre théorique proposé par Charaudeau (1992), plus spécifiquement sur les modalités qui interviennent dans l'expression d'une évaluation : l'accord et le désaccord, l'acceptation et le refus ainsi que le jugement. Dans leur étude pour la catégorisation automatique des évaluations de billets de blogues multidomaines, Vernier et coll. (2009b) ont constitué un lexique à partir d'un corpus d'expérimentation annoté, contenant un ensemble de marques d'expression associées à la modalisation dans le discours. Les « structures évaluatives » servant à cette ressource ont exploré entre autres les modalités d'expression subjective (« je pense », « je doute »), les configurations énonciatives (exclamations) et le lexique axiologique (« beau », « laid », etc.) proposé par Charaudeau (1992). Les auteurs ont d'abord créé un

composant d'extraction de « structures évaluatives », contenant les marques d'expression qui sont répertoriées dans l'ouvrage de Charaudeau, puis ils ont généralisé les règles de ces structures par apprentissage supervisé, afin de prédire le type d'évaluation contenue dans un corpus de texte à propos d'une entité (en l'occurrence « sushi » et « Sarah Palin »). Les résultats, qui sont encourageants du point de vue de la tâche, présentent cependant certaines limites qui ont affecté surtout le taux de rappel de la classification. D'une part, ils ont constaté que les unités lexicales évaluatives présentent des polarités différentes selon le domaine, observation qui est récurrente dans les études basées sur des approches symboliques. D'autre part, ils ont observé que pour certains types de domaines, comme l'évaluation au sujet d'une personne (« Sarah Palin »), les opinions ne sont pas modalisées par des marques d'expression qui inscrivent l'énonciateur dans le discours (par exemple, « je pense »), mais par des phrases plutôt objectives (« Sarah Palin est… »), qui n'étaient pas repérées par l'ensemble des marques prédéfinies.

En choisissant le cadre théorique de Charaudeau (1992) dans le contexte de notre recherche, la modélisation des textes numériques dans le format vectoriel se confronte à la sélection de bonnes marques qui puissent décrire de façon assez exhaustive les procédés linguistiques typiques de situations de communication où il est question de la confrontation. L'appareil méthodologique de Charaudeau permet de décrire les rapports logiques entre les assertions d'une phrase grâce aux marques qui signalent l'enchaînement entre les énoncés : par exemple, le rapport d'opposition ou le rapport d'équivalence entre deux assertions (p. 789). Une voie possible de l'emploi de ce cadre serait de modéliser les procédés argumentatifs qui visent à défendre ou à pourfendre un argument (par exemple, les arguments pour ou contre la grève étudiante). Mais la diversité de procédés et de stratégies discursives qui sont employés dans le type de discussion controversée nous fait croire qu'une telle modélisation serait réductrice, à cause des caractéristiques du discours argumentatif et polémique que nous avons soulevés dans le cadre de cette revue de littérature : il peut contenir des stratégies explicites et repérables en fonction des marques correspondant à certaines catégories du discours (modalisation, prise en charge énonciative, axiologisation accentuée), mais aussi implicites, qui sont liées au découpage référentiel opéré par les locuteurs et très dépendant des choix lexicaux et des références culturelles partagées par la communauté de

lecteurs. La prédominance de certains thèmes, la lexicalisation spécifique de ces derniers et les stratégies liées à des références culturelles comme l'ironie ou l'humour font partie de cette diversité propre du discours argumentatif et polémique, selon les auteurs étudiés.

Dans le cadre du développement du système de filtrage de sites racistes sur l'Internet, Valette (2004) a relevé l'importance des néologismes pour discriminer les sites racistes des sites antiracistes, en évaluant la précision de morphèmes associés à ces néologismes pour la performance du système de filtrage (par exemple « judéophobie » pour les antiracistes et « judéophilie » pour les racistes). Son approche, inspirée de la sémantique interprétative, a utilisé des techniques textométriques d'analyse de corpus pour faire une étude préalable de ce dernier et pour découvrir, à partir des contrastes révélés par ces techniques, comment chaque groupe emploie le langage pour défendre ses points de vue. Une approche qui ne prend comme point de départ que quelques catégories morphosyntaxiques correspondant à des opérations de mise en discours ne pourrait pas voir ce genre de spécificité. Autrement dit, si nous adoptons le cadre méthodologique de Charaudeau (1992), l'univers de marques linguistiques permettant de caractériser les opinions véhiculées dans les textes risque fort de se limiter à certaines marques linguistiques qui sont propres à la modalisation de phrases évaluatives et ignorer d'autres qui peuvent être pourtant fortement caractéristiques et d'un type opinion à caractériser.

Pincemin (1999b) dans sa thèse parle de deux approches pour aborder les corpus : l'approche par dotation et l'approche par érosion. Dans la première, l'analyse se fait à l'aide de dictionnaires et de catégories prédéfinies qui extraient les données textuelles censées correspondre à la représentation des phénomènes à caractériser dans les textes. L'approche par érosion, par contre, cherche à faire émerger les mots pertinents, de façon à construire une représentation acceptable des textes à l'égard de l'objectif à atteindre. Dans cette dernière approche, sur lequel se base la textométrie, les observables linguistiques pertinentes sont construites par la connaissance de leur comportement dans un contexte global qui préside la constitution du corpus, et fait appel à l'activité interprétative de l'analyste (Rastier et Pincemin, 1999). Il n'y a pas des *a priori* permettant d'associer certaines catégories grammaticales à des phénomènes : c'est l'interprétation qui identifie les observables linguistiques pertinentes pour décrire ces phénomènes, en lien avec les genres et les corpus.

> Ce serait encore à rapprocher de l'opposition entre stratégie de conquête et stratégie d'apprivoisement, que nous avions proposée comme lecture des pratiques en linguistique de corpus (…). La stratégie de conquête consiste à aborder le corpus à travers une grille d'analyse (fondée par la théorie). On ne retient du corpus que ce qu'on y trouve comme éléments correspondants à la grille d'analyse. Une certaine part du corpus échappe donc à l'analyse, et l'effort est mis pour affiner et enrichir le modèle, pour obtenir des représentations de plus en plus complètes. On vise à capter, à enrégimenter, les données selon les vues du modèle théorique. La stratégie d'apprivoisement part d'une représentation grossière, mais couvrant à sa manière l'ensemble du corpus. On la fait ensuite évoluer en l'affinant progressivement, à partir des régularités observées dans le corpus. Le modèle qui se dégage est alors par construction en affinité avec le corpus (Pincemin, 1999b, p. 321).

D'un point de vue méthodologique, la perspective que Rastier et Pincemin (1999) préconisent est la pratique d'une analyse globale et structurée des corpus, sans partir des catégories *a priori*. Le programme méthodologique de la sémantique interprétative, que nous allons présenter dans le prochain chapitre, se développe autour d'outils d'analyse statistique textuelle pour l'interprétation de grands corpus, en préconisant des démarches spécifiques permettant de découvrir dans les textes les données textuelles qui peuvent être pertinentes pour la fouille d'opinions. Le fait que les outils technologiques de la textométrique ont été élaborées pour répondre des questions théoriques propres à la sémantique interprétative, nous a motivés à choisir ce cadre théorique.

Nous voulons expliquer ce cadre et démontrer la justesse de la démarche textométrique pour répondre à notre question de recherche.

Chapitre 2. Cadre théorique

1. Introduction

L'objectif de notre recherche est de *systématiser et de valider une démarche méthodologique de fouille d'opinions basée sur la textométrie pour l'identification d'articles véhiculant des opinions divergentes dans une controverse et d'explorer, par une analyse de la progression chronologique du vocabulaire du corpus, l'applicabilité des résultats observés dans le développement des SRAP.* Pour atteindre cet objectif, nous proposons la réalisation de deux étapes distinctes.

La première étape vise à formaliser la démarche. Elle consiste à : 1) identifier et sélectionner des critères textuels discriminants et interprétables du point de vue thématique, dialectique et dialogique [17]et 2) vérifier si l'utilisation de l'ensemble des critères sélectionnés est efficace pour prédire la classe d'un article selon le type d'opinion qu'il défend.

La seconde étape de notre recherche vise à examiner l'applicabilité des résultats obtenus dans un contexte de recommandation. Nous avons constitué un corpus d'articles d'opinion provenant d'une controverse médiatique qui a duré 9 mois et nous allons vérifier à quel moment il est possible d'élaborer les critères textuels les plus performants pour la tâche de classification automatique.

Les concepts théoriques que nous allons présenter dans ce chapitre concernent le premier objectif de notre recherche et vise à expliciter les éléments sur lesquels repose la *démarche d'identification et de sélection de critères textuels*. Cette démarche s'inspire en grande partie des travaux réalisés par Valette (2004) et Eensoo et Valette (2012, 2014a, 2014b, 2015) et a été brièvement introduite dans la partie sur la justification empirique et théorique de la recherche (p.6). Elle préconise l'utilisation d'un ensemble particulier de calculs statistiques relevant du champ de la textométrie pour effectuer une étude préalable du corpus visant à repérer les critères textuels linguistiquement interprétables et statistiquement robustes pour la classification automatique de textes d'opinion.

[17] Ces concepts ont été introduits dans la page 6 et seront expliqués plus en détails dans cette partie.

Une clarification importante sur la sémantique interprétative s'impose : elle n'est pas une théorie sémantique formelle, établissant des règles ou des lois relatives au fonctionnement de la langue. Son application, qui s'impose peu à peu dans les tâches de classification de textes, envisage principalement l'exploitation de certains concepts de la théorie, mais sans prétendre à établir un modèle de la langue qui serait transposable en algorithme (Pincemin, 2012a). Les démarches méthodologiques sont suggérées pour guider trois choix principaux : 1) la constitution du corpus ; 2) le choix d'outils et de traitements statistiques ; et 3) la description des observables. À propos de ce dernier, Rastier (2015) affirme que les concepts descriptifs doivent être choisis en tentant compte de l'application :

> La SI [Sémantique interprétative] se voulant applicable, sa méthodologie entend concilier trois exigences : les principes de choix de corpus, la définition des hypothèses, le choix des concepts descriptifs (certaines distinctions pouvant être neutralisées en fonction des applications). (Rastier, 2015, p. 10).

Dans ce chapitre, nous voulons expliquer le cadre théorique de la sémantique interprétative (Rastier, 1987, 1989, 2001a, 2011 ; Rastier et coll., 1994) et associer aux concepts et principes énoncés les trois démarches méthodologiques mentionnées ci-haut : 1) la constitution du corpus ; 2) le choix d'outils et de traitements statistiques ; 3) la description des observables. Nous nous limiterons à exposer les notions théoriques les plus pertinentes pour notre recherche et qui ont été déjà exploitées dans d'autres travaux.

Nous divisons l'explication des concepts théoriques de la SI en trois parties. Dans la première, nous expliquons ce qu'est la textométrie et nous présentons un panorama des principales applications, incluant son exploitation dans le contexte de la fouille d'opinions. Dans la deuxième, nous exposons les concepts fondamentaux de la théorie (notamment la situation de communication, le discours, le genre, le texte et l'étude du sens) et leur rapport avec les démarches méthodologiques *1) constitution du corpus* et *2) choix d'outils* et de traitement statistique. Dans la troisième, nous exposons les concepts qui orientent la démarche méthodologique *3) description d'observables*, plus spécifiquement relative à la démarche de sélection et de catégorisation de critères textuels qui sera adoptée dans notre recherche. Cette description vise à détailler les stratégies de lecture interprétatives du corpus que les calculs

textométriques rendent possibles, ainsi que la qualification sémantique des critères qui émergent des calculs.

2. Textométrie : une introduction

La textométrique, aussi connue comme statistique textuelle, lexicométrie ou logométrie est une approche informatique destinée à l'analyse de corpus numériques et qui articule des calculs statistiques et des procédures d'interprétation des textes (Pincemin, 2012b). À la différence de l'ancienne lexicométrie, la textométrie peut porter sur des dimensions non seulement lexicales, mais aussi grammaticales, phonétiques ou prosodiques (Pincemin, 2012b). Les calculs textométriques (calcul des spécificités [Lafon, 1980], calcul de collocations (aussi connu comme n-grammes) et calcul de cooccurrences [Lafon, 1981], analyse factorielle des correspondances) mettent en œuvre des principes différentiels, en opérant dans une modélisation du corpus de manière contextuelle et contrastive : les corpus peuvent être partitionnés en fonction de différentes unités de contexte (textes, sous-corpus, fenêtres concurrentielles), ce qui permet une caractérisation des convergences et des différences statistiques d'un texte à l'autre ou d'un sous-corpus à l'autre, en fonction des distributions des mots présents.

Différemment de la linguistique de corpus, la textométrie n'est pas utilisée dans un contexte de description de la langue, mais dans d'autres domaines des sciences humaines pour l'étude comparative de corpus numériques (Pincemin, 2012b). Elle a été notamment appliquée dans l'étude comparative des genres (Bourion, 2001; Brunet, 2009; Malrieu et Rastier, 2001) et l'étude thématique (Mayaffre, 2008). Elle peut viser l'application informatique en conjonction avec la validation d'hypothèses théoriques, comme dans ce cas des travaux de Valette (2004) et Eensoo et Valette (2012, 2014a, 2014b, 2015), qui visent à déterminer les possibilités d'amélioration de la classification de textes véhiculant des opinions au moyen d'une caractérisation sémantique de corpus.

L'analyse textométrique des corpus et donc indépendante de la constitution de ressources linguistiques comme les dictionnaires et les ontologies. Elle permet ainsi de mettre en évidence les différences entre le vocabulaire employé dans les unités de contextes (par exemple, les sous-corpus), rendant compte des différentes lexicalisations d'un thème dans un

corpus, dans une partie de ce dernier, ou dans différents genres textuels. Sur le plan de la description des corpus numériques, la textométrie offre des solutions à des problèmes communs auxquels se heurte la majorité des traitements basés sur le TAL, comme la résolution d'ambiguïtés, car la description des phénomènes linguistiques observés est toujours contextualisée et effectuée en fonction des partitions constituées. Par exemple, la caractérisation d'un mot ne part pas de la forme graphique isolée et en rapport à une description référentialiste (par exemple, l'attribution d'une valeur positive à un mot par un dictionnaire de sentiments), mais elle part de la comparaison entre les distributions observées dans les unités de contextes constitués.

Quant aux outils de la textométrie, ils ont été développés en tenant compte des problématiques interprétatives, avec une importance donnée aux fonctionnalités de retour au texte, connu comme concordancier. Les principaux logiciels textométriques comme *Lexico3* et *TXM* accordent une place centrale à la fonction de « retour au texte », afin d'outiller l'analyste dans la tâche d'interprétation (Pincemin, 2012a). Cette fonction permet de cliquer sur un mot quelconque et de repérer tous les contextes où il apparaît, d'observer directement ses voisinages, ses successions et ses localisations. Dans ce sens, les logiciels textométriques offrent la possibilité de pratiquer l'analyse de corpus instrumentée (Pincemin, 2012b).

Dans le cadre des expérimentations menées par Valette (2004) et Eensoo et Valette (2012, 2014a, 2014b, 2015) pour la fouille d'opinions, la textométrie est préconisée comme démarche méthodologique de sélection de critères textuels pour la classification. Elle se place sur un cadre théorique linguistique particulier, qui est la sémantique interprétative, et vise à valider l'hypothèse que la qualification sémantique contrastive de critères provenant des différents niveaux de description des textes des sous-corpus (thématique, organisation argumentative et temporelle, représentation énonciative et modalisation) permet d'obtenir des critères textuels plus robustes pour la tâche de classification d'articles d'opinion que les traits discriminants trouvés par d'autres méthodes statistiques et symboliques (Eensoo et Valette, 2014b). La démarche de sélection de critères comprend des étapes clés, qui seront expliquées avec plus de détails dans notre méthodologie. Étapes que nous résumons ici cependant, afin de faciliter la lecture de notre exposé et comprendre le rapport entre les concepts qui seront

exposés dans ce chapitre et les démarches méthodologiques proposées pour la sélection de critères pour la fouille d'opinions.

1. Constitution de sous-corpus représentant chacun le type d'opinion à caractériser.

2. Application de calculs textométriques permettant de relever les contrastes entre les sous-corpus constitués.

3. Analyse des résultats visant sa qualification sémantique et la sélection des critères textuels.

3. Concepts structurels de la sémantique interprétative

Dans cette section, nous exposons ce que nous désignons ici de concepts structurels de la sémantique interprétative. Nous nous limitons à présenter le schéma de communication chez Rastier (1989), les concepts de discours, de genre et de texte, ainsi que les principes d'analyse du sens préconisés par la théorie. Cette section se clôt avec des exemples concrets d'application, ainsi que des explications sur les affinités entre ces concepts et principes théoriques et les démarches textométriques en ce qui a trait à : 1) la constitution du corpus et 2) le choix d'outils et de traitement statistique.

Avant de continuer, nous voulons donner quelques précisions sur les termes employés dans notre texte. La sémantique interprétative distingue le morphème, le mot et la lexie et considère que l'étendue du signe linguistique va du morphème au texte. L'unité d'analyse supérieure au morphème est la lexie, que Rastier définit comme « groupement stable de morphèmes constituant une unité fonctionnelle » (Rastier et coll., 1994, p. 223). À la différence du mot, qui se caractérise par un « groupement de morphèmes complètement intégré » (Rastier et coll., 1994, p. 223), la lexie peut manifester une intégration sémantique formelle composée de plusieurs mots (par exemple « rez-de-chaussée ») ou non formelle (par exemple « rendre compte », « tour Eiffel », « dès lors que », « guerre froide ») (Hébert, 2001). Pour ne pas alourdir notre exposé, nous allons utiliser le terme « mot » comme synonyme de lexie, tout en étant conscients du statut complexe de cette notion dans la théorie. Ainsi, nous employons le terme « mot » pour nous référer à la lexie formelle ou non formelle. De plus, nous distinguons, pour certaines explications, les « mots lexicaux » des « mots grammaticaux », tels que proposés par Polguère (2008) : les mots lexicaux regroupent les unités appartenant aux classes lexicales ouvertes (verbes, noms, adjectifs et adverbes) et les mots grammaticaux, ceux qui appartiennent aux classes lexicales fermées (pronoms, déterminants, conjonctions et prépositions).

Nous privilégions le terme « donnée textuelle » ou « données textuelles » au pluriel pour parler des données extraites de la chaîne textuelle, autant pour les lexies simples que pour celles composées de plusieurs mots. Les données textuelles citées sont mises entre guillemets simples. Par contre, le terme critère textuel est réservé à la donnée textuelle sélectionnée après

la démarche d'analyse textométrique du corpus. Nous convenons également de citer les critères textuels entre guillemets simples.

D'autres conventions d'écriture adoptées dans cette partie de notre exposé et dans d'autres chapitres se trouvent dans le Glossaire, vers lequel nous renvoyons le lecteur.

3.1 Le schéma de communication chez Rastier (1989)

Rastier (1989) postule que le sens n'est pas immanent au texte comme simple message, mais à la situation de communication, qui implique un échange entre l'émetteur et le récepteur à l'intérieur d'un système de normes sociales, instituées par la langue, par les genres et par les pratiques sociales. Ainsi, l'auteur explique que son schéma (figure 5 ci-dessous) n'est pas une représentation archétype de la communication entre l'émetteur et le récepteur, dans le sens de la transmission du message par un canal physique, mais plutôt un inventaire des conditions de la communication : celles-ci incluent l'émetteur et le récepteur, le texte et les systèmes de normes, qui se superposent aux deux premières instances.

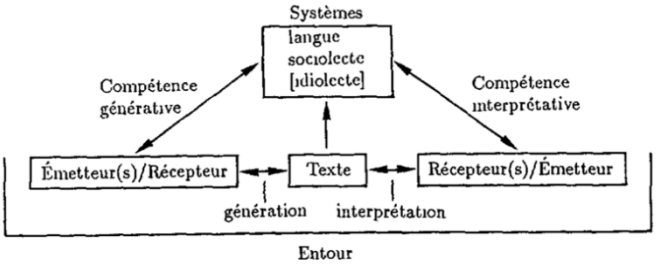


Figure 5. Schéma des conditions de la communication (Rastier, 1989, p. 47)

Dans le schéma de communication de Rastier (1989), l'émetteur et le récepteur ne sont pas uniquement des personnes. D'une part, Rastier englobe dans ces deux notions l'énonciation proprement dite pour se référer aux individus à l'origine et à la destination de la communication (des personnes, des groupes et même des systèmes informatiques) ; d'autre part, il inclut aussi dans ces deux notions l'énonciation représentée, entendue comme l'image que l'émetteur fait de soi et de son récepteur et qui est repérable dans les textes. Par cette distinction, l'auteur veut souligner que la représentation énonciative n'est pas nécessairement liée à la personne à l'origine du processus. Par exemple, le « je » d'un roman renvoie à son narrateur et pas à son auteur.

En tant que personnes réelles, l'émetteur et le récepteur sont définis en termes de compétences, le premier ayant une compétence générative et le dernier, une compétence interprétative. Les relations qu'ils entretiennent avec le texte ne sont pas unilatérales. Le texte agit aussi sur son auteur (l'émetteur) parce que ce dernier l'interprète en même temps qu'il le produit. De même, le texte agit sur le récepteur lorsque celui-ci anticipe son sens, le réinterprète ou le met en rapport avec d'autres textes précédents. C'est pour cette raison que dans le schéma, l'émetteur occupe également la place d'un récepteur et vice-versa. Ils sont placés dans un environnement linguistique plus large (l'entour) qui comprend l'ensemble des productions textuelles dans une culture, dans une société et dans une langue.

L'émetteur et le récepteur sont définis en fonction des places qu'ils occupent dans le processus de communication et les rapports qu'ils entretiennent d'une part avec le texte, et d'autre part avec les « systèmes ». Différemment du modèle de Charaudeau (1994) présenté dans le chapitre 1 (p. 100), le texte n'est pas un sous-produit de l'activité de communication. Il est au centre de cette activité puisqu'il est le seul objet empirique permettant d'en attester[18]. Dans la perspective de la SI, le texte est défini comme « une suite linguistique empirique attestée, produite dans une pratique sociale déterminée, et fixée sur un support quelconque » (Rastier et coll., 1994, p. 168).

Ainsi, si dans le modèle communicationnel de Charaudeau (1994), la communication entre l'émetteur et le récepteur se passe à un niveau abstrait de médiation subjective, d'un « monde à signifier » vers « un monde interprété », chez Rastier (1989), elle se réalise par la médiation du texte et de son rapport à une troisième instance, celle des systèmes : le dialecte (ou langue fonctionnelle), le sociolecte et l'idiolecte. Ceux-ci sont définis à la fois comme des systèmes sémiotiques et des systèmes de normes, car ils sont chargés de l'organisation des signes de communication entre les individus ou entre les collectivités, instituant différents registres de l'usage de la langue. Dans le modèle, ces systèmes sont connectés à l'émetteur et au récepteur par un rapport de compétence : tant l'émetteur que le récepteur doivent avoir les

[18] Par ailleurs, cette notion de texte n'exclut pas les échanges oraux, puisque ceux-ci peuvent être convertis également dans des objets d'analyse empirique de l'activité communicative.

compétences relatives aux systèmes qui sont exigées par la situation de communication. Les systèmes se définissent ainsi :

1. Le dialecte ou langue fonctionnelle : aussi appelé système fonctionnel de la langue, il établit les règles d'une langue parlée, à l'exemple des grammaires.

2. Le sociolecte : détermine l'usage d'une langue en rapport avec une pratique sociale donnée. Il se réfère aussi aux compétences linguistiques que les personnes acquièrent au sein d'une pratique (par exemple, le vocabulaire spécialisé d'un domaine) et à la maîtrise des genres textuels qui sont propres à cette pratique. Le sociolecte structure le lexique des domaines sémantiques relatifs aux diverses pratiques sociales.

3. L'idiolecte : plus présent dans le genre littéraire, il est caractérisé par l'usage d'une langue propre à un énonciateur, parfois transgressant le sociolecte et les règles d'un genre, voire d'une langue (Rastier et coll., 1994, p. 222). L'idiolecte a trait aussi au style d'un auteur particulier.

Les systèmes se superposent aux autres instances du schéma de Rastier (1989). En distinguant cette composante, l'auteur veut souligner que la langue « n'est jamais le seul système sémiotique à l'œuvre dans une suite linguistique » (p.37) et que d'autres systèmes, en particulier le sociolecte (qui comprend les genres et les pratiques associées) ont une influence sur la génération et l'interprétation de textes. Ces systèmes ne sont pas des réalités extralinguistiques, car les genres et les discours incarnent pour ainsi dire les données externes provenant des pratiques sociales en imposant un espace de contraintes linguistiques sur la génération et l'interprétation de textes.

Le schéma de communication présenté ici formule une des critiques principales de la sémantique interprétative par rapport à la notion du texte comme le « reflet » de représentations sociales et de situations objectives, tel que préconisé par l'AD : la sémantique interprétative met en cause la nécessité d'une composante extralinguistique comme nécessaire à l'interprétation, par exemple, la position sociale des sujets impliqués dans l'interlocution. Elle prône une théorie sémantique qui traite le contexte extralinguistique et linguistique

comme un seul objet théorique. La réunification de ces deux aspects tient à la centralité du texte dans le cadre général de la théorie ainsi qu'à l'importance accordée aux systèmes, plus particulièrement au sociolecte et au genre textuel, ce dernier étant théorisé dans une perspective sémiotique comme « l'empreinte significative et caractéristique d'une situation socioculturelle et du rapport au texte qu'elle prévoit » (Rastier et Pincemin, 1999, p. 89).

La SI propose des définitions pour le discours et les genres, ainsi que les pratiques sociales. Ces définitions sont expliquées dans les sections suivantes.

3.2 Le discours dans la perspective de la sémantique interprétative

Dans la perspective de la SI, le discours n'est pas assimilé à un mode d'organisation textuel lié à une intention, à l'exemple de la conception de Charaudeau (1994). Il est conçu dans son rapport au sociolecte (Rastier, 1989, 2008) et assimilé à l'utilisation de la langue dans les domaines d'activité humaine. Dans ce sens, le discours est une composante sémiotique qui a trait aux pratiques sociales, lesquelles sont reliées à la division du travail dans les sociétés (par exemple le discours politique, religieux, journalistique, etc.). Selon l'auteur, le discours renvoie à une « strate sémiotique d'une pratique sociale » (Rastier, 2008, p.15) constituée de signes (symboles, icônes et signaux) qui y sont échangés ou mis en jeu et qu'instaurent, à l'intérieur des pratiques sociales, les domaines sémantiques : les pratiques fixent les codifications linguistiques, comme la lexicalisation d'un domaine, et imposent l'acquisition de compétences discursives de la part des individus. Les indicateurs lexicographiques des dictionnaires (comme *agric*. pour agriculture) constituent selon l'auteur une « typologie empirique des domaines » (Rastier, 1989, p.39) et également des discours.

Dans la perspective de la SI, les textes se répartissent dans les discours, en fonction des pratiques sociales dont ils sont issus. Chaque discours se subdivise en plusieurs genres, qui en tant qu'objets décrits par la linguistique, font le lien entre les textes et la situation de communication. « Chaque texte procède d'un genre, et chaque genre est relatif à un discours (politique, religieux, scientifique, littéraire, etc.) » (Rastier, 1996b, p.16).

> 1. Les textes attestés dans des conditions réelles de communication constituent l'objet empirique de la linguistique. 2. Les textes sont produits et interprétés au sein

> de pratiques sociales. 3. À chaque type de pratique sociale correspond un type de discours (ex. politique, technique, littéraire). 4. Chaque discours se subdivise en genres (ex. dans le discours médical : l'article scientifique, le résumé d'observation, la lettre au collègue). Tout texte procède d'un genre, et par là relève d'un discours et d'une pratique sociale (Rastier et coll., 1994, p. 4).

3.3 Le concept de genre pour la sémantique interprétative

Dans la perspective théorique de la sémantique interprétative, le genre est chargé de rétablir la relation entre la dimension extralinguistique et le texte dans la mesure il constitue l'élément de liaison entre le texte et la pratique sociale. À cet égard, Rastier et Pincemin (1999) caractérisent le genre comme « le répondant, en tant que pôle intrinsèque du texte, des circonstances de rédaction et de lecture, qui font partie des pôles extrinsèques du texte » (p.89). Dans cette perspective, il n'est pas simplement un principe d'organisation d'œuvres, mais il constitue un modèle intériorisé par les émetteurs et les récepteurs au sein de leurs pratiques, orientant autant la production que l'interprétation des textes.

Pour Rastier (2001a), le genre offre un point de vue global sur les textes et contraint la portée sémantique des mots en présence. Le lexique, l'acceptabilité d'un énoncé et l'emploi des règles syntaxiques constituent des exemples d'éléments contraints par les genres (Rastier, 2001a, p. 230). Rastier (1989) propose une définition du genre comme un ensemble de prescriptions pour la production de textes, lesquelles sont projetées par les récepteurs lors de l'interprétation :

> (…) un programme de prescriptions positives ou négatives, et de licences qui règlent aussi bien la génération d'un texte que son interprétation ; elles ne relèvent pas du système fonctionnel de la langue, mais d'autres normes sociales (Rastier, 1989 p. 36).

Cette définition pose que les genres, en tant que « programme de prescriptions positives et négatives », interdisent ou encouragent certains procédés linguistiques lors de l'élaboration de textes, de même qu'ils interdisent ou encouragent certaines interprétations dans l'instance de réception. La relation entre ce qui est autorisé ou interdit par le genre est plus au moins contraignante, dépendamment des domaines d'activité. Ainsi, un article

scientifique impose des règles d'écriture plus rigoureuses qu'un roman littéraire au niveau de l'utilisation de la langue par exemple, ou de la structuration des chapitres, car les pratiques dont elles proviennent répondent à des objectifs différents (Rinck, 2006). Les genres contraignent également les relations sémantiques sur le plan de l'organisation interne des textes. Par exemple, l'introduction d'un article scientifique est généralement obligatoire et doit représenter un résumé des principaux thèmes abordés, au même titre qu'un chapitre ou une section sont utilisés pour cerner un sujet plus au moins cohésif. En établissant ce rapport entre l'organisation interne de textes et les contenus, les genres gèrent les attentes interprétatives, ils autorisent certaines relations sémantiques, tout en en interdisant d'autres, et ils délimitent certaines objectivités à l'égard du sens qu'un texte peut porter.

Par la caractérisation des textes et des genres, la sémantique interprétative veut souligner que le signe n'est pas l'objet de l'interprétation, puisque l'isoler, c'est le couper de ses conditions d'interprétation. Ainsi, en tant que programme méthodologique d'analyse du sens, elle se penche sur les conditions d'interprétation des signes fournies par une contrainte globale exercée par les discours et les genres.

> Comme la véritable unité sémiotique des langues est le morphème, le mot est déjà un syntagme, c'est-à-dire une unité de « discours ». Toutefois, en sémantique référentielle notamment, on considère les mots comme des unités élémentaires, et l'on discute sur des mots isolés pour souligner leur polysémie ; à cela, nous opposons que les mots restent indissociables des textes dont ils sont tirés, car les textes demeurent les seuls *objets empiriques* de la linguistique (Rastier, 2011, p. 24)

Rastier énonce le principe herméneutique de la détermination du global sur le local comme la pierre angulaire de sa théorie sémantique : le sens d'un texte subit des déterminations provenant d'un contexte plus large et global à l'origine de sa production. Ce contexte global est formé par les genres textuels et par les pratiques sociales correspondantes, autant que par les discours qui encadrent ces pratiques. Dans ce contexte, l'étude du sens d'un texte est tributaire de l'activité interprétative, laquelle ne peut pas être dissociée du corpus, des genres et des discours. Le programme d'analyse du sens préconisé par la sémantique interprétative ne cherche pas un sens caché ou l'intentionnalité des émetteurs, mais cherche à

expliciter les conditions d'interprétation permettant d'établir le sémantisme des mots. Cette question sera expliquée dans la section suivante.

3.4 L'analyse du sens comme parcours interprétatif

La sémantique interprétative stipule que le sens est construit par l'activité interprétative et c'est sur le phénomène de l'interprétation et des opérations interprétatives qu'elle se penche pour fonder l'étude du sens textuel. Dans la perspective de la SI, l'activité interprétative n'est pas forcément linéaire, dans le sens où l'on attribue mécaniquement un signifié à un signifiant au cours de la lecture. En fonction du genre, le lecteur aborde le texte avec une certaine attente. Au cours de la lecture, il formule des présomptions sur le sens, dans la mesure où il identifie des rapports entre les mots selon les divers contextes dans lesquels ceux-ci prennent leur sens : leurs voisinages, les phrases, le texte jusqu'aux genres et les pratiques auxquelles ils sont rattachés. Là repose la perspective méthodologique de la SI, qui préconise la description des parcours interprétatifs, comme le repérage de relations sémantiques qui peuvent déterminer une interprétation plausible d'un texte, parmi d'autres interprétations possibles.

Il faut néanmoins souligner que l'objectivité du sens n'est pas à l'image d'un objet des sciences pures (Rastier, 1989). À propos de l'objectivité du sens, Rastier (1989) fait remarquer que pour les sciences humaines, ainsi que pour les sciences du langage, il faut relativiser le postulat d'une objectivité ou d'une subjectivité absolue (Rastier, 1989, p. 16). La perspective préconisée par Rastier (1989) est d'envisager l'objectivation du sens textuel comme un travail de description de phénomènes linguistiques qui confirment et démontrent, par des « conjectures rationnelles » (Rastier, 1996a, p. 20), une lecture plausible d'un texte ou d'un ensemble de textes, parmi d'autres lectures possibles, en fonction des conditions de la communication, des pratiques et des genres. Selon Rastier (1989), « il y a une objectivité du sens dans la mesure où le texte contraint — sans pourtant déterminer entièrement — les lectures plausibles qu'on peut en faire » (Rastier, 1989, p.15).

> L'unité empirique du texte ne préjuge pas de la fixité de sa signification, de l'intention de son ou ses auteurs, de ses références non ostensives, ni de l'interprétation qui en est donnée par ses destinataires ; au sein de la sémiotique des

> cultures, il revient à la linguistique, considérée comme science des textes, de caractériser ces quatre pôles par des conjectures rationnelles sinon formelles. L'analyse du sens permet des conjectures sur le rapport du texte au monde physique (tel qu'il trouve des corrélats dans ses références non ostensives), comme sur son rapport au monde des représentations (tel qu'il trouve ses corrélats dans les intentions de son auteur et dans les interprétations de ses destinataires). (Rastier, 1996a, p. 20)

Les parcours interprétatifs sont définis comme une « suite d'opérations permettant d'assigner un ou plusieurs sens à une suite linguistique » (Hébert, 2001). Ils sont le fruit de deux opérations interprétatives qui permettent d'établir des liens de similitudes et de différences entre les mots dans un texte, en fonction des contraintes exercées par les genres et les discours.

Afin de comprendre davantage les opérations interprétatives et les formes d'objectivation du sens textuel dans le contexte de la SI, nous exposons dans les sections suivantes les principes qui définissent le sens à l'intérieur de la théorie.

3.5 La question du sens pour la sémantique interprétative

La sémantique interprétative est une théorie qui s'intéresse à la question du sens linguistique comme étant distincte du sens conceptuel et du sens référentiel (Hébert, 2001). Elle se positionne comme une critique de l'objectivisme en sémantique et du positivisme en linguistique. L'objectivisme postule que le sens réside dans le rapport entre les propositions et l'état de choses, ainsi que les conditions qui font qu'un énoncé est vrai ou faux (sémantique vériconditionnelle). Le positivisme postule que le sens est dans le rapport entre les propositions et les représentations mentales (sémantique cognitive, mentaliste). Ces deux visions défendent une ontologie substantielle, c'est-à-dire que le sens est calqué sur la propriété essentielle des choses, soit l'objet réel lui-même, soit la représentation de ces objets dans l'esprit des individus (Rastier, 1994). La sémantique interprétative prône une démarche d'analyse différentielle du sens : le sens serait le produit de différences et de similitudes qui ressortent de l'interaction sémantique entre les signes et leurs signifiés au sein des textes et au sein des discours.

La sémantique interprétative se positionne dans la tradition saussurienne et reprend de celle-ci la notion de la valeur du signe linguistique pour l'étude de la signification. Selon la perspective de Saussure, la valeur du signe ne peut être appréhendée que par la différence que ce dernier entretient avec d'autres signes à l'intérieur du système de la langue et dans les situations d'usage de ce système. L'opération qui permet aux locuteurs de reconnaître la valeur du mot n'est pas d'ordre mécanique, n'est pas une association entre la manifestation physique de ce signe (son image acoustique) et son concept. En renouant avec ce fondement saussurien, la sémantique interprétative postule que « Rien ne préexiste à la détermination de la valeur par le système [linguistique] » (Rastier, 2011, p. 29). Ce qui fixe la valeur d'un signe est la relation entre les signifiés, à l'intérieur d'un contexte plus global qui peut inclure l'usage fonctionnel de la langue pour des fins de communication, mais aussi les textes, les genres et les discours.

> Dans son principe, la sémantique interprétative est un développement de la sémantique différentielle de tradition saussurienne. Dans cette problématique, le sens est fait de différences et le concept fondamental est celui de valeur. (i) la valeur est la véritable réalité des unités linguistiques. (ii) Elle est déterminée par la position des unités dans le système (donc par les différences). (iii) Rien ne préexiste à la détermination de la valeur par le système. 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. (…). 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. 29-30).

Pour la sémantique interprétative, les signes sont composés de traits sémantiques appelés sèmes. Ce sont les relations d'identité et d'altérité observées entre les sèmes d'un signe à l'autre lors de l'activité interprétative qui détermine ce que Rastier et coll. (1994) appellent l'impression référentielle, la « représentation mentale contrainte par l'interprétation d'une suite linguistique » (p. 222). Ces traits sémantiques ne sont pas des abstractions d'idées ni d'objets existant dans le monde à propos desquels on pourrait dresser un inventaire exhaustif, puisqu'ils sont attestés seulement dans la confrontation avec d'autres signifiés au sein des textes et des discours. Par exemple, si le mot « rouge » dans le roman de Stendhal « Le Rouge et le Noir » représente métaphoriquement les concepts d'« armée » et que

le mot « noir » représente « église », ce n'est pas parce que ces signifiés sont hérités par défaut de ces couleurs, mais parce que dans le contexte du roman, ces traits sémantiques sont activés par l'activité interprétative, qui prend en compte la trame complexe d'autres signifiés présents et les références culturelles propres à la communauté linguistiExque (Hébert, 2001, p. 90-91). Les sèmes sont d'ordre métalinguistique, ils sont relatifs aux usages courants d'une langue, aux pratiques sociales, aux genres, aux textes et aux corpus et leur analyse résulte d'une validation humaine effectuée, par exemple, par un linguiste (Valette, 2010).

L'interaction sémantique qui peut se dégager entre les signes et les signifiés dans les textes est le fruit de l'activité interprétative et des opérations d'assimilation et de dissimilation. Dans la première, l'interprétant établit un lien sémantique entre les mots, en assimilant leurs rapports d'identité. Dans la deuxième, il identifie les contrastes et les différences existantes entre les mots en établissant entre eux une relation d'altérité. Sur le plan opérationnel, l'assimilation et la dissimilation effectuent l'activation ou l'inhibition de traits sémantiques appartenant aux mots, en fonction du contexte linguistique immédiat (syntagme, phrase, paragraphe), ainsi que des genres, discours et pratiques sociales.

Les genres et les discours fonctionnent toujours comme un point d'ancrage pour l'activité interprétative, fournissant un type de « validation » ultime du sens. Par exemple, l'opération d'assimilation permet d'établir entre les mots « chaise » et « fauteuil » le trait commun /pour s'asseoir/. L'opération de dissimilation permet de voir leur distinction par le trait sémantique /avec accoudoir/, présent seulement dans « fauteuil ». Les opérations interprétatives permettent d'attribuer la notion de confort à « fauteuil » et la notion d'inconfort à « chaise », dépendamment de la présence d'autres signifiés dans le texte qui contribuent à l'activation de ces traits.

Nous allons approfondir la notion de sèmes et expliquer davantage les opérations interprétatives dans la section suivante.

3.6 La théorie du sème

Dans le « Dictionnaire de Linguistique », le sème est défini comme l'unité sémantique minimale résultant de l'analyse des signifiés (Mounin, 1993). Il faut cependant éviter d'assimiler les sèmes à des atomes appartenant aux mots, car, pour la sémantique

interprétative, un mot n'est pas décomposable en sèmes. Cette décomposition constituerait une entreprise sans fin à cause même de la diversité de contextes où un mot peut apparaître (Pincemin, 1999a). Pour la sémantique interprétative, les sèmes sont des éléments de signification qui ressortent de la confrontation des signifiés de plusieurs mots, dans les différents paliers de contextualisation et dépendamment des genres et des discours (Pincemin, 1999a). Pour décrire le statut des sèmes en présence, la théorie se fonde sur les différences observées entre les mots en contexte et non par rapport à la langue en général :

> (…) L'utilisation du paradigme différentiel en langue consiste à décrire le contenu d'un mot par sa position dans le système fonctionnel de la langue, habituellement dans une perspective lexicographique, sans le lier à un contexte déterminé. En revanche, dans une perspective interprétative, il est indispensable de prendre en compte le contexte : ainsi les oppositions qui définissent les valeurs sont-elles décrites par rapport aux contextes d'interprétation dans lesquels elles prennent place. C'est le contexte qui détermine à la fois l'existence des oppositions (…). Une infinité de contextes potentiels permet ainsi de décrire une infinité de contenus, bien que de nombreuses régularités liées à des normes de genre (et parfois incluses dans la langue elle-même), et qui reflètent des propriétés du monde social ou physique, tendent à limiter le nombre de sèmes dans un corpus donné. (Rastier et coll., 1994, p. 85)

Valette (2010, p.25) illustre le concept de sème avec le mot « chien » en opposant la perspective lexicographique à la perspective de la sémantique interprétative. En prenant la première voie nous pouvons dire que le mot « chien » possède des traits sémantiques définitoires constitués par des propriétés physiques comme /mammifère/, /carnivore/, /canidé/ et /domestique/. Par contre, si nous demandons à un enfant de parler d'un chien, nous allons probablement retrouver des sèmes tels que /bête/, /poilue/, /qui a de grandes dents/, /qui aboie/, /qui mord/. L'auteur défend que dans une perspective interprétative, aucune de ces catégories de signifiés n'est meilleure ou plus vraie que l'autre, puisque le sens activé par le mot « chien » dépend du contexte de construction. La réalité décrite par ces exemples ne signifie pas que les mots sont dépossédés d'une référence, mais elle démontre que, considérés du point de vue de l'interprétation, les traits sémantiques ne sont pas uniquement déterminés par le système linguistique.

3.7 Opérations interprétatives

Comme nous l'avons mentionné, l'interaction sémantique qui peut se dégager des signes et des signifiés dans les textes est le fruit de deux opérations interprétatives désignées assimilation et dissimilation. Dans la première, l'interprétant[19] établit un lien sémantique entre les mots, en assimilant leurs rapports d'identité. Dans la deuxième, il identifie les contrastes et les différences existantes entre les mots en établissant une relation d'altérité. Sur le plan opérationnel, l'assimilation et la dissimilation effectuent l'activation ou l'inhibition de traits sémantiques (ou sèmes) appartenant aux mots, en fonction de leur usage en contexte, dépendamment des genres, des discours et des pratiques sociales.

Ainsi, l'assimilation d'un ensemble de mots comme « école », « étudiant », « rentrée scolaire » au sein d'un guide d'étudiant par exemple, active des traits de valeur typique, ou dénotative, qui sont prescrits par l'usage de ces mots dans le monde social. Rastier et coll. (1994) appellent ces traits sémantiques des sèmes inhérents. Un sème inhérent correspond généralement à la définition donnée par les dictionnaires de langue, puisque ces derniers définissent les mots en fonction des champs ou des domaines d'activité.

D'autres associations peuvent inhiber les traits inhérents de mots et activer des traits afférents (ou traits connotatifs) qui sont actualisés en contexte. Par exemple, le mot « éducation » dans un discours politique peut être interprété comme une valeur fondamentale, ou un droit lié à l'épanouissement de l'être, plutôt que dans le sens de l'instruction ou de l'institution.

Pour illustrer le concept que nous traitons ici, nous reprenons l'exemple donné par Hébert (2001, p. 90-91) de l'analyse sémique du titre du roman de Stendhal « Le rouge et le noir » (tableau 1). Ce titre reflète le conflit vécu par le personnage Julien Sorel, qui hésite entre la carrière militaire, représentée métaphoriquement par le rouge, et les ordres ecclésiastiques, représentés par le noir des habits. Le trait /couleur/, inhérent des mots

[19] L'interprétant est défini comme une « unité du contexte linguistique ou sémiotique permettant d'établir une relation sémique pertinente entre des unités reliées par un parcours interprétatif » (Hébert, 2001, p.212).

« rouge » et « noir » sont inhibés et la trame du roman produit l'afférence présente dans le titre : rouge pour l'armée et noir pour l'église.

Tableau 1. Analyse sémique du titre
« Le rouge et le Noir », Inspiré de Hébert (2001, p. 91)

	« rouge »	**« noir »**
Sèmes inhérents	/couleur/	/couleur/
Sèmes afférents	/armée/	/église/

3.8 Démarches méthodologiques : principes de choix du corpus et calculs statistiques de la textométrie

3.8.1 Sommaire des concepts structurels

La sémantique interprétative s'oppose à la notion de « compositionnalité du sens » et préconise le principe herméneutique de l'influence du global sur le local : le sens des mots dépend des régulations exercées par le contexte global, qui comprend le texte en soi dans son corpus d'accueil, son genre et son discours. Le discours n'est pas lié à la performance d'un acte de la part d'émetteurs dotés d'une intention, mais relié à une strate sémiotique de la réalité sociale, qui définit les pratiques associées à la vie professionnelle et qui dans ce sens, instaure des domaines sémantiques. Chaque champ de pratique instaure des genres spécifiques qui sont des actes typifiés à l'intérieur des pratiques et qui établissent un espace de contraintes tant pour la production que pour l'interprétation des textes. Le sens textuel est donc le résultat de ces surdéterminations : le rapport de signification entre le signifiant et le signifié n'est pas premièrement établi en fonction d'un référent externe, physique ou psychique, il est interdéfini en fonction d'autres mots en présence, à l'intérieur d'un ensemble de contraintes régulées par les textes, les genres et les discours.

Avec la notion de sème, la sémantique interprétative souligne que le sens est fait à partir de différences. Le lecteur aborde un texte avec une certaine présomption sur la signification des mots qu'ils contiennent, puisque les significations sont prescrites par la

pratique dans laquelle il s'insère. L'interprétation n'est pas un processus mécanique, elle est faite à partir d'un processus de construction qui stabilise le sens au cours de la lecture. Ce processus implique l'activation et l'inhibition de traits sémantiques, en fonction de l'assimilation et de la dissimilation qui peuvent être établies entre les mots, et considérant les différents contextes qui exercent une influence sur leur sens.

3.8.2 Implications des concepts structurels sur la constitution du corpus

La démarche méthodologique de la textométrie présente des compatibilités avec les concepts structurels exposés dans cette section. Une de ces compatibilités concerne les principes de rassemblement de textes dans le corpus pour les calculs textométriques. Rastier (2015) explique que pour être compatible avec la théorie, ce rassemblement doit refléter le principe de la détermination du global sur le local. À un niveau global, il faut connaître les pratiques, les discours et les genres auxquels les textes du corpus se rattachent, de façon à établir un point d'ancrage commun sur lequel les observables linguistiques doivent être interprétées.

> Un corpus est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages et rassemblés : (i) de manière théorique réflexive, en tenant compte des discours et des genres et (ii) de manière pratique en vue d'une gamme d'applications. (Rastier, 2011, p. 33-34).

Dans la démarche informatique de la textométrie, la transposition du principe herméneutique de l'influence du global sur le local se fait par la manière de constituer le corpus et par l'emploi de calculs statistiques (Valette, 2004 ; Pincemin, 2012a). Puisque « le sens d'un texte s'étudie en contrastant ses différentes parties (…) [et aussi] en le contrastant avec d'autres textes » (Rastier, 2008, p.20), le projet méthodologique de la sémantique interprétative va privilégier une approche comparative entre des corpus.

Rastier et Pincemin (1999) préconisent de faire d'une part la distinction entre le corpus de référence, c'est-à-dire le corpus constitué en tentant compte des objectifs de l'application, et qui doit être pris comme le contexte global d'analyse, et d'autre part les sous-corpus, constitués à partir du corpus de référence, à des fins comparatives. Le corpus de référence

rassemble les textes en fonction d'un critère générique et les sous-corpus sont composés de partitionnements du premier, qui reflètent une distinction métatextuelle (Rastier et Pincemin, 1999). Par exemple, un corpus de référence peut être constitué par l'ensemble des discours des chefs d'État et les sous-corpus, constitués des discours de chaque politicien, le but étant de comparer leurs différences et de faire ressortir leurs spécificités.

> L'analyste qui procède à des statistiques textuelles assume cela, à juste raison, en distinguant en pratique : (i) un corpus existant, correspondant aux textes accessibles dont il peut disposer (ii) un corpus de référence, constituant le contexte global de l'analyse, ayant le statut de référentiel représentatif, et par rapport auquel se calcule la valeur de paramètres (pondérations…) et se construit l'interprétation des résultats (iii) un corpus de travail [ou sous-corpus], ensemble des textes pour lesquels on veut obtenir une caractérisation (…) Ainsi, le corpus de référence, le corpus de travail (…) se distingue non pas tant par leur constitution (interne) que par leur rôle, leur position respective dans l'analyse : la première ancre une référence interprétative, le second définit le champ à explorer. (Rastier et Pincemin,1999, p. 84-85)

3.8.3 Implication des concepts structurels sur le choix d'outils et calculs statistiques

Puisque la signification d'un mot n'est pas donnée, mais construite relativement à un cadre de contraintes, les approches informatiques visées par la sémantique interprétative ne préconisent pas l'utilisation de ressources lexicales (comme les dictionnaires de sentiments) pour analyser les textes numériques. Quant à cet aspect, la textométrie s'assure d'être pleinement compatible avec la sémantique interprétative, puisque les calculs de la textométrie sont conçus et appliqués de manière à faire ressortir les contrastes et les différences entre les sous-corpus constitués, respectant le principe différentiel préconisé par la théorie, ainsi que la place accordée à l'investigation empirique :

> Enfin, la sémantique peut être conçue comme pleinement linguistique, sans dépendre d'une représentation du monde physique ou mental, d'une réalité externe. Certes, elle tourne ainsi ostensiblement le dos aux sémantiques référentielles, massivement dominantes dans les traitements automatiques des langues et en intelligence artificielle, via le recours aux ontologies. (Pincemin, 2012a, p. 75)

La textométrie est fondée sur des calculs principaux destinés à faire émerger les contrastes entre les sous-corpus constitués : 1) le calcul des spécificités et 2) le calcul de cooccurrence :

1. Le calcul des spécificités mesure la surreprésentation de certains mots dans un sous-corpus, en prenant en compte le nombre total d'occurrences dans le corpus de référence et le nombre total d'occurrences dans une partie du corpus (sous-corpus). Elle rend donc compte des données textuelles qui se trouvent surreprésentées ou au contraire de leur rareté, en mesurant l'écart entre la répartition aléatoire de ces mots et leur comportement effectif observé dans les sous-corpus.
2. Le calcul de cooccurrence mesure les affinités lexicales entre des couples de mots afin de calculer leur degré d'attraction, et rend accessible l'étude des associations. Différentes techniques sont préconisées dans le cadre de la textométrie, dont le calcul de pairs cooccurrents par des mesures d'écart réduit ou encore le calcul de cooccurrents autour de mots pôles, visant à repérer les mots qui sont en relation dans un contexte textuel plus restreint (par exemple, dans le même paragraphe, dans la même phrase ou dans une fenêtre avec un nombre déterminé de mots).

La particularité associée aux deux principaux calculs de la textométrie, c'est la possibilité qu'ils offrent de faire émerger des données textuelles provenant d'un niveau de contextualisation globale, relativement au corpus de référence, et d'un niveau de contextualisation locale, qui considère des zones contextuelles de localités plus restreintes, comme le voisinage des mots (Pincemin, 2012a).

Sur le plan global, le calcul des spécificités permet de repérer des données textuelles qui traversent les sous-corpus, car l'identification d'une donnée textuelle suremployée se fait en fonction de sa mise en rapport avec la fréquence totale dans le corpus de référence (Lafon, 1980). Ce dernier constitue selon Mayaffre (Mayaffre, 2008, p. 59) la « norme statistique globale » permettant d'effectuer les décomptes. En effet, comme explique Lafon (1980) le calcul des spécificités mesure l'écart entre une répartition aléatoire de mots, présente dans le corpus de référence et son comportement effectif dans une partie du corpus. Pincemin (2012b) explique que ce principe opératoire considère le principe théorique selon lequel l'usage de la langue dans un contexte déterminé (en l'occurrence, le sous-corpus), est réglé par des

contraintes linguistiques. En comparant la distribution des données textuelles dans le corpus de référence et les sous-corpus pour relever les contrastes, les calculs des spécificités de la textométrie mettent en œuvre le principe différentiel de la SI (Pincemin, 2012b). La pratique en textométrie démontre que le résultat de la démarche a l'avantage de révéler de nouvelles observables autrement imperceptibles, permettant d'établir des corrélations entre les spécificités rencontrées et les choix particuliers que font les émetteurs lors de l'élaboration du texte (Pincemin, 2012a ; Rastier, 2011).

Sur le plan local, le calcul de cooccurrence se réalise dans une région contextuelle plus réduite, au niveau de la phrase, du paragraphe ou d'une fenêtre de mots de taille variable, de façon à considérer tous les mots dans leurs contextes minimaux. L'étude de cooccurrences a été spécialement explorée pour l'analyse thématique de corpus (Pincemin, 2012b) et en rapport avec certains concepts théoriques de la SI que nous allons traiter dans les prochaines sections. Mayaffre a remarqué en 2008 l'affinité du calcul de cooccurrence, en définissant la cooccurrence comme le contexte minimal dans lequel un mot prend sa signification.

> Le contexte, linguistique comme statistique, est donc au palier supérieur du corpus. Au palier inférieur, nous voulons poser que le contexte minimal d'un terme est la cooccurrence. Nous considérons en effet, en corpus, que la forme minimale du contexte d'un terme, nécessaire à sa compréhension-interprétation, n'est pas le syntagme ou la phrase, mais la cooccurrence ; (…) qui présente l'avantage de se trouver accessible de manière systématique, étant entendu que nous saurions considérer, même avec un concordancier, un par un, tous les mots dans toutes leurs chaînes. (Mayaffre, 2008, p. 59-60)

3.8.4 Applications dans le contexte de la fouille d'opinions

Dans le cas de la démarche méthodologique proposée pour la fouille d'opinions, les études réalisées par Valette (2004) et Eensoo et Valette (2012, 2014a, 2014b, 2015) préconisent la constitution d'un corpus de référence composé d'articles d'opinions sur un thème précis, et la constitution des sous-corpus reflètent les différentes opinions véhiculées par les articles du corpus de référence. Les calculs textométriques sont proposés pour faire ressortir les données textuelles spécifiques de chaque sous-corpus qui pourraient servir comme critères textuels pour la fouille d'opinions. Le choix d'une donnée textuelle comme un critère

textuel est fait en aval des calculs textométriques et validé par l'analyse humaine, qui prend le soin de vérifier les contextes d'apparition des spécificités rencontrées. Cette analyse construit les critères textuels pertinents par un processus d'interprétation comprenant l'assimilation et la dissimilation de sèmes en fonction des régularités et des différences rencontrées dans les sous-corpus. Dans ce sens, les calculs textométriques permettent la qualification des critères qui seront utilisés pour représenter les articles du corpus dans le format vectoriel pour les algorithmes de fouille. Les critères textuels ainsi sélectionnés sont considérés comme des critères interprétables (Eensoo et Valette, 2012, 2014a, 2014b, 2015).

Dans Valette (2004), les critères textuels sélectionnés par la méthode de spécificités sont désignés globaux et ceux sélectionnés par la méthode de cooccurrences, locaux. La sélection de critères globaux et locaux présente une « solidarité d'échelle entre niveaux de complexité » (Rastier, 2006b, p.84), puisqu'elle rend accessible la sélection de critères provenant de différents paliers de complexité textuelle, soit le voisinage des mots, la phrase, le texte et le sous-corpus. Selon Valette (2004), les résultats expérimentaux de classification de textes d'opinion démontrent une compatibilité effective entre l'approche textométrique et le principe herméneutique de la SI selon lequel le global détermine le local : la combinaison de critères globaux et locaux démontre une performance supérieure et confirment la complémentarité des deux calculs.

La démarche interprétative présente un gain qualitatif dans le contexte général de la méthode de fouille d'opinions. À l'aide des concepts de la théorie (lesquels seront exposés dans la section sur les concepts descriptifs à la page 146), l'analyste cherche à rendre cohérente l'hétérogénéité apparente de ces observables, il interprète le corpus et les contrastes entre les sous-corpus et il justifie la sélection des critères textuels en fonction des interprétations et de l'hypothèse théorique qu'il souhaite valider. L'instrumentalisation technologique des logiciels de la textométrie lui est utile dans cette tâche, dans la mesure où ils offrent des outils permettant de repérer les contextes d'apparition de toutes les données textuelles du corpus (concordancier des applications *TXM* et *Lexico3*) (Pincemin, 2012a).

La comparaison entre sous-corpus par des calculs statistiques permet de repérer les particularités linguistiques qui supportent l'hypothèse interprétative ayant précédé la constitution des sous-corpus. C'est dans ce sens que l'approche méthodologique préconisée

par la SI parle de la description des parcours interprétatifs comme du repérage d'un ensemble cohérent d'observables linguistiques qui confirment une hypothèse interprétative en même temps qu'elle décrit objectivement les « cheminements » d'interprétation entrepris par l'analyste. Sur le plan méthodologique, la démarche textométrique appuyée sur les concepts de la SI s'assume comme éminemment empirique.

Rastier (2015) préconise que les concepts descriptifs de la SI soient choisis en tenant compte de l'application souhaitée. Les sections suivantes présenteront les concepts sur lesquels se base la sélection de critères textuels pour la fouille d'opinions ainsi que leur catégorisation (thématique, dialectique et dialogique), qui s'appuie sur la description sémantique des résultats du calcul.

4. Concepts descriptifs de la sémantique interprétative

La SI propose un ensemble de concepts permettant de décrire le sens des textes. Le sens est décrit en fonction de classes sémantiques qui peuvent être dégagées à partir de l'analyse sémique. Dans cette section, nous voulons expliquer la typologie des sèmes et les opérations permettant d'entamer cette analyse. Nous expliquerons également le concept de composante sémantique. Les composantes sémantiques sont également analysables, dans le contexte de la théorie, par la description des classes sémantiques. Dans le cas spécifique de la démarche méthodologique de fouille d'opinions préconisée par notre étude, ces concepts permettent de qualifier sémantiquement les critères textuels qui sont identifiés par le truchement des calculs et de les associer aux différentes composantes sémantiques qui sont définies par la théorie (thématique, dialectique et dialogique). Au fur et à mesure de l'exposé sur les composantes sémantiques, nous fournissons des exemples d'application des concepts dans la démarche d'élaboration de critères textométriques pour la fouille d'opinions tels que préconisés par les travaux de Valette (2004) et Eensoo et Valette (2012, 2014a, 2014b, 2015). Cette section se clôt avec un résumé sur l'utilisation des concepts descriptifs, en lien avec la démarche méthodologique qui sera adoptée.

4.1 Typologie des sèmes et des classes sémantiques

Les sèmes actualisés en contexte par des opérations interprétatives sont de deux types : 1) les sèmes génériques, communs à différents mots et 2) les sèmes spécifiques, qui mettent les mots en contraste. Par exemple, les mots « fourchette », « cuillère » et « couteau » ont en commun le sème générique /couvert/ et se distinguent respectivement par les sèmes spécifiques /pour prendre/, /pour contenir/ et /pour couper/. Un sème spécifique peut être un sème générique dans un autre groupement. Par exemple si les mots « cuillère », « fourchette », « couteau » et « ciseau » sont présents, les trois premiers peuvent être regroupés par le sème générique /couvert/. Le mot « couteau » a comme sème spécifique /pour couper/ par le premier regroupement, mais partage ce sème avec « ciseau ». Ainsi « couteau » et « ciseau » peuvent être regroupé dans le sème générique /pour couper/. De ce fait, le caractère générique ou spécifique d'un sème est relatif, le spécifique n'ayant du sens que par rapport au générique.

Les rapports d'équivalence et d'opposition dégagés des sèmes imposent une organisation au lexique présent dans un texte : les sèmes génériques regroupent les mots dans des classes. Les sèmes spécifiques imposent les différences entre les classes (Pincemin, 1999a).

Les classes formées par le regroupement de sèmes génériques sont au nombre de trois :

1. Taxème : classe sémantique à l'intérieur de laquelle les éléments ont un rôle équivalent. Par exemple, les mots « cuillère », « fourchette » et « couteau », partagent le sème /couvert/ et s'inscrivent dans le taxème //couvert//. Hébert (2001) explique que la composition de taxèmes reflète des situations de choix. Nous pourrions aussi opposer « autobus » et « métro » d'une part, et « autocar » et « train » d'autre part, sur la base du critère transports intra-urbains et extra-urbains, en les organisant ainsi dans les taxèmes //intra-urbains// et //extra-urbains//, mais nous pourrions également regrouper train et métro dans le taxème //sur rails//, ainsi qu'autobus et autocar dans le taxème //sur roues//.

2. Domaine : un domaine est une classe supérieure au taxème. Rastier le définit comme un « groupe de taxèmes liés à une pratique sociale » (Rastier et coll., 1994, p.222). Il explique que « Dans un domaine déterminé, il n'existe généralement pas de polysémie » (p.222), car les mots sont employés de façon assez homogène dans un domaine. Par exemple, « canapé » peut appartenir aux domaines //alimentation// ou //ameublement//. La composition et l'inventaire des domaines relèvent des normes sociales et des discours. Hébert (2001) donne l'exemple des mots « hostie » et « ciboire », qui appartiennent au domaine //religion// et non //alimentation//.

3. Dimension : la dimension est définie comme une classe sémantique de généralité supérieure, indépendante des domaines. Les dimensions sont groupées dans de petites catégories fermées et opposées. Elles peuvent diviser les domaines en deux classes. Par exemple, dans le domaine //culinaire//, les cuisiniers et les instruments de cuisine se différent par les dimensions //animé// et //inanimé//, respectivement. Les dimensions entretiennent des relations de

disjonction exclusives et obligatoires, comme //négatif// *versus* //positif//. Sur le plan sémantique, elles reflètent des « catégories *a priori* qui structurent tout l'univers d'une culture » (Rastier et coll., 1994, p. 63) puisqu'elles renvoient à des systèmes de valeurs tels que //bon// *versus* //mauvais//, //mélioratif// *versus* //péjoratif//, mais aussi à des réalités physiques comme //concret// et //abstrait//.

Suivant cette typologie, l'analyse sémique des mots « fourchette », « couteau » et « cuillère » résulterait dans le classement suivant : les trois ont en commun le sème générique (sg) /couvert/, qui forme un taxème //couvert//, et appartiennent au domaine //alimentation// ; ils s'inscrivent aussi dans la dimension : //inanimé// et //concret// et se distinguent par les sèmes spécifiques (sp) /piquer/, /couper/ et /contenir/, respectivement. La figure 6 ci-dessous illustre l'organisation du taxème //couvert// :

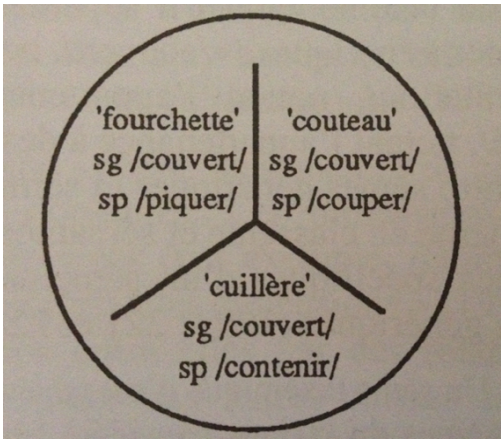


Figure 6. Le taxème //couvert// tiré de Hébert (2001, p.74)

La récurrence des sèmes génériques ou des sèmes spécifiques produit un type de structure différente selon le cas, soit l'isotopie ou la molécule sémique, respectivement. Nous les expliquerons dans la section suivante.

4.2 La structuration des sèmes : isotopies et molécules sémiques

Une isotopie est constituée de la récurrence de sèmes (spécifiques ou génériques), sur un empan de longueur variable (une phrase, un paragraphe, un texte ou même un corpus) (Rastier et coll., 1994 ; Valette, 2010). La récurrence de sèmes génériques est responsable de

l'homogénéité sémantique d'un texte (Rastier et coll., 1994; Pincemin, 1999a). Les isotopies signalent la présence de taxèmes, de domaines et de dimensions.

La molécule sémique est la récurrence groupée de sèmes spécifiques[20], qui peuvent appartenir à des classes sémantiques différentes (taxèmes, domaines et dimensions). Dans ce sens, la molécule sémique forme des classes sémantiques non lexicalisées, autrement dit, dont le sens n'est pas stabilisé en langue. Ces associations peuvent se stabiliser comme un syntagme nominal au fil du temps. Par exemple, la cooccurrence de « tabac » et « choix » ainsi que de « fumer » et « liberté » dans un texte, signale le réseau associatif des sèmes spécifiques /liberté/ et /fumer/ (Slodzian et Valette, 2009). Cette récurrence peut éventuellement créer un figement lexical, comme c'est d'ailleurs le cas dans l'expression « liberté de fumer », n'étant apparu que très récemment avec le débat sur l'interdiction de fumer dans les lieux publics (Slodzian et Valette, 2009).

Les isotopies et les molécules sémiques permettent de décrire la thématique des textes et c'est sur le plan de cette description que nous pouvons comprendre plus en profondeur leurs différences. Nous les présenterons dans les deux points suivants.

4.2.1 Description de thèmes par les isotopies et les molécules sémiques

Sur le plan thématique, les isotopies fournissent « des informations sur le domaine principal d'un texte (économie, politique, science…) » et structurent son « thème générique » (Rastier et coll., 1994, p. 207). La représentation d'un état de choses que produirait la dénotation est en fait liée à une particularité des isotopies (p.69). Par exemple, dans « L'amiral

[20] La sémantique interprétative distingue entre l'isotopie générique, formée par la récurrence de sèmes génériques et l'isotopie spécifique, formée par la récurrence de sèmes spécifiques. Comme nous avons vu, le caractère générique ou spécifique d'un sème est relatif, et la spécificité d'un sème est seulement perçue en fonction des classes sémantiques formées. Dans *Sémantique pour analyse* (Rastier et coll., 1994), les molécules sémiques sont identifiées comme étant des isotopies spécifiques. Pour des fins de simplification et en fonction de nos objectifs de recherche, nous avons neutralisé ces distinctions. Nous avons décidé d'utiliser le terme « isotopie » pour la récurrence de sèmes génériques formant les classes (taxème, domaine et dimensions) et molécule sémique pour la récurrence groupée de sème spécifiques, c'est-à-dire des sèmes qui peuvent appartenir à des classes sémantiques différentes.

ordonna de carguer les voiles », nous distinguons une isotopie induite par la récurrence du sème générique /navigation/ indexant le domaine //navigation//, qui constitue aussi son thème générique. En revanche, les molécules sémiques fournissent des informations sur des classes sémantiques à moindre généralité, par la présence de « groupements structurés de traits saillants » (p. 130). Elles sont signalées par un réseau associatif de sèmes spécifiques qui peuvent être indexés par des mots appartenant à des taxèmes, des domaines ou des dimensions différents. Pour cela, les molécules sémiques sont responsables de la formation de « thèmes spécifiques » (Ballabriga, 2005 ; Rastier, 2001a ; Rastier et coll., 1994).

Dans « L'Assommoir » de Zola, dont le thème générique est consacré au monde ouvrier et à la pauvreté des travailleurs, une molécule sémique qui regroupe les sèmes spécifiques /jaune/, /chaud/, /visqueux/ et /néfaste/ est lexicalisée dans les mots « alcool », « sauce », « morve », « huile », « pipi », selon l'exemple donné par Hébert (2001). L'auteur explique que pour signaler l'existence d'une molécule sémique, il faut que deux sèmes soient actualisés au moins deux fois dans un même mot. Penchons-nous sur l'exemple précité à l'aide du tableau sémique ci-dessous (tableau 2). Chaque mot de la molécule sémique dans « L'Assommoir » indexe au moins deux sèmes (le « + » signale la présence du sème) :

Tableau 2. Molécule sémique dans « L'Assommoir » de Zola

	/jaune/	**/chaud/**	**/visqueux/**	**/néfaste/**
alcool		+		+
sauce		+	+	
morve	+	+	+	+
huile	+		+	
pipi	+	+		+

Un exemple plus près de notre corpus qui nous permet de comprendre davantage la différence entre les isotopies et les molécules sémiques et par là entre thèmes génériques et thèmes spécifiques, c'est le débat public au sujet de l'avortement. Dans certains textes à ce sujet, il est possible de remarquer une isotopie qui indexe le sème générique /reproduction sexuée/, en fonction de la présence de mots comme « grossesse », « conception », « fœtus » et « avortement ». Le regroupement de certains sèmes spécifiques permet de voir la façon dont la

question de l'avortement est traitée. Par exemple, dans les textes favorables au droit à l'avortement, nous pourrions retrouver des mots formant une molécule sémique avec les sèmes /liberté/ + /reproduction/ associé à l'idée de droits reproductifs et actualisés par des mots comme « choix », « individu », « corps », « contraception ». Dans les textes contre l'avortement, nous pourrions retrouver la récurrence groupée de sèmes /mort/ + /humain/ actualisée dans des mots comme « crime », « assassinat », « être humain », « vie », etc. L'exemple ci-dessous, tiré d'un article d'opinion sur l'avortement, illustre la formation d'une molécule sémique regroupant les sèmes /liberté/ et /corps/ actualisée dans les mots « indépendants », « priorités », « aspirations », « corps » et « personne » :

> Forcer une femme, sous la menace de sanction criminelle, à mener le fœtus à terme, à moins qu'elle ne remplisse certains critères indépendants/liberté/ de ses propres priorités/liberté/ et aspirations/liberté/, est une ingérence profonde à l'égard de son corps /corps/ et donc une atteinte à la sécurité de sa personne/corps/ / [21].

4.2.2 Fond et formes sémantiques

Sur le plan perceptif, l'ensemble des isotopies crée le fond sémantique et les molécules sémiques constituent les formes sémantiques qui se détachent de ce fond[22]. Par rapport à la caractérisation thématique des textes, la distinction entre fonds sémantiques et formes

---

[21] Décriminalisation de l'avortement - Vingt ans de liberté et d'égalité. Louise Desmarais https://www.ledevoir.com/opinion/idees/173611/decriminalisation-de-l-avortement-vingt-ans-de-liberte-et-d-egalite. Consulté le 6 août 2018.

[22] Dans *Sémantique et recherches cognitives*, Rastier (2001c, p. 222) associe les fonds sémantiques aux isotopies génériques et les formes sémantiques aux isotopies spécifiques : « les rapports entre formes et fonds sémantiques, décrits comme des rapports entre isotopies génériques et isotopies spécifiques ». Par contre, dans *Sémantique pour Analyse*, nous constatons une autre homologation des concepts : « (…) les isotopies spécifiques sont ainsi un facteur de la cohésion des périodes (et, au-delà, des textes). Elles participent ainsi à la définition des fonds perceptifs » (Rastier et coll., 1994, p. 128). Une autre explication donnée par Rastier sur les formes sémantiques explique qu'une isotopie spécifique, lorsqu'elle se retrouve diffuse, peut participer à un « fond perceptif » (Rastier, 2006a). La relativité à l'égard des notions spécifique et générique dans le cadre théorique nous amène à neutraliser ces distinctions, et à mettre l'accent sur l'aspect perceptif de la notion de fond, que nous considérons ici comme l'unification de sèmes d'une classe lexicale, liée à un domaine. Dans cette perspective, les formes sémantiques correspondent aux molécules sémiques, c'est-à-dire à la récurrence groupée de sèmes spécifiques.

sémantiques est la même que celle élaborée par l'école linguistique de Prague pour la distinction entre thème et rhème, ou entre thème et focus (Rastier et coll., 1994). Le thème est l'information qui est déjà connue par les participants à la communication (Flament, 2006). Le rhème, est l'information nouvelle, ce qui est dit à propos du thème (Forest et Meunier, 2004). Suivant cette opposition entre fonds sémantiques et formes sémantiques, l'activité interprétative consisterait à relever les formes des fonds et à rendre compte de leurs transformations dans le temps textuel (Belghanem, 2009).

Dans Rastier et coll. (1994) le fond sémantique est associé aux isotopiques génériques et à des classes dans la langue socialement normées (comme les dimensions), tandis que les formes sémantiques sont associées aux molécules sémiques.

4.3 Isosémies : description sémantique des mots grammaticaux

Les explications et les exemples que nous avons fournis jusqu'ici mettent l'accent sur l'aspect thématique des textes. Malgré cela, la SI ne se limite pas à décrire la thématique et les types d'observables linguistiques auxquelles elle s'intéresse. Elles ne se limitent pas non plus aux mots lexicaux. Les mots grammaticaux ont leur statut sémantique mis en valeur dans le cadre de la théorie. La récurrence de mots qui relèvent de la morphosyntaxique des textes, incluant les morphèmes intégrant les flexions verbales, constitue également des fonds sémantiques qui peuvent être décrits (Rastier, 2015, p.7).

À la différence des taxèmes et des domaines, où l'actualisation de traits sémantiques peut s'appuyer sur la connaissance de l'usage de la langue par rapport à une pratique sociale ou une culture, l'actualisation de traits sémantiques des mots grammaticaux relève exclusivement du système fonctionnel de la langue. Leur récurrence atteste d'un type particulier d'isotopie, nommée isosémie, désignation que Rastier (Rastier et coll., 1994) emprunte lui-même de Pottier (1987). Une isosémie est ainsi définie par la récurrence de sèmes prescrits par le système fonctionnel de la langue, dont l'interprétation dépend des connaissances de ses règles grammaticales.

Les isosémies selon Rastier et coll. (1994) sont plus généralement présentes dans les relations de concordance et de rection entre syntagmes. Dans la langue française, ces relations se trouvent grammaticalisées par les morphèmes. Par exemple, en français, l'accord relevant

du nombre est grammaticalisé par le « s » qui est apposé à la fin des mots, désignant le sème générique /pluriel/. L'accord des mots au féminin dans « la grande montagne » (Rastier et coll, 1994, p. 120) est actualisé par les morphèmes « -a » et « -e » associés à l'article « le » et à l'adjectif « grand », respectivement. Ces morphèmes ont le trait inhérent /genre féminin/ et leur récurrence indexe les mots dans l'isosémie //genre féminin//.

À notre connaissance, Rastier n'a pas décrit de façon systématique la morphosyntaxique française et sur ce point, les apports théoriques de la SI ne nous permettent pas d'établir un ensemble de catégories descriptives. Cependant, dans « Sémantique pour l'analyse *»,* Rastier et coll. (1994) suggèrent certaines directions de recherche lorsqu'ils analysent les isosémies aspectuelles, relatives aux modalités verbales /perfectif/ et /imperfectif/ qui se trouvent présentes dans les verbes et dans d'autres mots grammaticaux du français. Sont cités en particulier les travaux de Pottier (1974, 1987), qui semble être le premier à essayer de systématiser une grammaire sémantique. Dans « Théorie et analyse en linguistique », Pottier (1987) dresse un ensemble de catégories sémantico-grammaticales ayant trait à la description sémantique du temps, de l'aspect, de la modalité et de la détermination par le biais des mots grammaticaux de la langue.

Rastier (2014) propose par contre la théorie des zones anthropiques permettant de situer la description de traits sémantiques des mots grammaticaux de la langue.

Rastier (2014) explique que la caractéristique principale de la communication humaine et qui nous distingue par rapport aux systèmes de communication des autres animaux, c'est la possibilité de parler de ce qui n'est pas là. La possibilité de dépasser le *hic et nunc* dans l'acte de la communication, ce que Benveniste (1967) désigne comme « débrayage », est l'effet de la reconnaissance d'un « soi » et aussi d'un « autre » dans la communication ou encore, d'un « ici » et d'un « ailleurs ». Cette reconnaissance, qui rend possible l'interlocution, est attestée par la deixis des langues sur quatre axes fondamentaux : l'axe personnel, l'axe temporel, l'axe spatial et l'axe modal (Rastier, 2014, p.37). Chacun de ces axes relève de catégories grammaticales présentes dans la plupart des langues. L'axe personnel est lié aux phénomènes de la pronominalisation et aux manières de représenter linguistiquement les personnes de l'interlocution (première, seconde et troisième personne) ; l'axe temporel et spatial relève des recours linguistiques qui permettent de situer les événements dans le temps et dans l'espace

(présent, passé et futur proche ou lointain, ici, là, ailleurs) ; et l'axe modal relève des catégories grammaticales qui sont associées aux verbes et qui traduisent « l'attitude du sujet parlant à l'égard de ses propres énoncés » (Dubois et coll., 2007, p. 306). En fonction de ces distinctions, Rastier (2014) définit trois zones dites anthropiques : la zone identitaire, proximale et distale. Chaque axe se trouve en relation avec ces zones, en créant différentes oppositions (tableau 3) :

1. Sur l'axe personnel se trouvent opposés les protagonistes de l'interlocution.
2. Sur l'axe temporel se trouvent opposés les repères temporels.
3. Sur l'axe local se trouvent opposés les repères spatiaux.
4. Sur l'axe modal se trouvent opposés les modalisateurs.

Tableau 3. Les trois zones anthropiques (Rastier, 2014, p.19)

	Zone identitaire	**Zone proximale**	**Zone distale**
Personne	JE, NOUS	TU, VOUS	IL, ON
Temps	MAINTENANT	NAGUÈRE, BIENTÔT	PASSÉ, FUTUR
Espace	ICI	LÀ	LÀ-BAS, AILLEURS
Mode	CERTAIN	PROBABLE	POSSIBLE/IRRÉEL

La zone *identitaire* est définie comme une zone de coïncidence entre l'énonciateur et l'interprète, tandis que la zone *proximale* présuppose l'altérité. La zone *distale* pour sa part est définie comme une zone d'étrangeté et réfère à ce qui est absent de l'interlocution. La structure grammaticale de la plupart des langues existantes reflète la démarcation de zones,

ainsi que les axes de la personne, du temps, de l'espace et du mode. Sur l'axe personnel, la paire JE/TU n'est pas nécessairement liée aux personnes de l'interlocution, mais peut être associé(e) par exemple à un groupe ou à une nation (Rastier, 2014). Sur l'axe du temps et de l'espace, la troisième zone démarque plus fortement ce qui est absent du *hic et nunc* de l'interlocution : les récits passés, les lieux lointains, l'histoire ou les événements projetés dans le futur font tous partie de la zone distale. Dans les langues, les flexions temporelles des verbes ainsi que les adverbes d'espace et de temps attestent grammaticalement de ces ruptures et permettent de structurer les événements, leurs successions et les processus impliqués (Dubois et coll., 2007, p. 53). Sur l'axe modal, les zones séparent plus proprement l'espace modal de l'énonciateur et de son rapport avec le monde qui l'entoure. Certaines formes d'expression communément utilisées pour introduire la pensée de l'énonciateur permettent de caractériser les positionnements modaux. Les expressions « je dis », « j'affirme », « je confirme », « je postule » par exemple, révèlent d'un positionnement certain, ancré dans la zone identitaire. Par contre, des expressions comme « il se peut », « il est probable », « il est improbable », renvoient à un positionnement plutôt incertain et hypothétique (Dubois et coll., 2007), ancrés dans la zone distale.

L'importance de cette proposition théorique de Rastier (2014) tient au fait que certaines propriétés des langues, comme la pronominalisation ou la référentialisation spatio-temporelle, gardent des traits sémantiques inhérents permettant d'inférer comment l'énonciateur se situe dans le continuum de l'espace-temps et aussi par rapport à tous les autres êtres du monde, qu'ils soient réels ou imaginaires. Dans les textes, l'identification de ces récurrences sémiques permet de décrire la manière particulière dont l'énonciateur se projette sur chacun de ces axes. L'élaboration textuelle peut ainsi être conçue comme une articulation particulière de chacun de ces axes, caractérisée par le point de vue de celui qui se représente dans le texte en tant qu'énonciateur. Nous allons développer un peu plus cette idée dans les prochaines sections où nous allons parler des composantes sémantiques.

4.4 Les composantes sémantiques

La sémantique interprétative propose d'étudier le contenu des textes, qu'elle définit comme l'interaction entre quatre composantes sémantiques : la thématique, la dialectique, la dialogique et la tactique.

> 1 — La thématique rend compte des contenus investis, c'est-à-dire du secteur de l'univers sémantique mis en œuvre dans le texte. Elle en décrit les unités. Par analogie, et bien qu'elle ne décrive pas spécifiquement le lexique, on peut dire qu'elle traite du « vocabulaire » textuel (…).
>
> 2 — La dialectique rend compte des intervalles temporels dans le temps représenté, de la succession des états entre ces intervalles et du déroulement aspectuel des processus dans ces intervalles
>
> 3 — La dialogique rend compte des modalités, notamment énonciatives et évaluatives, ainsi que des espaces modaux qu'elles décrivent. Dans cette mesure, elle traite de l'énonciation représentée (l'énonciation réelle ne relevant pas de la linguistique, mais de la psycholinguistique).
>
> 4 — La tactique rend compte de la disposition séquentielle du signifié, et de l'ordre (linéaire ou non) selon lequel les unités sémantiques à tous les paliers sont produites et interprétées. (Rastier et coll., 1994, p. 40)

Les composantes sémantiques sont analysables par le biais des sèmes et plus spécifiquement par la récurrence ou co-récurrence de sèmes. Inversement, chaque mot (lexical ou grammatical) peut être caractérisé en fonction des composantes dans la mesure où chaque mot peut être situé par sa position dans l'univers sémantique (thématique), par un repérage temporel (dialectique), modal (dialogique) ou distributionnel (tactique) (Rastier et coll., 1994, p. 40).

Un des objectifs de la SI est de décrire sémantiquement l'interaction entre les composantes dans les textes. Cette description répond, dans le cadre de la théorie, à la question de la caractérisation des genres et par là, à la caractérisation de textes organisés dans des corpus. En effet, le texte - tel que le genre dont il constitue une occurrence - atteste d'une interaction particulière entre les composantes. Par exemple, le genre de l'opinion (tel que nous

l'avons défini dans notre recherche) présente une thématique ouverte, parce qu'il n'impose pas un lexique circonscrit à un univers sémantique particulier comme c'est le cas des recettes de cuisine. Par rapport à la composante dialogique, le genre de l'opinion est marqué par un engagement subjectif plus prononcé de l'énonciateur, mais cela n'est pas une règle obligatoire : un article d'opinion peut être écrit en « je », mais aussi en « il », ou en « nous ». En gardant la perspective du genre comme une interaction entre les composantes, la SI offre l'avantage de caractériser, par rapport à un genre particulier ou par rapport à un ensemble de textes regroupé en corpus, l'usage que fait l'auteur de celui-ci, et de découvrir les procédés d'élaboration textuelle qui sont privilégiés.

Cette section expliquera comment la sémantique interprétative définit chacune des composantes. Dans cet exposé, nous souhaitons fournir quelques exemples d'application de ces concepts dans la démarche de sélection de critères textuels interprétables, tels que le proposent les travaux de Valette (2004) et Eensoo et Valette (Eensoo et Valette, 2012, 2014a, 2014b, 2015).

4.4.1 Composante thématique

La composante thématique rend compte des contenus, ou comme dit Rastier et coll. (1994, p. 40), de « l'univers sémantique mis en œuvre dans le texte ». La description thématique repose sur la description des sèmes en présence, qu'ils soient génériques ou spécifiques (Hébert, 2001) : « Au sens plus général, un thème est défini par toute récurrence d'un sème ou co-récurrence de sèmes, qu'ils soient génériques et/ou spécifiques » (Hébert, 2001, p. 122). De cette distinction, un thème générique est défini par un sème ou une structure de sèmes génériques récurrents, c'est-à-dire une isotopie. Rastier le considère comme le « sujet » du texte, puisque les isotopies génériques « induisent les impressions référentielles dominantes » (Rastier et coll., 1994, p. 177). Pour sa part, le thème spécifique est défini par « un groupement récurrent de sèmes spécifiques, c'est-à-dire une molécule sémique » (Hébert, 2001, p. 123). Sa particularité, comme nous avons expliqué précédemment, tient au fait qu'elle actualise plus d'un sème (co-récurrence de sèmes) et que l'association sémantique formée ne privilégie pas une lexicalisation particulière.

> La récurrence d'un sème générique induit une isotopie générique. Et parfois, dans son acception générale, le mot thème est employé pour désigner le sujet » d'un texte, c'est-à-dire son isotopie générique dominante, ordinairement un domaine sémantique. […]. En revanche, un thème spécifique peut se définir comme une molécule sémique, c'est-à-dire un groupement structuré de sèmes spécifiques. (Rastier, 2001a, p. 197).

Dans « Arts et Sciences du Texte », Rastier (2001a) explique que les thèmes génériques, qui sont formés par la récurrence de sèmes génériques, décrivent principalement les domaines sémantiques des textes. Outre le domaine, la récurrence de sèmes génériques peut décrire les taxèmes. À un autre niveau, le thème spécifique signale la récurrence d'un sème spécifique ou la co-récurrence de sèmes spécifiques. Il s'agit plutôt d'un réseau associatif avec des mots qui peuvent appartenir à des domaines ou à des taxèmes différents.

> (…) le domaine d'un texte peut être assimilé à un thème générique alors que son sujet peut être représenté par un thème spécifique (rappelons que les thèmes génériques sont habituellement des *isotopies* alors que les thèmes spécifiques sont des *molécules sémiques.* (Rastier et coll., 1994, p. 207).

Les travaux de Valette (2004) et Eensoo et Valette (2012, 2014a, 2014b, 2015) explorent la dimension thématique des textes en employant d'une part le calcul des spécificités et d'autre part, le calcul de cooccurrence. Le premier calcul permet de contraster les sous-corpus au corpus de référence et de ressortir les données textuelles qui sont surreprésentées dans chaque sous-corpus. Elles sont sélectionnées comme critères textuels sur la base des scores statistiques obtenus et sur certains critères linguistiques établis par les chercheurs. Les critères spécifiquement lexicaux sont associés à la composante thématique et leurs récurrences sémiques sont décrites en utilisant le concept d'isotopie. Le calcul de cooccurrence est également employé, permettant de dégager des couples de mots qui se trouvent fortement associés au sein de chaque sous-corpus : ces cooccurrences sont assimilées aux molécules sémiques et aussi décrites dans une perspective thématique. L'objectif de la démarche est de qualifier les critères textuels de façon contrastive, expliquant au même temps comment chaque sous-corpus véhicule des valorisations différentes sur les thèmes qui sont abordés dans la discussion.

Dans le but de trouver des critères textuels pour la construction d'un système de filtrage de textes racistes et antiracistes pour la plateforme PRINCIP[23], Valette (2004) a proposé d'appliquer des calculs textométriques à un corpus constitué de textes de sites web racistes et antiracistes repérés sur le Web. Le calcul des spécificités a permis de dégager un ensemble de critères textuels utilisés par les auteurs racistes et antiracistes, comprenant des mots lexicaux et grammaticaux. Parmi ces critères, les mots lexicaux sont ceux qui ont permis de décrire de façon plus précise la thématique des deux sous-corpus. Un travail de description des critères provenant du calcul des spécificités a permis d'identifier et de contraster les isotopies génériques de chaque sous-corpus, que l'auteur a nommé thème générique. L'auteur a également exploré le calcul de cooccurrence pour sélectionner d'autres critères proprement lexicaux qui permettaient de dégager les molécules sémiques des sous-corpus, en sélectionnant des cooccurrents spécifiques autour d'un ensemble de mots pôles récurrents dans le corpus de référence. Les cooccurrents sont associés à des molécules sémiques en fonction de la récurrence simultanée d'un sème qui provient du mot pôle et d'un sème qui provient de l'ensemble de cooccurrents spécifiques, formant un réseau autour du mot pôle. Les molécules sémiques ont été associées aux thèmes spécifiques de chaque sous-corpus.

Dans le sous-corpus raciste, Valette (2004) a découvert la spécificité de mots comme « femelle », « mâle », « bipède », « macaque » et il a décrit la présence d'une isotopie assez particulière chez les racistes, l'isotopie //animalité// qui était absente des textes antiracistes (Valette, 2004).

Par la méthode de cooccurrence, il a repéré un ensemble de molécules sémiques spécifiques de chaque sous-corpus. Le tableau 4 présente les cooccurrents spécifiques dans les sous-corpus raciste et antiraciste qui sont associés au mot pôle « immigration ». Nous pouvons observer que dans les textes racistes, un thème spécifique sur le contrôle et la croissance de l'immigration, ainsi que l'association de l'immigration au mot « invasion » marquait une différence avec les textes du sous-corpus antiraciste, où le mot est plutôt associé à « frontière » et « flux ».

[23] Plateforme pour la Recherche, l'Identification et la Neutralisation des Contenus Illégaux et Préjudiciables sur l'Internet (http://www.princip.net).

Tableau 4. Cooccurrences spécifiques associées au mot « immigration »
(Valette, 2004)

Raciste	Antiraciste
immigration-incontrôlée	immigration-frontière
immigration-croissante	immigration-flux
immigration-invasion	

Dans les travaux plus récents (Eensoo et Valette, 2012, 2014a, 2014b, 2015), la distinction conceptuelle entre thème générique et thème spécifique n'est pas explicitement mentionnée, mais le calcul de cooccurrence est encore retenu pour sélectionner des critères thématiques. Les critères relevant des domaines des textes sont particulièrement privilégiés.

4.4.2 Composante dialectique

La composante dialectique « articule la succession des intervalles dans le temps textuel, comme les états qui y prennent place et les processus qui s'y déroulent » (Rastier, 1989, p. 278). Cette composante rend compte de la représentation du temps ainsi que les structures de logique argumentative et en ceci, elle se rapproche des théories du récit et de la narrativité (Ballabriga, 2005, Rastier et coll., 1994).

> Comme elle traite des intervalles de temps représenté, la dialectique rencontre les théories du récit, qui ne participent pas de théories générales du texte, mais dont les acquis doivent être sauvegardés. (Rastier et coll., 1994, p. 178)

Selon Rastier (2014), le temps textuel est signalé par des « discontinuités qualitatives » au sein des textes (Rastier, 2014, p.3), et à l'intérieur du récit il décrit le sens des actions. Le temps textuel permet également de percevoir la connexion entre les événements, comme le début, le déroulement et la fin des actions, mais aussi leur transformation et leurs changements. La temporalité n'est pas seulement chronologique. Elle peut être sociale, liée à l'histoire humaine et à ses représentations collectives, comme la succession des années, des âges, des cycles ou des saisons. Elle peut être aussi relative au vécu individuel, se présentant sous forme de moments ponctuels entre un début et une fin du récit : il s'agit par exemple du changement des conditions qui transforment un personnage en un héros, ou un personnage

présumé innocent qui se révèle coupable à la fin du récit. Le temps textuel signale également la relation logique entre les syntagmes (Rastier et coll., 1994), par exemple la relation d'opposition entre deux contenus propositionnels : « Vous ne me croyez pas et *pourtant,* j'ai raison ».

Dans la perspective de la SI, la structure actancielle relève de la composante dialectique. Les fonctions syntaxiques sont décrites comme des sèmes génériques ou des molécules sémiques. Rastier désigne comme acteurs les unités dans les textes qui sont chargées de l'actance, impliquant autant les personnages du récit que les processus verbaux reliant les sujets et les prédicats. L'acteur est défini comme un complexe sémique qui peut être nommé ou recevoir des descriptions diverses dans le texte, à l'exemple d'un personnage ou du narrateur d'un texte ou même d'une entité non physique comme « baisse du dollar » (Rastier, 1989, p.73).

Rastier (1989) fait la distinction de deux fonctions dialectiques entre les acteurs, la fonction irénique et la fonction polémique. Les fonctions iréniques sont celles qui expriment des interactions de contrat ou d'échange entre les acteurs, ainsi que la réconciliation ou le consensus. Les fonctions polémiques en revanche, signalent des interactions d'affrontement et de lutte. Dans un récit par exemple, le DÉFI serait une fonction polémique, composée de deux interactions, l'attaque et la contre-attaque. Dans un roman, il pourrait présenter un personnage qui contrôle une ville et un peuple qui se lève contre l'invasion. Mais cette structure peut également être présente dans une phrase, par exemple : « Paul frappe Philippe, mais il riposte ». Dans l'hypothèse d'une interaction de DÉFI entre deux acteurs, chacun de ces derniers assumerait un rôle : le rôle d'agent, ou sujet de l'action, celui qui déclenche le DÉFI (ergatif) et le rôle de destinataire, celui qui reçoit la transmission (datif). Rastier et coll. (1994) proposent une formalisation des fonctions entre les acteurs par le biais de graphes sémantiques. Dans l'exemple ci-dessous, les acteurs A et B sont liés par des nœuds représentant les rôles et les fonctions. Dans l'hypothèse d'un DÉFI de A envers B, où A est ergatif (ERG) et B datif (B), le graphe serait représenté ainsi :

$$[A] \leftarrow (ERG) \leftarrow [D\acute{E}FI] \rightarrow (DAT) \rightarrow [B]$$

Dans la perspective de la SI, la structure actancielle de la composante dialectique fonctionne comme un système de propagation de sèmes dans le texte et sa description permet de faire l'inventaire des rôles et des fonctions entre les acteurs, ainsi que de rendre compte de leurs transformations dans le récit (Rastier et coll., 1994). Les objectifs de notre recherche nous amènent à neutraliser ces concepts, en fonction de la nature même du traitement informatique que nous proposons ici. Dans les travaux de Valette (2004) et Eensoo et Valette (2012, 2014a, 2014b, 2015), l'identification de critères textuels pertinents de la composante dialectique est basée sur la sélection de certaines unités morphsyntaxiques ressorties du calcul des spécificités qui sont sémantiquement liées au temps textuel et à l'argumentation, comme les adverbes de temps et les marqueurs argumentatifs.

Ordinairement, les traits temporels sont ceux qui permettent de situer les actions dans le temps. Des exemples de repérage temporel dans les textes sont donnés par des locatifs comme « maintenant », « hier », « toujours », ainsi que par les relateurs « avant », « quand », « pendant », « au cours de » (Adam, 2008 ; Pottier, 1987, p. 160). Dans le français, comme dans d'autres langues latines, les terminaisons verbales sont des morphèmes qui ont des traits inhérents relatifs au temps (passé, présent, futur). Les traits aspectuels sont ceux qui permettent d'ordonner les processus et de structurer le récit en épisodes, transitions et intervalles (Rastier, 1989, p. 67). Les processus peuvent être caractérisés comme accomplis ou non accomplis selon qu'ils mettent l'accent sur le caractère initiateur des actions (inchoatif), leur caractère de progression ou de continuité (duratif) ou encore d'achèvement (terminatif). Dubois et coll. (2007) expliquent que l'aspect « exprime la représentation que se fait le sujet parlant du procès exprimé par le verbe tandis que les temps, les modaux et les auxiliaires de temps expriment les caractères propres du procès indiqué par le verbe » (p. 53). Ainsi, les phrases « Pierre mangeait » et « Pierre a mangé » sont toutes les deux situées dans le temps passé, mais se distinguent par rapport à l'aspect, la première phrase mettant en relief l'aspect de déroulement (duratif) et la deuxième, l'aspect achevé ou accompli des processus (terminatif). Parmi les exemples d'aspects cités par Dubois et coll. (2007) se trouvent les verbes auxiliaires et les verbes semi-auxiliaires. Les verbes auxiliaires et les verbes semi-auxiliaires comme « commencer à », « en train de » ou « venir de » décrivent aussi les caractéristiques aspectuelles /inchoatif/, /duratif/ et /terminatif/ respectivement. La récurrence

de verbes perfectifs ou imperfectifs peut exprimer le caractère /singulatif/ ou /itératif/ des actions.

Une des particularités de la sémantique interprétative est que la théorie n'impose pas de catégories linguistiques *a priori* pour caractériser les phénomènes réputés à la composante dialectique. Par exemple, dans le vers d'Éluard « l'aube allume la source », le sème aspectuel /inchoatif/ se retrouve présent dans les mots « aube » et « allume » respectivement, même si d'autres sèmes de statuts différents s'y retrouvent également (c.-à-d., /clarté/) (Rastier, 2006a). L'actualisation de sèmes afférents peut également informer sur une interaction dialectique entre des acteurs. Par exemple, le contexte peut instruire l'afférence des sèmes /irénique/ pour « colombe » et /polémique/ pour « corbeau » dans un roman ou un poème (Belghanem, 2009). La pertinence des traits s'évalue dépendamment du corpus choisi et des objectifs de la description. Ce caractère non restrictif de la théorie, déliée de la préoccupation de faire correspondre un ensemble de catégories grammaticales à des structurations spécifiques, concède l'avantage de faire une description adaptée au corpus en question, évitant de créer des conceptualisations complexes qui ne tiennent pas compte de ce qui est effectivement attesté dans les genres et les textes considérés. Aussi, elle évite de présupposer l'absence d'une structuration dialectique en fonction de l'absence de certains marqueurs textuels.

En fonction de la place laissée à l'investigation empirique à l'intérieur de la théorie, l'application de la SI, dans le contexte de la fouille d'opinions, ne préjuge pas les types particuliers de données textuelles pour caractériser les composantes sémantiques. Il peut s'agir de critères typographiques, de mots, d'expressions lexicales figées ou semi-figées, de catégories morphosyntaxiques ou même de morphèmes. C'est l'interprétation de ces données en conformité avec la théorie et en aval des calculs textométriques qui permet de statuer sur leur valeur descriptive et sur le type de structuration sémantique qu'elles attestent relativement aux textes du corpus (Pincemin, 2012a).

Le nombre de critères caractérisés par Valette (2004) et Eensoo et Valette (2012, 2014a, 2014b, 2015) comme dialectiques relèvent de catégories grammaticales diverses et attestent les normes particulières des genres retenus pour les expérimentations. Ils sont des critères qui font état de l'organisation temporelle et de la logique argumentative. Il s'agit de marqueurs de structuration argumentative (« par contre », « car »), marqueurs de structuration

temporelle (« après », « puis »), de ponctuations de phrases (emphases, points d'interrogation, point d'exclamation, mots interrogatifs), de verbes modaux (« il faut », « il doit »), d'énumératifs (« Enfin », « Il serait également souhaitable que »), de marques de négation (« non », « ne », « pas », « jamais », « rien »), d'adverbes (« comment », « quand »), de marqueurs narratifs (« depuis des années », « puis ») et de locutions disjonctives (« alors que »). Plus généralement, ces indices font état de la manière dont les événements sont racontés et interreliés.

Dans les travaux de fouille d'opinions inspirés de la SI (Eensoo et Valette, 2012, 2014a, 2014b, 2015 ; Valette, 2004), la description des observables dialectiques est faite non seulement sur le plan des sèmes temporels (/passé/, /présent/, /futur/) et aspectuels (/inchoatif/, /duratif/, /terminatif/), mais aussi sur le plan des stratégies rhétoriques adoptées par les auteurs, afin de les élucider. Ainsi, des stratégies rhétoriques agressives et polémiques sont attestées par certains adverbes d'interrogation et marqueurs argumentatifs d'opposition (« or », « mais », « pourtant »), marqueurs d'emphase (« dire que », « honteux », « ! »), de saturation (« il y en a marre », « encore », « nombreux ») ou de polémicité (« jamais », « absolument »).

Les zones anthropiques de Rastier sont également utilisées comme catégories descriptives des critères dialectiques. En utilisant un corpus de commentaires d'articles de presse rédigés par les lecteurs internautes, Eensoo et Valette (2014a) ont constaté que la présence de critères relevant de la structuration argumentative était typique d'un profil spécifique de commentateur plus distancié et impersonnel. Ces critères ont été décrits comme indicateurs d'un ancrage dans la zone anthropique distale et relative à un registre de langage plus intellectuel, marqué par l'abstraction et la mise à distance :

> Les critères argumentatifs (*mais, comme, comment, dont*), caractéristiques de la composante dialectique, sont ici statistiquement significatifs. C'est l'indice d'un ancrage dans la zone anthropique distale (construction intellectuelle, abstraction, mise à distance). (Eensoo et Valette, 2014a, p. 117)

4.4.3 Composante dialogique

La composante dialogique rend compte de l'énonciation représentée et de la modalisation (Rastier et coll., 1994). L'énonciation représentée se distingue de la

représentation réelle, c'est-à-dire de la personne qui énonce. Elle est désignée dans la théorie comme un acteur ou un foyer énonciatif. Dans le récit, il s'agit du narrateur, mais dans certains genres, les foyers énonciatifs (ou acteurs) peuvent se multiplier. C'est le cas par exemple des textes scientifiques dans lesquels les nombreuses citations s'intègrent comme des « voix déléguées », considérés aussi comme des foyers énonciatifs (Rastier et coll., 1994, p. 181).

La modalisation pour sa part est décrite par Rastier (Rastier, 2015 ; Rastier et coll., 1994) comme un rapport entre l'acteur et son univers, ce dernier étant désigné comme « l'ensemble d'unités linguistiques associé à un acteur ou à un foyer énonciatif » (Rastier, 2015, p. 8) et qui détermine son point de vue, son jugement par rapport à ce qu'il exprime. Pour expliquer le concept d'univers, Rastier et coll. (1994) donnent l'exemple du narrateur de la « Cousine Bette » : « quand le narrateur (…) parle d'une mauvaise bonne action, « bonne » renvoie à l'univers des deux acteurs, et « mauvaise » à son propre univers » (p. 181). D'une façon générale, les mots qui signalent le jugement de l'énonciateur par rapport à son énoncé (Paveau et Safarti, 2003) relèvent de l'univers de l'énonciateur.

La SI propose des catégories de modalisation qui remontent à la logique aristotélicienne. Elles expriment la position de l'énonciateur sur la « réalité du monde » (Laurendeau, 2004) et les données objectives de cette réalité en ce qui a trait sa véracité, sa validité, sa probabilité ou sa possibilité. Sur un plan plus subjectif, la modalisation peut encore signaler les jugements des énonciateurs d'un point de vue appréciatif, par exemple bon et mauvais, positif et négatif, etc. (Laurendeau, 2004). La SI propose 4 catégories de modalisation : la modalité ontique (possible/impossible, factuel/non factuel), véridictoire (vrai/faux), épistémique (certain/incertain) et la modalité thymique (euphorique/dysphorique, c'est-à-dire du positif/négatif) (Hébert, 2001, p. 147). Par exemple, la phrase « Si les étudiants avaient voté contre la grève », le conditionnel présent « si les étudiants avaient », place le narrateur dans un univers hypothétique et non factuel.

Dans une perspective d'analyse statistique de corpus, plusieurs facteurs peuvent entrer en jeu lorsqu'il s'agit de caractériser la composante dialogique. Valette (2004) et Eensoo et Valette (2012, 2014a, 2014b, 2015) retiennent comme indices dialogiques principalement les pronoms personnels et les possessifs, des expressions subjectives, dans lesquelles

l'énonciateur exprime explicitement son point de vue, ou qui font état d'un récit personnel comme « je crois », « je m'aperçois », « je doute », « mon expérience », « pour ma part ».

L'interprétation de ces critères est basée sur des concepts de la SI, principalement en rapport avec les zones anthropiques. La description tient compte de l'ancrage de l'énonciateur dans le texte, dans une échelle qui varie entre un engagement prononcé, lorsque celui-ci prend en charge le propos en se représentant par un « je » ou un « nous », et une non prise en charge, caractérisée par l'absence de l'énonciateur et la présence d'autres acteurs auxquels il fait référence (« il », « ils »). L'élaboration énonciative égocentrée est caractéristique d'un énonciateur qui s'investit dans une zone anthropique identitaire (Eensoo et Valette, 2015), tandis que sa mise à distance caractérise l'énonciation comme relevant de la zone anthropique distale. Ces catégories peuvent également être des sources de description du registre du langage. Ainsi, l'utilisation de « tu » peut signaler un style plutôt familier, tandis que « vous » aurait la caractéristique d'établir une relation plus formelle ou un registre de langage soutenu. De même, le « nous » dénote un foyer énonciatif composé d'autres personnes, et démontre que le collectif emporte sur l'individuel (« je »).

Eensoo et Valette (2014b) proposent une méthode de sélection des critères textométriques pour la classification de sentiments, concernant un corpus constitué de 300 témoignages publiés par les internautes dans le site SAMESTORY (http://www.same-story.com) sur le thème de la santé. Les textes sont classés dans les catégories « gaies » et « tristes », puis soumis à une analyse outillée par les calculs textométriques. Pour la composante dialogique, les auteurs ont constaté un contraste important par rapport à l'énonciation. Les auteurs des textes « gais » s'adressaient explicitement à un « tu » (attesté par la fréquence de pronoms de deuxième personne), auquel ils donnaient des conseils sous la forme d'hyperliens, ou encore des encouragements (« bon courage »), tandis que les témoignages « tristes » se structuraient autour d'un « je » rapportant une expérience subjective.

4.4.4 Composante tactique

La tactique est une composante qui rend compte de la disposition linéaire des unités sémantiques dans le texte, précisant le rapport entre la linéarité du signifiant et du signifié. Rastier et coll. (1994) expliquent que les unités sémantiques sont restituées par des inférences

réalisées en fonction du palier textuel (le texte, le genre et le discours) et que leur sens n'est pas nécessairement lié à leur position linéaire dans le texte. Par exemple, dans les textes procéduraux, comme dans les recettes de cuisine, la disposition des unités sémantiques coïncide avec celle du signifiant, car l'ordre des événements permet une inférence des unités sémantiques à partir de leur position : préparation, cuisson, finition. Par contre, dans les romans policiers, où les événements qui expliquent les faits advenant au début sont seulement révélés à la fin du texte, le sémantisme des mots présents au début est seulement restitué à la fin. Autrement dit, les signifiés ne sont pas ordonnés selon la linéarité du signifiant (Hébert, 2001).

La composante tactique est la moins développée au sein de la sémantique interprétative (Hébert, 2001) et n'a pas été explorée encore dans le domaine de la fouille d'opinions. Nous la mentionnons ici à titre d'information, mais nous n'envisagerons pas d'exploiter les concepts de cette composante dans le cadre de notre recherche.

4.5 Démarche textométrique de sélection de critères textuels pour la fouille d'opinions : compatibilités avec la sémantique interprétative

Dans le contexte de la fouille d'opinions, le calcul des spécificités et le calcul de cooccurrence sont appliqués pour permettre l'analyse des données textuelles du corpus et la sélection de critères textuels. Ces calculs mettent en évidence les contrastes existants entre les sous-corpus constitués en fonction des types d'opinions qui doivent être caractérisés. La sélection de critères textuels est effectuée à la suite du calcul, par la qualification sémantique des données textuelles rencontrées. Elle comprend trois étapes principales :

1. Analyse des contextes d'apparition des données textuelles spécifiques à l'aide de l'outil de retour au texte (concordancier) fourni par les logiciels textométriques, afin d'assurer que la donnée textuelle est employée avec le même sens et la même fonction dans les contextes repérés.

2. Assimilation et dissimilation des récurrences rencontrées, et identification des isotopies, des isosémies et des molécules sémiques, dorénavant désignées structures sémantiques.

3. Catégorisation des critères par rapport aux composantes sémantiques, en fonction des structures sémantiques observées.

Dans la perspective de la démarche méthodologique proposée, nous désignons par *catégorisation* le processus qui associe un critère textuel à une des composantes sémantiques. Un critère est catégorisé comme thématique, dialectique ou dialogique si en fonction des récurrences qu'il décrit, il permet le repérage des thèmes, de la structuration argumentative et temporelle, ou de la représentation énonciative et de la modalisation présente dans les textes. Le tableau 5 ci-dessous présente une simplification des concepts de la théorie relatifs aux composantes (qui a été d'ailleurs proposée par les travaux de Eensoo et Valette [2012, 2014a, 2014b, 2015]). À chaque composante sémantique est associé un type d'organisation textuelle qui peut, elle, être décrite par la récurrence sémique identifiée.

Tableau 5. Types d'organisation textuelle associés aux composantes sémantiques

Composante sémantique	**Types d'organisation**
Thématique	Thèmes génériques et spécifiques.
Dialectique	Organisation temporelle et structuration argumentative.
Dialogique	Représentation énonciative et modalisation.

L'application de calculs textométriques permet de relever les contrastes entre les sous-corpus et de rendre compte des différences entre ces derniers aussi. Dans ce sens, elle permet une « lecture » des deux sous-corpus contrastés. Par le biais des opérations d'assimilation et de dissimilation, l'analyste observe la récurrence et la co-récurrence de sèmes et fait une première interprétation des résultats, observant les contrastes d'un sous-corpus à l'autre. Ensuite, l'analyste peut réorganiser, par processus interprétatif, les contenus sémantiques présents dans chaque sous-corpus séparément. Les opérations d'inhibition et d'activation de sèmes construisent les régularités pertinentes, et organisent les contenus sémantiques en

représentations de niveaux supérieurs, soit les isotopies, les isosémies et les molécules sémiques. L'analyse identifie les données textuelles linguistiquement pertinentes, c'est-à-dire qui ont un emploi uniforme dans l'ensemble de contextes, en utilisant le concordancier disponible dans les logiciels textométriques (Eensoo et Valette, 2012, 2014a, 2014b, 2015).

L'observation des régularités sémantiques construit des hypothèses sur l'interprétation elle-même et les valide au fur et à mesure, en opérant un va-et-vient entre les données textuelles observées et le contexte où elles apparaissent. Elle permet de décrire les sous-corpus de manière contrastive et de comprendre leurs particularités, respectant le principe différentiel de la théorie selon lequel le sens est fait de différences. L'analyse des critères et leur catégorisation est en consonance avec la perspective théorique de la sémantique interprétative, laquelle définit l'activé interprétative dans les termes d'une « conjecture rationnelle », soit une hypothèse qui ne reçoit pas de démonstration définitive, mais qui s'appuie sur la stabilisation des indices trouvés, d'autant plus que ceux-ci sont hétérogènes et peuvent en coalition agir sur la vraisemblance d'une lecture, parmi d'autres qui sont possibles (Rastier, 1996a, p. 19).

Les récurrences sémiques identifiées lors de l'interprétation sont décrites par l'analyse, c'est-à-dire les isotopies, les isosémies et les molécules sémiques. Sur le plan thématique, les isotopies attestent principalement les thèmes génériques (les domaines et les taxèmes des sous-corpus) et les molécules sémiques, les thèmes spécifiques. Les isosémies décrivent particulièrement les composantes dialectiques et dialogiques, et les critères considérés comme relevant de ces composantes sont dans la majorité des cas les mots grammaticaux, comme les pronoms, les adverbes, etc. Les concepts exposés plus haut à l'égard des composantes sémantiques offrent des éléments conceptuels pour décrire les régularités rencontrées dans le corpus.

D'un point de vue théorique, il n'y a pas de catégories grammaticales exclusives à une composante. Le sème n'est pas une propriété du référent et potentiellement un morphème ou un mot pourrait actualiser un sème exprimant une propriété dialectique (nous avons vu l'exemple du sème inchoatif dans le vers « l'aube allume la source » du poète Paul Éluard à la page 163). Cependant, la démarche de catégorisation proposée par les travaux analysés établit une correspondance entre certaines catégories morphosyntaxiques et les composantes sémantiques. Par exemple, les connecteurs logiques sont généralement associés à une

structuration sur le plan dialectique des textes. Au niveau thématique, les mots lexicaux aident à caractériser les textes dans le point de vue des thèmes abordés. De même, les pronoms sont plus susceptibles d'être catégorisés dans la composante dialogique, puisque leur fonction est assez homogène et ils font état de la position de l'énonciateur face à son propos.

Cependant, la correspondance de certaines catégories morphosyntaxiques aux composantes dans le processus de catégorisation ne préjuge pas le type de données textuelles qui peuvent mieux décrire une composante. Cela révèle une autre compatibilité de la SI avec la textométrie. Selon la sémantique interprétative, le sens ne se trouve pas dans des structures localisables du texte, il se manifeste diffusément dans tout le texte, et même les morphèmes peuvent être utilisés pour décrire le sens textuel. Puisque la méthode de spécificité fait émerger les données textuelles pertinentes en fonction du principe de relation entre les sous-corpus et le corpus de référence, elle permet de révéler des données textuelles qui seraient élaguées par d'autres techniques de filtrage du vocabulaire qui sont communément appliquées à la fouille d'opinions, comme les pronoms, les négations ou même la ponctuation. Dans le cas de la textométrie, ces données ont leur importance reconsidérée. Valette (2004) a démontré que ces variables, dites de « bas niveau » (Rastier, 2006a ; Pincemin, 2012a), et qui peuvent inclure d'ailleurs des données comme la case typographique, des balises HTML et des URL, peuvent devenir d'importants facteurs de caractérisation et description interprétatives.

La description comparative des régularités thématiques, dialectiques et dialogiques rencontrées permet d'identifier les stratégies argumentatives utilisées par les auteurs de chaque sous-corpus constitué et confère une valeur ajoutée dans le cadre général de la démarche méthodologique : elle démontre la pertinence des critères textuels retenus dans la mesure où ces derniers contribuent à expliciter la démarche interprétative de l'analyste lors de la sélection de critères. Elle est aussi compatible avec le principe de la SI selon lequel le sens n'est pas une propriété du signifiant, il est construit par l'activité interprétative.

5. Conclusion du chapitre

Dans ce chapitre, nous avons présenté le cadre théorique de la sémantique interprétative, les concepts principaux et l'application de ces concepts dans la démarche de fouille d'opinions. Nous avons décrit la relation entre les concepts structuraux et les démarches méthodologiques associées à la constitution du corpus et aux calculs textométriques. Nous avons également expliqué les concepts descriptifs de la SI, en les reliant à la démarche de sélection de critères textuels interprétables et compatibles avec les composantes sémantiques.

La textométrie, méthode de la linguistique de corpus calquée sur des calculs différentiels, s'affirme compatible avec les concepts de la SI. La méthode de spécificité permet de faire émerger les données globales surreprésentées de chaque sous-corpus, en tenant compte de l'ensemble du corpus de référence. La méthode de cooccurrence permet de voir les associations sémantiques locales, en analysant les voisinages des mots communs et fréquents qui traversent les deux sous-corpus. De ces calculs comparatifs et contrastifs, les données textuelles émergent, permettant à l'analyste de détecter les récurrences de traits sémantiques (isotopies) et les associations sémantiques (molécules sémiques) et de construire les critères textuels pertinents, par processus interprétatif.

Nous proposons la figure 7 ci-dessous, inspirée du schéma créé par Gérard (2004), pour résumer les concepts traités dans ce chapitre, en rapport avec la méthodologie de fouille d'opinions adoptée par les travaux analysés. Les concepts sont placés sur deux plans dans l'axe vertical, celui des « concepts structurels » et celui des « concepts descriptifs ». Dans la partie sur les concepts structurels se trouvent les processus et les opérations interprétatives que nous avons traités dans ce chapitre (p. 127) (assimilation/dissimilation ; inhibition/activation). Dans la partie sur les concepts descriptifs se trouvent les concepts correspondant aux éléments objectivés par les opérations interprétatives, que nous avons aussi traités ici (p.147). Dans la figure 7, nous avons également distribué les concepts sur l'axe horizontal en fonction de trois plans : le plan perceptif, le plan interprétatif et le plan descriptif.

Nous avons défini ces plans pour signaler une progression de l'activité interprétative : elle commence par une perception des différences entre le fond sémantique (les récurrences

globales des sous-corpus) et les formes sémantiques (les récurrences thématiques locales des sous-corpus) par les opérations d'assimilation et de dissimilation de sèmes. Ensuite, les opérations interprétatives d'inhibition et activation de sèmes permettent d'identifier les isotopies, les isosémies et les molécules sémiques. Finalement, sur le plan descriptif, les isotopies, les isosémies et les molécules sémiques identifiées sont décrites en rapport avec les composantes sémantiques : les thèmes génériques et spécifiques (composante thématique) l'organisation temporelle et argumentative (composante dialectique) et l'énonciation et la modalisation (composante dialogique).

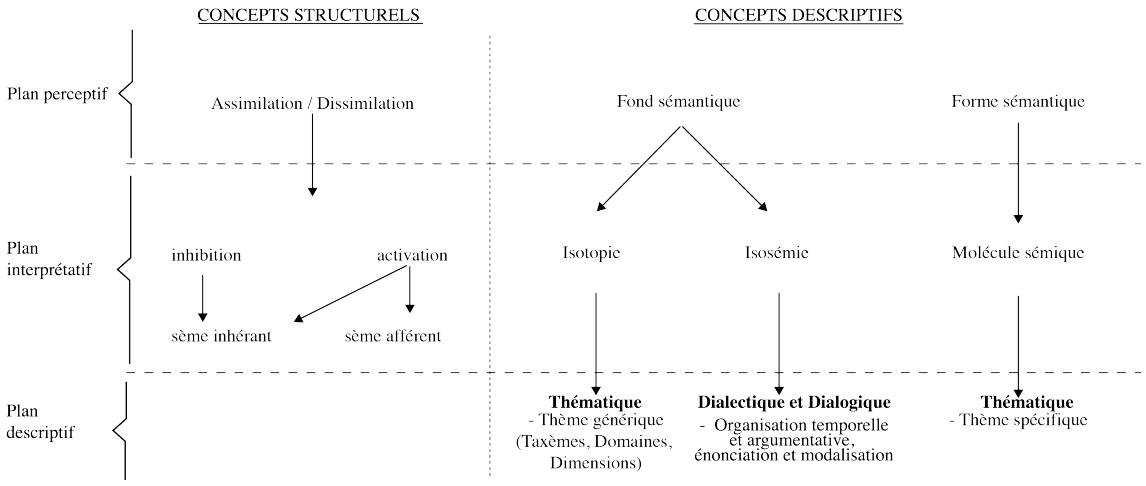


Figure 7. Les plans de la SI (inspirés de Gérard, 2004, Chapitre 1, p.3)

Encore dans la figure 7, les flèches signalent les liens de correspondance entre les concepts. Elles traversent les plans, indiquant une corrélation entre les concepts et les plans. Par exemple, à l'opération assimilation/dissimilation sur le plan perceptif correspondent les opérations d'inhibition/d'activation sur le plan interprétatif, permettant d'actualiser les sèmes inhérents et afférents. De même, le fond sémantique est corrélé aux isotopies et isosémies, et les formes sémantiques sont corrélées aux molécules sémiques. Sur le plan descriptif, chacun de ces éléments permet de décrire chacune des composantes sémantique, la thématique, la dialectique et la dialogique.

Afin d'illustrer l'agencement entre la démarche de sélection de critères textométriques pour la fouille d'opinions et les concepts de la sémantique interprétative, nous proposons le

tableau de correspondance ci-dessous (tableau 6). Nous associons les démarches méthodologiques aux concepts structurels et descriptifs exposés dans cette section. La colonne « Observables linguistiques » présente la résultante de la démarche interprétative et signale la progression de la première à la dernière étape, où est marquée l'obtention de critères textuels interprétables, c'est-à-dire des critères qui explicitent les stratégies argumentatives des auteurs de chaque sous-corpus sur le plan thématique, dialectique et dialogique.

Tableau 6. Correspondance entre les concepts de la SI et les démarches textométriques

Étapes	**Démarche méthodologique**	**Opérations**	**Concepts**	**Observables linguistiques**
Étape 1	Calculs contrastifs	Dissimilation et assimilation	Fonds sémantiques et formes sémantiques	Données textuelles spécifiques
Étape 2	Identification des structures sémantiques	Inhibition et activation de sèmes	Isotopies, isosémies et molécules sémiques	Critères textuels
Étape 3	Description des composantes	Interprétation	Thématique, Dialogique et Dialectique.	Critères textuels interprétables

Dans le chapitre suivant, nous allons formaliser la démarche méthodologique de recherche de fouille d'opinions qui a été appliquée dans le cadre de notre recherche.

Chapitre 3 : Méthodologie

1. Introduction

La méthodologie de fouille d'opinions réalisée dans cette recherche s'inscrit dans la famille des méthodes supervisées, basées sur les techniques d'apprentissage automatique. Elle s'appuie sur une démarche classique de fouille de textes (Forest, 2009 ; Forest et coll., 2009), en reprenant de celle-ci les étapes principales. Elle se fonde aussi sur des travaux expérimentaux dans le domaine de la fouille d'opinions (Eensoo et Valette, 2012, 2014a, 2014b, 2015; Valette, 2004), et comprend une étape préalable de sélection de critères textuels, par le truchement de calculs textométriques. Elle inclut par ailleurs une deuxième étape d'apprentissage sur les critères textuels sélectionnés pour la création d'un classifieur. Par ces démarches, nous voulons démontrer la pertinence de la sélection de critères textuels pour distinguer les articles qui sont thématiquement analogues, mais qui véhiculent des opinions divergentes.

Dans la deuxième section de ce chapitre, nous allons présenter la démarche de fouille de textes sur laquelle se base la méthodologie de fouille d'opinions proposée. Cette démarche sera présentée de manière synthétique, en expliquant davantage le processus d'apprentissage de la méthode supervisée. La méthodologie de fouille d'opinions repose sur cette démarche générale, mais compte certaines particularités qui incluent l'analyse textométrique pour la sélection de critères textuels.

Dans la troisième section, nous allons expliquer les démarches méthodologiques particulières qui seront adoptées, afin de répondre aux deux premières questions de notre recherche, à savoir, *quels sont les critères textuels permettant de caractériser et de distinguer les opinions exprimées dans les textes sur le plan thématique, dialectique et dialogique* et *quels types de critères issus de l'analyse textométrique entamée à la question 1 sont les plus performants pour prédire la classe des articles d'opinion lors d'un processus de classification automatique*. Ces questions correspondent à des étapes précises de la méthodologie de fouille d'opinions qui seront présentées. Nous allons également expliquer les calculs statistiques

relevant de la textométrie qui nous permettront de sélectionner les critères textuels pour la fouille d'opinions. À chaque calcul correspond également un type de critère : les critères globaux sont obtenus par le calcul des spécificités et les critères locaux par le calcul de cooccurrence. Cette typologie, ainsi que le fonctionnement des calculs seront expliqués dans cette section. Il sera question également des algorithmes d'apprentissage et des démarches permettant d'évaluer la performance des critères sélectionnés dans la prédiction du type d'opinion véhiculé par les articles.

Notre corpus de référence a été constitué d'un ensemble d'articles d'opinion au sujet de la Grève des étudiants de 2012 au Québec, tous publiés dans les journaux québécois. Ils ont été classés dans des classes opposées par la chercheuse selon un plan élaboré à cette fin. Les classes élaborées pour l'expérimentation reflètent l'opposition et l'antagonisme de groupes d'auteurs sur la discussion au sujet de la grève des étudiants de 2012 contre la hausse des frais de scolarité au Québec. Cette classification a été validée par une évaluation interjuges, et les résultats seront présentés dans la troisième section.

Les démarches pour atteindre le deuxième objectif de recherche, soit vérifier la pertinence de la démarche dans le contexte des SRAP, seront déclinées dans la quatrième section du présent chapitre. Nous proposons de diviser notre corpus en différents sous-corpus, chacun correspondant à un intervalle chronologique. Avec ce découpage et à l'aide de techniques textométriques, nous voulons explorer à quel moment de la controverse sur la grève étudiante nous pouvons observer l'apparition des critères textuels dont la pertinence et l'efficacité ont été démontrées à l'étape de la classification avec la méthode supervisée. Nous allons également tester la performance de l'ensemble de critères retrouvés dans le début de la controverse pour prédire la classe des articles qui ont été publiés postérieurement. En supposant que les critères textuels les plus pertinents pour la tâche de classification font leur apparition de manière anticipée dans le corpus et permettent d'atteindre un bon résultat de classification, il est possible de statuer sur l'applicabilité de la démarche de sélection de critères textuels dans un contexte de recommandation en temps réel, puisqu'il serait possible de développer un classifieur à partir des premiers articles parus dans la controverse.

2. Démarche méthodologique de fouille de textes

La démarche informatique que nous adoptons ne diffère pas fondamentalement de la démarche typique de plusieurs travaux en fouille de textes (Forest, 2009 ; Forest et coll., 2009) et repose sur le modèle vectoriel (Memmi, 2000 ; Salton, 1988). Ce modèle permet de transformer les textes du corpus en un ensemble de données textuelles structurées pour les rendre analysables par les machines. La démarche de fouille de textes comprend cinq étapes : 1) la constitution d'un corpus de textes ; 2) le filtrage ; 3) la transformation vectorielle ; 4) l'application des algorithmes (fouille) ; et 5) l'évaluation, l'interprétation et l'intégration. Nous allons expliquer dans les sections suivantes de façon sommaire chaque étape de la démarche méthodologique de la fouille de textes avant d'introduire celle de la fouille d'opinions qui sera adoptée dans le cadre de cette recherche.

2.1 Constitution du corpus

La constitution d'un corpus de textes est la première étape du processus de fouille de textes. Elle consiste à collecter un ensemble de textes pertinents pour la réalisation de la tâche. À cette étape, il faut spécifier les stratégies et les critères pour rassembler les textes. Selon Forest et coll. (2009) le choix de textes doit tenir compte des caractéristiques 1) générales (provenance, taille, date de création, etc.) ; 2) technologiques (support, format, etc.) ; 3) informationnelles (thèmes) ; 4) linguistiques (langue, genre, registre, etc.).

2.2 Filtrage

L'analyse statistique de textes repose sur des règles de comptage et de comparaison entre des variables discrètes provenant de la chaîne textuelle, que nous appelons ici « données textuelles ». Ces données sont obtenues à l'aide de techniques de segmentation appliquées aux textes numérisés. L'étape de filtrage comprend les techniques de segmentation, mais aussi des techniques qui permettent de discriminer les données textuelles les plus représentatives, communément appelées « traits discriminants ». À cette étape, il est préférable de définir une norme permettant d'isoler les données textuelles qui ont un poids informatif discriminant pour la tâche de fouille afin de réduire le nombre de variables linguistiques intervenant dans le traitement (Lebart et Salem, 1994).

Le filtrage comprend plusieurs sous-opérations de nature linguistique ou statistique. D'une part, elles servent à segmenter les textes et éliminer du lexique les mots qui ne sont pas pertinents pour la tâche. Dans l'étape de filtrage, les mots grammaticaux tels que les articles et les prépositions peuvent être éliminés, car leurs hautes fréquences et distributions dans les textes diminuent leur poids discriminant. Des ressources linguistiques, comme les « anti-dictionnaires », employés pour répertorier ces mots grammaticaux, sont utilisées pour les filtrer et les éliminer. D'autre part, le filtrage comprend des techniques visant à réduire la taille des données, éliminant celles dont la fréquence dépasse certains seuils (qui sont déterminés de façon heuristique) et en gardant celles qui sont les plus discriminantes. Des techniques telles que TF-IDF permettent d'éliminer les données textuelles qui sont fréquentes, mais dont la distribution est très uniforme, n'ayant pas pourtant un poids discriminant dans le corpus. Des analyseurs morphosyntaxiques comme les EPD peuvent également être utilisés à cette étape pour rassembler les variantes syntaxiques de mots qui ont la même racine sémantique. Ce processus est connu comme lemmatisation et permet d'obtenir des lemmes, qui sont des unités lexicales minimales regroupant les variantes d'un mot. Par exemple, les fréquences de toutes les conjugaisons d'un verbe sont ramenées à un seul lemme du verbe à l'infinitif (Forest, 2009).

2.3 Transformation

L'objectif de l'étape de transformation est d'obtenir une représentation numérique des textes organisés en corpus en fonction des traits discriminants. La représentation de l'ensemble de textes numériques[24] *T* d'un corpus se fait dans un espace vectoriel comportant *n* dimensions, dont *n* est le nombre d'attributs de *T,* c'est-à-dire, le nombre de traits discriminants (*TD*) de l'ensemble *T*. Chaque texte numérique *t* est un vecteur qui comporte l'information sur la fréquence ou la présence des traits discriminants *TD* (*td1…tdn*). La figure 8 (Forest et coll., 2009) est un exemple schématique du modèle vectoriel de

[24] Nous utilisons le terme « texte numérique » ici pour généraliser les types de documents textuels qui peuvent être représentés dans le modèle vectoriel. Il peut s'agir d'articles, mais aussi de segments textuels, comme les phrases et les paragraphes extraits des documents.

représentation des textes numériques. Les lignes représentent les textes numériques (qui peuvent être des articles ou des segments textuels comme des paragraphes ou des phrases) et les colonnes représentent les traits discriminants. Dans les cellules, le symbole § représente une valeur indiquant soit l'absence/présence du trait discriminant (représentation binaire), soit leur fréquence absolue (valeurs entières).

Traits discriminants (TD)

Textes numériques (T)

	td1	td2	td3	td4	td5
t1	§	§	§	§	§
t2	§	§	§	§	§
t3	§	§	§	§	§
t4	§	§	§	§	§
t5	§	§	§	§	§

Figure 8. Matrice de textes (T)/traits discriminants (TD)
(inspiré de Forest et coll., 2009)

2.4 Fouille

Les opérations visant le développement de classifieurs sont réalisées à l'étape de la fouille, qui englobe un ensemble d'opérations destinées à structurer les informations contenues dans la matrice vectorielle pour des fins de classification. À cette étape s'effectue le choix d'un algorithme de fouille adéquat à la nature du problème posé. Dans le cadre de la méthode de classification supervisée, cet algorithme peut être, par exemple, le classifieur bayésien naïf (NB), l'arbre de décision, la machine à vecteurs de support (SVM) ou encore les réseaux neuronaux (Ibekwe-SanJuan, 2007). La méthode de classification supervisée appliquée à la classification automatique de textes consiste à prédire la classe d'un texte numérique, à partir de la mémorisation des propriétés d'autres textes qui ont servi d'exemple pour la construction du classifieur. L'apprentissage automatique s'opère dans une logique d'induction : « on

généralise les valeurs de classification à partir de l'observation d'un nombre limité d'exemples » (Ibekwe-SanJuan, 2007, p. 90). L'objectif du classifieur est d'atteindre une fonction de classification qui sera performante tout en minimisant les risques d'erreur. Les étapes de la méthode supervisée peuvent être résumées ainsi :

1. Constitution d'un corpus d'apprentissage (avec $^2/_3$ du corpus)
2. Choix d'un algorithme d'apprentissage (par exemple : bayésien naïf, machine à vecteurs de support, réseaux neuronaux)
3. Apprentissage sur les traits discriminants du corpus et construction du classifieur
4. Évaluation du modèle à partir d'un échantillonnage du corpus d'apprentissage
5. Validation du classifieur sur la partie du corpus qui n'a pas servi à l'entraînement (corpus de test, constitué par $^1/_3$ du corpus)

2.5 Évaluation, interprétation et intégration

La cinquième étape de la démarche réside dans l'interprétation, l'évaluation et l'intégration des résultats générés par les algorithmes. Dans le cas de la classification automatique, les résultats sont évalués en utilisant des mesures qui estiment l'efficacité du classifieur à prédire correctement les classes des textes numériques appartenant au corpus de test, qui n'a pas servi à l'entraînement de l'algorithme.

Les mesures d'évaluation déterminent le niveau de justesse de la classification en considérant le nombre de textes numériques correctement affectés et correctement non affectés à une classe (taux de vrais positifs et taux de vrai négatifs) ainsi que le nombre de textes numériques erronément affectés ou non affectés à une classe (taux de faux positifs et taux de faux négatifs) (Sokolova et Lapalme, 2009). Les mesures classiques d'évaluation de tâches de classification de textes sont le rappel, la précision et le *fscore* (Forest et coll., 2009 ; Sokolova et Lapalme, 2009).

Dans les expérimentations avec la méthode de classification supervisée, il est usuel de valider la fiabilité du classifieur en réalisant des tests avec le corpus d'apprentissage (Ibekwe-SanJuan, 2007). Pour les corpus expérimentaux plus petits, il est fréquent d'utiliser la validation croisée à k itérations, qui consiste à répartir le corpus d'apprentissage en k échantillons et à utiliser l'un des k pour le test, et les k-1 restants pour l'apprentissage. L'opération est répétée pour chaque k. La procédure de validation croisée permet d'estimer les erreurs survenant du classifieur, en observant si les mesures d'évaluation se maintiennent constantes avec le croisement d'échantillons (Liu, 2001). Une fois que le classifieur démontre une bonne performance, il est projeté dans le corpus de test constitué pour l'application. Nous pouvons garantir de cette façon un plus grand niveau de fiabilité du classifieur.

Nous avons vu dans cette section l'approche générale de fouille de textes, utilisée dans plusieurs applications et destinée à assister l'analyse et la classification de textes numériques. La fouille d'opinions relève de la fouille de textes, dans le sens qu'elle porte aussi sur l'analyse sur des données textuelles et utilise plusieurs traitements communs de cette dernière. Dans la section suivante, nous allons présenter les étapes générales de notre démarche méthodologique de fouille d'opinions, qui de manière globale suit le même enchaînement d'étapes et de tâches que la fouille de textes. Nous donnerons également plus de détails au sujet de la segmentation de textes, de l'élaboration de traits discriminants dans le cadre de notre recherche (que nous appelons ici « critères textuels ») ainsi que des mesures d'évaluation qui seront utilisées dans l'étude.

3. Démarche méthodologique de fouille d'opinions

Les étapes de la démarche méthodologique de fouille d'opinions dans le cadre de notre recherche correspondent à des étapes précises de la démarche de fouille de textes présentée dans la section précédente. La figure 9 ci-dessous présente chacune des étapes impliquées dans notre démarche méthodologique. Dans notre cas, l'étape « Sélection de critères textuels » supplée celle du filtrage de la démarche de fouille de textes en proposant des techniques textométriques pour la sélection de critères textuels interprétables. Les critères sont dits interprétables, car leur sélection explicite les démarches de lecture interprétative opérées au sein du corpus, à la suite des résultats des calculs textométrique. Les critères textuels sélectionnés seront utilisés en entrée à l'étape « Transformation ». Le développement du classifieur s'effectue dans l'étape « Fouille », dans laquelle différents tests sont réalisés en variant le nombre et le type de critères, ainsi que les algorithmes d'apprentissage de la méthode supervisée, afin de produire un résultat optimal pour la tâche de classification. Le schéma de la démarche méthodologique de fouille d'opinions (figure 9) ajoute également des particularités concernant la constitution du corpus qui sont propres aux considérations théoriques de la SI. Ce corpus est constitué d'articles d'opinions sur un thème particulier (corpus de référence) et est par la suite organisé en différents sous-corpus, chacun constitué en fonction du type d'opinion véhiculée.

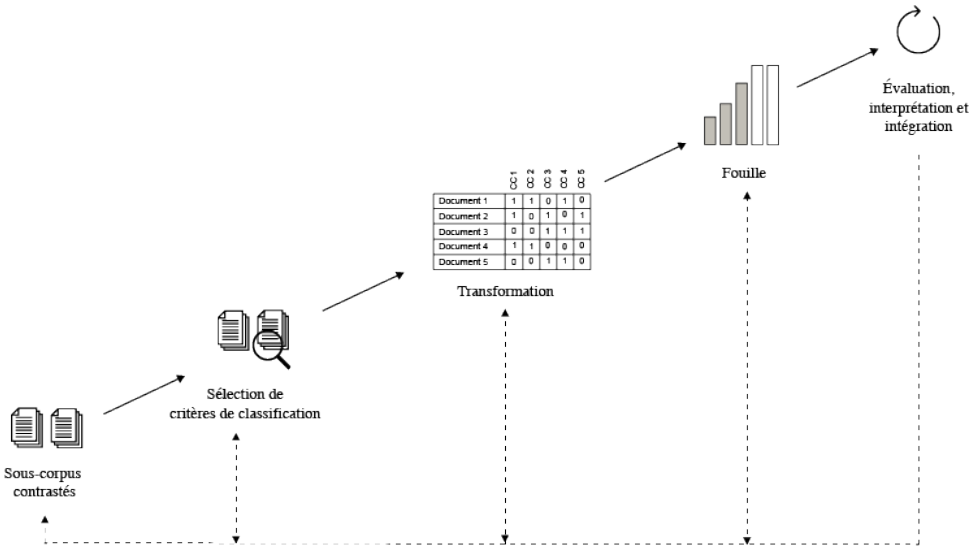


Figure 9. Démarche méthodologique de la fouille d'opinions

Nous allons détailler dans les prochaines sections les choix méthodologiques de chacune de ces étapes pour la réalisation de notre expérimentation.

3.1 Corpus de la recherche et constitution de sous-corpus contrastés

La constitution du corpus de la recherche et des sous-corpus contrastés constitue la première étape de la démarche méthodologique de la fouille d'opinions présentée dans cette recherche (figure 9, p.181)

La sémantique interprétative donne certaines orientations quant au choix du corpus : il n'est pas nécessaire qu'il soit représentatif de la langue générale, mais il doit refléter la pratique sociale dont les textes émanent (Pincemin, 1999a). La constitution de sous-corpus doit expliciter le critère de rassemblement des textes afin d'assister les démarches d'exploration interprétative.

Dans le cadre de cette recherche, nous avons circonscrit notre champ d'intérêt aux controverses débattues dans la presse écrite. Nous avons caractérisé l'opinion comme l'expression linguistique d'un acte communicationnel inscrit dans une pratique sociale institutionnalisée, le journalisme, et nous avons assimilé la controverse à un type particulier de débat d'opinions dans la presse, dont la particularité est la formation de camps explicitement opposés quant à la façon de voir et de comprendre une question d'importance pour la collectivité.

En fonction du type de phénomène qui nous intéresse ici, nous avons choisi pour notre expérimentation une controverse qui a occupé une place très importante dans les journaux québécois durant l'année 2012 : la mobilisation étudiante au Québec contre la hausse des droits de scolarité annoncée par le gouvernement de Jean Charest. Cette mobilisation, marquée par une intense polarisation entre deux camps — le gouvernement et les étudiants —, caractérise le type de conflit que nous appelons ici une controverse, c'est-à-dire des conflits ayant une structure triadique qui « renvoient à des situations où un différend entre deux parties est mis en scène devant un public, tiers placé dès lors en position de juge » (Lemieux, 2007, p. 195). L'aspect fort médiatisé du conflit étudiant a contribué à une intense discussion publique qui a même débordé de la question de la hausse des droits de scolarité, gagnant l'ampleur d'une crise sociale. Connu dans les médias comme le « printemps érable », le conflit

étudiant a mis en évidence de fortes oppositions sociales et politiques à l'intérieur de la société québécoise.

Les articles d'opinion provenant de la controverse sur la grève étudiante ont été rassemblés pour constituer le corpus de référence. Par la suite nous avons effectué une classification manuelle afin de caractériser ces articles selon le type d'opinion défendue. Nous avons constitué deux sous-corpus, chacun rassemblant les articles de chaque camp qui s'est opposé à l'autre dans la dispute. Les sections suivantes présentent les détails de ces démarches.

3.1.1 Thématique du corpus

Cette section a pour objectif de présenter la thématique du corpus de référence retenu dans le cadre de notre recherche. Des mots clés ont été ajoutés à chaque paragraphe pour aider au repérage des principaux thèmes.

Le début de la crise étudiante remonte au 18 mars 2011, quand le gouvernement libéral dirigé par Jean Charest a annoncé sa décision d'augmenter les droits de scolarité des universités québécoises de 2 168 $ à 3 793 $ par année en l'espace de cinq ans (soit une augmentation de 75 %). Réunis dans diverses assemblées, les étudiants ont voté, le 13 février 2012, pour une grève générale et illimitée dans les établissements d'enseignement supérieur. Depuis le début du conflit, leur objectif était très clair : l'abolition de la hausse des frais de scolarité. La grève, coordonnée principalement par les fédérations d'associations étudiantes québécoises (Coalition Large de l'Association pour une Solidarité Syndicale Étudiante (CLASSE) ; Fédération étudiante collégiale du Québec (FECQ) et Fédération étudiante universitaire du Québec (FEUQ), a touché plus de 400 000 étudiants postsecondaires québécois. Par son ampleur, sa durée et par la quantité de mobilisations, elle est considérée comme la plus importante grève étudiante du Québec (Blouin Genest, 2012a). [Grève étudiante ; droits de scolarité ; augmentation ; hausse ; grève générale ; grève illimitée ; enseignement supérieur ; Québec, gouvernement ; 13 février 2012 ; CLASSE ; FECQ ; FEUQ ; universités ; cégeps]

De nombreux moyens de pression ont été utilisés par les étudiants québécois pour forcer le gouvernement à suspendre la hausse des droits de scolarité. En plus des piquets de

grève organisés dans les établissements d'enseignement, les manifestations nationales (tenues le 22e jour de chaque mois) ont réuni plus de 200 000 personnes dans les rues de Montréal. Malgré la forte mobilisation, le gouvernement Charest refusait de négocier avec les étudiants sur la décision d'augmenter les droits de scolarité, argumentant que les universités québécoises se trouvaient dans un état critique de sous-financement. Les étudiants ont continué les manifestations pour sensibiliser l'opinion publique et pour dénoncer la position du gouvernement, qui refusait le dialogue depuis le début du conflit. [Piquets de grève ; manifestations ; Montréal ; manifestations nationales du 22 ; financement ; opinion publique]

Symbole de la résistance aux politiques néolibérales, le carré rouge, d'abord adopté par les étudiants en grève contre les compressions de l'aide financière en 2005 refait son apparition au printemps 2012 sur les vêtements, sacs, fenêtres et édifices de plusieurs partisans. Des mouvements s'opposant à la grève et en faveur de la hausse adoptent le carré vert pour exprimer leur point de vue. D'autres symboles ont été également adoptés, comme le carré blanc pour un armistice entre les étudiants et les grévistes, le carré jaune, en faveur d'une hausse, mais sur une période plus longue, et le carré bleu, refusant la hausse des frais de scolarité, mais refusant aussi la grève (Bonenfant et coll., 2013) [Carré rouge ; carré vert ; carré blanc ; carré jaune ; carré bleu].

La grève étudiante a été marquée par une construction médiatique très polarisée, qui a mis en évidence une forte opposition entre les deux camps en désaccord, soit les étudiants et le gouvernement. S'affranchissant peu à peu du problème de l'augmentation des droits de scolarité, l'opposition entre ces deux groupes a mis plusieurs questions à l'ordre du jour. On a alors assisté à la naissance d'un débat de fond sur le système d'éducation, sur le droit à la mobilisation, sur la notion de démocratie et sur l'avenir de la société québécoise (Blouin Genest, 2012b). Contrastant le discours du gouvernement et des grévistes, Julien (2012) montre une différence très accentuée dans la façon dont les parties du conflit s'exprimaient dans les médias :

> (...) les parties au conflit s'exprimaient en des termes suffisamment étrangers l'un à l'autre pour empêcher tout point de rencontre, notamment concernant leurs conceptions de l'éducation, des services publics et du processus démocratique (Julien, 2012, p. 152).

[Couverture médiatique ; polarisation ; système d'éducation ; droit à la mobilisation, démocratie ; société québécoise ; avenir]

En ce qui concerne le débat sur l'éducation, le choc d'idéologies a opposé d'une part un discours centré sur la vision de « l'utilisateur-payeur », défendu par le gouvernement, et d'autre part une vision de l'éducation comme un fondement de la société, soutenue par les étudiants en grève. Essentiellement, ces deux perspectives divergeaient à propos de la participation de l'État dans la conduite des services publics : d'un côté, une vision néolibérale prônait l'instrumentalisation de ces services par les lois du marché et, de l'autre côté, une vision sociale-démocrate soutenait la vision d'un État carrément garant de l'accessibilité à ces services. La CLASSE a joué un rôle prépondérant à l'intérieur de ce conflit, militant pour la gratuité scolaire au niveau universitaire, vision qui ne faisait pas consensus parmi les étudiants en grève. [Conception/philosophie de l'éducation ; utilisateur-payeur ; néolibéralisme ; sociale démocratie ; gratuité scolaire ; CLASSE]

Outre la question de l'éducation, le mouvement étudiant a également suscité un vif débat sur le droit à la mobilisation. En argumentant que les étudiants n'étaient pas des « employés » des universités, le gouvernement Charest refusait de nommer le mouvement de « grève », le qualifiant plutôt de « boycott ». Cette stratégie a remis en question la légitimité démocratique du mouvement, en créant une tension entre les étudiants dans les assemblées qui revendiquaient la reconnaissance de leurs droits collectifs, et les étudiants dissidents qui évoquaient leurs droits individuels d'assister aux cours pour lesquels ils avaient payé. Encouragés par ce discours, plusieurs étudiants opposés à la grève ont réussi à obtenir auprès des tribunaux des injonctions légales pour forcer la tenue de certains cours (Julien, 2012 ; Langlois, 2012). Ces recours juridiques n'ont toutefois pas continué, en raison de l'intense mobilisation et des piquets de grève organisés par les grévistes. [Droit de mobilisation, droit individuel ; droits collectifs ; injonctions ; grève ; boycottage ; recours juridiques]

La difficulté du gouvernement à établir un dialogue avec les étudiants a suscité des réactions diverses, incluant d'autres mouvements de protestation comme des artistes, écologistes, féministes et souverainistes. La manifestation du 22 avril, par exemple, organisée auparavant par des artistes écologistes du Jour de la Terre, a rassemblé une foule de plus de 250 000 personnes autour de la cause étudiante, ce qui démontre l'inclusion d'autres questions

citoyennes, comme le débat environnemental autour du dossier Plan Nord, qui prévoyait la construction de mines dans des régions nordiques. Nous retrouvons également la dénonciation des scandales de corruption impliquant le gouvernement québécois et l'industrie de la construction (Bonenfant et coll., 2013) [Mouvement de protestation ; artistes ; écologistes ; féministes ; souverainistes ; Jour de la Terre ; Plan Nord ; dénonciation de la corruption ; gouvernement Charest]

En réaction aux moyens de pression des étudiants et des manifestants et pour forcer le retour en classe, l'Assemblée nationale du Québec a voté en mai 2012 le projet de loi 78 qui, en plus de suspendre la session pour les étudiants en grève, imposait des distances minimales pour la tenue de piquets de grève près des institutions d'enseignement. La loi a également imposé des restrictions aux manifestations comptant plus de 50 participants. Dénoncée par plusieurs organisations, dont le Barreau du Québec, Amnistie internationale et le Conseil des droits de l'homme de l'Organisation des Nations Unies, la loi 78 a intensifié le nombre de manifestations nocturnes dans les rues de Montréal, connues comme « manifs de casseroles ».[Projet de loi 78 ; piquets de grève ; Barreau du Québec, Amnistie internationale ; Conseil des droits de l'homme de l'ONU ; manifestation de casseroles]

Le conflit étudiant a pris fin avec l'élection du gouvernement minoritaire dirigé par la chef du Parti Québécois, Pauline Marois, le 4 septembre 2012, et par l'annulation par décret de la hausse des droits de scolarité. La loi 78 a aussi été abrogée par décret par le gouvernement péquiste. Le Parti Québécois a organisé en 2013 le Sommet sur l'enseignement supérieur pour élaborer les principes directeurs des universités québécoises. [Élections ; Parti Québécois ; Sommet sur l'enseignement supérieur]

En 2014, avec les compressions budgétaires de la part du gouvernement libéral dirigé par Philippe Couillard, lesquelles ont été imposées aux universités, les dirigeants de l'Université McGill, de l'Université de Montréal (UdeM) et de l'Université du Québec à

Montréal (UQAM) ont indiqué qu'un déficit budgétaire est inévitable, ce qui fait du débat sur le financement et sur l'accessibilité de l'éducation un enjeu toujours important et actuel.[25]

La discussion médiatique autour de la grève étudiante présente les caractéristiques de ce qui configure une controverse, au sens adopté dans le cadre de notre recherche. Elle constitue un conflit triadique, qui oppose d'une part les étudiants grévistes et d'autre part le gouvernement et ses partisans. Puisque l'enjeu de la grève a été amplement diffusé par les médias, les deux groupes qui s'opposaient dans le débat cherchaient constamment à rallier le public. Sur le fond, la question qui traversait le débat sur la grève était la vision de l'éducation ainsi que le rôle de l'État et de l'individu à l'égard du financement de sa formation professionnelle. Comme dans d'autres types de controverses, cet objet de débat n'a pas fait consensus, mais a suscité un mouvement très important qui a débouché sur des élections et qui a changé le parti au pouvoir.

3.1.2 Stratégie de rassemblement des textes et classification manuelle du corpus

Dans la constitution du corpus, seulement les genres de l'opinion ont été retenus. Nous avons présenté le modèle de Grosse (2001), qui traite de l'évolution des genres textuels dans la presse, pour montrer que le genre de l'opinion est apparu dans le sillage d'un processus historique d'évolution des genres informatifs (les « hard-news », le récit, le reportage, etc.). Grosse distingue le genre de l'opinion du genre de l'information, qui ont une visée plus informative et plus axée sur le rapport de faits : le genre de l'opinion est caractérisé par l'emploi des figures rhétoriques et des techniques d'argumentation pour persuader les lecteurs d'un point de vue déterminé.

[25] « Les universités s'attendent à un déficit budgétaire ». Article paru dans le Journal de Montréal en date du 24/10/2014http://www.journaldemontreal.com/2014/10/24/les-universites-sattendent-a-un-deficit-budgetaire

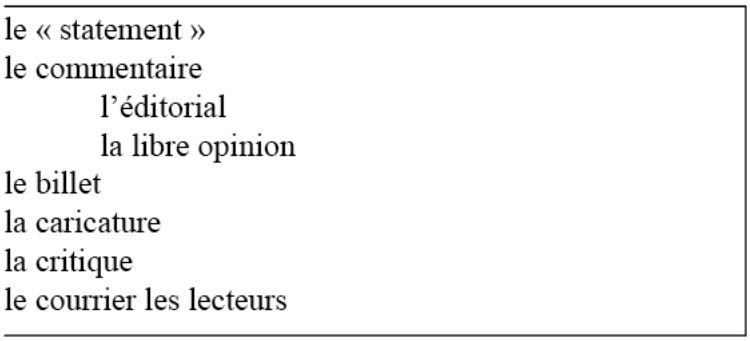


Figure 10. Les sous-genres de l'opinion (extrait de Grosse, 2001)

Nous avons retenu de ce modèle les sous-genres suivants : Éditorial, Article d'opinion et Courrier des lecteurs. Nous nous référons à l'Office québécois de la langue française (OLF) et à la Banque de données terminologique et linguistique du gouvernement du Canada (Termium) pour la définition de chacun de ces sous-genres :

1. Éditorial (domaine : journalisme/contenu du journal) : 1) Article écrit par une personnalité importante du journal et engageant la responsabilité morale de l'équipe tout entière ; 2) Article (généralement de première page) signé ou non signé, mais qui exprime le point de vue collectif du journal représenté par le rédacteur en chef (*editor-in-chief* en anglais). (Office québécois de la langue française [OLF], 2013)

2. Courrier des lecteurs (domaine : journalisme/contenu du journal) : Lettres envoyées par des lecteurs et publiées en raison de leur intérêt général. (Office québécois de la langue française [OLF], 2013)

3. Article d'opinion : Un article d'opinion est comme un essai persuasif. L'auteure ou l'auteur a une opinion ou un point de vue sur une question quelconque et désire convaincre la lectrice ou le lecteur du bien-fondé de son argument. (Bureau de la Traduction du Canada, s.d.)

Nous présentons au tableau 7 ci-dessous les caractéristiques 1) générales, 2) technologiques, 3) linguistiques et 4) informationnelles des articles du corpus (Forest et coll., 2009).

Tableau 7. Caractéristiques du corpus (inspiré de Forest et coll., 2009)

Caractéristiques	**Choix**	**Justification**
1. Générales (Genre, sources)	Articles d'opinion publiés dans des journaux : Éditorial, Libre opinion, Courrier des lecteurs.	Retenir les articles qui présentent clairement un positionnement à l'égard de la controverse.
2. Technologiques	Articles en format numérique (HTML ou XML).	Faciliter les traitements automatiques.
3. Linguistiques	Français.	La majorité des sources recensées dans la base *Eureka.cc* ayant traité de la crise étudiante sont en français.
4. Informationnelles	Thème : grève étudiante Date : 13 février 2012 (date du déclenchement de la grève des étudiants) jusqu'au 30 octobre 2012 (annulation par décret de la hausse de scolarité). Lieu : Québec.	Utiliser les articles publiés dans les journaux depuis le début du conflit jusqu'à la date de l'annulation de la hausse.

a) Requêtes

Le corpus nécessaire à notre recherche a été constitué à partir de requêtes lancées dans la base de données *Eureka.cc*. Nous avons premièrement élaboré une stratégie de recherche (tableau 8) comprenant des mots clés qui décrivaient de manière générale le thème du corpus (grève étudiante, boycottage et mobilisation étudiante, printemps érable) ainsi que l'enjeu débattu (frais de scolarité). Nous avons également ajouté les mots « Québec » et « Québécois » pour restreindre la requête à la grève du Québec. Avec cette stratégie, nous avons voulu atteindre un haut taux de rappel dans les résultats pour bien connaître le contenu de notre corpus et par la suite éliminer des articles non pertinents pour la tâche. La requête booléenne ci-dessous a été lancée sur le moteur. Nous avons sélectionné appliqué un filtre de type de sources, pour circonscrire nos articles à la catégorie « Éditorial et Opinion » du moteur

Eureka.cc et ne repérer que des éditoriaux, chroniques et lettres des lecteurs[26]. Cette stratégie de recherche a permis de repêcher un total de *452* articles.

Tableau 8. Requête 1

Filtres	**Requête**	**Limite chronologique**
SECTION= Éditorial et Opinion	TEXT= [[(grève OU conflit OU boycottage*OU crise OU mobilisation OU piquet*) $4 (étudiant*OU cours OU université+ OU cégep*)] OU [(droit*ou frais) $2 (scolarité OU éducation)] ET (Québec ou québécois*)] OU (‘printemps québécois’ OU ‘printemps érable’)	13 février 2012 à 30 octobre 2012

Nous avons réalisé une classification manuelle de la totalité des articles repêchés par cette première requête, afin de constituer les deux sous-corpus qui représentent les opinions divergentes sur la grève étudiante du Québec. Nous avons créé ainsi deux classes nommées ETUD et GOUV. La classe ETUD regroupe les articles qui sont favorables à la grève tandis que GOUV regroupe les articles qui sont contre la grève (les détails de cette classification sont donnés à la page 195). Cela nous a permis de constater que certains articles du total considéré n’étaient pas pertinents pour notre tâche. Un total de 117 articles des 452 initiaux a été rejeté en fonction des observations suivantes :

1. Articles au sujet de la grève étudiante qui proposent des solutions pour la question du financement des études ou pour dénouer le conflit, mais qui ne prennent pas parti explicitement pour un camp de la dispute : cela signifie que ce sont des articles

[26] L’outil d’indexation du site *Eureka.cc* catégorise les articles selon le balisage provenant des documents numériques fournis par les journaux. Pour Éditoriaux et Opinion, l’indexation utilise l’information présente dans le nom du cahier. Ainsi, tous les articles parus dans les cahiers Éditorial, Opinion, Lettres de lecteur, Point de vue Perspective, etc, sont indexés sous la rubrique qui nous a servi de filtre.

difficiles à trancher parce que leurs propos ne favorisent ni un camp ni l'autre de la dispute. Notre idée initiale était de classer ces articles dans une classe appelée « recadrage », dans le but d'explorer une troisième classe dans la démarche méthodologique. Mais le taux d'articles initialement classés dans recadrage était très faible par rapport aux autres classes (41 articles soit 9 % du corpus). Ainsi, nous avons pris la décision de les rejeter.

2. Discussions politiques très écartées sur la grève ou qui parlent de la grève étudiante de façon secondaire : plusieurs articles qui parlent de la succession électorale ont été rejetés, car ils traitent de la question étudiante en marge du thème principal. Il en va de même pour certains articles qui citent la question étudiante à l'intérieur d'un autre sujet, mais sans aborder les enjeux propres à la grève.

3. Documents contenant plusieurs articles sur des sujets différents : plusieurs lettres des lecteurs repérées dans *Eureka*.cc sont rassemblées dans un seul document, ce qui empêche la catégorisation du document entier, car chaque lettre a un positionnement différent ou traitent d'un sujet différent. Nous avons pris la décision de rejeter ces articles.

À la suite de la classification manuelle des 335 articles restants, nous avons constaté un déséquilibre dans la répartition d'articles favorables et défavorables à la grève étudiante, avec 215 articles pour les premiers et 120 articles pour les seconds.

Très peu de phénomènes présentent des corpus parfaitement équilibrés entre les classes. Ce problème propre aux études sur les algorithmes d'apprentissage est connu comme *Imbalance problem* (He et Garcia, 2009) et peut avoir un impact direct sur la précision des classifieurs : puisque les algorithmes construisent des classifieurs et les valide à partir de corpus qui ne sont pas représentatifs de la distribution empirique du phénomène étudié, les règles d'induction qui servent à prédire les objets de la classe minoritaire peuvent s'avérer plus faibles, car le nombre de traits ayant servi à l'apprentissage est moindre comparativement à la classe majoritaire. Cela est particulièrement le cas de certains corpus médicaux, comme le cas de l'incidence d'une maladie rare sur une population, dont le ratio peut atteindre un déséquilibre significatif de l'ordre de 100:1, 1000 : 1 ou 10000:1. Il s'agit là de déséquilibres

extrêmes dont le chercheur doit tenir compte lors de l'échantillonnage du corpus et de la validation de son classifieur.

Dans le cas de notre recherche, le déséquilibre retrouvé n'est pas considéré comme extrême, se situant dans un ratio de 1 : 2. Par contre, nous ne pouvons pas déterminer si lors de l'étape d'analyse textométrique, ce déséquilibre peut avoir un impact sur le nombre ou sur la qualité des critères textuels qui seront retenus pour les algorithmes d'apprentissage. Ainsi, nous avons pris la décision de construire des requêtes supplémentaires pour observer, d'une part, si le ratio de déséquilibre entre les classes persistait et, d'autre part, pour augmenter le nombre d'articles de notre corpus.

Deux autres stratégies de recherche sur la base *Eureka.cc* ont été élaborées. Nous avons d'abord exclu le filtre « Éditorial et Opinion » de notre requête et nous avons réalisé une recherche dans tous les autres types de sources de la base *Eureka.cc*. Nous avons par contre ajouté un filtre d'auteurs pour pouvoir sélectionner les articles écrits par un ensemble de chroniqueurs connus pour leur partisannerie progrève et antigrève (tableaux 9 et 10). Nous avons pris soin de sélectionner, pour chaque requête, le même nombre d'auteurs. Cette stratégie était destinée à repérer les articles d'opinion qui auraient pu être indexés dans d'autres types de sources de la base (par exemple, sous la rubrique « Nouvelle »).

Tableau 9. Requête 2 : Groupe des pros-grève

Filtres	**Requête**	**Limite chronologique**
SECTION=Éditorial et Opinion AUTHOR=[(Stéphane Baillargeon OU Michel David OU Josée Blanchette OU Lise Payette OU Michelle Ouimet OU Rma Elkouri)]	TEXT= [[(grève OU conflit OU boycottage*OU crise OU mobilisation OU piquet*) $4 (étudiant*OU cours OU université+ OU cégep*)] OU [(droit*ou frais) $2 (scolarité OU éducation)] ET (Québec ou québécois*)] OU ('printemps québécois' OU 'printemps érable')	13 février 2012 à 30 octobre 2012

Tableau 10. Requête 3 : Groupe des antigrèves

Filtres	**Requête**	**Limite chronologique**
SECTION=Éditorial et Opinion AUTHOR=[(Richard Martineau OU Mathieu Bock Côté OU Sophie Durocher OU Joseph Facal OU J. Jacques Samson OU Alain Dubuc OU Lysiane Gagnon)]	TEXT= [[(grève OU conflit OU boycottage*OU crise OU mobilisation OU piquet*) $4 (étudiant*OU cours OU université+ OU cégep*)] OU [(droit*ou frais) $2 (scolarité OU éducation)] ET (Québec ou québécois*)] OU (‘printemps québécois’ OU ‘printemps érable’)	13 février 2012 à 30 octobre 2012

La Requête 2 (Groupe des progrèves) a retourné un total de 78 articles et la Requête 3 (Groupe des antigrèves), a repêché un total de 188 articles, soit un total de 266 articles. Nous avons ajouté ces derniers aux 335 déjà classés.

Nous avons par la suite procédé à la vérification des 266 articles repérés par les requêtes 2 et 3. Dans la catégorie des progrèves, pour laquelle nous avons attribué la classe ETUD, presque la moitié des articles ont été rejetés, car ils correspondaient aux critères d’exclusion que nous avons cités précédemment (p.191 et 192). Parmi les 78 articles retrouvés, un total de 34 a été rejeté ; 39 articles ont été retenus pour la classe ETUD et 5 articles pour la classe GOUV. Dans le corpus des antigrèves, correspondant à la classe GOUV, 188 articles ont été revus. Parmi ceux-ci, 69 ont été rejetés. La majorité des articles, soit 115, a été retenue dans la classe GOUV et 4 dans la classe ETUD. Des exemples d’articles classés dans les deux classes sont donnés dans l’Annexe A à la fin de ce document.

Nous avons exclu également les doublons retrouvés par les requêtes. Pour identifier les articles doublons, nous avons utilisé le coefficient Jaccard à partir d’un programme développé dans le langage Python (Moore, 2012).

3.1.3 Statistiques du corpus

Au total nous avons lu et classé 715 articles (doublons escomptés) (452 de la Requête 1 et 266 des Requêtes 2 et 3). Nous avons rejeté un nombre total de 220 articles (117 de la Requête 1 et 103 des Requêtes 2 et 3). Notre corpus comprend donc 495 articles classés soit dans ETUD, soit dans GOUV. Le tableau 11 ci-dessous présente le résultat de la classification.

Tableau 11. Résultat de la classification

Actions	**Nombre d'articles**
Classé sur ETUD	258
Classé sur GOUV	237
Rejetés	220
Total d'articles analysés	715
Total d'articles retenus (ETUD + GOUV)	495

3.1.4 Classification des articles

Nous avons établi les critères pour la classification manuelle de notre corpus au fur et à mesure de la lecture des articles. Nous avons retenu comme critères un ensemble de thèmes et d'arguments propres à chaque groupe sur la base des événements qui se sont déroulés pendant le mouvement. L'objectif a été d'organiser les articles en deux classes distinctes, tel que préconisé par l'approche textométrique. Nous avons remarqué que certains articles favorables à l'un ou à l'autre camp de la dispute abordaient un ou plusieurs critères de classification illustrés sur le tableau 12 ci-dessous.

La classification a été réalisée à l'aide de l'outil *Eureka Analytik*[27] du moteur de recherche *Eureka.cc* Cet outil permet de créer des corpus et d'attribuer des étiquettes aux articles.

[27] http://eureka.cc/fr/analyser-linformation/

Tableau 12. Critères de classification du corpus de référence

Classe	Critères à considérer
ETUD	▪ Arguments majoritairement contre la hausse. ▪ Arguments favorables à la continuité de la mobilisation. ▪ Dénonciations de la violence de la police envers les étudiants. ▪ Plaidoyers pour l'accessibilité à l'éducation ou sur l'État comme garant des services de base pour tous les citoyens. ▪ Arguments soulignant l'importance de respecter le droit collectif des étudiants de faire la grève. ▪ Arguments critiquant l'attitude du gouvernement envers les étudiants en grève. ▪ Arguments critiquant le projet de loi 78.
GOUV	▪ Arguments majoritairement en faveur de la hausse. ▪ Arguments contre les demandes des étudiants. ▪ Arguments sur le caractère raisonnable des propositions du gouvernement et contre la suite du mouvement de grève. ▪ Plaidoyers pour une vision marchande de l'éducation ou qui défendent l'idée de l'éducation comme un investissement individuel. ▪ Arguments critiquant l'inflexibilité des étudiants dans les négociations. ▪ Arguments dénonçant la violence des manifestants ou les inconvénients provoqués par les étudiants en grève. ▪ Arguments soulignant l'importance des droits individuels des étudiants. ▪ Arguments en faveur du projet de loi 78.

3.1.5 Fidélité de la classification manuelle du corpus de référence

La réalisation de tests de fidelité à la suite d'une classification permet de réduire les sources de biais provenant du caractère subjectif propre à la tâche. Deux approches s'offrent pour évaluer la fidélité : la fidélité interjuges et la fidélité intrajuges. La première repose sur la vérification de l'accord entre deux ou entre plusieurs juges. La seconde repose sur une approche de test et de retest, dans laquelle le chercheur vérifie la stabilité et l'accord de son observation (la classification effectuée) entre différentes occasions dans le temps. Ensuite il faut calculer le coefficient de corrélation entre les diverses observations, afin de déterminer la constance des estimations et ainsi assurer la reproductibilité de la procédure. Une des statistiques les plus utilisées pour mesurer le taux d'accord est la statistique kappa de Cohen (k) (Fortin, 2010).

Pour déterminer le pourcentage d'accord avec la statistique kappa, il faut diviser le nombre de jugements coïncidant entre les juges par le nombre total d'observations, puis multiplier ce résultat par 100 pour obtenir un pourcentage. Selon P.T. Higgins et Green (2008), les valeurs de kappa se situant entre 40 à 59 % indiquent un taux d'accord raisonnable. Les valeurs entre 60 % et 74 % sont considérées comme bonnes tandis que les valeurs au-dessus de 75 % reflètent un taux d'accord excellent.

Avec le support du *Centre Statistique de l'Université de Montréal*, nous avons élaboré une stratégie d'évaluation de la fidélité de la classification réalisée par la chercheuse sur les articles de l'expérimentation. Nous avons décidé d'engager deux assistants de recherche pour réaliser le test interjuges, sur un total de 250 articles (plus de 1/2 du corpus de référence). Cet échantillon d'articles a été choisi de façon aléatoire selon la méthode d'échantillonnage probabiliste (Fortin, 2010) faite à l'aide de la fonction « Random » du logiciel *Microsoft Excel*. L'échantillon a été formé par des articles classés comme GOUV et ETUD, ainsi que les articles rejetés par la classification (en fonction des critères déclinés à la page 190 et 191).

Afin de mesurer l'accord interjuges, nous avons décidé de retenir autant les articles classés que les articles rejetés.

À l'aide de l'outil *Analytik*, les assistants de recherche ont classé manuellement chaque article dans les classes GOUV et ETUD ou REJETÉ, en suivant un plan de classification

(Annexe B) qui contenait les critères exposés au tableau 12 (p.196). Pour le premier assistant, nous avons obtenu un score kappa de 77,2 %. Pour le deuxième, nous avons obtenu un taux d'accord de 78,4 %. Les taux obtenus sont excellents. La plupart des désaccords retrouvés dans les classifications concernaient les articles qui ont été rejetés du corpus final. Si nous considérons seulement les articles initialement classés par la chercheuse dans les classes GOUV et ETUD, nous obtenons un taux d'accord de 98 % et 95,2 % pour les 135 et les 138 articles respectivement. Cela démontre que l'accord quant à la classification des articles faisant partie du corpus de référence est très significatif, ce qui garantit la fidélité de notre classification initiale.

3.2 Sélection de critères textuels textométriques

La sélection de critères textuels textométriques constitue la deuxième étape de la démarche méthodologique de la fouille d'opinions présentée dans cette recherche (figure 9, p.181). Elle consiste à sélectionner les critères textuels en employant des méthodes textométriques. Nous allons dans cette section décrire les étapes de l'analyse textométrique du corpus. Les critères textuels sélectionnés seront utilisés pour donner une représentation informatique aux articles du corpus. Ils sont typés en fonction des démarches qui amènent leur obtention, lesquelles seront expliquées dans cette section aussi.

La première démarche pour la sélection de critères textuels consiste à réaliser la segmentation des textes numériques afin d'obtenir des données textuelles discrètes. Les différentes techniques de segmentation permettent de diversifier les types de données en fonction de leur variation lexicographique. Les techniques de segmentation sont décrites plus en détail dans cette section (p. 200).

Après la segmentation des textes, deux calculs statistiques textométriques sont employés pour l'analyse du corpus : le calcul des spécificités et le calcul de cooccurrence à partir d'un mot pôle. Nous avons présenté dans le chapitre 2 des explications sur chacun des calculs et nous allons l'expliquer davantage dans cette section. L'objectif de cette démarche est de sélectionner les critères textuels appartenant à un contexte global, constitué du corpus de référence et des sous-corpus, ainsi que d'un contexte local, situé sur les voisinages des mots. Le calcul des spécificités mobilise le contexte global, puisqu'il est déployé en tenant

compte des données textuelles présentes dans le corpus de référence et les sous-corpus. Le calcul de cooccurrences mobilise le contexte local dans la mesure où les associations sémantiques avec les mots pôles sont recherchées dans des unités de contexte plus restreintes, comme les phrases ou les fenêtres contextuelles contenant un nombre déterminé de mots à gauche et à droite du mot pôle.

La démarche de sélection de critères comprend également une étape de catégorisation des critères textuels sélectionnés en fonction des composantes sémantiques thématique, dialectique et dialogique. Cette étape sera expliquée aussi dans cette section. Nous voulons caractériser les critères textuels retenus en fonction des récurrences sémantiques et des stratégies argumentatives adoptées par chaque classe d'auteurs (ETUD et GOUV), afin de savoir comment les différents groupes impliqués dans le débat public sur la grève étudiante ont abordé le sujet et par quelles stratégies argumentatives ils ont essayé de convaincre leur lectorat. Cette partie comprend aussi la description des structures sémantiques identifiées, en s'appuyant sur les concepts de la sémantique interprétative exposés dans le chapitre 2. La description des critères retenus cherche à découvrir les stratégies argumentatives des auteurs situés dans chaque camp et vérifie dans quelle mesure la structure actantielle polémique proposée par Garand (1998) se manifeste dans la discussion. Cette description présente une valeur ajoutée au cadre général de la recherche, dans la mesure où elle rend cohérent le choix des critères, à l'égard de la démarche interprétative qui explicite les régularités sémantiques dans tous les niveaux de structuration textuelle : thématique, dialectique et dialogique.

Les démarches de segmentation, d'application de calculs et de catégorisation de critères nous permettront de créer une typologie des critères textuels qui sera testée dans la recherche.

Les opérations effectuées pour la sélection de critères textuels sont réalisées à l'aide de deux logiciels textométriques — *Lexico3* (Salem et coll., 2003) et *TXM* (Heiden et coll., 2010). Ces logiciels fournissent un ensemble d'outils permettant d'effectuer les opérations ici décrites, comme la segmentation du texte en formes graphiques, le partitionnement du corpus de référence en sous-corpus, l'implémentation de calculs de spécificités (Lafon, 1980) et le calcul de cooccurrences à partir d'un mot pôle (Lafon, 1981). Nous allons privilégier l'utilisation de *TXM* pour la plupart des tâches et utiliser *Lexico3* pour un type particulier de

segmentation textuelle qui n'est pas disponible sur *TXM*. Les critères motivant le choix de ces outils sont expliqués dans l'Annexe C.

3.2.1 Techniques de segmentation

La textométrie n'accorde pas d'avantages à des types particuliers de données textuelles pour étudier les propriétés linguistiques des textes et recommande également que l'on varie les techniques pour l'obtention de telles données (Pincemin, 2012a). Dans le cadre des expérimentations en fouille d'opinions, Eensoo et coll. (2015) définissent trois types de critères textuels, correspondant à différents types de données textuelles :

1. Critères unitaires : formes graphiques (mots simples), lemmes ou catégories morphosyntaxiques ;
2. Critères composites adjacents : segments répétés ;
3. Cooccurrences : mots et lemmes cooccurrents.

Dans le but d'obtenir le maximum de variation sur les types de critères et de comparer la performance de chaque type lors de l'étape de fouille, nous allons appliquer les techniques de segmentation suivantes : a) segmentation en formes graphiques pour l'obtention de mots simples ; b) lemmatisation ; c) extraction de segments répétés ; et d) l'extraction de cooccurrents. L'analyse textométrique portera sur l'ensemble de ces unités extraites des textes.

Les sections suivantes expliquent la différence entre chacun de ces types de segmentation et indiquent les démarches amenant à leur obtention.

a) Choix de formes graphiques

La segmentation du texte en formes graphiques est une des techniques les plus simples de segmentation du texte et consiste à puiser dans ce dernier un ensemble de mots simples et isolés, appelés « forme graphique ». Une forme graphique est une suite de caractères bornée par un caractère délimiteur (comme le point ou l'espace). Elle peut avoir plusieurs occurrences dans un texte. Deux suites identiques de caractères sont considérées comme deux occurrences de la même forme graphique. Les formes graphiques ne sont pas des unités de sens et cette

segmentation ne permet pas d'extraire des mots composés comme « pomme de terre », « garde-robe » ou « par exemple ». En outre, cette technique n'est pas sensible aux différentes flexions d'un mot : les déclinaisons ou conjugaisons d'une même forme graphique sont décomptées comme des variables distinctes dans les traitements (Lebart et Salem, 1994).

Par défaut, le logiciel *TXM* réalise la segmentation des textes en formes graphiques discrètes. Le segmenteur lexical de *TXM* effectue l'opération de segmentation en fonction de paramètres spécifiés lors de l'importation du corpus, en utilisant certains délimiteurs comme les signes de ponctuation, les caractères d'élision et de fin de phrase. La liste de délimiteurs par défaut de *TXM* se trouve dans l'Annexe C. Au total nous avons obtenu 21 077 formes graphiques, c'est-à-dire des mots simples.

Nous garderons le terme « mot simple » pour parler de la donnée textuelle constituée par une forme graphique. Un critère constitué par un mot simple est un « critère unitaire simple ».

b) Lemmatisation

La lemmatisation est une opération permettant de regrouper dans les mêmes formes graphiques les différentes flexions d'un mot. Ainsi, on ramène des formes verbales à l'infinitif, les substantifs au singulier, les adjectifs au masculin singulier et les formes élidées à leur forme sans élision (Lebart et Salem, 1994). Cette procédure rend plus aisées les inférences sur les fréquences cumulées des lemmes obtenus (Brunet, 2000), puisque ceux-ci sont transformés en des variables discrètes (séparées et indivisibles). La lemmatisation peut être optimisée avec une procédure d'étiquetage de la partie du discours (EPD), qui consiste à attribuer une étiquette à chaque mot d'une phrase en fonction de sa catégorie linguistique (syntaxique ou morphologique) afin de découper ses composantes dans des parties du discours (pronom, verbe, adjectif, etc.). Cette stratégie permet de réduire le phénomène d'ambiguïté intervenant dans la transformation de données textuelles en lemmes. Par exemple, dans une phrase où la forme « avions » est étiquetée comme un verbe par l'EPD, la lemmatisation ramène cette forme au lemme « avoir » et non pas à la forme singulière du substantif « avion ».

TXM offre la possibilité d'installer un outil d'analyse morphosyntaxique appelé *TreeTagger*, disponible dans diverses langues, dont le français. Lors de l'importation du

corpus, il est possible également de lancer le processus de marquage morphosyntaxique qui attribue une étiquette à chaque mot correspondant à sa catégorie morphosyntaxique. La lemmatisation avec *TreeTagger* ramène à un même lemme les variations verbales et normalise les formes au pluriel pour les substantifs, adjectifs et prépositions. Au total, nous avons obtenu 12 201 lemmes avec *TreeTagger*.

Nous garderons le terme « lemme » pour parler de la donnée textuelle constituée par un lemme. Un critère constitué par un lemme est un « critère unitaire lemmatisé ».

c) Extraction de segments répétés

L'extraction de segments répétés est une opération basée sur des méthodes probabilistes et permet d'obtenir des données textuelles plus larges que des simples mots isolés. En réalisant des coupures dans le texte, il est possible d'obtenir des séquences contiguës, correspondant à des expressions ayant plus d'un mot, comme des mots composés (pomme de terre), des locutions nominales, verbales ou adverbiales (« niveau de vie », « problèmes financiers », « mettre le feu », etc.) (Lebart et Salem, 1994). L'opération consiste à établir des délimiteurs de séquences, plus généralement des signes de ponctuation et d'analyser la fréquence des occurrences juxtaposées. « Les occurrences non séparées par un délimiteur de séquence sont des occurrences de segments » (Lebart et Salem, 1994, p. 60). L'extraction de segments répétés permet d'obtenir des cooccurrents qui entretiennent des relations syntaxiques entre eux, ce qui rend possible l'étude du système idiomatique d'une langue. Par exemple, la découverte de syntagmes figés comme « pomme de terre », « boîte à lettres » ou semi-figés comme « casser les oreilles », « prendre l'air », etc.) (Mayaffre, 2008).

Nous avons utilisé la fonctionnalité de segments répétés de *Lexico3* pour l'obtention de segments répétés. Par défaut, la fonctionnalité de segmentation du logiciel propose un seuil minimal de fréquence de 10 pour la recherche de segments répétés. Après des tests effectués en variant ce seuil minimal, nous avons remarqué que les résultats apportés par les seuils plus petits étaient bruités par l'apparition de plusieurs segments répétés contenant les mêmes mots et relevant des mêmes locutions (par exemple : 'niveau de vie' ; 'niveau de'). Cette observation nous a motivés à utiliser le seuil proposé par la fonctionnalité de segmentation du logiciel.

Nous garderons le terme « segment répété » pour parler de la donnée textuelle constituée par un segment répété. Un critère constitué par un segment répété est un « critère adjacent ».

d)Extraction de cooccurrents

L'extraction de cooccurrents fait état de la proximité récurrente de certaines formes graphiques. Selon Lafon (1981, p.97), « les cooccurrences constituent (…) la manifestation matérielle des rapports et des relations diverses qui se nouent dans la chaîne syntagmatique d'un texte ». Ces relations peuvent être très variables, à l'exemple du rapport de verbes avec leurs compléments ou le rapport sémantique entre des mots et des signifiés dans un texte.

L'extraction de cooccurrents consiste à repérer les formes graphiques qui apparaissent fréquemment ensemble dans une unité de contexte qui peut être, par exemple, une phrase ou une fenêtre d*e n* mots. Cette technique rend compte de liaisons sémantiquement motivées entre des mots et est aujourd'hui notamment utilisée pour explorer les thématiques des corpus à un niveau plus généralisé (Pincemin, 1999a ; Sjöblom et Leblanc, 2012 ; Guaresi, 2016 ; Vanni et Mittmann, 2016). Dans le cadre cette recherche, le repérage de cooccurrents consiste à rechercher un ensemble de mots se retrouvant dans le voisinage d'un mot pôle déterminé.

Le *TXM* offre parmi ses fonctionnalités le calcul de cooccurrences. La recherche des cooccurrences se développe à partir d'un mot particulier, le mot pôle, qui peut être un mot simple ou un lemme par exemple. Ensuite, il faut spécifier les seuils et les unités de contexte à partir desquels les calculs sont lancés. Il faut choisir un seuil de fréquence minimale d'apparition de la cooccurrence dans le corpus. L'unité de contexte pour la recherche des cooccurrents peut être la phrase, le paragraphe ou une fenêtre de longueur fixe entourant le mot ou le lemme recherché et le seuil de co-fréquence est le nombre de rencontres entre les couples de mots considérés.

Dans le contexte de notre travail, la recherche de cooccurrences est liée à la sélection de critères locaux. Les démarches spécifiques qui touchent cette sélection de critères locaux et de détails sur le calcul de cooccurrences sont exposées à la page 211.

Nous garderons le terme « cooccurrent » pour parler de la donnée textuelle constituée par un cooccurrent associé à un mot pôle. Les critères constitués par les cooccurrents sont appelés « critères locaux » dans le cadre de la recherche.

d) Données textuelles obtenues

Le tableau 13 ci-dessous présente le nombre de chaque type de donnée textuelle obtenue suite à la segmentation du corpus par les logiciels. Nous allons garder cette terminologie pour nous référer à chacune de ces données textuelles. De cette façon, un critère textuel sélectionné est une donnée textuelle du type mot simple, lemme, segment répété ou cooccurrent.

Tableau 13. Résultat de la segmentation : données textuelles obtenues

Type	**Nombre de données**
Mots simples	21 077
Lemmes	12 201
Segments répétés	36 085
Cooccurrents	1888

3.2.2 Calcul des spécificités et sélection de critères textuels globaux

La segmentation des textes numériques est une technique permettant d'extraire toutes les données textuelles du corpus et est une étape nécessaire des analyses textométriques. À la suite de la segmentation, des outils statistiques de la textométrie permettent de ressortir les spécificités de chaque sous-corpus et rend possible l'étude des régularités sémantiques présentes. Les données textuelles devenues déstructurées recouvrent leur contexte au sein des divisions opérées dans les corpus par les logiciels : elles peuvent être indexées dans des unités de contexte divers comme le texte, le paragraphe, la phrase, etc. Une des tâches centrales dans la démarche méthodologique de fouille d'opinions proposée dans le cadre de cette recherche est la possibilité de découper le corpus en partitions, de façon à cibler la recherche de ces motifs en fonction des classes qui ont été créées pour caractériser les types d'opinions véhiculées.

En tenant compte des problématiques propres à la textométrie, les logiciels *TXM* et *Lexico3* offrent une fonctionnalité de création de partitions (ou de sous-corpus, pour reprendre la terminologie utilisée dans cette étude) à partir du corpus de référence, permettant de déployer un niveau de contextualisation qui prend en compte la globalité des textes organisés dans les sous-corpus. L'acquisition de cette vue globale offre la possibilité de comparer et de caractériser, par des outils statistiques, les données textuelles des sous-corpus. Fondé sur la distribution en probabilité des mots dans le corpus (Lafon, 1984), le calcul des spécificités rend compte des données textuelles qui ont un emploi localisé dans une certaine partie du corpus, en considérant le nombre total d'occurrences de ces données dans le corpus complet.

Dans le cadre de notre travail, les régularités observées dans les sous-corpus quant aux données textuelles ressorties par le calcul des spécificités sont décrites en fonction des concepts d'isotopie et d'isosémie, présentés dans le cadre théorique (chapitre 2). En regardant les données surreprésentées, il est possible d'identifier les traits récurrents et de caractériser les régularités rencontrées sur le plan sémantique (Pincemin, 2012a).

Dans les travaux visant l'application de méthodes textométriques à la construction de critères textuels pour la fouille d'opinions, une démarche de sélection des critères globaux est proposée (Eensoo et Valette, 2015, p.6 ; Eensoo et coll. (2015). Les données textuelles ressortant des calculs doivent répondre à trois exigences pour être sélectionnées comme critères textuels :

a) Caractère spécifique à un sous-corpus : choix d'un seuil de spécificité représentatif.

b) Répartition uniforme dans le sous-corpus : les données textuelles qui ont un emploi très localisé dans le sous-corpus (par exemple, qui sont présents dans une très petite quantité de textes) doivent être éliminées.

c) Pertinence linguistique : les données textuelles qui ont une seule fonction ou une seule signification sont privilégiées comme critères.

Dans le cadre de notre recherche, nous avons décidé de retenir les trois étapes, mais de définir le seuil de répartition (étape b) de façon empirique, afin de pouvoir évaluer

expérimentalement le meilleur seuil de répartition. Les prochaines sections expliquent ces étapes et le fonctionnement des calculs utilisés pour l'obtention des critères textuels globaux.

a) Calcul des spécificités

Dans une étude sur la variabilité de mots dans les corpus, Lafon (1980) a insisté sur le fait que les mots présents dans un corpus de textes n'ont pas de distribution normale, à l'exemple de la loi de Gausse. Par conséquent, il est trompeur d'estimer la probabilité de leur apparition par une moyenne arithmétique, dont se sert le calcul de fréquence relative. Lafon a démontré que la distribution des mots en corpus suivait une loi hypergéométrique et que, par rapport à la cloche gaussienne, cette distribution serait déformée par la grande quantité de mots qui présente de basses fréquences (ENS de Lyon et Université de Franche-Comté, 2017).

Lafon (1980) a donc proposé la méthode de spécificités pour identifier, dans une partie du corpus, les mots qui seraient surreprésentés (ou au contraire, sous-représentés) en considérant le nombre total de ses occurrences dans le corpus complet. La méthode de spécificité est fondée sur la distribution en probabilité des mots dans le corpus, suivant la loi hypergéométrique (Lafon, 1984), et mesure si la fréquence que l'on observe dans une partie du corpus est due ou non au hasard. Le calcul des spécificités considère les paramètres suivants (Lebart et Salem, 1994) :

- T = la longueur du corpus T
- t = la taille du sous-corpus
- f = la fréquence du mot dans le corpus
- k = la fréquence du mot dans le sous-corpus

Avec ces paramètres, le calcul des spécificités porte un jugement sur une fréquence observée d'un mot quelconque dans le sous-corpus, afin de savoir si celle-ci est normale ou anormale par rapport à la fréquence la plus vraisemblable de ce mot dans le corpus de référence. Autrement dit, il mesure l'écart existant entre une répartition aléatoire de mots dans les corpus de référence et leur comportement effectif observé dans le sous-corpus (Pincemin, 2012a). Le calcul procède par un tirage aléatoire sans remise pour prélever du corpus de référence des échantillons contenant exactement le nombre de mots correspondant à la taille *t* du sous-corpus. Il construit ensuite une distribution de probabilités à partir des paramètres *T*, *t*

et *f* sur l'ensemble des fréquences possibles du mot (les valeurs contenues entre 0 et *f*). Le mode de cette distribution, étant la valeur la plus vraisemblable d'apparition du mot, sert à porter un jugement sur la valeur *k* observé dans le sous-corpus, en utilisant l'équation :

$$P(x = K) = \frac{f!\,(T-f)!\,t!\,(T-t)!}{k!\,(f-k)!\,(t-k)!\,(T-f-t+k)!\,T!}$$

Sur le logiciel *TXM*, le calcul des spécificités est lancé sur une partition créée à partir du corpus de référence (dans notre terminologie, des sous-corpus). Les résultats du calcul sont donnés sur un tableau qui spécifie, pour chaque donnée textuelle présente dans le corpus de référence, un score conventionnel de valeur positive ou négative indiquant soit la surreprésentation de cette donnée dans le sous-corpus, soit sa sous-représentation. Puisque le modèle hypergéométrique produit des variations exponentielles, la spécificité est représentée par des logarithmes de base 10 permettant de comparer des ordres de grandeur plutôt que des probabilités. Par exemple, un score +3 représente une valeur 1/1000 (3 étant le nombre de zéros, l'exposant de 10^3 résultats de log10 [1000]) et indique qu'il y a moins d'une chance sur mille que la donnée textuelle considérée soit retrouvée par hasard avec une fréquence aussi élevée dans la partition considérée, relativement à l'ensemble du corpus (ENS de Lyon et Université de Franche-Comté, 2017). Plus le score de spécificité est élevé, plus la chance est faible de retrouver par hasard une fréquence aussi élevée pour la donnée textuelle dans la partition considérée, relativement à la taille du corpus. Plus le score est bas, plus la chance est faible de retrouver la donnée par hasard avec une fréquence aussi basse dans la partition considérée, relativement à l'ensemble du corpus.

Le logiciel *TXM* propose un seuil de représentativité fixé à 2 (+ 2 pour la surreprésentativité et - 2 pour la sous-représentativité. Les valeurs entre -2 et +2 sont considérées comme banales. Le Manuel de *TXM* n'explique pas le choix de ce seuil (ENS de Lyon et Université de Franche-Comté, 2017). Il est par contre présent dans le graphique de visualisation des spécificités du logiciel, qui délimite, par des lignes horizontales, des zones de spécificités dans lesquelles les probabilités d'apparition sont supérieures à 1 % (> +2 ou < -2), et les zones de banalité (< + 2 et > -2), correspondant à des valeurs qui ne sont pas très importantes du point de vue du modèle.

Le calcul des spécificités ne détermine pas un seuil de signifiance, il cherche plutôt à mesurer l'écart de l'équirépartition, faisant intervenir tous les données textuelles du corpus (Lafon, 1980, note 11 p. 141). À la sortie du calcul, une lecture de listes permet de caractériser chacune des données textuelles du corpus par ses spécificités positives ou négatives. Comme il s'agit d'un cas de comparaisons multiples, un seuil est fixé arbitrairement pour permettre d'effectuer les comparaisons (Lebart et Salem, 1994 p. 176). Néanmoins, le choix de ce seuil se base sur des valeurs plus sévères que les seuils classiques, plutôt 1 % ou 1 pour 1000, comme l'explique Lebart et Salem (1994) :

> Le calcul simultané de plusieurs valeurs-tests ou de plusieurs seuils de probabilités se heurte à l'écueil des comparaisons multiples, bien connu des statisticiens.
>
> Supposons que les parties de texte soient parfaitement homogènes et donc que l'hypothèse d'indépendance entre les formes et les parties soit réalisée. Les valeurs-tests attachées aux spécificités, pour une partie donnée, sont alors toutes les réalisations de variables aléatoires normales centrées réduites indépendantes. Dans ces conditions, en moyenne, sur 100 spécificités calculées, 5 seront en dehors de l'intervalle [-1,96, + 1,96], et 5 dépasseront la valeur 1.65 (test unilatéral). Le seuil de 5 % n'a de sens en fait que pour un seul test, et non pour des tests multiples.
>
> Autrement dit, l'utilisateur non averti trouvera presque toujours « de quoi s'étonner » au seuil de 5 %... on résout de façon pragmatique cette difficulté en choisissant un seuil plus sévère (…) : le seuil 1 % correspond à une valeur supérieure à 2.33 et le seuil 1 pour 1000 à une valeur supérieure à 3,09. (Lebart et Salem, 1994, p. 183)

La taille du corpus et les ordres de grandeur des fréquences sont aussi à prendre en compte (Lafon, 1980, p. 162). Dans un corpus de grande taille (Lafon [1980] mentionne un corpus de 300 000 occurrences), lorsque les fréquences plus petites sont considérées (fréquences plus grandes que 0 et 1), la fréquence observée ne dépasse pas celle du modèle (spécificité négative), ce qui fait que les petites fréquences sont toujours considérées comme sous-employées. Dans les corpus de petite taille, l'écart pour les basses fréquences peut s'avérer plus grand et dans ce cas, il peut être judicieux de choisir un seuil de spécificité de +2 (probabilité de 1 %), puisque le seuil de +3 peut s'avérer très sévère.

Afin d'éviter de choisir un seuil exigeant, nous allons considérer, pour la sélection de critères, les données textuelles ayant un seuil de spécificité plus grand que +2, même si notre corpus est considéré de grande taille selon l'exemple fourni par Lafon (1980, p.162). Nous voulons valider le seuil de spécificité le plus performant lors de la tâche de classification.

b) Répartition dans le corpus

Parfois, la spécificité d'une donnée textuelle est fonction de sa fréquence importante dans seulement quelques articles du sous-corpus et elle se trouve mal distribuée dans ce dernier. Dans ce cas, la donnée peut être spécifique au sous-corpus, mais pas représentative de ce dernier. Analyser la répartition de données textuelles a comme objectif l'élagage de celles qui sont présentes dans une portion très réduite des articles considérés et pour cela, ne sont pas discriminantes de l'ensemble du sous-corpus. Dans ces conditions, il peut être utile de définir un seuil minimal d'articles dans lequel la donnée textuelle doit être présente pour qu'elle soit sélectionnée comme critère textuel. Cette procédure vise à pondérer la spécificité produite par un ensemble de données surreprésentées dans un sous-corpus, mais mal réparties dans ce dernier.

La fonctionnalité « Carte de Sections » de Lexico3 est mentionnée dans quelques travaux (Eensoo et Valette, 2012, 2014a, 2014b, 2015) comme outil privilégié de vérification de la répartition de données textuelles dans le sous-corpus. Il s'agit d'une interface de visualisation dans laquelle il est possible de voir le nombre d'articles (représentés par des cartes) affectés par une donnée textuelle quelconque.

Les figures 11 et 12 ci-dessous présentent l'histogramme des affectations des mots simples les plus spécifiques dans chaque sous-corpus ETUD et GOUV, respectivement. Sur l'axe vertical des deux graphiques se trouve l'effectif des mots spécifiques dont le score de spécificité est plus grand que +2. Sur l'axe horizontal se trouve le nombre d'articles affectés par chaque mot spécifique, rangé dans des intervalles de 20. Il est à noter que la majorité des mots spécifiques apparaissent au moins 1 fois dans un petit nombre d'articles, plus spécifiquement entre 20 et 100 articles, dans les deux sous-corpus, ce qui correspond à environ 12 % du total d'articles du corpus de référence. Les affectations se dispersent à partir de

l'intervalle 120 dans les deux cas, avec une petite quantité de mots affectés à une grande quantité d'articles.

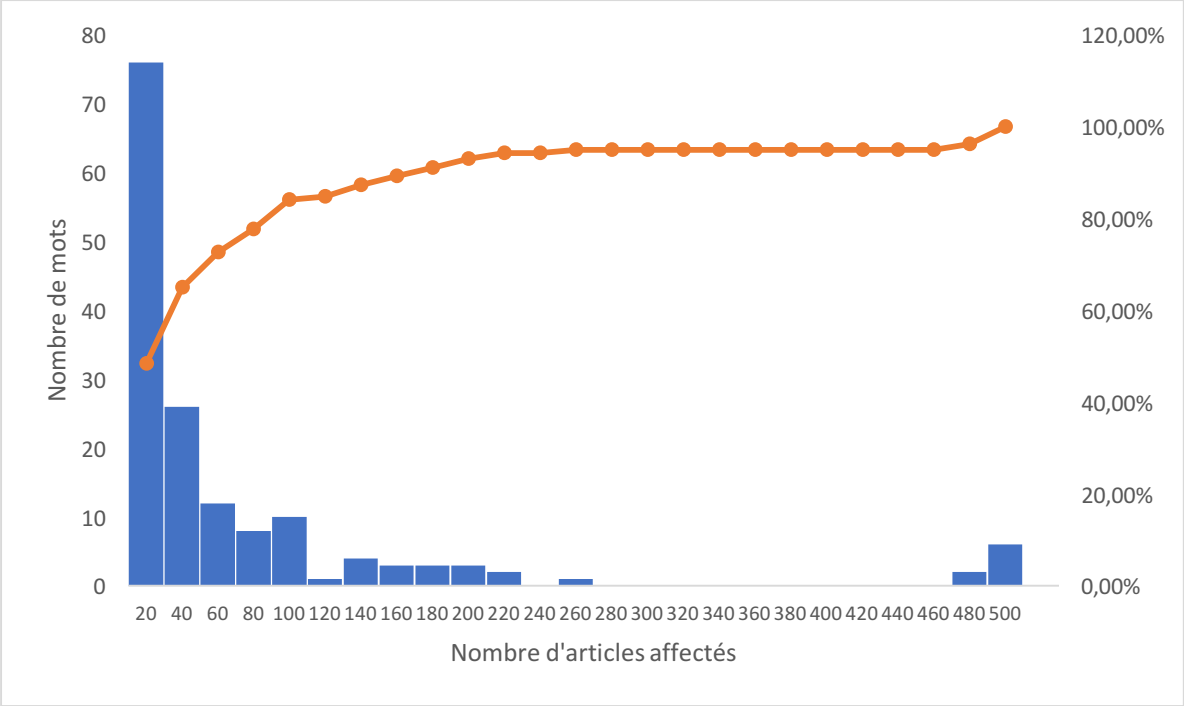


Figure 11. Répartition de la fréquence des mots simples les plus spécifiques dans ETUD

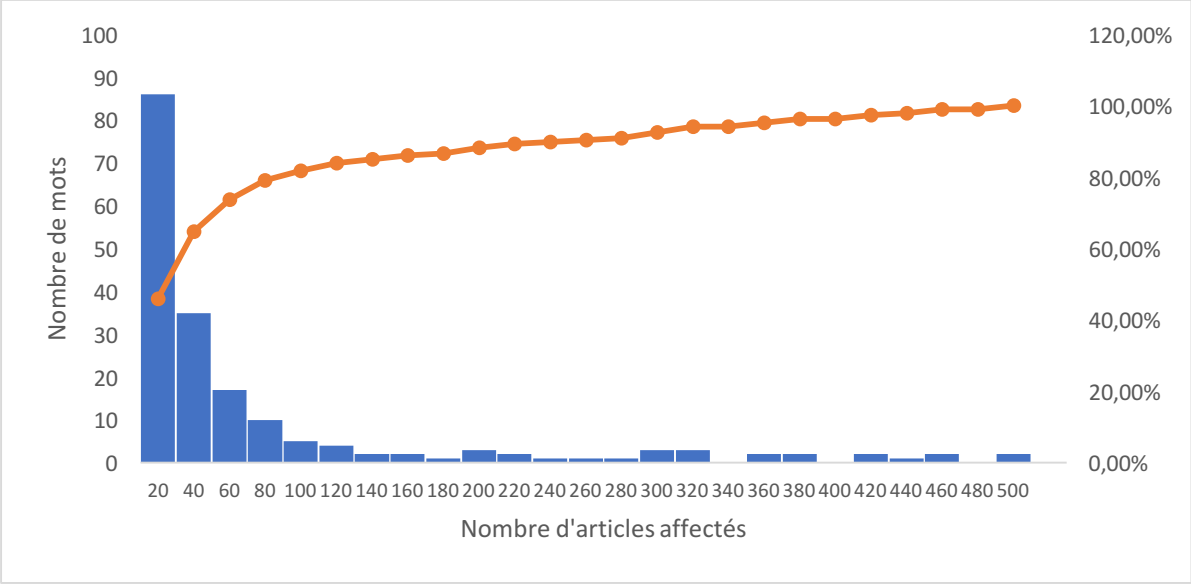


Figure 12. Répartition de la fréquence des mots simples les plus spécifiques dans GOUV

Dans les graphiques ci-haut, la densité maximale de la distribution est la colonne la plus élevée. Elle indique qu'un nombre maximal de 20 articles sont affectés par presque la moitié des mots simples spécifiques des deux sous-corpus, soit 48,41 % pour ETUD et 45,99 % pour GOUV. Ces valeurs sont détaillées sur le tableau des fréquences cumulées ci-dessous :

Tableau 14. Répartition des mots simples spécifiques pour ETUD et GOUV avec cumul

ETUD			**GOUV**		
Classes	**Mots spécifiques**	**% cumulé**	**Classes**	**Mots spécifiques**	**% cumulé**
20	**76**	**48,41 %**	**20**	**86**	**45,99 %**
40	26	64,97 %	40	35	64,71 %
60	12	72,61 %	60	17	73,80 %
80	8	77,71 %	80	10	79,14 %
100	10	84,08 %	100	5	81,82 %
200	3	92,99 %	200	3	88,24 %
300	0	94,90 %	300	3	92,51 %
400	0	94,90 %	400	0	96,26 %
500	6	100,00 %	500	2	100,00 %

Dans le cadre de notre recherche, nous avons décidé de définir des seuils d'affectation de manière empirique, afin de tester la performance que produirait, dans la tâche de classification, la variation de ces paramètres. Par contre, nous allons éliminer les données textuelles spécifiques qui apparaissent dans un seul document.

c) Pertinence linguistique

Le critère de pertinence linguistique a trait à l'uniformité de l'emploi linguistique des données textuelles spécifiques dans les textes du sous-corpus. À partir du concordancier du logiciel *TXM* (figure 13), il est possible d'extraire les contextes à gauche et à droite d'une donnée textuelle et d'évaluer comment celle-ci est employée dans les textes. L'analyse des contextes vise à élaguer les données textuelles qui apparaissent avec des fonctions

grammaticales différentes où avec un emploi sémantiquement diversifié dans les contextes repérés.

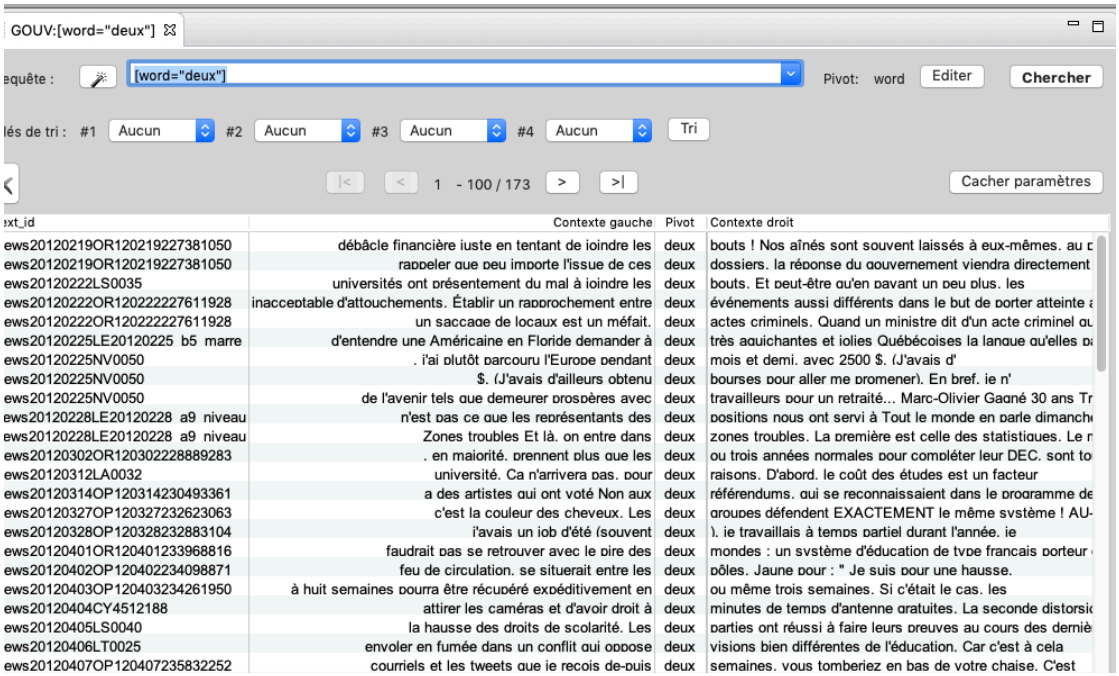


Figure 13. Recherche par le mot simple « deux » dans le concordancier

Par exemple, le mot « deux » apparaît avec un haut score de spécificité dans le corpus GOUV (+3,3), mais son emploi est très divers comme atteste la figure 13 ci-dessus : sa spécificité ne semble pas liée à un contexte d'emploi spécifique, mais l'effet d'un hasard. Par contre, le mot « nous », très spécifique du sous-corpus ETUD (+11,1) est un pronom qui ne présente pas de polysémie et se retrouve employé dans tous les contextes avec le même sens.

Dans la prochaine section, nous allons expliquer le deuxième calcul utilisé dans le cadre de la recherche, le calcul de cooccurrence, ainsi que les démarches pour la sélection de critères textuels locaux.

3.2.3 Calcul de cooccurrence et sélection de critères textuels locaux

Dans le cadre de notre recherche, nous allons utiliser le calcul de cooccurrence autour d'un mot pôle proposé par Lafon (1981) pour repérer les critères locaux spécifiques à chaque sous-corpus. Ces critères sont dits locaux, car le calcul permettant leur repérage porte sur des zones délimitées de la chaîne textuelle, ordinairement une fenêtre contenant un nombre

déterminé de mots à gauche et à droite du mot pôle. Selon la perspective théorique que nous proposons ici, les cooccurrents spécifiques de chaque corpus permettent de caractériser les thèmes spécifiques de ce dernier. Les cooccurrents spécifiques rencontrés sont homologués conceptuellement aux molécules sémiques. Nous proposons d'explorer seulement les associations entre des mots lexicaux (noms substantifs, verbes et adjectifs qualitatifs), afin de rester dans l'exploration thématique du corpus. Nous voulons également faire une analyse comparative entre les cooccurrents retrouvés dans chaque sous-corpus pour mettre en évidence les stratégies utilisées par chaque camp de la dispute et justifier la pertinence de la démarche.

La démarche de cooccurrence inspirée d'Eensoo et Valette (2015) comprend les étapes suivantes, qui seront détaillées dans les prochaines sections :

a) Sélection de mots pôles.

b) Calcul des paires de mots cooccurrents pour les mots pôles dans chaque sous-corpus ETUD et GOUV, séparément.

c) Analyse des contextes d'apparition de ses cooccurrences et évaluation de leur pertinence linguistique.

d) Sélection de cooccurrents spécifiques à un sous-corpus.

a) Sélection de mots pôles

De pair avec la méthodologie proposée par Eensoo et Valette (2015), nous allons sélectionner comme mot pôle les mots qui présentent une haute fréquence et dont la distribution sur l'ensemble des articles du corpus de référence est considérée uniforme. Cela revient à dire que les mots pôles ne peuvent pas être discriminants d'un seul sous-corpus. Au sens de la théorie, les mots pôles sont considérés comme appartenant au fond sémantique du corpus de référence : leur caractère non discriminant et leur répartition uniforme homologuent les isotopies génériques du corpus de référence et par conséquent, les « informations sur le domaine principal » de ce dernier (Rastier et coll., 1994, p. 207).

> Ainsi, un mot tel qu'« étranger » est considéré comme un fond sémantique d'une part, parce que, sur notre corpus de test, les mesures de rappel antiraciste et raciste sont respectivement de 55,96 % et 44,04 %, et d'autre part, parce que ce mot est actualisé dans 59,24 % des textes dans leur ensemble. Autrement dit, le mot « étranger » est très fréquent dans les textes racistes et dans les textes antiracistes (dans plus d'un texte sur deux), mais il n'est pas discriminant dans la mesure où il apparaît à peu près autant dans les deux sous- corpus. (Valette et Slodzian, 2008, p.127)

Afin de trouver les mots pôles pertinents, nous avons effectué des analyses en utilisant le logiciel *Wordstat* de *Provalis Research.*[28].

Nous avons ressorti les 300 lemmes les plus discriminants du corpus, en utilisant la mesure TF-IDF (Salton et McGill, 1983). Cette mesure statistique pondère les hautes fréquences des mots du corpus en fonction de leur présence locale dans les articles. Elle permet ainsi d'élaguer les mots qui malgré leur haute fréquence ne sont pas représentatifs du corpus puisqu'elles se trouvent réparties également dans la majorité des articles. Cela est notamment le cas de certains verbes comme les auxiliaires être et avoir.

À partir de la liste obtenue avec *WordStat,* nous avons constaté qu'un total de 11 lemmes ayant les scores TF-IDF les plus élevés figuraient sur plus de 50 % des textes (tableau 15). Ils sont : « étudiant » (92,93 %), « Québec » (83,43 %), « scolarité » (76,57 %), « gouvernement » (73,94 %), « même » (67,88 %), « autre » (66,46 %), « faire » (64,65 %), « hausse » (61,21 %), « cela » (55,96 %), « aussi » (55,15 %) et « Québécois » (53,94 %).

Tableau 15. Liste des mots les plus fréquents et les mieux répartis dans le corpus de référence

Lemmes	**Fréquence**	**Nombre d'articles affectés**	**% d'articles**	**TD * IDF**
étudiant	2451	459	81 %	78,1
Québec	1021	430	86 %	80,3

[28] Adresse URL : https://provalisresearch.com/products/content-analysis-software/

scolarité	809	391	78 %	93,8
gouvernement	1243	367	72 %	163
droit	916	326	659 %	149,4
même	782	317	59 %	131,6
autre	625	322	66 %	110,9
faire	624	441	64 %	118,2
hausse	720	305	59 %	153,5
cela	487	273	55 %	122,8
aussi	488	273	55 %	126,1

Nous avons pris la décision de limiter la sélection de mots pôles aux mots lexicaux qui sont plus significatifs du thème du corpus, excluant ainsi les mots grammaticaux comme « même » ainsi que les verbes. Le tableau 16 ci-dessous présente les mots pôles sélectionnés, ainsi que leur répartition dans les deux sous-corpus. Comme nous pouvons constater, les mots sélectionnés présentent une répartition uniforme dans ETUD et GOUV et ne sont pas discriminants de l'un ou de l'autre.

Tableau 16. Mots pôles sélectionnés avec leur distribution dans les sous-corpus ETUD et GOUV

Mots pôles	**ETUD**	**GOUV**
étudiant	91 %	93 %
Québec	87 %	85 %
scolarité	76 %	81 %
gouvernement	75 %	72 %
droit	70 %	60 %
hausse	62 %	61 %

b) Calcul des paires de mots cooccurrents pour les mots pôles dans les sous-corpus ETUD et GOUV

Le calcul de cooccurrence (Lafon, 1981) considère un mot pôle *P* et deux expansions situées avant et après le pôle. Deux indicateurs sont pris en compte pour calculer la relation

d'un cooccurrent *C* au mot pôle *P* à l'intérieur de ces expansions : la cofréquence $C \rightarrow P$ ou $P \rightarrow C$ selon que *C* précède *P* ou vice-versa. Ensuite, la distance en nombre de mots simples entre *C* et *P*, qui traduit le degré d'attraction entre les deux : si les distances sont trop faibles (par exemple, 1 ou 2), une distance moyenne peut indiquer une relation étroite entre le mot pôle et le cooccurrent. En tenant compte de ces deux valeurs, le calcul de cooccurrence génère un coefficient qui hiérarchise les cooccurrences rencontrées, en faisant intervenir la cofréquence de *C* et *P* et d'autre part, la distance moyenne séparant les deux. Les coefficients plus grands sont plus importants et considèrent à la fois les cooccurrences qui ont une cofréquence élevée et une distance moyenne faible. Le résultat du calcul permet d'obtenir la liste de cooccurrents les plus attirés par le mot pôle *P*.

Le calcul de cooccurrence doit être réitéré pour chaque mot pôle *P* sélectionné en fonction des besoins de l'étude. Nous allons appliquer le calcul pour les mots pôles sélectionnés qui figurent sur le tableau 16.

Sur le choix concernant le contexte environnant le mot pôle, Lafon (1981) explique que la « limitation des expansions est variable : elle peut correspondre, soit à la première ponctuation forte en soit à un nombre fixe d'occurrences de part et d'autre du pôle » (p. 101). Nous avons choisi d'utiliser des expansions à la droite et à la gauche avec une taille maximale de 10 mots simples. Nous avons aussi varié la taille de la fenêtre pour vérifier si le changement de cette taille permet de retrouver des résultats très différents.

Dans *TXM*, le calcul de la cooccurrence repose également sur le modèle hypergéométrique (Lafon, 1980) et ordonne la liste de cooccurrents d'un corpus ou d'un sous-corpus selon l'ordre de spécificité. Par rapport au calcul des spécificités mentionné à la page 205, le calcul de cooccurrents fait une adaptation de paramètres : il prend d'abord la taille T du sous-corpus et la taille *t* comprenant l'assemblage de toutes les fenêtres contextuelles dans lesquelles figure le mot pôle. Ainsi, pour déterminer les cooccurrents spécifiques, le calcul mesure si la fréquence des mots simples retrouvés à côté des mots pôles dans les fenêtres considérées est aléatoire par rapport au sous-corpus, ou si au contraire ces fréquences sont spécifiques. Il porte alors un jugement sur la valeur *k* observée dans les fenêtres, en utilisant l'équation :

$$P(x = K) = \frac{f!\,(T-f)!\,t!\,(T-t)!}{k!\,(f-k)!\,(t-k)!\,(T-f-t+k)!\,T!}$$

Les paramètres du modèle sont :

- T = taille du sous-corpus
- t = taille de toutes les fenêtres contextuelles rassemblées contenant le mot pôle
- f = fréquence du mot dans le sous-corpus
- k = fréquence du mot dans toutes les fenêtres contextuelles qui contiennent le mot pôle

c) Contexte d'apparition et pertinence linguistique

La pertinence linguistique des critères locaux est analysée en fonction des contextes d'apparition, de la même façon que nous avons proposé pour la sélection de critères textuels globaux. Ainsi, pour le calcul de cooccurrences, seulement les cooccurrents employés avec la même fonction grammaticale et le même sens seront sélectionnés comme critères textuels.

La démarche de sélection de critères locaux vise à repérer d'autres critères thématiques pour la tâche de classification. De cette manière, nous voulons à l'aide du calcul de cooccurrence découvrir les critères textuels qui n'ont pas été repérés dans la démarche de sélection de critères globaux par le calcul des spécificités. Ainsi, il est important de vérifier si les cooccurrences qui sont ressorties dans la démarche proposée ne sont pas parmi les critères globaux. Nous voulons avec cela nous assurer que la démarche de sélection de critères locaux est pertinente du point de vue méthodologique, en démontrant qu'elle peut efficacement apporter d'autres critères textuels non repérés par la méthode de spécificité.

d) Sélection de cooccurrents spécifiques à un sous-corpus

La sélection de cooccurrents spécifiques est effectuée par la comparaison entre les résultats du calcul de cooccurrence pour chaque sous-corpus, à la suite des traitements précédents. Seront éliminés les cooccurrents communs aux sous-corpus ETUD et GOUV afin de ne retenir que les spécifiques non communs.

3.2.4 Catégorisation et description de critères textuels dans les composantes thématique, dialectique et dialogique

La catégorisation de critères textuels en fonction des composantes sémantiques a deux objectifs :

1. Permettre la description des régularités sémiques observées dans chaque sous-corpus sur chacune des composantes sémantiques.

2. Déployer un cadre expérimental pour comparer la performance des trois types des critères (thématique, dialectique et dialogique), ainsi que de la coalition de ces trois types.

Dans cette section, nous allons expliquer le processus de catégorisation des critères dans les composantes thématique, dialectique et dialogique, ainsi que leur description.

Pour le travail de catégorisation de critères, nous avons défini un ensemble de catégories de marqueurs linguistiques qui peuvent être associés à des phénomènes particuliers à chaque composante, en considérant les particularités du genre de l'opinion. Nous nous sommes inspirés de la classe de marqueurs proposée par la « Linguistique textuelle » d'Adam (2008), par la « Grammaire du sens et de l'expression » de Charaudeau (1992) et par les exemples donnés par les travaux Valette (2004) et Eensoo et Valette (2012, 2014a, 2014b, 2015). L'association de certaines catégories de marqueurs linguistiques aux composantes sémantiques n'est pas nouvelle, comme nous avons pu le voir dans les exemples cités dans le chapitre 2 (p.157 à 167), mais la prise en compte du contexte d'emploi des critères textuels à caractériser est le véritable gage de pertinence de la démarche de catégorisation.

La description de critères thématique, dialectique et dialogique repose sur les récurrences sémiques identifiées (isotopies, isosémies et molécules sémiques) et l'identification des stratégies argumentatives adoptées par chaque groupe en dispute. Par stratégie argumentative, nous voulons signifier l'ensemble des procédés qui visent à invalider et à délégitimer les arguments du groupe adverse et de renforcer et légitimer un ensemble d'arguments en faveur d'une opinion. Nous utilisons certains concepts exposés dans notre revue de la littérature au sujet de la caractérisation du genre de l'opinion et de l'argumentation

polémique (Garand 1998). Il nous intéresse dans ce sens de repérer à partir des critères textuels retenus et catégorisés dans les composantes les types d'arguments suivants :

- Arguments qui dénoncent des vices de raisonnement de l'opposant : contradictions ou incompatibilités, inconséquences ou confusion, simplifications ou omissions, etc.
- Arguments qui contestent le découpage référentiel effectué par l'adversaire consistant à produire un contre-exemple, à déplacer ou à hiérarchiser le problème ou qui contestent la compréhension des concepts utilisés.
- Arguments qui font appel au *pathos* de l'auditeur concernant des catégories axiologiques comme le beau et le laid, le bien et le mal, l'admirable ou l'exécrable.
- Arguments logiques avec lesquels l'énonciateur attaque son adversaire par le biais d'une contradiction logique de son discours.
- Arguments circonstanciels avec lesquels l'énonciateur met en rapport les arguments de l'adversaire et ses actes pour lui imputer une inconséquence ou une contradiction.
- Arguments personnels dont l'énonciateur se sert pour attaquer frontalement son adversaire, l'attaque en question pouvant aller jusqu'à l'insulte ou à l'injure.

Les sections suivantes présentent la liste de marqueurs linguistiques qui seront retenus pour chaque composante sémantique.

a) Critères thématiques

Pour la composante thématique, nous allons retenir des mots lexicaux (mots simples, lemmes, segments répétés et cooccurrents) c'est-à-dire des substantifs, verbes et adjectifs permettant de caractériser l'univers sémantique des textes. Nous sommes particulièrement intéressés par les critères permettant de décrire les domaines, les taxèmes et les dimensions de chaque sous-corpus.

Sur le plan de la composante thématique, l'analyse sémique cherche à identifier les taxèmes, les domaines ou les dimensions qui font état des thèmes présents dans les sous-corpus.

Il est à noter que les critères textuels locaux obtenus par la méthode de cooccurrence sont assimilés aux thèmes spécifiques, puisqu'ils correspondent sur le plan théorique à la co-récurrence de sèmes, c'est-à-dire aux molécules sémiques. Notre objectif ici est de décrire les molécules sémiques retrouvées dans ETUD et GOUV et les thèmes spécifiques de ces deux sous-corpus.

b) Critères dialectiques

Nous allons retenir dans la composante dialectique tous les marqueurs temporels et aspectuels ainsi que ceux faisant état de la structure logique et argumentative du texte (mots simples, lemmes et segments répétés). Au sens de la sémantique interprétative, les critères dialectiques permettent de décrire « de la succession des intervalles dans le temps textuel, comme les états qui y prennent place et les processus qui s'y déroulent » (Rastier, 1989, p. 278). Nous nous inspirons de la distinction d'Adam (2008) pour dresser une liste marqueurs permettant d'ordonner le texte sur l'axe du temps et de logique argumentative :

- Organisateurs temporels : ils organisent les parties du texte de façon séquentielle. Ils peuvent s'organiser sur un ordre d'information de plus en plus grande : « d'abord », « ensuite », « (et) puis », « depuis », « après », « la veille », « le lendemain », « trois jours plus tard », etc.
- Organisateurs énumératifs : ils peuvent être additifs (« et », « ou », « aussi », « ainsi que », « avec cela », « de même », « également », « en plus »), ouvrir une série (« d'une part », « d'abord », « premièrement », « en premier lieu », « d'un côté »), signaler la poursuite d'une action (« ensuite », « puis », « en second lieu », « et ») et aussi sa fermeture (« d'autre part », « enfin », « de l'autre », « en dernier lieu », « et », « c'est tout », « pour terminer »).
- Marqueurs de changement de topicalisation : ils ordonnent les actions en signalant un passage d'un thème à l'autre (« quant à », « en ce qui concerne », « concernant »).
- Organisateurs spatiaux (« à gauche/à droite », « devant/arrière », « au-dessus/dessous », « plus loin », « d'un côté/de l'autre », etc.).

- Marqueurs d'illustration et d'exemplification : ils introduisent des exemples en donnant à l'énoncé un statut d'illustration d'une assertion principale (« par exemple », « notamment », « en particulier », « comme », « entre autres » et « ainsi »).
- Marqueurs de reformulation : ils soulignent la modification de point de vue (« c'est-à-dire », « autrement dit », « c'est », « en un mot », « en d'autres termes »). Comparables aux marqueurs d'intégration conclusifs (« bref », « en somme »).
- Connecteurs argumentatifs : retraitement du contenu propositionnel, soit comme un argument, soit comme une conclusion, soit encore comme un argument chargé d'étayer ou de renforcer une inférence ou comme un contre-argument. Il y a les connecteurs argumentatifs et concessifs (« mais », « pourtant », « cependant », « certes », « toutefois », « quand même »), explicatifs et justificatifs (« car », « parce que », « puisque », « si », « c'est que »), et hypothétiques (« même », « d'ailleurs », « de plus », « non seulement »). Certains connecteurs contre-argumentatifs marquent une opposition forte (« mais », « pourtant », « néanmoins », « cependant », « quand même »), et d'autres connecteurs marquent une opposition plutôt nuancée (« certes », « bien que », « malgré », « quoique »).
- Modalités objectives (« devoir », « falloir »).
- Indicateurs rhétoriques (emphases, points d'interrogation, mots interrogatifs).

c) Critères dialogiques

Les critères dialogiques rendent compte de la représentation de l'énonciation et des modalités. Il nous intéresse de retenir pour cette composante les marques de l'énonciation et de l'interlocution, c'est-à-dire la représentation de l'énonciateur ainsi que de la personne de l'interlocution (son lecteur ou son public). Les critères à retenir (mots simples, lemmes et segments répétés) doivent correspondre aux pronoms personnels, ainsi, des expressions explicitement subjectives marquant la modalité, qui sont, au sens préconisé par Rastier, « l'ensemble de données textuelles associé à un acteur ou à un foyer énonciatif » (Rastier, 2015, p. 8). Il s'agit notamment de 4 modalités identifiées par la sémantique interprétative :la

modalité ontique (possible/impossible, factuel/non-factuel), véridictoire (vrai ou faux), épistémique (certain ou incertain) et la modalité thymique (euphorique/dysphorique, c'est-à-dire du positif / négatif) (Hébert, 2001, p. 147).

- Pronoms personnels (« je », « nous », « tu », « vous », « il(s) », « elle(s) », « on »), possessifs (« mon », « mien », « ma », « son », « ton », « sa »), et de traitement (« monsieur », « madame », etc.).
- Indice de personnes : appellatifs (noms propres, substituts lexicaux, noms de qualité).
- Marqueurs de cadre médiatif ou de source de savoir : ces marqueurs signalent qu'une portion du texte n'est pas prise en charge par celui qui parle, mais médiatisée par une autre voix (« selon », « d'après », « pour », « de source sûre »).
- Indications d'un support de perceptions et de pensées rapportées : effets de point de vue reposant sur une focalisation perceptive (« voir », « entendre », « sentir », « toucher », « goûter ») ou sur une focalisation cognitive (« savoir »).
- Marqueurs de structuration de la conversation : (« bon », « ben », « pis », « alors », etc.) et les phatiques (« tu sais », « tu vois », « euh », etc.).
- Modalités subjectives (« vouloir », « penser », « espérer »).
- Verbes exprimant une opinion (« croire », « savoir », « se douter », « ignorer », « convenir », « prétendre »).
- Adverbes modaux (« peut-être », « sans doute », « probablement », « certainement »).

3.2.5 Résumé de la démarche textométrique de sélection de critères textuels

Nous avons présenté dans notre cadre théorique le tableau 6 (p. 174) dans lequel figure la correspondance entre les concepts de la sémantique interprétative et les étapes de la démarche textométrique de sélection de critères adoptés par les travaux Valette (2004) et d'Eensoo et Valette (2012, 2014a, 2014b, 2015). Ce tableau contient également les observables linguistiques obtenues dans chaque étape. Le tableau 17 suivant associe les opérations proposées dans la méthodologie à chacune de ces étapes.

Tableau 17. Opérations associées aux étapes de la démarche textométrique

Étapes	Démarche méthodologique	Observables linguistiques	Opérations
Étape 1	Calculs contrastifs	Données textuelles spécifiques (mots, segments répétés et cooccurrents).	Techniques de segmentation pour l'obtention des données textuelles. Application des calculs des spécificités et de cooccurrences.
Étape 2	Identification des structures sémantiques	Critères textuels	Identification de critères globaux et locaux en fonction de la pertinence linguistique. Identification de critères thématique, dialectique et dialogique récurrents.
Étape 3	Description des composantes	Critères textuels interprétables	Catégorisation et description des critères textuels récurrents associés aux composantes.

Nous présentons dans la prochaine section la troisième étape de la démarche méthodologique de fouille d'opinions, qui est la Transformation.

3.3 Transformation : représentation vectorielle des critères textuels textométriques

À la suite de la sélection des critères textuels par une analyse textométrique dans l'étape de filtrage vient l'étape de Transformation, dans laquelle nous représentons chaque article du corpus dans une matrice de vecteurs (Salton, 1988). Dans cette dernière, chaque article est représenté par un vecteur où figure l'absence ou la présence de l'ensemble de critères textuels sélectionnés (pondération binaire), ou leur fréquence absolue (pondération par fréquences). Nous allons constituer différentes matrices vectorielles pour notre expérimentation en variant les types de critères en fonction : 1) de la segmentation (critères

unitaires simples, critères unitaires lemmatisés et critères adjacents) ; 2) du calcul (critères globaux et locaux) ; 3) de la catégorisation effectuée (critères thématiques, dialectiques et dialogiques). En dernier, nous voulons également faire une quatrième expérimentation avec tous les types de critères confondus.

Les tableaux 18 à 21 ci-dessous affichent les principales matrices constituées dans le cadre de chaque expérimentation et précisent sur quel type de critère textuel la classification sera effectuée.

Tableau 18. Expérimentation 1 : critères par type de segmentation

	Type de critère	**Type de donnée textuelle**
M1_Mots_simples	Critères unitaires simples	Mots simples
M2_Lemmes	Critères unitaires lemmatisés	Lemmes
M3_Segments	Critères adjacents	Segments répétés

Tableau 19. Expérimentation 2 : critères globaux et locaux

	Type de critère	**Type de donnée textuelle**
M4_Globaux	Critères globaux	Mots simples, lemmes et segments répétés obtenus par le calcul des spécificités
M5_Locaux	Critères locaux	Cooccurrents obtenus par le calcul de cooccurrence

Tableau 20. Expérimentation 3 : critères catégorisés

	Type de critère	**Type de donnée textuelle**
M6_Thématiques	Critères thématiques	Mots simples, lemmes, segments répétés et cooccurrents catégorisés comme thématiques
M7_Dialectiques	Critères dialectiques	Mots simples, lemmes et segments répétés catégorisés comme dialectiques
M8_Dialogiques	Critères dialogiques	Mots simples, lemmes et segments répétés catégorisés comme dialogiques

Tableau 21. Expérimentation 4 : tous les critères confondus

	Type de critère	Type de donnée textuelle
M9_Tous	Tous les critères	Mots simples, lemmes, segments répétés et cooccurrents

Pour évaluer la performance de notre démarche de fouille d'opinions, nous effectuons des tests qui sont confrontés aux résultats de classification obtenus à partir d'une matrice servant de ligne de comparaison. La constitution de cette matrice, désignée désormais comme ligne de comparaison (LC), exige un effort minimum sur le plan du traitement linguistique. Dans le cadre de cette recherche, nous avons constitué 2 lignes de comparaison : la LC1_motsSimples, contenant des mots simples extraits du corpus, sans aucun traitement computationnel ; la LC2_Lemmes, contenant des lemmes extraits du corpus. Nous avons également varié la pondération de ces deux lignes de comparaison afin d'évaluer l'effet de la pondération binaire et de la pondération par fréquence sur le résultat de la classification.

Dans chacune de ces expérimentations, nous voulons aussi tester la variation du score de spécificité pour connaître si le choix de ce dernier a un impact sur les résultats. Malgré le fait que les critères textuels que nous avons retenus ont des scores qui varient de +2 à +16, nous avons constaté que la plupart des critères (91 %) ont des scores variant entre +2 et +5. Ainsi, nous avons pu tester seulement la variation de scores se situant dans l'intervalle +2 et +5 dans les matrices.

3.4 Classification : choix des algorithmes

La Classification constitue la quatrième étape de la démarche méthodologique de la fouille d'opinions présentée dans cette recherche (figure 9, p.181)

Dans le cadre de ce travail, ces deux algorithmes ont été employés et leurs performances comparées pour la tâche de classification, l'objectif étant d'observer lequel est le plus adapté à la classification : l'algorithme machine à vecteurs de support (SVM) ou bayésiens naïfs (NB).

Le SVM construit un modèle probabiliste permettant de déterminer le seuil d'appartenance d'un objet à une classe. Opérant à l'intérieur d'un espace où les vecteurs sont représentés selon leur position (figure 14 ci-dessous), l'algorithme essaye de trouver une ligne séparant les vecteurs de façon optimale dans l'espace, appelée hyperplan (H). L'hyperplan doit maximiser la distance *d* qui l'écarte des vecteurs ayant servi à le trouver (les vecteurs supports), de manière à former deux sous-ensembles de vecteurs éloignés dans l'espace de représentation. Dans ce sens, la classification opérée par le SVM est binaire, car cette ligne sépare les objets en deux classes — les exemples positifs (+) et les exemples négatifs (*) (Ibekwe-SanJuan, 2007).

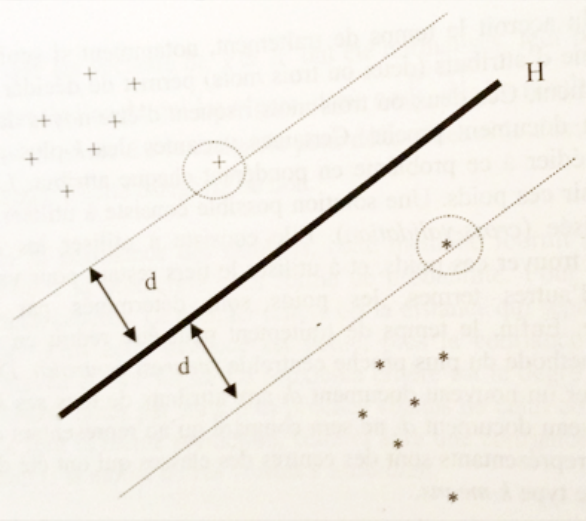


Figure 14. Exemple de classification de deux types d'objets (+, *) (Ibekwe-SanJuan, 2007)

L'algorithme affecte une classe (dans l'exemple de la figure ci-haut, les classes sont représentées par + ou *) à un nouvel objet selon la position que celui-ci occupe dans l'espace de représentation par rapport à l'hyperplan. La procédure confère un bon niveau de généralisation du classifieur si les objets à classer sont suffisamment différents des vecteurs supports. Dans le cas où les objets à classer ne sont pas linéairement séparables (les vecteurs sont très similaires), le SVM va les représenter dans un espace de dimension plus élevée (par exemple, dans un espace tridimensionnel), afin de retrouver une séparation linéaire optimale (Ibekwe-SanJuan, 2007).

Les classifieurs bayésiens naïfs « cherchent à prédire la valeur d'un nouvel objet, à partir d'une estimation des probabilités prenant en compte des connaissances ou des observations existantes » (Ibekwe-SanJuan, 2007, p. 99). Dans le cas de la classification de textes, les articles sont représentés par des vecteurs qui tiennent compte de la fréquence des données textuelles. La présence et l'absence de chaque mot sont traitées comme un attribut booléen indépendant et le modèle calcule la probabilité qu'un article appartienne à une classe donnée en fonction des fréquences existantes (Witten et Frank, 2005). Les modèles générés à partir du classifieur bayésien naïf produisent des estimations probabilistes pour affecter une classe à un article. Le théorème de Bayes, appliqué à la classification de texte, se formule comme suit :

$$P\ (R = r\ |e) = \frac{P(e|R = r)\ .P(R = r)}{P(e)}$$

Pour calculer la probabilité d'un article *R* d'appartenir à la classe *r* étant donné les évidences *e*, c'est-à-dire les mots qui se trouvent représentés dans le vecteur de l'article (P (R = r | e)), il faut multiplier la probabilité de *e* étant donné R = r ($P(e|R = r)$) par la probabilité qu'un article tiré au hasard appartienne à la classe r ($P(R = r)$), le tout divisé par la probabilité générale des évidences (P (e)), c'est-à-dire par la probabilité qu'un article sélectionné au hasard soit représenté par le vecteur de l'article *R* (Ibekwe-SanJuan, 2007).

Le choix de ces deux algorithmes dans le cadre de la recherche se base premièrement sur la nécessité d'avoir une comparabilité avec les études menées par Eensoo et Valette (2012, 2014a, 2014b, 2015), qui privilégient l'utilisation de SVM et NB.

Dans le cadre de notre expérimentation, nous avons utilisé le logiciel *Weka* (*Waikato Environment for Knowledge Analysis*) (Hall et coll., 2009), développé par la *Waikato University* en Nouvelle-Zélande. Les détails sur le fonctionnement du logiciel ainsi que les critères motivant son choix sont explicités sur l'Annexe C.

3.5 Évaluation du classifieur

L'évaluation du classifieur constitue la cinquième étape de la démarche méthodologique de la fouille d'opinions présentée dans cette recherche (figure 9, p.181)

L'évaluation de la performance du classifieur détermine s'il détient un niveau raisonnable de généralisation, c'est-à-dire s'il est capable de prédire correctement les classes de nouveaux articles. L'étape d'évaluation dans le cadre de notre recherche se constituera des étapes suivantes :

1. Constitution de corpus d'entraînement et corpus de test par la technique de validation croisée.

2. Mesurer la performance du classifieur pour chaque matrice constituée.

Ces étapes sont décrites dans les deux prochaines sections.

3.5.1 Validation croisée pour l'évaluation du classifieur

Pour chaque matrice constituée dans le cadre de cette recherche, nous avons employé la technique de validation croisée (*k-cross-validation*) lors de l'étape d'apprentissage, afin d'assurer que le classifieur ne soit pas biaisé par un manque de représentativité du corpus d'apprentissage. Cette technique consiste à répéter le processus de classification en variant le corpus de test et le corpus d'apprentissage à plusieurs reprises avec différents échantillons aléatoires afin de vérifier le taux d'erreur survenant dans la classification (Ibekwe-SanJuan, 2007 ; Witten et Frank, 2005). Le logiciel *Weka* (Hall et coll., 2009) offre l'option de choisir le type d'échantillonnage pour la validation. La fonctionnalité divise le corpus d'apprentissage en *k* sous-ensembles de taille égale, respectant la distribution de classes à l'intérieur de chaque sous-ensemble. Ensuite il échange ces sous-ensembles dans les fonctions de test et d'apprentissage : chaque sous-ensemble est utilisé comme corpus de test et les quatre sous-ensembles restants comme corpus d'apprentissage. En évaluant le résultat de chacun des cinq tests, le logiciel donne le résultat global de la classification dans les classes ETUD et GOUV en utilisant les mesures de rappel et de précision. Nous allons expliquer les mesures dans la prochaine section.

Dans le cadre de la recherche, nous avons choisi de faire une validation croisée à 5 itérations, dans le but d'avoir une bonne comparabilité avec des travaux qui ont aussi utilisé l'approche de sélection de critères textuels pour la fouille d'opinions (Eensoo et Valette, 2012, 2014a, 2014b, 2015).

3.5.2 Mesures d'évaluation

La plupart de travaux en fouille d'opinions font usage de mesures classiques pour évaluer la performance des algorithmes d'apprentissage de la méthode supervisée, comme la précision, le rappel et le *fscore*. Ces mesures ont été conçues initialement pour mesurer la performance de systèmes de repérage d'information, en évaluant la capacité du système à identifier l'ensemble des documents pertinents par rapport à la formulation d'une requête (Sokolova et Lapalme, 2009 ; Witten et Frank, 2005).

Le calcul de rappel et la précision mettent en rapport le nombre d'articles pertinents et non pertinents retournés par la requête par rapport au nombre d'articles pertinents et non pertinents existants. La matrice de confusion (tableau 22 ci-dessous) est un outil permettant de représenter visuellement ce rapport, en prenant les résultats d'un algorithme de recherche d'information ou de classification. Dans la matrice, la relation « oui » et « oui » entre classe prédite et classe réelle est un cas vrai positif (vp) : il s'agit du nombre d'objets retournés qui sont pertinents. Par conséquent, un vrai négatif (vn) est le nombre d'objets non retournés qui ne sont pas pertinents. Les faux positifs et les faux négatifs sont les objets retournés qui ne sont pas pertinents et les objets pertinents qui ne sont pas retournés.

Tableau 22. Matrice de confusion

		Classe prédite	
		Oui	**Non**
Classe réelle	**oui**	vrai positif (vp)	faux négatif (fn)
	non	faux positif (fp)	vrai négatif (vn)

Transportées dans le contexte de la classification automatique de textes, ces mesures ont été proposées pour évaluer la capacité du classifieur à identifier, pour chaque classe, le nombre d'articles correctement affectés à la classe. L'exactitude de la classification est évaluée en considérant tous les articles qui ont été reconnus comme membres ou non membres de la classe (vp et vn) et les articles qui ont été erronément attribués ou non attribués à la classe (fp et fn). Les mesures de rappel et de précision sont obtenues de la manière suivante :

$$\text{Rappel} = \frac{vp}{vp + fn}$$

$$\text{Précision} = \frac{vp}{vp + fp}$$

Le rappel mesure le taux d'articles correctement affectés à une classe par rapport au nombre d'articles existants pour la classe. La précision mesure le nombre d'articles correctement affectés à une classe par rapport à tous les articles existants.

Sokolova et Lapalme (2009) soulignent que ces mesures négligent la classification d'objets qui ont été correctement reconnus comme non membres de la classe (vn) :

> For topic classification (e.g., identification of documents about a given city or documents about bands and artists, etc.) documents are simply classified as being relevant to the topic or not; hence, classes are built as positive vs everything else. Retrieval of relevant documents being the more important task, the focus in this case is on true positive classification. (…) The formulas for these measures neglect the correct classification of negative examples, they instead reflect the importance of retrieval of positive examples in text/document classification. (Sokolova et Lapalme, 2009, p. 428)

Ainsi, les mesures de rappel et de précision sont efficaces lorsqu'il s'agit d'évaluer l'attribution d'une seule classe aux objets, puisqu'elles considèrent seulement les vrais positifs dans le calcul. Pourtant, lorsqu'il y a plus de deux classes à prédire, les objets qui n'ont pas été reconnus comme membres de la classe en question, mais qui ont été classés correctement comme non membres de la classe sont importants pour évaluer l'exactitude du modèle.

À partir d'un travail comparatif des principales mesures d'évaluation appliquées aux tâches de classification automatique de textes, Solokova et Lapalme (2009) ont démontré que dans une classification où il y a deux classes bien définies (par exemple la classification binaire de sentiments dans les classes positives et négatives), la mesure de l'AUC (*Area Under the Curve*) peut fournir une évaluation plus fiable que le rappel et la précision. L'AUC est une mesure capable d'identifier l'habileté du classifieur à prédire correctement les membres et les non membres d'une classe. Autrement dit, il permet d'évaluer la capacité du classifieur à éviter la fausseté de la classification :

$$AUC = (\frac{vp}{vp + fn} + \frac{vn}{vn + fp})$$

Dans le cadre de notre recherche, nous allons vérifier en plus du rappel et de la précision, la mesure AUC, afin d'obtenir une mesure plus fiable pour comparer nos résultats aux lignes de comparaisons.

4. Exploration de critères textuels dans le corpus pour des fins de recommandation

Dans cette section, nous traitons des démarches méthodologiques pour vérifier la pertinence de notre méthodologie dans le développement des SRAP.

Afin que la démarche méthodologique de fouille d'opinions proposée dans cette recherche soit applicable dans un contexte réel, il faut que l'ensemble de critères textuels qui ont démontré une meilleure performance pour la tâche de classification puisse être élaboré à partir d'un ensemble de textes parus au début de la controverse. De cette façon, il serait possible tout au début de la controverse de construire un classifieur capable de prédire la classe des articles qui font leur parution subséquemment dans le conflit.

À cette étape, nous voulons explorer, à l'aide de calculs textométriques, à quel moment de la controverse nous pouvons observer l'apparition des critères textuels dont la pertinence et l'efficacité ont été démontrées à l'étape de classification. En supposant que les critères textuels les plus performants pour la tâche de classification font leur apparition de manière anticipée dans le corpus, nous pouvons statuer sur l'applicabilité de notre démarche méthodologique dans un contexte de recommandation. Cette étape correspond à l'objectif de recherche 3 :

Objectif 3 : Effectuer des analyses textométriques sur différentes périodes du corpus et explorer à quel moment il est possible de retrouver l'ensemble des critères textuels plus performants pour la tâche de classification.

La démarche méthodologique pour atteindre cet objectif est d'ordre exploratoire et est constituée des étapes suivantes :

1. Sélectionner la matrice qui a démontré une meilleure performance en termes de rappel et de précision pour la tâche de classification. Dresser la liste des critères textuels qui composent cette matrice.

2. Diviser le corpus de la recherche en sous-ensembles. Ces petits sous-corpus seront construits sous forme cumulative, suivant un ordre chronologique (tableau 23 ci-dessous).

3. Réaliser une étude textométrique de chaque sous-corpus et observer l'apparition des critères textuels, afin de déterminer dans quel moment les critères de la matrice ayant démontré une meilleure performance surviennent dans la controverse.

4. Tester la performance des critères textuels qui surviennent au début de la controverse pour prédire la classe des articles parus postérieurement.

Tableau 23. Découpage chronologique du corpus

Sous corpus	**Période**	**Nombre d'articles**
Février	15/02/2012 à 29/02/2012	41
Mars	02/03/2012 à 28/03/2012	15
Avril	01/04/2012 à 30/04/2012	216
Mai	01/05/2012 à 31/05/2012	58
Juin	01/06/2012 à 30/06/2012	135
Juillet	02/07/2012 à 18/07/2012	5
Août	01/08/2012 à 29/08/2012	13
Septembre	05/09/2012 à 27/09/2012	9
Octobre	16/10/2012 à 25/10/2012	3

Nous avons retenu les 3 premiers mois de discussion sur le conflit étudiant, qui a pris fin le 4 septembre 2012 avec l'annulation par décret de la hausse des droits de scolarité et la suppression du projet de loi 78 par le gouvernement de Pauline Marois. Les dates correspondent aux articles présents dans notre corpus. La majorité de ces articles sont parus au mois d'avril 2012.

5. Conclusion du chapitre

Notre recherche propose de 1) développer et de valider une démarche méthodologique de fouille d'opinions basée sur les techniques de la textométrie pour l'identification de textes véhiculant des opinions divergentes dans une controverse et 2) d'explorer la pertinence de cette démarche méthodologique pour le développement d'un SRAP. Pour le premier objectif, nous avons proposé un modèle de fouille d'opinion qui comprend deux étapes clés : l'étape « Filtrage », pour effectuer l'analyse textométrique de deux sous-corpus dont les textes véhiculent des opinions divergentes, afin de sélectionner les critères textuels qui différencient chaque camp du débat sur les plans thématique, dialectique et dialogique ; et l'étape « Transformation », dans laquelle les critères issus de l'analyse textométrique sont utilisés pour donner une représentation des textes dans le modèle vectoriel. Pour répondre au deuxième objectif, nous avons proposé d'effectuer une analyse chronologique du corpus, afin de savoir si les critères textuels les plus performants pour la classification d'articles dans le temps peuvent être sélectionnés par une approche textométrique au début de la controverse.

Les éléments théoriques de la sémantique interprétative ont des implications sur la démarche méthodologique adoptée, orientant le choix des outils et des étapes du traitement (Pincemin, 1999a). La constitution du corpus et la sélection de critères à partir de l'analyse textométrique des sous-corpus contrastés sont en consonance avec les concepts exposés dans le cadre théorique. La description des parcours interprétatifs par le biais des statistiques vise à donner une vision globale et synthétique des propriétés sémantiques en œuvre dans le corpus, permettant de caractériser les stratégies argumentatives élaborées par chaque groupe dans la discussion. Cette description confère une valeur ajoutée dans le cadre de la démarche, dans la mesure où elle révèle les contrastes existants dans cette élaboration argumentative et justifie la sélection de critères textuels en fonction de l'interprétation de ces derniers.

Le corpus dans notre recherche est constitué des articles d'opinion traitant de la grève étudiante au Québec en 2012 publiés dans les principaux journaux de la province. Les textes ont été placés dans les classes ETUD ou GOUV selon leur tendance à favoriser les enjeux défendus d'un des groupes d'auteurs qui s'oppose à l'autre dans le débat. Nous entendons faire par la suite l'étude contrastive de ces deux sous-corpus pour connaître les critères

textuels spécifiques à chacun des groupes d'auteurs et analyser les récurrences sémantiques permettant de caractériser le contenu des sous-corpus sur le plan des trois composantes sémantiques qui ont été retenues dans notre cadre théorique. Cette analyse a pour objectif d'expliciter les différences entre les textes de chaque groupe d'auteurs et d'obtenir des critères textuels interprétables et linguistiquement pertinents pour le développement du classifieur.

La prochaine étape vise à donner une représentation informatique des textes du corpus en fonction des critères textuels retenus dans l'étape descriptive. Nous allons par la suite tester différents algorithmes d'apprentissage de la méthode supervisée pour prédire la classe des articles. L'étude entend également connaître le type de critère et l'algorithme d'apprentissage les plus performants (machine à vecteurs de support ou bayésien naïf). Pour l'évaluation des algorithmes, nous avons utilisé les mesures de rappel et de précision, ainsi que l'AUC.

Afin de déterminer l'applicabilité de notre démarche méthodologique dans un SRAP, nous avons réalisé des analyses sur le corpus pour connaître à quel moment il est possible de trouver, par la méthode textométrique, les critères textuels qui ont démontré la meilleure performance pour la tâche de classification automatique.

Chapitre 4. Résultats

1. Introduction

Ce chapitre présente les résultats de la recherche, apportant des réponses aux trois questions de recherche posées. Premièrement, les résultats de la première question sont présentés, concernant la sélection des critères textuels interprétables et compatibles avec les composantes thématique, dialectique et dialogique. Ensuite, nous allons présenter les résultats de la deuxième question, qui porte sur les expérimentations avec les critères textuels sélectionnés pour la classification d'articles d'opinion provenant du corpus de référence constitué. Finalement, les résultats de la troisième question sur l'exploration chronologique de notre corpus sont présentés, permettant de se prononcer sur l'adaptabilité de la démarche de sélection de critères textuels pour le développement des SRAP.

2. Question de recherche 1 : Critères textuels permettant de caractériser et de distinguer les opinions exprimées dans les corpus ETUD et GOUV sur le plan thématique, dialectique et dialogique

2.1. Introduction

La première question de la recherche est formulée de la manière suivante : *dans un corpus d'articles sur la Grève étudiante au Québec en 2012, lequel a été classé préalablement par un humain dans des classes opposées, quels sont les critères textuels permettant de caractériser et de distinguer les opinions exprimées dans les articles de chaque classe sur le plan thématique, dialectique et dialogique ?*

Dans le schéma de fouille d'opinions proposé dans le chapitre 3 (figure 9, p. 182), cette question renvoi à la sélection de critères textuels interprétables et compatibles avec les composantes sémantiques, expliquées dans le même chapitre (p. 157 à 168). L'objectif de cette étape est de faire l'analyse textométrique de notre corpus, en contrastant les articles favorables et défavorables à la grève des étudiants (ETUD et GOUV), afin de trouver les critères textuels qui les différencient sur le plan thématique, dialectique et dialogique.

Nous avons proposé d'explorer deux calculs textométriques pour la sélection de critères textuels. Le calcul des spécificités explore le niveau de contextualisation globale, puisqu'il statue sur la représentativité des données textuelles dans un sous-corpus, en tentant compte des probabilités d'apparition de ces mêmes données dans l'ensemble des articles regroupés dans le corpus de référence. Le calcul de cooccurrence explore le niveau de contextualisation locale de chaque sous-corpus, dans la mesure où il identifie les affinités lexicales entre des couples de mots dans des zones plus circonscrites de la chaîne textuelle, comme les phrases et les paragraphes.

Dans cette section, nous allons d'abord présenter les résultats de l'application du calcul des spécificités et le calcul de cooccurrence pour l'élaboration des critères textuels globaux et

locaux. Nous voulons dans cette première partie expliciter les résultats et les décisions prises pour la sélection de ces critères. Il est question également de l'élimination de critères considérés comme non pertinents, selon le cadre défini dans notre méthodologie. Par la suite nous allons caractériser les critères textuels sélectionnés, en décrivant les récurrences sémiques identifiées sur le plan thématique, dialectique et dialogique. Nous allons décrire les sous-corpus ETUD et GOUV de façon contrastive en fonction des composantes sémantiques et des structures sémantiques identifiées (isotopies, isosémies et molécules sémiques).

La présente section est organisée en fonction des étapes de la démarche de sélection de critères textuels tels que nous l'avons exposée dans la méthodologie (tableau 17, p. 222) : l'étape 1 comprend l'application des calculs contrastifs, l'étape 2 comprend l'identification des structures sémantiques et l'étape 3, la description des composantes.

2.2 Étape 1 : application des calculs de spécificités et de cooccurrence

Cette section décrit les résultats des démarches effectuées pour la sélection de critères textuels à partir du calcul des spécificités et du calcul de cooccurrences, permettant d'élaborer l'ensemble de critères globaux et locaux pour notre expérimentation. Les résultats de chacun de ces calculs seront présentés dans les deux sections suivantes.

2.2.1 Résultats du calcul des spécificités pour l'élaboration de critères globaux

En utilisant les logiciels *TXM* et *Lexico3,* nous avons calculé les spécificités des sous-corpus ETUD et GOUV pour trois différents types de segmentation : mots simples, lemmes et segments répétés. Nous n'avons retenu pour l'analyse que les données textuelles dont le score de spécificité était plus grand ou égal à +2 dans chacun des sous-corpus considérés. Le tableau 24 ci-dessous présente, pour chaque type de segmentation, le nombre total de données textuelles considérées dans notre analyse.

Tableau 24. Effectif de données textuelles avec score de spécificité ≥ +2
dans les sous-corpus ETUD et GOUV

Type de segmentation	**Nombre de données textuelles ≥ +2**
Mot simple	344
Lemme	309
Segment répété	577
Total	**1230**

Nous avons ensuite procédé à l'élimination de données textuelles qui ne satisfaisaient pas les conditions de sélection définies dans notre recherche. Nous avons ainsi exclu :

1. Les données textuelles présentes dans un seul document du sous-corpus
2. Les données textuelles dont la fonction grammaticale ou la signification en contexte étaient inconstantes et par conséquent, non pertinentes du point de vue linguistique.

Pour identifier les données textuelles qui n'étaient pas pertinentes du point de vue linguistique, nous avons utilisé le concordancier de *TXM*. Celle-ci permet de voir le contexte immédiat entourant la donnée textuelle et aider à comprendre comment cette dernière est employée dans le texte.

Dans la section suivante, nous exposons d'autres détails motivant l'élimination de données textuelles non pertinentes.

a) Éliminations de données textuelles non pertinentes provenant du calcul des spécificités

La plupart des éliminations effectuées sur le résultat du calcul des spécificités correspondaient aux critères établis dans le cadre de cette recherche. En premier lieu, les données textuelles présentes dans un seul document ont été éliminées. Par exemple, le mot

simple ‘marre’ a été éliminé malgré son score de spécificité relativement élevé (+5), puisqu’il figurait sur un seul document dans le sous-corpus GOUV. Deuxièmement, les mots simples, les lemmes ou les segments répétés dont la fonction grammaticale ou la signification n’était pas constante ont été éliminés. Deux des exemples les plus évidents de ces cas sont les lemmes ‘avoir’ et ‘être’, survenus parmi les spécificités positives des sous-corpus. Comme il s’agit aussi de verbes auxiliaires, ces données figuraient sur d’innombrables contextes et n’étaient pas pertinentes d’un point de vue linguistique.

Nous avons remarqué la présence de quelques mots simples spécifiques en majuscule dans le sous-corpus GOUV, tels que ‘LES’, ‘DEUX’ et ‘DANS’. Dans la plupart des cas, la spécificité de ces mots était le résultat de leur présence dans les titres d’articles de certains journaux qui présentaient ce style graphique particulier. Dans d’autres cas, le mot en majuscules figurait dans le texte lui-même et faisait état d’un style d’écriture polémique. Par exemple, le mot simple ‘TOUS’, spécifique au sous-corpus GOUV (+2,71), figurait dans des contextes où l’auteur voulait renforcer le sens du mot « tous » (voir les exemples ci-dessous). Nous avons décidé de conserver les mots majuscules uniquement lorsque ceux-ci étaient liés à une stratégie argumentative et d’éliminer les mots simples spécifiques qui relevaient du style graphique des journaux[29].

> [news20120314OP120314230493361] (…) artiste porte un carré rouge ? Ils vont *TOUS* porter un carré rouge. So-so-solidarité, les amis. Et marchons (…)

> [news20120327OP120327232623063] Tu acceptes TOUT ce qu’ils font et *TOUS* les moyens qu’ils utilisent pour arriver à leurs fins, ou (…)

[29] Une des caractéristiques de la plateforme *Eureka.cc,* que nous avons utilisée dans le cadre de notre recherche, nous a amenés à ne pas explorer les données textuelles relevant du style graphique des documents. En effet, les documents qui sont intégrés dans cette plateforme sont souvent soumis à un traitement visant à l’uniformisation du format des articles, afin de correspondre au format de présentation de la plateforme et non de la source originelle. Ainsi, le fait de considérer des informations de surface dans le cadre de notre expérimentation pourrait nous induire à des conclusions erronées sur l’utilisation de certains styles graphiques dans les composantes d’articles comme les titres.

[news20120425OP120425239243689] (…) Dans la poche des contribuables. De *TOUS* les contribuables. 100 % des contribuables paient pour que 28 % (…)

En ce qui concerne les lemmes, certaines éliminations ont été motivées par des problèmes de traitement du lemmatiseur *TreeTagger*, intégré dans *TXM*. Par exemple, nous avons observé dans GOUV la spécificité du lemme 'frai | frais'. La recherche de concordances du lemme 'frai' retourne, en plus de 'frais', le mot simple 'fraîche'. Cela montre une erreur de lemmatisation du mot « frais », lequel a été interprété dans certains contextes comme pluriel de « frai ». Cette même observation s'applique au lemme 'cour | cours', employé avec une diversité de sens dans les textes GOUV où il se trouve spécifique ('cour juridique', 'dans ma cour', 'cours universitaire', 'au cours des années', etc.). Nous n'avons pas pris en compte les lemmes qui présentaient ce problème lors de la sélection de critères textuels.

Du côté de la spécificité de segments répétés, nous avons observé des résultats très divers et non pertinents du point de vue linguistique. Certains segments répétés correspondaient à des expressions lexicales figées comme 'grève étudiante' ou 'droit de manifester'. Cependant, plusieurs mots lexicaux suivis ou précédés de prépositions ou d'articles sont survenus parmi les segments répétés spécifiques. Dans ce dernier cas, nous avons pris la décision de les éliminer et de ne conserver que les expressions lexicales figées ou semi-figées. Ainsi, des segments comme 'la gauche', 'la majorité des' et 'pour dénouer' ont été éliminés. Par contre, lorsqu'il s'agissait de mots grammaticaux, ce critère restrictif n'a pas été considéré. Nous avons sélectionné comme critères textuels tous les segments répétés qui étaient révélateurs de la dialogique et de la dialectique des textes, en fonction des contextes analysés et aussi en fonction de la récurrence des traits sémantiques qui confirmait la présence d'un certain phénomène. Par exemple, la spécificité de 'ils' chez GOUV, référant fréquemment aux étudiants et signalant une position de distanciation de l'énonciateur, a motivé la sélection de segments faisant état du même phénomène, tels que 'et ceux', 'leurs membres', 'leurs cours', 'ils ont', 'ils sont'. Ces segments, en plus de référer textuellement dans la plupart des cas aux étudiants, indiquent également une dialogique centrée sur la troisième personne du pluriel.

Notre calcul des spécificités pour les segments répétés a pris comme contexte une fenêtre contenant jusqu'à 10 mots, afin de laisser au calcul statistique le soin de repérer le plus grand nombre possible d'expressions employées par les auteurs des textes. Plusieurs résultats étaient ainsi répétés et il nous a fallu faire un choix entre des segments qui étaient porteurs du même sens. Nous avons préféré les segments plus courts et sans la présence d'articles ou de prépositions. Ainsi, entre 'la hausse des frais de scolarité' et 'la hausse des frais', nous avons gardé seulement le dernier. Même si la spécificité du premier segment s'avérait plus grande, le deuxième a été retenu comme critère, puisque la chaîne de caractères 'hausse des frais' se trouve incluse dans 'hausse des frais de scolarité'. Pareillement, entre 'des prêts et bourses' et 'prêts et bourses', le dernier a été retenu comme critère textuel.

Nous avons remarqué, parmi les mots, les lemmes et les segments répétés spécifiques, plusieurs noms propres. Certains de ces noms apparaissaient deux fois dans les données textuelles ressorties, soit le prénom, soit le nom de famille. Par exemple, parmi les mots simples spécifiques d'ETUD se trouvait 'Martine' et 'Desjardins', les deux renvoyant à la même personne. Lors de la sélection des critères de type mot simple ou lemme, nous avons pris la décision de garder seulement les noms qui avaient un score de spécificité plus haut. Ainsi, même si 'Pellerin' et 'Fred' renvoie à la même personne, 'Pellerin' a été conservé en raison de son score de spécificité plus élevé.

Nous avons également éliminé les entêtes des articles qui sont apparus comme spécifiques dans la liste de mots simples et de segments répétés. Les mots simples 'libre' et 'lettres', ainsi que le segment répété 'Conflit étudiant' n'ont pas été considérés.

L'Annexe D contient la liste de toutes les données textuelles éliminées, soit pour les raisons mentionnées dans cette section, soit à cause du critère de distribution et de pertinence linguistique définies dans le cadre de notre recherche. Nous avons spécifié dans le tableau qui figure dans l'annexe les raisons motivant chacune de ces éliminations.

b) Total de critères globaux retenus

Le Tableau 25 ci-dessous présente l'effectif total de critères textuels globaux sélectionnés en fonction de leur pertinence linguistique et de leur distribution sur les sous-corpus.

Tableau 25. Effectif total de critères textuels sélectionnés
avec score de spécificité ≥ +2

Type de segmentation	**Nombre**
Mot simple	241
Lemme	208
Segment répété	194
Total	643

2.2.2 Résultats du calcul de cooccurrences pour l'élaboration de critères locaux

Nous avons effectué le calcul de cooccurrence pour chaque mot pôle choisi dans les sous-corpus ETUD et GOUV, comme nous l'avons expliqué au chapitre 3 (p. 212). Le calcul a été réalisé avec *TXM* sur les lemmes du corpus. Afin de repérer un maximum de cooccurrents, nous avons défini un seuil de fréquence minimale de 2, permettant de ressortir les lemmes cooccurrents qui apparaissent au moins 2 fois dans les sous-corpus. Nous avons également défini une fenêtre contextuelle contenant 10 mots à la droite et 10 mots à la gauche du mot pôle. Ce seuil a été défini de façon expérimentale, en observant si la variation de la fenêtre contextuelle apportait des différences significatives sur le nombre de cooccurrents retrouvés en forte association. Nous avons observé qu'en dessous de 10, le nombre de cooccurrents ayant un score négatif augmente, et les résultats ne permettent pas de repérer de nouveaux cooccurrents.

TXM permet d'ordonner les résultats du calcul de cooccurrence selon la fréquence d'apparition du mot cooccurrent, la co-fréquence de ce dernier avec le mot pôle, l'indice de spécificité et la distance moyenne entre les cooccurrents. L'indice de spécificité est un indicateur statistique de présence, faisant intervenir la surreprésentation du cooccurrent sur l'ensemble des fenêtres contextuelles où se retrouvent les mots pôles, la co-fréquence du cooccurrent avec le mot pôle ainsi que la distance moyenne entre les deux (Lafon, 1980). Afin de pouvoir analyser le contexte d'un plus grand nombre de cooccurrents, nous avons seulement éliminé ceux qui ont présenté un indice de spécificité négatif.

Des résultats générés, nous avons seulement conservé les lemmes étiquetés comme substantifs, verbes, adverbes ou adjectifs par *TreeTagger*. Ce choix d'écarter d'autres catégories morphosyntaxiques est motivé par l'objectif de la démarche de sélection de critères locaux, expliqué dans la méthodologie, et qui consiste à ne repérer que des mots lexicaux liés à la thématique du corpus.

Le tableau 26 ci-dessous présente l'effectif total de cooccurrents retenus dans chaque sous-corpus et qui ont été considérés dans notre analyse. Au total, 354 cooccurrents ont été repérés pour chaque mot pôle dans les deux sous-corpus.

Tableau 26. Effectif de cooccurrents dans les sous-corpus ETUD et GOUV pour chaque mot pôle

Mot pôle	**ETUD**	**GOUV**
étudiant	31	26
Québec	24	24
scolarité	29	26
gouvernement	39	40
droit	32	24
hausse	32	27
Total	**187**	**167**

a) Éliminations de cooccurrents non pertinents

La démarche de sélection de critères thématiques locaux considère d'une part seulement les cooccurrents qui sont uniques à chaque sous-corpus et qui sont en plus différents des critères globaux sélectionnés par la méthode de spécificité. Le calcul de cooccurrence a permis en effet de repérer de nouveaux critères thématiques spécifiques à chaque sous-corpus et qui ne figuraient pas parmi les critères globaux sélectionnés, ce qui confirme la pertinence de la démarche.

Pour la sélection des critères, nous avons d'une part éliminé les cooccurrents pour un même mot pôle qui étaient communs aux deux sous-corpus. Par exemple, le mot pôle « scolarité » était fortement associé à « augmentation » dans les sous-corpus ETUD et GOUV.

Finalement, pour nous assurer de la pertinence linguistique des cooccurrents spécifiques de chaque sous-corpus, nous avons évalué leurs contextes d'apparition en utilisant le concordancier de *TXM*. Nous avons éliminé les cooccurrents qui étaient présents dans un seul document et aussi ceux qui étaient employés avec des fonctions grammaticales ou des significations différentes dans les textes. Nous avons également éliminé des concurrents associés à des noms d'organisations ou à des noms propres, comme 'scolarité-perron' et 'Québec-sûreté'. La liste de cooccurrents éliminés ainsi que les raisons motivant leur élimination se retrouvent également dans l'Annexe D.

b) Total de critères locaux retenus

Le tableau 27 ci-dessous présente l'effectif total de critères locaux sélectionnés pour chaque mot pôle dans les sous-corpus ETUD et GOUV. Un total de 108 cooccurrents a été retenu comme critères locaux. Au total, nous avons éliminé 77 cooccurrents du sous-corpus ETUD et 69 cooccurrents du sous-corpus GOUV.

Tableau 27. Effectif total de critères locaux sélectionnés

Mot pôle	**ETUD**	**GOUV**
'étudiant'	22	21
'Québec'	13	10
'scolarité'	20	11
'gouvernement'	22	25
'droit'	17	13
'hausse'	16	18
Total	**110**	**98**

2.3 Étape 2. Identification de structures sémantiques

Le concordancier de *TXM* a rendu possible l'analyse des contextes pour procéder à l'élimination de données textuelles non pertinentes ainsi qu'à la sélection de critères textuels compatibles avec les composantes thématiques, dialectiques et dialogiques. Dans le cadre de

notre travail, la sélection de critères a été accompagnée d'un travail d'interprétation et de catégorisation de chaque critère dans la composante, afin d'élucider leur pertinence et de mettre au clair les stratégies d'élaboration textuelle des sous-corpus ETUD et GOUV.

Nous avons rédigé des mémos qui décrivent les contextes d'apparition de chaque critère textuel retenu, lesquels figurent sur l'Annexe E. Nous avons essayé, dans cette description, d'expliquer comment le critère textuel participait à la structuration du type de phénomène reliée à la composante, que ce soit par exemple la présence d'un thème particulier (thématique), ou par la récurrence d'un mode de structuration argumentative et temporelle (dialectique) ou énonciative (dialogique). Nous avons commencé à décrire les contextes des critères textuels plus spécifiques des sous-corpus en passant par ceux dont les scores de spécificité positive étaient plus bas, et ce pour chaque type de segmentation (mot simple, lemme et segment répété). Cette description a été faite de façon cumulative et corrélationnelle, dans la mesure où les récurrences qui étaient observées et qui renforçaient une hypothèse interprétative identifiée au début de l'analyse ont été décrites subséquemment en fonction de cette hypothèse.

La catégorie grammaticale a été l'un des critères pour la sélection et la catégorisation des critères textuels, tel que nous l'avons défini dans la méthodologie (p. 218). Les critères de nature lexicale comme les substantifs, les adjectifs et les verbes ont été majoritairement catégorisés comme thématiques. Cela a été le cas des critères globaux spécifiquement lexicaux comme 'grève', 'boycott' et 'artistes'. Comme défini dans notre méthodologie (p. 218), tous les critères locaux repérés par le calcul de cooccurrence ont été catégorisés comme thématiques.

Nous avons catégorisé comme critères dialectiques tous les indices de temporalité, ainsi que les critères textuels permettant de caractériser les styles argumentatifs de chaque groupe. Plusieurs adverbes de temps ont été obtenus dans les résultats et décrits dans les mémos en fonction de leur aspect (inchoatif, duratif ou terminatif). Nous avons noté également la présence de dates liées à des événements historiques cités par les auteurs du sous-corpus ETUD. D'autres indices plus particuliers à la structuration de l'argumentation ont été catégorisés comme dialectiques. Nous avons remarqué la présence d'adverbes et de conjonctions de coordination ('Mais', 'mais aussi', 'Bref', 'encore'), mais aussi des indices de

questionnement (‘?’) et de phrases négatives (‘pas’). La description de ces phénomènes dans les mémos rédigés vise à rendre compte des caractéristiques de l’argumentation de chaque groupe, de la différence entre les stratégies adoptées pour défendre leurs opinions et du caractère polémique qui a été révélé par la récurrence de certains traits sémantiques, selon les caractéristiques du discours polémique identifiées par Garand (1998) et que nous avons présentés dans la méthodologie (p.219).

Parmi les critères dialogiques retenus se trouvent notamment les pronoms personnels. Dans le corpus GOUV, nous avons remarqué la mise à distance de l’énonciateur et la spécificité de certains segments répétés qui faisaient référence à la troisième personne, représentant les étudiants : ‘eux-mêmes’, ‘leurs cours’, ‘leurs membres’. Nous avons considéré comme dialogiques certains accords verbaux liés à la troisième personne lorsque ceux-ci faisaient référence aux étudiants dans le sous-corpus GOUV. Ainsi, ‘qui ne sont’ ‘ont pas’ ‘les’, ‘par les étudiants’, ‘les étudiants eux’, ont été catégorisés comme dialogiques. Des verbes ou expressions qui communiquent le positionnement ou la pensée de l’énonciateur ont été catégorisés comme dialogiques, parmi lesquels se trouvent certains modes verbaux (le conditionnel présent de la troisième personne ‘seraient’), les conjonctions de subordination exprimant l’hypothèse (‘si les’, ‘si on’), les formulations attribuées à la perception subjective de l’énonciateur (‘je pense’, ‘je vois’, ‘suis sûr’) ou les attitudes de l’énonciateur envers son interlocuteur (impératif : ‘il ne faut pas’).

Certains critères de nature lexicale se sont avérés plus caractéristiques de la composante dialogique. Le sous-corpus GOUV présentait le critère textuel ‘blogue’, associé à des liens vers le blogue du chroniqueur qui a signé l’article. Dans ETUD, les auteurs avaient une proximité avec leur public, comme démontre la spécificité des liens vers des adresses courriel, ainsi que la spécificité du mot ‘joindre’ (‘michelle.ouimet@lapresse.ca’, ‘joindre notre chroniqueur’). Nous avons caractérisé ces critères comme dialogiques puisqu’ils renseignent sur la volonté des énonciateurs d’ETUD à être en contact avec leur public de manière plus directe que celle observée dans le corpus GOUV. Cette proximité de l’énonciateur dans ETUD est aussi mise en évidence par les critères textuels spécifiques figurant sur la signature des articles. La spécificité de ‘Université’ et ‘Faculté’ à côté des noms

des chroniqueurs montre le souhait des auteurs d'identifier leur appartenance à une institution, un phénomène qui est plutôt rare dans le sous-corpus GOUV.

Même si certaines catégories grammaticales étaient plus caractéristiques d'une composante que de l'autre, nous avons quand même gardé une souplesse dans la catégorisation, en nous basant avant tout sur les contextes d'emploi. La catégorisation de signes de ponctuation représente un cas typique de la prééminence donnée à la prise en compte du contexte. Par exemple, la spécificité de '% en' dans ETUD était liée au thème de la hausse et illustrait une préoccupation de la proportion de l'augmentation au cours des années (75 % en cinq ans). Par contre, la spécificité des guillemets '« »' démontrait une prédominance du discours rapportée dans les textes, phénomène plus caractéristique de la dialogique. La liste des marqueurs que nous avons définis dans la méthodologie (p. 218) nous a guidés dans la catégorisation des critères, mais nous avons privilégié l'analyse du contexte d'emploi de la donnée textuelle analysée, pour la catégoriser comme critère textuel d'une composante sémantique.

2.3.1 Effectif de critères textuels retenus par type

Le tableau 28 ci-dessous résume le nombre de critères textuels retenus correspondants aux types définis dans notre recherche. Nous avons créé 8 types de critères en fonction du type de segmentation textuelle (critères unitaires simples, critères unitaires lemmatisés, critères adjacents), du type de calcul appliqué pour l'obtention du critère (critères globaux, critères locaux) et de la catégorisation effectuée dans les composantes sémantiques (critères thématiques, dialectiques et dialogiques). La catégorie « Tous les critères » est le dernier type et rassemble tous les types de segmentation (mots simples, lemmes et segments répétés) et par conséquent, il comprend tous les autres types de critères. À l'exception des critères unitaires simples, des critères unitaires lemmatisés et des critères adjacents, tous les autres types de

critères textuels contiennent tous des mots, des lemmes et des segments répétés. Nous avons spécifié cette information dans la colonne « Type de donnée textuelle » du tableau 28.[30]

Tableau 28. Typologie de critères textuels avec le nombre de critères sélectionnés

Types de critères	**Type de donnée textuelle**	**Nombre de critères**
Critères unitaires simples	Mot simple	241
Critères unitaires lemmatisés	Lemme	208
Critères adjacents	Segment répété	184
Critères globaux	Mot simple, lemme et segment répété	633
Critères locaux	Cooccurrent	204
Critères thématiques	Mots simple, lemme, segment répété et cooccurrent	620
Critères dialectiques	Mot simple, lemme et segment répété	107
Critères dialogiques	Mot simple, lemme et segment répété	110
Total de critères	Mots simple, lemme, segment répétés et cooccurrent	837

2.4 Étape 3 : description des composantes thématiques, dialectiques et dialogiques

La catégorisation des critères textuels a été faite de façon à expliciter la lecture interprétative opérée au sein du corpus. Nous avons identifié les récurrences sémiques de chaque sous-corpus et nous avons essayé de contraster les observations sur ces récurrences d'un sous-corpus à l'autre, de façon à tirer au clair leurs oppositions sur le plan des composantes sémantiques. Cette section présente le résultat de cette analyse. Nous voulons présenter les structures sémantiques identifiées (isotopies, isosémies et molécules sémiques) et

[30] Parmi les critères globaux de type lemmes et mot simple, il existe plusieurs mots qui peuvent être retrouvés dans certains lemmes repérés et qui sont porteurs de la même signification. Nous présentons dans ces tableaux le nombre de critères textuels total retenus, sans décompter ceux qui peuvent se trouver en double.

décrire chaque sous-corpus sur le plan thématique, dialectique et dialogique en fonction de ces structures et des critères textuels qui leur sont associés.

Comme énoncé dans le chapitre 3 (p. 218), nous avons appuyé la description de critères textuels sur les concepts descriptifs de la sémantique interprétative exposée dans le cadre théorique, mais aussi sur la « Grammaire du sens et de l'expression » de Charaudeau (1992). L'utilisation de cette dernière a été particulièrement pertinente pour aider à l'interprétation de récurrences de critères textuels dialectiques et dialogiques. Nous utilisons les catégories du discours de Charaudeau pour aider le travail d'interprétation, tout en posant que le principal gage de cette interprétation est le contexte dans lequel le critère textuel apparaît. Sur le plan thématique, nous avons décrit les récurrences de sèmes rencontrés afin de dégager les différences sur la façon de décrire les événements reliés à la grève étudiante et pour rendre explicite le découpage référentiel opéré par les auteurs favorables et défavorables à la grève étudiante. Nous avons également associé les stratégies argumentatives rencontrées dans les sous-corpus à celles décrites par Garand (1998) à propos du discours polémique.

La description que nous allons présenter dans cette section vise à apporter une interprétation sur les résultats observés, qui justifie la sélection de critères textuels en fonction des récurrences rencontrées et du rapport de ces récurrences avec les stratégies argumentatives élaborées par chaque groupe dans la discussion. Nous n'avons pas cherché à ce que cette interprétation soit validée par des tiers. Nous avons voulu par contre que la rigueur scientifique de notre démarche soit appuyée par les mémos qui explicitent l'analyse et l'interprétation qui ont été faites.

Nous renvoyons le lecteur à l'Annexe E pour comprendre davantage les regroupements effectués. Dans les sections suivantes, ce travail d'interprétation deviendra plus clair et sera illustré avec des exemples concrets tirés des textes et des concordances du logiciel *TXM*[31].

[31] Les extraits tirés du concordancier de *TXM* sont identifiés avec un code qui renvoie à l'article de notre corpus. Dans l'Annexe F figurent les références bibliographiques de ces articles avec leur titre, date de publication et source.

Dans les sections suivantes, nous décrivons de manière comparative les trois composantes sémantiques en fonction des structures sémantiques repérées dans chaque sous-corpus.

2.4.1 Thématique

Nous avons divisé la description de critères thématiques en deux parties, celle des critères thématiques globaux et celle des critères thématiques locaux. Dans la première partie, l'analyse sémique a identifié les isotopies principales de chaque sous-corpus à partir des critères obtenus par le calcul des spécificités. Dans ce sens, nous identifions les classes sémantiques (domaines et taxèmes) et nous décrivons les thèmes génériques d'ETUD et GOUV. La seconde partie présente les thèmes spécifiques des sous-corpus, qui correspondent aux molécules sémiques identifiées à partir du résultat du calcul de cooccurrence.

a) Critères thématiques globaux : isotopies et thèmes génériques

Les tableaux 29 et 30 ci-dessous présentent pour les sous-corpus ETUD et GOUV respectivement, l'ensemble de structures sémantiques identifiées à la suite de l'analyse des critères textuels globaux qui ont été retenus. Nous avons regroupé les isotopies identifiées par domaines, de manière à déceler les thèmes les plus génériques traités dans chaque sous-corpus. Nous avons cherché à constituer les isotopies en fonction de la récurrence d'un sème générique partagé par les critères textuels qui les composent.

Dans ETUD et dans GOUV (tableau 29 et 30), nous avons identifié trois domaines communs: //éducation//, //politique// et //économie//. Dans ETUD (tableau 29) nous avons identifié deux taxèmes, //valeurs// et //mobilisation//. Dans GOUV (tableau 30), nous avons identifié 4 autres taxèmes: //négociation//, //communication//, //intimidation// et //diffamation//. Dans les tableaux, la colonne « Classes sémantiques » affiche les domaines et taxèmes identifiés. La colonne « Isotopies » présente un ensemble d'isotopies associées aux domaines et aux taxèmes. La colonne « Critères » présente les critères textuels de chaque isotopie et par là, aux domaines et taxèmes concernés. Elle inclut tous les critères textuels de type mots simples, lemmes et segments répétés repérés par le calcul des spécificités, avec leur score de spécificité. Nous avons cherché à ce que les classes sémantiques et les isotopies

identifiées puissent aider à mettre en évidence les différences thématiques observées entre les deux sous-corpus.

Tableau 29. Thématique d'ETUD

Classes sémantiques	**Isotopies**	**Spéc.**	**Critères**
//éducation//	/éducation supérieure/	+16	'éducation'
		+3	'communauté universitaire'
		+2	'étudiants universitaires'
		+2	'enseignement supérieur'
		+2	'supérieur'
		+2	'étudiant'
	/connaissance/	+5	'savoir'
		+3	'enseignement'
		+3	'instruire'
		+2	'instruction'
	/institutions d'éducation/	+4	'éducation'
		+4	'université'
		+3	'administrations'
		+3	'droits de scolarité'
//politique//	/médiateurs politiques/	+6	'Courchesne'
		+6	'ministre'
		+4	'Madame Beauchamp'
		+4	'Beauchamp'
		+4	'Mme Beauchamp'
		+4	'Mme Courchesne'
		+4	' Michelle Courchesne'
		+3	'instance(s)'
		+3	'gouvernement du Québec'
		+3	'gouvernement québécois'
		+3	'premier ministre'
		+2	'délégués'
		+2	'ministère'
		+2	'Bourassa'

	/autoritarisme/	+3	violence’
		+3	‘matraque’
		+3	‘guerre’
		+3	‘ligne dure’
		+2	‘matraques’
		+2	‘policier’
		+2	‘répressif’
		+2	‘imposé’
		+2	‘images’
		+2	‘procéder’
	/mépris/	+8	‘mépris’
		+3	‘refuser’
		+2	‘arrogance’
		+2	‘iPhone’
	/désintégration/	+3	‘diviser’
		+3	‘dérive’
		+2	‘pourrir’
		+2	‘corruption’
		+2	‘indigne’
//économie//	/économie de marché/	+4	‘économie’
		+4	‘développement’
		+4	‘entreprise(s)’
		+4	‘système économique’
		+4	‘inégalité’
		+3	‘niveau de vie’
	/gaspillage/	+5	‘milliard de dollars’
		+2	‘salarial’
	/construction/	+4	‘construction(s)’
		+3	‘immobilier’
		+2	‘campus’
		+2	‘UdeM’

	/charge/	+5	‘@card@’[32]
		+3	‘en cinq ans’
		+3	‘75’,
		+3	‘%en’
		+3	‘sur le dos’
		+3	‘quart’
		+2	‘temps plein’
		+2	‘travailler’
		+2	‘contribuables’,
		+2	‘endettés’
//valeurs//	/expression/	+3	‘point de presse’
		+2	‘critique’
	/avenir/	+6	‘jeunesse’
		+5	‘jeune(s)’
		+5	‘génération(s)’
		+3	‘avenir’
		+2	‘réussite’
	/égalité/	+5	‘communauté’
		+4	‘accessible’
		+4	‘accès’
		+4	‘humain’
		+3	‘commun’
		+3	‘pour tous’
		+3	‘juste’
		+3	‘valeurs’
		+2	‘partie de la’
		+2	‘à tous’
	/droit/	5	‘juste part’
		+4	‘fondamental’

[32] Le symbole @card@ est généré par le lemmatisateur *TreeTagger* et regroupe les chiffres repérés dans les textes.

		+3	‘droit de grève’
		+3	‘adoption’
		+3	‘commission’
		+2	‘loi 78’
		+2	‘ONU’
	/moment historique/	+3	‘souvenir’
		+3	‘souviens’
		+2	‘réveil’
	/conciliation/	+5	‘pacifique(s)’
		+3	‘asseoir’
		+3	‘reconnaître’
	/société/	+11	‘société’
		+8	‘autochtone(s)’
		+3	‘classe moyenne’
		+3	‘citoyenne’
		+3	‘homme(s)’
		+3	‘famille’
		+3	‘population’
		+2	‘citoyen’
//mobilisation//	/mobilisation/	+5	‘grève’
		+3	‘solidarité’
		+3	‘le 22’
		+2	‘mobilisation’
		+2	‘mobiliser’
		+2	‘revendication’
		+2	‘unir’
	/leaders/	+5	‘Martine Desjardins’
		+4	‘Desjardins’
		+4	‘présidente de la FEUQ’
		+3	‘présidente’
		+2	‘président’
	/partisans/	+5	‘Réjean Parent’
		+3	‘chefs syndicaux’,

		+2	‘Rocher’
	/lutte/	+3	‘art’
		+3	‘stratégie(s)’
		+3	‘résistance’

Tableau 30. Thématique de GOUV

Classes sémantiques	**Isotopies**	**Spéc.**	**Critères**
//éducation//	/activité scolaire/	+4	‘cours’
		+4	‘session’
	/éducation supérieure/	+3	‘des études supérieures’
		+3	‘études universitaires’
	/accessibilité/	+5	‘bas’
		+3	‘accessibilité aux études’
		+3	‘plus bas’
		+2	‘accessibilité’
	/frais/	+10	‘frais de scolarité’
		+7	‘hausses’
		+7	‘frais’
		+4	‘hausse des frais’,
//politique//	/support/	+3	‘appui’
		+3	‘Pellerin’
		+2	‘appuyer’
		+2	‘sympathie’
		+2	‘camarades’
		+2	‘partisan’
	/organisations/	+5	‘centrales syndicales’
		+4	‘syndicales’
		+4	‘centrales’
		+3	‘groupe(s)’
	/gauche/ /péjoratif/	+15	‘artiste(s)’
		+8	‘gauche’

	+7	‘carré’
	+5	‘carré(s) rouge(s)’
	+5	‘humoriste(s)’
	+5	‘idéologique’
	+4	‘rouge(s)’
	+2	‘comédien’
	+2	‘boutonnière’
/militantisme/	+5	‘militants’
	+3	‘lutte contre’
	+3	‘manifestant(s)’
	+3	‘leaders’
/polarisation/	+4	‘côté’
	+4	‘du côté’
	+3	‘position’
	+2	‘bord’
	+2	‘deux parties’
	+2	‘défendu’
	+2	‘autre côté’
/médiateurs politiques/	+3	‘gouvernement Charest’
	+3	‘ministre Line Beauchamp’
	+3	‘Assemblée nationale’
	+3	‘libéraux’
/ordre/	+3	‘responsabilité(s)’
	+3	‘compromis’
	+3	‘modèle québécois’
	+3	‘social-démocratie’
	+2	‘démocratie’
	+2	‘droit de manifester’
/désordre/	+3	‘perturbations’
	+3	‘crise étudiante’
	+2	‘anarchie’
	+2	‘descendre’

/radicalisme/	+6	‘boycott’
	+5	‘gel’
	+5	‘boycotter’
	+4	‘gel des frais’
	+4	‘boycottage’
	+3	‘geler’
	+3	‘gel des droits’
	+3	‘annuler’
	+3	‘annulation’
	+3	‘Dubois’
	+2	‘faction’
	+2	‘boycottent’
/représentativité/	+5	‘associations étudiantes’
	+4	‘majorité’
	+4	‘Québécois’
	+4	‘associations’
	+4	‘votent’
	+3	‘petit’
	+3	‘majorité des Québécois’
	+2	‘majorité des étudiants’
/élection/	+4	‘péquiste’
	+4	‘promettre’
	+3	‘PQ’
	+3	‘campagne électorale’
	+3	‘première ministre’
	+3	‘candidats’
	+2	‘campagne’
	+2	‘promis’
/candidats/	+8	‘Marois’
	+5	‘Pauline Marois’
	+5	‘Mme Marois’
	+3	‘Parti québécois’
	+2	‘CAQ’

//économie//	/finances publiques/	+4	‘salaire minimum’
		+4	‘inflation’
		+3	‘minimum’
		+3	‘revenu’
		+3	‘proportionnel au revenu’
		+3	‘budget’
		+3	‘chiffre’
		+2	‘indexation’
		+2	‘$’
		+2	‘contribuables’
		+2	‘chiffres’
		+2	‘paient’
		+2	‘dividendes’
		+2	‘fardeau’
		+2	‘salaire’
		+2	‘rattrapage’
	/fédéralisme/	+5	‘province’
		+3	‘autres provinces’
		+3	‘péréquation’
		+2	‘Ouest’
//négociation//	/proposition/	+3	‘annoncées’
		+3	‘propositions’
		+2	‘proposées’
	/obstacle/	+3	‘fermeté’
		+3	‘camper’
		+3	‘impasse’
		+2	‘camp’
	/résolution/	+3	‘dénouer’
		+3	‘mécanisme’
		+3	‘piste’
		+2	‘concrets’
	/compensation/	+7	‘remboursement’

		+4	‘prêts et bourses’,
		+4	‘bonification’
		+3	‘atténuer’
		+3	‘ajuster’
		+3	‘bourse(s)’
		+3	‘parental(e)’
		+3	‘régime des prêts’
		+2	‘familial’
//communication//	/monologue/ /négatif/	+5	‘Twitter’
		+3	‘médias sociaux’
		+3	‘micro’
		+3	‘affiche’
	/presse écrite/ /positif/	+3	‘écrire’
		+3	‘chronique’
//intimidation//	/menace/	+4	‘intimider’
		+3	‘méfaits’
		+3	‘menacer’
		+2	‘pression’
	/accès public/	+4	‘de la rue’
		+3	‘pont’
		+3	‘journaliste(s)’
		+2	‘photographe’
//diffamation//	/immaturité/	+5	‘beurre’
		+2	‘maman’
		+2	‘disproportionner’
		+2	‘test’
		+2	‘bacon’
		+2	‘empresser’
		+2	‘répéter’
	/insolence/	+4	‘fasciste’
		+3	‘salope’
		+3	‘fascisme’

	+2	‘bêtise’
	+2	‘Hitler’
	+2	‘nazi’
	+2	‘sexiste’
/diffamateur/	+3	‘Barbe’
	+2	‘Brin’

Domaine //éducation// dans ETUD et GOUV

Le domaine //éducation// se trouve présent dans les deux sous-corpus. Dans ETUD nous retrouvons l’isotopie /éducation supérieure/ **[/éducation supérieure/ : ‘étudiants universitaires’, ‘enseignement supérieur’, ‘supérieur’, ‘étudiant’, ‘communauté universitaire’]** et l’isotopie /connaissance/, qui met en évidence des notions valorisantes de l’éducation, en actualisant des sèmes en rapport avec l’accès à la connaissance et au savoir **[/connaissance/ : ‘éducation’, ‘savoir’, ‘enseignement’, ‘instruire’, ‘instruction’**]. L’isotopie /institutions d’éducation/ dans ETUD est attestée par la récurrence d’un sème générique regroupant les administrations éducatives officielles **[/institutions d’éducation/ : ‘Éducation’, ‘université’, ‘administrations’, ‘droits de scolarité’]** : ‘Éducation’ en majuscule apparaît fréquemment associée à des noms d’institutions (par exemple, ‘ministère de l’Éducation’) aussi bien que ‘université’, ‘administrations’ et ‘droits de scolarité’. Différemment du sous-corpus ETUD, où l’isotopie /éducation supérieure/ est attestée par ‘enseignement supérieur’, dans GOUV c’est le critère ‘études’ qui apparaît avec une spécificité plus grande **[/éducation supérieure/ : ‘des études supérieures’, ‘études universitaires’]**. Cela souligne une différence entre la notion de l’éducation préconisée par chaque groupe. En effet, le critère ‘enseignement’ actualise le sème /recevoir/ tandis que ‘études’ actualise /faire/ (une personne reçoit de l’enseignement, mais elle fait des études). Du côté d’ETUD, l’éducation est perçue comme un droit ou comme une valeur permettant l’épanouissement de l’individu. Dans GOUV, la notion de ‘études universitaires’ est plutôt associée à une étape permettant l’accès à la vie professionnelle dont les frais doivent être assumés par les bénéficiaires, comme montre les extraits ci-dessous :

[news20120605OP12060524524162](…) n'y a aucun problème d'accessibilité aux *études universitaires* au Québec, mais nous avons un grave problème d'orientation professionnelle (…)

[news20120302OR120302228889283] (…) contribution d'un étudiant au coût de ses *études universitaires*. Les frais de scolarité au Québec sont les plus bas en

[news20120323OP120323232002931](…), pour refléter davantage la rentabilité économique des *études universitaires*. 2— Un ajustement correspondant des prêts et bourses. 3— Une différenciation des (…)

[news20120407TB0027](…) à des liquidités additionnelles pendant leurs années d'*études universitaires.* (…)

Le sous-corpus GOUV aborde la question d'accessibilité aux études **[/accessibilité/ : 'accessibilité aux études', 'accessibilité', 'bas', 'plus bas']** thème qui constitue un des arguments les plus importants des étudiants contre la hausse. Cependant, le traitement donné à ce thème dans GOUV se fait dans la perspective du coût de l'éducation. Les auteurs attaquent l'argument des étudiants et défendent que le prix des frais universitaires (ou frais de scolarité) au Québec se trouve parmi les plus bas comparativement aux autres provinces canadiennes et défendent la hausse des frais **[/frais/ : 'hausse(s)', 'hausse des frais', 'frais', 'frais de scolarité']**. Encore par rapport au domaine //éducation//, la thématique de GOUV aborde les effets du mouvement sur la session en cours. L'isotopie /activité scolaire/ **[/activité scolaire/ : 'cours', 'session']** se rapporte au calendrier universitaire, thématique qui est absente du sous-corpus ETUD.

Domaine //politique// dans ETUD et GOUV

Dans le sous-corpus ETUD, le domaine //politique// se trouve associé à une thématique sur l'autoritarisme **[/autoritarisme/ : 'violence', 'matraque', 'matraques', 'guerre', 'policier', 'répressif', 'imposé', 'ligne dure', 'images', 'procéder']**, sur le mépris du gouvernement dans la conduction de la crise **[/mépris/ : 'mépris', 'arrogance', 'refuser', 'iPhone']** ainsi que l'isotopie /désintégration/ regroupant des critères qui qualifient négativement le gouvernement et ses actions **[/désintégration/ : 'pourrir, 'indigne',**

‘diviser’, ‘corruption, ‘dérive’]. L’isotopie /médiateurs politiques/ regroupe le nom des personnes responsables de négocier avec les étudiants, notamment les ministres de l’Éducation et le premier ministre **[/médiateurs politiques/ : ‘Courchesne’, ‘ministre’, ‘instance(s)’, ‘Bourassa’, ‘Madame’, ‘Beauchamp’, ‘Madame Beauchamp’, ‘Mme Beauchamp’, ‘Mme Courchesne’, ‘Michelle Courchesne’, ‘premier ministre’, ‘délégués, ministère’, ‘gouvernement du Québec’, ‘gouvernement québécois’]**.

Le domaine //politique// dans GOUV est articulé autour de différents thèmes génériques. D’une part, nous retrouvons une critique sur le support politique que les étudiants ont reçu de la part de certains groupes politiques et organisations. Ce thème regroupe les isotopies /support/ **[/support/ : ‘appui’, ‘appuyer’, ‘sympathie’, ‘camarades’, ‘partisan’, ‘Pellerin’]** /organisations/ **[/organisations/ : ‘syndicales’, ‘centrales syndicales’, ‘centrales’, ‘groupe(s)’]**. L’alliance entre les centrales syndicales et d’autres personnes nominalement mentionnées dans les textes GOUV, est perçue de façon péjorative par les auteurs de ce sous-corpus et est souvent critiquée. D’autre part, une critique envers les symboles et les idées de la lutte étudiante est signalée par les isotopies /militantisme/ **[/militantisme/ : ‘militant(s)’, ‘lutte contre’, ‘manifestant(s)’, ‘leader(s)’)** et /gauche/ **[/gauche/ : ‘artiste(s)’, ‘boutonnière’, ‘carré (s)’, ‘carré(s) rouge(s)’, ‘comédiens’, ‘gauche’, ‘humoriste(s)’, ‘idéologique’, ‘rouge(s)’]**, qui visent à caractériser le mouvement et ses tenants d’un point de vue idéologique et péjoratif. Un thème visant à qualifier le mouvement étudiant comme radical et chahuteur est attesté par les isotopies /radicalisme/ **[/radicalisme/ : ‘Dubois’, ‘faction’, ‘gel’, ‘gel des frais’, ‘geler’, ‘gel des droits’, ‘annuler’, ‘annulation’, ‘boycott’, ‘boycotter’, ‘boycottage’, ‘boycottent’]** et /désordre/ **[/désordre/ : ‘anarchie’, ‘perturbation’, ‘crise étudiante’, ‘descendre’**]. Dans celles-ci, nous retrouvons les critiques sur les déclarations du leader de la CLASSE (‘Dubois’) dont l’organisation est qualifiée de ‘faction’ ; sur le caractère déraisonnable et extrême des demandes des étudiants (‘gel’) ; et finalement sur les troubles causés par les manifestations. Un thème centré sur le rétablissement de l’ordre est manifesté par l’isotopie /ordre/ **[/ordre/ : ‘responsabilité(s)’, ‘compromis’, ‘démocratie’, ‘modèle québécois’, ‘droit de manifester’, ‘social-démocratie’]**, dans laquelle un vocabulaire lié à la responsabilité et au respect à la démocratie est signalé.

D'autres thèmes du domaine //politique// du corpus GOUV sont attestés par l'isotopie /polarisation/ **[/polarisation/ : 'côté', 'du côté', 'bord', 'position', 'deux parties', 'défendu', 'autre côté']**, où l'opposition entre les groupes impliqués dans le conflit est thématisée. L'isotopie /représentativité/ révèle la critique dirigée au manque de représentativité des étudiants en grève par rapport à l'univers d'étudiants existants : les auteurs GOUV se plaignent souvent que le mouvement des étudiants n'a pas l'appui de la majorité des Québécois et qu'ils sont sous-représentés par un petit nombre d'associations qui votent dans les assemblées **[/représentativité/ : 'majorité', 'Québécois', 'petit', 'majorité des Québécois', 'majorité des étudiants', 'associations étudiantes', 'associations', 'votent']**. Dans l'isotopie /médiateurs politiques/, un thème sur les actions et le positionnement du gouvernement face au conflit est signalé par la présence de certaines entités nommées, ainsi que le nom d'institutions et de partis politiques directement impliqués dans les négociations avec les étudiants **[/médiateurs politiques/ : 'gouvernement Charest', 'ministre Line Beauchamp', 'Assemblée nationale', 'libéraux']**. Le thème sur les élections provinciales était également présent dans le sous-corpus GOUV **[/élection/ : 'péquiste', 'PQ', 'campagne électorale', 'candidats', 'promettre', 'promis', 'première ministre')**, et les attentes par rapport à un éventuel changement de pouvoir qui pourrait remplacer le premier ministre Jean Charest par la leader du Parti québécois, Pauline Marois, dont la principale promesse électorale était de geler les droits de scolarité **[/candidats/ : 'Mme Marois', 'Pauline Marois', 'Marois', 'CAQ', 'Parti québécois']**.

Domaine //économie// dans ETUD et GOUV

Nous avons aussi repéré le domaine //économie//, dans lequel nous pouvons remarquer la différence entre deux thèmes génériques récurrents dans les sous-corpus ETUD et GOUV. Tandis que dans ETUD le domaine //économie// regroupe des isotopies qui présentent des thèmes sur le modèle économique et sur la critique de la gestion maladroite de l'argent public **[/économie de marché/, /gaspillage/, /construction/, /charge/]**, dans GOUV il se structure autour de propositions pour mieux gérer le budget public afin de diminuer la charge d'impôt sur les contribuables **[/finances publiques/, /fédéralisme/]**.

Dans ETUD, les isotopies du domaine //économie// structurent une thématique critique au modèle économique basé sur l'économie du marché et du capitalisme, dont les conséquences sont l'aggravation des inégalités et du sous-financement, ce dernier pouvant mener à l'extinction des services publics de base, comme l'éducation **[/économie du marché/ : 'économie', 'développement', 'entreprises', 'système économique', 'niveau de vie', 'inégalité']**. D'autres critiques sont émises sur la gestion du budget publique, dénonçant le gaspillage d'argent, avec des critiques dirigées contre l'augmentation des salaires des cadres et des recteurs **[/gaspillage/ : milliard de dollars', 'salarial']**, ainsi que la charge liée aux coûts de l'éducation pour les étudiants **[/charge/ : 'en cinq ans', 'quatre ans', '75', '%en', 'temps plein', 'travailler', 'sur le dos', 'contribuables', 'endettés', 'quart', '@card@']**. L'isotopie /construction/ évoque la critique sur le gaspillage d'argent, particulièrement dans le secteur de la construction, lequel a investi beaucoup d'argent dans les travaux de rénovation sur le campus et l'achat immobilier, décisions soupçonnées d'être les vraies motivations politiques derrière l'annonce du gouvernement de hausser les droits de scolarité **[/construction/ : 'UdeM', 'campus', 'immobilier', 'constructions', 'construction']**. L'isotopie /charge/, pour sa part, fait état du thème sur l'étalement de la hausse dans le temps, de la lourdeur financière que la hausse représente pour les étudiants plutôt que pour les contribuables d'aujourd'hui et le risque que la hausse condamne la génération des jeunes à l'endettement. Dans GOUV, la récurrence du sème /finances publiques/ **[/finances publiques/ : 'salaire minimum', 'inflation', 'minimum', 'revenu', 'indexation', 'proportionnel au revenu', 'budget', 'chiffres', '$', 'contribuables', 'chiffre', 'paient', 'dividendes', 'fardeau', 'salaire', 'rattrapage']** atteste des préoccupations reliées à la gestion de l'argent public, tandis que /fédéralisme/ **[/fédéralisme/ : 'province(s)', 'autres provinces', 'péréquation', 'Ouest']** aborde le système de péréquation, qui avantage le Québec dans le transfert d'impôts comparativement aux autres provinces et sans lequel le système d'éducation devient insoutenable d'un point de vue financier, selon l'argument des défenseurs de la hausse.

Taxème //valeurs// dans ETUD

L'allusion à des valeurs à défendre est très caractéristique de la thématique du sous-corpus ETUD. Le taxème //valeurs// regroupe les isotopies qui renvoient à un ensemble

d'arguments en faveur du gel des droits de scolarité et pour la continuation du mouvement. La récurrence du sème générique /avenir/ témoigne de l'inquiétude des étudiants face aux conséquences néfastes que la hausse pourrait entraîner chez les générations futures **[/avenir/ : 'avenir', 'génération', 'jeunesse', 'réussite', 'jeune(s)'**]. Ce thème gagne de l'ampleur dans l'argumentaire pour la défense des valeurs démocratiques, d'égalité d'opportunités et d'accès aux services publics. L'isotopie /égalité/, par exemple, est actualisée par des critères textuels relevant de valeurs humanistes de citoyenneté **[/égalité/ : 'commun', 'pour tous', 'accessibilité, 'accessible', 'accès', ' partie de la', 'communauté', 'humain', 'juste', 'à tous', 'valeurs']**. La récurrence du sème générique /société/ renforce la vision du collectif et l'appartenance à une société **[/société/ : 'classe moyenne', 'société', 'citoyen', 'citoyenne', 'homme(s)', 'autochtone(s)', 'famille', 'population']**. Pareillement, la présence de flexions au masculin et au féminin pour certains mots **['citoyen', 'citoyenne']** renforce la notion d'appartenance et d'inclusion. Les critères textuels inclus dans cette isotopie se trouvent employés à l'intérieur d'un récit sur la nécessité de l'inclusion, comme l'attestent les exemples d'emploi du critère 'autochtone' :

> [news20120416LE20120416_a7_peuples](…)combler l'écart entre les niveaux de s colarité des *autochtones* et des non-
> autochtones permettrait d'injecter 179 milliards de dollars (…)

> [news20120416LE20120416_a6_vlan] (…)l'impact néfaste d'une hausse des droits de scolarité au Québec sur de potentiels étudiants *autochtones* déjà découragés (…)

Encore dans le taxème //valeurs// d'ETUD, l'isotopie /moment historique/ atteste un thème d'exaltation du mouvement des étudiants, soulignant son caractère historique et son potentiel transformateur **[/moment historique/ : 'souvenir', 'réveil', 'souviens']**. L'isotopie /conciliation/ exprime le caractère pacifique des manifestations et la disposition des étudiants à négocier avec le gouvernement **[/conciliation/ : 'pacifique', 'pacifiques', 'asseoir', 'reconnaître']**. À son tour, l'isotopie /droits/ met en évidence des thèmes reliés aux droits fondamentaux qui sont menacés par les actions du gouvernement. Ici nous retrouvons l'isotopie /expression/, avec un thème sur le droit de la parole exprimant l'encouragement à l'esprit critique et à la liberté d'expression **[/expression/ : 'critique', 'point de presse']**, ainsi

qu'à la dénonciation de l'adoption de la loi 78 par le gouvernement Charest, qui empêche la tenue de manifestations de plus de 40 personnes **[/droits/ : 'fondamental', 'juste part', 'droit de grève', 'loi 78', 'ONU', 'adoption', 'commission']**. Nous retrouvons également une discussion sur l'injustice associée à la hausse, avec la récurrence de 'juste part' dans ETUD. Cette expression a été d'abord employée par la ministre Line Beauchamp pour demander aux étudiants de « faire leur juste part », c'est-à-dire d'assumer une partie des coûts reliés à leur formation universitaire. Elle est remise en question par plusieurs auteurs ETUD, qui se demandent si le Québec a aussi fait sa « juste part » pour garantir l'accès à l'éducation supérieure.

> [news20120404LA0007] On traite les étudiants d'enfants gâtés qui refusent de payer leur *juste part*, de futurs médecins qui vont rouler sur l'or et qui osent demander aux travailleurs de payer pour leur éducation. Honte à eux.

> [news20120407NV0026] Je ne suis pas contre le fait de demander à chacun de faire sa *juste part*, mais je me demande sérieusement si le Québec fait sa juste part pour les étudiants.

Taxème // mobilisation// dans ETUD

Du côté ETUD, un thème valorisant la lutte étudiante est révélé par le taxème **//mobilisation//**. Ce thème se présente comme un contre-argument au cadrage politique et idéologique diffusé dans le groupe adversaire et il caractérise le mouvement étudiant comme une union d'individus qui luttent en faveur d'un intérêt commun. Les étudiants insistent pour parler de leur mobilisation comme d'une grève, dans le but d'y conférer un caractère organisé et institutionnel dont l'objectif est de légitimer le droit de manifester et de revendiquer sur la scène publique **[/mobilisation/ : 'grève', 'mobilisation', 'mobiliser', 'revendication', 'unir', 'solidarité', 'le 22'**]. La mobilisation des étudiants est également caractérisée dans son aspect stratégique, comme le montre l'isotopie /lutte/. Dans ce dernier exemple, les critères textuels 'art', 'stratégie(s)' et 'résistance' signalent un thème sur la nécessité d'une organisation tactique contre le gouvernement **[/lutte/ : 'art', 'stratégie(s)', 'résistance']**. Nous remarquons également une isotopie regroupant les partisans du mouvement **[/partisans/ : 'Réjean Parent', 'chefs syndicaux', 'Rocher']** ainsi que les personnes

engagées dans l'organisation et la direction de cette mobilisation, dont les mentions sont faites d'une façon plutôt formelle et respectueuse **[/leaders/ : 'Martine Desjardins', 'présidente de la FEUQ', 'président']**. Cela s'oppose au traitement accordé au leader étudiant Gabriel Nadeau-Dubois dans GOUV, qui est souvent associé au radicalisme.

Taxème // négociation// dans GOUV

Dans le sous-corpus GOUV, nous observons la tentative des auteurs d'imputer aux étudiants la raison de l'impasse qui mène à l'aggravation de la crise, en les accusant de faire peu d'efforts dans les négociations proposées par le gouvernement, spécialement à l'égard des compensations offertes au programme de prêts et bourses.

Le taxème //négociation// dans GOUV renvoie à une thématique sur les propositions offertes par le gouvernement québécois et qui dénonce le manque d'ouverture des étudiants à faire des concessions ou à négocier avec les représentants. L'isotopie /proposition/ regroupe le vocabulaire caractéristique à la communication officielle **[/proposition/ : 'annoncés', 'propositions', 'proposées'**]. Le souhait de trouver une solution pour la fin du conflit est exprimé par la récurrence du sème /résolution/ **[/résolution/ : 'dénouer', 'mécanisme', 'piste', 'concrets']**. La récurrence du sème générique /compensation/ est remarquable et elle est attestée par la présence de divers critères textuels relatifs aux propositions du gouvernement, comme le programme de prêt et bourses, la bonification parentale, ainsi que la proposition d'indexation de la bourse en fonction du revenu familial **[/compensation/ : 'remboursement', 'prêt et bourse(s)', 'régime de prêts', 'bonification', 'ajuster', 'parental(e)', 'familial', 'atténuer']**. Ces propositions constituent, dans l'argument élaboré par les auteurs, des alternatives raisonnables servant à atténuer les effets économiques de la hausse des droits et à mettre fin au conflit. Une critique du positionnement des étudiants face à ces propositions se retrouve dans l'isotopie /obstacle/, laquelle exprime la fermeté des étudiants et leur refus de s'ouvrir à la négociation [**/obstacle/ : 'fermeté', 'camp', 'camper', 'impasse'**].

Taxème //communication// dans GOUV

Nous avons relevé dans le taxème //communication// dans GOUV, un thème sur les moyens de communication employés par les étudiants qui a la particularité de souligner

l'aspect monologal du type de communication qu'ils ont privilégié. Les billets publiés par les manifestants dans Twitter et à la prise de parole par les représentants des étudiants dans les assemblées **[/monologue/ /négatif/ : 'Twitter', 'médias sociaux', 'micro', 'affiche'] sont souvent mentionnés et de façon à les rendre risibles**. Contrastant avec la prise de parole étudiante, on remarque les moyens de communication privilégiés par les auteurs de GOUV qui sont, selon les auteurs de ce sous-corpus, davantage « cautionnés » ou « reconnus » **[/presse écrite/ : écrire', 'chronique']**.

Taxème //diffamation// dans GOUV

Dans GOUV, nous observons la présence plus prononcée de thèmes visant à disqualifier l'argumentaire adversaire. D'une part, les auteurs GOUV dénoncent le manque de maîtrise argumentative des étudiants et leur incapacité à mener les négociations avec maturité. Le taxème //diffamation// rassemble certaines isotopies que nous avons identifiées pour ce thème. Il se constitue par la récurrence du sème générique /immaturité/ **[/immaturité/ : 'maman', 'beurre', 'disproportionner', 'test', 'bacon', 'empresser', 'répéter']** et regroupe aussi certaines injures et accusations prononcées par les étudiants et contre lesquels les auteurs GOUV se défendent **[/insolence/ : 'fasciste', 'salope', 'fascisme', 'bêtise', 'Hitler', 'nazi', 'sexiste']**. D'autre part, nous notons la présence d'un thème sur l'intimidation (taxème //intimidation//) où se retrouvent les isotopies /menace/ **[/menace/ : pression', 'intimider', 'méfaits', 'menacer']** et /accès public/, cette dernière actualisant un thème au sujet du blocage de lieux publics pendant les manifestations **[/accès public/ : 'de la rue', 'pont(s)', 'entrée', 'journaliste(s)', 'photographe']**.

Dans la prochaine section, nous allons présenter l'analyse des thématiques spécifiques en fonction du repérage de cooccurrents, lesquels sont homologués dans le cadre de notre recherche comme des molécules sémiques.

b) Critères thématiques locaux: molécules et thèmes spécifiques

Dans cette section, nous soumettons sous la forme de tableaux les critères locaux sélectionnés pour chaque mot pôle. Afin de faciliter la comparaison des thèmes spécifiques des sous-corpus ETUD et GOUV, les tableaux affichent les cooccurrences de chacun côte à

côte, ordonnés en fonction du score de spécificité. À la suite des tableaux se trouve la description des thèmes spécifiques en rapport avec les critères locaux sélectionnés.

Mot pôle : étudiant

Tableau 31. Cooccurrents du mot pôle ‘étudiant’

ETUD		**GOUV**	
Cooccurrence	**Indice**	**Cooccurrence**	**Indice**
étudiant-détermination	5,08	étudiant-collégial	6,61
étudiant-opposer	4,40	étudiant-porte-parole	4,55
étudiant-entente	3,99	étudiant-cause	3,41
étudiant-échouer	2,92	étudiant-déclencher	3,27
étudiant-recruter	2,92	étudiant-injonction	3,04
étudiant-début	2,71	étudiant-empêcher	2,89
étudiant-doctorat	2,67	étudiant-millier	2,88
étudiant-postsecondaire	2,67	étudiant-cégep	2,81
étudiant-blâmer	2,66	étudiant-exiger	2,78
étudiant-minimiser	2,66	étudiant-contribution	2,75
étudiant-rejeter	2,58	étudiant-profiter	2,73
étudiant-bonifier	2,55	étudiant-établissement	2,62
étudiant-réveiller	2,55	étudiant-manif	2,58
étudiant-dialogue	2,54	étudiant-boursier	2,50
étudiant-assujettir	2,48	étudiant-éterniser	2,50
étudiant-échanger	2,48	étudiant-retourner	2,44
étudiant-allié	2,24	étudiant-gréviste	2,38
étudiant-dénigrer	2,24	étudiant-saisir	2,29
étudiant-négocier	2,19	étudiant-protester	2,26
étudiant-terrain	2,13	étudiant-réduction	2,04
étudiant-collège	2,11		
étudiant-affirmer	2,03		

Les cooccurrents autour du mot pôle ‘étudiant’ nous permettent de déceler différents thèmes spécifiques et de percevoir les contrastes entre ETUD et GOUV. Nous remarquons par

exemple que dans ETUD, ce mot pôle se retrouve relié à l'éducation supérieure **['étudiant-doctorat', 'étudiant-post-secondaire']** tandis que dans GOUV, il est lié à l'éducation postsecondaire **['étudiant-collégial', 'étudiant-cégep', 'étudiant-établissement']**. Dans ETUD, ce rapport révèle l'intention d'associer aux étudiants en grève des notions valorisantes comme l'expérience et la maturité, tandis que dans GOUV l'argumentaire consiste au contraire à démontrer l'immaturité des étudiants en grève, évoquant plus fréquemment les étudiants collégiaux ainsi que leur appartenance à la CLASSE, association considérée la plus radicale du mouvement. Dans GOUV, cette tentative de dévaloriser l'image des étudiants en grève ou de l'associer à des intérêts purement politiques est aussi présente dans **'étudiant-gréviste'** et dans ses porte-paroles **['étudiant-porte-parole]**.

Le discours de valorisation des étudiants dans ETUD est également attesté par les cooccurrents qui évoquent la détermination et la résistance des étudiants en grève [**'étudiant-détermination', 'étudiant-opposer', 'étudiant-rejeter'**], ainsi que leur vision et leur capacité de mobilisation **['étudiant-réveiller', 'étudiant-mobilisation', 'étudiant-recruter', 'étudiant-allié]**. Dans GOUV, un thème spécifique centré sur la protestation, le manque de flexibilité des étudiants et la durée du conflit contraste avec les positions défendues par les auteurs d'ETUD **['étudiant-protester',' étudiant-manif', 'étudiant-empêcher',' étudiant-millier', 'étudiant-éterniser']**.

Une autre différence entre les thèmes spécifiques des sous-corpus se trouve dans la tentative des auteurs des deux camps d'imputer à l'un ou à l'autre la responsabilité par le déclenchement et l'intensification de la crise. Dans GOUV, cette tentative est attestée par les cooccurrents **'étudiant-déclencher'** et **'étudiant-cause'**. Les auteurs GOUV insistent également sur l'argument selon lequel les étudiants ne seraient pas ouverts aux propositions de bonification offertes par le gouvernement pour mettre fin à la crise **['étudiant-exiger']**, alors qu'ils devraient saisir l'opportunité de négociation **['étudiant-profiter', 'étudiant-saisir', 'étudiant-réduction', 'étudiant-contribution', 'étudiant-boursier']**. Dans ETUD, nous remarquons une stratégie de défense à cette accusation, attestée par les cooccurrents **['étudiant-affirmer' 'étudiant-dénigre' 'étudiant-blâmer', 'étudiant-minimiser', 'étudiant-assujettir']**. Les auteurs ETUD veulent également déconstruire le point de vue qui fait croire qu'ils sont inflexibles dans les négociations **['étudiant-entente', 'étudiant-**

dialogue', 'étudiant-négocier', étudiant-terrain, 'étudiant-échanger', 'étudiants-bonifier', 'étudiant-échouer'], puisqu'ils font valoir leur position depuis le déclenchement du conflit **['étudiant-début']**.

Nous retrouvons également dans GOUV un thème sur la question des injonctions demandées par des étudiants opposés au mouvement et qui veulent retourner en classe **['étudiant-injonction', 'étudiant-retourner]**.

Mot pôle : Québec

Tableau 32. Cooccurrents du mot pôle 'Québec'

ETUD		**GOUV**	
Cooccurrence	**Indice**	**Cooccurrence**	**Indice**
Québec-milliard	5,00	Québec-révélateur	3,62
Québec-diriger	4,20	Québec-comparer	3,30
Québec-conservateur	3,58	Québec-généreux	2,94
Québec-progressiste	3,36	Québec-habiter	2,38
Québec-député	3,01	Québec-décennie	2,16
Québec-francophone	2,96	Québec-gouverne	2,42
Québec-entier	2,65	Québec-problème	2,67
Québec-insulte	2,57	Québec-région	2,35
Québec-aréna	2,35	Québec-canadien	2,19
Québec-fouet	2,35		
Québec-surnombre	2,35		
Québec-dynamique	2,35		
Québec-fiscal	2,19		

Nous avons remarqué que le mot pôle 'Québec' est associé majoritairement au gouvernement du Québec et à l'administration de la province et dans une moindre mesure à la région géographique. Dans ETUD, 'Québec' apparaît dans des arguments critiques face au gouvernement **['Québec-diriger, Québec-député', Québec-dynamique]**, à son inhabileté dans la gestion de l'argent public **['Québec-milliard', 'Québec-fiscal', 'Québec-aréna']**, et en lien avec des mesures conservatrices **['Québec-conservateur']** ou austères **['Québec-**

insulte', 'Québec-fouet'] qui sont adoptées. Les auteurs d'ETUD préconisent la mise en place d'un gouvernement plus progressiste **['Québec-progressiste']**. Le mot pôle 'Québec' se trouve également associé à la notion population **['Québec-francophone'],** et évoque l'idée de totalité et de majorité, comme le démontrent les cooccurrents **[Québec-surnombre', 'Québec-entier']**.

Dans GOUV, 'Québec' est aussi associé au gouvernement **[Québec-gouverne]**, mais évoque plutôt une critique face à la générosité des programmes sociaux du Québec **[Québec-généreux]**, ainsi qu'il évoque les questions économiques centrales pour la province et pour ses habitants **['Québec-région', 'Québec-habitant']** qui ont été soulevées par la grève des étudiants **['Québec-révélateur', 'Québec-problème', 'Québec décennie']**. Se trouve également présent le thème de la comparaison entre le Québec et les autres provinces du Canada **['Québec-comparer', 'Québec-canadien']**.

Mot pôle : scolarité

Tableau 33. Cooccurrents du mot pôle 'scolarité'

ETUD		**GOUV**	
Cooccurrence	**Indice**	**Cooccurrence**	**Indice**
scolarité-dégel	9,65	scolarité-annoncer	4,93
scolarité-imposer	4,45	scolarité-réclamer	3,91
scolarité-afférent	4,28	scolarité-graduel	3,06
scolarité-concerner	3,39	scolarité-lutter	2,86
scolarité-parler	3,35	scolarité-boycotteur	2,68
scolarité-entêtement	2,54	scolarité-garderie	2,66
scolarité-embûche	2,52	scolarité-justifier	2,56
scolarité-loi-cadre	2,52	scolarité-accroître	2,40
scolarité-tarif	2,52	scolarité-payer	2,12
scolarité-impôt	2,46	scolarité-gratuité	2,08
scolarité-dossier	2,31	scolarité-choix	2,07
scolarité-matériel	2,31		
scolarité-applicable	2,06		

scolarité-controverse	2,06	
scolarité-élevé	2,06	
scolarité-foulée	2,06	
scolarité-montée	2,06	
scolarité-récurrent	2,06	
scolarité-amener	2,01	

Dans le corpus de référence, le mot scolarité se réfère dans la plupart des cas à l'expression « frais de scolarité ». Les cooccurrents spécifiques d'ETUD et GOUV mettent en évidence les différences entre les perspectives par rapport à la hausse des frais. Nous remarquons que dans ETUD, 'scolarité' se trouve associée à l'idée d'imposition et d'entêtement **['scolarité-imposer', 'scolarité-entêtement', 'scolarité-embûche']** ce qui concorde avec la perspective des auteurs de ce sous-corpus par rapport à la décision du gouvernement. Aussi, les auteurs d'ETUD insistent sur la nécessité d'une discussion portant sur la hausse des frais **['scolarité-dossier', 'scolarité-concerner', 'scolarité-parler']**. Par contre, dans GOUV, la notion d'imposition est nuancée et 'scolarité' est plutôt liée à 'annoncer' **['scolarité-annoncer']** et à la nécessité de faire un choix **['scolarité-choix']**.

Des différences peuvent être aussi perçues dans la manière dont la question de la hausse est abordée par chaque groupe en fonction des cooccurrents retrouvés. Dans ETUD, le mot pôle scolarité est associé à dégel **['scolarité-dégel']**, à la nécessité d'une loi-cadre qui empêcherait le gouvernement d'imposer des augmentations par décret dans les services publics **['scolarité-loi-cadre']**, ou encore aux coûts impliqués dans l'éducation **['scolarité-impôt', 'scolarité-tarif', 'scolarité-afférent', 'scolarité-matériel']**. Dans GOUV, les thèmes sont plus en relation avec les bénéfices qui justifieraient la hausse **['scolarité-justifier', 'scolarité-accroître']**. Le cooccurrent **'scolarité-garderie',** par exemple, apparaît dans le contexte de justification de la hausse, évoquant les conséquences du sous-financement des universités pour d'autres services publics essentiels à la société. Contrairement à la défense d'un dégel, les auteurs de GOUV préconisent une hausse graduelle et modérée et critiquent la position des étudiants par rapport à la hausse **['scolarité-gratuité', 'scolarité-payer]** et par rapport au mouvement **['scolarité-boycotteurs', 'scolarité-lutter', 'scolarité-réclamer']**.

Mot pôle : gouvernement

Tableau 34. Cooccurrents du mot pôle ‘gouvernement’

ETUD		**GOUV**	
Cooccurrence	**Indice**	**Cooccurrence**	**Indice**
gouvernement-fédéral	5,49	gouvernement-décréter	3,87
gouvernement-accepter	4,36	gouvernement-succéder	3,70
gouvernement-argument	3,86	gouvernement-impopularité	3,25
gouvernement-décider	3,53	gouvernement-ouverture	3,00
gouvernement-actuel	3,24	gouvernement-gérer	2,96
gouvernement-scandale	3,21	gouvernement-reprocher	2,96
gouvernement-sérieux	2,97	gouvernement-signifier	2,91
gouvernement-preuve	2,90	gouvernement-souhaiter	2,76
gouvernement-dupe	2,67	gouvernement-dénouement	2,70
gouvernement-attitude	2,59	gouvernement-céder	2,52
gouvernement-intransigeance	2,38	gouvernement-tort	2,43
gouvernement-irresponsable	2,38	gouvernement-réduire	2,43
gouvernement-suspendre	2,38	gouvernement-concession	2,41
gouvernement-corrompre	2,30	gouvernement-renverser	2,41
gouvernement-incomber	2,30	gouvernement-consister	2,28
gouvernement-argent	2,29	gouvernement-subventionner	2,22
gouvernement-choisir	2,28	gouvernement-augmentation	2,21
gouvernement-artisan	2,17	gouvernement-déplait	2,18
gouvernement-convier	2,17	gouvernement-décision	2,10
gouvernement-mentir	2,17	gouvernement-offre	2,07
gouvernement-saboter	2,17	gouvernement-muscler	2,06
gouvernement-minier	2,03	gouvernement-user	2,06
		gouvernement-fermer	2,05
		gouvernement-adopter	2,03

Dans les contextes analysés, nous avons remarqué que le mot pôle ‘gouvernement’ réfère à l’administration du premier ministre Jean Charest et, dans une moindre mesure, au gouvernement fédéral. Dans ETUD, les cooccurrents de ‘gouvernement’ révèlent les critiques

et les accusations des auteurs contre le gouvernement actuel **['gouvernement-actuel', 'gouvernement-irresponsable', 'gouvernement-corrompre', 'gouvernement-mentir', 'gouvernement-saboter', 'gouvernement-dupe']**, incluant des critiques précises quant à la mauvaise gestion de l'argent public **['gouvernement-minier']**. D'autres cooccurrents mettent en évidence les demandes que les auteurs d'ETUD dirigent au gouvernement ainsi que les actions qu'il devrait entreprendre par rapport à la crise **['gouvernement-incomber', 'gouvernement-accepter', 'gouvernement-décider', 'gouvernement-suspendre', 'gouvernement-choisir', 'gouvernement-sérieux']**.

Les auteurs de GOUV formulent aussi des critiques contre le gouvernement, mais par rapport à la gestion de la crise **['gouvernement-gérer', 'gouvernement-reprocher', 'gouvernement-décréter', 'gouvernement-fermer', 'gouvernement-muscler', 'gouvernement-déplait']** et mettent en évidence l'image du gouvernement éclaboussée par son attitude face à la grève **['gouvernement-user']**. Dans ce sous-corpus, nous remarquons également un thème sur les concessions faites par le gouvernement **['gouvernement-subventionner']**, l'effort d'expliquer ou même de justifier les actions de ce dernier **['gouvernement-consister', 'gouvernement-signifier', 'gouvernement-adopter']**, et sur le bien-fondé des actions entreprises par le gouvernement **['gouvernement-souhaiter', 'gouvernement-réduire']**. La récurrence de certains cooccurrents reliés aux élections atteste également de la présence de ce thème spécifique dans le sous-corpus GOUV **['gouvernement-renverser', 'gouvernement-succéder', 'gouvernement-céder'].**

Mot pôle : droit

Tableau 35. Cooccurrent du mot pôle 'droit'

ETUD		**GOUV**	
Cooccurrence	**Indice**	**Cooccurrence**	**Indice**
droit-augmenter	6,92	droit-hausser	5,25
droit-renoncer	5,78	droit-brimer	3,94
droit-privilège	4,24	droit-abolition	3,31
droit-pacte	4,17	droit-régressif	2,56
droit-international	4,06	droit-majorer	2,54

droit-expression	3,68	droit-veto	2,54
droit-contractuel	3,63	droit-parole	2,46
droit-garantir	3,53	droit-dépasser	2,36
droit-global	2,81	droit-admissible	2,08
droit-indexer	2,67	droit-démocrate	2,08
droit-primauté	2,67	droit-efficacité	2,08
droit-obligation	2,65	droit-explorer	2,08
droit-vote	2,58	droit-injecter	2,08
droit-bafouer	2,48		
droit-faveur	2,47		
droit-inaliénable	2,42		
droit-défendre	2,37		

Le mot pôle ‘droit’ réfère à « droits de scolarité », mais aussi aux droits dans le sens de l’ensemble des règles juridiques. Dans ETUD, une thématique sur les droits juridiques rassemble des thèmes divers sur la liberté d’expression **[‘droit-expression’]**, le droit de vote dans les assemblées étudiantes **[‘droit-vote’]**, les principes fondamentaux reliés aux droits de l’homme **[‘droit-inaliénable’, ‘droit-obligation’, ‘droit-primauté’]**, les pactes internationaux **[‘droit-international’, ‘droit-pacte’]**, le droit à l’éducation **[‘droit-éducation’]** et l’idée que l’éducation n’est pas un privilège **[‘droit-privilège’]**. Quelques cooccurrents mettent en évidence également l’attitude que les étudiants doivent prendre par rapport à ces droits **[‘droit-défendre’, ‘droit-garantir’]**. En ce qui concerne les droits de scolarité, droit apparaît lié à la question de l’augmentation **[‘droit-augmenter’]**, au droit contractuel des étudiants **[‘droit-contractuel’]**, aux personnes favorables à la hausse **[‘droit-faveur’],** et aussi favorables à ce que le gouvernement renonce à la hausse **[‘droit-renoncer’]**.

Dans GOUV, le mot pôle droit est plutôt associé à des contextes reliés à la hausse de droits de scolarité et à la défense de la hausse **[‘droit-hausser’, ‘droit-majorer’, ‘droit-admissible’]** pour injecter de l’argent dans les universités et pour améliorer son efficacité **[‘droit-efficacité’, ‘droit-injecter’]**. D’autres cooccurrents révèlent une critique sur les

propositions des étudiants relativement à la hausse **[‘droit-abolition’, ‘droit-régressif’]** et au mouvement étudiant, accusant ce dernier de dépasser les enjeux initiaux **[‘droit-dépasser’]**. Les auteurs de GOUV proposent d’explorer de solutions pour l’augmentation des droits qui pourraient entraîner la sortie de la crise **(‘droit-explorer’)**. Quelques cooccurrents de ‘droit’ dans le sous-corpus GOUV sont aussi actualisés avec une notion de droits juridiques et construisent un thème sur la nécessité de respecter les droits des étudiants qui ne veulent pas participer au mouvement **[‘droit-brimer’, ‘droit-veto’, ‘droit-parole’, ‘droit-démocratie’]**.

Mot pôle : hausse

Tableau 36. Cooccurrents du mot pôle ‘hausse’

ETUD		**GOUV**	
Cooccurrence	**Indice**	**Cooccurrence**	**Indice**
hausse-moratoire	4,98	hausse-proposer	6,97
hausse-approuver	4,56	hausse-maintenir	6,81
hausse-favorable	4,52	hausse-raisonnable	5,95
hausse-affecter	4,01	hausse-décret	4,39
hausse-brutal	3,16	hausse-objet	3,94
hausse-débat	2,90	hausse-accompagner	2,87
hausse-prêcher	2,75	hausse-baguette	2,54
hausse-scander	2,75	hausse-objecter	2,54
hausse-tenant	2,64	hausse-contester	2,45
hausse-annonce	2,45	hausse-reculer	2,45
hausse-privatisation	2,45	hausse-abolir	2,33
hausse-mentalité	2,29	hausse-équivaloir	2,33
hausse-modéré	2,29	hausse-concret	2,15
hausse-automne	2,02	hausse-prévoir	2,09
hausse-baser	2,00	hausse-étirer	2,08
hausse-coalition	2,00	hausse-étalement	2,08
		hausse-incontournable	2,08
		hausse-majoration	2,08

Dans ETUD, les cooccurrents autour du mot pôle ‘hausse’ attestent d’un thème portant sur les personnes qui supportent la hausse **[‘hausse-favorable’, ‘hausse-prêcher’, ‘hausse-scander’, ‘hausse-tenant’, ‘hausse-coalition’]**, sur les propositions des étudiants qui sont contre la hausse **[‘hausse-moratoire’]** et sur la nécessité d’un débat relativement à ce sujet **[‘hausse-débat’]**. Les auteurs ressortent les arguments sur lesquels se base la hausse **[‘hausse-baser’]** et sur la proposition d’une hausse modérée **[‘hausse-modérer’]**, en disant qu’elle est brutale **[‘hausse-brutal’]**. Ils dénoncent également les dangers de la prééminence d’une mentalité favorable à la privatisation des services publics **[‘hausse-privatisation’, ‘hausse-affecter’, ‘hausse-mentalité’]** et critiquent la manière dont la hausse a été proposée, c’est-à-dire sans faire de discussion sociale plus approfondie **[‘hausse-annonce’, ‘hausse-approuver’]**. Dans GOUV, la hausse est plutôt proposée **[hausse-proposer]** et les auteurs sont déjà convaincus qu’elle rentrera en vigueur bientôt **[‘hausse-prévoir’]** malgré l’opposition des étudiants **[‘hausse-contester’, ‘hausse-objecter’]**. Contrairement à la proposition de moratoire défendue par les étudiants, les auteurs de GOUV soutiennent le maintien de la hausse **[‘hausse-maintenir’]** et défendent l’idée qu’elle est incontournable **[‘hausse-incontournable’]** et qu’il n’est pas possible pour le gouvernement de revenir sur sa décision **[‘hausse-reculer’]**. Selon les auteurs GOUV, le gouvernement devrait répondre aux problèmes concrets soulevés par les étudiants **[‘hausse-concret’]**. D’autres cooccurrents dans GOUV font état des propositions défendues par Pauline Marois, candidate aux élections, qui préconise l’abolition de la hausse **[‘hausse-abolir]**. Pauline Marois est durement critiquée par les auteurs GOUV pour cette position qu’on juge déraisonnable et trompeuse, et on l’accuse de soumettre cette position uniquement dans une visée électoraliste **[‘hausse-baguette’]**. De plus, par rapport au mot pôle ‘hausse’ nous retrouvons un thème dans GOUV qui justifie l’aspect raisonnable de la hausse **[‘hausse-raisonnable’, ‘hausse-équivaloir’]** et qui valorise les propositions gouvernementales visant à atténuer ses effets à long terme **[‘hausse-accompagner’, ‘hausse-étirer’, ‘hausse-majoration’, ‘hausse-étalement’]**.

2.4.2 Dialectique

Cette section présente les critères textuels dialectiques retenus à la suite du calcul des spécificités.

Dans le tableau 37 ci-dessous figurent les critères dialectiques retenus pour les sous-corpus ETUD et GOUV. Nous avons départagé les critères en deux catégories : ceux qui sont caractéristiques de l'organisation temporelle du récit et ceux qui sont caractéristiques de la structuration argumentative de chaque sous-corpus. La quantité de critères dialectiques retrouvés dans les deux sous-corpus est faible comparativement au nombre de critères des composantes thématiques et dialogiques. Mais certaines différences remarquables ont été observées entre les deux sous-corpus.

Tableau 37. Critères dialectiques

Catégorie	**ETUD**		**GOUV**	
Structuration argumentative	**Spéc.**	**Critère**	**Spéc.**	**Critère**
	+4	'-il'	+11	'Mais'
	+4	'-nous'	+8	'ne'
	+4	','	+8	'pas'
	+4	'de nouveaux'	+6	'Mais il'
	+4	'et plus'	+5	'…'
	+4	'et'	+5	'n''
	+3	'alors qu''	+4	'Mais elle'
	+3	'car il'	+4	'même si'
	+3	'de plus en plus'	+4	'Parce que'
	+3	'dire que les'	+3	'cela ne'
	+3	'En effet'	+3	'comme ça'
	+3	'est encore'	+3	'dans ce dossier'
		'est là'		'est pas parce'
		'et surtout'		'Mais la'
	+3	'manifestement'	+3	'Mais les'
	+3	'non pas'	+3	'mais'
	+3	'non seulement'	+3	'par ailleurs'
	+3	'plus de'	+3	'par contre'
	+3	'plus grand'	+3	'par la suite'
	+3	'près de'	+3	'parce'
	+3	'quelle'	+3	'pas besoin'

	+3	‘question de’	+3	‘Pas’
	+3	‘tout comme’	+3	‘Quand on’
	+3	‘véritable’	+3	‘si on’
	+3	‘alors qu’’	+3	‘Sur le’
	+2	‘à cette’	+3	‘TOUS’
	+2	‘arriver à’	+3	‘cela ne’
	+2	‘au détriment’	+3	‘comme ça’
	+2	‘Au lieu’	+2	‘--’
	+2	‘encore’	+2	‘-ils’
	+2	‘notamment’	+2	‘?’
	+2	‘pourtant’	+2	‘.’
	+2	‘Toutefois’	+2	‘Bref’
	+2	‘voilà’	+2	‘et donc'
			+2	‘là-dessus’
			+2	‘Parce’
			+2	‘si elle’
			+2	‘si je’
			+2	‘si vous’
			+2	‘très’
			+2	‘voulez’
Organisation temporelle	+5	‘de demain’	+4	‘pendant la’
	+4	‘2012’	+3	‘au cours des’
	+3	‘aujourd'hui’	+3	‘au terme’
	+3	‘demain’	+3	‘cette année’
	+3	‘Depuis des’	+2	‘hier’
	+3	‘dès le’	+2	‘y a quelques’
	+3	‘moment’		
	+2	‘1960’		
	+2	‘2005’		
	+2	‘ce temps’		
	+2	‘Depuis’		
	+2	‘depuis’		
	+2	‘Lors’		

a) Organisation temporelle des sous-corpus ETUD et GOUV

Relativement à l’organisation temporelle, nous avons retrouvé dans le sous-corpus ETUD des adverbes et de mots lexicaux signalant un ancrage dans le temps présent, décrivant des processus qui se déroulent au moment de l’énonciation (Charaudeau, 1992, p.452). En plus de l’effet d’actualité et de référence au temps chronologique, ces critères font également allusion au temps vécu et expérimenté historiquement (Charaudeau, 1992, p. 465), lequel est souvent comparé à des événements antérieurs **[‘aujourd’hui’, ‘moment’]**. Le critère ‘moment’ en particulier évoque les événements présents dans une perspective valorisante. L’allusion à des événements historiques est aussi indiquée par la spécificité de critères liés à des moments ponctuels présents ou passés **[‘2005’, ‘1960’, ‘2012’],** renvoyant tantôt au moment de la grève **(‘2012’)**, tantôt à des événements historiques précis, comme l’année de mise en place du programme fédéral d’aide aux étudiants **[‘1960’]** et la modification du régime de prêts et bourses en 2005. Le critère **‘au sein’** corrobore la notion de ponctualité, indiquant le lieu physique où se situent les événements. Par ailleurs, des critères spécifiques d’ETUD indiquent la visée rétrospective de ces processus, en remontant à l’origine des événements **[‘Depuis des’, ‘dès le’, ‘depuis’]**. (Charaudeau, 1992, p. 479). Voici quelques contextes dans lesquels ces critères ont été retrouvés :

> [news20120405MO0057] (…) de vie et de travail que nous avons *aujourd’hui*, les mesures de solidarité sociale qui nous protègent tous de la (…)

> [news20120420LT0022] (…) Madame Beauchamp Vous dormez sans doute mal en ce *moment* Y’a ben du monde Qui attendent un dénouement(…)

> [news20120409NV0019] (…) donner un cours d’histoire, j’indiquerai quand même deux *moments* importants, en 1963 et en 1976,

> [news20120428TB0029] (…) sur les hausses de frais de scolarité. *Lors* d’entretiens avec certains étudiants, cette suggestion semblait une alternative intéressante (…)

[news20120428TB0025] (…) débattent, argumentent, critiquent. *Depuis* des semaines, les appels au débat se multiplient dans les journaux (…)

[news20120611VE0014] (…) Ce ton a été donné, d'ailleurs, *dès le* début du conflit par la ministre de l'Éducation elle-même, Line Beauchamp, (…)

[news20120224LE20120224_a8_moral] (…) l'engagement moral qui balise la conduite de la démocratie étudiante *depuis* longtemps. Il est vrai que le droit de grève des étudiants (…)

L'effet d'anticipation (Charaudeau, 1992, p. 471) provoqué par **'demain'** et aussi **'de demain'** est aussi caractéristique de l'organisation temporelle d'ETUD**.** Les auteurs ETUD anticipent les conséquences néfastes qu'une hausse pourrait avoir pour les générations futures.

[news20120227MO0031] (…) juger des choses par nous-mêmes, c'est la liberté du Québec *de demain* qui en dépend. Nous serons ainsi des citoyens (…)

[news20120427LT0028] (…) unis par leurs seules revendications. Mais ils façonnent le Québec *de demain*, et doivent le faire en citoyens responsables (…)

La temporalité de GOUV est caractérisée également par des événements ponctuels, mais qui sont plutôt relatifs au passé récent **['hier', 'y a quelques', 'cette année'].** Quelques critères liés à la temporalité indiquent une vision durative des processus, lesquels sont bornés par des limites plus ou moins précises **['pendant la', 'au cours des', 'derniers jours']** (Charaudeau, 1992, p. 478-479).

[news20120229MO0036] (…) M. Lagacé a soulevé un très bon point *hier* en écrivant que beaucoup de personnes ne savent pas comment les leaders du Mouvement (…)

[news20120417LA0039] (…) Comment se fait-il que 6400 étudiants français ont choisi *cette année* de s'inscrire dans nos universités « payantes » a (…)

[news20120414NV0028] (…) l'autoroute 20 *durant* des heures. Il y a quelques mois, au port de Trois-Rivières, des grévistes ont coupé l'alimentation (…)

[news20120419LT0027] (…) À l'université a régné au cours des *derniers jours* un climat totalement à l'opposé de celui que l'on associe aux études supérieures (…)

[news20120912LA004]1 (…) Il n'y a pas vraiment eu de débat sur ces mesures *pendant la* campagne. Ce silence s'explique (…)

[news20120601TB0017] (…) jusqu'au prochain scrutin, qui aura lieu *au cours de* la prochaine année, très probablement dès l'automne.

a) Structuration argumentative des sous-corpus ETUD et GOUV

En ce qui concerne la structuration argumentative, la caractéristique qui ressort le plus dans le sous-corpus ETUD est l'effet de saturation et d'accumulation provoqué par la présence d'adverbes qui expriment la valeur de surenchère (Charaudeau, 1992, p. 504). La récurrence de ces traits sémantiques correspond à une stratégie argumentative consistant à utiliser plusieurs arguments de preuve. Font partie de ces critères la virgule (','), signe de ponctuation qui est ressorti parmi les critères les plus spécifiques et qui renforce la notion d'accumulation, aussi bien que la conjonction 'et' **['de nouveaux', 'et plus', 'de plus en plus', 'est encore', 'non seulement', 'plus de', 'plus grand', 'tout comme', 'et', 'mais aussi', 'encore', 'près de', 'et surtout', ',' 'mais aussi', 'arriver à']**. Les intensificateurs **['véritable', 'manifestement']** font également état de cette notion de saturation et d'intensité, en exprimant une valeur de totalité (Charaudeau, 1992, p. 253)

[news20120221LS0043] (…) Le moins que l'on puisse dire est que le torchon brûle et brûlera *encore* entre la ministre Line Beauchamp et les différentes associations étudiantes (…)

[news20120217LE20120217_a9_vivement] (…) qui se pratique pas mal en ce moment et ça fait *encore* plus mal quand les entreprises quittent le pays (…)

[news20120404LE20120404_a9_sous] (…) 'inquiéter, pas vis-à-vis de ses hésitations) d'un autre monde *plus* juste et *plus* démocratique. Dans ces conditions (…)

[news20120404QT0019] (…) vous écrivons *non seulement* à titre d'anciens étudiants, mais aussi à titre d'enseignants, et *surtout* en tant que citoyens. (…)

[news20120410MO0042] (…) Secor, près de 150 milliards de dollars *et plus* de 20 milliards en revenus fiscaux, sur 25 ans. (…)

[news20120410MO0042] (…) le Plan Nord doit rapporter autant que le dit le gouvernement ? Un montant qui comblerait ce dont les universités ont besoin *et plus encore*. (…)

[news20120418VE0018] (…) frais de scolarité. Le débat se glisse alors vers *de nouveaux* horizons. Cela place les étudiants en opposition (…)

[news20120419TB0020] (…) aux problématiques en jeu. *Tout comme* vous, nous sommes anxieux devant la somme de travail et d'organisation (…)

[news20120619LE2012-06-19_352769] (…) l'intimidation me semblent bien capricieux, *tout comme* je trouve les sujets de vos contrariétés (…)

[news20120226LS0035] (…) C'est pourtant un *véritable* débat de société qui devrait être refait autour de cette question. (…)

[news20120418VE0019] (…) preuve, notre système de santé gargantuesque qui engloutit les milliards sans *véritable* progrès. Ne serait-il pas plus pertinent de (…)

La présence de la formulation 'dire que les' dans ETUD fait état d'une mise à distance visant à atténuer les effets agressifs d'une interpellation directe (Charaudeau, 1992, p. 149). En effet, la stratégie argumentative des auteurs ETUD n'est pas marquée par l'agressivité, elle cherche plutôt à fournir des explications sur les positions défendues, comme montre la présence de certains adverbes **['voilà'', est là', 'à cette', 'en matière', 'question de', 'En effet', 'notamment', 'car il']** (Charaudeau, 1992, p. 790) et des marqueurs de comparaison **['au détriment', 'au profit', 'Au lieu']**(Charaudeau,1992, p. 822).

[news20120406LE20120406_a8_etincelle] (…) l'étude à crédit est proposée comme rempart de l'accessibilité aux études, *alors qu*'il n'en est rien. *Voilà* un discours (…)

[news20120405LE20120405_a8_nombrils] (…) Dans cette grève qui s'étire en longueur sans *pourtant* ébranler les croyances gouvernementales, (…)

[news20120605NV0021] (…) Parce que notre plus grande richesse, c'est eux et *non pas* quelques dollars de plus (…)

[news20120619LE2012-06-19_352769] (…) Devant votre attitude si formidablement intraitable *en matière de* violence et d'intimidation (…)

[news20120324LE20120324_b3_pas] (…) pour le gouvernement et son chef, mais il ne s'agissait pas d'une *question de* principe fondamentale. (…)

[news20120427VE0017] (…) personnes qui devront rembourser les emprunts. Et c'est là que le mot « régressif » doit être utilisé. (…)

[news20120425LT0023] (…) par les intérêts respectifs de ces deux générations au détriment de considérations plus larges sur un véritable projet de société (…)

[news20120605NV0021] (…) les plus démunis qui en paient d'abord le prix, *au profit* d'un petit nombre de privilégiés, (…)

Nous avons également retrouvé dans ETUD des critères textuels qui révèlent une stratégie d'interrogation rhétorique (Charaudeau, 1992, p.133) et qui expriment de l'indignation **['-il', '-nous', 'voulons-nous', 'quelle']** :

[news20120223NV0027] (…) Comment pourrons-*nous* expliquer à ces étudiants que selon le gouvernement Charest, faire leur juste part signifierait de ne pas entrer à l'université ? (…)

[news20120216LE20120216_a8_avant] (…) La bourde de 2004 avait soulevé une redoutable et efficace vague étudiante, avec une grève générale percutante. Cet épisode peut-*il* être répété ? (…)

[news20120412MO0045] (…) L'autre question dont il faudrait débattre, c'est : quel avenir voulons-*nous* assurer à ces étudiants ? (…)

[news20120404QT0019] (…) Les écoles ne sont pas saines. Elles rendent les enfants malades. Mais *quelle* importance, puisqu'ils parleront tous l'anglais fluently ? (…)

Nous avons remarqué dans le sous-corpus GOUV la spécificité d'adverbes qui marquent l'opposition entre deux assertions et qui sont utilisés pour réfuter des arguments adverses **['alors qu', 'pourtant', 'Toutefois']**. Ces critères signalent la présence d'une contradiction entre deux assertions (Charaudeau, 1993, p. 523). Nous avons remarqué aussi des oppositions restrictives, c'est-à-dire qui admettent quelque chose de vrai dans l'assertion qui les précèdent **['Toutefois']** (Charaudeau, 1992, p. 790). La présence de négations du type « réplique » **['non pas', 'cela ne']** fait également partie de cette stratégie de réfutation, tout en apportant dans l'assertion précédente une explication (Charaudeau, 1992, p. 790).

Nous avons également relevé dans GOUV des démonstratifs dénotant un registre familier du langage et qui apportent un effet de familiarité conversationnelle **('là-dessus', 'comme ça', 'Sur le', 'dans ce dossier']** (Charaudeau, 1992, p. 232). Ces démonstratifs sont aussi liés à des propos qui expriment l'indignation, comme on peut le voir dans les exemples ci-dessous :

[news20120927LA0052] (…) quand le gouvernement aura plié *là-dessus* pour acheter la paix, ne vous satisfaites pas de si peu. (…)

[news20120424NV0047] (…) Il y a un principe de base qu'on ne comprend pas ici. Ce n'est pas *comme ça* que ça marche dans une démocratie. (…)

En ce qui concerne la structuration argumentative de GOUV, nous avons observé plusieurs critères relevants de phrases négatives **['n', 'pas', 'ne', 'Pas', 'pas besoin', 'ne', 'est pas parce']**. En analysant le contexte de ces phrases, nous constatons que la stratégie argumentative des auteurs GOUV consiste à nier les arguments des étudiants et de leurs partisans ou à rendre illégitimes leurs affirmations (Charaudeau, 1992, p. 563-565).

[news20120502LA0047] (…) En effet, si on *n*'augmente *pas* le financement des universités, pas besoin de demander davantage aux étudiants (…)

[news20120215OP120215226595245] (…) pour les services de garde. Aucun gel *ne* peut plus être maintenu. Ils ne sont qu'une monnaie pour (…)

[news20120215OP120215226595245] (…) Les membres de celles-ci *ne* sont *pas* des employés de l'État, mais des bénéficiaires de la subvention (…)

[news20120219OR120219227381050] (…) aux générations à venir et je *ne* peux *pas* croire qu'elles ne pourront pas bénéficier de ce que les générations (…)

[news20120504LE2012-05-04_349172] (…) Line Beauchamp a décidé d'exclure la CLASSE de la table de discussions, ce *n*'est *pas parce* qu'elle a découvert, avec stupeur, que cette association étudiante radicale (…)

L'argumentation de GOUV est aussi marquée par des traits indiquant la polémicité (Charaudeau, 1992, p. 253-260), comme les formulations en majuscules qui expriment de l'intensité **['TOUS']** et des intensificateurs forts **['très']** :

[news20120314OP120314230493361](…) artiste porte un carré rouge ? Ils vont *TOUS* porter un carré rouge. So-so-solidarité, les amis. Et marchons (…)

[news20120708OP120708249728045](…) pas le mandat de refléter l'opinion de *TOUS* les Québécois, mais celle de leurs lecteurs. Comme Le Figaro (…)

[news20120225VE0018](…) Car ces derniers ont placé la barre *très* haute et chauffé leurs troupes à blanc avec une cassette d'arguments (…)

[news20120402QT0016](…) passent leurs mandats de grève avec presque toujours de *très* faibles majorités, donc ils ne peuvent affirmer qu'ils représentent (…)

[news20120609LA0047](…) Une meilleure répartition de la richesse, avez-vous dit ? *Allez* donc demander aux camarades bien nantis qui affichent fièrement un carré rouge (…)

Parmi les adverbes d'opposition, nous avons relevé dans GOUV la spécificité de critères indiquant une opposition plus prononcée que celle retrouvée dans ETUD **['Mais', 'par contre', 'Mais il', 'Mais elle', 'Mais la', 'Mais les**']. L'opposition dans GOUV a un

caractère restrictif : **‘**mais’ et ‘par contre’, par exemple, mettent en opposition une assertion à une autre, mais dans laquelle quelque chose de vrai ou de plausible est néanmoins admis (Charaudeau, 1992, p.512). Nous avons remarqué également la présence de certaines formulations avec la conjonction ‘si’ **[‘si elle’, ‘si je’, ‘si on’, ‘si vous’**]. Ces formulations n’expriment pas d’hypothèses, mais signalent des relations de cause et de conséquence entre les assertions.

> [news20120221NV0022] (…) à déclencher une grève illimitée pour protester contre la hausse des frais de scolarité. *Mais* à qui peut nuire une grève des étudiants sinon à eux-mêmes ? (…)

> [news20120302OR120302228889283] (…) L’appui à la présente grève est *par contre* mitigé l’Université Laval où pas moins de 45 000 (…)

> [news20120222LS0035]Une hausse de 325 $ par année sur cinq ans ne me paraît pas démesurée *si on* sait que l’État paiera toujours 80 %

> [news20120407OP120407235832252] qui chiera le plus gros étron verbal. *Si vous* voyiez les courriels et les tweets que je reçois depuis deux semaines

> [news20120504LE2012-05-04_349172] J’aurais pu m’épanouir comme mécano *si j*’avais aimé les moteurs, aussi. L’université ne garantit pas le bonheur.

> [news20120613OP120613246255078] avais rien écrit de tout cela. Même *si j’*applaudissais à deux mains les politiciens corrompus et tripais sur Stephen Harper (…)

Les formes interrogatives sont également assez présentes dans GOUV comme montre la spécificité du point d’interrogation et d’autres indicateurs de phrases interrogatives **[‘ ?’, ‘-ils’]**. Les auteurs répondent généralement eux-mêmes aux questions qu’ils ont posées dans le texte, comme le montre la spécificité de **[‘Parce que’ ‘-ils’, ‘pas parce que’, ‘parce’]**. La présence de ces phrases interrogatives révèle une stratégie de provocation et de dénégation dans GOUV : le questionnement consiste à proposer un argument qui est rejeté à l’avance (Charaudeau, 1992, p. 136)

[news20120221NV0022] (…) Les étudiants peuvent-*ils* se permettre de perdre un an de salaire ? (…)

[news20120314LA0046] (…) ce débat reposait sur des bases fumeuses. *Parce que* la hausse des droits ne pénalisera pas les pauvres. *Parce que* le gel des droits n'est pas progressiste (…)

[news20120323LA0040] (…)Mais ce n'est *pas parce que* le gouvernement n'accepte pas de négocier avec les manifestants qu'il ne doit pas écouter ce qu'ils ont à dire. (…)

[news20120314LA0046] (…) les artisans du cinéma sont des alliés naturels des étudiants, *parce que* le monde des arts est plus à gauche, (…)

D'autres critères relevant de la structuration argumentative de GOUV sont des traits terminatifs, exprimant la conclusion (Charaudeau, 1992, p. 788) et impliquant une assertion qui doit être acceptée du fait de l'assertion de départ **['au terme' 'par la suite', 'Bref', 'et donc']**. Nous avons remarqué également un certain nombre d'indices qui caractérisent la tentative de fournir des explications relatives à l'argumentaire qui est fait. Par exemple, la présence de marqueurs explicatifs comme le double tiret **['--']** introduisent des explications à l'intérieur d'une phrase, et la présence de certains marqueurs explicatifs comme [**'même si']** et **'[quand on']** (Charaudeau, 1992, p.514). . Nous avons détecté également la spécificité du point de suspension **['…']**, qui est souvent utilisé pour exprimer la perplexité ou la dérision à l'égard d'un argument du groupe adverse. Le point final a été aussi retenu comme critère ['.'], puisqu'il contrastait avec le suremploi de la virgule dans le sous-corpus ETUD et parce qu'il peut être associé également à une structuration argumentative terminative.

[news20120430LA0033] (…)Assez pour conclure, *au terme* de ce vaste débat public, que la balance penche clairement en faveur des partisans de cette hausse. (…)

[news20120428OR120428239768866] (…) financière. Un tel précédent voudrait être imité *par la suite* par tous les groupes qui confrontent un jour ou l'autre (…)

[news20120417NV0018] la viande halal, et bla-bla-bla. *Bref*, si on veut chialer, ce ne sont pas les choix qui manquent. (…)

[news20120507OR120507240980335] (…) faible participation à la manifestation démontrait bien *par ailleurs* que la mobilisation s'était effilochée dans les rangs des boycotteurs (…)

[news20120403OP120403234261946] (…) colossal nous permet d'empocher via la péréquation -- argent que nous utilisons pour nous payer des programmes en or que (…)

[news20120326OP120326232493039] (…) veut annuler la hausse des frais de scolarité ... Mais où va-t-elle prendre tout cet argent ? Mais (…)

[news20120328OP120328232883104] (…) toujours attendre que l'argent vienne des autres ... UNE HONTE ! Imaginez ... (…)

2.4.3 Dialogique

Cette section présente les critères textuels dialogiques retenus à la suite du calcul des spécificités.

Nous présentons dans le tableau 38 ci-dessous l'ensemble de critères dialogiques spécifiques aux sous-corpus ETUD et GOUV repérés par notre analyse. Nous avons départagé les critères en deux catégories, ceux qui font état de l'énonciation représentée et ceux qui sont liés à la modalisation.

Tableau 38. Critères dialogiques

Catégorie	**ETUD**		**GOUV**	
Énonciation	**Spéc.**	**Critère**	**Spéc.**	**Critère**
	+12	'nous'	+5	'étudiants qui'
	+10	'que nous'	+5	'ils n''
	+9	'«'	+5	'Les'
	+9	'»'	+5	'on'
	+7	de notre'	+5	'ont'

+4	‘à nous’	+4	‘et ceux’
+4	‘Madame’	+4	‘leur’
+4	‘nos jeunes’	+4	‘leurs cours’
+4	‘notre jeunesse’	+4	‘leurs membres’
+4	‘notre’	+4	‘qui ne sont’
+4	‘nous les’	+4	‘et ceux’
+4	‘votre gouvernement’	+4	‘leur’
+3	‘aux jeunes’	+3	‘avaient pas’
+3	‘de votre’	+3	‘ces derniers’
+3	‘J’’	+3	‘ceux’
+3	‘je vois’	+3	‘dans leur’
+3	‘je’	+3	‘des gens’
+3	‘joindre notre’	+3	‘et on’
+3	‘les étudiants n’’	+3	‘gens’
+3	‘Les étudiants ont’	+3	‘ils ont’
+3	‘M’	+3	‘ils sont’
+3	‘Michèle’	+3	‘ne paient’
+3	‘nous avons’	+3	‘on a’
+3	‘nous nous’		‘On a’
+3	‘nous pouvons’	+3	‘on est’
+3	‘nous sommes’	+3	‘on fait’
+3	‘nous voulons’	+3	‘On n’’
+3	‘Ouimet’	+3	‘ont pas’
+3	‘voulons’	+3	‘quelqu'un’
+2	‘avec eux’	+3	‘quelqu'un’
+2	‘ces étudiants’	+3	‘son gouvernement’
+2	‘ceux et celles qui’	+3	‘veulent’
+2	‘joindre’	+3	‘avaient pas’
+2	‘La Presse’	+2	‘blogue’
+2	‘Professeur’	+2	‘de leurs’

	+2	‘votre’	+2	‘est que’
			+2	‘eux-mêmes’
			+2	‘Ils’
			+2	‘les étudiants eux’
			+2	‘les gens’
			+2	‘leur’
			+2	‘leurs’
			+2	‘ne sont’
			+2	‘par les étudiants’
			+2	‘se sont’
			+2	‘tu’
			+4	‘Il y a’
Modalisation	+3	‘sentiment’	+3	‘dans le cas’
			+3	‘il devrait’
			+3	‘il est’
			+3	‘Il ne faut’
			+3	‘je pense’
			+3	‘semble que’
			+3	‘si les’
			+3	‘suis sûr’
			+3	‘serait’
			+3	‘sent’
			+2	‘allez’
			+2	‘normalement’

a) Critères dialogiques liés à l’énonciation

Les formes pronominales se voient affectées d’importantes spécificités dans les sous-corpus ETUD et GOUV. La figure 15 présente une comparaison de la fréquence relative des

pronoms nominatifs les plus spécifiques choisis par les auteurs du sous-corpus ETUD et du sous-corpus GOUV : le pronom ‘nous’ du côté ETUD et ‘ils’ dans GOUV.

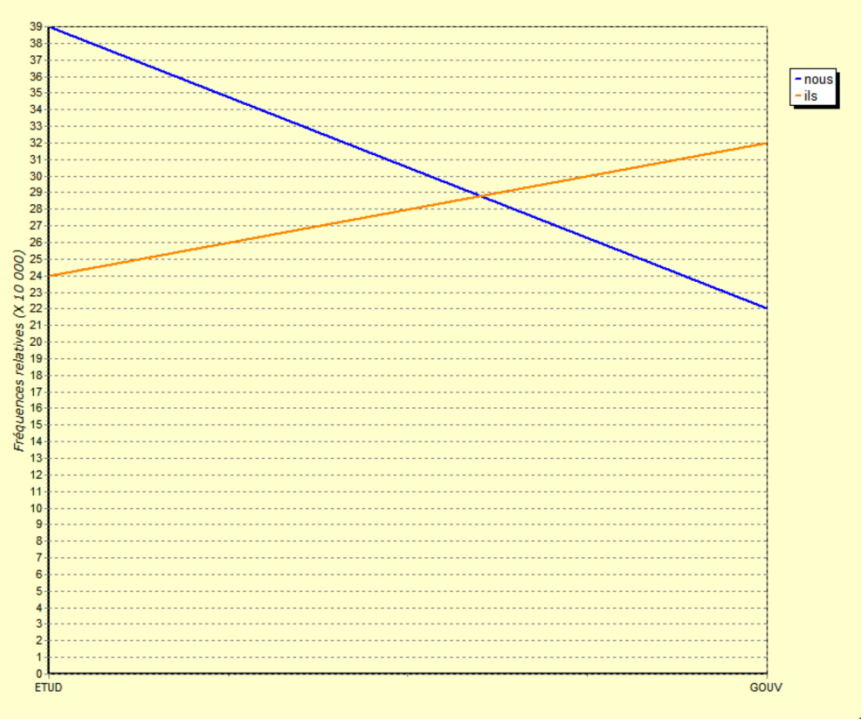


Figure 15. Comparaison de la fréquence relative des pronoms ‘nous’ et ‘ils’

La spécificité du pronom ‘nous’ dans le sous-corpus ETUD caractérise un foyer énonciatif plus égocentré comparativement au discours des auteurs défavorables à la grève, avec un haut degré de prise en charge de l’énonciation. Le ‘nous’ dans le discours des auteurs favorables à la grève fait référence à un énonciateur représenté par l’ensemble des étudiants. La présence de cet énonciateur collectif est renforcée par la spécificité d’autres critères textuels retenus pour la composante dialogique **[‘notre’, ‘notre jeunesse’, ‘nos jeunes’]** et par des indices autoréférentiels **[‘nous nous’].** Avec une moindre spécificité, d’autres critères **(‘je’, ‘J’’, ‘je vois’)** attestent que la prise en charge énonciative est une caractéristique du sous-corpus ETUD. Nous avons constaté également que l’énonciateur d’ETUD se manifeste d’une manière assertive et certaine **[‘nous sommes’, ‘nous voulons’, ‘nous avons’, ‘nous pouvons’, ‘je vois’ ‘les étudiants ont’]**. Voici quelques exemples de l’emploi du pronom ‘nous’ tirés du sous-corpus ETUD :

[news20120217LE20120217_a9_vivement] (…) Partout le ton est le même, celui de *nous* convaincre que nos élus savent où ils s'en vont (…)

[news20120218LE20120218_b5_sens](…) c'est un moyen d'urgence pour la chose la plus précieuse que *nous* possédions : le savoir.

[news20120428VE0015] (…) D'ailleurs, *nous* devons considérer ceux qui vont bénéficier d'un enseignement universitaire (…)

[news20120406QT0017] (…) De toute façon*, je vois* là une dérive de plus en plus présente dans notre société (…)

[news20120404QT0019] (…) Elle ne réglera rien, et *nous sommes* convaincus que vous le savez. (…)

[news20120616LE2012-06-16_352606](…) comme le font des milliers de citoyens, sur le genre de société que *nous voulons* construire au Québec. (…)

[news20120405MO0057] (…) les conditions de vie et de travail que *nous avons* aujourd'hui, les mesures de solidarité sociale qui nous protègent (…)

Dans le corpus GOUV, le pronom 'ils' par contre, est très souvent utilisé pour référer aux étudiants, ce qui marque la mise à distance de l'énonciateur et la teneur accusative des propos des auteurs envers les étudiants. Effectivement, ce 'ils' au pluriel s'adresse dans la grande majorité des cas aux étudiants et à leurs actions. D'autres critères textuels retenus pour la composante dialogique de GOUV confirment cette hypothèse. Par exemple, la spécificité du verbe avoir conjugué [**'ont'**), l'article au pluriel généralement associé à étudiants et grévistes **('Les')**, ainsi que d'autres critères liés à la troisième personne du pluriel [**'Ceux', 'leur', 'leurs membres', 'leurs cours', 'étudiants eux-mêmes', 'ceux qui', ' ces derniers', 'veulent', 'ils ont', 'ont pas', 'ont fait', 'ils sont', 'ils n', 'étudiants qui', 'dans leur', 'de leurs']**. La spécificité des critères comme 'on' **['on a', 'on est', 'On le', 'et on']** confirme l'existence d'un énonciateur qui ne s'identifie pas dans le discours et qui reste pourtant plus dissocié de ce qu'il énonce. Voici quelques exemples tirés du sous-corpus GOUV et de l'emploi du pronom 'ils' :

[news20120215OP120215226595245]. Réclamer une négociation équivaut de la part des étudiants à une semblable demande qui viendrait des regroupements de divers prestataires pour négocier l'aide gouvernementale qu'*ils* recevront.

[news20120327OP120327232623063] Ça t'apprendra à les avoir critiqués quand *ils* bloquaient des ponts !

[news20120327OP120327232623063] Voilà pourquoi les étudiants te conspuent même si tu les défends : *ils* ne veulent RIEN savoir de tes idées.

La mise à distance énonciative s'observe également par la façon de représenter l'interlocution. Plusieurs exemples donnés par les auteurs GOUV utilisent des formules générales pour référer aux personnes **('gens, 'des gens', 'les gens', 'quelqu'un')**.

[news20120222TB0019] (…) Il est normal pour des *gens* de défendre leurs acquis sur l'espace public (…)

[news20120323LA0040] (…) des hausses des droits de scolarité avec *des gens* qui veulent les baisser ! Le gouvernement Charest n'a pas non (…)

[news20120402OP120402234098871] (…), mais un sabot de Denver. Dès que *quelqu'un* propose une nouvelle idée pour faire bouger les choses (…)

[news20120607VE0023] (…) Il faut donc que *quelqu'un* paie. Ce *quelqu'un* s'appelle la société. Nous en sommes tous, y compris (…)

Dans GOUV, nous avons retrouvé également le 'tu' présent dans des lettres adressées à une personne spécifique, où 'tu' réfère le plus souvent à la personne attaquée, mais aussi utilisée pour donner des exemples, dans un style plutôt informel et plus proche du langage ordinaire :

[news20120327OP120327232623063] (…) les étudiants qui manifestent dans les rues : *tu* crois que plusieurs sont dans ta gang ?

[news20120515OP120515242254453] Quand la maison est en feu, *tu* ne discutes pas de la couleur du divan.

Dans le sous-corpus ETUD, nous avons repéré la spécificité du guillemet **['« »']**, indiquant la présence du discours rapporté. Les paroles citées proviennent de sources variées : dans la plupart de cas il s'agit de notions ou d'arguments utilisés par le groupe opposant et qui sont repris par les étudiants, provoquant un effet ironique. L'intention des auteurs d'ETUD est de ridiculiser et de contester le découpage référentiel fait par le gouvernement pour traiter de la question du conflit. Dans certains cas, il s'agit de déclarations données par les détracteurs du mouvement étudiant et qui sont réfutées ou contestées. Cette stratégie sert d'une part à conférer aux paroles citées un effet d'évidence, plus efficace que le discours indirect. Elle cherche également à provoquer l'étonnement chez le lecteur. Dans une moindre mesure, on retrouve les guillemets dans les paroles des tenants du mouvement qui ont été publiées dans les médias ou affichées sur des pancartes pendant les manifestations.

[news20120406LE20120406_a8_etincelle] (…), mais en plus, elle plaira dans la forme au « contribuable », lui qui en effet risquait de se retrouver par ricochet avec cette hausse démesurée (…)

[news20120414LE20120414_b5_appel] (…) approuver pareille augmentation de la part d'un gouvernement selon lequel « chacun doit faire sa part » ? (…)

Un aspect propre de l'interlocution d'ETUD est la présence de courriels. Les auteurs d'ETUD démontrent une intention visible de se rapprocher de leur lectorat, en proposant, à la fin de l'article, le courriel pour joindre le chroniqueur ou la chroniqueuse. Cela est mis en évidence par la spécificité de critères présents dans les courriels des chroniqueurs **['michele', 'la presse']** (Michèle Ouimet, chroniqueuse du journal La Presse) et le message associé à ces courriels [**'joindre'**]. Également, l'activité professionnelle de la personne qui signe l'article [**'Professeur']** indique un énonciateur plus proche, qui s'identifie nominalement et professionnellement, de manière à conférer plus de légitimité à sa parole. Ce genre d'identification est absent dans GOUV.

[news20120514LA0004] Pour joindre notre chroniqueuse : michele.ouimet@ lapresse.ca

[news20120222TB0023] Marc Frappier, Ph. D. *Professeur* titulaire

[news20120410LE20120410_a6_recherche] Vincent Larivière — *Professeur* à l'École de bibliothéconomie et de sciences de l'information de l'Université de Montréal et chercheur associé à l'Observatoire des sciences et des technologies de l'UQAM

Nous avons retenu aussi comme critères dialogiques d'ETUD les pronoms de traitements **['M.', 'Madame']**. Ces pronoms sont souvent mis vis-à-vis le nom de personnes politiques ou d'autorités. Par exemple, dans plusieurs cas, 'M.' précède le nom du premier ministre québécois Jean Charest et 'Madame', le nom de la ministre de l'Éducation, Line Beauchamp. Le choix d'utiliser le pronom de traitement plutôt que le poste occupé par la personne crée un effet ironique qui vise à détracter la légitimité des personnes concernées. D'ailleurs, le critère 'votre gouvernement', spécifique d'ETUD, exprime cette mise à distance et la non-identification au pouvoir établi.

Si, dans le corpus ETUD, les chroniqueurs proposaient à leurs lecteurs de prendre contact par courriel, dans GOUV nous remarquons une autre stratégie de proximité. Le critère 'blogue' et 'blogues' est associé à la signature de certains articles et à des adresses web (par exemple, « blogues. journaldemontreal.com »), avec un lien invitant le lecteur à visiter le blogue du chroniqueur. Les auteurs GOUV citent souvent les articles qu'ils ont écrits dans le passé dans leurs blogues et dans certains cas, ce mot apparaît aussi dans les textes.

[news20120612LA0045] (…) Au rayon des comparaisons débiles, Lise Ravary nous en raconte une bonne, dans son *blogue* du Journal de Montréal. (…)

[news20120623OP120623247664276] (…) je sais... * * * Visitez le *blogue* de Richard Martineau *blogues*. journaldemontreal.com/martineau

[news20120611OP120611245992973] (…) électrique. * * * Visitez le *blogue* de Sophie Durocher *blogues*.journaldemontreal.com/sophiedurocher

b) Critères dialogiques liés à la modalisation

Nous avons remarqué que la dialogique de GOUV était marquée par une modalisation hypothétique, en fonction de la présence de plusieurs marques renvoyant au mode conditionnel **['seraient' 'serait', 'si les']** :

> [news20120225NV0050]. D'autre part, *si les* gouvernements qui se sont succédé depuis 20 ans avaient indexé les frais de scolarité annuellement (…)

> [news20120404CY4512188] Cette exploitation du médium n'aurait toutefois pas aussi bien réussi *si les* médias n'avaient pas fait preuve d'une certaine complaisance.

> [news20120215OP120215226595245] (…) qui disent appuyer les éducatrices en garderie *seraient* sans doute moins généreux si les demandes exorbitantes de ces dernières se (…)

> [news20120323LA0040] (…) Ce ne *serait* pas une bonne idée. Négocier quoi ? Et avec qui ? (…)

Aussi, la dialogique des textes GOUV présente une certaine variation, avec des modalités exprimant la certitude **['Il y a', 'Y en a', 'qui ne sont', 'il ne faut'],** mais aussi des formulations plus proches d'un avis personnel **['je pense', 'semble que'**], ainsi que de modes à l'impératif **['il devrait']**.

Dans ETUD, nous avons aussi retenu comme critère dialogique 'sentiment', utilisé dans des contextes où l'énonciateur évoque un sentiment collectif et partagé par l'ensemble des tenants du mouvement étudiant.

> [news20120412TB0015] (…) Dans ce dossier, j'ai le *sentiment* que le gouvernement québécois n'écoute pas suffisamment (…)

> [news20120416LE20120416_b7_manifester] (…) Les communautés virtuelles très hétéroclites nourrissent la masse et stimulent le *sentiment* d'appartenance, la solidarité. (…)

[news20120416LT0027] (…) Et pourtant, je devrais avoir le *sentiment* de payer ma « juste part » (…)

2.5 Synthèse des résultats de la 1re question de recherche

Cette section a présenté le résultat de la démarche de sélection et de catégorisation des critères textuels dans les composantes sémantiques (thématique, dialectique et dialogique) Cette démarche nous a permis de créer une typologie de critères dans le cadre de notre recherche qui figure dans le tableau 28 (p.249). Les critères textuels retenus ont été sélectionnés en fonction de leur score de spécificité, de leur distribution dans les sous-corpus et de l'analyse des contextes d'apparition. Nous n'avons pas retenu comme critères textuels les données textuelles avec un score de spécificité plus petit que +2 et apparaissant dans un seul document. Par la suite nous avons effectué la sélection des critères en fonction de sa pertinence linguistique. Nous avons songé à ce que le critère soit employé avec la même signification dans la majorité des contextes repérés.

Le travail de catégorisation des critères textuels dans les composantes a été effectué de façon progressive et concomitante à l'analyse des contextes d'apparition, lesquels ont été analysés à l'aide du concordancier de l'application *TXM*. Nous avons validé progressivement les hypothèses interprétatives formulées à partir des spécificités positives les plus élevées qui ont été observées dans les deux sous-corpus. À mesure de l'avancement de l'analyse, certaines impressions sont confirmées et d'autres récurrences sémantiques sont également perçues et qualifiées en fonction des composantes. La description des contextes a été faite dans les mémos figurants sur l'Annexe E.

La description des sous-corpus ETUD et GOUV en fonction des critères textuels retenus montre les différences entre les stratégies argumentatives élaborées par chaque groupe d'auteurs sur le plan thématique, dialectique et dialogique. Cette description a été faite de manière contrastive. Ainsi, nous avons regroupé certains critères thématiques dans des domaines et taxèmes, pour expliciter les différences entre des thématiques semblables dans les deux sous-corpus. Les domaines semblables comme l'éducation, l'économie et le politique présentent des isotopies assez diverses, révélant le découpage référentiel adopté dans les sous-corpus. Les molécules sémiques identifiées à partir de l'analyse des cooccurrents spécifiques

autour de certains mots pôles nous ont permis de décrire les thématiques spécifiques et aussi de signaler les différences perçues entre des questions semblables traitées dans les deux sous-corpus.

Nous avons également montré les différences entre la structuration argumentative et temporelle des sous-corpus, en décrivant les critères retenus dans la composante dialectique et dialogique. Sur le plan dialectique, le sous-corpus ETUD présente une structuration cumulative du récit et la temporalité fait référence à des événements ponctuels, mais aussi historiques. GOUV présente la récurrence de formulations négatives et réfutative, avec une tentative plus explicite d'invalider les arguments défendus par le groupe adversaire, et une organisation temporelle durative qui explicite une impression de perpétuement des événements. Sur le plan dialogique, la différence de la représentation énonciative dans les deux sous-corpus est remarquable, avec la présence d'un énonciateur engagé, collectif et égocentré dans le sous-corpus ETUD et d'un énonciateur distancé dans GOUV, qui s'adresse particulièrement aux personnes qu'il tente d'accuser.

3. Question de recherche 2 : recherche de critères plus performants pour la classification automatique de textes d'opinion

3.1 Introduction

Dans cette section, nous présentons les résultats de la classification automatique des articles de notre corpus, en utilisant les critères textuels sélectionnés par les calculs textométriques. L'objectif de cette démarche est de connaître les types de critères les plus performants pour prédire l'opinion véhiculée par les articles. Cette étape de la recherche vise à répondre la question de recherche suivante : *quels types de critères issus d'une analyse textométrique contrastive dans la classification automatique de textes d'opinion provenant de la controverse sur la grève étudiante au Québec en 2012 sont les plus performants pour prédire la classe des articles d'opinion lors d'un processus de classification automatique ?*

Dans le chapitre 3, nous avons présenté les matrices vectorielles qui ont été constituées pour les expérimentations (p.224 à 225). Nous avons disposé deux lignes de comparaison avec tous les mots simples et les lemmes ayant au moins deux occurrences sur l'ensemble du corpus (LC1_motsSimples et LC2_Lemmes). Nous avons constitué également 9 autres matrices vectorielles en utilisant les critères sélectionnés par la démarche textométrique. Nous avons proposé de typer les matrices en fonction des distinctions suivantes : 1) segmentation (critères unitaires simples, critères unitaires lemmatisés et critères adjacents) ; 2) calcul (critères globaux et locaux) ; 3) catégorisation effectuée (critères thématiques, dialectiques et dialogiques) ; 4) tous les types de critères confondus :

1. **Type de segmentation des critères (critères unitaires simples, critères unitaires lemmatisés et critères adjacents)** : M1_CritèresUnitairesSimples ; M2_CritèresUnitairesLemmatisés ; M3_CritèresAdjacents.

2. **Type de calcul (critères globaux et locaux) :** M4_Globaux et M5_Locaux.

3. **Catégories de critères (critères thématiques, dialectiques ou dialogiques) :** M6_Thématiques, M7_Dialectiques et M7_Dialectiques.

4. **Tous les types de critères confondus :** M9_Tous.

Nous présentons dans les tableaux ci-dessous le nombre de mots et de lemmes dans les deux matrices servant de comparaison (LC1 et LC2) ainsi que le nombre de critères textuels dans chaque matrice (M1 à M9) :

Tableau 39. Lignes de comparaison

Matrice	**Type de critère**	**Nb. de critères**
LC1_motsSimples	Mots simples	10866
LC2_Lemmes	Lemmes	6016

Tableau 40. Matrices constituées pour les expérimentations avec le nombre total de critères

Matrice	**Type de critère**	**Nb. de critères**
M1_CritèresUnitairesSimples	Critères unitaires simples (mots simples)	241
M2_CritèresUnitairesLemmatisés	Critères unitaires lemmatisés (lemmes)	208
M3_CritèresAdjacents	Critères adjacents (segments répétés)	184
M4_Globaux	Critères globaux	447
M5_Locaux	Critères locaux	204
M6_Thématiques	Critères thématiques	471
M7_Dialectiques	Critères dialectiques	90
M8_Dialogiques	Critères dialogiques	90

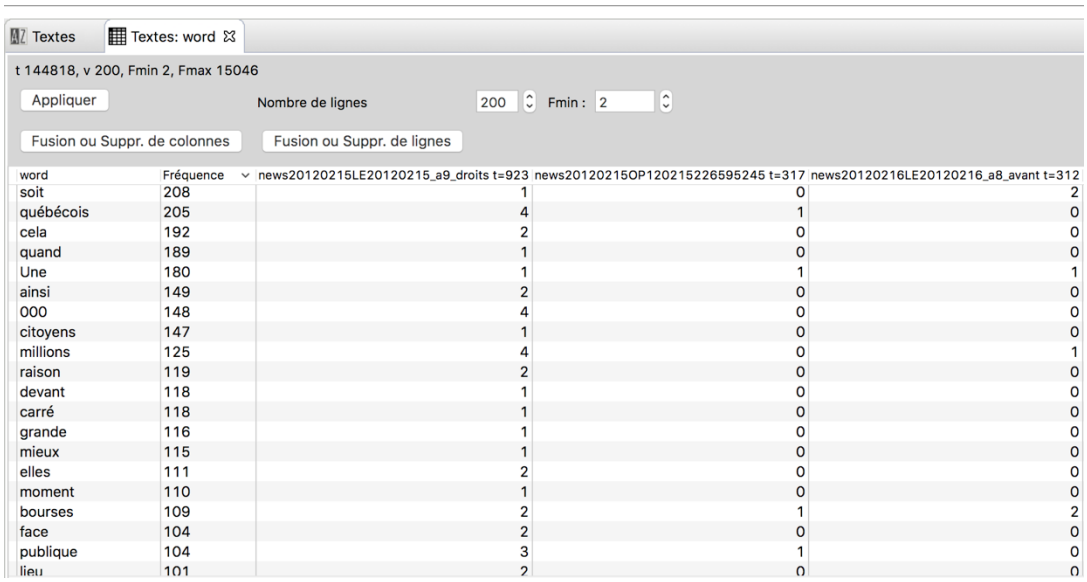
M9_Tous	Tous les critères	648

Dans notre expérimentation, nous avons effectué des tests en variant pour chaque matrice le nombre de critères en fonction du score de spécificité (de +2 à +5). Nous avons également testé la différence des résultats en comparant les performances des matrices pondérées de façon binaire ou avec les fréquences absolues.

Dans les sections suivantes, nous allons présenter les traitements effectués pour l'exportation des données de *TXM* et pour la construction des classifieurs dans *Weka* (Hall et coll., 2009). Par la suite, nous présentons les résultats de la classification pour chacune des matrices constituées.

3.2 Exportation de données et création des matrices

Les données pour l'expérimentation (les critères textuels et leurs fréquences) ont été obtenues à partir du logiciel *TXM*. La figure 16 affiche la sortie de données de la fonction « Table lexicale » de *TXM*. Cette fonction permet d'afficher tous les mots simples ou lemmes du corpus ou d'un sous-corpus. Afin de pouvoir afficher les fréquences de chaque donnée textuelle dans chacun des articles du corpus, nous avons créé une partition appelée « Textes » qui nous a permis d'obtenir l'affichage des données dans une matrice, illustrée dans la figure 16. Nous pouvons voir dans la première colonne les mots simples extraits du corpus avec leur fréquence totale (F). Dans les colonnes subséquentes figurent les fréquences de ces mots dans chaque article du corpus. Afin de générer les matrices pour l'expérimentation, nous avons effectué l'exportation de ce tableau de données au format CSV.



word	Fréquence	news20120215LE20120215_a9_droits t=923	news20120215OP120215226595245 t=317	news20120216LE20120216_a8_avant t=312
soit	208	1	0	2
québécois	205	4	1	0
cela	192	2	0	0
quand	189	1	0	0
Une	180	1	1	1
ainsi	149	2	0	0
000	148	4	0	0
citoyens	147	1	0	0
millions	125	4	0	1
raison	119	2	0	0
devant	118	1	0	0
carré	118	1	0	0
grande	116	1	0	0
mieux	115	1	0	0
elles	111	2	0	0
moment	110	1	0	0
bourses	109	2	1	2
face	104	2	0	0
publique	104	3	1	0
lieu	101	2	0	0

Figure 16. Exportation de données dans *TXM*

Les matrices contenant les critères textuels sélectionnés ont été générées par le biais de la fonction « Index » de *TXM*. Cette fonction permet de construire des requêtes combinant plusieurs termes et de filtrer les résultats en fonction du type de donnée textuelle : mot simple, lemme, partie du discours (sur le fonctionnement de la recherche dans *TXM*, voir l'Annexe C). Nous avons construit des requêtes contenant tous les critères sélectionnés en exportant dans des fichiers séparés les mots simples, les lemmes, les segments répétés et les cooccurrents. Nous avons par la suite intégré à ce fichier CSV d'autres variables relatives aux critères sélectionnés : la fréquence dans le sous-corpus (*f*), le score de spécificité (S), la composante (thématique, dialectique, dialogique), et la classe dans laquelle le critère apparaît avec un score de spécificité positif (ETUD ou GOUV).

À partir de ce fichier global contenant toutes les informations sur les critères sélectionnés, nous avons généré les autres matrices de l'expérimentation en utilisant un programme développé en langage *Python* conçu spécialement pour la réalisation de cette tâche. Ce programme nous a permis de filtrer chacune des variables associées aux critères (fréquence, score de spécificité, composante d'appartenance, etc.) et d'obtenir comme sortie les matrices au format ARFF, qui est le format de représentation vectorielle utilisé comme entrée pour les algorithmes du logiciel *Weka* (Hall et coll., 2009).

Les matrices LC1_motsSimples et LC2_Lemmes, qui sont les lignes de comparaisons pour l'expérimentation, ont été créées à partir de la fonction « Table lexicale » de TXM et exportées dans le format CSV. Le travail de conversion appliqué pour l'obtention du fichier ARFF a été le même que pour les matrices avec les critères textuels sélectionnés.

3.2.1 Gestion de critères textuels en double pour certaines matrices vectorielles

À l'exception des matrices M1_CritèresUnitairesSimples, M2_CritèresUnitairesLemmatisés et M3_CritèresAdjacents, qui ont été constituées avec un seul type de segmentation (mots, lemmes, segments répétés), toutes les autres matrices ont été constituées en combinant des critères de différentes segmentations, ce qui a eu pour effet de générer plusieurs cas de critères en double qu'il a fallu gérer. Cela a été le cas principalement pour la constitution des matrices M4_Globaux, M6_Thématiques, M7_Dialectiques, M8_Dialogiques et M9_Tous. Afin de ne pas utiliser des critères textuels répétés qui pourraient fausser les résultats de la classification, nous avons dû faire certains choix à l'égard du type de segmentation à retenir.

Deux cas spécifiques sont survenus et touchaient principalement les lemmes et les mots simples : lorsqu'un mot simple se retrouvait dans sa forme lemmatisée sur la liste de lemmes retenus avec un score de spécificité proche, seulement le lemme était gardé dans la liste finale de critères pour la matrice. Par contre, les mots simples ayant un score de spécificité plus grand que sa forme lemmatisée étaient privilégiés comme critères.

Aussi, dans la liste de cooccurrents élaborée pour chaque mot pôle, nous avons remarqué la présence de certains doublons. Il s'agissait de cooccurrents communs qui étaient associés à des mots pôles différents. Pour éliminer les cas doubles, nous avons appliqué deux mesures différentes selon les cas suivants :

1. Mêmes cooccurrents appartenant à des mots pôles différents dans le même sous-corpus : conserver le cooccurrent comme critère textuel dans la matrice.

2. Mêmes cooccurrents appartenant à des mots pôles différents dans des sous-corpus différents : si les indices sont proches, éliminer le cooccurrent de la

matrice. Si les indices sont différents, choisir le cooccurrent comme un critère du sous-corpus qui présente l'indice de cooccurrence le plus élevé.

Par rapport aux segments répétés, nous avons enlevé tous ceux dont la composition comportait des mots simples ou lemmes déjà repérés dans une matrice déterminée. Par exemple, nous avons exclu de la matrice avec les critères thématiques le segment répété 'centrales syndicales' puisque le lemme 'central' et le lemme 'syndical' figuraient déjà comme critères textuels dans la matrice.

La prochaine section explique les traitements effectués avec *Weka* (Hall et coll., 2009) pour les tests et la construction des classifieurs.

3.3 Création et évaluation de classifieurs sur *Weka*

Nous avons utilisé l'outil « Explorer » de *Weka* (Hall et coll., 2009) pour réaliser des tests sur les différentes matrices vectorielles. Le logiciel offre plusieurs algorithmes de classification supervisée, parmi lesquels le classifieur machine à vecteurs de support (SVM) (SMO dans la terminologie de *Weka*) et le classifieur bayésien naïf (NB). Avec *Weka*, nous pouvons créer des classifieurs basés sur différents algorithmes d'apprentissage et comparer leur performance. Un module d'évaluation commun pour tous les classifieurs constitués permet de faire cette comparaison, en utilisant les mesures de rappel, précision et *fscore*. Le logiciel fournit également la matrice de confusion, permettant d'obtenir les données nécessaires pour le calcul de la performance de l'algorithme par la mesure AUC, présentée dans notre méthodologie (p.229).

Nous avons effectué l'entraînement des classifieurs SVM et NB sur les bases de comparaisons (LC1_motsSimples et LC2_Lemmes) et sur chaque matrice crée (M1_CritèresUnitairesSimples, M2_CritèresUnitairesLemmatisés, M3_CritèresAdjacents, M4_Globaux, M5_Locaux, M6_Thématiques, M7_Dialectiques, M8_Dialogiques, M9_Tous) afin de comparer leurs résultats.

Pour le test des classifieurs, nous avons utilisé la méthode de la validation croisée à *k* itérations. Comme nous l'avons expliqué (p.228), cette méthode permet de varier les échantillons du corpus servant d'apprentissage et de test pour les algorithmes, en divisant le

corpus en *k* échantillons de taille égale et en utilisant un des *k* pour le test et les *k-1* restants pour l'apprentissage. Le processus est itéré pour le nombre *k* choisi. La méthode fait par la suite la moyenne des taux d'erreurs de chaque itération, afin de produire une estimation du taux d'erreur général du classifieur (Witten et Frank, 2005). Notre recherche a utilisé la validation à 5 itérations.

3.4 Résultats des tests sur les classifieurs

Nous présentons dans cette section les résultats des tests effectués. La première section présente les résultats obtenus avec les lignes de comparaison LC1_motsSimples et L2_Lemmes avec pondération binaire et par fréquences. Les sections subséquentes présentent les résultats avec les matrices constituées des critères textuels sélectionnés par notre analyse. Nous avons divisé cette dernière partie de la présentation de résultats en 4 expérimentations :

1. **Expérimentation 1 :** M1_CritèresUnitairesSimples, M2_CritèresUnitairesLemmatisés, M3_CritèresAdjacents.
2. **Expérimentation 2 :** M4_Globaux et M5_Locaux.
3. **Expérimentation 3 :** M6_Thématiques, M7_Dialectiques, M8_Dialogiques.
4. **Expérimentation 4 :** M9_Tous.

Les résultats des expérimentations de 1 à 4 sont présentés sous la forme d'un tableau qui contient les informations suivantes : un chiffre indiquant l'expérimentation, l'identifiant de la matrice (M1 à M9), le type de pondération utilisée (F pour fréquence ou B pour binaire), le nombre de critères textuels utilisés, le seuil de spécificité minimale des critères textuels présents (+2 à +5), l'algorithme utilisé pour la constitution du classifieur, (SVM ou NB) et les résultats dans trois mesures évaluatives : rappel, précision et AUC. Afin de faciliter la lecture des résultats dans le texte, nous ajoutons aux identifiants des matrices un chiffre indiquant le score de spécificité minimal des critères textuels qui les constituent (2 à 5), et aussi une lettre indiquant le type de pondération (F pour fréquence et B pour binaire). Les chiffres au début des lignes des tableaux seront également utilisés dans le texte pour faire référence à une expérimentation spécifique. Finalement, la dernière colonne du tableau affiche une flèche vers

le haut pour certains résultats, qui indique une performance supérieure de la classification par rapport à la ligne de comparaison. Nous indiquerons cette dernière dans le texte précédant le tableau pour chacune des expérimentations.

3.4.1 Lignes de comparaison : LC1_motsSimples et LC2_Lemmes

Les tests effectués avec la ligne de comparaison démontrent que le classifieur SVM performe mieux dans les tests effectués avec pondération par fréquence, tandis que NB présente les meilleurs résultats avec les matrices à pondération binaire. Pour la ligne de comparaison LC1_motsSimples, le meilleur résultat a été obtenu avec la matrice binaire, en utilisant le classifieur NB (rappel et précision : 0,79; AUC : 1,58). Dans l'expérimentation avec pondération par fréquence, LC1_motsSimples obtient de meilleurs résultats avec SVM (rappel et précision : 0,75; AUC : 1,50). Les tests avec la LC2_Lemmes pondérée par fréquence montrent que les résultats obtenus pour les deux classifieurs sont inférieurs à ceux obtenus avec LC1_motsSimples, indiquant que l'utilisation de mots simples présente en général une meilleure performance que l'utilisation de lemmes pour la prédiction. Par contre, lorsque la pondération binaire est utilisée, LC2_Lemmes a une performance proche de celle de LC1_motsSimples.

Tableau 41. Résultats : lignes de comparaison

Matrice	**Pond.**	**Critères**	**Classifieur**	**Rap.**	**Pr.**	**AUC**
LC1_motsSimples	F	10866	NB	0,63	0,63	1,26
			SVM	0,75	0,75	1,50
LC2_Lemmes	F	7044	NB	0,59	0,59	1,19
			SVM	0,74	0,74	1,48
LC1_motsSimples	B	10866	NB	0,79	0,79	1,58
			SVM	0,75	0,75	1,51
LC2_Lemmes	B	7044	NB	0,78	0,78	1,56
			SVM	0,76	0,76	1,52

3.4.2 Expérimentation 1 : type de segmentation

Dans les tests effectués avec les matrices M1_CritèresUnitairesSimples, M2_CritèresUnitairesLemmatisés et M3_CritèresAdjacents (tableau 42), la performance des classifieurs diminue au fur et à mesure que la quantité de critères diminue. Cela montre que la quantité de critères est plus importante que le score de spécificité des critères retenus dans les matrices. Les résultats des deux classifieurs sont meilleurs lorsque nous utilisons tous les critères dont le score de spécificité est plus grand ou égal à +2, soit un total de 241 critères. Le test 1.1 effectué avec la M1_CritèresUnitairesSimples (M1_2_F) sur un total de 241 critères à pondération par fréquence a été celui avec les meilleurs résultats, avec un taux de rappel de 0,85 et un taux de précision de 0,80 pour le SVM et 0,85 en rappel et 0,89 précision pour NB.

D'après les résultats, nous observons que le choix de critères unitaires lemmatisés n'a pas eu d'impact significatif sur les résultats lorsque les tests effectués avec ces derniers et avec les critères unitaires simples sont comparés. Ils présentent des résultats comparables à ceux obtenus dans les tests avec les critères adjacents.

Comparativement aux lignes de comparaison LC1_motsSimples et LC2_Lemmes, nous pouvons constater que la sélection de critères apporte un gain significatif dans la performance des classifieurs, indépendamment du type de pondération. Les résultats de 1.1 obtenus avec les 241 critères textuels sélectionnés pour la matrice M1_CritèresUnitairesSimples (M1_2_F) dépassent de 0,39 points l'AUC de LC1_motsSimples avec le classifieur NB et 0,25 point avec SVM. Ces résultats s'avèrent constants lorsque la pondération est variée (test 1.5 avec M1_2_B), même avec une quantité réduite de critères textuels (test 1.6 avec M1_3_B et test 1.7, avec M1_4_B).

Dans le tableau 42 suivant, les flèches indiquent les expérimentations qui ont obtenu une performance supérieure par rapport aux lignes de comparaison. Nous comparons, en tenant compte des pondérations, les matrices M1_CritèresUnitairesSimples à LC1_motSimples, M2_CritèresUnitairesLemmatisés à LC2_Lemmes et M3_CritèresAdjacents à LC1_motsSimples, qui se trouve elle à être la matrice avec la meilleure performance générale, considérant la pondération binaire et par fréquence.

Tableau 42. Résultats par type de segmentation

#	Matrice	Spéc.	Pond.	N.Critères	Class.	Rap.	Pr.	AUC	
1.1	M1_2_F	≥ +2	F	241	NB	0,85	0,80	1,65	↗
					SVM	0,85	0,89	1,75	↗
1.2	M1_3_F	≥ +3	F	96	NB	0,87	0,75	1,60	↗
					SVM	0,82	0,83	1,66	↗
1.3	M1_4_F	≥ +4	F	47	NB	0,73	0,78	1,54	↗
					SVM	0,77	0,77	1,54	↗
1.4	M1_5_F	≥ +5	F	23	NB	0,77	0,64	1,37	↗
					SVM	0,66	0,76	1,47	
1.5	M1_2_B	≥ +2	B	241	NB	0,88	0,88	1,75	↗
					SVM	0,80	0,80	1,59	↗
1.6	M1_3_B	≥ +3	B	96	NB	0,81	0,81	1,61	↗
					SVM	0,80	0,80	1,60	↗
1.7	M1_4_B	≥ +4	B	47	NB	0,78	0,78	1,55	
					SVM	0,78	0,78	1,55	↗
1.8	M1_5_B	≥ +5	B	23	NB	0,73	0,73	1,46	
					SVM	0,72	0,72	1,44	
1.9	M2_2_F	≥ +2	F	208	NB	0,79	0,79	1,57	↗
					SVM	0,83	0,83	1,65	↗
1.10	M2_ 3_F	≥ +3	F	94	NB	0,77	0,77	1,54	↗
					SVM	0,81	0,82	1,62	↗
1.11	M2_4_F	≥ +4	F	48	NB	0,74	0,75	1,49	↗
					SVM	0,78	0,78	1,55	↗
1.12	M2_ 5_F	≥ +5	F	28	NB	0,70	0,70	1,40	↗
					SVM	0,73	0,74	1,46	
1.13	M2_2_B	≥ +2	B	208	NB	0,85	0,85	1,70	↗
					SVM	0,79	0,79	1,57	↗
1.14	M2_3_B	≥ +3	B	94	NB	0,81	0,81	1,61	↗

					SVM	0,79	0,79	1,59 ↗
1.15	M2_4_B	≥ +4	B	48	NB	0,76	0,76	1,52
					SVM	0,78	0,78	1,56 ↗
1.16	M2_5_B	≥ +5	B	28	NB	0,73	0,73	1,46
					SVM	0,74	0,74	1,47
1.17	M3_2_F	≥ +2	F	184	NB	0,78	0,78	1,55 ↗
					SVM	0,81	0,81	1,62 ↗
1.18	M3_3_F	≥ +3	F	149	NB	0,78	0,78	1,55 ↗
					SVM	0,82	0,82	1,63 ↗
1.19	M3_4_F	≥ +4	F	46	NB	0,73	0,73	1,46 ↗
					SVM	0,73	0,74	1,45
1.20	M3_5_F	≥ +5	F	14	NB	0,66	0,67	1,30 ↗
					SVM	0,68	0,69	1,35
1.21	M3_2_B	≥ +2	B	184	NB	0,82	0,82	1,63 ↗
					SVM	0,77	0,77	1,53 ↗
1.22	M3_3_B	≥ +3	B	149	NB	0,81	0,81	1,61 ↗
					SVM	0,78	0,78	1,56 ↗
1.23	M3_4_B	≥ +4	B	46	NB	0,76	0,76	1,51
					SVM	0,73	0,73	1,45
1.24	M3_5_B	≥ +5	B	14	NB	0,69	0,69	1,38
					SVM	0,69	0,69	1,37

3.4.3 Expérimentation 2 : critères globaux et locaux

Les résultats de l'expérimentation 2 présentés dans le tableau 43 ci-dessous indiquent une très bonne performance en général des critères globaux (M4_Globaux), spécialement ceux dont des scores de spécificité sont ≥ +2 et ≥ +3 (test 2.1, M4_2_F; test 2.2, M4_3_F; test 2.5, M4_2_B; et test 2.6, M4_3_B), soit un total de 447 et 253 critères textuels sélectionnés. SVM s'avère plus performant dans tous les tests avec pondération par fréquence. Les expérimentations démontrent aussi que la pondération binaire sur les 447 critères choisis pour

la matrice M4_Globaux (test 2.5 avec M4_2_B) apporte une nette amélioration à la performance du classifieur NB (rappel : 0,91 ; précision : 0,91 ; AUC : 1,82). Ce résultat est supérieur à ceux obtenus avec les lignes de comparaison LC1_motsSimples et LC2_Lemmes à pondération binaire. Par contre, le classifieur SVM opère de façon plus satisfaisante sur la M4_Globaux (test 2.1, M4_2_F) avec 447 critères pondérés par fréquence (rappel : 0,86; précision : 0,86; AUC : 1,71).

Les résultats avec les critères locaux de la matrice M5_Locaux ont démontré une performance inférieure relativement aux lignes de comparaison, dans tous les tests effectués. Les meilleurs résultats ont été obtenus avec le classifieur SVM sur la matrice M5_4_F (test 2.11), contenant 24 critères pondérés par fréquence (rappel et précision : 0,54 ; AUC : 1,07). Le test 2.16 sur seulement 10 critères locaux à pondération binaire a donné un résultat semblable, avec un taux de rappel de 0,52 et 0,53 de précision.

Dans le tableau 43 ci-dessous, les flèches indiquent les résultats supérieurs à la ligne de comparaison L1_Mots, tenant compte des pondérations.

Tableau 43. Résultats par type de critères: globaux et locaux

#	**Matrice**	**Spéc.**	**Pond.**	**N.Critères**	**Class.**	**Rap.**	**Pr.**	**AUC**	
2.1	M4_2_F	≥ +2	F	447	NB	0,83	0,83	1,67	↗
					SVM	0,86	0,86	1,71	↗
2.2	M4_3_F	≥ +3	F	253	NB	0,83	0,83	1,67	↗
					SVM	0,85	0,85	1,69	↗
2.2	M4_4_F	≥ +4	F	96	NB	0,79	0,79	1,58	↗
					SVM	0,82	0,82	1,63	↗
2.4	M4_5_F	≥ +5	F	42	NB	0,79	0,79	1,58	↗
					SVM	0,82	0,82	1,63	↗
2.5	M4_2_B	≥ +2	B	447	NB	0,91	0,91	1,82	↗
					SVM	0,83	0,83	1,66	↗
2.6	M4_3_B	≥ +3	B	253	NB	0,86	0,86	1,71	↗
					SVM	0,85	0,85	1,70	↗

2.7	M4_4_B	≥ +4	B	96	NB	0,81	0,81	1,62 ↗
					SVM	0,81	0,81	1,62 ↗
2.8	M4_5_B	≥ +5	B	42	NB	0,78	0,78	1,56
					SVM	0,78	0,78	1,56
2.9	M5_2_F	≥ +2	F	204	NB	0,49	0,49	0,98
					SVM	0,51	0,51	1,01
2.10	M5_3_F	≥ +3	F	51	NB	0,50	0,51	1,01
					SVM	0,49	0,49	0,97
2.11	M5_4_F	≥ +4	F	24	NB	0,48	0,48	0,96
					SVM	0,54	0,54	1,07
2.12	M5_5_F	≥ +5	F	10	NB	0,49	0,48	0,97
					SVM	0,51	0,49	0,99
2.13	M5_2_B	≥ +2	B	204	NB	0,49	0,49	0,98
					SVM	0,53	0,53	1,06
2.14	M5_3_B	≥ +3	B	51	NB	0,50	0,50	1,00
					SVM	0,50	0,50	0,99
2.15	M5_4_B	≥ +4	B	24	NB	0,51	0,51	1,02
					SVM	0,53	0,53	1,05
2.16	M5_5_B	≥ +5	B	10	NB	0,53	0,53	1,05
					SVM	0,52	0,51	1,02

3.4.4 M3 : Expérimentation 3 : critères thématiques, dialectiques et dialogiques

Dans l'expérimentation 3, nous avons testé les critères catégorisés comme thématiques, dialectiques et dialogiques, afin de connaître la catégorie de critères la plus performante pour la classification. D'après les tests effectués et présentés dans le tableau 44 ci-dessous, nous avons observé que la matrice M6_Thématiques présentait les meilleurs résultats relativement aux lignes de comparaison LC1_motsSimples et L1_Lemmes, indépendamment du type de pondération utilisé. Dans le test 3.1 effectué avec pondération par fréquence, la classification avec 471 critères thématiques choisis dépasse en 0,32 point l'AUC de la matrice

LC1_motsSimples en utilisant le classifieur NB (rappel et précision : 0,79; AUC : 1,58). Le résultat est supérieur aussi avec le classifieur SVM, mais moins important, dépassant de 0,09 points l'AUC de LC1_motsSimples (rappel et précision : 0,80; AUC : 1,59). Comparée à L1_Lemmes dans la même pondération, la différence entre les résultats est encore plus significative. Dans le test 3.5 avec pondération binaire (M6_2_B), le résultat de M6_Thématiques est supérieur à celui de LC1_motsSimples binaire, avec 0,14 point de différence en AUC pour le classifieurs NB (rappel et précision : 0,86; AUC : 1,72) et 0,8 pour SVM (rappel et précision : 0,81; AUC : 1,61). Par rapport à la ligne de comparaison LC2_Lemmes, les résultats obtenus par M6_Thématiques sont également supérieurs lorsque les 471, 193 ou 84 critères thématiques sont utilisés, indépendamment de la pondération choisie.

Le test 3.9 (M7_2_F), effectué avec 90 critères textuels dialectiques de la matrice M7_Dialectiques à pondération par fréquence présente de meilleurs résultats avec NB relativement à LC1_motsSimples (rappel : 0,73; précision : 0,74 ; AUC : 1,47). La pondération binaire a un effet nuisible à la performance des classifieurs dans le cas des critères dialectiques. Les meilleurs résultats avec pondération binaire ont été obtenus dans le test 3.13 (M7_2_B) avec l'utilisation des 90 critères, autant avec NB (rappel : 0,76; précision : 0,76; AUC : 1,52) qu'avec SVM (rappel : 0,72; précision 0,72; AUC : 1,45). Mais ces résultats sont inférieurs à ceux obtenus par les lignes de comparaison LC1_motsSimples et LC2_Lemmes à pondération binaire pour les deux classifieurs.

Dans le test avec les critères dialogiques de la matrice M8_Dialogiques, la matrice qui a démontré une meilleure performance dans la classification est la M8_2_F, avec 90 critères dialogiques, qui a obtenu un AUC de 1,43 pour NB et 1,51 pour SVM (test 3.17). Par rapport à la ligne de comparaison LC1_motsSimples et LC2_Lemmes pondérée par fréquence, les résultats obtenus par ces 90 critères dialogiques s'avèrent meilleurs. Par contre, les lignes de comparaison à pondération binaire performent mieux que le meilleur résultat obtenu avec les 90 critères dialogiques sélectionnés à pondération binaire (test 3.21, M8_2_B).

Dans tous les cas, nous observons que la diminution du nombre de critères (en fonction du score de spécificité) a un impact négatif sur la performance du classifieur.

Dans le tableau 44 ci-dessous, les flèches indiquent les résultats supérieurs à la ligne de comparaison L1_Mots, en tenant compte des pondérations.

Tableau 44. Résultats par type de critères: thématiques, dialectiques et dialogiques

#	**Matrice**	**Spéc.**	**Pond.**	**N.Critères**	**Class.**	**Rap.**	**Pr.**	**AUC**	
3.1	M6_2_F	≥ +2	F	471	NB	0,79	0,79	1,58	↗
					SVM	0,80	0,80	1,59	↗
3.2	M6_3_F	≥ +3	F	193	NB	0,77	0,77	1,54	↗
					SVM	0,79	0,79	1,57	↗
3.3	M6_4_F	≥ +4	F	84	NB	0,79	0,75	1,50	↗
					SVM	0,78	0,79	1,56	↗
3.4	M6_5_F	≥ +5	F	40	NB	0,70	0,71	1,39	
					SVM	0,73	0,73	1,45	
3.5	M6_2_B	≥ +2	B	471	NB	0,86	0,86	1,72	↗
					SVM	0,81	0,81	1,61	↗
3.6	M6_3_B	≥ +3	B	193	NB	0,83	0,83	1,66	↗
					SVM	0,74	0,74	1,49	
3.7	M6_4_B	≥ +4	B	84	NB	0,79	0,79	1,57	↗
					SVM	0,75	0,75	1,51	
3.8	M6_5_B	≥ +5	B	40	NB	0,74	0,74	1,47	
					SVM	0,71	0,71	1,42	
3.9	M7_2_F	≥ +2	F	90	NB	0,73	0,74	1,47	↗
					SVM	0,73	0,73	1,45	
3.10	M7_3_F	≥ +3	F	51	NB	0,70	0,71	1,40	↗
					SVM	0,68	0,68	1,36	
3.11	M7_4_F	≥ +4	F	13	NB	0,64	0,64	1,27	↗
					SVM	0,66	0,66	1,32	
3.12	M7_5_F	≥ +5	F	4	NB	0,60	0,61	1,19	
					SVM	0,59	0,60	1,17	
3.13	M7_2_B	≥ +2	B	90	NB	0,76	0,76	1,52	

					SVM	0,72	0,72	1,45	
3.14	M7_3_B	≥ +3	B	51	NB	0,72	0,72	1,43	
					SVM	0,70	0,70	1,40	
3.15	M7_4_B	≥ +4	B	13	NB	0,62	0,62	1,24	
					SVM	0,67	0,67	1,34	
3.16	M7_5_B	≥ +5	B	4	NB	0,60	0,60	1,19	
					SVM	0,60	0,60	1,20	
3.17	M8_2_F	≥ +2	F	90	NB	0,71	0,72	1,43	↗
					SVM	0,76	0,76	1,51	↗
3.18	M8_3_F	≥ +3	F	60	NB	0,68	0,71	1,37	↗
					SVM	0,74	0,75	1,48	
3.19	M8_4_F	≥ +4	F	24	NB	0,60	0,65	1,22	
					SVM	0,66	0,66	1,30	
3.20	M8_5_F	≥ +5	F	8	NB	0,57	0,62	1,17	
					SVM	0,63	0,63	1,25	
3.21	M8_2_B	≥ +2	B	90	NB	0,74	0,74	1,48	
					SVM	0,75	0,75	1,49	
3.22	M8_3_B	≥ +3	B	60	NB	0,71	0,71	1,42	
					SVM	0,71	0,71	1,42	
3.23	M8_4_B	≥ +4	B	24	NB	0,64	0,64	1,27	
					SVM	0,66	0,66	1,31	
3.24	M8_5_B	≥ +5	B	8	NB	0,62	0,63	1,25	
					SVM	0,61	0,61	1,23	

3.4.5 Expérimentation 4 : tous les critères

Les résultats de l'expérimentation 4 avec la matrice M9_Tous permettent de constater que l'utilisation conjointe des critères a un effet global positif pour la classification relativement aux lignes de comparaison, et ce indépendamment des classifieurs testés et du type pondération utilisé. Pour les 648 critères sélectionnés et pondérés par fréquences (test 4.1, M9_2_F), le classifieur SVM atteint 0,85 de rappel et précision et un AUC de 1,69. La différence de ce résultat par rapport au test réalisé avec LC1_motsSimples est de 0,19 point et

0,21 par rapport à LC2_Lemmes. La variation du nombre de critères améliore légèrement la performance du classifieur NB, comme nous pouvons le constater dans le test 4.2 avec les 302 critères du test 4.2 (M9_3_F) : ce test atteint 0,83 et 0,84 de rappel et précision respectivement et un AUC de 1,67. Par ailleurs, le classifieur NB s'avère plus performant pour la pondération binaire, comme le démontrent les résultats obtenus dans le test M9_2_B avec 648 critères (rappel et précision : 0,89; AUC : 1,77), qui dépasse l'AUC des lignes de comparaison LC1_motsSimples et LC2_Lemmes de 0,19 et de 0,21 point respectivement. Cette amélioration est observée également dans le test 4.6 (M9_3_B) avec un nombre de 302 critères et en utilisant toujours le classifieur NB (rappel et précision : 0,85; AUC : 1,69). Nous observons aussi que l'utilisation de 120 critères et de 52 critères à pondération par fréquence dans le test 4.3 (M9_4_F) et 4.4 (M9_5_F) présente un meilleur résultat comparativement aux lignes de comparaison avec le même type de pondération.

Tableau 45. Résultat avec tous les critères confondus

#	**Matrice**	**Spéc.**	**Pond.**	**N.Critères**	**Class.**	**Rap.**	**Pr.**	**AUC**	
4.1	M9_2_F	≥ +2	F	648	NB	0,82	0,82	1,63	↗
					SVM	0,85	0,85	1,69	↗
4.2	M9_3_F	≥ +3	F	302	NB	0,83	0,84	1,67	↗
					SVM	0,81	0,81	1,61	↗
4.3	M9_4_F	≥ +4	F	120	NB	0,77	0,77	1,54	↗
					SVM	0,79	0,79	1,58	↗
4.4	M9_5_F	≥ +5	F	52	NB	0,71	0,71	1,42	↗
					SVM	0,78	0,78	1,55	↗
4.5	M9_2_B	≥ +2	B	648	NB	0,89	0,89	1,77	↗
					SVM	0,83	0,83	1,64	↗
4.6	M9_3_B	≥ +3	B	302	NB	0,85	0,85	1,69	↗
					SVM	0,84	0,84	1,67	↗
4.7	M9_4_B	≥ +4	B	120	NB	0,81	0,81	1,61	↗
					SVM	0,81	0,81	1,61	↗
4.8	M9_5_B	≥ +5	B	52	NB	0,77	0,77	1,54	

		SVM	0,75	0,75	1,50

3.5 Synthèse des résultats de la 2e question de recherche

Cette partie des résultats a révélé que la sélection d'un ensemble de critères interprétables et compatibles avec les composantes sémantiques est plus efficace pour la classification des articles de notre corpus. Nous avons remarqué que les critères globaux, lesquels englobent les critères thématiques, dialectiques et dialogiques provenant du calcul des spécificités, présentent les meilleurs résultats en considérant les deux classifieurs utilisés (NB et SVM) et les deux types de pondération (par fréquence absolue et binaire) : le test 2.1 effectué avec la M4_Globaux sur un nombre total de 447 critères textuels pondérés par fréquence (M4_2_F) présente un score AUC de 1,67 avec le classifieur NB, et un score AUC de 1,71 avec SVM. Lorsque nous utilisons la pondération binaire, ce résultat performe encore mieux, avec un score AUC de 1,82 pour NB et de 1,66 pour SVM. Le meilleur résultat obtenu parmi toutes les expérimentations effectuées a été avec la M4_Globaux en pondération binaire (M4_2_B) dans le test 2.5 et en utilisant le classifieur NB (rappel : 0,91 précision : 0,91; AUC : 1,82). Les critères sélectionner pour la matrice M4_Globaux figurent sur l'Annexe G.

Par rapport à la ligne de comparaison LC1_motsSimples pondérée par fréquence, les 447 critères de la M4_Globaux du test 2.1 (M4_2_F) présentent un résultat plus élevé en termes de rappel et précision en utilisant NB et SVM. Cette différence en AUC est de 0,40 pour NB et de 0,21 pour SVM. Les expérimentations avec pondération binaire démontrent également la pertinence des critères globaux pour la classification des articles. Avec NB, M4_2_B (test 2.5) présente 0,24 point de plus en AUC. Avec SVM, cette différence est de 0,14.

Des résultats comparables avec les critères globaux ont été aussi retrouvés dans les expérimentations 1 et 4. Dans la première, nous avons observé que les critères thématiques de la matrice M6_Thématiques sont performants pour la classification automatique, avec un AUC de 1,75 pour le test avec SVM et pondération par fréquence et un AUC de 1,75 pour le test avec NB et pondération binaire. Dans la seconde, où nous avons varié les paramètres de la

matrice M9_Tous, les résultats démontrent que la diversité de critères utilisés a un impact positif sur la classification automatique, et ce, même avec l'utilisation des cooccurrents de la matrice M5_Locaux qui n'ont pas démontré une bonne performance isolément.

4. Question de recherche 3 : évolution des critères textuels dans le corpus

4.1. Introduction

Les expérimentations réalisées dans la section précédente révèlent que les critères globaux apportent les meilleurs résultats pour la prédiction du type d'opinion des articles de notre corpus. Dans cette section, nous voulons évaluer la pertinence de la démarche de sélection de ces critères textuels dans un contexte de recommandation. Nous voulons spécifiquement savoir comment évoluent les spécificités dans notre corpus et s'il est possible d'identifier les critères textuels les plus performants avec un ensemble d'articles publiés au début de la controverse, de manière à prédire la classe des articles qui sont publiés postérieurement. Nous voulons répondre à la question de recherche suivante : *à quel moment dans la controverse apparaissent les critères textuels les plus performants pour la tâche de classification ?*

Nous avons proposé d'effectuer des analyses textométriques sur différentes périodes du corpus et d'explorer à quel moment il est possible de retrouver l'ensemble des critères qui ont démontré la meilleure performance pour la tâche de classification automatique. À titre de rappel, nous reprenons les étapes définies dans notre méthodologie pour effectuer ces analyses :

1. Sélectionner la matrice qui a démontré la meilleure performance en termes de rappel et précision pour la tâche de classification. Dresser la liste des critères textuels qui composent cette matrice.

2. Diviser le corpus de la recherche en sous-ensembles. Ces petits sous-corpus seront construits sous forme cumulative, suivant un ordre chronologique.

3. Réaliser une étude textométrique de chaque sous-corpus et observer l'apparition des critères textuels, afin de déterminer à quel moment les critères

de la matrice ayant démontré la meilleure performance surviennent dans la controverse.

4. Tester la performance des critères textuels qui surviennent au début de la controverse pour prédire la classe des articles parus postérieurement.

La matrice sélectionnée pour cette expérimentation a été M4_Globaux avec 447 critères textuels. Cette matrice a été constituée à partir des critères globaux ayant un score de spécificité ≥ +2. Les critères textuels de cette matrice se trouvent dans Annexe G.

4.2 Analyse textométrique des sous-corpus chronologiques

En suivant le découpage chronologique par mois expliqué dans la méthodologie (p.233), nous avons constitué 9 sous-corpus de façon cumulative, dont les caractéristiques sont présentées dans le tableau 46 ci-dessous. Le corpus SC_Octobre contient le nombre total d'articles du corpus de référence de notre recherche et, dans ce sens, il est équivalent au corpus de référence. Ainsi, ce corpus sera considéré comme ligne de comparaison pour les autres expérimentations menées.

Tableau 46. Sous-corpus constitués pour l'analyse de l'évolution des critères textuels les plus performants

Sous-corpus	**Nombre d'articles**	**Nombre de mots simples**	**Nombre d'occurrences**
SC_Février	41	4470	23608
SC_Mars	56	5866	34518
SC_Avril	272	14291	157198
SC_Mai	330	16849	205181
SC_Juin	465	20222	290065
SC_Juillet	470	20410	293259
SC_Août	483	20724	302096
SC_Septembre	492	20980	308194
SC_Octobre	495	21077	310202

Nous pouvons remarquer que la distribution du nombre d'articles dans les mois où se sont déroulés les événements de la grève est assez variable. Elle commence avec la publication de 41 articles au mois de février et continue au mois de mars avec seulement 15 articles de plus. Par contre, une augmentation considérable est observée à partir du mois d'avril. Comme nous pouvons voir dans le tableau 46, le sous-corpus SC_Avril comporte presque le triple des articles publiés dans les deux mois précédents. Nous pouvons également observer que le nombre de publications ralentit à partir du mois de juin.

Nous avons représenté visuellement dans les figures 17 et 18 l'évolution chronologique du vocabulaire du corpus et l'évolution du nombre d'occurrences, respectivement. La figure 17 montre que le vocabulaire triple au mois d'avril et cette même observation peut être faite à partir de la figure 18 pour le nombre d'occurrences. L'augmentation observée dans les deux graphiques coïncide avec l'accroissement du nombre de publications que nous avons exposé dans le tableau 46. Les figures montrent que l'évolution du vocabulaire aussi bien que le nombre d'occurrences plafonnent à partir du mois de juin, ce qui fait état d'une certaine stabilisation terminologique du corpus. En effet, l'apparition de nouveaux mots entre juin et septembre dans le corpus est relativement faible comparé aux quatre mois précédents.

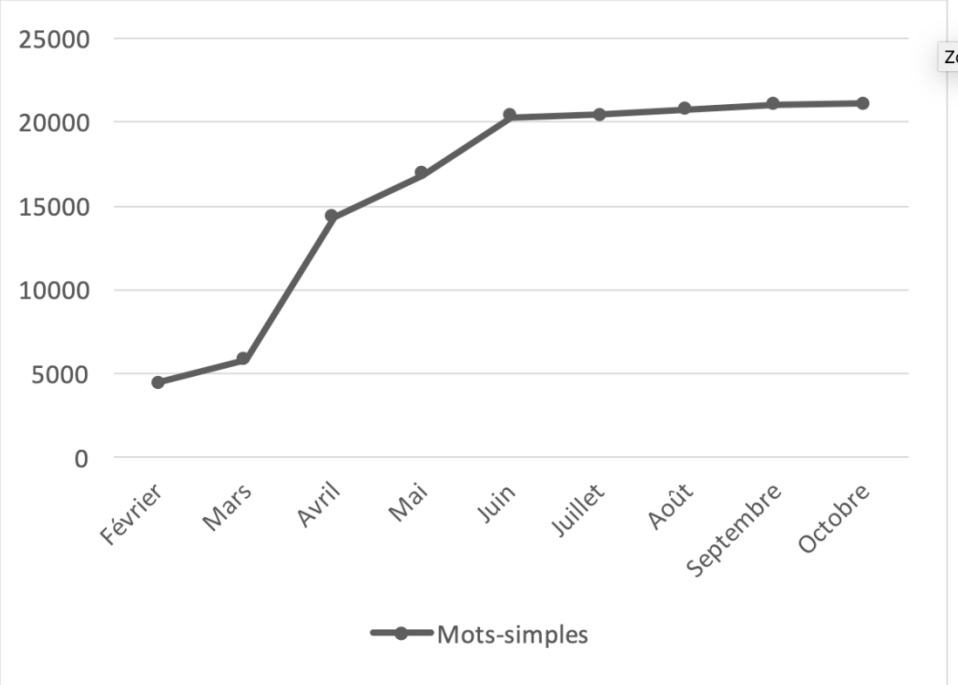


Figure 17. Évolution chronologique des mots simples dans le corpus

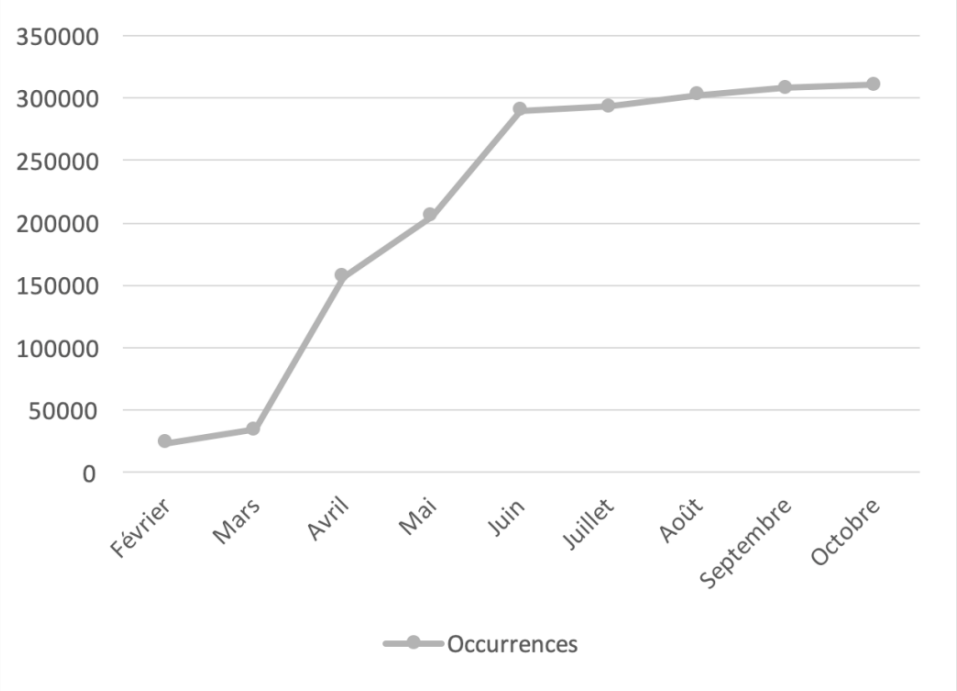


Figure 18. Évolution chronologique des occurrences dans le corpus

Pour faire l'analyse textométrique dans *TXM* et *Lexico3*, chaque sous-corpus du tableau 46 a été partitionné, de manière à former encore deux sous-corpus, ETUD et GOUV. Cette procédure doit être faite pour découvrir la liste des spécificités positives des sous-corpus chronologiques constitués. Dans le tableau 47 ci-dessous, nous présentons la répartition d'articles dans chaque classe à l'intérieur des sous-corpus chronologiques.

Tableau 47. Répartition des articles dans les classes ETUD et GOUV des sous-corpus chronologiques

Sous-corpus	**Nombre d'articles**	
	ETUD	**GOUV**
SC_Février	26	15
SC_Mars	29	27
SC_Avril	152	120
SC_Mai	173	157
SC_Juin	250	215
SC_Juillet	250	220
SC_Août	258	233
SC_Septembre	258	234
SC_Octobre	258	237

Nous avons par la suite calculé les spécificités des sous-corpus chronologiques pour chaque type de segmentation qui a été pris en compte dans la matrice M4_Globaux : mots simples, lemmes et segments répétés. Nous avons comparé les critères textuels globaux de la matrice M4_Globaux avec le résultat du calcul sur les sous-corpus chronologiques, en considérant les types de segmentation séparément. Pour cette comparaison, nous avons considéré dans les sous-corpus chronologiques seulement les spécificités égales ou supérieures à +2. Nous avons fait le décompte du nombre de critères textuels de la matrice M4_Globaux qui figurait dans la liste de spécificités obtenue comme résultat. Cette opération nous a permis de calculer le pourcentage de critères globaux avec un score de spécificité d'au moins +2 qui pourraient être retrouvés dans chaque sous-corpus chronologique constitué.

La progression de l'apparition de critères textuels dans le temps est présentée dans le graphique en courbes à la figure 19 ci-dessous. À la figure 20, nous avons représenté la progression d'apparition des critères globaux dans les sous-corpus chronologiques en fonction du nombre d'occurrences. Dans les deux graphiques, l'axe des ordonnées représente le pourcentage de critères globaux rencontrés dans les sous-corpus chronologiques. Chaque série du graphique représente les types de segmentation considérés (mot simple, lemme et segment répété).

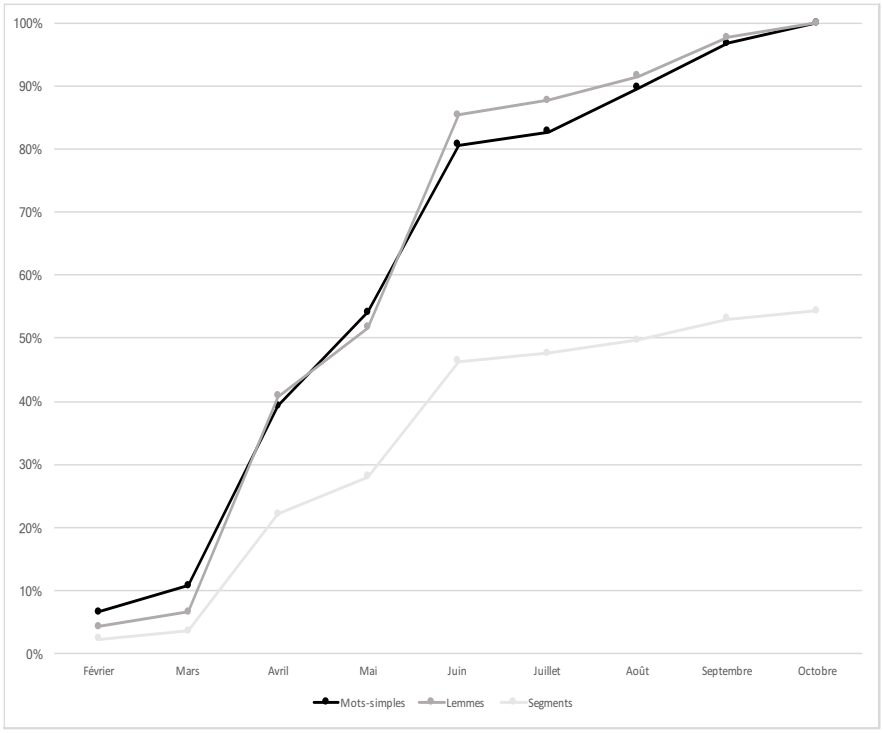


Figure 19. Progression de l'apparition des critères globaux dans le temps

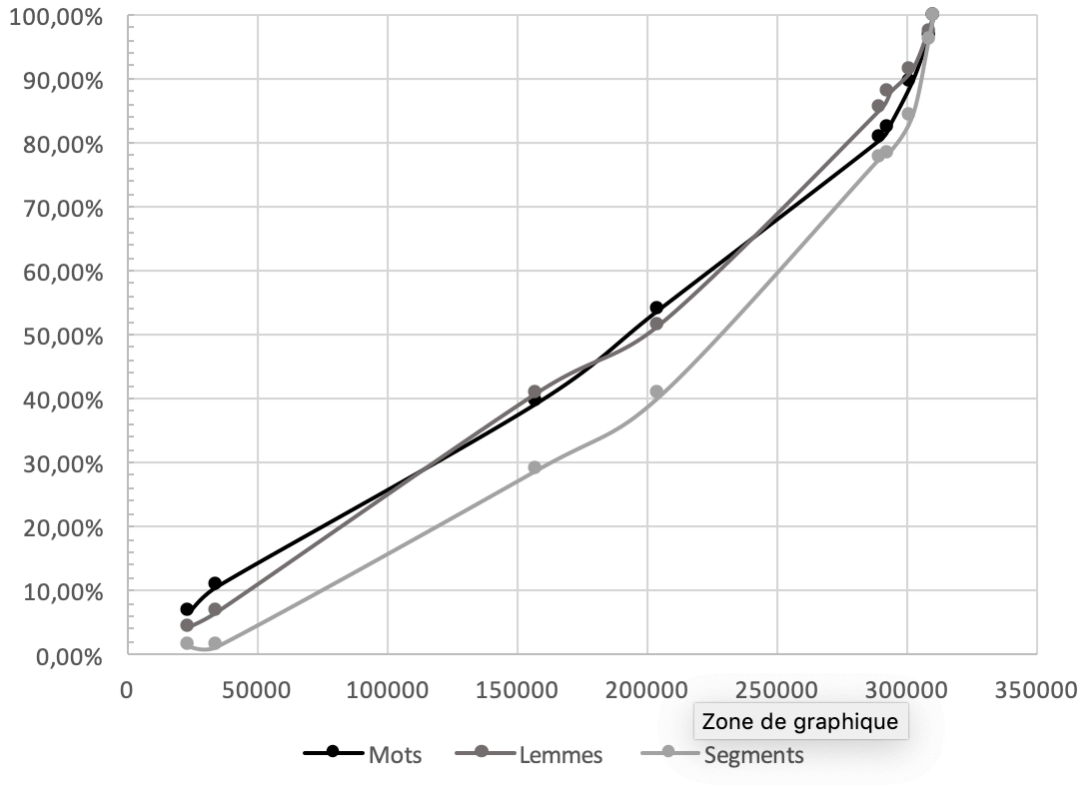


Figure 20. Progression de l'apparition des critères globaux
en fonction du nombre d'occurrences

Nous pouvons remarquer à la figure 19 qu'au début de la controverse, le nombre de critères globaux de la matrice M4_Globaux apparaît en très petit nombre. L'augmentation observée sur le nombre des critères globaux rencontrés coïncide avec une couverture plus intense de la grève dans les médias, à partir du mois d'avril. Entre les mois de mai et juin, le nombre d'articles (465) atteint presque la totalité du nombre d'articles total du corpus de référence (495), avec 93% du nombre d'occurrences total (290065 occurrences dans un total de 310202). À la figure 20, nous pouvons remarquer qu'environ 50% des critères globaux les plus performants apparaissent lorsqu'il existe de 100.000 à 150.000 occurrences, environ la moitié d'occurrences de notre corpus de référence. Cela correspond aussi aux sous-corpus SC_Avril et SC_Mai. SC_Avril contient 272 articles, un peu plus de la moitié des articles du corpus de référence, et SC_Mai contient 330 articles, soit 2/3 du nombre total d'articles du corpus de référence (495).

Nous avons découvert que certains critères globaux qui ont une grande spécificité dans la M4_Globaux font déjà leur apparition de manière anticipée dans les sous-corpus SC_Février. C'est le cas de 'éducation', 'leur', 'université', '@card', entre autres. Nous remarquons que le score de spécificité de ces critères augmente dans les résultats obtenus pour les sous-corpus chronologiques subséquents. Également, à mesure que nous avançons dans les résultats des sous-corpus dans le temps, certaines données textuelles qui ressortent comme spécifiques au début de la controverse (par exemple, dans SC_Février et SC_Mars), se révèlent moins spécifiques plus tard. D'autres ne sont plus repérés comme spécifiques lorsque les mois avancent. Nous avons aussi remarqué que certains critères de la M4_Globaux apparaissent comme spécifiques au début de la controverse, mais disparaissent quelques mois plus tard, même si à la fin ils se trouvent comme spécifiques.

En regardant les résultats du calcul des spécificités du sous-corpus SC_Avril, nous avons remarqué qu'une grande partie des données textuelles plus spécifiques sont parmi les critères globaux de matrice M4_Globaux_2. Par exemple, parmi les 161 lemmes les plus spécifiques de SC_Avril, 90 sont aussi des critères globaux de la M4_Globaux, ce qui représente 55% du total des lemmes spécifiques retrouvés.

Afin de pousser plus loin notre analyse et vérifier la performance des critères globaux retrouvés au début de la controverse pour la prédiction d'articles d'opinion publiés postérieurement, nous avons réalisé certains tests de classification automatique. Les résultats sont présentés dans la section suivante.

4.3 Tests sur la quantité des critères globaux retrouvés dans les sous-corpus chronologiques

L'objectif de ce test est de savoir si l'utilisation du nombre de critères globaux rencontrés dans un sous-corpus chronologique déterminé est suffisante pour exécuter la classification des articles d'opinion publiés postérieurement avec une efficacité comparable à celle obtenue par la totalité de critères globaux qui ont été utilisés dans les tests avec la matrice M4_Globaux. Nous avons ainsi réalisé 7 tests, en variant le nombre de critères globaux dans les matrices. Le tableau 48 présente les matrices constituées et le nombre de critères textuels utilisés dans chacune. Chaque matrice a été constituée avec le nombre des

critères globaux de la M4_Globaux_2 qui ont été retrouvés dans l'analyse des sous-corpus chronologiques. Comme nous avons mentionné, la matrice M_Octobre constitue la ligne de comparaison pour l'expérimentation, puisqu'elle contienne la totalité de critères textuels qui sont présents dans M4_Globaux.

Tableau 48. Corpus pour le test de classification
en variant le nombre de critères globaux rencontrés

Matrices	**Nb.de critères**	**Nombre de documents**	
		Corpus d'entraînement	**Corpus de test**
M_Avril	164	330	165
M_Mai	218	330	165
M_Juin	361	330	165
M_Juillet	367	330	165
M_Août	397	330	165
M_Septembre	433	330	165
M_Octobre (ligne de comparaison)	447	330	165

Pour ce test, nous n'avons pas cherché à constituer des matrices avec les critères globaux retrouvés dans les mois de février et mars (SC_Février et SC_Mars), dû à la faible quantité d'articles présents dans ces deux sous-corpus chronologiques. Nous avons donc testé seulement les critères globaux élaborés à partir du mois d'avril. Nous n'avons pas non plus procédé à l'analyse de toutes les spécificités rencontrées dans les sous-corpus chronologiques, afin de sélectionner d'autres critères textuels qui pourraient être considérés. Par exemple, les critères globaux de la M4_Globaux qui sont ressortis du calcul des spécificités de SC_Avril correspond à environ 50% des spécificités totales (≥ +2) rencontrées dans ce sous-corpus. D'autres critères textuels auraient pu être sélectionnés, mais nous n'avons voulu utiliser que ceux qui étaient présents dans M4_Globaux seulement. Dans la figure 21, nous avons représenté la proportion de critères globaux de type mot simples qui sont présents dans chaque sous-corpus chronologique par rapport au nombre de mots simples de spécificité ≥ +2 qui pourraient être considérés comme critères.

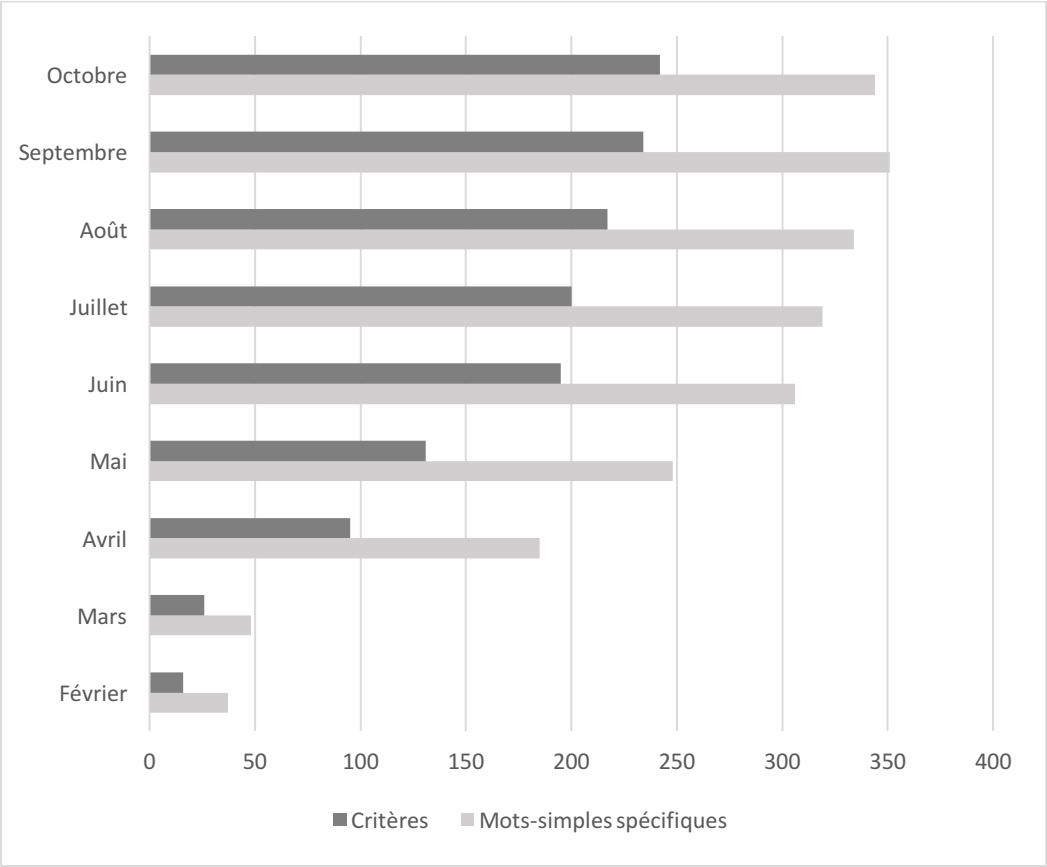


Figure 21. Proportion de critères globaux du type mot simples rencontrés dans les sous-corpus chronologiques par rapport au nombre de mots simples spécifiques

Le deuxième objectif de ce test est de savoir si l'entraînement du classifieur fait à partir des articles qui sont apparus en premier dans la controverse, avec les critères globaux de la M4_Globaux présents à chaque période, est efficace pour prédire les articles qui ont été publiés postérieurement. Ainsi, nous avons cherché à ce que les corpus d'entraînement soient constitués en fonction de la date de publication de l'article. Nous avons respecté la répartition préconisée par la méthodologie de fouille de textes (2/3 pour le corpus d'entraînement et 1/3 pour le corpus de test), puisque le nombre d'articles publiés à partir de mai correspond au 2/3 du corpus de référence, soit un nombre de 330 articles. La seule exception est la matrice M_Avril, dans laquelle le nombre d'articles publiés est de 272. Pour avoir des bases de comparaison raisonnables, nous avons décidé de maintenir la même répartition pour cette matrice, même si le nombre d'articles publiés jusqu'à avril est moindre.

Le tableau 49 ci-dessous présente les résultats de la classification. Nous avons varié la pondération (fréquence et binaire), ainsi que le classifieur utilisé (NB et SVM). Les résultats sont exprimés avec les mesures de rappel, précision et AUC.

Tableau 49. Résultats de la classification chronologique

Matrice	**Pondération**	**N.Critères**	**Classifieur**	**Rap.**	**Pr.**	**AUC**
M_Avril	Fréquences	164	NB	0,74	0,74	1,48
			SVM	0,81	0,81	1,62
M_Mai	Fréquences	218	NB	0,78	0,80	1,56
			SVM	0,85	0,85	1,57
M_Juin	Fréquences	361	NB	0,87	0,88	1,75
			SVM	0,90	0,90	1,68
M_Juillet	Fréquences	367	NB	0,87	0,88	1,75
			SVM	0,89	0,89	1,67
M_Août	Fréquences	397	NB	0,86	0,87	1,72
			SVM	0,89	0,89	1,67
M_Septembre	Fréquences	443	NB	0,86	0,86	1,71
			SVM	0,92	0,92	1,69
M_Octobre	Fréquences	447	NB	0,86	0,86	1,71
			SVM	0,90	0,90	1,68
M_Avril	Binaire	164	NB	0,78	0,78	1,56
			SVM	0,82	0,82	1,63
M_Mai	Binaire	218	NB	0,85	0,85	1,69
			SVM	0,78	0,78	1,57
M_Juin	Binaire	361	NB	0,90	0,90	1,80
			SVM	0,82	0,82	1,65
M_Juillet	Binaire	367	NB	0,89	0,89	1,78
			SVM	0,82	0,82	1,64
M_Août	Binaire	397	NB	0,89	0,89	1,78
			SVM	0,81	0,81	1,62
M_Septembre	Binaire	443	NB	0,92	0,92	1,83

			SVM	0,80	0,80	1,60
M_Octobre	Binaire	447	NB	0,90	0,90	1,81
			SVM	0,81	0,81	1,61

Dans les expérimentations avec pondération par fréquence et binaire, nous observons que les résultats s'améliorent à mesure qu'augmente le nombre de critères globaux pris en compte dans les matrices.

Quant aux résultats obtenus par l'expérimentation menée avec la matrice M_Octobre pondérée par fréquence, nous observons des résultats comparables avec M_Juin. Elle performe un peu mieux que M_Octobre avec le classifieur NB (Rappel : 0,87; Précision : 0,88; AUC : 1,71) et présente un résultat pareil pour SVM (Rappel et précision : 0,90; AUC : 1,68). Les matrices M_Juillet, M_Août et M_Septembre présentent également un meilleur résultat avec le classifieur NB dans la pondération par fréquence, résultats qui sont aussi proches de ceux de M_Octobre. Pour ce qui est des expérimentations menées avec les matrices à pondération binaire, nous observons aussi qu'à partir de la matrice M_Juin, les résultats sont plus proches de ceux de M_Octobre. Avec NB, M_Juin obtient 0,90 en rappel et précision et un AUC de 1,80, résultats assez proches de ceux obtenus par M_Octobre (Rappel et Précision : 0,90; AUC : 1,81). Les matrices M_Juillet, M_Août et M_Septembre présentent la même tendance, avec des scores AUC de 1,78, 1,78 et 1,83 respectivement. Les matrices M_Avril et M_Mai présentent des résultats inférieurs au regard de la ligne de comparaison. Dans la pondération par fréquences, le meilleur résultat est obtenu par M_Mai avec SVM, mais le taux de rappel et de précision se trouve à 5 points au-dessous de M_Octobre.

4.4 Synthèse des résultats de la 3^{e} question de recherche

Les résultats démontrent que nous ne pouvons pas retrouver une quantité significative de critères globaux dans les sous-corpus chronologiques de manière anticipée. La survenance des critères globaux qui sont dans la M4_Globaux est progressive et influencée par la quantité d'articles considérés et le nombre d'occurrences total accumulé dans les sous-corpus chronologiques. Il y a donc une corrélation entre les variables. D'après les tests de classification effectués, nous avons remarqué que la performance des classifieurs est meilleure

à partir de la matrice M_Juin, laquelle comporte 361 critères globaux des 447 critères compris dans la ligne de comparaison (M_Octobre).

5. Conclusion chapitre

Ce chapitre présente les réponses aux questions posées dans notre recherche. Dans un premier temps, nous avons exposé les résultats de la démarche d'élaboration et de sélection de critères textuels. Nous avons détaillé les décisions prises en fonction de l'observation du résultat obtenu par les calculs textométriques, principalement en ce qui concerne les éliminations effectuées. Par la suite, nous avons explicité la lecture interprétative des critères textuels retenus en fonction des concepts de la sémantique interprétative. Nous avons aussi montré comment la catégorisation des critères dans les composantes sémantiques a été effectuée et nous avons décrit les sous-corpus ETUD et GOUV de manière contrastive, considérant les différentes stratégies argumentatives élaborées par les auteurs de chaque sous-corpus sur les plans thématique, dialectique et dialogique. Cette description s'est appuyée sur la récurrence sémique observée sur l'ensemble des critères textuels spécifiques retenus.

Dans un deuxième temps, nous avons présenté les résultats de la classification automatique des articles de notre corpus. Les étapes de l'exportation et de la constitution de lignes de comparaison ainsi que des matrices pour les expérimentations ont été décrites. Les résultats statistiques ont été présentés sous forme de tableaux et organisés en fonction du type de critères textuels. Nous avons décrit les résultats découverts, en indiquant les matrices et les paramètres qui ont présenté les meilleurs résultats en termes de rappel, de précision et d'AUC. Nous avons démontré que les critères globaux, qui comprennent les critères thématiques, dialectiques et dialogiques élaborés à partir du calcul des spécificités, performent mieux que toutes les autres expérimentations, lorsque nous considérons l'ensemble des résultats obtenus sur les classifieurs utilisés (NB et SVM) et sur le type de pondération (par fréquence ou binaire). La matrice qui a obtenu le meilleur résultat dans la classification a été la matrice M4_Globaux, avec 447 critères globaux avec pondération binaire et en utilisant le classifieur NB.

Finalement, nous avons exploré la pertinence de la démarche de sélection de critères textuels dans un contexte de recommandation en considérant l'évolution chronologique du vocabulaire de notre corpus de référence et en évaluant à partir de quel moment il était

possible de retrouver les critères textuels qui ont démontré la meilleure performance pour la classification des articles d'opinion, c'est-à-dire les critères de la matrice M4_Globaux avec un score de spécificité ≥ +2. Dans cette démarche, nous avons exposé la corrélation entre la survenance des critères globaux en fonction des mois dans lesquels les articles de la controverse ont été publiés, et aussi en fonction de l'effectif d'occurrences accumulé au cours de la controverse. Nous avons observé que les critères globaux de la M4_Globaux qui ont une spécificité plus élevée apparaissent de manière anticipée dans notre corpus de référence.

Par la suite, nous avons effectué différents tests de classification, en considérant seulement la quantité de critères globaux qui pourraient être retrouvés lors de l'analyse textométrique des sous-corpus chronologiques. Les corpus d'entraînement ont été constitués en considérant un critère chronologique, avec des articles publiés au début de la controverse, afin de prédire les articles des mois subséquents. Nous avons découvert que pour atteindre des résultats comparables à la classification effectuée avec la totalité de critères la matrice M4_Globaux (ligne de comparaison M_Octobre), il a été nécessaire d'utiliser 361 critères globaux retrouvés à partir de l'analyse textométrique du sous-corpus SC_Juin. La matrice avec moins de critères dont les résultats s'approchent plus à ceux obtenus par la ligne de comparaison est la M_Mai, avec 218 critères textuels.

Au prochain chapitre, nous discutons les principaux résultats obtenus dans le cours de notre étude.

Chapitre 5. Discussion

1. Introduction

Ce chapitre a pour objectif de discuter les résultats saillants obtenus par la recherche. Dans les prochaines sections, les résultats pour chacune des 3 questions de recherche posées sont analysées. La première section porte sur la démarche d'élaboration de critères textométriques pour la tâche de fouille d'opinions, et analyse cette dernière du point de vue de sa pertinence, tout en discutant les concepts théoriques traités dans la recherche. La deuxième section examine les expérimentations menées avec les critères textuels sélectionnés, en présentant les facteurs qui ont contribué au succès de la classification des articles. La troisième section discute les résultats obtenus pour l'analyse chronologique du corpus et aborde les questions concernant l'application de la méthodologie de recherche proposée dans le contexte des SRAP.

2. Élaboration de critères textuels pour la tâche de classification d'articles d'opinion

Dans la méthode de fouille d'opinions systématisée dans le cadre de cette recherche, nous avons proposé une étape de constitution de deux sous-corpus, ETUD et GOUV, lesquels comportent des articles d'opinion classés en fonction de la divergence d'opinions qu'ils véhiculent. À la suite de cette étape, nous avons proposé l'application de calculs textométriques. En comparant et en contrastant les distributions des données textuelles dans les sous-corpus, ces calculs ont permis d'identifier celles qui se retrouvaient surreprésentées ou sous-représentées dans chacun des sous-corpus. Une variété de données, incluant des mots lexicaux, des mots grammaticaux, des syntagmes, des signes de ponctuation et des chiffres, ont été par la suite interprétées et sélectionnées comme critères textuels, en fonction de leur spécificité dans un des sous-corpus et de leur pertinence linguistique. L'interprétation des critères ainsi que la description effectuée a mis en évidence la façon dont les camps opposés dans la discussion ont essayé d'élaborer leur positionnement sur les plans thématique, dialectique et dialogique.

La description des critères textuels à partir de l'observation de régularités sémantiques s'avère de grande valeur pour mettre au clair les différences observées d'un sous-corpus à l'autre. Quelques variantes entre les sous-corpus qui étaient connues au départ et qui ont été décrites dans le chapitre 3 (p.184) ont été confirmées par les résultats des calculs textométriques. D'autres différences plus subtiles ont été révélées par l'analyse textométrique. Un des apports les plus importants de la démarche de sélection de critères textuels est de révéler des observables linguistiques étonnantes du point de vue des attentes qui pourraient être formulées par une personne possédant une connaissance préalable de la discussion dans les médias. Par exemple, le critère 'artistes' apparaît comme le mot le plus spécifique de GOUV, révélant une stratégie de dépréciation des partisans des étudiants, alors que personne n'aurait pensé à faire de ce mot un critère textuel caractéristique de la thématique de ce sous-corpus.

Après la réalisation des calculs textométriques sur les sous-corpus ETUD et GOUV, la principale tâche consistait à regarder le contexte d'apparition des données textuelles dont le score de spécificité était égal ou supérieur au seuil de +2. L'élimination de données textuelles non pertinentes a été la première étape de ce travail de sélection. D'abord, nous avons éliminé les données en fonction de leur distribution : toutes celles qui étaient présentes dans un seul document, indifféremment de leurs fréquences dans le corpus ou dans les sous-corpus, ont été éliminées d'emblée. Par la suite, l'analyse des contextes le concordancier de *TXM*, nous a permis d'identifier facilement les données textuelles qui étaient employées avec des fonctions différentes ou des significations différentes, ainsi que celles qui démontraient une pertinence linguistique et qui pourraient décrire, de manière corrélative avec d'autres données textuelles spécifiques, la thématique, la dialectique et la dialogique des articles regroupés dans les sous-corpus. La majorité des données textuelles non éliminées ont été retenues comme critères textuels, sauf quelques exceptions qui ne faisaient pas état d'une régularité sémantique observée.

La rédaction de mémos décrivant les contextes des critères textuels au fur et à mesure de leur analyse, en partant des plus spécifiques jusqu'aux moins spécifiques, nous a permis de mettre en relation les observations sur le contexte d'emploi des critères retenus et d'identifier les récurrences sémiques. Cette procédure s'est avérée la plus efficace, dans la mesure où nous avons élaboré les hypothèses interprétatives à partir des spécificités positives les plus importantes dans les sous-corpus, et ensuite validé ces hypothèses en fonction de la récurrence des données textuelles dont les scores de spécificité étaient moins importants.

La description des composantes sémantiques dans le cadre de notre travail, appuyé par les concepts de la sémantique interprétative, a cherché à rendre compte des régularités présentes dans chaque sous-corpus ainsi que des différences significatives sur la façon dont les auteurs d'ETUD et GOUV ont élaboré leur opinion au sujet de la grève étudiante. La principale conclusion qui peut être tirée de la description des composantes renvoie à la diversité d'indices qui permettent de caractériser les positions défendues par les camps dans la dispute au sujet de la grève. Il est intéressant de voir par exemple que les critères textuels identifiés dans les sous-corpus ne sont pas nécessairement liés à des modalités d'expression qui sont communément attribuées à l'évaluation (Charaudeau, 1992) comme les adjectifs. La

faible quantité d'adjectifs parmi les critères textuels les plus spécifiques invalide le regard trop restrictif de certaines recherches en fouille d'opinions qui cherchent à caractériser l'opinion (ou les différences entre les opinions véhiculées par les textes) en fonction d'un vocabulaire évaluatif. L'étude scientifique du corpus par les calculs textométriques a l'avantage de révéler les choix individuels opérés par les locuteurs pour défendre leurs arguments et de découvrir que les stratégies argumentatives révélatrices du caractère polémique de la discussion ne sont pas uniquement décrites par un ensemble « d'indices de polémicité » cernables dans la langue, mais par des choix linguistiques relevant d'un cadre culturel partagé. Les critères 'beurre' et 'bacon' par exemple, employés dans les textes des auteurs GOUV pour représenter les comportements des étudiants dans la grève, sont à la fois étonnants d'un point de vue de la description de ce sous-corpus et emblématique de l'incidence de l'idiolecte (Rastier, 1989) dans sa structuration argumentative.

L'observation et la description des régularités sémantiques en fonction des composantes ont également permis de considérer comme critères textuels des observables linguistiques qui sont généralement qualifiées d'insignifiantes du point de vue de la représentation informatique de textes pour les tâches de classification automatique. Cela est notamment le cas des signes de ponctuation qui ont été qualifiés et sélectionnés comme critères textuels dans le cadre de notre démarche ou encore de certaines conjonctions. Ce genre de variables est généralement rejeté par des approches statistiques de sélection de critères textuels pour les applications de fouille de textes et de fouille d'opinions. Ces dernières préconisent la réduction de la dimensionnalité de la représentation vectorielle par le biais de l'utilisation de ressources comme les anti-dictionnaires destinés à éliminer ces variables, qui sont considérées non pertinentes du point de vue sémantique. Dans notre démarche, nous avons qualifié sémantiquement les critères textuels à retenir, en fonction des récurrences sémantiques observées. Ainsi, la pertinence de la conjonction 'et' comme critère textuel d'ETUD se justifiait par la présence de la structuration cumulative révélée par d'autres critères, même si dans certains contextes repérés dans ce sous-corpus, la conjonction 'et' n'exprimait pas la notion d'addition : elle était aussi employée dans un contexte restrictif (comme dans « L'éducation est un droit et non un privilège ») ou encore se trouvait associée à une interrogation (« Et alors? »). L'observation de la récurrence de critères sémantiques

exprimant la notion de cumulation nous a aussi motivés à retenir comme critère la virgule (‘,’), qui était spécifique dans le sous-corpus ETUD. En étant un signe servant à isoler et à séparer les éléments de la phrase, sa récurrence corroborait les effets d’accumulation et de surenchère observés dans la structuration argumentative de ce sous-corpus.

Les regroupements effectués à partir de la récurrence sémique identifiée par les critères textuels ont des portées variables. Le nombre de critères qui permettent de décrire l’organisation cumulative caractéristique de l’argumentation du sous-corpus ETUD était plus important que le nombre de critères qui décrivaient la présence d’interrogations rhétoriques. Mais nous avons explicitement gardé tous les critères textuels, même si certains phénomènes décrits étaient moins saillants que d’autres. La description des critères textuels en rapport avec les composantes révèle que les regroupements effectués sont cohérents, dans la mesure où ils permettent aussi de contraster les stratégies des camps en discussion. Les exemples tirés des textes éclairent sur les opérations interprétatives réalisées, démontrant la plus-value de la démarche de qualification interprétative des critères.

Dans les sections suivantes, nous allons discuter les résultats obtenus par rapport à chacune des composantes sémantiques.

2.1 Sélection et catégorisation de critères textuels thématiques

Sur le plan thématique, la description des thèmes génériques et des thèmes spécifiques a mis en évidence le traitement donné aux diverses questions relatives à la hausse des frais de scolarité et au mouvement des étudiants dans les rues. La catégorisation de critères textuels thématiques provenant du calcul des spécificités nous a permis d’abord d’identifier les domaines communs dans les sous-corpus. Le choix de regrouper sous le même domaine les thèmes semblables des deux sous-corpus ainsi que le choix de les distinguer en fonction des isotopies retrouvées nous a permis de comprendre plus en détail les différences relatives au découpage référentiel opéré par chaque groupe. Nous pouvons percevoir qu’en effet, la discussion sur la grève étudiante dans les journaux québécois s’insérait dans un contexte discursif controversé et polémique (Garand, 1998) : les stratégies argumentatives employées cherchaient non seulement la défense de points de vue sur l’enjeu discuté, mais aussi à mettre

au point les questions qui devraient être discutées, c'est-à-dire imposer les thèmes susceptibles de faire adhérer des partisans.

Par rapport aux critères textuels thématiques provenant du calcul de cooccurrence, nous avons remarqué que la méthode permettait également d'observer les différences thématiques importantes entre les défenseurs et les détracteurs du mouvement étudiant. La démarche consistant à calculer les cooccurrences autour de mots pôles, qui sont à la fois fréquents et distribués dans l'ensemble du corpus, a eu l'avantage de révéler de nouveaux critères qui n'ont pas été repérés par le calcul des spécificités et qui étaient reliés à des mots représentatifs du thème du corpus de référence. Cependant, certains thèmes qui ont été décrits par cette analyse n'étaient pas forcément nouveaux, et les critères locaux repérés renforçaient parfois les thèmes génériques déjà décrits par les critères globaux. Par exemple, le thème à propos des accusations de corruption contre le gouvernement, associé au mot pôle 'gouvernement' dans le sous-corpus ETUD, ou encore la tentative d'associer les étudiants à la notion d'immaturité, révélée par certains cooccurrents spécifiques autour du mot pôle 'étudiant', ont été aussi décrits dans l'analyse des critères globaux regroupés dans les isotopies /désintégration/ et /immaturité/ respectivement.

Cette découverte nous amène à nous interroger sur la pertinence de certaines distinctions théoriques que nous avons choisi de retenir, en rapport avec la définition de thèmes génériques et thèmes spécifiques. Elle concerne principalement l'homologation d'isotopies génériques aux thèmes génériques d'une part, et l'homologation entre les molécules sémiques et les thèmes spécifiques d'autre part. Sur le plan des critères globaux, nous avons proposé de décrire toutes les récurrences de sèmes génériques comme des isotopies et des thèmes génériques ; nous avons aussi proposé d'homologuer les cooccurrents à des molécules sémiques, qui dans le contexte de la sémantique interprétative sont caractérisées comme des thèmes spécifiques ou plus particulièrement, comme la co-récurrence de deux sèmes spécifiques indexés par des mots appartenant à des taxèmes, des domaines ou des dimensions différentes. Or, nous avons remarqué que l'identification de domaines communs aux deux sous-corpus permet de relativiser ces distinctions. Les domaines identifiés (//politique//, //éducation// et //économie//, etc.) sont des thèmes génériques présents dans le corpus de référence pris dans son ensemble. Mais dans le contexte des sous-corpus, ces

domaines présentent des différences sémantiques significatives et révélatrices des stratégies argumentatives spécifiques qui ont été employées. Ainsi, les isotopies identifiées en rapport avec ces domaines dans chaque sous-corpus peuvent être appréciées comme thèmes génériques du sous-corpus en question, mais aussi comme des thèmes spécifiques relatifs au corpus de référence, puisque certaines isotopies qui lui sont associées actualisent des sèmes spécifiques à l'intérieur d'un domaine. L'exemple le plus évident de ce phénomène a été décrit dans le domaine //éducation// dans ETUD et GOUV. Nous avons remarqué que l'isotopie /éducation supérieure/ était liée à une spécificité bien contrastée dans chaque groupe. Dans ETUD, l'éducation supérieure était associée à la notion d'accès et de réception dû à la présence de critères comme 'enseignement supérieur'. « Enseignement » renvoie à la notion de réception, puisqu'on reçoit l'enseignement. Par contre, dans GOUV, l'emphase était mise à la notion d'exécution et de réalisation, comme révèle les critères 'des études supérieures' et 'études universitaires' (on fait des études).

Le rapport entre générique et spécifique a aussi été révélé par le calcul de cooccurrence. Nous avons remarqué que le choix des mots pôles 'hausse', 'droit' et 'scolarité' correspondaient dans la majorité des cas à l'expression 'hausse des droits de scolarité', présente dans les mots clés de la requête qui ont servi pour constituer le corpus de référence. Les mots pôles faisaient état d'un thème générique traversant le corpus de référence, mais le fait qu'ils apparaissaient ensemble dans plusieurs contextes a produit certains inconvénients par rapport au repérage de cooccurrents. En effet, nous avons remarqué dans nos résultats que les cooccurrents associés à un de ces trois mots pôles se recoupaient fréquemment, ce qui a réduit le choix de cooccurrents pertinents.

2.2 Sélection et catégorisation de critères textuels dialectiques

Le calcul des spécificités sur des mots simples, des lemmes et des segments répétés nous a permis de découvrir plusieurs critères dialectiques dans les deux sous-corpus. Dans notre méthodologie, nous avons défini certaines catégories de marqueurs susceptibles d'être catégorisés dans la composante dialectique, dans la mesure où ils attestent de l'organisation temporelle, ainsi que de la structuration argumentative des textes. Il s'agit des marqueurs temporels, énumératifs, spatiaux, argumentatifs, etc. Les résultats du calcul des spécificités ont

démontré que plusieurs mots grammaticaux surreprésentés dans les sous-corpus correspondaient en effet à une de ces catégories de marqueurs. Mais nous avons gardé quand même une certaine latitude pour la catégorisation de critères, afin de donner le privilège à l'observation empirique et à l'analyse des contextes dans lesquels les critères textuels étaient employés.

Ce choix méthodologique, d'ailleurs préconisé par la sémantique interprétative (Rastier et coll., 1994, p. 123) s'est montré en effet utile pour la description de la composante dialectique des sous-corpus. Par exemple, le repérage d'une temporalité historique dans le sous-corpus ETUD, a motivé le choix du critère 'moment', qui est un mot lexical absent des catégories de marqueurs répertoriés. Également, la formulation de phrases interrogatives comme stratégie d'argumentation et de polémisation était attestée par des critères comme '-nous' et '-ils'. Avec le trait d'union, ces pronoms – qui pourraient être catégorisés comme dialogiques - s'avèrent plus caractéristiques de la dialectique des textes regroupés dans les sous-corpus.

Par rapport à la structuration argumentative, la description des critères textuels retenus a démontré des régularités importantes dans les deux sous-corpus, révélant dans le camp ETUD la récurrence de critères textuels qui expriment l'idée de saturation (**'**de nouveaux', 'et plus', 'de plus en plus', 'est encore', 'non seulement', 'plus de', 'plus grand'), caractéristique du sentiment d'impatience et de débordement des étudiants pendant le confit. Du côté GOUV, nous avons remarqué une stratégie de dénégation du discours de l'adversaire, exprimé par des critères qui indiquent la formulation de réfutations ('alors qu', 'pourtant', 'Toutefois'**)**. L'interprétation des critères et la consultation du contexte d'emploi dans les textes eux-mêmes nous a permis de voir que la spécificité des mots grammaticaux retrouvés par les calculs et retenus comme critères textuels n'est pas un phénomène aléatoire : elles sont significatives du point de vue de l'élaboration argumentative de chaque groupe. Par ailleurs, nous remarquons que les régularités les plus importantes qui ont été retrouvées sur le plan de la structuration argumentative sont attestées par les spécificités positives les plus importantes. La haute spécificité du mot simple 'Mais' dans GOUV (+11) constitue un cas exemplaire de la stratégie de dénégation du discours adversaire, qui a été confirmé par la récurrence d'autres critères

associés à des contextes où il était question de la réfutation d'un argument ('pas', 'ne', 'est pas parce que', 'par contre', etc.).

Le repérage de critères dialectiques relevant de l'organisation temporelle a démontré que les actions et les événements abordés dans les sous-corpus ne faisaient pas seulement état d'une succession chronologique, mais qu'ils étaient aussi ancrés dans une temporalité sociale (Rastier, 2014), liée à l'histoire et à certaines représentations collectives particulières de la société québécoise, comme démontre l'évocation des années reliées à des événements historiques. D'ailleurs, la représentation du temps dans le récit différait dans ces deux aspects, dépendamment des sous-corpus. Nous observons que dans GOUV, l'organisation temporelle cite des événements plus courants liés aux manifestations organisées par les étudiants tandis que dans ETUD, nous observons une récurrence plus importante d'événements historiques et un rapport sur l'expérience du temps vécu, comme l'indique la présence des années. Par exemple, les critères 'aujourd'hui' et 'demain' dans ETUD, n'actualisent pas nécessairement les isosémies //présent// et //futur// qui seraient reliées au déroulement chronologique des actions et des événements, mais ils font état sémantiquement d'un sentiment d'anticipation de l'avenir de la société dans le temps historique, lié à l'importance qu'il faut accorder aux événements « d'aujourd'hui », que sont les manifestations.

L'analyse du contexte d'emploi des critères dialectiques avec le concordancier a été moins importante dans le cadre de notre démarche interprétative, puisque la récurrence sémique de ces critères, qui sont pour la plupart des mots grammaticaux, pouvait être envisagée en fonction de l'actualisation de sèmes inhérents particuliers au système de la langue française. Le recours à l'ouvrage « Grammaire du sens et de l'expression » de Charaudeau (1992) et aussi à certains concepts de la sémantique interprétative s'est avéré important dans le cadre de notre analyse et de notre interprétation des récurrences observées, permettant non seulement de les décrire, mais aussi de rendre compte des différences existantes entre les sous-corpus.

2.3 Sélection et catégorisation de critères textuels dialogiques

Le calcul des spécificités a aussi rendu possible l'élaboration de critères dialogiques qui décrivent l'énonciation représentée et les aspects modaux. La plupart des critères retenus

dans la composante dialogique font partie des catégories de marqueurs définis dans notre méthodologie (p.219 à 222) comme les pronoms personnels (‘nous’ et ‘ils’), les marqueurs signalant les pensées rapportées (‘« »’), les modalités exprimant la pensée subjective (‘je pense’, ‘semble que’), et la formulation d’hypothèses (seraient’ ‘serait’, ‘si les’). La plupart des critères dialogiques retrouvés sont liés à l’énonciation et au rapport entre l’énonciateur et ses interlocuteurs (lecteurs auxquelles il se dirige). Quelques critères dialogiques liés à la modalisation ont été retrouvés, principalement dans le sous-corpus GOUV.

Comme dans le cas de l’élaboration de critères dialectiques, nous avons laissé de la place à l’investigation empirique et à l’analyse du contexte d’emploi pour catégoriser certains critères textuels comme dialogiques. Le positionnement énonciatif a été également évalué et décrit sur le plan de la relation entre l’énonciateur et ses interlocuteurs, ce qui a eu pour effet la sélection de certains mots lexicaux tels que ‘blogue’ et ‘Professeur’, dans cette catégorie de critères.

Par rapport à l’énonciation représentée, la présence marquée de pronoms personnels et de pronoms possessifs rend clair le positionnement énonciatif de chaque groupe. Le sous-corpus ETUD est marqué par la récurrence de ‘nous’, qui révèle un positionnement identitaire fort et qui fait appel à la notion de collectivisme. Ce critère est d’ailleurs un des plus spécifiques de ce sous-corpus. Il contraste avec l’impersonnalité du sous-corpus GOUV, dans lequel l’énonciateur se distancie pour se positionner contre une cible représentée par le ‘ils’ et associée, dans la plupart des cas, aux étudiants eux-mêmes. L’articulation de ce ‘nous’ contre ‘ils’ illustre l’opposition d’un groupe qui se positionne pour la défense d’une cause et d’un autre qui choisit la voie de l’attaque frontale (Garand, 1998).

3. Facteurs contribuant au succès de la classification automatique d'articles d'opinion

Les résultats de la classification automatique effectuée sur notre corpus démontrent que la sélection de critères textuels, linguistiquement orientés, à partir d'une étude contrastive de corpus, est efficace pour la prédiction de la classe des articles d'opinion provenant d'une controverse médiatique. Avec le classifieur NB et en utilisant une pondération binaire, nous avons réussi à classer correctement 91 % des articles du corpus d'apprentissage avec la validation croisée à 5 itérations. Ces articles ont été représentés par 447 critères textuels globaux qualifiés et sélectionnés à partir du calcul des spécificités. L'ensemble de ces critères comprend les trois catégories de critères définis par notre recherche (thématique, dialectique et dialogique) et combine aussi des critères textuels typés en fonction de la segmentation effectuée sur les textes du corpus (critères unitaires simples, critères unitaires lemmatisés et critères adjacents). De manière générale, la démarche de sélection de critères démontre une bonne efficacité pour la tâche de classification automatique. Sur un total de 144 tests, 80 ont surpassé les résultats obtenus par les meilleures lignes de comparaison auxquelles ils ont été comparés, soit 56 % des tests.

Dans le cadre de cette recherche, nous avons testé différents paramètres pour évaluer la classification automatique des articles du corpus, l'objectif étant de découvrir expérimentalement quels sont les paramètres optimaux. Souvent dans les recherches de fouille de textes (et par extension en fouille d'opinions), les questions concernant les modalités d'application des algorithmes, ainsi que les seuils à partir desquels les critères textuels sont considérés comme discriminants pour une tâche spécifique, n'ont pas de réponses uniques et complètement satisfaisantes. Les réponses sont le plus souvent obtenues de manière exploratoire, en procédant par évaluations et hypothèses successives (Forest et coll., 2009). Dans le cadre de notre recherche, nous avons évalué l'impact produit par le choix des algorithmes, par le type de pondération (binaire ou par fréquence) ainsi que par le nombre et les types de critères textuels. Dans les sections suivantes, nous avons évalué les conséquences de chacun de ces choix sur la performance de la classification automatique, soulignant les

facteurs qui contribuent aux bonnes performances obtenues ou au contraire, qui nuisent aux performances des classifieurs.

3.1 Impact du choix du nombre de critères spécifiques pour la classification des articles

Les expérimentations effectuées dans le cadre de cette recherche montrent qu'il est possible d'obtenir de très bons résultats pour la classification des articles en utilisant une petite quantité de critères textuels, comparativement aux deux lignes de comparaison constituées, qui comportent chacune 10 866 et 7044 mots simples et lemmes, respectivement. Si nous prenons par exemple l'expérimentation 1 avec pondération par fréquence, dans laquelle nous avons varié les types segmentation (mots simples, lemmes et segments répétés), nous observons que seulement 48 critères textuels de type unitaire simple (avec juste des mots simples sélectionnés) permettent d'obtenir un résultat supérieur à celui obtenu par la ligne de comparaison la plus performante (LC1_motsSimples). Lorsque les tests sont effectués sur les mêmes matrices, mais à pondération binaire, nous observons également qu'un nombre de 48 critères textuels de type unitaire simple permet déjà d'obtenir des résultats proches. En comparant aussi les tests effectués dans les expérimentations 2, 3 et 4, nous observons aussi que la prise en compte d'un petit nombre de critères textuels dont le score de spécificité est plus grand ou égal à +4 permet de dépasser les résultats de la ligne de comparaison dans la majorité des tests effectués.

D'une part, ces résultats montrent que les critères sélectionnés sont très discriminants pour la classification automatique. Lorsque nous comparons les tests qui ont utilisé la totalité de critères sélectionnés (ceux avec le score de spécificité à partir de +2), nous remarquons que tous les résultats dépassent ceux obtenus par les lignes de comparaison, à l'exception seulement des critères locaux, dont les résultats sont inférieurs dans tous les tests effectués.

Les figures 22 et 23 représentent la corrélation entre le nombre de critères textuels sélectionnés pour nos expérimentations et le taux de rappel et de précision obtenus pour la classification automatique, en utilisant le classifieur NB et la pondération binaire. D'après la courbe de tendance, nous pouvons voir que la performance du classifieur augmente lorsque la quantité de critères textuels utilisés augmente aussi. Par contre, dans les deux graphiques, nous

pouvons remarquer certaines valeurs aberrantes (des points qui sont écartés de la courbe). Ces valeurs sont celles obtenues par la matrice M5_Locaux, laquelle comporte seulement des critères locaux (cooccurrents spécifiques sélectionnés).

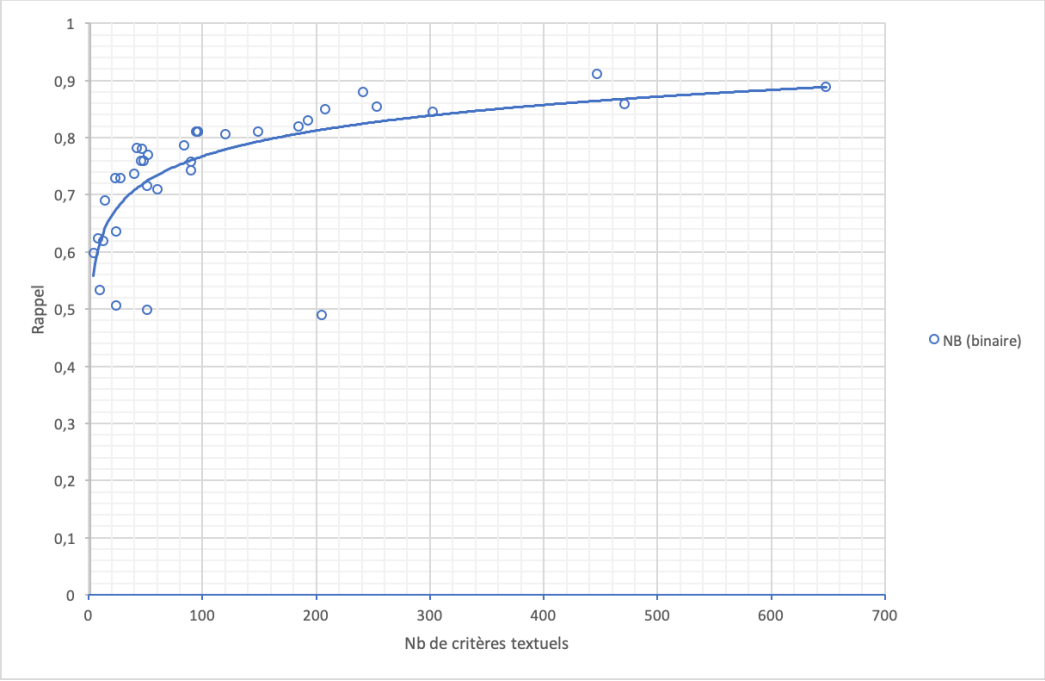


Figure 22. Taux de rappel en fonction du nombre de critères textuels sélectionnés

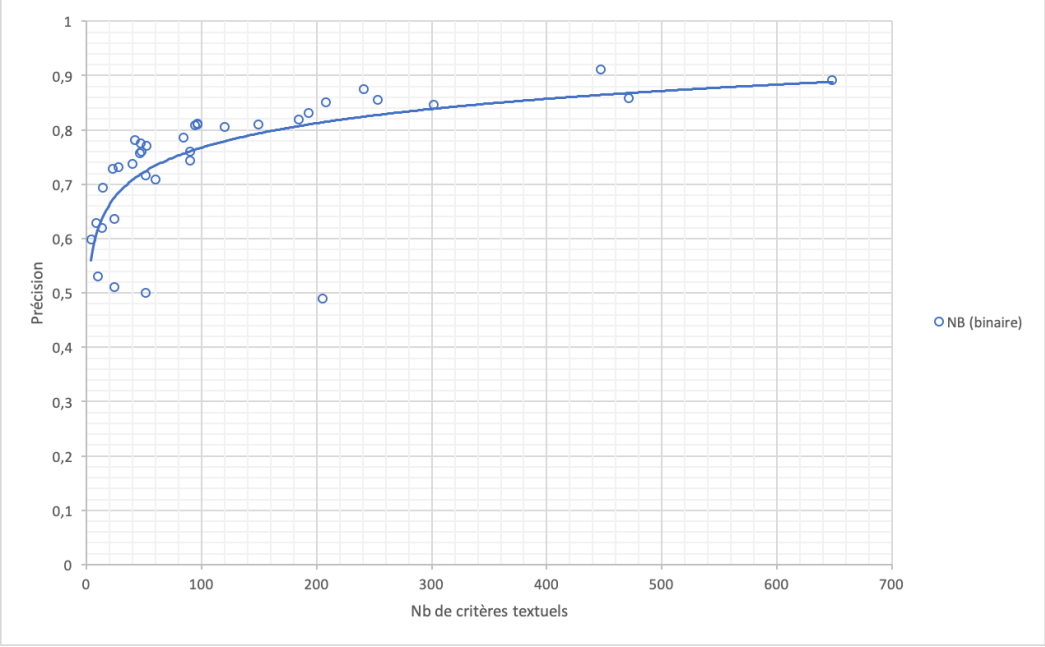


Figure 23. Taux de précision en fonction du nombre de critères textuels sélectionnés

Les résultats obtenus dans les expérimentations sont comparables à ceux obtenus par les études menées par Eensoo et Valette (2012, 2014a, 2014b, 2015) dans le domaine de la fouille d'opinions, malgré l'obtention de la mauvaise performance des critères locaux pour la classification.

Nous avons aussi remarqué que le choix de retenir des critères avec un seuil de spécificité égal ou supérieur à +2 est le choix le plus approprié dans le cadre de notre expérimentation. Dans plusieurs cas, nous notons que les expérimentations utilisant des critères dont le score est égal ou supérieur à +3 démontrent aussi une performance, mais qui n'est pas supérieure à celle obtenue dans les expérimentations avec une plus grande quantité de critères spécifiques. Plus les critères sont spécifiques, plus la performance de la classification tend à diminuer. Mais ce résultat peut aussi être lié au nombre de critères qui ont été utilisés dans l'expérimentation et pas nécessairement à la spécificité des critères, car la quantité de critères varie de façon importante selon le score de spécificité considéré.

3.2 Impact de la pondération et du choix des classifieurs sur la performance des classifieurs

Les expérimentations menées dans le cadre de notre recherche démontrent une différence de la performance des classifieurs NB et SVM en fonction de la pondération choisie dans les matrices vectorielles. Les figures ci-dessous montrent la comparaison entre les taux de rappel et précision obtenus pour chacun des classifieurs dans chaque type de pondération. Comme nous pouvons observer, le classifieur SVM performe mieux avec la pondération par fréquence (figures 24 et 25). Le classifieur NB performe de manière plus satisfaisante lorsque nous utilisons une pondération binaire (figures 26 et 27). Les meilleurs résultats sont obtenus avec NB en pondération binaire.

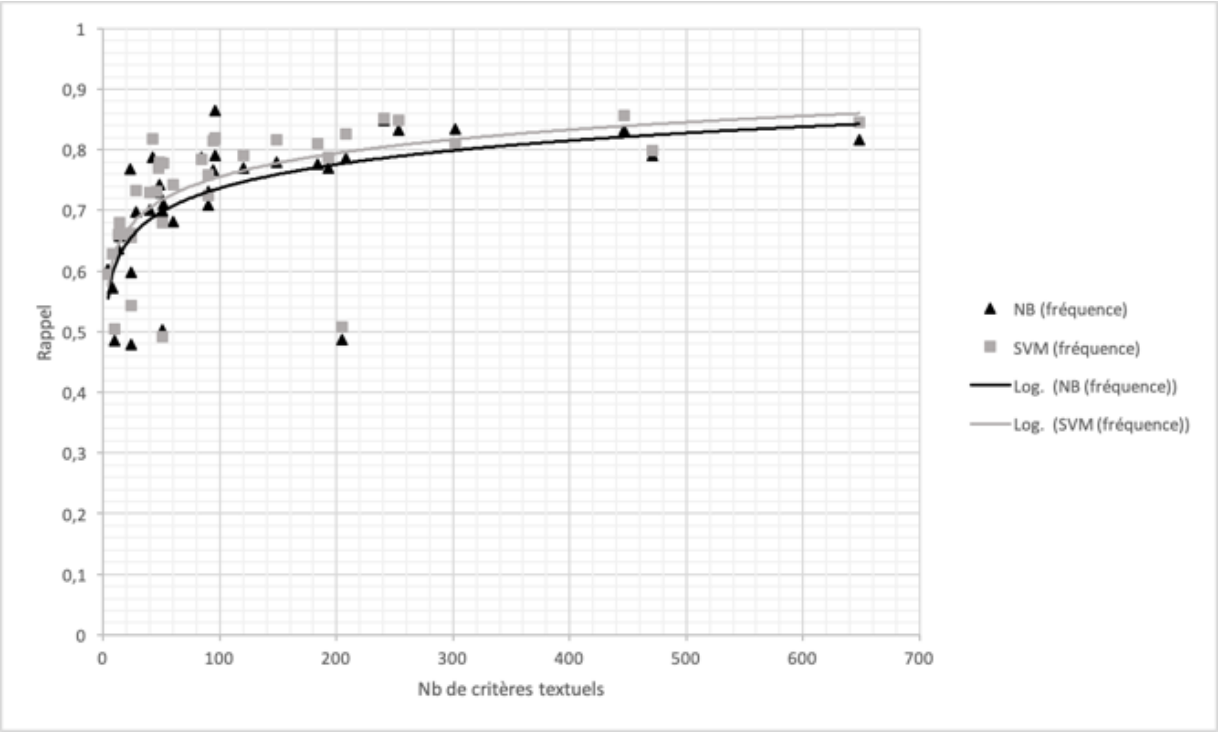


Figure 24. Taux de rappel obtenu par NB et SVM en pondération par fréquence

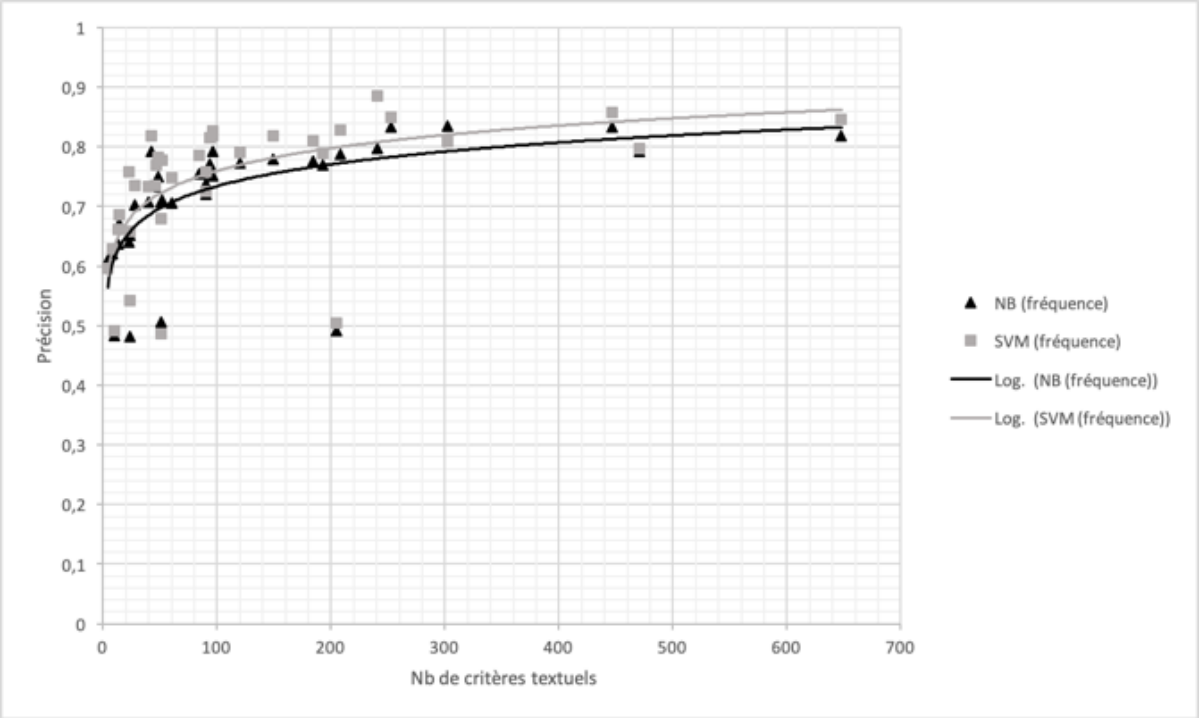


Figure 25. Taux de précision obtenu par NB et SVM en pondération par fréquence

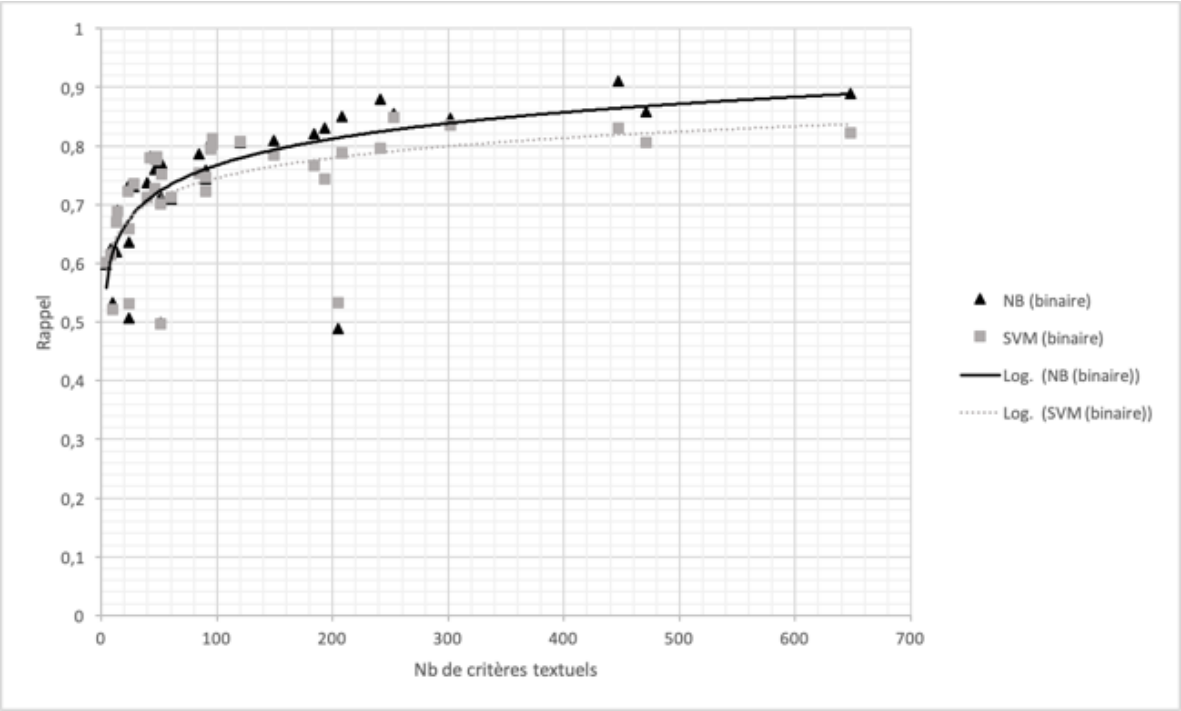


Figure 26. Taux de rappel obtenu par NB et SVM en pondération binaire

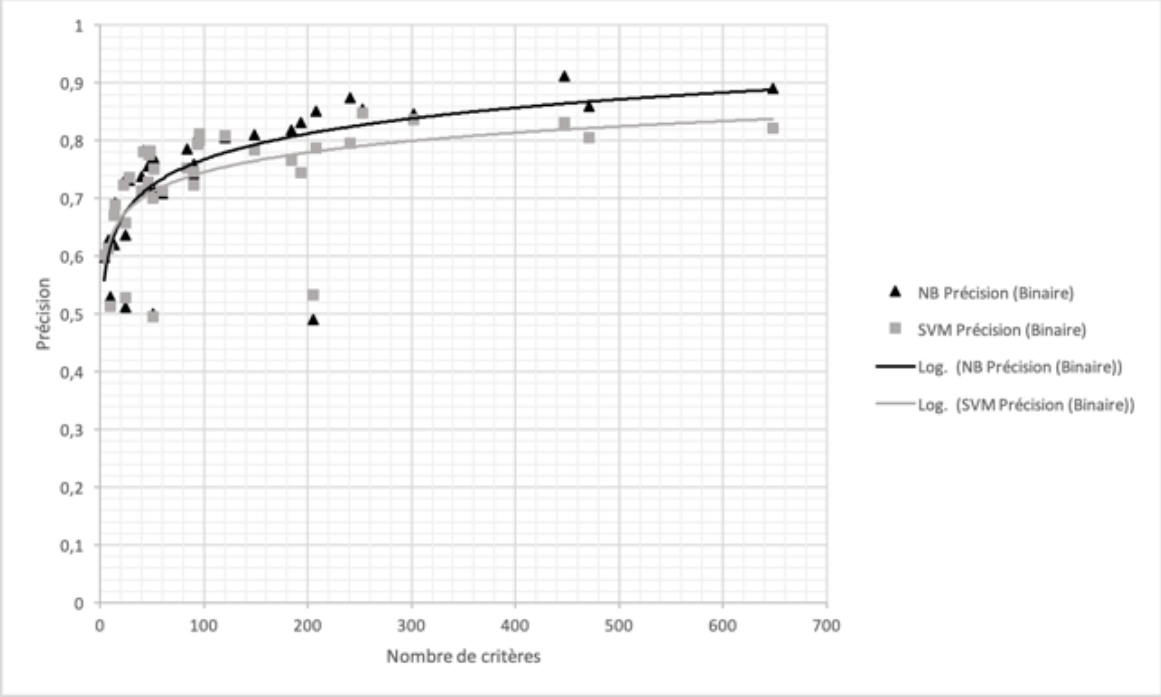


Figure 27. Taux de précision obtenu par NB et SVM en pondération binaire

En se basant sur les expérimentations de fouille d'opinions précédentes, notamment celle de Pang et coll. (2002), Eensoo et Valette (2012, 2015) préconisent une représentation des documents en fonction de la fréquence absolue des mots. Dans les expérimentations qu'ils ont menées, ces auteurs constatent que le classifieur SVM démontre une meilleure performance que NB pour la classification de témoignages et de récits publiés par les internautes (Eensoo et Valette, 2012) et aussi de tweets (Eensoo et Valette, 2015).

Les facteurs influant sur la performance des classifieurs et sur le rapport avec le type de pondération utilisé sont difficiles à déterminer. Dans *Weka*, deux types de classifieurs sont disponibles, le bayésien naïf multinomiale (*Naive Bayes Multinomial)* et le bayésien naïf tout court (*Naive Bayes*). La différence entre ces deux classifieurs se trouve sur le plan du modèle de probabilité que chacun utilise (McCallum et Nigam, 1998) : le premier est basé sur le modèle multinomial et spécifie qu'un texte est représenté par l'ensemble des occurrences de mots. Lorsque le classifieur calcule les probabilités, il considère seul les mots ayant des occurrences, en ignorant les valeurs nulles. Le deuxième est basé sur le modèle de Bernouille (McCallum et Nigam, 1998 ; Schneider, 2004) et considère que les textes sont représentés par un vecteur d'attributs binaires, qui prend en compte les présences et les absences des mots. Lorsqu'il calcule les probabilités sur les valeurs non binaires (fréquences absolues), il multiplie la probabilité de tous les attributs, incluant les mots qui sont absents. McCallum et Nigam (1998) expliquent que ce dernier est plus approprié pour des textes représentés avec des valeurs nominales c'est-à-dire, qui ont des valeurs fixes (0 et 1). Dans *Weka*, le classifieur *Naive Bayes* fonctionne aussi avec des valeurs multinomiales, comme les fréquences absolues (Witten et Frank, 2005). Cependant, lorsque l'algorithme est appliqué à des matrices comportant une telle représentation, le calcul considère aussi les valeurs nulles et pondère les fréquences en calculant la moyenne dans la distribution. Des tests supplémentaires avec le modèle multinomial permettraient de valider ces suppositions par rapport à notre expérimentation.

3.3 Impact du type de segmentation sur la performance des classifieurs

Les résultats montrent que la lemmatisation ne permet pas d'améliorer les résultats. Nous pouvons d'une part observer que les lignes de comparaison constituées par des mots

simples (LC1_motsSimples) sont plus performantes que celles constituées de lemmes (LC2_Lemmes) pour la classification. D'autre part, même si les résultats des expérimentations avec les critères unitaires lemmatisés sélectionnés sont meilleurs que la ligne de comparaison composée de lemmes, ils sont inférieurs par rapport à la performance obtenue par les critères unitaires simples, composés seulement de mots simples.

En observant les résultats de l'expérimentation 1, nous pouvons observer que les critères unitaires simples sont les plus performants. La figure ci-dessous permet de voir que M1_CritèresUnitairesSimples performe de façon supérieure, en comparant aux autres matrices, dans les différents seuils de critères considérés.

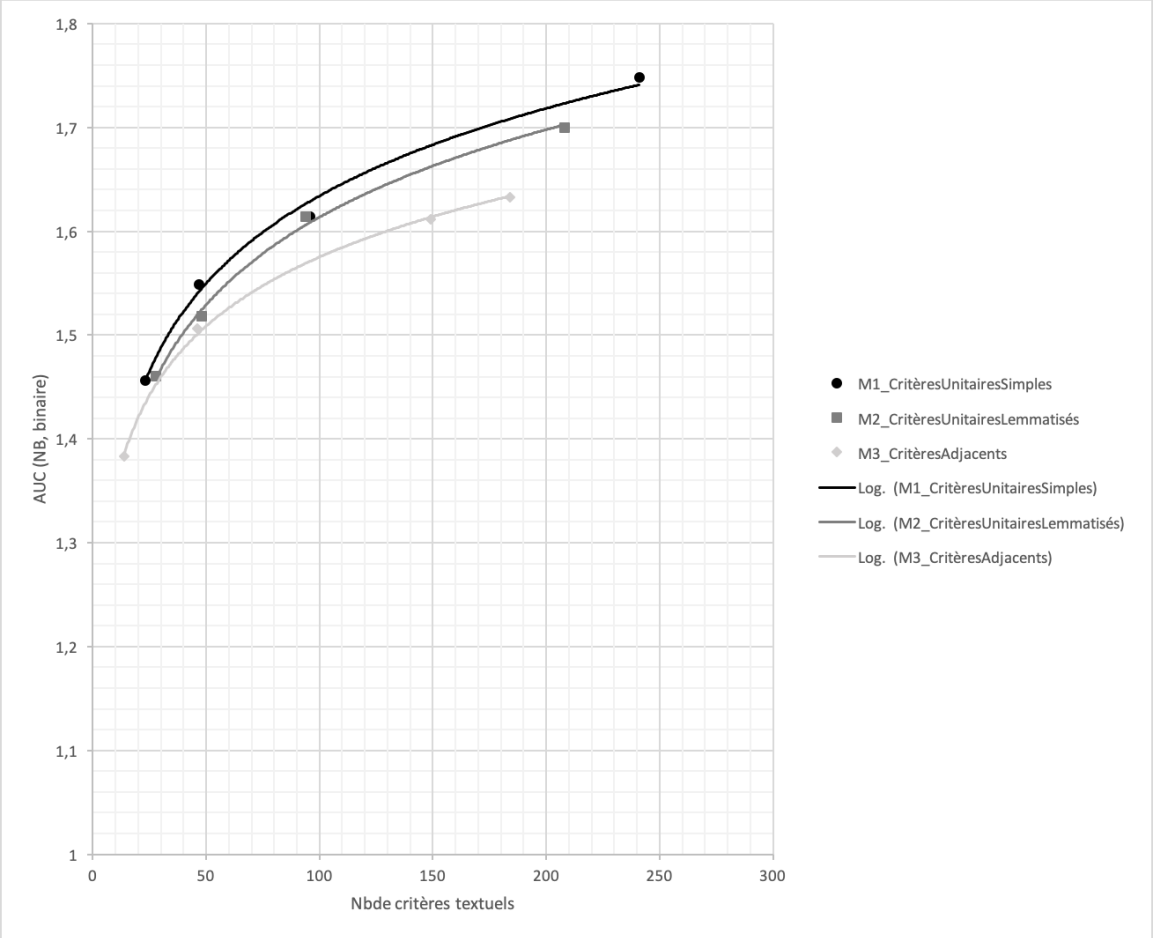


Figure 28. Comparatif entre les matrices de l'expérimentation 1 (résultats en AUC, avec le classifieur NB et la pondération binaire)

Nous avons également observé que la combinaison de critères unitaires simples, de critères unitaires lemmatisés et de critères adjacents ne permet pas d'améliorer significativement la performance des classifieurs. Cependant, cette combinaison n'a pas nécessairement d'effet négatif sur la classification. La figure 29 montre une comparaison entre les résultats en AUC du classifieur NB obtenu pour différentes quantités de critères par les matrices binaires composées de critères unitaires simples (M1_UnitairesSimples), critères globaux (M4_Globaux) et tous les critères (M9_Tout). Nous pouvons voir que les résultats avec les critères unitaires simples (M1_MotsSimples) sont comparables au niveau de la performance aux critères globaux (M4_Globaux) et à tous les critères (M9_Tous), qui combinent différents les types de segmentation (mots simples, lemmes et segments répétés).

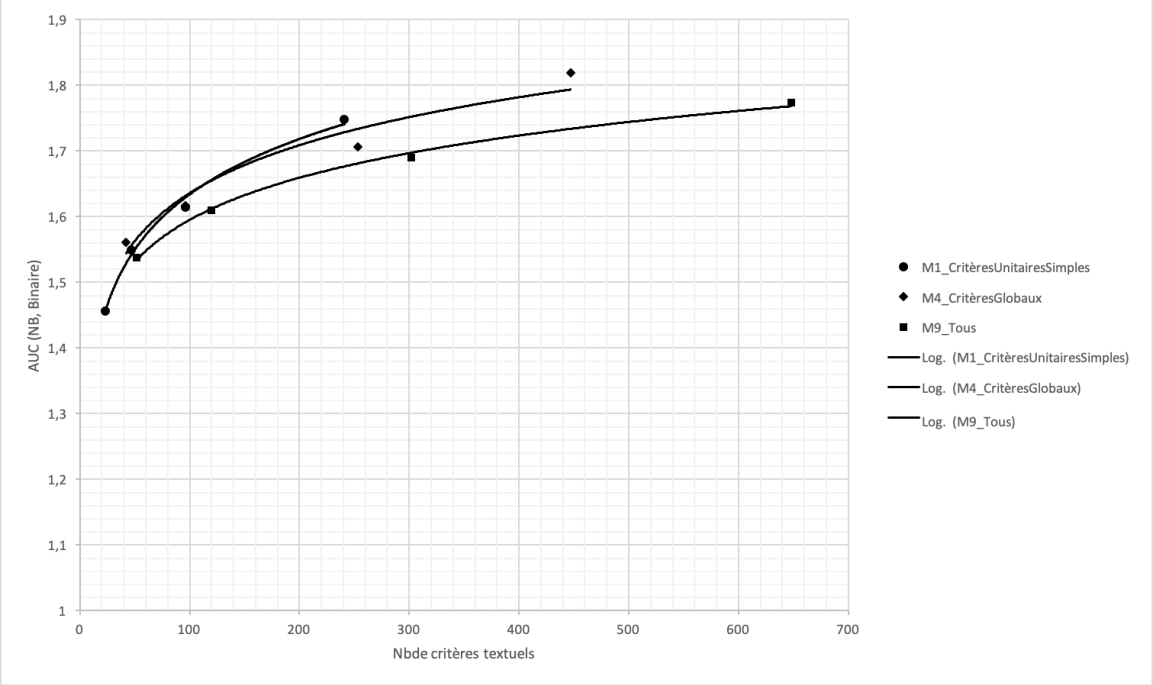


Figure 29. Comparaison entre la performance des critères unitaires simples (M1_MotsSimples) et les critères globaux et tous les critères (M4_Globaux et M9_Tous).

Lors de la combinaison de critères dans les matrices, nous avons effectué certains choix par rapport aux critères unitaires lemmatisés qui se retrouvaient également parmi les critères unitaires simples. Nous avons pris la décision de conserver les critères unitaires simples qui avaient un score de spécificité plus grand que son correspondant sous une forme lemmatisée ou normalisée. Par exemple, nous avons trouvé dans le sous-corpus GOUV le mot

simple ‘Mais’ en majuscule, avait un score de spécificité de +11. Dans le même sous-corpus, nous avons aussi retrouvé la forme normalisée ‘mais’, mais avec un score de +2. Même si ce dernier s’avérait plus fréquent que le mot simple ‘Mais’ dans le sous-corpus, nous avons pris la décision de conserver le critère le plus spécifique, puisqu’il était rare dans le sous-corpus ETUD. Par ailleurs, lorsqu’un mot simple avait un score de spécificité proche de son correspondant lemmatisé, nous avons décidé de plutôt conserver le lemme. D’après nos résultats, nous pouvons inférer que cette stratégie de sélection s’avère efficace.

3.4 Coalition de critères thématiques, dialectiques et dialogiques

Les expérimentations montrent qu’il existe un impact positif entre le choix de l’ensemble de critères thématiques, dialectiques et dialogiques et la performance de la classification. Nous pouvons observer cette corrélation en comparant l’expérimentation 3, dont les tests sont faits avec les catégories des critères thématiques, dialectiques et dialogiques pris isolément (M6_Thématiques, M7_Dialectiques et M8_Dialogiques), et les autres expérimentations, dont les matrices combinent ces trois catégories indistinctement : M1_CritèresUnitairesSimples, M2_CritèresUnitairesLemmatisés, M3_CritèresAdjacents, M4_CritèresGlobaux, M9_Tous). Nous avons représenté dans la figure 30 ci-dessous les résultats en AUC obtenus pour les tests effectués sur chacune de ces matrices, en utilisant le classifieur NB et la pondération binaire. Les lignes pointillées sont les matrices qui présentent les trois catégories de critères combinés. Nous pouvons observer qu’à l’exception de M6_Thématiques, les matrices utilisant une seule catégorie de critères (M8_Dialogiques et M9_Dialectique) présentent un résultat inférieur pour la classification, lorsque comparé à toutes les autres matrices, considérant la même quantité de critères textuels. M6_Thématique présente des résultats voisins à M1_Mots et comparables à M3_Segments et M2_Lemmes, lesquels combinent des critères thématiques, dialectiques et dialogiques. Cependant, les matrices M4_Globaux et M9_Tous, qui englobent les trois catégories de critères avec différents types de segmentation (mot simple, lemme, segments répétés et cooccurrents) présentent les meilleurs résultats.

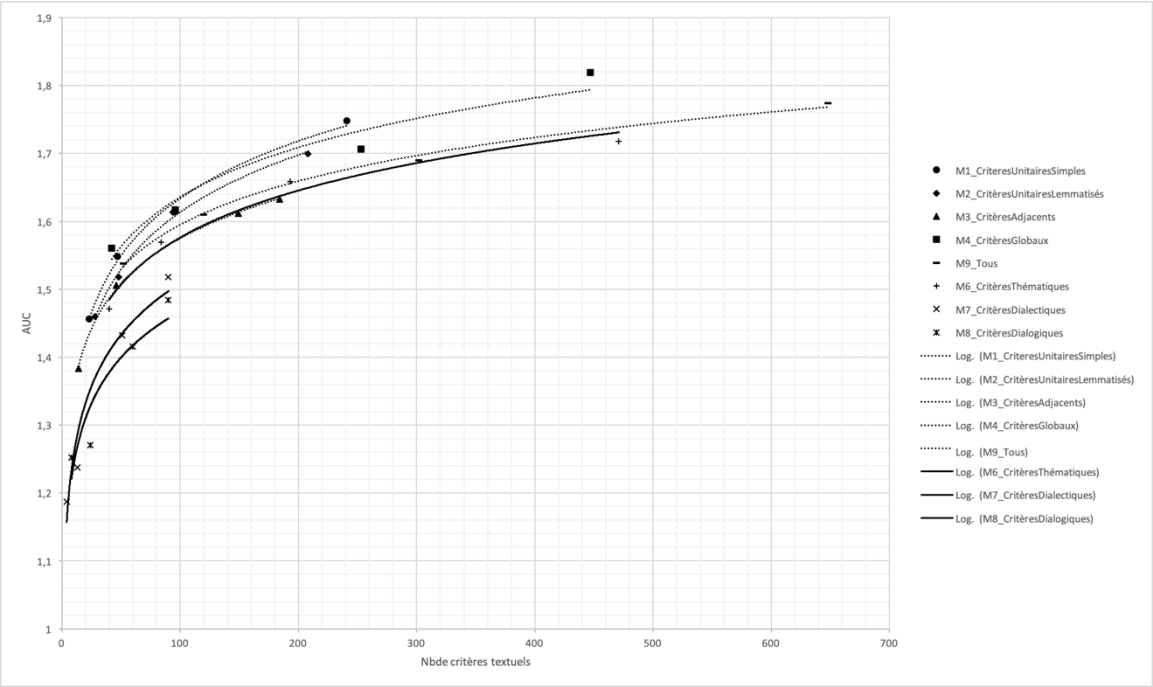


Figure 30. Comparaison entre les matrices comportant une seule catégorie de critères et les matrices avec les critères thématiques, dialectiques et dialogiques combinés

Ces résultats sont comparables à ceux obtenus par les études de Eensoo et Valette (2012, 2014a, 2014b, 2015), qui ont aussi démontré que la coalition de ces critères est plus performante que lorsqu'ils sont pris isolément. Nous avons trouvé par ailleurs que les critères thématiques sont ceux qui permettent d'atteindre les meilleurs résultats pour la classification. Nos résultats nous amènent aussi aux mêmes conclusions que ces auteurs : la prise en compte des niveaux de description thématique, dialectique et dialogique emporte sur le résultat de la classification, confirmant la pertinence de la caractérisation différentielle des sous-corpus de sélection de critères textuels par un processus d'identification et de qualification des récurrences sémantiques.

3.5 Faible performance des critères locaux

Les tests avec les critères locaux de la matrice M5_Locaux (tableau 43, p. 313) se sont avérés très peu performants dans toutes les expérimentations. Par ailleurs, la combinaison de tous les types de critères dans la matrice M9_Tous (tableau 43, p. 318) n'apporte pas les meilleurs résultats pour la classification. Il est intéressant de noter que la seule différence entre M9_Tous et M4_Globaux, qui est la matrice la plus performante, est la présence des critères

locaux dans la première. Par conséquent, l'utilisation de critères locaux en amont d'autres types de critères affecte négativement la performance de la classification.

La faible performance des critères locaux dans la tâche de classification automatique exige un examen sur l'efficacité de la démarche de sélection de ces cooccurrents. Dans notre démarche, nous avons calculé les cooccurrents pour chaque mot pôle choisi dans les deux sous-corpus et nous avons éliminé tous les cooccurrents qui se retrouvaient en double dans ceux-ci, afin de repérer ceux qui étaient uniques. Cependant, les cooccurrents uniques, même s'ils étaient aussi spécifiques à un sous-corpus particulier, présentaient de faibles fréquences dans l'ensemble du corpus. Cela est possiblement une des causes affectant la faible performance de ces critères textuels pour la classification des articles dans notre expérimentation.

4. Application de la démarche dans le contexte des SRAP

Le troisième objectif de notre recherche était d'effectuer des analyses textométriques sur différentes périodes du corpus, afin d'explorer à quel moment il était possible de retrouver l'ensemble des critères qui ont démontré la meilleure performance pour la tâche de classification automatique. Pour atteindre cet objectif, nous avons réalisé l'analyse textométrique de 9 sous-corpus qui rassemblent cumulativement les articles d'opinion publiés sur la grève étudiante au cours de la période où l'événement était au centre de l'attention médiatique. Nous avons analysé l'apparition chronologique des critères globaux sélectionnés (critères de la matrice M4_Globaux), en regardant le pourcentage de critères qui pourraient être sélectionnés à partir des sous-corpus chronologiques.

Aussi, afin de comprendre la performance des critères au fil du temps pour la prédiction de la classe des articles, nous avons effectué des tests de classification automatique, en constituant des corpus d'apprentissage et des corpus de test en fonction d'une répartition chronologique des articles. Notre objectif était d'évaluer la possibilité d'atteindre une bonne performance pour la classification des articles, en entraînant les classifieurs avec la quantité de critères globaux qui ont été retrouvés au début de la controverse et en utilisant aussi, pour l'entraînement, les articles publiés dans les premiers mois du conflit, afin de prédire la classe de ceux qui ont été publiés postérieurement.

Dans cette section, nous discutons les résultats de ces tests, afin de nous prononcer sur la pertinence de l'élaboration de critères textuels à partir de la méthode textométrique pour le développement des SRAP. Cependant, nous ne prétendons pas donner une solution à la modélisation des systèmes de recommandation à partir de la méthodologie de fouille d'opinions que nous avons proposée dans le cadre de cette recherche. Nous voulons seulement fournir des points de réflexion pour la modélisation de tels systèmes, à partir de l'évaluation sur l'évolution du vocabulaire du corpus au fil du temps et du pouvoir des critères textuels pour la prédiction d'articles futurs. Nous exposons néanmoins certaines possibilités d'application de nos découvertes et de généralisation de nos résultats, tout en considérant les limites posées par le cadre expérimental de notre recherche, autant que par les particularités du développement des SRAP dont il faut tenir compte.

4.1 Étude chronologique des sous-corpus

Une première analyse des sous-corpus constitués de façon chronologique nous a permis de constater que la répartition du nombre d'articles dans le temps est inégale et qu'elle connaît des pics, suivis de moment de stabilisation. Cela n'est pas surprenant du point de vue du phénomène étudié, puisque l'attention que les médias consacrent à un certain événement peut varier de façon importante avec le temps. Il faut surtout considérer que dans la presse, l'opinion au sujet d'un événement quelconque vient après une certaine maturation de ce dernier dans l'actualité (Jacobs et Townsley, 2011). Ainsi, même si la grève étudiante faisait la une depuis le début du mouvement, le sujet est apparu dans les tribunes libres des journaux un peu plus tard, et à mesure que l'événement gagnait de l'ampleur et l'attention dans les médias.

Par ailleurs, les nouveaux événements rapportés dans l'actualité peuvent faire revivre d'anciennes controverses, suscitant des débats marqués par le conflit et par la polarisation. Par exemple, le changement de température dans une région peut faire réveiller la controverse sur le changement climatique, et réveiller également la discussion au sujet de l'impact de l'activité économique sur l'environnement (Charaudeau, 2017). Dans ce sens, les controverses peuvent supporter de grandes variations pendant les années, avec des pics de publications et des moments de « silence », dans lesquels d'autres événements prennent le relais. Le corpus de la grève étudiante que nous avons constitué est quand même petit, et concerne seulement la grève étudiante de 2012. Mais la controverse sur la question du financement des universités remonte à quelques années : en 2005, les modifications du régime de prêts et bourses, ainsi que le projet de coupure de 103 M$ dans le programme d'aide financière aux études ont suscité également une grève de 4 mois, qui a aussi soulevé des discussions publiques sur l'enjeu du financement de l'éducation (Pétry et coll., 2006).

Dans la perspective de la méthodologie de fouille d'opinions proposée dans le cadre de notre travail, la quantité d'articles du corpus et les ordres de grandeur de fréquences sont des facteurs importants dans l'élaboration des critères textuels. Plus la taille du corpus est importante, plus il est possible de former des critères textuels à partir des spécificités rencontrées. Ainsi, le fait que très peu de critères globaux aient été retrouvés dans les mois de février et mars est aussi lié à la taille du vocabulaire de ces sous-corpus, comparativement aux autres mois succédant. Nous observons un premier pic dans le nombre de publications au mois

d'avril, suivi par une augmentation de l'ordre de 122 680 occurrences au mois précédent. Le nombre de mots simples est aussi plus important, comptant 14 291 mots simples, soit 8425 de plus qu'au mois de mars. En avril, nous comptons aussi 272 articles, ce qui représente plus que la moitié du nombre d'articles compris dans le corpus de référence de notre recherche (495).

Si la probabilité de retrouver les critères globaux accroît en fonction de la taille des sous-corpus chronologiques, nous avons constaté que cette corrélation n'est pas exactement proportionnelle à la quantité d'articles rassemblés dans les sous-corpus. Par exemple, même si le sous-corpus du mois avril comptait 54 % du nombre d'articles du corpus de référence, nous avons pu retrouver seulement 36 % du nombre de critères textuels globaux. Au mois de mai, le sous-corpus comptait 75 % du corpus total et c'est en mai que nous observons que la quantité de critères globaux retrouvée est la plus proche du seuil de 50 % de la quantité totale de critères globaux sélectionnés dans l'expérimentation avec la matrice M4_Globaux.

Par ailleurs, il est intéressant de voir que seulement 50 % des critères globaux qui ont été retrouvés dans le sous-corpus en mai sont déjà assez performants pour la classification des articles, en considérant le facteur chronologique dans la constitution du corpus d'entraînement et des tests. Avec les 218 critères globaux retrouvés en mai (SC_Mai), nous avons pu atteindre un taux de rappel et précision de 85 % dans la classification d'articles avec le classifieur SVM pour les quatre mois suivants, même si ce résultat est inférieur à celui obtenu lorsque tous les critères globaux sont considérés (l'expérimentation avec le SC_Octobre a obtenu 90% en rappel et précision). Après le mois de mai, nous observons un autre pic relatif au nombre de publications au mois de juin (avec un accroissement de 135 nouvelles publications, et un ajout de 3373 nouveaux mots simples), après lequel le vocabulaire du corpus se stabilise, ce qui explique également l'amélioration et la stabilisation de la performance des classifieurs dans les tests effectués pour les sous-corpus chronologiques subséquents.

Il est aussi important de souligner que les résultats des tests de classification effectués pourraient atteindre de meilleurs résultats si d'autres critères textuels étaient pris en compte, ce qui pourrait être le cas dans tous les sous-corpus chronologiques. Notre recherche n'a pas évalué la possibilité de sélectionner d'autres critères textuels à partir du résultat du calcul des spécificités dans les sous-corpus, mais les résultats que nous avons obtenus pour la

classification d'articles laisse croire que la prise en compte de critères textuels en fonction des paramètres définis (score de spécificité, pertinence linguistique et distribution) est efficace. Dans tous les résultats obtenus pour la classification d'articles avec les différents types de critères (à l'exception des critères locaux), les critères textuels linguistiquement pertinents avec un score de spécificité égal ou supérieur à +2 et présents dans plus de deux documents des sous-corpus ETUD et GOUV sont assez discriminants pour la classification des articles. Cette observation nous amène à conclure que l'accumulation de critères textuels obéissants à ces paramètres peut améliorer le résultat de la classification.

Tout en gardant les caractéristiques propres à notre expérience, surtout relativement aux étapes de constitution de notre corpus de référence, nous pouvons conclure d'après l'analyse chronologique menée qu'il est possible d'élaborer des critères textuels pertinents pour la classification automatique d'articles d'opinion lorsque la controverse atteint un pic sur le plan du nombre de publications, permettant ainsi de prédire 85 % des articles d'opinions publiés postérieurement. Il faut considérer par contre que dans notre expérimentation, il y a une certaine stabilisation terminologique du corpus après les pics observés. Les pics qui se produisent au cours des controverses s'avèrent de bonnes sources pour la découverte et la sélection de critères textuels pour la classification automatique.

4.2 Application de la méthodologie pour le développement d'un SRAP : limites et possibilités

Les résultats relatifs à notre troisième question de recherche nous permettent de constater que l'élaboration de critères textuels suffisamment robustes pour la classification automatique d'articles d'opinion peut être effectuée de manière relativement anticipée, ce qui peut assurer la reproductibilité de la démarche dans un corpus comparable à celui que nous avons constitué dans le cadre de cette recherche. Il faut néanmoins pondérer sur la généralisation de nos observations dans d'autres contextes, puisqu'elle se confronte aux caractéristiques du corpus de référence constitué, autant sur le plan de la taille de ce dernier que de la progression de son vocabulaire dans le temps. Néanmoins, il est encourageant de percevoir qu'un premier pic dans le nombre de publications permet d'élaborer des critères

textuels qui peuvent être performants pour la prédiction de la classe des articles au fil de quelques mois.

Les résultats renforcent l'idée que l'approche textométrique peut contribuer à la modélisation d'un SRAP, dans un contexte de recommandation d'articles d'opinion issus d'une controverse médiatique, dans la mesure où l'apprentissage sur un ensemble d'articles apparus précédemment est efficace pour prédire des articles publiés postérieurement dans la controverse. Une analyse chronologique du même genre que nous avons menée sur une controverse qui s'étale sur une période plus importante aurait permis d'obtenir un niveau de confiance plus grand sur ce constat. Principalement si la controverse présente dans le temps plusieurs pics correspondant à des moments où la discussion gagne de l'ampleur, et dans lesquels nous pouvons observer une augmentation du nombre de publications et aussi du vocabulaire.

Si l'utilisation de la textométrie comme préoutillage pour la fouille d'opinions se démontre profitable dans le cadre de notre expérimentation, son utilisation dans un contexte applicatif comme celui de la recommandation doit considérer plusieurs facteurs importants. En premier lieu, nous avons restreint notre expérimentation à des genres journalistiques spécifiques. Notre méthode a évalué la classification d'articles d'opinion dans un contexte particulier, impliquant le choix de certains genres de l'opinion, et la sélection d'un thème bien délimité. La méthode contrastive de la textométrique présuppose une conception assez sophistiquée du corpus, qui requiert d'une part l'élaboration d'un corpus de référence en fonction de certaines exigences (genres particuliers, thème, langue, etc.) et d'autre part par la caractérisation de sous-corpus contrastés, afin permettre l'opérabilité des calculs. Or, les SRAP sont généralement implémentés dans un contexte où plusieurs sources d'informations sont agrégées et donc, proviennent de différents genres journalistiques. Dans ce sens, la méthodologie ne pourrait être applicable qu'en considérant certains aspects concernant l'organisation de la collection dans le système et les genres qui sont recommandés.

En second lieu, l'élaboration de critères textuels à partir de l'analyse textométrique est très dépendante du corpus constitué et elle décrit bien le corpus en question. Par conséquent, les critères qu'elle permet d'élaborer sont bien adaptés pour l'application avec ce corpus, mais difficilement généralisable dans d'autres contextes. Si la méthode ne peut être pas

transposable directement dans une application, elle permet néanmoins d'observer les caractéristiques liées au corpus qui sont susceptibles d'être appliquées. Dans ce contexte, quelques possibilités s'ouvrent : la méthode pourrait être appliquée à la constitution de ressources destinées à filtrer le lexique de nouveaux articles provenant d'une controverse particulière, afin de donner une représentation optimale des données textuelles présentes. Elle pourrait être aussi appliquée au raffinement des techniques de traitement existants, par exemple, dans le choix des valeurs discriminantes sélectionnées automatiquement par les algorithmes.

L'évolution de la performance des classifieurs dans le temps en fonction de l'ajout de critères textuels provenant du calcul des spécificités démontre que l'analyse textométrique des sous-corpus contrastifs, effectués dans les périodes de pics de publication, peut améliorer le rappel et la précision dans la détection de l'opinion véhiculée par les articles. Si les spécificités démontrent une bonne valeur discriminante pour la classification automatique d'articles d'opinion, il faut considérer que cette spécificité peut aussi changer avec le temps. Pour illustrer cela, nous avons représenté dans la figure 31 la variation subie par 4 mots simples de notre corpus dans le temps, illustrant trois cas de variation des spécificités : 1) critère qui apparaît comme spécifique au début de la controverse et devient plus spécifique au fil du temps; 2) critère qui apparaît comme spécifique au début et qui devient banal au fil de temps (avec le score de spécificité plus bas que +2) ; 3) critère spécifique au début qui varie sa spécificité dans le temps, et redevient spécifique à la fin et 4) critère qui maintient sa spécificité au fil du temps. La figure permet de voir que la stabilisation des scores de spécificité des critères textuels est aussi dépendante de la stabilisation du vocabulaire du corpus de référence.

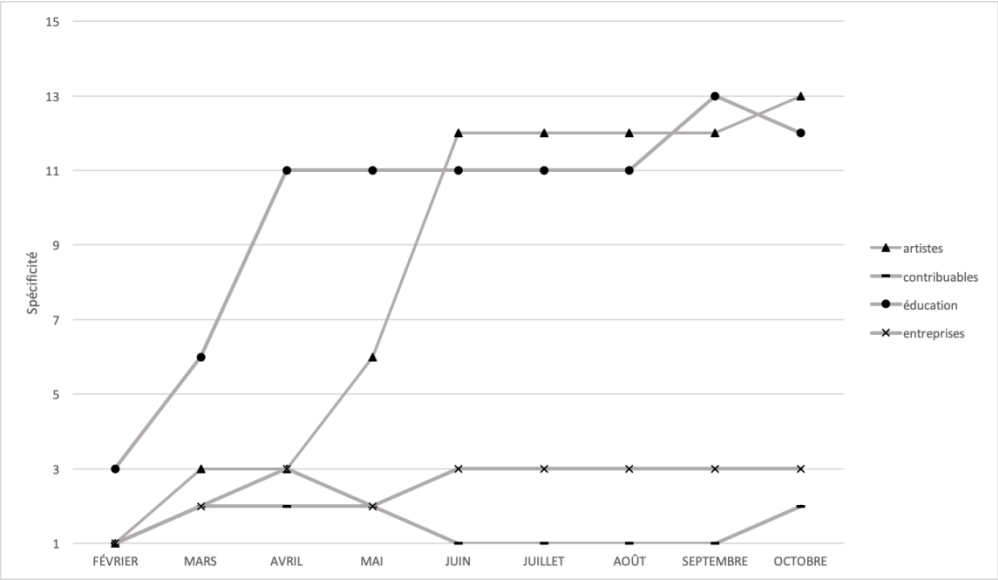


Figure 31. Évolution des spécificités de ‘artistes’, ‘contribuables’, ‘éducation’ et ‘entreprises’

Si l’analyse textométrique de sous-corpus contrastés semble à première vue fastidieuse dans le cas d’une application en temps réel, il faut se rappeler que les controverses ont une valeur documentaire non négligeable pour l’étude historique des phénomènes sociaux, et que les articles publiés à propos de controverses n’ont pas le caractère périssable associé à d’autres genres journalistiques comme les nouvelles. D’ailleurs, quelques initiatives qui exploitent la valeur documentaire des controverses pour l’étude des phénomènes sociaux ont inspiré la création de plusieurs plateformes offertes par les bibliothèques. Nous pouvons citer à ce titre la plateforme *World News Digest* de la *Princeton University*.

Dans cette recherche, nous n’avons pas problématisé les aspects plus ciblés des systèmes de recommandation, comme la modélisation du profil d’usager, ni les techniques permettant de recommander des articles qui s’opposent aux patrons de lecture d’un usager donné. Par la proposition d’une démarche méthodologique de fouille d’opinions basée sur la sélection de critères textométriques, nous avons étudié les critères textuels les plus performants pour la tâche de classification, en les caractérisant en fonction d’une théorie sémantique.

Conclusion

Cette recherche vise à donner une contribution au problème de la diversification des systèmes de recommandation d'articles de presse, et explore les apports scientifiques de champs disciplinaires divers comme les sciences de l'information, les études en journalisme, les théories discursives et sémantiques, la fouille d'opinions et la textométrie. Ces différents sujets ne sont pas épuisés dans le cadre de la recherche, mais dialoguent sur le plan théorique et méthodologique dans la poursuite d'un objectif commun, qui est de dépasser l'émergence d'un savoir basé sur une simple exploitation automatisée de données numériques, cherchant à soumettre ces données à des questionnements multiples qui peuvent nourrir et enrichir les modèles scientifiques proposés.

Le but de ce chapitre est de donner un bilan de la recherche réalisée. Dans un premier temps, nous présentons un résumé, en mettant en évidence les résultats obtenus pour chacune des questions posées. Par la suite, nous présentons les contributions de la recherche pour le champ des sciences de l'information, en soulignant les aspects théoriques et méthodologiques qui ont été travaillés. Le chapitre se clôt avec un ensemble de recommandations pour les recherches futures.

1. Résumé de la recherche

Nous avons posé une question centrale dans le cadre de cette recherche qui est de savoir comment il est possible de distinguer automatiquement des articles d'opinion qui sont similaires sur le plan thématique, mais qui sont différents par rapport à l'opinion qu'ils véhiculent. Nous avons également proposé d'explorer la pertinence de cette méthodologie comme solution pour le problème de la diversification des systèmes de recommandation d'articles de presse (SRAP), qui présente des articles véhiculant des opinions divergentes à propos d'un thème spécifique que l'usager de ce système cherche ou consulte. À travers l'élaboration d'une expérimentation utilisant un corpus d'articles d'opinion provenant de la discussion sur la grève étudiante au Québec en 2012, nous avons proposé l'objectif général de *systématiser et de valider une démarche méthodologique de fouille d'opinions basée sur la textométrie pour l'identification de textes véhiculant des opinions divergentes dans une controverse et d'explorer, par une analyse de la progression chronologique du vocabulaire du corpus, l'applicabilité des résultats observés dans le développement des SRAP*.

La démarche de fouille d'opinions qui fait l'objet de cette recherche a été proposée en premier par les travaux de Valette (2004) et Eensoo et Valette (2012, 2014a, 2014b, 2015) et s'inscrit dans le domaine général de la linguistique de corpus. Elle préconise l'étude préalable du corpus afin de sélectionner des critères textuels interprétables pour représenter les textes numériques dans le format vectoriel. Elle cherche à valider l'hypothèse selon laquelle l'utilisation d'un ensemble de critères textuels sélectionnés en fonction d'une analyse interprétative effectuée sur le corpus est plus performant pour la classification de textes que le choix d'un ensemble de termes discriminants sélectionnés en fonction d'un critère statistique, mais sans relation sémantique pertinente. Aussi, elle préconise la qualification sémantique des contrastes qui peuvent être observés dans des sous-corpus constitués en fonction du type d'opinion que le classifieur devra prédire, et en utilisant dans cette qualification un ensemble de concepts proposés par la sémantique interprétative pour décrire le sens textuel.

Dans le cadre de cette méthodologie, la textométrie constitue une démarche d'importance considérable, notamment l'emploi de deux calculs, le calcul des spécificités et le

calcul de cooccurrence. Les résultats fournis par ces derniers permettent d'attester les contrastes entre les sous-corpus et d'accuser les données textuelles spécifiques des sous-corpus représentant les opinions. Celles-ci sont ensuite qualifiées comme critères textuels par des opérations interprétatives permettant de décrire le mode dont chaque groupe d'auteurs a élaboré leur stratégie argumentative sur le plan des contenus abordés (thématique), de la structuration temporelle et argumentative (dialectique) et de la représentation de l'énonciation (dialogique). Ainsi, la méthodologie cherche à valider l'hypothèse qu'il existe une structuration globale imprégnée dans le texte et attestée dans le corpus qui peut être mis au profit de la classification automatique. Puisque sur le plan théorique, les textes sont une interaction entre les composantes sémantiques, la représentation des textes par un ensemble de critères textuels qui décrivent ces composantes peut être efficace pour la fouille d'opinions.

Pour atteindre l'objectif général proposé par la recherche, nous avons posé les questions suivantes :

1. Dans un corpus d'articles sur la Grève étudiante au Québec en 2012 qui a été classé préalablement par un humain dans des classes opposées, quels sont les critères textuels permettant de caractériser et de distinguer les opinions exprimées dans les textes sur le plan thématique, dialectique et dialogique ?

2. Quels types de critères issus de l'analyse textométrique entamée à question 1 sont les plus performantes pour prédire la classe des articles d'opinion lors d'un processus de classification automatique ?

3. À quel moment dans la controverse apparaissent les critères les plus performants pour la tâche de classification ?

Le travail de systématisation de la démarche méthodologique de fouille d'opinions a été effectuée par la formulation d'une méthodologie composée d'étapes précises, qui inclut l'étude préalable du corpus à partir de techniques textométriques et qui est en consonance avec les concepts et présupposés exposés dans le cadre théorique de la sémantique interprétative. Notre recherche a contribué à distinguer deux types de concepts dans la théorie (concepts structuraux et concepts descriptifs) et de les associer aux démarches spécifiques de l'analyse

textométrique. Nous avons également élaboré une liste de marqueurs linguistiques dans le but d'associer les résultats des calculs aux composantes sémantiques (thématique, dialectique et dialogique). La pertinence et la performance de la démarche méthodologique a été validée par l'expérimentation menée avec un corpus constitué de 495 articles d'opinion publiés dans la presse québécoise au sujet de la grève étudiante au Québec en 2012. Notre recherche a aussi contribué à pousser la réflexion au sujet de l'application de cette démarche, en faisant l'étude chronologique du corpus utilisé, qui visait à découvrir si la classification automatique des articles pourrait se faire de façon performante au fil du temps, en s'appuyant sur des critères textuels sélectionnés à partir d'un ensemble d'articles parus au tout début du débat dans la presse.

Les sections suivantes résument les résultats obtenus pour chacune des questions posées.

1.1 Critères textuels distinguant les opinions des groupes

Dans un premier temps, nous avons appliqué les calculs textométriques pour ressortir les contrastes présents dans les sous-corpus ETUD et GOUV, constitués dans la recherche. Le choix des critères pour donner une représentation des textes pour les algorithmes de fouille a été fait en amont de l'exploitation du corpus. La description contrastive des critères textuels rencontrés dans chaque sous-corpus appuient le présupposé de la théorie selon lequel les textes se structurent par une interaction thématique, dialectique et dialogique particulière. Les critères textuels ont été catégorisés dans chaque composante, en tenant compte d'une liste préliminaire établie par l'étude, qui associait certaines catégories syntaxiques à chacune des composantes. Cette catégorisation a été souple, afin de rester dans l'esprit de la théorie. Par exemple, la liste de marqueurs linguistiques établit que les substantifs doivent être catégorisés comme critères thématiques, mais l'analyse du contexte d'emploi de certains substantifs démontrent qu'ils sont plus descriptifs de phénomènes liés à la dialectique ou la dialogique des textes.

Les résultats obtenus par l'exploitation descriptive et interprétative du corpus ont mis en évidence l'opinion défendue par les groupes en dispute, et ont permis de comprendre comment ces derniers s'engagent dans un usage particulier du genre de l'opinion.

L'identification de domaines communs dans les deux sous-corpus a révélé l'adoption de systèmes lexicaux semblables, mais qui véhiculaient des valorisations différentes. L'utilisation de 'grève' par les étudiants et de 'boycott' par les défenseurs de la hausse est le cas le plus emblématique de ce phénomène. Aussi, la fréquence de mots grammaticaux particuliers a montré les différences entre la structuration du temps textuel et de l'argumentation ainsi que de la représentation de l'énonciateur.

Nous avons démontré la pertinence de faire une analyse préalable du corpus afin de découvrir des indices qui servent à caractériser les stratégies argumentatives qui sont effectivement employées par les groupes et qui peuvent être exploitées pour la représentation des textes dans le format vectoriel. Ce choix méthodologique ne se basait pas sur des hypothèses linguistiques *a priori*, et l'originalité de la démarche consiste à considérer la pertinence d'indices qui ne sont pas usuellement pris en compte dans la représentation de textes d'opinion dans les études du domaine, comme les signes de ponctuation et les pronoms personnels. L'analyse interprétative du corpus a détecté les phénomènes récurrents sur le plan sémantique, permettant de qualifier les critères comme pertinents pour la représentation des textes dans le format vectoriel. L'analyse a aussi révélé les critères textuels qui attestent le caractère fortement polémique de la discussion, et qui sont axés sur des références culturelles et historiques propres à la communauté où elle a été produite.

La description interprétative des spécificités rencontrées dans le corpus a aussi permis d'identifier les caractéristiques du discours polémique dans la discussion sur la grève étudiante, et de déceler les positions idéologiques, politiques et personnelles des groupes impliqués dans la discussion. L'analyse a explicité la structure actantielle qui est établie par les auteurs de chaque classe constituée, tels que décrit par Garand (1998), lorsque cet auteur caractérise le discours polémique. L'évocation de l'adversaire discursif dans les sous-corpus constitue un indice de cette structure actantielle (par exemple, 'Madame' dans ETUD et 'carrés rouges' dans GOUV). Un autre aspect polémique identifié dans nos résultats était la tentative de fausser la parole du groupe opposant, particulièrement présent dans la dialectique de GOUV. D'autres stratégies polémiques ont été révélées par l'engagement passionnel et le discours d'exaltation bien caractéristique d'ETUD. L'attaque à des personnes cibles est

présent dans les deux sous-corpus, ainsi que l'évocation d'alliés qui sont cités pour supporter leur opinion.

1.2 Types des critères les plus performants pour prédire la classe des opinions dans le débat

La pertinence et la performance de la démarche méthodologique ont été validées par l'expérimentation menée avec un corpus constitué de 495 articles d'opinion publiés dans la presse québécoise au sujet de la grève étudiante au Québec en 2012. Après la sélection et la qualification sémantique des critères textuels spécifiques rencontrées dans le corpus, la recherche a testé l'utilisation de ces derniers pour le développement d'un classifieur, permettant de prédire la classe de l'article selon l'opinion qu'il défend dans la controverse (plus favorable à ETUD ou à GOUV). L'expérimentation a voulu tester les différentes catégories de critères constitués dans le cadre de la recherche pour voir quelle catégorie était la plus performante et aussi pour voir si la considération de toutes les catégories de critères permettait d'atteindre de meilleurs résultats pour la classification automatique.

Dans le but de varier les expérimentations, nous avons aussi effectué deux autres catégorisations de critères, une en fonction du type de calcul appliquée et l'autre en fonction du type de segmentation (mots simples, lemmes et segment répétés). Ainsi, les critères globaux ont été obtenus par le biais du calcul des spécificités et de critères locaux ont été obtenus avec l'application du calcul de cooccurrence. La recherche a testé si les critères provenant de différents niveaux de contextualisation (global et local) étaient plus performants dans la classification que les critères provenant d'un seul niveau de contextualisation.

Les meilleurs résultats ont été obtenus avec le classifieur bayésien naïf, utilisant la totalité de critères globaux sélectionnés par la démarche textométrique (447 critères) et une pondération binaire pour la matrice (rappel : 0,91 précision : 0,91 ; AUC : 1,82). Les critères globaux incluent tous les critères thématiques, dialectiques et dialogiques qui ont été sélectionnés à partir du calcul des spécificités et dont le score de spécificité était égal ou plus grand que +2. Nous avons donc constaté que cette coalition est plus performante que lorsque chaque catégorie de critères est prise isolément. Cela démontre que les sous-corpus sont mieux caractérisés avec des critères provenant des différentes composantes sémantiques et lorsqu'ils

sont obtenus à partir du calcul des spécificités. Nous avons aussi constaté que les critères obtenus par le calcul de cooccurrence ne sont pas performants et ils peuvent même nuire les résultats de la classification lorsque toutes les catégories de critères sont combinées. Les résultats obtenus sont compatibles avec ceux obtenus par les recherches antérieures et plus performantes que les lignes de comparaison constituées, ce qui permet de valider le bien-fondé de la démarche.

1.3 Apparition des critères les plus performants pour la tâche de classification

La recherche a contribué à pousser la réflexion au sujet de l'application de la méthodologie de fouille d'opinions basé sur la sélection de critères textuels interprétables dans les SRAP, en faisant l'étude chronologique du corpus utilisé. Cette étude visait à découvrir si la classification automatique des articles pourrait se faire de façon performante au fil du temps, en s'appuyant sur des critères textuels sélectionnés à partir d'un ensemble d'articles parus au tout début du débat dans la presse. Les résultats révèlent que pour atteindre des résultats comparables à la meilleure performance obtenue avec les critères globaux (90 % avec pondération binaire et classifieur NB), il a été nécessaire l'atteinte d'un nombre important d'articles dans la controverse. La matrice composée avec 361 critères sélectionnés à partir de 465 articles publiés jusqu'au mois de juin (M_Juin), obtient les mêmes résultats que la matrice qui utilise la totalité de critères globaux (M_Octobre). Toutefois, l'utilisation de seulement 164 critères sélectionnés au début de la controverse permettrait d'obtenir des résultats proches avec le classifieur SVM et en utilisant la pondération binaire (82 % de rappel et précision). Les résultats sur les tests chronologiques effectués renforcent l'idée que l'approche textométrique peut contribuer à la modélisation d'un SRAP, dans la mesure où l'apprentissage sur un ensemble d'articles parus précédemment et représenté avec les critères textuels sélectionnés est efficace pour prédire des articles publiés postérieurement dans la controverse. Nous avons constaté que cette efficacité peut accroître à la mesure que la discussion gagne une certaine maturation dans les médias.

2. Contributions au domaine des sciences de l'information

2.1 Apports théoriques et méthodologiques

Nous avons effectué dans cette recherche une revue de la littérature pour comprendre d'une part comment le problème de la diversification dans les systèmes de recommandation était traité dans les études plus récentes du domaine et d'autre part pour comprendre comment l'adoption d'une démarche méthodologique de fouille d'opinions qui s'appuie sur le choix d'un cadre linguistique explicite (Eensoo et Valette, 2012 ; 2014a ; 2014b ; 2015) pourrait répondre au problème de la diversification dans le contexte de recommandation d'articles d'opinion dans les SRAP.

Certaines lacunes concernant les recherches courantes sur les systèmes de recommandation de presse qui proposent de donner une réponse appropriée au problème de la diversité ont été présentées. Le constat émergent de cette revue est que la diversification est souvent définie comme une mesure de distance statistique qui cherche à maximiser la différence entre les articles à être recommandés, en fonction des mots présents dans ces derniers. Cette différence n'est pas pourtant relative à la manière dont un être humain perçoit l'opposition entre les opinions exprimées dans débat public. Elle est inférée en fonction de traits discriminants présents, qui peuvent être un ensemble de mots statistiquement pertinents, mais sans relation explicite entre eux.

Il y a par contre des recherches dans les systèmes de recommandation qui proposent d'explorer la différence d'opinions véhiculées par les articles sur une question publique débattue c'est-à-dire, circonscrit sur un thème spécifique. Ces derniers caractérisent l'opinion véhiculée par un texte de manière réductrice, et cherchent à représenter les textes par un ensemble d'indices subjectifs, tels quels les sentiments et les émotions exprimées par les journalistes. Cette démarche s'avère moins pertinente pour les articles à teneur argumentative, puisque l'élaboration d'une opinion sur une question d'intérêt public n'est pas nécessairement liée à l'expression émotive de celui qui rédige.

Pour parvenir à combler un manque de conceptualisation de l'opinion dans les recherches de fouille d'opinions, notre étude a analysé des écrits sur l'opinion journalistique et le concept d'opinion publique, en analysant en particulier la contribution de Habermas (1991) sur l'évolution historique du concept d'opinion publique, ainsi que l'ouvrage de Jacobs et Townsley (2011). Cette dernière nous a permis de dresser un panorama sur comment la discussion publique contemporaine se configure dans un espace médiatique de plus en plus diversifié. Dans notre recherche, l'opinion a été définie comme l'expression linguistique d'un acte communicationnel, liée à une pratique institutionnelle et que se configure comme une critique envers le pouvoir constitué. Cette définition exclut les conceptions qui tentent de réduire l'expression linguistique de l'opinion comme un ensemble de phrases évaluatives à propos d'un objet. Elle implique la considération de facteurs liés à la situation de communication qui influencent l'élaboration linguistique de l'opinion et qui peuvent être repérés par des indices textuels autres que le seul vocabulaire subjectif ou émotif : la manière de cadrer la discussion, le choix de thèmes abordés, la proximité ou la mise à distance que le locuteur établi avec son interlocuteur reflètent les choix linguistique des auteurs et peuvent être pourtant des sources pour découvrir les indices permettant de déterminer à quel type d'opinion le locuteur se rallie à l'intérieur d'une discussion publique et par quelle stratégie il cherche à convaincre son lectorat sur son opinion.

En abordant le concept de controverse, la revue de la littérature a montré que la diffusion d'opinions dans la presse est souvent menée sous la forme d'une discussion entre des groupes qui défendent des positions contraires et qui élaborent des stratégies visant à invalider l'argument l'un de l'autre. Les stratégies argumentatives utilisées par ces groupes rappellent dans plusieurs aspects le discours polémique, tel qu'il est caractérisé par les auteurs abordés dans la littérature (Amossy, 2014 ; Charaudeau, 2017 ; Garand, 1998) ; Kerbrat-Orecchioni, 1980).

Dans le but de dépasser la vision de l'expression linguistique de l'opinion dans laquelle se basent les recherches courantes et aussi de connaître les méthodes pour repérer d'autres indices linguistiques pertinents, notre recherche a exploré des études sur les genres textuels et sur les théories discursives. D'une part, le genre de l'opinion a été caractérisé en fonction de l'enjeu persuasif qui est posé par la situation de communication spécifique où il est utilisé, et

dans laquelle une question est mise à l'assentiment d'un public, amené à débattre et aussi à prendre position. D'autre part, la recherche a abordé l'approche de l'analyse du discours, qui propose d'étudier l'incidence de l'énonciation (en tant que représentation discursive du locuteur) sur le sens textuel à l'intérieur d'une situation de communication déterminée.

L'analyse de la littérature sur l'analyse du discours nous a amené à constater que cette dernière préconise l'étude du sens textuel à partir du repérage d'un ensemble de marques linguistiques qui peuvent mettre en évidence la manière dont le locuteur s'inscrit dans le discours. Dans le cas d'une situation où il est enjeux de la persuasion, certaines marques peuvent aider à comprendre comment le locuteur cherche à convaincre son interlocuteur pour lui faire adhérer à une opinion. Les choix linguistiques qui sont explicités par l'emploi de ces marques peuvent faire apparaître aussi les systèmes de valeurs et de représentations du locuteur. La recherche a évoqué les études de Charaudeau (1992), qui propose une syntaxe discursive, dans laquelle ces marques peuvent être associées à des objets syntaxiques déterminés, et associés aux intentionnalités des locuteurs.

L'adoption de l'appareil méthodologique proposée par l'analyse du discours se confronte au choix de bonnes marques pour caractériser l'expression de l'opinion et demande une caractérisation préalable du genre textuel. À cette dernière vision, nous avons contrasté la perspective de la sémantique interprétative de François Rastier (1987, 1993, 2001) qui est adoptée dans les travaux de Valette (2004) et Eensoo et Valette (2012, 2014a, 2014b, 2015). Elle préconise une approche non référentielle du langage et l'adoption d'une méthodologie pour étudier le sens textuel qui procède par une description systématique des textes en fonction des genres textuels et les corpus. Elle ne cherche pas à décrire le sens textuel par un ensemble de marques qui sont corrélées à des intentions a priori, ni à caractériser le genre par un univers clôt de marques linguistiques. Elle cherche au contraire à caractériser les signifiants d'un texte en fonction de récurrences sémantiques qui peuvent être observées lorsque celui-ci est comparé à d'autres textes, en tentant compte de l'espace de contraintes que le genre textuel, et le corpus impose sur l'interprétation. Ainsi, les bonnes « marques » pour caractériser l'empreinte de l'énonciateur et les stratégies mises en place sont déterminées a posteriori, en fonction d'un processus d'interprétation qui prend le genre textuel et le corpus comme ancrage de ce processus.

Le problème de recherche que nous avons posé concerne une partie très spécifique du processus de recommandation basé sur le contenu. Dans le modèle de filtrage d'information proposé par Belkin et Croft (1992, p. 31) et présenté dans la revue de la littérature, nous avons cerné cette partie, qui correspond à la composante « représentation » dans le modèle vectoriel. C'est cette représentation qui permet le système de recommandation de détecter la similarité entre les documents à recommander et les profils d'usagers, qui sont eux aussi représentés par un ensemble de termes extraits des documents qu'ils ont consultés auparavant. Avec modèle d'information décisionnelle ERIn proposé par Furner (2002), nous avons vu que la modélisation du fonctionnement d'un système de recommandation se base sur des décisions conscientes des concepteurs et développeurs de ces systèmes. Le choix de représenter les documents par un ensemble de termes statistiquement importants extraits de ces derniers est un choix qui apporte un jugement sur la pertinence de ces termes pour générer un ordre de préférence relativement aux documents à être proposé par le système, suite à une recherche ou à un processus de filtrage.

Cet aspect théorique permet d'intégrer le travail réalisé dans cette recherche à l'intérieur d'un domaine plus large dans les sciences de l'information qui a trait à la l'indexation, et à l'utilisation du langage pour décrire les documents et pour en faciliter l'accès et le repérage. Dans le champ des sciences de l'information, l'indexation a trait aux formes de représentation abstraites du contenu des documents et qui servent à les organiser dans les catalogues et dans les bases de données. La représentation des documents s'appuie sur une structure formalisée (système de classification, thésaurus) qui reflète les relations entre les concepts d'un champ disciplinaire particulier (Hjørland, 2013). Les termes d'indexation, les descripteurs et les concepts constituent des points d'accès aux documents et fonctionnent comme éléments de base pour la conception de ces structures d'organisation.

L'action de regrouper les documents semblables, autant pour les organiser dans des systèmes informatisés ou physiques, est le résultat d'une décision de nature intellectuelle et qui relève d'un questionnement scientifique. Elle doit aussi considérer des aspects qui sont importants pour usagers finaux, les personnes qui auront accès aux documents. Par exemple, la décision de classer un ouvrage sur « Pshychologie » dans le domaine des sciences humaines ou dans les sciences naturelles est liée aux discussions épistémologiques de ce domaine de

connaissances et à des choix paradigmatiques du champ scientifique (Hjørland, 2013). : devons-nous étudier la psyché humaine en mesurant des faits observables (positionnement empiriciste, en sciences naturelles) ou à partir de l'analyse des facteurs historiques et culturels de l'individu (positionnement historiciste et herméneutique, en sciences humaines) ? Dans quelle mesure ce choix affecte-t-il l'efficacité sur le repérage de ce document par une communauté d'usagers dans un système d'organisation de documents ?

La représentation d'un ensemble de textes numériques dans le format vectoriel pour faciliter son repérage relève également d'une décision intellectuelle qui se rattache à un paradigme scientifique. Nous avons vu dans cette recherche que les études sur la fouille d'opinions cherchent à choisir les termes discriminants d'une opinion en fonction de constats sémantiques associés à certains genres. Ainsi, la proposition de caractériser l'opinion d'un article par la fréquence de mots positifs ou négatifs, relève d'un constat empirique observé dans certains genres textuels comme les critiques sur les produits : une opinion négative est souvent exprimée avec certains termes d'axiologie négative (mauvais) tandis qu'une opinion positive est exprimée au contraire avec des termes positifs (bon). Il s'agit d'une vision logiciste du fonctionnement de langue qui présuppose l'invariabilité sémantique des mots (Eensoo et coll., 2011). Notre travail a remis en question cette vision du langage, en proposant d'envisager l'opinion comme l'expression d'un acte communicationnel, lié à un enjeu argumentatif et qui établit une structure actantielle qui peut être étudiée, analysé et caractérisé par plusieurs indices textuels. Le cadre théorique choisi définit le sens textuel à partir d'une perspective interprétative, en soumettant le choix des « termes d'indexation » par une contextualisation qui est fait au niveau du corpus (choix du corpus de référence et des sous-corpus en fonction des opinions d'une controverse) et par la caractérisation contrastive de ces derniers, dans le but de dégager les différences significatives sur le plan de l'élaboration thématique, dialectique et dialogique.

En ce qui concerne l'application de la méthodologie dans le cadre des systèmes de recommandation d'articles de presse, notre recherche a contribué à découvrir que l'utilisation de critères sélectionnés au début d'une controverse est efficace pour représenter les textes dans le format vectoriel. Dans une perspective prospective, ces critères peuvent servir à entraîner des classifieurs pour prédire les articles à venir dans la controverse, tout en explicitant les

raisons linguistiques de leur efficacité. Nous avons confirmé les découvertes effectuées par des travaux antécédents puisque nous avons démontré que l'ensemble de critères textuels thématiques, dialectiques et dialogiques sélectionnés par l'approche interprétative sont efficaces pour la fouille d'opinons. Les résultats indiquent que l'attention portée par les médias à propos d'une controverse constitue une opportunité pour élaborer les critères textuels qui seront performants pour la prédiction d'articles.

3. Recherches futures

Notre recherche a comparé la démarche de sélection de critères textuels interprétables avec une représentation simple du corpus de textes servant comme ligne de comparaison. Ce dernier a été constitué avec un ensemble de termes extraits du corpus, sans aucun traitement spécial. D'autres études pourraient notamment explorer le bien-fondé de la méthode de fouille d'opinions que nous préconisons, en la comparant avec d'autres approches adoptées dans les travaux revus, telles comme le filtrage du vocabulaire avec la sélection de traits discriminants par des mesures statistiques, l'utilisation de dictionnaire de sentiments, ou encore, le choix de représenter les textes par l'extraction des phrases évaluatives présentes.

Plusieurs facteurs sont aussi à considérer pour l'application de la méthode présentée ici dans le contexte d'un SRAP. D'abord, il faut que les documents de la collection dans les systèmes soient organisés par thématiques, en fonction de différentes controverses. En deuxième lieu, nous avons envisagé un scénario de recommandation dans lequel le document que l'usager consulte puisse constituer la base sur laquelle la recommandation est faite. Ainsi, dans le cas où l'usager lit un article sur la grève étudiante classé comme ETUD, le système lui proposerait des articles classés comme GOUV, et qui sont aussi publiés dans à peu près la même période. Nous n'avons pas dans cette recherche élaboré le fonctionnement de ce système de façon détaillée, mais il n'en demeure pas moins que les résultats obtenus permettent de voir la pertinence de considérer une approche linguistique pour aider à obtenir une représentation optimale des textes numériques, considérant le cas particulier des controverses médiatiques. Il faut considérer en tout cas les questions liées à l'automatisation de ce processus dans les systèmes de recommandation. Les résultats de la recherche laissent entrevoir qu'un ensemble de critères de spécificité plus grande que +2 obtenus par le calcul des spécificités, et qui caractérisent les composantes sémantiques que nous avons étudié, est efficace pour la classification automatique. Cela indique que la création d'un processus semi-automatisé pour représenter les articles d'une controverse peut être efficace dans le contexte d'un système de recommandation, mais cette hypothèse est encore à explorer.

Bibliographie

Abbar, S., S. Amer-Yahia, P. Indyk, et S. Mahabadi. « Real-Time Recommendation of Diverse Related Articles. » In *WWW '13 Proceedings of the 22nd International Conference on World Wide Web in Rio de Janeiro, Brazil, May 13 - 17, 2013*, 1–12. New York, NY, USA : ACM, 2013. doi: 10.1145/2488388.2488390.

Acosta, A., et A. Bittar. 2007 [en ligne]. « La gratounette : classification automatique générique de textes d'opinion. » In *Actes du troisième Défi Fouille de Textes, DEFT2007, Grenoble, France, 3 juillet 2007*, 23–34. Disponible sur : https://deft.limsi.fr/actes/actes_deft2007.pdf. Consulté le15 août 2019.

Adam, J.-M. 2008. *La linguistique textuelle : introduction à l'analyse textuelle du discours*. 3e édition. Paris: Armand Colin.

Ahn, J., P. Brusilovsky, G. Jonathan, D. He, et S.-Y. Syn. 2007. « Open User Profiles for Adaptive News Systems: Help or Harm? » In *Proceedings of the 16th International Conference on World Wide Web*, édité par C.L. Williamson, M.E. Zurko, P.J. Patel-Schneider, et P.J Shenoy, 11–20. New York : ACM.

Amossy, R. 2014. *Apologie de la polémique*. Paris : Presses universitaires de France.

Amossy, R. et R. Koren. 2009 [en ligne]. « Rhétorique et argumentation : approches croisées. » *Argumentation et analyse du discours* 2. Disponible sur : http://journals.openedition.org/aad/561. Consulté le 7 janvier 2019.

Andersen, J. 2008. « The Concept of Genre in Information Studies. » *Annual Review of Information Science and Technology* 42 (1): 339–67.

Anscombre, J.-C., et O. Ducrot. 1983. *L'argumentation dans la langue*. Mardaga. Philosophie et Langage. Bruxelles.

Aristote. 1991. *Rhétorique*. Paris : Le Livre de Poche.

Bachimont, B. 2000. « Engagement sémantique et engagement ontologique : Conception et réalisation d'ontologies en ingénierie des connaissances. » In *Ingénierie des*

connaissances, évolutions récentes et nouveaux défis., édité par J. Charlet, M. Zacklad, G. Kassel, et D. Bourigault, 305–24. Paris : Eyrolles.

Bakhtin, M. 1986. « The Problem of Speech Genres. » In *Speech Genres and Other Late Essays*, édité par V. W. McGee, C. Emerson, et M. Holquist, 60–102. Austin : University of Texas Press.

Bakshy, E., S. Messing, et L.A. Adamic. 2015. « Exposure to Diverse Information on Facebook. » *Science* 348 (6239) : 1130–33.

Balahur, A., et R. Steinberger. 2012. « Rethinking Sentiment Analysis in the News: From Theory to Practice and Back. » In *Proceedings of the 1st Workshop on Opinion Mining and Sentiment Analysis, Sevilla, Espagne*, 1–12.

Balahur, A., R. Steinberger, M. Kabadjov, V. Zavarella, E. Goot, M. Halkia, B. Pouliquen, et E. Belyaeva. 2013. « Sentiment Analysis in the News Alexandra. » In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010)*, 2216–20. Valletta, Malta.

Ballabriga, M. 2005 [en ligne]. « Sémantique textuelle. » Texto ! Mars 2005. Disponible sur : http://www.revue-texto.net/Reperes/Cours/Ballabriga2/index.html. Consulté le 7 janvier 2019.

*Bally, C. 1951. *Traité de stylistique française*. 3e édition. Vol. 1. Paris : Klincksieck.

Belghanem, A. 2009 [en ligne]. « La Sémantique interprétative : du mot au corpus et du sème aux formes sémantiques ». *Texto ! Textes et Culture* XIX (1). Disponible sur : http://www.revue-texto.net/docannexe/file/3434/belghanem_seminterpretative.pdf. Consulté le 7 janvier 2019.

Belkin, N. et B. Croft. 1992. « Information Filtering and Information Retrieval: Two Sides of the Same Coin? » *Communications of the ACM* 35 (12): 29–38.

Belkin, N.J. 2000. « Helping People Find What They Don't Know ». *Communications of the ACM* 43 (8): 58–61.

Benveniste, E. 1967. *Problèmes de linguistique générale*. Tome.1. Paris : Gallimard.

Bethard, S., U. Hong, A. Thornton, V. Hatzivassiloglou, et D. Jurafsky. 2004. « Automatic Extraction of Opinion Propositions and Their Holders. » In *2004 AAAI Spring Symposium on Exploring Attitude and Affect in Text, Palo Alto, California, January 2004*.

Blei, D.M., A.Y. Ng et M. I. Jordan. 2003. « Latent Dirichlet Allocation ». *Journal of Machine Learning Research* 3 : 993–1022.

Blondiaux, L. 1998. *La fabrique de l'opinion. Une histoire sociale des sondages*. Paris : Le Seuil.

Blouin Genest, G. 2012a. « Le conflit étudiant québécois : une « épidémie » de sens pour un Québec politiquement malade ». » *Cultures & Conflits* 3 (87) : 147–151.

Blouin Genest, G. 2012b. « Le (dé) goût d'un printemps : la construction sociale de la violence et de l'extrémisme politique lors du conflit étudiant québécois. » *Cultures & Conflits* 3 (87) : 160–166.

Bonenfant, M., A. Glinoer, et M.-E. Lapointe. 2013. *Le printemps québécois : une anthologie*. Les Éditions Écosociété. Montréal (Québec).

Bourdieu, P. 1973. « L'opinion publique n'existe pas. » *Les temps modernes* 29 (318) : 1292–1309.

Bourion, E. 2001 [en ligne]. « L'aide à l'interprétation des textes électroniques. » Thèse de doctorat, Université de Nancy II. Disponible sur : http://www.revue-texto.net/1996-2007/Corpus/Publications/Bourion/Bourion_Aide.html. Consulté le15 août 2019.

Brunet, E. 2000 [en ligne]. « Qui lemmatise dilemme attise ». *Lexicometrica 2*. Disponible sur : http://lexicometrica.univ-paris3.fr/article/numero2/brunet2000.PDF. Consulté le15 août 2019.

Brunet, E. 2009. *Comptes d'auteurs. Études statistiques de Rabelais à Gracq*. Paris : Champion.

Bureau de la Traduction du Canada. s.d. [en ligne] *Banque de données terminologiques et linguistiques du Canada (Terminus)*. Disponible sur : http://www.btb.termiumplus.gc.ca/. Consulté le 15 août 2019.

Burke, R. 2002. « Hybrid Recommender Systems: Survey and Experiments. » *User Modeling and User-Adapted Interaction*, 4:331–70. 12.

Chandler, D. 1997. « An Introduction to Genre Theory. » *The Media and Communications Studies Site*. Disponible sur : http://www.aber.ac.uk/media/Documents/intgenre/chandler_genre_theory.pdf. Consulté le 7 janvier 2019.

Chaovalit, P., et L. Zhou. 2005. « Movie Review Mining : A Comparison between Supervised and Unsupervised Classification Approaches. » In *Proceedings of the 38th Annual Hawaii International Conference on System Science, Big Island, HI, 6 January 2005*. IEEE. doi : 10.1109/HICSS.2005.445.

Charaudeau, P. 1992. *Grammaire du sens et de l'expression*. Hachette. Paris.

Charaudeau, P. 1994 [en ligne]. « Les conditions de compréhension du sens de discours. » *Langage en FLE Texte et Compréhension, Revue ICI et LÀ*. Disponible sur : http://www.patrick-charaudeau.com/Les-conditions-de-comprehension-du,62.html. Consulté le 21 novembre 2016.

Charaudeau, P. 1995. « Une analyse sémiolinguistique du discours. » *Langages*, no. 117 : 96–111.

Charaudeau, P. 2006a [en ligne]. « Discours journalistique et positionnements énonciatifs. Frontières et dérives ». *Semen* 22. Disponible sur : http://semen.revues.org/2793. Consulté le 7 janvier 2019.

Charaudeau, P. 2006b. « Un modèle socio-communicationnel du discours. Entre situation de communication et stratégies d'individuation ». *Médias et Culture. Discours, outils de communication, pratiques : quelle(s) pragmatique (s) ?* 15–40. Paris : L'Harmattan.

Charaudeau, P. 2007 [en ligne]. « De l'argumentation entre les visées d'influence de la situation de Communication ». *Argumentation, Manipulation, Persuasion*. Disponible sur : http://www.patrick-charaudeau.com/De-l-argumentation-entre-les.html. Consulté le 7 janvier 2019.

Charaudeau, P. 2017. *Le débat public : entre controverse et polémique, enjeu de vérité, enjeu de pouvoir*. Limoges : Lambert-Lucas.

Charton, E., et R. Acuna-Agost. 2007. "Quel modèle pour détecter une opinion ? Trois propositions pour généraliser l'extraction d'une idée dans un corpus." In *Actes du Troisième Défi Fouille de Textes, Grenoble, France*, 35–50.

Chu, W., et S.-T. Park. 2009. « Personalized Recommendation on Dynamic Content Using Predictive Bilinear Models. » In *WWW '09 Proceedings of the 18th International Conference on World Wide Web, Madrid, Spain*, 691–700. ACM Press.

Dave, K., S. Lawrence, et D.M. Pennock. 2002. « Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. » In *WWW '03*

Proceedings of the 12th International Conference on World Wide Web, Budapest, Hungary, May 20 - 24, 2003, 519–28. New York : ACM.

Desarkar, M., et N. Shinde. 2014. « Diversification in News Recommendation for Privacy Concerned Users. » In *Proceedings of the 2014 International Conference on Data Science and Advanced Analytics, Shanghai, China, 30 October - 2 November*, 135–141. IEEE.

Dick, Archie L. 1999. « Epistemological Positions and Library and Information Science. » *Library Quarterly* 69 (3) : 305–23.

Dubois, J., M. Giacomo, L. Guespin, C. Marcellesi, J.-B. Marcellesi et J.-P. Mével. 2007. *Grand Dictionnaire linguistique & Sciences du Langage*. Paris : Larousse.

Ducrot, O., et T. Todorov. 1972. *Dictionnaire encyclopédique des sciences du langage*. Paris : Éditions du Seuil.

Eensoo, E., D. Nouvel, A. Martin et M. Valette. 2015. « Combiner analyses textométriques, apprentissage supervisé et représentation vectorielle pour l'analyse de la subjectivité. » *Actes du 11e Défi Fouille de Textes (DEFT'2015), Caen (France)*. Disponible sur : http://www.atala.org/taln_archives/ateliers/2015/DEFT/deft-2015-long-010.pdf. Consulté le 7 janvier 2019.

Eensoo, E., E. Bourion, M. Slodzian et M. Valette. 2011. « De la fouille de données à la fabrique de l'opinion. Enjeux épistémologiques et propositions. » *Les Cahiers du Numérique* 7 (2) : 15–36.

Eensoo, E., et M. Valette. 2012. « Sur l'application de méthodes textométriques à la construction de critères de classification en analyse des sentiments. » *Actes de la conférence conjointe JEP-TALN-RECITAL*, Volume 2 : TALN :367–74. Grenoble, France : ATALA & AFCP.

Eensoo, E., et M. Valette. 2014a. « Approche textuelle pour le traitement automatique du discours évaluatif. » *Langue française* 184 : 107–22.

Eensoo, E., et M. Valette. 2014b. « Sémantique textuelle et TAL : un exemple d'application à l'analyse des sentiments ». *Documents, textes, œuvres. Perspectives sémiotiques*. Collection rivages linguistiques : 75–89. Presses Universitaires de Rennes

Eensoo, E., et M. Valette. 2015. « Une méthodologie de sémantique de corpus appliquée à des tâches de fouille d'opinion et d'analyse des sentiments : étude sur l'impact de

marqueurs dialogiques et dialectiques dans l'expression de la subjectivité. » *Texto! Textes et Culture* XX (3) : 2015.

Ekstrand, M.D., F.M. Harper, M.C. Willemsen, et J.A. Konstan. 2014. « User Perception of Differences in Recommender Algorithms. » In *Proceedings of the 8th Conference on Recommender Systems (RecSys'14), Silicon Valley, CA, USA, October 06 - 10, 2014*, 161–168. New York: ACM Press.

Emediato, Q. 2011. « L'argumentation dans le discours d'information médiatique. » Édité par Ida Lucia Machado et Emilia Mendes. *Argumentation et analyse du discours*, Approches de l'AD et de l'argumentation au Brésil, n°7 (Octobre). Disponible sur : http://aad.revues.org/1209. Consulté le 7 janvier 2019.

ENS de Lyon, et Université de Franche-Comté. 2017 [en ligne]. « Manuel de TXM. (Version 0.7.8). » Disponible sur : http://textometrie.ens-lyon.fr/files/software/TXM/0.7.8/. Consulté le 7 janvier 2019.

Esuli, A., et F. Sebastiani. 2006. « SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. » In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, 417–22. European Language Resources Association (ELRA).

Flament, D. 2006. « L'entrée thème/rhème du glossaire de Comenius. » *Revue des linguistes de l'Université Paris Ouest Nanterre La Défense* 55 : 61–71.

Forest, D. 2009. « Vers une nouvelle génération d'outils d'analyse et de recherche d'informations. » *Documentation et Bibliothèque* 55 (3) : 77–89.

Forest, D. et J-G. Meunier. 2004. « Classification et catégorisation automatiques : application à l'analyse thématique de données textuelles. » *JADT 2004 : 7 Journées internationales d'analyse statistique de données textuelles*, 434–44.

Forest, D., A. Hoeydonck, D. Létourneau, et M. Bélanger. 2009. "Impacts de la variation du nombre de traits discriminants sur la catégorisation des documents." In *Actes du Cinquième Défi Fouille de Textes, DEFT2009, Paris, France, 22 Juin 2009.*, 77–90.

Fortin, M.-F. 2010. *Fondements et étapes du processus de recherche : méthodes quantitatives et qualitatives*. 2e édition. Chenelière Éducation.

Fortuna, B., C. Galleguillos, et N. Cristianini. 2009. “Detecting the Bias in Media with Statistical Learning Methods.” In *Text Mining: Classification, Clustering and Applications*, 27–50. Text Mining: Theory and Applications. USA : Chapman & Hall/CRC.

Fraser, N. 1990. « Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy. » *Social Text*, n°. 25/26 : 56–80.

Furner, J. 2002. « On Recommending. » *Journal of the American Society for Information Science* 53 (9): 747–63.

Garand, A. 1998. « Propositions méthodologiques pour l’étude du polémique. » In *États du polémique*, édité par A. Hayward et D. Garand, 22 :211–68. Les cahiers du centre de recherche en littérature québécoise. Québec : Éditions Nota Bene.

Ge, M., C. Delgado-Battenfeld, et D. Jannach. 2010. « Beyond Accuracy: Evaluating Recommender Systems by Coverage and Serendipity. » In *RecSys ’10 Proceedings of the Fourth ACM Conference on Recommender Systems, Barcelona, Spain, September 26 - 30, 2010*, 257–60. New York : ACM. doi:.1145/1864708.1864761.

Gemmis, M., P. Lops, G. Semeraro et C. Musto. 2015. « An Investigation on the Serendipity Problem in Recommender Systems. » *Information Processing & Management* 51 (5) : 695–717.

Gérard, C. 2004. « Contribution à une sémantique interprétative des styles : étude de deux œuvres de la modernité poétique : Jacques Dupin et Gérard Macé. » Thèse de doctorat, Université de Toulouse Le Mirail.

Goodnight, T. 1991. « Controversy. » In *Argument in Controversy: Proceedings of the Seventh SCA/AFA Conference on Argumentation*, édité par D.W. Parson, 1–13. Annandal : Speech Communication Association.

Govier, T. 1999. *The Philosophy of Argument*. Newport News (VA) : Vale Press.

Grosse, E. U. 2001 [en ligne]. « Évolution et typologie des genres journalistiques ». *Semen - Revue sémio-linguistique des textes et discours* 13. Disponible sur : http://semen.revues.org/2615?lang=en. Consulté en 7 janvier 2019.

Grouin, C., B. Arnulphy, J.-B. Berthelin, S. El Ayari, A. García-Fernandez, A. Grappy, M. Hurault-Plantet, P. Paroubek, I. Robba et P. Zweigenbaum. 2009. « Présentation de

l'édition 2009 du Défi Fouille de Textes (DEFT'09). » *Actes du Cinquième Défi Fouille de Textes, DEFT2009, Paris*, 35–50.

Guaresi, M. 2016. « Cooccurrences, contrastes et caractérisation textuels. Applications à un corpus de professions de foi électorales (1958 -2007). » In *13th International Conference on Statistical Analysis of Textual Data, Nice, France, 7-10 Juin 2016*, 439–51.

Habermas, J. 1991. *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. Cambridge, MA. : MIT Press.

Habermas, J. 1996. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. Studies in Contemporary German Social Thought. Cambridge, MA : MIT Press.

Hall, M., E. Frank, G. Holmes, P. Bernhard, P. Reutemann, et I. H. Witten. 2009. « The WEKA Data Mining Software: An Update. » *SIGKDD Explorations* 11 (1).

He, H., et E.A. Garcia. 2009. « Learning from Imbalanced Data. » *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING* 21 (9) : 1263–85.

Hébert, L. 2001. *Introduction à la sémantique des textes*. Paris : Éditions Champions.

Heiden, S., J.-P. Magué et B. Pincemin. 2010. « TXM : une plateforme logicielle open-source pour la textométrie – conception et développement. » *10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*, 2 :1021–32. Rome : Edizioni Universitarie di Lettere Economia Diritto.

Herlocker, J.L., J.A. Konstan, L.G. Terveen, et J.T. Riedl. 2004. « Evaluating Collaborative Filtering Recommender Systems. » In *ACM Transactions on Information Systems (TOIS)*, 168–77.

Hermet, G., B. Badie, P. Birnbaum, et P. Braud. 2000. « Opinion publique. » *Dictionnaire de la science politique et des institutions politiques*. Paris : Colin.

Hjørland, B. (2013). « User-based and Cognitive Approaches to Knowledge Organization: A Theoretical Analysis of the Research Literature. » *Knowledge Organization*, 40 (1), 11–26.

Hu, M., et B. Liu. 2004. « Mining and Summarizing Customer Reviews. » In *KDD '04 Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge*

Discovery and Data Mining, Seattle, WA, USA, August 22 - 25, 2004, 168–77. New York : ACM.

Ibekwe-SanJuan, F. 2007. *Fouille de textes : méthodes, outils et applications*. Paris : Lavoisier.

IJntema, W., F. Goossen, F. Frasincar, et F. Hogenboom. 2010. « Ontology-Based News Recommendation. » In *Proceedings of the 2010 EDBT/ICDT Workshops, Lausanne, Switzerland, March 22-26, 2010*. ACM International Conference Proceeding Series. New York : ACM. doi: 10.1145/1754239.1754257.

Jacobs, R.N. et E. Townsley. 2011. *The Space of Opinion - Media Intellectuals and the Public Sphere*. New York : Oxford.

Jacques, M.-P. 2005. « Pourquoi une linguistique de corpus ? » *La linguistique de corpus*, Rivages linguistiques : 21–29.

Jannach, D., M. Zanker, A. Felfernig et G. Friedrich. 2011. *Recommender Systems: An Introduction*. Cambridge University Press.

Jindal, N., et Liu Bing. 2006. « Identifying Comparative Sentences in Text Documents. » In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006*, 244–51. New York : ACM. doi: 10.1145/1148170.1148215.

Julien, F. 2012. « Le printemps érable comme choc idéologique. » *Cultures & Conflits* 3 (87) : 152 – 159.

Kaiser, C., S. Schlick, et F. Bodendorf. 2011. « Warning System for Online Market Research -- Identifying Critical Situations in Online Opinion Formation. » *Knowledge-Based Systems* 24 (6): 824–36.

Karimi, M., D. Jannach et M. Jugovac. 2018. « News Recommender Systems – Survey and Roads Ahead. » *Information Processing and Management*, no. 58: 1203–27.

Kawai, Yukiko, Tadahiko Kumamoto et Katsumi Tanaka. 2007. « Fair News Reader: Recommending News Articles with Different Sentiments Based on User Preference ». *Computer Science*, 4692 : 612–622.

Kellner, D. 2014. "Habermas, the Public Sphere, and Democracy." In *Re-Imagining Public Space: The Frankfurt School in the 21st Century*, édité par D. Boros et J. Glass, 19–43. Palgrave Macmillan US.

Kerbrat-Orecchioni, C. 1980. *Le discours polémique*. Lyon : Presses universitaires de Lyon.

Lafon, P. 1980. « Sur la variabilité de la fréquence des formes dans un corpus. » *Mots* 1 : 127–65.

Lafon, P. 1981. « Analyse lexicométrique et recherche des cooccurrences. » *Mots* 3 : 95–148.

Lafon, P. 1984. *Dépouillements et statistiques en lexicométrie*. Genève-Paris : Slatkine – Champion.

Langlois, P. 2012. « Révolte contre le néolibéralisme, riposte contre la liberté d'association. » *Cultures & Conflits* 3 (87) : 167 – 173.

Latour, B. 2006. *Changer de société, refaire la sociologie*. Éditions La Découverte.

Laurendeau, P., et R. Delamotte-Legrand. 2004. « Modalité, Opération de Modalisation et Mode Médiatif. » In *Des faits de langue aux discours*, 1:83–95. Les médiations langagières. Rouen : Publications de l'Université de Rouen.

Lebart, L. et A. Salem. 1994. *Statistique textuelle*. Paris : Dunod.

Lee, Y., S. Na, J. Kim, S. Nam, H. Jung, et J. Lee. 2008. « KLE at TREC 2008 Blog Track : Blog Post and Feed Retrieval. » In *Conference: Proceedings of The Seventeenth Text Retrieval Conference, TREC 2008, Gaithersburg, Maryland, USA, November 18-21, 2008*. National Institute of Standards and Technology.

Lei, L., W. Ding-Ding, Z. Shun-Zhi et L. Tao. 2011. « Personalized News Recommendation: A Review and an Experimental Investigation ». *Journal of Computer Science and Technology* 26 (5): 754–766.

Lemieux, C. 2007. « À quoi sert l'analyse des controverses ? » *Mil neuf cents. Revue d'histoire intellectuelle* 1 (25) : 191–212.

Li, L., et T. Li. 2013. « News Recommendation via Hypergraph Learning: Encapsulation Ofuser Behavior and News Content. » In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, 305–314. New York : ACM. doi : 10.1145/2433396.2433436.

Li, L., T. Li, D. Knox, et B. Padmanabhan. 2011. « SCENE : A Scalable Two-Stage Personalized News Recommendation System. » In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, 125–134. New York : ACM. doi : 10.1145/2009916.2009937.

Liu, B. 2001. *Web Data Mining : Exploring Hyperlinks, Contents, and Usage Data*. New York : Springer.

Liu, B. 2012. *Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies*. California: Morgan & Claypool.

Lun-Wei, K., L. Yu-Ting, et C. Hsin-Hsi. 2006. « Opinion Extraction, Summarization and Tracking in News and Blog Corpora. » In *Proceedings of AAAI-CAAW-06, the Spring Symposia on Computational Approaches to Analyzing Weblogs, January 2006*, 100–107.

Lynch, C.A. 2001. « Personalization and Recommender Systems in the Larger Context: New Directions and Research Questions. » In *Proceedings of the Second DELOS Network of Excellence Workshop on Personalisation and Recommender Systems in Digital Libraries, 2001, Dublin, Ireland, June 18-20, 2001*. ERCIM Workshop Proceedings. ERCIM.

Malrieu, D., et F. Rastier. 2001. « Genres et variations morphosyntaxiques. » *Traitement automatique de langues* 42 (2) : 548–77.

Martin, F.-J., J. Donaldson, A. Ashenfelter, M. Torrenset et R. Hangartner. 2011. « The Big Promisses of Recommender Systems ». *AI Magazine* 32 (3) : 19–27.

Maurel, S., P. Curtoni, et L. Dini. 2007. « Classification d'opinions par méthodes symbolique, statistique et hybride. » In *Actes du Troisième Défi Fouille de Texte, DEFT2007*, 121–27.

Mayaffre, D. 2008. « De l'occurrence à l'isotopie. Les co-occurrences en lexicométrie. » *Syntaxe et sémantique,* 53–72.

McCallum, A., et K. Nigam. 1998. « A Comparison of Event Models for Naive Bayes Text Classification. » In *Papers from the AAAI Workshop*, 41–48. AAAI Press.

McNee, S.M., J. Riedl, et Konstan. 2006. « Being Accurate Is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems. » In *Extended Abstracts Proceedings of the 2006 Conference on Human Factors in Computing Systems, CHI 2006, Montréal, Québec, Canada, April 22-27, 2006*, 1097–1101. New York : ACM. doi : 10.1145/1125451.1125659.

Memmi, D. 2000. « Le modèle vectoriel pour le traitement de documents ». *Cahiers Leibniz*. 2000-14, INPG

Montes-Garcia, A., J.M. Álvarez-Rodríguez, J.E. Labra-Gayo et M. Martínez-Merino. 2013. « Towards a Journalist-Based News Recommendation System: The Wesomender Approach. » *Expert Systems with Applications* 40 (17): 6735–41.

Moore, P. 2012 [en ligne]. « NearDuplicatesDetection 0.2.0. » https://pypi.org/project/NearDuplicatesDetection/0.2.0/. Consulté le 5 janvier 2019.

Mounin, G. 1993. *Dictionnaire de la linguistique*. 1e édition. Paris : Quadrige/PUF.

Musto, C., P. Basile, P. Lops, M. Gemmis, et G Semeraro. 2017. « Introducing Linked Open Data in Graph-Based Recommender Systems. » *Information Processing & Management* 53 (2) : 405–35.

Mutz, D.C. et L. Young. 2011. « Communication and Public Opinion. Plus ça change ? » *Public Opinion Quaterly* 75 (5) : 1018–1044.

Nageswara Rao, K. et V. Talwar. 2008. « Application Domain and Functional Classification of Recommender Systems a Survey. » *Desidoc Journal of Library and Information Technology* 28. ACM Press: 17–36.

Newman, N., R. Fletcher, A. Kalogeropoulos, D. Levy, et R.K. Nielsen. 2017 [en ligne]. « Reuters Institute Digital News Report 2017. » Disponible sur : https://ssrn.com/abstract=3026082. Consulté le 15 août 2019.

Office québécois de la langue française (OLF). 2013 [en ligne]. *Grand Dicionnaire Terminologique*. http://www.granddictionnaire.com.

Ohana, B., et B. Tierney. 2009. « Sentiment Classification of Reviews Using SentiWordNet. » In *9th. IT&T Conference, Dublin Institute of Technology, Dublin, Ireland, 22-23 October*. doi : 10.21427/D77S56.

Özgöbe, O., J.A. Gulla, et R.C Edur. 2014. « A Survey on Challenges and Methods in News Recommendation. » In *Proceedings of the 10th International Conference on Web Information Systems and Technologies, Barcelona, Spain*, 2:278–85. doi : 10.5220/0004844202780285.

P.T. Higgins, J. et S. Green. 2008. *Cochrane Handbook for Systematic Reviews of Interventions*. Wiley-Blackwell. England.

Pang, B, L Lee, et S Vaithyanathan. 2002. « Thumbs up? Sentiment Classification Using Machine Learning Techniques. » In *Proceedings of the Conference on Empirical*

Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002, 79–86. Association for Computational Linguistics.

Pang, B. et L. Lee. 2008. *Opinion Mining and Sentiment Analysis : Foundations and Trends in Information Retrieval*. Boston : Now - The Essence of Knowledge.

Pang, B., et L. Lee. 2004. « A Sentimental Education : Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. » In *ACL '04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Barcelona, Spain, July 21 - 26, 2004*, 271–78. Stroudsburg, PA, USA : Association for Computational Linguistics. doi: 0.3115/1218955.1218990.

Parisier, E. 2011. *The Filter Bubble: What the Internet Is Hiding from You*. New York: The Penguin Press.

Park, D. H., K. Kyeong Hyea, Choi Il Young et Kim Jae Kyeong. 2012. « A Literature Review and Classification of Recommender Systems Research ». *Expert Systems with Applications* 39 (11) : 10 059–10 072.

Paveau, M.-A. 2003 [en ligne]. « L'entrée doxa : pour un traitement rigoureux d'une notion floue. » *Mots. Les langages du politique* 71. Disponible sur : http:// mots.revues.org/8683. Consulté le 16 août 2019.

Paveau, M.-A. et G.-E. Safarti. 2003. *Les grandes théories de la linguistique : de la grammaire comparée à la pragmatique*. Paris : Armand Colin.

Pazzani, M.J et D. Billsus. 2007. « Content-Based Recommendation Systems. » *The Adaptive Web*. Vol. 4321. Lecture Notes in Computer Science. Berlin, Heidelberg : Springer.

Perelman, C., et L. Olbrechts-Tyteca. 1988. *Traité de l'argumentation : la nouvelle rhétorique*. 5e édition. Éditions de l'Université de Bruxelles.

Pétry, F., É. Bélanger, et L.M. Imbeau. 2006. *Le Parti libéral. Enquête sur les réalisations du gouvernement Charest*. Prisme. Presses de l'Université Laval.

Pincemin, B. 1999a. « Sémantique interprétative et analyses automatiques de textes : que deviennent les sèmes ? » *Revue Sémiotiques* 17 : 71–120.

Pincemin, B. 1999b. *Diffusion ciblée automatique d'informations : conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents.* Thèse de doctorat, CNRS.

Pincemin, B. 2012a [en ligne]. « Sémantique interprétative et textométrie. » *Texto!* XVII (3). Disponible sur : https://halshs.archives-ouvertes.fr/halshs-00981180. Consulté le 7 janvier 2017.

Pincemin, B. 2012b. « Hétérogénéité des corpus et textométrie. » Langages 3 (187) : 13–26.

Piryani, R., D. Madhavi et V.K. Singh. 2017. « Analytical Mapping of Opinion Mining and Sentiment Analysis Research during 2000-2015. » *Information Processing & Management* 53 : 122–50.

Plutchik, R. 1991. *The Emotions*. University Press of America.

Poirer, D, F. Fessant, et I. Tellier. 2010. « De la classification d'opinions à la recommandation : l'apport des textes communautaires. » *TAL* 51 (3) : 19–46.

Polguère, A. 2008. *Lexicologie et sémantique lexicale : notions fondamentales*. 2e édition. Les Presses de l'Université de Montréal.

Pottier, B. 1974. *Linguistique générale : théorie et description*. Klincksieck. Paris.

Pottier, B. 1987. *Théorie et analyse en linguistique*. Langue Linguistique Communication. Hachette.

Pu, P., L. Chen et R. Hu. 2011. « A User-Centric Evaluation Framework for Recommender Systems. » In *Proceedings of the 5th Conference on Recommender Systems (RecSys'11)*, 157–64.

Rao, J., A. Jia, Y. Feng et D. Zhao. 2013. « Personalized News Recommendation Using Ontologies Harvested from the Web. » *WAIM 2013. Lecture Notes in Computer Science*. Vol. 7923. Berlin, Heidelberg: Springer.

Rastier, F. 1987. *Sémantique interprétative*. Formes sémiotiques. Paris : PUF.

Rastier, F. 1989. *Sens et textualité*. Paris : Hachette.

Rastier, F. 1994. « Sur l'immanentisme en sémantique. » *Cahiers de linguistique française* 15 : 325–335.

Rastier, F. 1996a. « Pour une sémantique des textes. » *Sens et Textes*, 9–35. Paris.

Rastier, F. 1996b [en ligne]. « La sémantique des textes : concepts et applications. » *Texto !* http://www.revue-texto.net/Inedits/Rastier/Rastier_Concepts.html. Consulté le 7 janvier 2019.

Rastier, F. 2001a. *Arts et sciences du texte*. Paris : PUF.

Rastier, F. 2001b [en ligne]. « Éléments de théorie des genres. » *Texto !* Disponible sur : <http://www.revue-texto.net/Inedits/Rastier/Rastier_Elements.html>. Consulté le 7 janvier 2019.

Rastier, F. 2001c. *Sémantique et recherches cognitives* (PUF). Paris.

Rastier, F. 2006a. « Formes sémantiques et textualité ». *Langages* 3 (163) : 99–114.

Rastier, F. 2006b. « Sémiotique des sites racistes ». *Mots. Les langages du politique* 80 : 73–85.

Rastier, F. 2008. Que cachent les « données textuelles » ? *Actes des 9e Journées internationales d'analyse statistique des données Textuelles (JADT 2008)*, 13–26. Lyon : Presses Universitaires de Lyon.

Rastier, F. 2011. *La mesure et le grain*. Paris : Honoré Champion.

Rastier, F. 2014 [en ligne]. « Action et Récit ». *Texto ! Textes et Culture* XIX (3). Disponible sur : http://www.revue-texto.net/index.php?id=3579. Consulté le 7 janvier 2019.

Rastier, F. 2015. « La sémantique interprétative. » *The Routledge Handbook of Semantics*, 534. Routledge Handbooks in Linguistics. New York.

Rastier, F. et B. Pincemin. 1999. « Des genres à l'intertexte. » *Sémantique de l'intertexte* 33 : 83–111.

Rastier, F., M. Cavazza et A. Abeillé. 1994. *Sémantique pour l'analyse : de la linguistique à l'informatique.* Paris : Masson.

Resnick, P. et H.R. Varian. 1997. « Recommender Systems. » *Communications of the ACM* 40 (3).

Rieffel, R. 2010. *Sociologie des médias*. Paris : Ellipses.

Rinck, F. 2006. « L'article de recherche en sciences du langage et en lettres. Figure de l'auteur et identité disciplinaire du genre. » Thèse de doctorat, Grenoble, France : Université de Grenoble III - Stendhal.

Russell, D.L. 1997. « Rethinking Genre in School and Society. An Activity Theory Analysis. » *Written Communication* 14 (4): 504–54.

Salem, A., C. Lamalle, W. Martinez, S. Fleury, B. Fracchiolla, A. Kuncova et A. Maisondieu. 2003 [En-ligne]. Lexico3 - « Outils de statistique textuelle. Manuel d'utilisation. » *Styled-CLA2T, Université de la Sorbonne Nouvelle - Paris 3*. Disponible sur : http://www.tal.univ-paris3.fr/lexico/. Consulté le 7 janvier 2019.

Salton, G. 1988. « Term-Weighting Approaches in Automatic Text Retrieval ». *Information Processing and Management: An International Journal* 24 (5) : 513–523.

Salton, G. et M.J. McGill. 1983. *Introduction to Modern Information Retrieval*. New York, NY, USA : McGraw-Hill, Inc.

Saussure, F. 2016. *Curso de linguística geral*. 28e édition. São Paulo : Cultrix.

Schneider, K. M. 2004. « On Word Frequency Information and Negative Evidence in Naive Bayes Text Classification. » In *International Conference on Natural Language Processing (in Spain)*, 474–85. Springer.

Schuldt, J.P., S.H. Konrath et N. Schwarz. 2011. « 'Global Warming' or 'Climate Change'? Wheter the Planet Is Warming Depends on Question Wording. » *Public Opinion Quarterly* 75 (1) : 115–124.

Shani, G., et A. Gunawardana. 2011. « Evaluating Recommendation Systems. » In *Recommender Systems Handbook*, édité par F. Ricci, L. Rokach, B. Shapira, et P.B Kantor, 257–297. Springer.

Shi, Y., X. Zhao, J. Wang, M. Larson et A. Hanjalic. 2012. « Adaptive Diversification of Recommendation Results via Latent Factor Portfolio. » In *Proceedings of the 35th International Conference on Research and Development in Information Retrieval (SIGIR'12*, 175–84.

Sjöblom, M.Katsberg et J.-M. Leblanc. 2012. « Extraction des isotopies d'un corpus textuel : analyse systématique des structures sémantiques et des cooccurrences, à travers différents logiciels textométriques. » *Texto !* XVII (3). Disponible sur : http://www.revue-texto.net/index.php?id=3059. Consulté le 7 janvier 2019.

Slodzian, M. et M. Valette. 2009. « Connaissances prescrites ou connaissances décrites ? L'apport de la sémantique des textes. » In *Actes du 12e colloque international sur le document électronique*, 129–141. Paris : Europia Productions.

Sokolova, M. et G. Lapalme. 2009. « A Systematic Analysis of Performance Mesures for Classification Tasks. » *Information Processing and Management* 45: 427–437.

Somasundaran, S., T. Wilson, J. Wiebe et V. Stoyanov. 2007. « QA with Attitude: Exploiting Opinion Type Analysis for Improving Question Answering in On-Line Discussions and the News. » *Conference on Weblogs and Social Media*. Boulder, Clorado, USA.

Soo-Min, K. et E. Hovy. 2005. « Identifying Opinion Holders for Question Answering in Opinion Texts. » In *AAAI Workshop on Question Answering in Restricted Domains*.

Statistique Canada. 2016 [en ligne]. « L'utilisation des médias pour suivre les nouvelles et l'actualité. » 2016. 89-652-X2016001. Disponible sur : https://www150.statcan.gc.ca/n1/pub/89-652-x/89-652-x2016001-eng.htm. Consulté le 7 janvier 2019.

Sunstein, C. 2007. *Republic.com 2.0*. Princeton, NJ : Princeton University Press.

Susen, S. 2011. « Critical Notes on Habermas's Theory of the Public Sphere. » *Social Analysis* 5 (1) : 37–62.

Torres-Moreno, S., P. Curtoni, et L. Dini. 2007. « Comment faire pour que l'opinion forgée à la sortie des urnes soit la bonne? Application au Défi DEFT 2007. » In *Actes du Troisième Défi Fouille de Texte, DEFT2007, Grenoble, France*, 121–27.

Turney, P.D. 2002. « Thumbs up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews. » In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, 417–24.

Valette, M. 2004. « Sémantique interprétative appliquée à la détection automatique de documents racistes et xénophobes sur Internet. » *Approches sémantiques du document numérique, Actes du 7e colloque international sur le document électronique*, 215–230.

Valette, M. 2010. « Approche textuelle du lexique. » Mémoire, Paris : Institut National des Langues et Civilisations Orientales.

Valette, M. et M. Slodzian. 2008. « Sémantique des textes et recherche d'information. » *Revue française de linguistique appliquée* XIII (1) : 119-133.

Van Alstyne, M., et E. Brynjolfsson. 2005. « Global Village or Cyber-Balkans? Modeling and Measuring the Integration of Electronic Communities. » *Management Science* 51 (6) : 851–68.

Vanni, L. et A. Mittmann. 2016. « Cooccurrences spécifiques et représentations graphiques, le nouveau 'Thème' d'Hyperbase. » In *13th International Conference on Statistical Analysis of Textual Data, Nice, France*, 295–305. JADT 2016.

Vargas, S., L. Baltrunas, A. Karatzoglou et P. Castells. 2014. « Coverage, Redundancy and Size-Awareness in Genre Diversity for Recommender Systems. » In *Proceedings of the 8th Conference on Recommender Systems (RecSys'14)*, 209–16.

Vechtomova, O. 2010. « Facet-Based Opinion Retrieval from Blogs. » *Information Processing and Management* 46 (1): 71–88.

Venturini, T. 2009. « Diving in Magma: How to Explore Controversies with Actor-Network Theory. » *Public Understanding of Science* 19 (3): 258–73.

Venturini, T. 2012. « Building on Faults: How to Represent Controversies with Digital Methods. » *Public Understanding of Science* 21 (7): 796–812.

Vernier, M., L. Monceaux et B. Daille. 2009a. « DEFT'09 : Détection de la subjectivité et catégorisation de textes subjectifs par une approche mixte symbolique et statistique. » *Actes de l'atelier de clôture de la 5e Édition du Défi Fouille de Textes*.

Vernier, M., L. Monceaux, B. Daille et E. Dubreil. 2009b. « Catégorisation des évaluations dans un corpus de blogs multi-domaine. » *Revue des nouvelles technologies de l'information (RNTI)*, 45–70.

Vernier, M., Y. Mathet, F. Rioult, T. Charnois, S. Ferrari, et D. Legallois. 2007. « Classification de textes d'opinion : une approche mixte n-grammes et sémantique. » *Atelier Défi Fouille de Textes*, 103–14.

Walter, H. (2001). « Axiologie et sémantique chez André Martinet ». *La linguistique*, *37*, 59–68.

Wei-Hao, L., et A. Hauptmann. 2006. « Are These Documents Written from Different Perspectives? A Test of Different Perspectives Based on Statistical Distribution Divergence. » *Proceedings of the 42th Conference on Association for Computational Linguistics, Sydney, Australia*, 1057–1064. Association for Computational Linguistics.

Wen, H., L. Fang et L. Guan. 2012. « A Hybrid Approach for Personalized Recommendation of News on the Web. » *Expert Systems with Applications* 39 (5): 5806–14.

Wilson, P. et N. Robinson. 1990. « Form Subdivisions and Genre. » *Library Resources & Technical Services* 34 (1): 36–43.

Winsor, D.A. 1999. « Genre and Activity Systems: The Role of Documentation in Maintaining and Changing Engineering Activity Systems. » *Written Communication* 16 (2): 200–224.

Witten, I. H. et E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Second Edition. Morgan Kaufmann.

Woods, J., et Walton. 2004 : 1982. *Argument : Critical Thinking, Logic and the Fallacies*. 2e édition. Toronto : Prentice Hall.

Zhang, J., Y. Kawai, T. Kumamoto et K. Tanaka. 2009. « A Novel Visualization Method for Distinction of Web News Sentiment. » *Lecture Notes in Computer Science*, 5802 :181–94.

Ziegler, C.N., S.M McNee, J.A. Konstan, et G. Lausen. 2005. « Improving Recommendation Lists through Topic Diversification ». In *WWW '05 Proceedings of the 14th International Conference on World Wide Web*, 22–32. New York: ACM.

Annexes

Annexe A. Exemples d'articles classés dans ETUD et GOUV

Classe ETUD

Journal :	Le Devoir
Section :	Idées, p. A9
Titre :	Grève étudiante : au-delà des sous
Auteur :	Diane Lamoureux — Professeure au Département de science politique de l'Université Laval
Date :	Mercredi 4 avril 2012
Critère de sélection :	Arguments favorables à la continuité de la mobilisation Plaidoyer pour l'accessibilité à l'éducation ou sur l'état comme garant des services de base pour tous les citoyens.
	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. Certes, la hausse des droits de scolarité n'est pas un prétexte. Comme se plaît à le rappeler la ministre de l'Éducation, Line Beauchamp, c'est avec une rare unanimité que la FECQ, la FEUQ et la CLASSE ont refusé de participer à des consultations bidon où il n'était pas question de remettre en cause la hausse, seulement d'en discuter les modalités d'application. En réitérant leur refus lors de la manifestation de novembre 2011, puis en déclenchant un mouvement de grève à partir de la mi-février, les étudiants ont clairement indiqué qu'ils refusaient la hausse des droits de scolarité. Le mouvement de grève a d'ailleurs (presque) débuté avec une action de solidarité entre les étudiants et la Coalition contre la hausse des tarifs afin de bloquer la tour de la Bourse, le 16 février. Contrairement à ce qu'affirme le gouvernement, le mouvement étudiant, en se solidarisant avec les groupes communautaires, féministes et syndicaux qui s'opposent notamment à la « taxe santé » uniforme et à

la hausse des tarifs d'électricité, voulait montrer qu'il n'entendait pas défendre de supposés privilèges, mais s'opposer à une mesure sectorielle (la hausse des droits de scolarité) qui s'inscrit dans un mouvement plus large de démantèlement de la nature publique des services gouvernementaux et de tarification de ceux-ci selon le principe de « l'usager-payeur ».

Depuis le début de cette vague de mobilisation étudiante, on peut voir un déplacement des enjeux : on reparle maintenant de plus en plus ouvertement de gratuité scolaire et on dénonce les gabegies administratives des directions universitaires, plus préoccupées de béton et de compétitivité (surtout celle de leurs salaires) que de formation intellectuelle. 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.

On aurait tort d'attribuer une telle transformation au caractère fallacieux de la revendication initiale ou à la fourberie de certains leaders étudiants. C'est plutôt l'œuvre du caractère politisant de la lutte. Si plusieurs avaient des espoirs en déclenchant le mouvement, personne ne pouvait prédire le cours qu'il prendrait et la formidable leçon de science politique qu'en tireront ses participants. En agissant collectivement, en inventant des slogans, en prenant le temps de discuter, en arpentant les rues des villes pour faire autre chose que se déplacer, en profitant du soleil de ce printemps inespéré, les militantes et les militants donnent chair à ces valeurs fondamentales des sociétés démocratiques que sont l'égalité, la liberté et la solidarité.

Face au tournant néolibéral accentué dans les politiques québécoisesdepuis le fameux « déficit zéro » de Lucien Bouchard et réitéré par les gouvernements successifs à Québec (et à Ottawa), la grève étudianterappelle que l'éducation n'est pas une marchandise que l'on débite à la pièce selon la capacité de payer du « client » et que l'on choisit en fonction de sa rentabilité supposée. Elle souligne également que l'université n'est pas une entreprise dont la gestion relève de son seul conseil d'administration.

Dans ce sens, on peut situer le mouvement étudiant actuel dans la foulée du printemps arabe revendiquant la démocratie, dans la logique du mouvement des indignés de Madrid ou d'Athènes contre la « discipline budgétaire », ou dans celle du mouvement Occupy, dénonçant l'accroissement des inégalités sociales. D'abord l'expression du refus d'une supposée fatalité (tous les prix augmentent, pourquoi pas lesdroits de scolarité ?). Ensuite une formidable expérience qui fera sentir ses effets à long terme : la

	recherche, non sans quelques tâtonnements (c'est si certains plaquaient une solution toute faite sur le mouvement qu'il faudrait s'inquiéter, pas vis-à-vis de ses hésitations) d'un autre monde plus juste et plus démocratique. 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...

Journal :	Le Devoir
Section :	Éditorial, p. A6
Titre :	Libre opinion —Des oubliées : les études supérieures et la recherche
Auteur :	Vincent Larivière — Professeur à l'École de bibliothéconomie et de sciences de l'information de l'Université de Montréal et chercheur associé à l'Observatoire des sciences et des technologies de l'UQAM
Date :	Mardi 10 avril 2012
Critère de sélection :	Arguments favorables à la continuité de la mobilisation Plaidoyer pour l'accessibilité à l'éducation ou sur l'état comme garant des services de base pour tous les citoyens.
	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).

Effet pervers

De récentes données sur le rôle des étudiants de doctorat dans le système québécois de la recherche montrent fortement qu'une réduction de l'accès aux études supérieures aura un effet négatif important sur les activités de recherche des universités. En effet, bien qu'en phase d'apprentissage, les étudiants de doctorat comptent pour une part importante de la main-d'œuvre en recherche des universités.

Le nombre d'articles scientifiques -- indicateur fiable de l'activité de recherche fondamentale des universités -- auxquels au moins un étudiant de doctorat a contribué est passé de 1500 à 3000 entre 2000 et 2007, et représente, en 2007, le tiers du total des publications des universités québécoises. Dans certaines disciplines, ce pourcentage est encore plus élevé : en 2007, les doctorants québécois ont contribué à 50 % des articles en physique, 43 % en recherche biomédicale, 40 % en chimie, 37 % en biologie et, finalement, 35 % en psychologie. Et ces chiffres ne concernent que les doctorants : l'inclusion des étudiants de maîtrise les augmenterait encore davantage.

Une baisse de l'accès aux études de cycles supérieurs et, par conséquent, du nombre de doctorants, non seulement réduirait la capacité de recherche actuelle des universités -- qui perdraient ainsi l'accès à une main-d'œuvre relativement « bon marché », stimulée et désireuse de contribuer à l'avancement des connaissances --, mais hypothéquerait également celle de demain, en réduisant le nombre de nouveaux chercheurs formés.

En somme, en réduisant l'attrait des études aux cycles supérieurs, toute hausse importante des droits de scolarité aurait aussi pour conséquence d'influer sur la capacité de recherche des universités québécoises. À l'heure où les universités et les gouvernements n'ont que les mots « société du savoir » et « innovation » à la bouche, on peut se demander s'ils ont réellement songé aux effets pervers de

	leurs décisions.

Journal :	Le Droit
Section :	Éditorial, p. 13
Titre :	Arguments contre la hausse
Auteur :	Guy Bellemare, Professeur, Relations industrielles, Université du Québec en Outaouais
Date :	Jeudi 5 avril 2012
Critère de sélection :	Plaidoyer pour l'accessibilité à l'éducation ou sur l'état comme garant des services de base pour tous les citoyens.
	En 1990, le gouvernement de Robert Bourassa procède au dégel des droits de scolarité ; ils passèrent en quatre ans de 581$ à 1630$. La conséquence fut immédiate : les inscriptions à l'UQO chutent de 6072 étudiants en 1992 à 4562 en 1998, remontant à 4987 en 2010. L'accessibilité universitaire fut détruite et le sera encore si le gouvernement persiste à vouloir hausser les frais de scolarité. La hausse des frais diminuera drastiquement les inscrits à l'UQO, provoquant, comme en 1992-1998, une chute des budgets, financée selon le nombre d'étudiants inscrits. À l'UQO, entre 1994 et 1996, le nombre de postes de professeurs passe de 149 à 127. Le gel des frais est la base de l'accessibilité. Les Universités du Québec comptent le plus d'étudiants universitaires de première génération. Pour eux, ça signifie que les frais et l'endettement sont de très grands risques financiers. Ils doivent travailler davantage pour payer leurs études, ce qui nuit à la réussite. Leurs chances de placement rapide sont plus faibles. Les promesses de bonification du régime des prêts et bourses sont des illusions

Journal :	Le Nouvelliste

Section :	Opinions, p. 15
Titre :	Continuez!
Auteur :	Marie-Josée Richard, Enseignante de français au secondaire, Mère de famille, Shawinigan
Date :	Jeudi 12 avril 2012
Critère de sélection :	Arguments favorables à la continuité de la mobilisation Plaidoyer pour l'accessibilité à l'éducation ou sur l'état comme garant des services de base pour tous les citoyens.

Depuis plusieurs semaines, les étudiants n'ont jamais tant marché de leur vie. Et je marche avec eux. Le débat actuel ne tourne qu'autour de l'argent, de la hausse des frais de scolarité ou de leur gel. Mais voilà, le débat ne se bat pas la gueule à la bonne place.

Selon moi, il serait grand temps que les Québécois se posent les vraies questions : quelles sont les valeurs qui nous définissent en tant que peuple ? L'éducation ? La santé ? L'économie ? Le Plan Nord ? L'entrepreneuriat ? L'avenir des jeunes ? Les gaz de schiste ? L'énergie nucléaire ?

Personnellement, j'aimerais que le gouvernement tienne un référendum pour déterminer quelles sont les valeurs à prioriser au Québec. Je suis à peu près certaine que l'éducation et la santé arriveraient ex aequo. À partir de ce constat, il serait plus facile de prendre des décisions politiques alignées sur ces valeurs.

Je suis complètement découragée quand j'entends dire que les jeunes étudiants sont riches parce qu'ils ont beaucoup de gadgets électroniques et qu'ils vont en voyage chaque année. Ce sont les enfants de parents qui les ont gâtés, qui leur ont fait voir du pays, ce sont les enfants de consommateurs et parfois de surconsommateurs endettés et même surendettés. Et vous voudriez que j'adhère à ces pauvres arguments ? L'éducation n'est pas un service qu'on doit indexer au coût de la vie, c'est un droit, c'est le fluide vital de notre société. C'est triste à mourir que la ministre de l'Éducation cautionne la marchandisation du savoir.

L'autre question dont il faudrait débattre, c'est : quel avenir voulons-nous assurer à ces étudiants? Ils sont démographiquement peu nombreux, ils sont déjà hypothéqués avant même d'avoir entrepris des études post-secondaires.

Ne nous le cachons pas : ils devront nous supporter financièrement quand nous serons vieux, mous de la caboche et que nous aurons les

	poches vides. Moi, ce que j'aimerais que les jeunes puissent faire, c'est étudier sans se casser trop la tête avec leurs finances scolaires et personnelles. J'aimerais que tous les élèves qui le désirent, même mes étudiants défavorisés, puissent rêver faire des études avancées. Ce que je veux que le Québec fasse pour les jeunes, c'est former des esprits penseurs capables de créer, d'innover, d'entreprendre, de gérer efficacement, de penser autrement que nous, qui visiblementn'arrivons qu'à nous enfoncer davantage. J'aimerais qu'ils prennent leur place, que nous leur ouvrions la voie, que nous les soutenions, que nous les guidions, que nous les accompagnions et que nous les laissions prendre leur envol. Pas que les néo-libéraux leur coupent les ailes avec une hausse soudaine et exagérée des frais de scolarité. Étudiants, continuez à sonner la cloche! À nous de les écouter.

Classe GOUV

Journal :	Le Quotidien
Section :	Chronique, p. 10
Titre :	L'accessibilité n'est pas menacée
Auteur :	Manoj Karivelil, Jonquière
Date :	Lundi 2 avril 2012
Critère de sélection :	Plaidoyer pour une vision marchande de l'éducation ou l'idée de l'éducation comme un investissement individuel Arguments contre les demande des étudiants. Arguments en faveur de la hausse. Texte dénonçant la violence ou les inconvénients provoqués par les manifestations.
	À la suite de l'écoute répétitive des leaders des fédérations collégiales et universitaires (et on peut le dire sans gêne, probablement des futurs leaders syndicaux) dans les médias, j'en suis venu à m'en faire un devoir de répliquer sur leurs revendications. Ainsi, on parle sans cesse

de l'accessibilité aux études supérieures, et comment une hausse des frais de scolarité diminuera celle-ci. Faux! Le Québec représente probablement l'endroit où l'accessibilité aux études supérieures est la meilleure en Amérique du Nord, et elle le restera. Un étudiant venant d'un milieu défavorisé ne diminuera aucunement ses chances d'avoir accès aux études supérieures. Le programme de prêts et bourses québécois permet à tous ceux qui désirent entreprendre des études supérieures de le faire. À partir de cela, le goût et le désir de poursuivre des études supérieures reviennent à lui-même et sa famille. À noter ici que les provinces ayant des frais de scolarité beaucoup plus élevés ont des taux de diplomation égaux, voire supérieurs aux Québec. On nous bombarde sans cesse ce montant de 1625$. On parle ici d'une augmentation de 163$ par session étalée sur cinq ans. En réalité, même si on parlait d'une hausse de 1625$ par année dès maintenant, le coût additionnel à l'étudiant représente moins de 14$ par jour d'étude. Rappelons ici qu'on parle de monter la contribution de l'étudiant à 17 % de la facture totale, 83 % de la vraie facture sera assumée par l'ensemble des contribuables. Les leaders étudiants expriment tous le fait qu'ils ont l'appui de la population dans leurs demandes. Faux! J'en conviens, ils ont l'appui de la population pour que le gouvernement retourne à la table des négociations avec eux. En effet, l'empêchement des contribuables d'aller travailler, et le blocage des rues et ponts, font que la classe moyenne désire que le conflit se règle. Ce n'est pas la même chose qu'un appui de leurs demandes idéologiques. La forte majorité de la population n'appuie pas leurs demandes de gel ou de gratuité scolaire, car ces augmentations affecteront directement la classe moyenne, qui croupit déjà avec le fardeau fiscal le plus élevé en Amérique du Nord.

On parle toujours des chiffres monstres de 100 000 à 300 000 étudiants en grève. Faux! On fait preuve ici d'une extrapolation massive de chiffres. Les leaders syndicalo-étudiants passent leurs mandats de grève avec presque toujours de très faibles majorités, donc ils ne peuvent affirmer qu'ils représentent tous les étudiants collégiaux et universitaires. En fait, ils en représentent légèrement plus que la moitié seulement.

Ayant étudié moi-même à l'université, et ayant fini avec le montant maximal de prêts, j'ai évidemment pris plusieurs années en début de carrière pour rembourser cette dette. J'ai toujours considéré mon prêt étudiant comme un investissement pour moi-même et ma famille. L'augmentation proposée est juste, et le Québec restera toujours l'endroit en Amérique du Nord où les frais de scolarité seront les plus bas.

Journal :	Le Soleil
Section :	Éditorial, p. 31
Titre :	Profiter de l'ouverture
Auteur :	Brigitte Breton
Date :	Jeudi 5 avril 2012
Critère de sélection :	Arguments sur la raisonnabilité des propositions du gouvernement et contre la continuation du mouvement de grève. Texte critiquant l'inflexibilité des étudiants dans les négociations.

Le gouvernement et les associations étudiantes sont rendus au point où ils n'ont pas le choix de s'asseoir pour trouver un dénouement à la crise qui perturbe les cégeps et les universités depuis février. Rester sur ses positions initiales est devenu un pari trop risqué, tant pour le gouvernement libéral que pour les étudiants opposés à la hausse des droits de scolarité.

Les deux parties ont réussi à faire leurs preuves au cours des dernières semaines. Le gouvernement Charest n'a pas fléchi devant le mécontentement suscité par sa décision d'augmenter de 325 $ par année les droits de scolarité, contrairement à d'autres dossiers où il a reculé devant les contestataires. De leur côté, les étudiants ont démontré qu'ils étaient capables de se mobiliser pour une cause et d'obtenir des appuis à l'extérieur des campus. Avertissement à ceux qui les croyaient amorphes, non politisés et individualistes.

Ce premier test franchi avec succès, reste l'étape suivante, qui fait appel au sens des responsabilités et au réalisme : conclure une entente satisfaisante aux yeux de chacune des parties afin que le cours normal des choses puisse reprendre rapidement dans les cégeps et les universités.

L'étau se resserre. Les manifestations se corsent et donnent lieu à des arrestations. Ni les étudiants ni Québec n'ont intérêt à ce qu'une manif se transforme en opération saccage et conduise à des blessures, voire à des décès. Le recours aux tribunaux s'accentue et rétrécit le champ d'action des opposants. Québec ne peut s'en réjouir, car les injonctions réservent des lendemains tendus sur les campus.

La pression se fait aussi plus forte chez les dirigeants des

établissements et le personnel, qui doivent allonger le calendrier scolaire et repousser les vacances. Qu'ils soient pour ou contre la grève, le décompte est également enclenché chez les étudiants, qui évaluent quel sera le coût du boycottage des cours, du travail rémunéré perdu ou retardé. Calcul aussi de la part des libéraux, qui doivent se demander ce qu'ils gagnent vraiment à garder des milliers de jeunes dans la rue plutôt qu'en classe et, si la session s'étire, en classe plutôt qu'à leur emploi d'été.

Des étudiants estiment que la ministre de l'Éducation Line Beauchamp ne montre pas suffisamment d'ouverture au dialogue en refusant de parler de gel des droits ou de gratuité scolaire. C'est rêver en couleurs que d'espérer qu'elle puisse avancer sur ce terrain.

Cependant, après avoir répété pendant des semaines que la hausse des droits était accompagnée d'une bonification du régime de prêts et bourses (donc que tout allait bien et que le plan gouvernemental n'avait pas besoin de retouches), voilà qu'elle réitère son offre de discuter d'accessibilité aux études et du programme d'aide financière.

Les étudiants doivent saisir cette occasion et tenter d'aller chercher le maximum d'aide pour atténuer les effets de la hausse des droits chez les étudiants plus vulnérables.

Une plus grande bonification du régime de prêts et bourses et des modifications aux règles de remboursement sont la voie privilégiée pour dénouer l'impasse entre Québec et les associations étudiantes. Elles permettent de rallier à la fois les intérêts des étudiants et de bien des citoyens qui, même s'ils reconnaissent la nécessité de hausser les droits de scolarité, s'inquiètent des effets que cette hausse pourrait avoir sur la poursuite d'études supérieures et l'endettement des jeunes.

Rappelons que c'est aussi l'esprit des recommandations formulées par le comité consultatif sur l'accessibilité aux études du Conseil supérieur de l'éducation et d'un comité de l'Université du Québec qui se sont penchés sur la hausse des droits.

Le moment est venu d'explorer des solutions qui sauront répondre à la fois aux intérêts des étudiants et à l'ensemble de la société.

Journal :	Le Quotidien
Section :	Chronique, p. 10
Titre :	Respecter la démocratie
Auteur :	Jean-Pierre Gauthier, Saguenay
Date :	Vendredi 6 avril 2012
Critère de sélection :	Texte soulignant l'importance des droits individuels des étudiants.

Chers mentors et professeurs à qui nous confions la responsabilité de nos enfants pour en faire des adultes responsables! Dans vos complaintes et critiques envers le recteur de l'Université du Québec à Chicoutimi, vous faites référence à la démocratie. Lorsqu'une population étudiante vote contre un boycott des cours, n'est-il pas démocratique de respecter le choix de la masse ? Empêcher ceux qui ont voté contre leconflit d'assister à leurs cours n'est-il pas directement une atteinte à cette chère démocratie que vous semblez tant chérir ?

En aucun cas le recteur n'a empêché les étudiants de manifester contre la hausse des frais de scolarité. Il protège la majorité qui désire poursuivre leurs études en toute quiétude. Il fait son travail de façon responsable, alors faite le vôtre! Cette tâche constitue à influencer pour le reste de leur vie des jeunes à l'esprit malléable. Il ne s'agit pas de jouer le rôle du professeur ultra-cool, anticonformiste qui utilise la démocratie lorsqu'elle lui sert et qui la met de côté si elle porte préjudice à des agissements indéfendables!

Journal :	Le Soleil
Section :	Éditorial, p. 29
Titre :	La hausse nuit-elle réellement ?
Auteur :	Michel Giroux, Québec
Date :	Mercredi 11 avril 2012

Critère de sélection :	Arguments contre les demandes des étudiants. Arguments favorables
	Depuis plus de six semaines, bon nombre d'étudiants de cégeps et d'universités réclament à grands cris le gel des droits de scolarité, et certains vont même jusqu'à exiger l'entière gratuité. Plus de 60 % des étudiants sont en classe pendant ce temps. Les hausses sont effectivement assez élevées, mais ayons la franchise de l'avouer, elles auraient dû être exigées il y a plusieurs années. À un point aussi crucial de l'année scolaire, il serait grand temps que les autorités gouvernementales interviennent. Il appartient au premier ministre Charest de s'adresser à la nation, démontrant, chiffres et tableaux à l'appui, les raisons légitimes de ces augmentations de même que les coûts engendrés par un recul de son gouvernement. Tout le monde convient que les hausses sont rapides et importantes. Mais demandons-nous si elles vont réellement empêcher des étudiants de poursuivre leurs études. Quand les prix à la consommation augmentent, les étudiants se privent-ils réellement de tout ce à quoi ils sont habitués ? Ils devront certes faire des choix comme nous tous, citoyens touchés (comme eux) par lesdites hausses.

Annexe B. Guide de classification manuelle des articles dans les classes ETUD et GOUV

Guide de classification

1. À propos de ce guide

Ce guide contient les instructions pour le travail de contre-codage pour la recherche doctorale « Recommandation d'articles de presse basée sur la fouille d'opinions : une approche pour assister la classification automatique de controverses », réalisé à l'École de Bibliothéconomie et sciences de l'information de l'Université de Montréal. Vous ferez une classification de 200 articles d'opinion au sujet de la grève étudiante parus dans la presse québécoise. Ces articles ont été choisis aléatoirement du corpus original de la recherche. La classification sera effectuée sur le site *Eureka* (www.eureka.cc) à l'aide de l'outil *Eureka Analytik*. En tant qu'annotateur, vous vous engagez à suivre les consignes présentes dans ce guide.

2. Information sur le corpus

Le corpus de documents à classer est formé par 250 articles d'opinion publiés dans la presse québécoise francophone. Il est composé des éditoriaux, courriers de lecteurs et d'articles de tribune libre au sujet de la grève étudiante de 2012 au Québec. Les articles sont parus entre 13 février 2012 au 30 octobre 2012.

3. Critères de classification

La classification consiste à attribuer une étiquette en fonction de l'opinion véhiculée dans l'article en lien avec la grève étudiante. Si l'article est plus favorable à la position défendue par les étudiants, le participant doit attribuer l'étiquette **ETUD** à l'article. Au contraire, si l'article est plus favorable à la position du gouvernement, il faudra attribuer l'étiquette **GOUV**. Le tableau ci-dessous une liste de critères à considérer pour décider l'étiquette à attribuer. La liste sert à guider votre travail de classification mais vous devez faire réaliser la tâche selon votre meilleur jugement.

Il se peut que certains articles ne présentent aucun des critères établis pour la classification dans **ETUD** et **GOUV** ou qu'ils présentent certaines particularités qui rendent difficile la tâche de classification. Afin d'identifier ces articles, nous avons inclus dans le tableau ci-dessous les critères permettant de rejeter un article. Si l'un ou l'autre de ces critères s'applique, vous devez attribuer l'étiquette **Rejeter** à l'article concerné.

Les articles à classer peuvent attester un seul critère de la liste ou en attester plusieurs. Les critères peuvent donc être cumulatifs pour un même article. Dans la section **4) Annexes**, vous trouverez un exemple d'article pour chaque classe, ainsi que les critères choisis pour les classer. Nous avons également mis un exemple d'article à rejeter.

Classe	**Critères à considérer**
ETUD	Arguments majoritairement contre la hausse. Arguments favorables à la continuité de la mobilisation. Dénonciations de la violence de la police envers les étudiants. Plaidoyers pour l'accessibilité à l'éducation ou sur l'État comme garant des services de base pour tous les citoyens. Arguments soulignant l'importance de respecter les droits collectifs des étudiants de faire une grève. Arguments critiquant l'attitude du gouvernement envers les étudiants en grève. Arguments critiquant le projet de loi 78.
GOUV	Arguments majoritairement en faveur de la hausse. Arguments contre les demandes des étudiants. Arguments sur le caractère raisonnable des propositions du gouvernement et contre la suite du mouvement de grève. Plaidoyers pour une vision marchande de l'éducation ou qui défendent l'idée de l'éducation comme un investissement individuel. Arguments critiquant l'inflexibilité des étudiants dans les négociations. Arguments dénonçant la violence des manifestants ou les inconvénients provoqués par les étudiants en grève. Arguments soulignant l'importance des droits individuels des étudiants.

	Arguments en faveur du projet de loi 78.
Rejeter	Textes au sujet de la grève étudiante qui proposent des solutions pour la question du financement des études ou pour dénouer le conflit, mais qui ne prennent pas parti explicitement d'un camp de la dispute. Discussions politiques très écartées du sujet de la grève ou qui parlent de la grève étudiante de façon secondaire. Textes qui rassemblent des sujets disparates dans un seul document.

4. Outil de classification : Eureka Analytik

Afin de réaliser la classification, vous devrez accéder à l'outil Eureka Analytik. Cet outil contient une section appelle Corpus qui contient les articles d'opinion qui doivent être classés. Tous les articles du corpus sont dotés d'une étiquette initiale nommée **Révision**. La classification consiste à attribuer une des étiquettes existantes (**ETUD**, **GOUV** ou **Rejeter**) aux articles listés dans le Corpus, et supprimer l'étiquette **Révision**.

Comment apposer les étiquettes

Voici les étapes à suivre dans l'outil de classification

- Accéder au site nouveau.eureka.cc
- Cliquer sur l'onglet « Déjà abonné »
- Entrez votre code d'usager et votre mot de passe : etudiantMarcela2 / etudiantMarcela2
- Cliquer sur le lien « Analytik » qui se retrouve dans le menu principal du site.
- Une fois sur l'Analytik, l'usager verra le contenu de l'onglet « Grève étudiante ».
- Dans le deuxième niveau d'onglets (ceux en beige), cliquer sur corpus.
- Cliquer sur le premier article de la liste. Le texte de l'article apparaîtra dans une composante à droite de la liste. Le site affiche également une troisième colonne avec les étiquettes **Révision**, **ETUD**, **GOUV** et **Rejeter**.

- Tous les articles possèdent l'étiquette **Révision**. Après la lecture de l'article, vous devez choisir l'étiquette que vous voulez attribuer et décocher l'étiquette **Révision**. Pour ce faire, il faut simplement cocher la case à côté de l'étiquette désirée à la troisième colonne et décocher la case **Révision.**
- Les flèches situées en haut de l'article permettent de passer d'un article à l'autre ou de revenir en arrière.

Attention: afin de pouvoir décocher la case de l'étiquette **Révision***,* il faut d'abord choisir l'étiquette que l'on veut attribuer à l'article.

Annexes

Vous trouverez ci-dessous des exemples d'articles classés et rejetés.

Classe ETUD

Journal :	Le Devoir
Section :	Idées, p. A9
Titre :	Grève étudiante : au-delà des sous
Auteur :	Diane Lamoureux - Professeure au Département de science politique de l'Université Laval
Date :	Mercredi 4 avril 2012
Critère de sélection :	Arguments favorables à la continuité de la mobilisation Plaidoyer pour l'accessibilité à l'éducation ou sur l'état comme garant des services de base pour tous les citoyens.
	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. Certes, la hausse des droits de scolarité n'est pas un prétexte. Comme se plaît à le rappeler la ministre de l'Éducation, Line Beauchamp, c'est avec une rare unanimité que la FECQ, la FEUQ et la CLASSE ont refusé de participer à des consultations bidon où il n'était pas question de remettre en cause la hausse, seulement d'en discuter les modalités d'application. En réitérant leur refus lors de la manifestation de novembre 2011, puis en

déclenchant un mouvement de grève à partir de la mi-février, les étudiants ont clairement indiqué qu'ils refusaient la hausse des droits de scolarité.

Le mouvement de grève a d'ailleurs (presque) débuté avec une action de solidarité entre les étudiants et la Coalition contre la hausse des tarifs afin de bloquer la tour de la Bourse, le 16 février. Contrairement à ce qu'affirme le gouvernement, le mouvement étudiant, en se solidarisant avec les groupes communautaires, féministes et syndicaux qui s'opposent notamment à la « taxe santé » uniforme et à la hausse des tarifs d'électricité, voulait montrer qu'il n'entendait pas défendre de supposés privilèges, mais s'opposer à une mesure sectorielle (la hausse des droits de scolarité) qui s'inscrit dans un mouvement plus large de démantèlement de la nature publique des services gouvernementaux et de tarification de ceux-ci selon le principe de « l'usager-payeur ».

Depuis le début de cette vague de mobilisation étudiante, on peut voir un déplacement des enjeux : on reparle maintenant de plus en plus ouvertement de gratuité scolaire et on dénonce les gabegies administratives des directions universitaires, plus préoccupées de béton et de compétitivité (surtout celle de leurs salaires) que de formation intellectuelle. 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.

On aurait tort d'attribuer une telle transformation au caractère fallacieux de la revendication initiale ou à la fourberie de certains leaders étudiants. C'est plutôt l'oeuvre du caractère politisant de la lutte. Si plusieurs avaient des espoirs en déclenchant le mouvement, personne ne pouvait prédire le cours qu'il prendrait et la formidable leçon de science politique qu'en tireront ses participants. En agissant collectivement, en inventant des slogans, en prenant le temps de discuter, en arpentant les rues des villes pour faire autre chose que se déplacer, en profitant du soleil de ce printemps inespéré, les militantes et les militants donnent chair à ces valeurs fondamentales des sociétés démocratiques que sont l'égalité, la liberté et la solidarité.

Face au tournant néolibéral accentué dans les politiques québécoisesdepuis

	le fameux « déficit zéro » de Lucien Bouchard et réitéré par les gouvernements successifs à Québec (et à Ottawa), la grève étudianterappelle que l'éducation n'est pas une marchandise que l'on débite à la pièce selon la capacité de payer du « client » et que l'on choisit en fonction de sa rentabilité supposée. Elle souligne également que l'université n'est pas une entreprise dont la gestion relève de son seul conseil d'administration. Dans ce sens, on peut situer le mouvement étudiant actuel dans la foulée du printemps arabe revendiquant la démocratie, dans la logique du mouvement des indignés de Madrid ou d'Athènes contre la « discipline budgétaire », ou dans celle du mouvement Occupy, dénonçant l'accroissement des inégalités sociales. D'abord l'expression du refus d'une supposée fatalité (tous les prix augmentent, pourquoi pas lesdroits de scolarité?). Ensuite une formidable expérience qui fera sentir ses effets à long terme : la recherche, non sans quelques tâtonnements (c'est si certains plaquaient une solution toute faite sur le mouvement qu'il faudrait s'inquiéter, pas vis-à-vis de ses hésitations) d'un autre monde plus juste et plus démocratique. 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...

Classe GOUV

Journal :	Le Quotidien
Section :	Chronique, p. 10
Titre :	L'accessibilité n'est pas menacée
Auteur :	Manoj Karivelil, Jonquière
Date :	Lundi 2 avril 2012
Critère de sélection :	Plaidoyer pour une vision marchande de l'éducation ou l'idée de l'éducation comme un investissement individuel Arguments contre les demande des étudiants. Arguments en faveur de la hausse.

	Texte dénonçant la violence ou les inconvénients provoqués par les manifestations.
	À la suite de l'écoute répétitive des leaders des fédérations collégiales et universitaires (et on peut le dire sans gêne, probablement des futurs leaders syndicaux) dans les médias, j'en suis venu à m'en faire un devoir de répliquer sur leurs revendications. Ainsi, on parle sans cesse de l'accessibilité aux études supérieures, et comment une hausse des frais de scolarité diminuera celle-ci. Faux! Le Québec représente probablement l'endroit où l'accessibilité aux études supérieures est la meilleure en Amérique du Nord, et elle le restera. Un étudiant venant d'un milieu défavorisé ne diminuera aucunement ses chances d'avoir accès aux études supérieures. Le programme de prêts et bourses québécois permet à tous ceux qui désirent entreprendre des études supérieures de le faire. À partir de cela, le goût et le désir de poursuivre des études supérieures reviennent à lui-même et sa famille. À noter ici que les provinces ayant des frais de scolarité beaucoup plus élevés ont des taux de diplomation égaux, voire supérieurs aux Québec. On nous bombarde sans cesse ce montant de 1625$. On parle ici d'une augmentation de 163$ par session étalée sur cinq ans. En réalité, même si on parlait d'une hausse de 1625$ par année dès maintenant, le coût additionnel à l'étudiant représente moins de 14$ par jour d'étude. Rappelons ici qu'on parle de monter la contribution de l'étudiant à 17% de la facture totale, 83% de la vraie facture sera assumée par l'ensemble des contribuables. Les leaders étudiants expriment tous le fait qu'ils ont l'appui de la population dans leurs demandes. Faux! J'en conviens, ils ont l'appui de la population pour que le gouvernement retourne à la table des négociations avec eux. En effet, l'empêchement des contribuables d'aller travailler, et le blocage des rues et ponts, font que la classe moyenne désire que le conflit se règle. Ce n'est pas la même chose qu'un appui de leurs demandes idéologiques. La forte majorité de la population n'appuie pas leurs demandes de gel ou de gratuité scolaire, car ces augmentations affecteront directement la classe moyenne, qui croupit déjà avec le fardeau fiscal le plus élevé en Amérique du Nord. On parle toujours des chiffres monstres de 100 000 à 300 000 étudiants en

	grève. Faux! On fait preuve ici d'une extrapolation massive de chiffres. Les leaders syndicalo-étudiants passent leurs mandats de grève avec presque toujours de très faibles majorités, donc ils ne peuvent affirmer qu'ils représentent tous les étudiants collégiaux et universitaires. En fait, ils en représentent légèrement plus que la moitié seulement. Ayant étudié moi-même à l'université, et ayant fini avec le montant maximal de prêts, j'ai évidemment pris plusieurs années en début de carrière pour rembourser cette dette. J'ai toujours considéré mon prêt étudiant comme un investissement pour moi-même et ma famille. L'augmentation proposée est juste, et le Québec restera toujours l'endroit en Amérique du Nord où les frais de scolarité seront les plus bas.

Rejeter

Journal :	Journal du Québec
Section :	Spectacles, vendredi 22 juin 2012, p. 46
Titre :	Paul piché parle...
Auteur :	Sophie Durocher
Date :	Vendredi 22 juin 2012
Critère de rejection :	Textes qui rassemblent des sujets disparates dans un seul document. Discussions politiques très écartées du sujet de la grève ou qui parlent de la grève étudiante de façon secondaire.

	1 .... DU PRINTEMPS QUÉBÉCOIS On parle de démocratie, mais la rue fait partie de la démocratie ! C'est sûr que les dérapages on s'en passerait. Ça nuit au message, ça nuit à l'opinion, ça empêche les gens de réfléchir comme du monde. Ça sert le statu quo. Mais les dérapages de l'autre côté aussi, on s'en passerait. Ceux des politiciens. On aurait bien voulu qu'ils soient plus adultes que les jeunes

dans la rue.

2 ... DE PAULINE MAROIS QUI ENLÈVE SON CARRÉ ROUGE

Je trouve ça très correct. Elle l'a porté pendant toute la session parlementaire, elle l'a défendu clairement, elle a expliqué pourquoi elle le portait. Elle n'a jamais reculé. Maintenant, la session est finie, on s'en va en élection. Il n'y a pas juste le conflit étudiant. Elle veut devenir première ministre de tous les Québécois, et il y a beaucoup d'autres causes qui ont été très oubliées (au grand plaisir de Jean Charest). Alors elle remet sa fleur de lys, et moi je trouve ça bien correct. Je ne vois pas du tout un désaveu. Je ne vois pas non plus de l'opportunisme. Elle ne va quand même pas porter le carré rouge ... sur la plage !

3 ... DE STAR ACADÉMIE

EN TOURNÉE

J'ai participé à tous les shows de Montréal et Québec, sauf un, pour chanter Sur ma peau avec Andrée-Anne. C'est une belle gang de jeunes. J'aime l'esprit de cette cuvée-là. C'est un show hyper nationaliste. Je trouve que le show de Star Académie est plus nationaliste que ce qu'on va faire à la St Jean. Avec l'hymne à Québec, on déroule le gros drapeau bleu et blanc sur la foule ! Et personne ne chiale, tout le monde est content, on se reconnaît là-dedans ! C'est le drapeau du Québec, ils sont nationalistes, ils chantent Piché et Desjardins. Ils remettent la chanson québécoise en avant. Ils chantent en anglais, et on ne vient pas fous. C'est un show très émotif.

Annexe C. Outils de textométrie et de fouille de textes

TXM

TXM est une plateforme logicielle construite dans le cadre du projet *ANR Textométrie*[33], financée par le programme *ANR Corpus et outils de la recherche en sciences humaines et sociales*. Le logiciel est destiné à l'analyse de grands corpus textuels et comprend les calculs textométriques comme l'analyse factorielle des correspondances, le calcul des spécificités et le calcul de cooccurrences. *TXM* permet de connaître le comportement de données textuelles dans le corpus telles que leur évolution temporelle, les contrastes qu'elles révèlent en fonction de comparaisons effectuées entre les textes ou sous-corpus, ainsi que leur attraction à l'intérieur d'un texte ou d'un corpus. L'interprétation des résultats de ces calculs est facilitée par le concordancier, qui permet de voir le contexte d'apparition de chaque donnée textuelle dans une fenêtre de mots de quantité paramétrable à gauche et à la droite de la donnée cible.

TXM possède également un moteur de recherche natif, le *Corpus Query Processor* (CQP), qui permet de construire des requêtes ciblées, en combinant plusieurs opérateurs avancés. À travers de la construction de requêtes avec le langage *Corpus Query Language* (CQL), il est possible de cibler la recherche en fonction des traitements effectués, par exemple, la recherche de lemmes, ou la recherche de données textuelles correspondant à une étiquette de la partie du discours.

Le choix du logiciel dans le cadre de notre recherche n'a pas été le résultat d'un travail comparatif, mais il est basé sur certains critères considérés importants par la chercheure. D'abord, il s'agit d'un logiciel libre et gratuit, disponible pour les systèmes opérationnels *Windows* et *Mac OS X*. La disponibilité de matériels explicatifs, comme des tutoriels, des forums de discussion et des vidéos gratuites et téléchargeables sur le web, a également motivé ce choix : plusieurs questions et problèmes éprouvés lors de l'utilisation ont été promptement répondus par la communauté d'usagers, qui est très active sur les réseaux sociaux. Par rapport

[33] 'Agence Nationale de la Recherche, sous la tutelle du ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation en France.

à d'autres logiciels textométriques comme *Lexico3*, nous avons trouvé d'autres avantages, comme la possibilité d'intégrer facilement un étiqueteur de la partie du discours, la mise à jour fréquente, la simplicité de l'interface, les modules d'exportation des données, et l'installation sur le système *Mac OS X* (qui n'est pas disponible pour *Lexico3)*. En plus, la littérature analysée dans notre travail cite fréquemment l'utilisation de *TXM*, et cela nous a permis d'avoir une base de comparaison entre nos résultats et ceux des travaux cités.

Nous avons installé sur un poste *Mac OS X Yosemite* version 10.10.5 la version 0.7.7 du logiciel *TXM*. Par défaut, le *TXM* réalise la segmentation des textes en formes graphiques discrètes (mots simples). Nous avons également intégré le dictionnaire français du logiciel *TreeTagger*, afin de lemmatiser le corpus et d'effectuer l'étiquetage morphosyntaxique des données textuelles. Le corpus importé dans *TXM* est composé par des fichiers en format texte (TXT), et chaque fichier contient le titre et le contenu de chaque article. Les métadonnées des articles sont définies dans un fichier séparé en format CSV. Nous avons identifié chaque article du corpus par la métadonnée ID qui correspond nom du fichier. Les autres métadonnées sont la DATE, contenant la valeur de la date de publication (DATE=00-00-000) et la CLASSE pour la classe de l'article (classe=ETUD ou classe=GOUV). Les paramètres de l'encodage de caractères choisi ont été le UTF-16 et l'option pour la langue française.

Pour le segmenteur lexicale, qui effectue la segmentation du corpus dans des mots simples, nous avons défini les paramètres par défaut suggérés par le logiciel : caractères blancs, caractères de ponctuation caractères d'élision, et caractères de fin de phrase.

Dans *TXM*, la création des sous-corpus pour le calcul textométrique se fait par le biais de la fonctionnalité partitions. Pour la création de la partition, le logiciel *TXM* demande de choisir comme paramètres « l'unité structurelle du texte » et la « propriété ». Les paramètres de sélection pour « unité structurelle » sont le texte entier, le paragraphe ou la phrase. Ceux qui concernent la « propriété » correspondent aux métadonnées définies pour le corpus. Nous avons créé les partitions ETUD et GOUV en utilisant comme propriété la métadonnée CLASSE et comme unité structurelle le texte.

Sur *TXM*, nous pouvons lancer le calcul des spécificités directement sur la partition créée ou en créant une table lexicale. L'avantage de la table lexicale c'est qu'elle permet de

visualiser les spécificités des deux partitions dans un seul tableau et de cibler le type d'unité textuel (mot, lemme ou étiquette morphosyntaxique) qui est désigné dans le logiciel par le terme « valeur de propriété ». Ainsi, nous avons créé trois tables lexicales, chacune correspondant à une valeur de propriété différente.

Lors de l'application du calcul des spécificités, le logiciel *TXM* présente les résultats dans un tableau (Figure 1). Les lignes affichent les données textuelles choisies comme valeur de propriété (mot, lemme et partie du discours) et les colonnes, les partitions créées et les scores attribués aux données textuelles, qui peuvent être positifs ou négatifs. Les scores positifs signalent la spécificité de l'unité sur la partition considérée tandis que le score négatif signale sa rareté. Pour chaque donnée textuelle, le tableau présente sa fréquence totale dans le corpus, sa fréquence dans chaque sous-corpus et les scores de spécificité.

Partition: word | Partition: word

Unités	Fréquence T 300039	ETUD t=162796	score	GOUV t=137243	score
éducation	312	232	12,7	80	-12,7
nous	796	526	11,1	270	-11,1
société	353	251	10,2	102	-10,2
«	323	228	8,9	95	-8,9
*	291	207	8,6	84	-8,6
»	324	227	8,4	97	-8,4
mépris	34	33	7,6	1	-7,6
notre	283	198	7,3	85	-7,3
Université	159	117	6,4	42	-6,4
Courchesne	41	37	6,1	4	-6,1
ministre	433	284	6,0	149	-6,0
jeunesse	94	73	5,7	21	-5,7
autochtones	24	23	5,0	1	-5,0
savoir	83	64	4,9	19	-4,9
grève	361	235	4,7	126	-4,7
Parent	38	33	4,7	5	-4,7
Outremont	17	17	4,5	0	-4,5
Éducation	103	76	4,4	27	-4,4
et	5033	2869	4,4	2164	-4,4
jeunes	277	183	4,4	94	-4,4
Desjardins	36	31	4,3	5	-4,3

Figure 1. Calcul des spécificités lancé sur les partitions ETUD et GOUV. Les scores de la classe ETUD triés en ordre décroissant.

Lexico3

Développé par André Salem, Serge Fleury, Cédric Lamalle et William Martinez, *Lexico3* est un logiciel qui permet de mener des analyses contrastives et chronologiques sur de grands corpus. Les principales fonctionnalités sont l'inventaire de segments répétés et le

concordancier, qui permet de relever les contextes environnant un mot simple ou un segment répété. Il est possible de lancer des calculs de spécificités sur les mots simples et sur des segments répétés en fonction de partitions créées sur le corpus d'articles. Le logiciel offre également une fonctionnalité pour effectuer des analyses factorielles. Des graphiques de ventilation en fréquences absolues, relatives ou histogrammes de spécificités peuvent être générés, pour avoir une visualisation du comportement du lexique sur le corpus entier ou sur les partitions créées à partir de ce dernier.

Dans le cadre de notre recherche, nous avons utilisé *Lexico3* seulement pour effectuer le calcul des spécificités sur les segments répétés, puisque ce type de segmentation textuelle n'était pas disponible sur *TXM*. *Lexico3* offre une fonction permettant de lancer la segmentation et d'obtenir des segments répétés en fonction de deux paramètres qui sont la longueur du segment en termes de nombre de mots, et la fréquence minimale à considérer pour le repérage du segment. Après la segmentation effectuée, Lexico3 présente la liste de segments répétés obtenus avec leur longueur et leur fréquence.

Pour effectuer le calcul des spécificités dans *Lexico3*, il faut d'abord constituer des partitions à partir du corpus d'articles. Une fois le corpus partitionné, le calcul peut être lancé en cliquant sur le bouton section « Principales caractéristiques lexicométriques ».

Nous avons organisé le corpus de notre recherche dans un fichier TXT contenant tous les articles, identifiés par un entête qui contenait les métadonnées. Pour la segmentation des textes dans des segments répétés, nous avons défini comme paramètres un seuil de probabilité de 5 et une fréquence minimale de 10, laquelle a été définie heuristiquement, en analysant les paramètres qui présentaient des résultats optimaux. Lorsque la fréquence minimale est plus basse, les segments répétés présentés par le logiciel sont plus nombreux, mais plus « bruités » : il y a des segments longueur plus petite qui se trouve dans la chaîne de caractères de segments plus longs. Puisqu'il n'y a pas une manière automatique de ramener des segments qui expriment les mêmes concepts, nous avons préféré d'utiliser une fréquence minimale qui pourrait faciliter le travail de nettoyage des segments non pertinents.

Les partitions ont été créées en fonction des classes ETUD et GOUV, lesquelles ont été définies dans les métadonnées des entêtes. Par défaut, Lexico3 organise les partitions en

fonction des textes. À la suite du calcul des spécificités *Lexico3* a généré comme résultat un tableau contenant les mots et les segments répétés, les valeurs de fréquence dans le corpus et dans la partie considérée, ainsi que le score de spécificité. Nous avons exporté cette liste dans un fichier XML, puis nous avons organisé les données dans un tableau sur *Excel*. Ensuite, nous avons fait l'élimination de segments répétés non pertinents.

Weka

Développé par l'Université Waikato en Nouvelle-Zélande, *Weka* propose une importante collection d'algorithmes d'apprentissage et d'outils pour la classification automatique. Il fournit également un support très efficace pour assister le processus de fouille de textes, qui inclut la préparation de données, une interface pour l'évaluation statistique de modèles d'apprentissage, et un module de visualisation des données. Le logiciel permet également de comparer les performances de différents algorithmes d'apprentissage. *Weka* est écrit en langage Java et est distribué gratuitement sous les termes de la *General Public License* (GNU) (Witten et Frank, 2005).

L'ouvrage *Data Mining : Practical Machine Learning Tools and Techniques* (Witten et Frank, 2005), utilisé dans notre recherche, se consacre à l'explication des principes fondamentaux de la fouille de données (incluant la fouille de textes) et offre un tutoriel assez exhaustif du fonctionnement de *Weka*. L'existence de cet ouvrage nous a particulièrement motivés dans le choix du logiciel, mais nous l'avons également comparé à une autre option de logiciel libre dans le marché, appelé *RapidMiner*. Par rapport à ce dernier, *Weka* offre quelques avantages : la disponibilité de filtres pour effectuer des traitements sur les données et un outil de visualisation qui permet de regarder la distribution de données dans un diagramme de dispersion.

D'autres critères particuliers à *Weka* nous ont motivé son choix dans le cadre de notre recherche : la facilité à importer les données dans le logiciels à partir du format CSV, qui est le format d'exportation d'autres logiciels utilisés dans le cadre de notre étude comme *TXM* et *WordStat* ; utilisation d'un format de représentation des données non propriétaire et facilement manipulable (format ARFF) ; disponibilité de filtres permettant de traiter ou de sélectionner les données, ce qui a été particulièrement utile pour la conversion des fréquences absolues en

fréquences binaires ; disponibilité des algorithmes choisis dans le cadre de la recherche (bayésien naïf et SVM) ; présence de modules d'évaluation de classifieurs assez puissants comme le temps pour performer la tâche, les mesures de rappel et de précision et la génération de la matrice de confusion, que nous a permis de calculer plus facilement l'autre mesure AUC utilisé dans notre recherche ; et finalement, la présence d'un outil pour faire la validation croisée sur le corpus.

Dans le cadre de notre recherche, nous avons utilisé comme entrée dans *Weka* des fichiers ARFF directement, sans passer par le module de conversion de fichiers CSV du logiciel. Nous avons également utilisé un filtre pour convertir les fréquences absolues en fréquences binaires. Les matrices converties ont été sauvegardées dans des fichiers ARFF séparés à la suite de la conversion. Par rapport à l'outil de classification (fonction *Classify*), nous avons gardé les paramètres par défaut du logiciel pour chaque algorithme de classification utilisé pour les tests effectués avec les lignes de comparaisons et les matrices constituées pour l'expérimentation. Dans les images ci-dessous figurent les paramètres par défaut du logiciel.

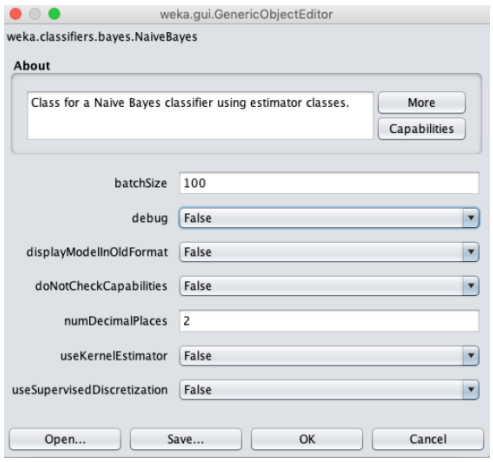


Figure 1. Paramètres par défaut pour l'algorithme bayésien naïf

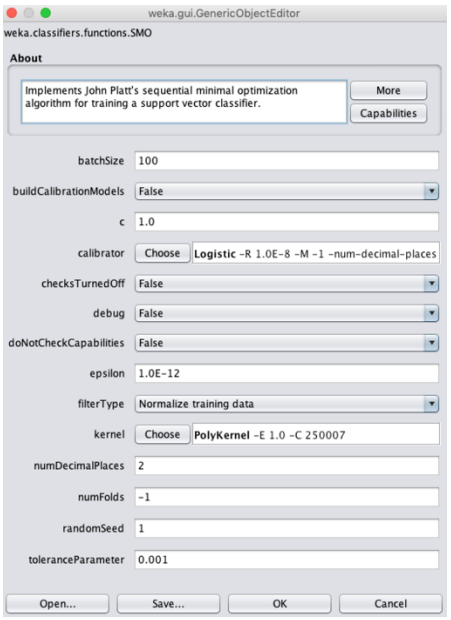


Figure 2. Paramètres par défaut pour l'algorithme machine à vecteurs de support

Annexe D. Liste de données textuelles éliminées

Mots

Donnée textuelle	Motivation pour l'exclusion
a	contextes différents
ça	contextes différents
Ce	contextes différents
des	contextes différents
deux	contextes différents
division	contextes différents
Outremont	contextes différents
piste	contextes différents
Renaud	contextes différents
terme	contextes différents
tiens	contextes différents
Université	contextes différents
y	contextes différents
Y	contextes différents
moment	contextes différents
1420	contextes différents
Mont-Royal	contextes différents
la	contextes différents
55	contextes différents
en	contextes différents
part	contextes différents
couches	contextes différents
occuper	contextes différents
En	contextes différents
d'	contextes différents
doctorat	contextes différents
classe	contextes différents
nouveaux	contextes différents
marché	contextes différents
ans	contextes différents
Elles	contextes différents
oeil	contextes différents
André	contextes différents
Faculté	contextes différents
27	contextes différents
projet	contextes différents
question	contextes différents
de	contextes différents
États	contextes différents
à	contextes différents

Donnée textuelle	Motivation pour l'exclusion
jaune	contextes différents
dégager	contextes différents
autre	contextes différents
sur	contextes différents
ou	contextes différents
suivre	contextes différents
même	contextes différents
crédits	contextes différents
A	contextes différents
vert	contextes différents
derniers	contextes différents
origine	contextes différents
lequel	contextes différents
complet	contextes différents
sous	contextes différents
Coalition	contextes différents
si	contextes différents
Martine	entité nommée repétée
Michelle	entité nommée repétée
Pauline	entité nommée repétée
Réjean	entité nommée repétée
Fred	entité nommée repétée
Nadeau-	entité nommée repétée
logiciels	présent dans un seul article
marre	présent dans un seul article
autochtone	présent dans un seul article
Bouret	présent dans un seul article
Brière	présent dans un seul article
Chevrier	présent dans un seul article
loteries	présent dans un seul article
Loto-Québec	présent dans un seul article
*	style de la source
DE	style des titres
DES	style des titres
ET	style des titres
l'	style des titres
LA	style des titres
LE	style des titres
LES	style des titres
PAS	style des titres

Il	contextes différents
St-Jean	contextes différents
relevé	contextes différents
remporté	contextes différents
résidents	contextes différents
régime	contextes différents
opposition	contextes différents
mener	contextes différents
Paris	contextes différents
cas	contextes différents

UN	style des titres
UNE	style des titres
DANS	style des titres
EST	style des titres
NE	style des titres
DEUX	style des titres
[	utilisé comme paranthèses
]	utilisé comme paranthèses

Lemmes

Donnée textuelle	Motivation pour l'exclusion
/	contextes différents
asseoir	contextes différents
avoir	contextes différents
cela	contextes différents
de	contextes différents
deux	contextes différents
du	contextes différents
en	contextes différents
joindre	contextes différents
libre	contextes différents
Michelle	contextes différents
moment	contextes différents
Outremont	contextes différents
projet	contextes différents
reconnaître	contextes différents
Renaud	contextes différents
Université	contextes différents
y	contextes différents
appeler	contextes différents
appartenir	contextes différents
Jeanne	contextes différents
États	contextes différents
total	contextes différents
peur	contextes différents
autre	contextes différents
ou	contextes différents
marché	contextes différents
agiter	contextes différents
être	contextes différents
André	contextes différents

Donnée textuelle	Motivation pour l'exclusion
régime	contextes différents
sous	contextes différents
sur	contextes différents
revenir	contextes différents
voter	contextes différents
profit	contextes différents
fait	contextes différents
jaune	contextes différents
résistance	contextes différents
il	contextes différents
opposition	contextes différents
retenir	contextes différents
fonder	contextes différents
terme	contextes différents
St-Jean	contextes différents
compléter	contextes différents
résident	contextes différents
LE	contextes différents
part	contextes différents
favoriser	contextes différents
nouveau	contextes différents
prochain	contextes différents
savoir	contextes différents
Mont-Royal	contextes différents
Pauline	entité nommée repétée
Fred	entité nommée repétée
Nadeau-	entité nommée repétée
cour| cours	erreur de lemmatisation
frai|frais	erreur de lemmatisation
logiciel	présent dans un seul article

monter	contextes différents
soin	contextes différents
rapporter	contextes différents
mère	contextes différents
complet	contextes différents
après	contextes différents
coalition	contextes différents
doctorat	contextes différents
classe	contextes différents
cas	contextes différents
Paris	contextes différents
distance	contextes différents
boîte	contextes différents
rang	contextes différents
technologie	contextes différents
immigrant	contextes différents
jour	contextes différents
espoir	contextes différents

loterie	présent dans un seul article
marre	présent dans un seul article
remporter	présent dans un seul article
Loto-Québec	présent dans un seul article
alliance	présent dans un seul article
concert	présent dans un seul article
Bouret	présent dans un seul article
Brière	présent dans un seul article
DE	style des titres
ET	style des titres
LA	style des titres
LES	style des titres
PAS	style des titres
DES	style des titres
DANS	style des titres
[	utilisé comme paranthèses
]	utilisé comme paranthèses

Segments répétés

Donnée textuelle	Motivation pour l'exclusion
de 30	contextes différents
Université de Montréal	contextes différents
beaucoup d	contextes différents
Université du Québec	contextes différents
toute la	contextes différents
comme s	contextes différents
moins d	contextes différents
quatre ans	contextes différents
un seul	contextes différents
pour ce	contextes différents
qui lui	contextes différents
celui de l	contextes différents
celui de	contextes différents
de ses	contextes différents
temps de	contextes différents
est en	contextes différents
à la	contextes différents
à tout	contextes différents
est tout	contextes différents
il dit	contextes différents
de toute	contextes différents
sur cette	contextes différents

Donnée textuelle	Motivation pour l'exclusion
que cette	non pertinent
que le gouvernement	non pertinent
rouge à	non pertinent
classe et	non pertinent
mettre de l	non pertinent
cela se	non pertinent
Québécois sont	non pertinent
Dans le	non pertinent
décidé d	non pertinent
tendance à	non pertinent
capables de	non pertinent
eau dans	non pertinent
intentions de	non pertinent
besoin de	non pertinent
compte de	non pertinent
nos institutions	non pertinent
de Québécois	non pertinent
de droite	non pertinent
de manifester	non pertinent
était la	non pertinent
de questions	non pertinent
nos universités	non pertinent

au même	contextes différents
pour qu	contextes différents
par jour	contextes différents
a rien de	contextes différents
fin de semaine	contextes différents
de moins	contextes différents
a rien	contextes différents
y en	contextes différents
Il ne	contextes différents
ou non	contextes différents
par les	contextes différents
les autres	contextes différents
les deux	contextes différents
à voir	contextes différents
et qui	contextes différents
en avant	contextes différents
le fond	contextes différents
les plus	contextes différents
en a	contextes différents
la violence	non pertinent
du Québec	non pertinent
de l	non pertinent
de société	non pertinent
la société	non pertinent
la classe	non pertinent
de grève	non pertinent
ministre de	non pertinent
ministre de l	non pertinent
du savoir	non pertinent
la jeunesse	non pertinent
la solidarité	non pertinent
accès à l	non pertinent
de savoir	non pertinent
la construction	non pertinent
la force	non pertinent
me souviens	non pertinent
le dos	non pertinent
Université de	non pertinent
de la violence	non pertinent
de droits	non pertinent
de la population	non pertinent
de la grève	non pertinent
de solidarité	non pertinent
de s	non pertinent
de nouvelles	non pertinent
de droits de	non pertinent
session d	non pertinent
hausses des	non pertinent
idée d	non pertinent
du même	non pertinent
du débat	non pertinent
rappeler que	non pertinent
du revenu	non pertinent
du Parti québécois	non pertinent
crédits d	non pertinent
du cégep	non pertinent
du coût	non pertinent
pour dénouer	non pertinent
régime de	non pertinent
contre le	non pertinent
aux yeux	non pertinent
sont devenus	non pertinent
gens qui	non pertinent
au revenu	non pertinent
majorité des	non pertinent
Au Québec	non pertinent
la gauche	non pertinent
la faute	non pertinent
la crise	non pertinent
la majorité des	non pertinent
la hausse	non pertinent
la hausse des	non pertinent
la contribution	non pertinent
de temps	non pertinent
la CAQ	non pertinent
la Coalition	non pertinent
la FTQ	non pertinent
la social	non pertinent
la même	non pertinent
la session	non pertinent
la prochaine	non pertinent
la province	non pertinent
de la province	non pertinent
de la CLASSE	non pertinent
de pression	non pertinent
de frais	non pertinent
de 325	non pertinent
des études	non pertinent
des deux	non pertinent
des journalistes	non pertinent
qui a été	non pertinent
qui doivent	non pertinent

de 75%	non pertinent
la chance	non pertinent
de comprendre	non pertinent
la communauté	non pertinent
droit à l	non pertinent
où on	non pertinent
éducation est	non pertinent
mais la	non pertinent
ministre a	non pertinent
droits de	non pertinent
place de	non pertinent
refuse de	non pertinent
baisse de	non pertinent
art de	non pertinent
permettra de	non pertinent
enquête sur	non pertinent
adoption de	non pertinent
commission d	non pertinent
jeunesse qui	non pertinent
santé et	non pertinent
niveau de	non pertinent
accès à	non pertinent
milliards de	non pertinent
projet de	non pertinent
parole de	non pertinent
police de	non pertinent
le sens	non pertinent
le savoir	non pertinent
des droits de	non pertinent
le développement	non pertinent
le premier	non pertinent
le ministre	non pertinent
et que nous	non pertinent
des mois	non pertinent
des entreprises	non pertinent
à la violence	non pertinent
les États	non pertinent
la police	non pertinent
la police de	non pertinent
la possibilité	non pertinent
la population	non pertinent
la guerre	non pertinent
la grève	non pertinent
la ministre de	non pertinent
la marchandisation	non pertinent
les générations	non pertinent
qui le	non pertinent
qui m	non pertinent
qui ne	non pertinent
que la hausse des	non pertinent
que le gel	non pertinent
un autre	non pertinent
est une	non pertinent
les mains	non pertinent
les manifestants	non pertinent
le plan	non pertinent
les libéraux	non pertinent
le fait	non pertinent
le budget	non pertinent
le carré rouge	non pertinent
le PQ	non pertinent
des casseroles	non pertinent
des choix	non pertinent
le système	non pertinent
les frais	non pertinent
les hausses	non pertinent
le remboursement	non pertinent
des contribuables	non pertinent
bonification du	non pertinent
dénouer l	non pertinent
les associations	non pertinent
La démocratie	non pertinent
la position	non pertinent
des hausses	non pertinent
que la hausse	non pertinent
des prêts et	non pertinent
le carré	non pertinent
le fardeau	non pertinent
des artistes	non pertinent
sur Twitter	non pertinent
de la session	non pertinent
un carré	non pertinent
de remboursement	non pertinent
de gauche	non pertinent
annuler la	non pertinent
étudiants qui	non pertinent
des prêts	non pertinent
prêts et	non pertinent
boycottage des	non pertinent
le gel	non pertinent
le gel des	non pertinent
gel des	non pertinent

les valeurs	non pertinent
les revendications	non pertinent
la sécurité	non pertinent
la question	non pertinent
les entreprises	non pertinent
la violence et	non pertinent
étudiants pour	non pertinent
dans la société	non pertinent
gouvernement du	non pertinent
Québec à	non pertinent
gouvernement qui	non pertinent
par la	non pertinent
un gouvernement qui	non pertinent
un étudiant	non pertinent
un véritable	non pertinent
du Québec en	non pertinent
une famille	non pertinent
du droit	non pertinent
une baisse de	non pertinent
une commission	non pertinent
une commission d	non pertinent
aux entreprises	non pertinent
les manifestations	non pertinent
la capacité	non pertinent
ministre du	non pertinent
la FEUQ	non pertinent
de la police	non pertinent
ministre de la	non pertinent
contre les	non pertinent
société qui	non pertinent
débat de	non pertinent
question de la	non pertinent
la retraite	non pertinent
éducation et	non pertinent
la ministre de l	non pertinent
et je	non pertinent
recteurs et	non pertinent

les artistes	non pertinent
Ce n	non pertinent
des cours	non pertinent
des frais	non pertinent
des frais de	non pertinent
* * *	non pertinent
frais de	non pertinent
* *	non pertinent
Y en a marre	présent dans un seul article
Y en a	présent dans un seul article
en a marre	présent dans un seul article
y en a	présent dans un seul article
sur le dos des	segment déjà sélectionne
de plus	segment déjà sélectionne
la classe moyenne	segment déjà sélectionne
de plus en	segment déjà sélectionne
Université du Québec en	segment déjà sélectionne
Pour joindre notre	segment déjà sélectionne
président de la	segment déjà sélectionne
président de	segment déjà sélectionne
dos des	segment déjà sélectionne
présidente de la	segment déjà sélectionne
étudiants n	segment déjà sélectionne
plus d	segment déjà sélectionne
hausse de	segment déjà sélectionne
la hausse des frais de	segment déjà sélectionne
accessibilité aux	segment déjà sélectionné : accessibilité aux études
les associations étudiantes	segment déjà sélectionné : associations étudiantes
Les associations étudiantes	segment déjà sélectionné : associations étudiantes
Les associations	segment déjà sélectionné : associations étudiantes
que les associations étudiantes	segment déjà sélectionné : associations étudiantes
cours des	segment déjà sélectionné : au cours des
terme de	segment déjà sélectionne : au terme
la crise étudiante	segment déjà sélectionné : crise étudiante
de ne pas	segment déjà sélectionné : de ne
plus en	segment déjà sélectionne : de plus en plus
la hausse des droits de	segment déjà sélectionne : droits

un droit	non pertinent
un point	non pertinent
une loi	non pertinent
ministère de l	non pertinent
une société	non pertinent
est un droit	non pertinent
est l	non pertinent
un Québec	non pertinent
début de la	non pertinent
juste et	non pertinent
des étudiants en	non pertinent
des policiers	non pertinent
la ministre	non pertinent
le PLQ	non pertinent
la circulation	non pertinent
pas se	non pertinent
la façon	non pertinent
La ministre	non pertinent
de scolarité	non pertinent
une forme	non pertinent
du régime	non pertinent
sur la gestion	non pertinent
le cas de	non pertinent

scolarité	de scolarité
	segment déjà sélectionné : droits
des droits de scolarité	de scolarité
	segment déjà sélectionné : droits
de droits de scolarité	de scolarité
	segment déjà sélectionné : en
en cinq	cinq ans
	segment déjà sélectionné : est
pas parce	pas parce
	segment déjà sélectionné : et
ceux qui	ceux
	segment déjà sélectionné :
aux études universitaires	études univesitaires
	segment déjà sélectionné : frais
que les frais de scolarité	de scolarité
	segment déjà sélectionné : frais
frais de scolarité à	de scolarité
	segment déjà sélectionné : frais
les frais de	de scolarité
	segment déjà sélectionné : frais
les frais de scolarité	de scolarité
hausse des frais de	segment déjà sélectionné : frais
scolarité	de scolarité
	segment déjà sélectionné : frais
hausse des frais de	de scolarité
	segment déjà sélectionné : frais
des frais de scolarité	de scolarité
	segment déjà sélectionné : gel
gel des frais de scolarité	des frais
	segment déjà sélectionné : gel
le gel des frais	des frais
	segment déjà sélectionné : gels
le gel des droits	des droits
	segment déjà sélectionné :
de son gouvernement	gouvernement
	segment déjà sélectionné :
que la hausse des droits	hausse des droits
la hausse des frais de	segment déjà sélectionné :
scolarité	hausse des frais
	segment déjà sélectionné :
la hausse des frais	hausse des frais
	segment déjà sélectionné : il ne
Il ne faut pas	faut
	segment déjà sélectionné : il ne
ne faut pas	faut
y a	segment déjà sélectionné : Il y a

de tous les	non pertinent
une majorité	non pertinent
la presse	non pertinent
les gens	non pertinent
la position du	non pertinent
étudiants et de	non pertinent
des militants	non pertinent
la suite	non pertinent
la raison	non pertinent
à 7	non pertinent
faire un	non pertinent
pour une	non pertinent
que tu	non pertinent
les médias	non pertinent
que les autres	non pertinent
que le	non pertinent

Il y	segment déjà sélectionné : Il y a
étudiants eux	segment déjà sélectionné : les étudiants eux-mêmes
de la loi 78	segment déjà sélectionné : loi 78
la loi 78	segment déjà sélectionné : loi 78
que nous sommes	segment déjà sélectionné : nous sommes
pas besoin de	segment déjà sélectionné : pas besoin
de la FEUQ	segment déjà sélectionné : présidente de la FEUQ
des prêts et bourses	segment déjà sélectionné : prêts et bourses
sur le dos des étudiants	segment déjà sélectionné : sur le dos
hausse des droits de scolarité	segment déjà sélectionné : droits de scolarité
de frais de scolarité	segment déjà sélectionné : frais de scolarité
il pas	Segment non trouvé dans *TXM*
parole étudiants	Segment non trouvé dans *TXM*
Conflit étudiant	titre de section
Grève étudiante	titre de section

Coocurrents

Donnée textuelle	Motivation pour l'exclusion
étudiant-demander	contextes différents
étudiant-compter	contextes différents
étudiant-exposer	contextes différents
étudiant-représenter	contextes différents
étudiant-contribuer	contextes différents
étudiant-issu	contextes différents
étudiant-mauvais	contextes différents
étudiant-technique	contextes différents
québec-paradis	contextes différents
québec-particulier	contextes différents
québec-élever	contextes différents
québec-juin	contextes différents
québec-moyen	contextes différents
québec-débrider	contextes différents
québec-continent	contextes différents
québec-traire	contextes différents

Donnée textuelle	Motivation pour l'exclusion
gouvernement-voler	contextes différents
droit-canadien	contextes différents
droit-favoriser	contextes différents
droit-ignorant	contextes différents
droit-cause	contextes différents
droit-maximum	contextes différents
droit-modeste	contextes différents
droit-relatif	contextes différents
droit-saint	contextes différents
hausse-concerner	contextes différents
hausse-constater	contextes différents
hausse-élever	contextes différents
hausse-principal	contextes différents
hausse-recul	contextes différents
hausse-principe	contextes différents
hausse-toucher	contextes différents

québec-compter	contextes différents
québec-étranger	contextes différents
québec-avancer	contextes différents
québec-international	contextes différents
québec-briller	contextes différents
québec-passer	contextes différents
québec-rappeler	contextes différents
scolarité-indispensable	contextes différents
scolarité-prétendre	contextes différents
scolarité-fond	contextes différents
gouvernement-accabler	contextes différents
gouvernement-an	contextes différents
gouvernement-casser	contextes différents
gouvernement-désastreux	contextes différents
gouvernement-engranger	contextes différents
gouvernement-glisser	contextes différents
gouvernement-renversement	contextes différents
gouvernement-peiner	contextes différents
gouvernement-populaire	contextes différents
gouvernement-pourchasser	contextes différents
gouvernement-pouvoir	contextes différents
gouvernement-stratégique	contextes différents
gouvernement-tenir	contextes différents

hausse-uniforme	contextes différents
québec-collégial	nom propre
québec-coalition	nom propre
québec-journal	nom propre
québec-sûreté	nom propre
scolarité-perron	nom propre
étudiant-accréditer	présent dans un seul article
étudiant-roi	présent dans un seul article
scolarité-changement	présent dans un seul article
scolarité-jeudi	présent dans un seul article
scolarité-mécanique	présent dans un seul article
gouvernement-sourd	présent dans un seul article
gouvernement-intégral	présent dans un seul article
gouvernement-minuit	présent dans un seul article
droit-emporter	présent dans un seul article
droit-légal	présent dans un seul article
droit-changement	présent dans un seul article
droit-jeudi	présent dans un seul article
hausse-scie	présent dans un seul article
hausse-mécanique	présent dans un seul article
hausse-rapide	présent dans un seul article
hausse-AVANT	style des titres

Annexe E. Mémos : analyse des concordances

Mots

Critère	Interprétation du contexte	Composante	Classe	Spéc.
artistes	Critique sur l'appui des artistes sur les étudiants, mépris envers cette classe, défenseur de causes de gauche, jeunesse, rêve, idéologie.	Thématique	GOUV	14
éducation	Beaucoup plus présente dans ETUD. On parle de l'éducation supérieure, de l'accessibilité à l'éducation, système d'éducation, éducation publique, éducation du Québec, l'éducation n'est pas une marchandise. Dans GOUV, coût de l'éducation, éducation de qualité, se payer, l'éducation comme service.	Thématique	ETUD	13
nous	Énonciateur proche, prise en charge.	Dialogique	ETUD	11
Mais	en majuscule, figure dans un contexte où un argument est présenté puis rejeté dans la phrase subséquente.	Dialectique	GOUV	11
société	Référence à la société québécoise, appel à une notion de collectivisme. Question ou choix de société, société québécoise, débat, modèle de société, démocratique, civile, solide, inclusif, équitable, notre société. Dans le contexte GOUV, société québécoise, notre société et aussi société démocratique.	Thématique	ETUD	10
«	Discours rapporté, pour marquer la différence entre l'énonciation et les autres « voix » présentes.	Dialogique	ETUD	9
»	Discours rapporté, pour marquer la différence entre l'énonciation et les autres « voix » présentes.	Dialogique	ETUD	8
gauche	Référence à « la gauche » en tant qu'entité idéologique. Aussi, idéologie de gauche.	Thématique	GOUV	8
Marois	Entité nommée, Pauline Marois, au sujet des élections, aussi critique à l'appui du PQ aux étudiants.	Thématique	GOUV	8
mépris	Une seule occurrence dans le sous-corpus GOUV. Mépris des politiciens et journalistes envers les étudiants et les manifestants.	Thématique	ETUD	8
notre	Énonciateur proche, prise en charge.	Dialogique	ETUD	7
hausses	La forme au pluriel est plus présente dans GOUV et se réfère à la hausse des frais.	Thématique	GOUV	7
frais	Frais de scolarité. Contraste avec droits, terme préféré dans le sous-corpus ETUD.	Thématique	GOUV	7
remboursement	Proposition de compensation à la hausse des frais. Remboursement de la dette.	Thématique	GOUV	6
boycott	Suremploi de « boycott » par opposition à « grève » dans ETUD.	Thématique	GOUV	6
Courchesne	Entité nommée. Ministre de l'Éducation qui a remplacé la précédente et qui poursuit les négociations avec les associations étudiantes.	Thématique	ETUD	6
ministre	Désignation fréquemment utilisée pour se référer aux ministres de l'Éducation.	Thématique	ETUD	6
jeunesse	Référence à la nouvelle génération d'étudiants, les étudiants universitaires.	Thématique	ETUD	6
beurre	Expression « Ils veulent le beurre et l'argent du beurre ». Les auteurs jugent les exigences des étudiants abusives.	Thématique	GOUV	5
pas	Phrases négatives.	Dialectique	GOUV	5
ont	Présent de l'indicatif, « Ils ont ». Référence à la 3e personne du pluriel (ils).	Dialogique	GOUV	5
n'	Phrases négatives.	Dialectique	GOUV	5
autochtones	Il y a une seule occurrence du mot dans GOUV. Non-accessibilité aux études	Thématique	ETUD	5

	des peuples autochtones. Employé dans un contexte où on parle du taux de diplomation chez les autochtones.			
militants	Manière de définir les étudiants. Beaucoup moins présent dans ETUD.	Thématique	GOUV	5
savoir	Économie du savoir, institutions de savoir, partage du savoir, commercialisation du savoir. Le sens du mot dans sa forme substantive dans GOUV est absent.	Thématique	ETUD	5
…	Points de suspension : utilisé pour exprimer la perplexité, la dérision, et aussi pour dire que l'argument contraire est vague. Exprime aussi le prolongement d'une réflexion.	Dialectique	GOUV	5
grève	Terme utilisé dans ETUD par opposition à « boycott » dans GOUV. Manière de qualifier le mouvement des étudiants.	Thématique	ETUD	5
gel	Le gel est une demande étudiante et se présente comme une préoccupation fréquente chez les défenseurs du gouvernement. On parle des conséquences néfastes du gel.	Thématique	GOUV	5
Twitter	Entité nommée. Utilisation du Twitter par les manifestants et tenants de la grève étudiante.	Thématique	GOUV	5
Les	Pronom démonstratif au pluriel et en majuscule au début des phrases, associé généralement à une personne (les étudiants, les membres, les mécontents, etc.).	Dialogique	GOUV	5
carré	Référence au carré rouge, symbole de la lutte étudiante. Mentionne d'autres carrés.	Thématique	GOUV	4
Éducation	Entité nommée associée à des noms d'institutions (par exemple, ministère de l'Éducation).	Thématique	ETUD	4
syndicales	Se réfère aux centrales ou organisations syndicales et sur leur appui financier au mouvement de grève.	Thématique	GOUV	4
humoristes	Absent dans le corpus étudiant. Exprime de l'ironie envers les humoristes à cause de leur appui à la grève et aux étudiants.	Thématique	GOUV	4
cours	Dans la majorité des cas analysés, se réfère au boycottage de cours et aux suspensions dues à la grève étudiante.	Thématique	GOUV	4
et	Structuration cumulative du récit.	Dialectique	ETUD	4
jeunes	Utilisé pour ces référer aux étudiants, aux personnes qui doivent avoir accès à l'éducation.	Thématique	ETUD	4
majorité	La plupart de contextes montrent qu'il s'agit de la majorité d'individus favorables aux arguments du gouvernement ou contre les étudiants. Moins présent dans ETUD.	Thématique	GOUV	4
boycottage	Qualification du mouvement comme un « boycottage » de cours par opposition à « grève ».	Thématique	GOUV	4
Desjardins	Référence à Martine Desjardins, leader étudiant de la FÉUQ, Fédération étudiante universitaire du Québec.	Thématique	ETUD	4
Madame	Désignation, pronom de traitement, utilisé dans un sens ironique.	Dialogique	ETUD	4
ne	Phrases négatives.	Dialectique	GOUV	4
-il	Présence de formulations interrogatives. Style polémique, exprime de l'indignation.	Dialectique	ETUD	4
économie	Le système économique québécois, l'économie du savoir, démocratisation de l'économie, critique à l'économie du marché et au capitalisme.	Thématique	ETUD	4
Québécois	Entité nommée, se réfère à des noms d'institutions et au peuple québécois.	Thématique	GOUV	4
centrales	Référence aux centrales syndicales et à leur support financier au mouvement de	Thématique	GOUV	4

	grève.			
associations	Utilisé plus dans GOUV pour « associations étudiantes ».	Thématique	GOUV	4
ils	Présence massive du pronom en troisième personne dans GOUV.	Dialogique	GOUV	4
bas	Plus présent dans GOUV. Frais de scolarité est qualifié comme bas, surtout par rapport à d'autres provinces canadiennes. Dans ETUD, « bas » est également associé à frais, mais dans ce contexte les auteurs reprennent l'argument du groupe contraire pour le rebattre.	Thématique	GOUV	4
-nous	Fonction argumentative : formulation de questions.	Dialectique	ETUD	4
Beauchamp	Entité nommée. Ministre de l'Éducation du Québec qui a participé au début des négociations avec les étudiants.	Thématique	ETUD	4
juste	Contestation sur la formule de la « juste part » dans le discours du gouvernement, plus souvent cité entre guillemets. Aussi, l'utilisation de l'adjectif « juste ». La formule « juste part » n'est pas assez présente dans GOUV.	Thématique	ETUD	4
accès	Accès à l'éducation. Les étudiants réclament un meilleur accès. Du côté GOUV, accès aux prêts et bourses, accès aux salles de cours, accès à l'éducation marquent un emploi plus varié.	Thématique	ETUD	4
entreprises	Thème sur la subvention et l'aide fiscale gouvernementale aux entreprises et en plus petit nombre, les universités gérées comme des entreprises. Cette thématique n'est pas abordée du côté GOUV.	Thématique	ETUD	4
inflation	Presque absent de ETUD. La hausse est qualifiée de raisonnable parce que les droits de scolarité n'ont jamais été ajustés en fonction de l'inflation. Aussi, proposition d'indexer les frais à l'inflation.	Thématique	GOUV	4
votent	Relatif au vote pour la grève des étudiants dans les assemblées et contestation de ce droit.	Thématique	GOUV	4
développement	Expression de la préoccupation du modèle de développement économique, ou contestation d'un modèle, préconisation d'un type de développement.	Thématique	ETUD	4
session	Se refaire à la perte de la session dans le cas de la continuation du mouvement de grève.	Thématique	GOUV	4
côté	Du côté des étudiants, du gouvernement — opposition de perspectives, introduction de l'opinion des parties du conflit.	Thématique	GOUV	4
on	Présence massive du pronom personnel on. Exclusion du sujet parlant.	Dialogique	GOUV	4
,	Structuration cumulative du récit.	Dialectique	ETUD	4
communauté	Désignation, ensemble d'étudiants, professeurs, dirigeants, etc.	Thématique	ETUD	4
2012	Mention à l'année du mouvement.	Dialectique	ETUD	4
accessible	Thème sur l'accès à l'éducation, la lutte des étudiants a pour objectif de rendre l'éducation plus accessible.	Thématique	ETUD	4
bonification	Plus présent dans GOUV. Parle de la proposition du gouvernement de bonification du programme de prêt et bourses. Dans ETUD, le terme est repris pour critiquer ladite proposition.	Thématique	GOUV	4
construction	Construction de campus, corruption dans le secteur de la construction, gaspillage d'argent.	Thématique	ETUD	4
rouges	Référence au symbole des étudiantes (carrés rouges).	Thématique	GOUV	3
inégalités	Absent dans GOUV. Il fait référence à l'agrandissement des inégalités entre les classes sociales.	Thématique	ETUD	3

sentiment	Absent dans GOUV. Sentiment des étudiants, parle d'un sentiment collectif et partagé.	Dialogique	ETUD	3
Barbe	Critique contre l'argument ad hominem utilisé par l'écrivain québécois Jean Barbe contre le ministre des Finances, Raymond Bachand, partisan des étudiants. Barbe a écrit sur Facebook et Twitter : « C'est Raymond Bachand qui a dit que bloquer un centre-ville était inacceptable ? Bachand, tu te souviens quand tu as essayé de pogner les fesses de ma blonde, à Paris, quand tu étais soul ? Tsé, pendant le party de la première de Notre-Dame-de-Paris ? C'était inacceptable. Je t'ai pas vargé dessus avec une matraque. Prends note, stp ».	Thématique	GOUV	3
Dubois	Entité nommée se référant à Gabriel-Nadeau Dubois, leader étudiant de la CLASSE, association étudiante la plus radicale du mouvement.	Thématique	GOUV	3
fasciste	Les auteurs dans GOUV se défendent d'être accusés de « fasciste » par quelques tenants de la grève étudiante.	Thématique	GOUV	3
minimum	Proposition pour indexer les frais de scolarité au salaire minimum.	Thématique	GOUV	3
commission	Les étudiants revendiquent une commission d'enquête pour faire une investigation sur la corruption dans le secteur de la construction. Aussi, enquête sur la question du financement des universités.	Thématique	ETUD	3
provinces	Plus présent dans GOUV pour comparer Québec avec d'autres provinces concernant la question de l'éducation publique et le financement des universités. Thème moins fréquent dans ETUD.	Thématique	GOUV	3
veulent	Présent de l'indicatif, ils veulent. Référence à la 3e personne du pluriel (ils).	Dialogique	GOUV	3
demain	Organisation temporelle : référence au temps futur.	Dialectique	ETUD	3
présidente	Désignation.	Thématique	ETUD	3
art	Utilisé dans le sens d'habileté, par exemple, « l'art de gouverner ».	Thématique	ETUD	3
PQ	Entité nommée. Critiques au Parti Québécois.	Thématique	GOUV	3
annuler	Discussion sur l'annulation de la hausse dans un éventuel changement de pouvoir.	Thématique	GOUV	3
idéologique	Beaucoup plus présent dans GOUV. On caractérise la grève comme un enjeu idéologique.	Thématique	GOUV	3
social-démocratie	Une seule occurrence dans ETUD. Discussion sur le modèle de social-démocratie.	Thématique	GOUV	3
instances	Instances supérieures, institutions réglementaires.	Thématique	ETUD	3
stratégies	Sens de combat, de manipulations, etc.	Thématique	ETUD	3
perturbations	Manière de se référer aux actes des étudiants. Presque absent de ETUD.	Thématique	GOUV	3
impasse	Par rapport à la grève et l'impasse existant entre étudiants et le gouvernement.	Thématique	GOUV	3
résistance	Au sujet de la résistance des étudiants face à la fermeté du gouvernement.	Thématique	ETUD	3
génération	Référence aux étudiants, à la génération d'étudiants, celle qui se bénéficiera du gel des droits de scolarité et de la mobilisation actuelle des étudiants.	Thématique	ETUD	3
population	Population du Québec.	Thématique	ETUD	3
budget	Discussions sur le budget provincial. Plus présent dans GOUV.	Thématique	GOUV	3
reconnaître	« Il faut reconnaître », formule qui exprime la nécessité d'avoir le bon sens.	Thématique	ETUD	3
annoncées	Mesures annoncées par le gouvernement, hausse, bonifications. Absent dans ETUD.	Thématique	GOUV	3
violence	Dénonciation de la violence contre les étudiantes.	Thématique	ETUD	3
manifestants	Utilisé pour se référer aux étudiants.	Thématique	GOUV	3

enseignement	Enseignement universitaire ou supérieur, vocabulaire plus légaliste, qualité de l'enseignement.	Thématique	ETUD	3
quelqu'un	Pronom indéterminé utilisé tantôt pour se référer à l'interlocuteur, tantôt à soi-même (« traiter quelqu'un de fasciste »). Effacement énonciatif.	Dialogique	GOUV	3
moment	Référence au présent, à la manifestation, ou aux évènements clés du mouvement de grève. Aussi présent dans le côté GOUV, mais moins nombreux.	Dialectique	ETUD	3
journalistes	Thème sur les insultes dirigées aux journalistes par les étudiants, protection aux journalistes et à liberté d'expression.	Thématique	GOUV	3
carrés	Référence au symbole des étudiants en grève. Manière ironique et dérisoire de se référer aux étudiants.	Thématique	GOUV	3
citoyenne	Le mot est employé dans le contexte de participation citoyenne et aussi comme une forme d'identification, par exemple, « je suis une mère citoyenne ».	Thématique	ETUD	3
responsabilités	Appel à la responsabilité et au respect à l'ordre : les étudiants doivent avoir des responsabilités.	Thématique	GOUV	3
véritable	Utilisé pour parler du véritable enjeu de la grève ou qualifier le mouvement et sa contribution à la société. Accentuation.	Dialectique	ETUD	3
leaders	Critiques aux leaders des étudiants.	Thématique	GOUV	3
bourses	Propositions pour améliorer le programme de prêts et bourses du gouvernement, défense d'un régime de bonification, présentation d'une solution alternative.	Thématique	GOUV	3
75	Montant en pourcentage de l'augmentation.	Thématique	ETUD	3
administrations	Le critère se réfère aux administrations universitaires (au pluriel) et aussi collégiales (cégeps). Propos critiques de l'attaque des administrations au mouvement, comme par exemple, la suspension de la session.	Thématique	ETUD	3
ceux	Référence aux étudiants dans la plupart, mais aussi à d'autres intervenants non nommés (par exemple, ceux qui dénoncent la hausse) — mise à distance de l'énonciateur.	Dialogique	GOUV	3
revenu	Propositions d'un remboursement proportionnel au revenu, ou propositions pour redéfinir le rapport entre le revenu et l'aide sociale.	Thématique	GOUV	3
aujourd'hui	Référence au temps présent, non seulement chronologique, mais de l'actualité, des jours actuels.	Dialectique	ETUD	3
dénouer	Proposition de solutions pour dénouer l'impasse, la crise. Exprime la volonté de cesser le mouvement.	Thématique	GOUV	3
famille	Revenus familiales, soutient de la famille, relations familiales avec les étudiants en grève.	Thématique	ETUD	3
adoption	Adoption de la loi 78 qui limite la tenue de manifestations.	Thématique	ETUD	3
voulons	Nous voulons, dans la plupart. Demandes des étudiants, modalité assertorique et positionnement certain. Conjugaison : nous.	Dialogique	ETUD	3
avenir	Sur les conséquences pour l'avenir de la société ou des étudiants.	Thématique	ETUD	3
gens	Référence générales à des personnes non identifiées, mise à distance et généralisation par rapport à l'interlocuteur.	Dialogique	GOUV	3
M	Abréviation de Monsieur. Façon de s'adresser à des politiciens.	Dialogique	ETUD	3
péréquation	Allusion au système de péréquation qui privilégie le Québec au détriment des autres provinces.	Thématique	GOUV	3
je	Énonciateur égocentrique, forme présente dans les lettres signées.	Dialogique	ETUD	3
affiche	À propos de l'exposition médiatique reçue par les étudiants.	Thématique	GOUV	3

méfaits	Manière de qualifier les manifestations, les perturbations causées par les étudiants	Thématique	GOUV	3
Pellerin	Mention sur Fred Pellerin, poète et compositeur québécois, qui se rallie aux étudiants.	Thématique	GOUV	3
salope	Injure dirigée à des femmes qui critiquent le mouvement des étudiants, particulièrement Janette Bertrand et Sophie Durocher.	Thématique	GOUV	3
TOUS	Utilisation du mot en majuscule (TOUS), polémique, idée de perplexité et d'indignation.	Dialectique	GOUV	3
serait	Verbe conjugué au conditionnel présent, dénotant l'hypothèse et la possibilité.	Dialogique	GOUV	3
hommes	Hommes et femmes reviennent fréquemment dans les occurrences. Pour se référer à la société, hommes et femmes, inclusivement.	Thématique	ETUD	3
souviens	À propos de la devise québécoise : je me souviens.	Thématique	ETUD	3
diviser	Au sujet des tentatives de diviser le mouvement étudiant, particulièrement par rapport à la non-reconnaissance des franges plus radicales du mouvement de grève (CLASSE).	Thématique	ETUD	3
pacifiques	Manière de qualifier les manifestations des étudiants.	Thématique	ETUD	3
libéraux	Façon de se référer au gouvernement.	Thématique	GOUV	3
groupes	Référence à des groupes organisés (étudiantes, gauche, association, groupes de pression, manifestants, etc.).	Thématique	GOUV	3
guerre	Manière de qualifier le conflit du point de vue des images et de l'opinion publique. Combat médiatique.	Thématique	ETUD	3
solidarité	Solidarité entre les personnes. Propos positifs sur cette solidarité démontrée par la société.	Thématique	ETUD	3
michele	Courriel pour contacter le chroniqueur qui a écrit l'article.	Dialogique	ETUD	3
ouimet	Courriel pour contacter le chroniqueur qui a écrit l'article.	Dialogique	ETUD	3
province	Utilisé pour parler du Québec, la situation économique de la province.	Thématique	GOUV	3
asseoir	Dans le sens de négocier. Asseoir sur la table de négociation.	Thématique	ETUD	3
chronique	Référence au genre journalistique écrit ou lu par l'auteur auparavant (« ma chronique », « la lecture de cette chronique »).	Thématique	GOUV	3
fermeté	Manière de qualifier la position du gouvernement. Les propos sont tantôt critiques, tantôt panégyriques.	Thématique	GOUV	3
sent	Avoir l'impression de quelque chose, ressentir. Se refaire à l'énonciateur (on sent).	Dialogique	GOUV	3
allez	Impératif. Manière d'interpeller les étudiants.	Dialogique	GOUV	3
candidats	Allusion à la campagne électorale.	Thématique	GOUV	3
propositions	Propositions du gouvernement et des étudiants.	Thématique	GOUV	3
Pas	Négation au début des phrases, polémique.	Dialectique	GOUV	3
valeurs	Thème sur les valeurs de la société qui doivent être priorisées.	Thématique	ETUD	3
compromis	Nécessité d'un compromis entre le gouvernement et les étudiants.	Thématique	GOUV	3
petit	Dans plusieurs contextes, se réfèrent aux actions des étudiants, à leur symbole, de leur nombre. Thème sur la faible représentativité des étudiants.	Thématique	GOUV	3
J'	Énonciateur égocentrique, présent dans les lettres principalement.	Dialogique	ETUD	3
quelle	Associé à des questions (quelle importance, quelle valeur, questionnement).	Dialectique	ETUD	3
images	Référence aux images des manfestations diffusées dans les médias pour dénigrer	Thématique	ETUD	2

	les étudiants.			
commun	Dans plusieurs contextes, le terme est utilisé pour les biens communs et patrimoines. Valorisation du service public.	Thématique	ETUD	2
générations	Exprime la préoccupation des conséquences de l'augmentation de frais sur les générations, sur l'avenir.	Thématique	ETUD	2
Depuis	Début, à partir de, structuration du temps.	Dialectique	ETUD	2
péquiste	Référence au Parti Québécois, l'opposition de ce parti au gouvernement des libéraux et leur défense du gel des droits de scolarité.	Thématique	GOUV	2
président	Se référant aux paroles des présidents de plusieurs institutions (syndicats, fédérations éducationnelles, etc.).	Thématique	ETUD	2
1960	Année de la mise en place du programme fédéral d'aide aux étudiants. Référence aussi à la Révolution tranquille.	Dialectique	ETUD	2
arrogance	Critique sur l'arrogance du gouvernement et des politiciens.	Thématique	ETUD	2
accessibilité	Les auteurs défendent que les propositions gouvernementales ne menacent pas ladite accessibilité, thème qui est fréquemment mentionné par les étudiants comme étant une conséquence néfaste de la hausse.	Thématique	GOUV	2
votre	Dans les lettres adressées à des personnes dans le conflit.	Dialogique	ETUD	2
annulation	Exprime l'enjeu de l'annulation de la session en fonction de la grève.	Thématique	GOUV	2
voulez	Phrases interrogatives, dans la plupart des cas dirigez aux étudiants.	Dialectique	GOUV	2
bacon	Comparaison entre les étudiants et les bacons cuisants dans la poêle : « ils se tortillent par terre pour obtenir ce qu'ils désirent », comme le bacon.	Thématique	GOUV	2
boutonnière	Critiques au carré rouge porté à la boutonnière, propos méprisants sur symbole de la grève étudiante.	Thématique	GOUV	2
défendu	Sur les défenses des différents acteurs impliqués dans la dispute, souligne les différences des positions.	Thématique	GOUV	2
proposées	Au sujet des solutions proposées par le gouvernement pour dénouer l'impasse, la crise. Exprime la volonté de cesser le mouvement.	Thématique	GOUV	2
test	Vocabulaire scolaire, utilisé dans un sens ironique : test que les étudiants doivent passer avec la grève.	Thématique	GOUV	2
tu	Attaques directes à une personne en particulier. Aussi utilisé dans le contexte d'exemples.	Dialogique	GOUV	2
leur	Mise à distance énonciatif, référence aux étudiants dans la plupart.	Dialogique	GOUV	2
chiffres	Au sujet des chiffres présentés par les deux parties en conflit, contestation de certains chiffres ou demande à présenter des chiffres réels.	Thématique	GOUV	2
camarades	Façon de se référer aux gens qui supportent les étudiants (leurs camarades).	Thématique	GOUV	2
parentale	Thème au sujet de la contribution parentale sur laquelle la bourse est calculée.	Thématique	GOUV	2
promis	Promesse du parti québécois d'abolir la hausse.	Thématique	GOUV	2
Hitler	Critique sur la comparaison de Jean Charest à Hitler retrouvé dans les réseaux sociaux.	Thématique	GOUV	2
idéologiques	Les étudiants sont plutôt concernés par des enjeux idéologiques, selon l'opinion des auteurs GOUV.	Thématique	GOUV	2
micro	Le micro ici est utilisé dans le contexte de « prendre la parole » et se réfère aux déclarations des étudiants dans les assemblés ou dans les médias. Utilisé surtout dans un contexte critique et méprisant.	Thématique	GOUV	2
Ouest	« Ouest » se réfère aux autres provinces canadiennes sur la question de la	Thématique	GOUV	2

	péréquation et la distribution de ressources et richesses dans le Canada, qui bénéficie de manière particulière le Québec.			
ponts	Dénonciation du blocage des ponts par les étudiants, thème sur les perturbations causées par le mouvement.	Thématique	GOUV	2
leurs	Mise à distance énonciatif, référence aux étudiants dans la plupart.	Dialogique	GOUV	2
Ils	Mise à distance énonciatif. Le pronom réfère aux étudiants dans la plupart.	Dialogique	GOUV	2
-ils	Polémique, présent dans les questions.	Dialectique	GOUV	2
Bourassa	Référence à la gestion du premier ministre Robert Bourassa, qui a procédé au dégel des frais de scolarité en 1990.	Thématique	ETUD	2
endettés	Sur le problème d'endettement des étudiants que la hausse pourrait causer.	Thématique	ETUD	2
lapresse	Courriel pour contacter le chroniqueur qui a écrit l'article.	Dialogique	ETUD	2
pourrir	Pourrir le conflit et les discussions. Le gouvernement a laissé pourrir le conflit par son refus de négocier avec les étudiants.	Thématique	ETUD	2
réussite	Réussite scolaire, réussite des étudiants dans leur vie individuelle.	Thématique	ETUD	2
campus	Employé à propos des projets immobiliers sur les campus, mais aussi pour exprimer le lieu où se passe le conflit.	Thématique	ETUD	2
--	Le double tiret atteste un style particulier des auteurs GOUV, pour donner une explication supplémentaire.	Dialectique	GOUV	2
?	Présence de phrases interrogatives.	Dialectique	GOUV	2
supérieur	À propos de l'enseignement supérieur, valorisation de cet enseignement et thématique sur l'accessibilité.	Thématique	ETUD	2
corruption	Propos sur les cas de corruption au niveau gouvernemental	Thématique	ETUD	2
démocratie	Thématique sur la nécessité de respecter la démocratie, explications sur ce qui signifie la démocratie, le respect de l'ordre.	Thématique	GOUV	2
contribuable	Contre-argument sur le thème sur le poids que l'éducation représente aux contribuables québécois. Aussi, les étudiants d'aujourd'hui deviennent de contribuables de demain.	Thématique	ETUD	2
salaire	Sur les salaires que les étudiants pourront avoir après avoir fréquenté l'université.	Thématique	GOUV	2
électorale	Les auteurs mentionnent la campagne électorale comme solution possible pour dénouer le conflit.	Thématique	GOUV	2
citoyen	Manière de se référer à un individu qui s'oppose à « gens » du côté GOUV.	Thématique	ETUD	2
joindre	Joindre courriel.	Dialogique	ETUD	2
critique	Thème sur la critique, de la nécessité d'avoir la pensée critique.	Thématique	ETUD	2
2005	Année des modifications du régime de prêts et bourses. Projet de coupure de 103 M $ dans le programme d'aide financière aux études.	Dialectique	ETUD	2
imposé	Impositions du gouvernement, ou de l'idéologie de ce dernier et de la hausse comme une imposition. Contraste avec proposition qui est plus fréquent du côté GOUV.	Thématique	ETUD	2
iPhone	Réaction des étudiants contre la critique disant que l'aide financière aux études leur permet d'acheter des iPhone (c'est-à-dire, qu'ils sont gâtés).	Thématique	ETUD	2
Lors	Marque de moments précis du mouvement étudiant, des décisions prises.	Dialectique	ETUD	2
pacifique	Manière de se référer à la manifestation.	Thématique	ETUD	2
Toutefois	Contre-argumentation, opposition.	Dialectique	ETUD	2
UdeM	Thème sur les investissements à l'UdeM (construction du nouveau campus).	Thématique	ETUD	2

eux-mêmes	Référence aux étudiants, mise à distance énonciatif.	Dialogique	GOUV	2
travailler	Dans la plupart des contextes, thème sur la nécessité de travailler pour payer les études ou travailler après les études pour payer la dette.	Thématique	ETUD	2
familial	Sur les modifications du régime de prêt offert par le gouvernement : aucune contribution parentale jusqu'à un revenu de 60 000 par année.	Thématique	GOUV	2
instruction	Comme synonyme d'éducation, spécialement dans des contextes où on parle de l'éducation des enfants.	Thématique	ETUD	2
matraques	Dénonciation de la violence contre les étudiants et contre les « gardes armées de matraques dans le campus ».	Thématique	ETUD	2
moments	Référence aux événements présents, valorisation des moments comme ceux de la mobilisation.	Dialectique	ETUD	2
ONU	Au sujet de la contestation de l'ONU sur l'inconstitutionnalité de la loi 78 qui vise à contrôler les manifestations.	Thématique	ETUD	2
Professeur	Signature à la fin de l'article, identification personnelle de l'auteur.	Dialogique	ETUD	2
réveil	Réveil politique, réveil de la génération.	Thématique	ETUD	2
Rocher	Mentions à l'appui de Guy Rocher, sociologiste québécois, à la grève étudiante, qui a siégé à la commission Parent.	Thématique	ETUD	2
paient	Thématique sur les coûts de l'éducation et sur les personnes qui devrait payer ces coûts. Les citoyens paient beaucoup d'impôts, alors les étudiants devraient payer leur juste part.	Thématique	GOUV	2
campagne	Les auteurs mentionnent la campagne électorale comme solution possible pour dénouer le conflit.	Thématique	GOUV	2
indexation	Au sujet de l'indexation de tarifs, une mesure qui peut éviter les crises politiques.	Thématique	GOUV	2
$	Présence de chiffres et montants, thématique sur les coûts de l'éducation.	Thématique	GOUV	2
étudiant	Au singulier, « étudiant » figure sur des exemples (« un étudiant, dépense par étudiant). Aussi utilisé comme forme d'identification (« je suis un étudiant »).	Thématique	ETUD	2
contribuables	Au sujet des conséquences de la continuité du gel sur les contribuables.	Thématique	GOUV	2
pression	La pression exercée par le mouvement, le climat de pression en général causé par la grève.	Thématique	GOUV	2
CAQ	Coalition Avenir Québec. Préoccupation sur les élections, la campagne électorale et la succession possible des libéraux.	Thématique	GOUV	2
Bref	Terminatif.	Dialectique	GOUV	2
fardeau	Associé au fardeau fiscal que la hausse pourrait représenter pour les contribuables.	Thématique	GOUV	2
seraient	En parlant des conséquences, positionnement de l'énonciateur à l'égard de son propos. Exprime la possibilité.	Dialogique	GOUV	2
notamment	Démonstration dans l'argument, utilisation d'exemples.	Dialectique	ETUD	2
hier	Réaction des auteurs à des événements et déclarations données par d'autres personnes dans un passé récent.	Dialectique	GOUV	2
Parce	Explications, réponses à des questions formulées par l'auteur du texte.	Dialectique	GOUV	2
anarchie	Accusation contre les étudiants de perturber l'ordre et les règles démocratiques.	Thématique	GOUV	2
descendre	Descendre dans la rue, manifester, dans un sens péjoratif.	Thématique	GOUV	2
blogues	Lien pour visiter le blogue du chroniqueur.	Dialogique	GOUV	2
boycottent	Les étudiants boycottent les cours.	Thématique	GOUV	2

Brin	Critique contre Collette Brin, professeure de l'U. de Laval qui a méprisé le comportement des médias.	Thématique	GOUV	2
concrets	Les auteurs disent qu'il faut considérer les vrais problèmes, façon de déqualifier le discours de l'opposant.	Thématique	GOUV	2
dividendes	Discussion sur l'impôt, le gel représenterait un fardeau sur les contribuables québécois.	Thématique	GOUV	2
faction	Manière de se référer à la CLASSE, comme une faction radicale du mouvement.	Thématique	GOUV	2
fascisme	Commentaires sur l'attitude des étudiants de qualifier le gouvernement ou d'autres personnes qui sont contre leur mouvement de fascistes, ou de pratiquer le fascisme.	Thématique	GOUV	2
là-dessus	Style indigné des auteurs, familier.	Dialectique	GOUV	2
photographe	Actes violents pratiqués contre les professionnels de la presse (photographes).	Thématique	GOUV	2
mobilisation	Mobilisation des étudiantes, commentaire sur l'ampleur de la mobilisation, sur son pouvoir. Aussi employé pour qualifier les actions des étudiants.	Thématique	ETUD	2
encore	Structuration cumulative du récit.	Dialectique	ETUD	2

Lemmes

Critère	Interprétation du contexte	Composante	Classe	Spéc.
éducation	Beaucoup plus présente dans ETUD. On parle de l'éducation supérieure, de l'accessibilité à l'éducation, système d'éducation, éducation publique, éducation du Québec, l'éducation n'est pas une marchandise. Dans GOUV, coût de l'éducation, éducation de qualité, se payer, l'éducation comme service. AUSSI : Entités nommées associées à des noms d'institutions (par exemple, ministère de l'Éducation).	Thématique	ETUD	16
artiste	Critique sur l'appui des artistes aux étudiants, mépris envers cette classe, association avec l'idéologie de gauche, immaturité, rêve.	Thématique	GOUV	15
nous	Énonciateur proche, prise en charge.	Dialogique	ETUD	12
société	Référence à la société québécoise, appel à une notion de collectivisme. Question ou choix de société, société québécoise, débat, modèle de société, démocratique, civile, solide, inclusif, équitable, notre société. Dans le contexte GOUV, société québécoise, notre société, société démocratique.	Thématique	ETUD	11
«	Discours rapporté, pour marquer la différence entre l'énonciation et les autres « voix » présentes.	Dialogique	ETUD	9
»	Discours rapporté, pour marquer la différence entre l'énonciation et les autres « voix » présentes.	Dialogique	ETUD	9
gauche	Référence à « la gauche » en tant qu'entité idéologique. Aussi, idéologie de gauche.	Thématique	GOUV	8
ne	Forte présence de phrases négatives.	Dialectique	GOUV	8
Marois	Entité nommée, Pauline Marois.	Thématique	GOUV	8
autochtone	Sur le manque d'accessibilité aux études chez les peuples autochtones.	Thématique	ETUD	8
mépris	Une seule occurrence dans le sous-corpus GOUV. Mépris des politiciens et journalistes envers les étudiants et les manifestants.	Thématique	ETUD	7

remboursement	Proposition de compensation à la hausse des frais. Remboursement de la dette.	Thématique	GOUV	7
carré	Référence au carré rouge, symbole de la lutte étudiante. Mentionne d'autres carrés. Identifie les étudiants par leur symbole. Manière ironique et dérisoire de se référer aux étudiants.	Thématique	GOUV	7
pas	Forte présence de phrases négatives.	Dialectique	GOUV	6
Courchesne	Entité nommée. Ministre de l'Éducation qui a remplacé la ministre précédente et qui poursuit les négociations avec les associations étudiantes.	Thématique	ETUD	6
jeunesse	Référence à la nouvelle génération d'étudiants, les étudiants universitaires.	Thématique	ETUD	6
boycott	Suremploi de « boycott » par opposition à « grève » dans ETUD.	Thématique	GOUV	6
notre	Énonciateur proche, prise en charge.	Dialogique	ETUD	6
ministre	Désignation fréquemment utilisée pour se référer aux ministres de l'Éducation.	Thématique	ETUD	6
@card@	Présence marquée de chiffres, reliés à des coûts, mais aussi à certaines dates.	Thématique	ETUD	5
beurre	Expression « Ils veulent le beurre et l'argent du beurre ». Les auteurs jugent les exigences des étudiants abusives.	Thématique	GOUV	5
province	Utilisé pour parler du Québec, la situation économique de la province.	Thématique	GOUV	5
président	Désignation pour se référer aux présidents de certaines institutions.	Thématique	ETUD	5
grève	Utilisé dans ETUD par opposition à boycott dans GOUV. Manière de qualifier le mouvement des étudiants.	Thématique	ETUD	5
majorité	La plupart de contextes montrent qu'il s'agit de la majorité d'individus favorables aux arguments du gouvernement ou contre les étudiants. Moins présent dans ETUD.	Thématique	GOUV	5
humoriste	Absent dans le corpus étudiant. Discours de mépris et d'ironie envers les humoristes à cause de leur appui à la grève et aux étudiants.	Thématique	GOUV	5
gel	Le gel est une demande étudiante et se présente comme une préoccupation fréquente chez les défenseurs du gouvernement. On parle des conséquences néfastes du gel.	Thématique	GOUV	5
idéologique	Les auteurs GOUV disent que les motivations de la lutte étudiante sont plutôt idéologiques.	Thématique	GOUV	5
génération	Référence aux étudiants, à la génération d'étudiants, celle qui se bénéficiera du gel des droits de scolarité et de la mobilisation actuelle des étudiants.	Thématique	ETUD	5
…	Points de suspension : utilisé pour exprimer la perplexité, la dérision, ou pour dire que l'argument contraire est vague. Exprime aussi le prolongement d'une réflexion.	Dialectique	GOUV	5
communauté	Désignation, ensemble d'étudiants, professeurs, dirigeants, etc.	Thématique	ETUD	5
on	Présence massive du pronom personnel on. Exclusion du sujet parlant.	Dialogique	GOUV	5
boycotter	Suremploi de « boycotter » par opposition à « grève » dans ETUD.	Thématique	GOUV	5
boycottage	Qualification du mouvement comme un « boycottage » de cours par opposition à « grève ».	Thématique	GOUV	5
Twitter	Entité nommée. Utilisation du Twitter par les manifestants et tenants de	Thématique	GOUV	5

	la grève étudiante.			
pacifique	Adjectif caractérisant les mobilisations étudiantes, absent de GOUV.	Thématique	ETUD	5
bas	Plus présent dans GOUV. Frais de scolarité est qualifié comme bas, surtout par rapport à d'autres provinces canadiennes. Dans ETUD, bas se réfère aussi à des frais, mais dans ce contexte les auteurs reprennent l'argument du groupe contraire pour le rebattre.	Thématique	GOUV	5
jeune	Utilisé pour ces référer aux étudiants, aux personnes qui doivent avoir accès à l'éducation, à la jeunesse.	Thématique	ETUD	5
côté	Du côté des étudiants, du gouvernement — opposition de perspectives, introduction de l'opinion des parties du conflit. Souligne la polarisation existante.	Thématique	GOUV	4
leur	Référence à 3e personne (les étudiants), sujet parlant distancé.	Dialogique	GOUV	4
Desjardins	Référence à Martine Desjardins, leader étudiant de la FÉUQ, Fédération étudiante universitaire du Québec.	Thématique	ETUD	4
accessible	Thème sur l'accès à l'éducation, la lutte des étudiants a pour objectif de rendre l'éducation plus accessible.	Thématique	ETUD	4
Madame	Désignation, pronom de traitement, utilisé dans un sens ironique.	Dialogique	ETUD	4
inégalité	Absent dans GOUV. Référence à l'agrandissement des inégalités entre les classes sociales.	Thématique	ETUD	4
fasciste	Les auteurs de GOUV se défendent d'être accusés de « fasciste » par quelques tenants de la grève étudiante.	Thématique	GOUV	4
Québécois	Entité nommée, se réfère à des noms d'institutions et au peuple québécois.	Thématique	GOUV	4
et	Structuration cumulative du récit.	Dialectique	ETUD	4
fondamental	Relatif aux droits, aux principes de la société. Valeurs et principes.	Thématique	ETUD	4
demain	Organisation temporelle : référence au temps futur.	Dialectique	ETUD	4
Beauchamp	Entité nommée. Ministre de l'Éducation du Québec qui a participé au début des négociations avec les étudiants.	Thématique	ETUD	4
développement	Expression de la préoccupation du modèle de développement économique, ou contestation d'un modèle, préconisation d'un type de développement.	Thématique	ETUD	4
rouge	Référence au symbole des étudiantes (carrés rouges).	Thématique	GOUV	4
bonification	Plus présent dans GOUV. Parle de la proposition du gouvernement de bonification du programme de prêt et bourses. Dans ETUD, le terme est repris moins fréquemment pour critiquer ladite proposition.	Thématique	GOUV	4
accès	Accès à l'éducation. Les étudiants réclament un meilleur accès. Du côté GOUV, accès aux prêts et bourses, accès aux salles de cours, accès à l'éducation marquent un emploi plus varié.	Thématique	ETUD	4
session	Se refaire à la perte de la session dans le cas de la continuation du mouvement de grève.	Thématique	GOUV	4
,	Structuration cumulative du récit.	Dialectique	ETUD	4
hausse	Comprends les formes au pluriel et au singulier. Hausses au pluriel est plus présent dans GOUV et fait référence à la hausse des frais.	Thématique	GOUV	4
université	Entité nommée, généralement associée au nom de l'université : par exemple Université de Montréal, Université de Québec à Montréal,	Thématique	ETUD	4

	Université de Sherbrooke, etc.			
péquiste	Référence aux tenants du Parti Québécois (PQ). Propos critiques dans la plupart, thématique sur le politique et les élections.	Thématique	GOUV	4
construction	Construction de campus, corruption dans le secteur de la construction, gaspillage d'argent.	Thématique	ETUD	4
promettre	Se refaire aux promesses tenues par les parties opposantes dans le conflit. Compromis assumés par des partis politiques, acteurs des mouvements, etc.	Thématique	GOUV	4
syndical	Se réfère aux centrales et aux organisations syndicales. Dénonce leur appui financier au mouvement de grève.	Thématique	GOUV	4
association	Utilisé plus dans GOUV pour « associations étudiantes ».	Thématique	GOUV	4
humain	Référence à l'être humain, les droits de l'homme.	Thématique	ETUD	4
intimider	Les auteurs GOUV se sentent intimidés par les étudiants.	Thématique	GOUV	4
geler	Thématique du gel, opposition à la demande de gel des droits de scolarité proposée par les étudiants.	Thématique	GOUV	3
pont	Référence au blocage des ponts pendant les manifestations des étudiants.	Thématique	GOUV	3
population	Population du Québec.	Thématique	ETUD	3
ajuster	Ajustement des droits de scolarités. L'augmentation est traitée dans le sens d'un ajustement.	Thématique	GOUV	3
Dubois	Entité nommée se référant à Gabriel-Nadeau Dubois, leader étudiant de la CLASSE, association étudiante la plus radicale du mouvement.	Thématique	GOUV	3
méfait	Référence aux actes de violence des étudiants. Absent du sous-corpus ETUD.	Thématique	GOUV	3
minimum	Proposition des défenseurs du gouvernement pour indexer les frais de scolarité au salaire minimum.	Thématique	GOUV	3
violence	Dénonciation de la violence contre les étudiantes.	Thématique	ETUD	3
responsabilité	Appel aux responsabilités des parties prenantes du mouvement de grève : parents, professeurs, étudiants, gouvernement. Dans ETUD, la responsabilité se réfère à celle du gouvernement à l'égard du conflit.	Thématique	GOUV	3
PQ	Entité nommée. Critiques au Parti Québécois dans le contexte du mouvement.	Thématique	GOUV	3
enseignement	Enseignement universitaire ou supérieur, qualité de l'enseignement.	Thématique	ETUD	3
quelqu'un	Pronom indéterminé utilisé tantôt pour se référer à l'interlocuteur, tantôt à soi-même (« traiter quelqu'un de fasciste »). Effacement énonciatif.	Dialogique	GOUV	3
instruire	Instruction éducationnelle, formation scolaire, associée à l'éducation des enfants.	Thématique	ETUD	3
écrire	On retrouve beaucoup de mentions sur des choses qui ont été écrites, soit dans les médias, soit dans des livres (par exemple, en citant la pensée de certains auteurs).	Thématique	GOUV	3
social-démocratie	Une seule occurrence dans ETUD. Discussion sur le modèle de social-démocratie sous-jacente à la lutte des étudiants.	Thématique	GOUV	3
économie	Le système économique québécois, l'économie du savoir, démocratisation de l'économie, critique à l'économie du marché et au capitalisme.	Thématique	ETUD	3
instance	Instances supérieures, institutions réglementaires.	Thématique	ETUD	3

inflation	Presque absent de ETUD. Critique sur le gel et qui considère l'augmentation raisonnable parce qu'ils n'ont jamais été ajustés en fonction de l'inflation. Aussi, proposition d'indexer les frais à l'inflation.	Thématique	GOUV	3
impasse	Par rapport à la grève et l'impasse existant entre étudiants et le gouvernement.	Thématique	GOUV	3
groupe	Référence aux groupes de pression, les étudiants, les personnes organisées contrent la hausse des frais.	Thématique	GOUV	3
juste	Critiques formulées contre la formule de la « juste part » proférée par le gouvernement, plus souvent cité entre guillemets. Aussi, l'utilisation de l'adjectif « juste ». La formule « juste part » n'est pas assez présente dans GOUV.	Thématique	ETUD	3
entreprise	Thème sur la subvention, sur aide fiscale gouvernementale décernée aux entreprises et en plus petit nombre, les universités gérées comme des entreprises. Cette thématique n'est pas abordée du côté GOUV.	Thématique	ETUD	3
petit	Présence de formulations diminutives dans la manière de caractériser les étudiants et sa faible représentativité (petit groupe, petit nombre) ou encore leurs actions (petite politique) ou la futilité de leurs demandes (petite argumentation).	Thématique	GOUV	3
Barbe	Critique contre l'argument ad hominem utilisé par l'écrivain québécois Jean Barbe contre le ministre des Finances, Raymond Bachand, partisan des étudiants. Barbe a écrit sur Facebook et Twitter : « C'est Raymond Bachand qui a dit que bloquer un centre-ville était inacceptable ? Bachand, tu te souviens quand tu as essayé de pogner les fesses de ma blonde, à Paris, quand tu étais soul ? Tsé, pendant le party de la première de Notre-Dame-de-Paris ? C'était inacceptable. Je t'ai pas vargé dessus avec une matraque. Prends note, stp ».	Thématique	GOUV	3
dénouer	Proposition de solutions pour dénouer l'impasse, la crise. Exprime la volonté de cesser le mouvement.	Thématique	GOUV	3
souvenir	Référence à la devise « Je me souviens » —Je me souviens / Que né sous le lys / Je croîs sous la rose.	Thématique	ETUD	3
centrale	Référence aux centrales syndicales et à leur support financier au mouvement de grève.	Thématique	GOUV	3
revenu	Propositions d'un remboursement proportionnel au revenu, ou propositions pour redéfinir le rapport entre le revenu et l'aide sociale.	Thématique	GOUV	3
art	Utilisé dans le sens d'habileté par exemple, « l'art de gouverner ».	Thématique	ETUD	3
commun	Dans plusieurs contextes, le terme est utilisé pour les biens communs et le patrimoine. Valorisation du service public.	Thématique	ETUD	3
manifestant	Manière de se référer aux étudiants, thématique sur les manifestations et les perturbations entraînées par le mouvement.	Thématique	GOUV	3
avenir	Sur les conséquences pour l'avenir de la société ou des étudiants.	Thématique	ETUD	3
immobilier	Thème sur les projets immobiliers dans les universités qui ont motivé la décision d'augmenter les frais.	Thématique	ETUD	3
mécanisme	Solutions : mécanisme de remboursement, de redistribution, d'évaluation	Thématique	GOUV	3
menacer	Menaces causées par les étudiants. Aussi, les étudiants se sentent menacés sans raison.	Thématique	GOUV	3

refuser	Refus du gouvernement de négocier et refus des étudiants d'accepter les propositions du gouvernement.	Thématique	ETUD	3
micro	Le micro ici est utilisé dans le contexte de « prendre la parole » et se réfère aux déclarations des étudiants dans les assemblés ou dans les médias. Plusieurs propos critiques.	Thématique	GOUV	3
bourse	Propositions pour améliorer le programme de prêts et bourses du gouvernement, défense d'un régime de bonification, présentation d'une solution alternative.	Thématique	GOUV	3
parce	Explications, réponses à des questions formulées par l'auteur du texte.	Dialectique	GOUV	3
mais	Coordination d'opposition.	Dialectique	GOUV	3
adoption	Critiques contre l'adoption de la loi 78 qui limite la tenue de manifestations.	Thématique	ETUD	3
matraque	Dénonciation de la violence contre les étudiants (gardes armées de matraques dans le campus).	Thématique	ETUD	3
journaliste	Thème sur les insultes dirigées aux journalistes par les étudiants. Protection aux journalistes et à la liberté d'expression.	Thématique	GOUV	3
M	Monsieur. Pronom de traitement.	Dialogique	ETUD	3
parental	Thème au sujet de la contribution parentale sur laquelle la bourse est calculée.	Thématique	GOUV	3
appui	Thématique sur l'appui à la grève, gagne et perte d'appui.	Thématique	GOUV	3
militant	Façon de se référer aux étudiants en grève ou à leurs actions.	Thématique	GOUV	3
péréquation	Allusion au système de péréquation qui privilégie le Québec au détriment des autres provinces.	Thématique	GOUV	3
piste	Piste pour la sortie du conflit, piste de solution.	Thématique	GOUV	3
camper	Thématique sur la position adoptée par chaque camp (camper sur leurs positions, étudiants et gouvernements. Souligne la difficulté de dépasser la polarisation du débat.	Thématique	GOUV	3
Pellerin	Mention sur Fred Pellerin, poète et compositeur québécois, qui supporte les étudiants.	Thématique	GOUV	3
salope	Injure dirigée aux femmes qui ont critiqué publiquement le mouvement des étudiants, particulièrement les journalistes Janette Bertrand et Sophie Durocher.	Thématique	GOUV	3
stratégie	Façon de qualifier les actions du gouvernement pour retenir les étudiants : stratégie d'antidialogue et de division.	Thématique	ETUD	3
homme	Hommes et femmes reviennent fréquemment dans les occurrences. Pour se référer à la société, hommes et femmes, inclusivement.	Thématique	ETUD	3
quart	Chiffres, portion de la population, ou des étudiants.	Thématique	ETUD	3
gens	Référence générales à des personnes non identifiées, mise à distance et généralisation par rapport à l'interlocuteur.	Dialogique	GOUV	3
chiffre	Au sujet des chiffres présentés par les deux parties en conflit, contestation de certains chiffres ou demande à présenter des chiffres réels.	Thématique	GOUV	3
annulation	Exprime l'enjeu de l'annulation de la session en fonction de la grève.	Thématique	GOUV	3
solidarité	Solidarité entre les personnes. Propos positifs sur cette solidarité démontrée par la société.	Thématique	ETUD	3

michele	Courriel pour contacter le chroniqueur qui a écrit l'article.	Dialogique	ETUD	3
ouimet	Courriel pour contacter le chroniqueur qui a écrit l'article.	Dialogique	ETUD	3
manifestement	De façon évidente.	Dialectique	ETUD	3
fermeté	Manière de qualifier la position du gouvernement. Les propos sont tantôt critiques, tantôt panégyriques.	Thématique	GOUV	3
atténuer	Les propositions pour atténuer les effets de la hausse, atténuer le fardeau fiscal sur les contribuables.	Thématique	GOUV	3
dérive	Façon de qualifier le gouvernement, la corruption, la situation économique du Québec, la gestion des universités.	Thématique	ETUD	3
accessibilité	Les auteurs défendent que les propositions gouvernementales ne menacent pas ladite accessibilité, thème qui est fréquemment mentionné par les étudiants comme étant une conséquence néfaste de la hausse.	Thématique	GOUV	2
démocratie	Thématique sur la nécessité de respecter la démocratie, explications sur ce qui signifie la démocratie, le respect de l'ordre.	Thématique	GOUV	2
leader	Critiques aux leaders des étudiants.	Thématique	GOUV	2
tu	Attaques directes à une personne en particulier. Aussi utilisé dans le contexte d'exemples.	Dialogique	GOUV	2
pourtant	Constatation qui s'oppose au contexte.	Dialectique	ETUD	2
revendication	Sur les revendications des étudiants.	Thématique	ETUD	2
arrogance	Critique sur l'arrogance du gouvernement et des politiciens.	Thématique	ETUD	2
citoyen	Manière de se référer à un individu, homme ou femme.	Thématique	ETUD	2
bord	Utilisé dans le sens de camp, côté, pour parler de ceux qui sont mis du côté des étudiants ou du gouvernement.	Thématique	GOUV	2
bacon	Comparaison entre les étudiants et les bacons cuisants dans la poêle : « ils se tortillent par terre pour obtenir ce qu'ils désirent », comme le bacon.	Thématique	GOUV	2
boutonnière	Critiques au carré rouge porté à la boutonnière, propos méprisants sur symbole de la grève étudiante.	Thématique	GOUV	2
TOUS	Utilisation du mot en majuscule (TOUS), polémique, idée de perplexité et d'indignation.	Dialectique	GOUV	2
nazi	Critiques contre certains commentaires dans Twitter qui compare Jean Charest à Hitler ou les actions du gouvernement comme nazistes.	Thématique	GOUV	2
photographe	Actes violents contre journalistes et photographes.	Thématique	GOUV	2
sexiste	Critique contre les accusations formulées par les étudiants contre leurs opposants. Les étudiants les accusent d'être sexistes.	Thématique	GOUV	2
aujourd'hui	Référence au temps présent, non seulement chronologique, mais de l'actualité, des jours actuels.	Dialectique	ETUD	2
depuis	Début, à partir de, inchoatif, pendant un certain temps.	Dialectique	ETUD	2
mobilisation	Mobilisation des étudiantes, commentaire sur l'ampleur de la mobilisation, sur son pouvoir. Aussi employé pour qualifier les actions des étudiants.	Thématique	ETUD	2
perturbation	Manière de se référer aux actes des étudiants. Presque absent de ETUD.	Thématique	GOUV	2
famille	Au sujet des revenus familiaux, du soutien de la famille, relations familiales avec les étudiants en grève.	Thématique	ETUD	2
Ouest	« Ouest » se réfère aux autres provinces canadiennes sur la question de la	Thématique	GOUV	2

	péréquation et la distribution de richesses dans le Canada, qui bénéficie de manière particulière le Québec.			
Hitler	Critiques contre certains commentaires dans Twitter qui compare Jean Charest à Hitler ou les actions du gouvernement comme nazistes.	Thématique	GOUV	2
normalement	Souligne le caractère anormal de la situation : normalement il serait autrement.	Dialogique	GOUV	2
budget	Budget du gouvernement, mais aussi familial dans quelques contextes. Parle aussi du budget des associations.	Thématique	GOUV	2
appuyer	Divers contextes qui parlent de l'appui autant aux étudiants qu'au gouvernement.	Thématique	GOUV	2
Bourassa	Référence au gouvernement de Robert Bourassa qui a décidé en 1990 d'augmenter les frais de scolarité. Conséquences de cette décision.	Thématique	ETUD	2
indigne	Façon de qualifier le gouvernement Charest et ses actions. Aussi relatif au mouvement des indignés.	Thématique	ETUD	2
lapresse	Courriel pour contacter le chroniqueur qui a écrit l'article.	Dialogique	ETUD	2
réussite	Réussite scolaire, réussite des étudiants dans leur vie individuelle.	Thématique	ETUD	2
campus	Employé à propos des projets immobiliers sur les campus, mais aussi pour exprimer le lieu où se passe la mobilisation des étudiants.	Thématique	ETUD	2
--	Le double tiret atteste un style particulier des auteurs GOUV, pour donner une explication supplémentaire.	Dialectique	GOUV	2
corruption	Propos sur les cas de corruption au niveau gouvernemental	Thématique	ETUD	2
?	Présence de phrases interrogatives.	Dialectique	GOUV	2
compromis	Nécessité d'un compromis entre le gouvernement et les étudiants.	Thématique	GOUV	2
ministère	Ministère de l'Éducation.	Thématique	ETUD	2
position	Position des personnes dans le conflit.	Thématique	GOUV	2
mobiliser	Thème sur la mobilisation, incitation à la mobilisation, jeunesse qui se mobilise.	Thématique	ETUD	2
votre	Présent dans les lettres adressées à des personnes dans le conflit, pour se référer aux interlocuteurs.	Dialogique	ETUD	2
iPhone	Réaction des étudiants contre la critique disant que l'aide financière aux études leur permet d'acheter des iPhone (c'est-à-dire, qu'ils sont gâtés).	Thématique	ETUD	2
UdeM	Thème sur les investissements à l'UdeM (construction du nouveau campus).	Thématique	ETUD	2
répressif	Thématique sur la violence et la répression de la part du gouvernement.	Thématique	ETUD	2
unir	Dans la plupart des cas, l'union entre les étudiants et la société.	Thématique	ETUD	2
entrée	Thème sur le blocage des entrées des lieux publics.	Thématique	GOUV	2
délégué	Se refaire aux délégués du conseil général du PLQ, politiciens. Thème sur la politique.	Thématique	ETUD	2
instruction	Comme synonyme d'éducation, spécialement dans des contextes où on parle de l'éducation des enfants.	Thématique	ETUD	2
ONU	Au sujet de la contestation de l'ONU sur l'inconstitutionnalité de la loi 78 qui vise à contrôler les manifestations.	Thématique	ETUD	2
réveil	Réveil politique, réveil de la génération.	Thématique	ETUD	2
Rocher	Mentions à l'appui de Guy Rocher, sociologiste québécois, à la grève étudiante. Guy Rocher a siégé à la commission Parent.	Thématique	ETUD	2

salarial	Thématique sur l'augmentation de salaire des recteurs, des cadres et fonctionnaires de l'état.	Thématique	ETUD	2
sympathie	Au sujet des sympathisants du mouvement étudiant.	Thématique	GOUV	2
policier	Critiques diverses sur l'intervention policière, spécialement sur la violence et l'intimidation causées par la présence de policiers sur les campus.	Thématique	ETUD	2
voilà	Interjection, dévoile la raison réelle des choses, selon les étudiants.	Dialectique	ETUD	2
indexation	Au sujet de l'indexation de tarifs, une mesure qui peut éviter les crises politiques.	Thématique	GOUV	2
rattrapage	Thème sur le rattrapage du déficit : le gouvernement n'a pas eu une bonne stratégie de proposer la hausse pour rattraper le déficit causé par des années de gel.	Thématique	GOUV	2
.	Ponctuation forte dans GOUV. Contraste avec la virgule dans le côté des auteurs ETUD.	Dialectique	GOUV	2
$	Présence de chiffres et montants, thématique sur les coûts de l'éducation.	Thématique	GOUV	2
CAQ	Coalition Avenir Québec. Préoccupation sur les élections, la campagne électorale et la succession des libéraux.	Thématique	GOUV	2
répéter	Ton d'avertissement, disant que plusieurs choses ont été répétées, mais n'ont pas été entendues.	Thématique	GOUV	2
camp	Se réfère au camp de la dispute et à l'idée de guerre. Souligne la polarisation existante dans la crise étudiante.	Thématique	GOUV	2
fardeau	Associé au fardeau fiscal que la hausse pourrait représenter pour les contribuables.	Thématique	GOUV	2
procéder	Dans la majorité des contextes, le mot est lié aux actions du gouvernement par rapport à la hausse.	Thématique	ETUD	2
anarchie	Accusation contre les étudiants de perturber l'ordre et les règles démocratiques.	Thématique	GOUV	2
bêtise	Se réfère aux propos proférés par les étudiants dans les médias.	Thématique	GOUV	2
comédien	Introduis la pensée d'un comédien nommé (Claude Legault, Jacques L'Heureux). Aussi utilisé dans un sens péjoratif, personnes subventionnées par l'état qui sont partisans du mouvement de grève.	Thématique	GOUV	2
blogue	Lien pour visiter le blogue du chroniqueur.	Dialogique	GOUV	2
Brin	Critique contre Collette Brin, professeure de l'U. de Laval qui a méprisé le comportement des médias.	Thématique	GOUV	2
disproportionner	La mobilisation est disproportionnée vis-à-vis aux genres de revendications, selon les opposants des étudiants.	Thématique	GOUV	2
dividende	Vocabulaire de finances, bonification.	Thématique	GOUV	2
faction	Manière de se référer à la CLASSE, comme une faction radicale du mouvement.	Thématique	GOUV	2
fascisme	Commentaires sur l'attitude des étudiants de qualifier le gouvernement ou d'autres personnes qui sont contre leur mouvement de fascistes, ou de pratiquer le fascisme.	Thématique	GOUV	2
empresser	Impulsivité, actions non réfléchies.	Thématique	GOUV	2
là-dessus	Style indigné des auteurs, registre de langage familier.	Dialectique	GOUV	2
maman	Lié à une thématique d'infantilisation des étudiants, manière de qualifier	Thématique	GOUV	2

	leurs actions.			
annuler	Discussion sur l'annulation de la hausse dans un éventuel changement de pouvoir.	Thématique	GOUV	2
très	Intensif.	Dialectique	GOUV	2
pression	La pression exercée par le mouvement, le climat de pression en général causé par la grève.	Thématique	GOUV	2
familial	Sur les modifications du régime de prêt offert par le gouvernement : aucune contribution parentale jusqu'à un revenu de 60 000 par année.	Thématique	GOUV	2
partisan	Les débats sur la grève se font dans un contexte partisan.	Thématique	GOUV	2
commission	Les étudiants revendiquent une commission d'enquête pour faire une investigation sur la corruption dans le secteur de la construction. Aussi, enquête sur la question du financement des universités.	Thématique	ETUD	2

Segments répétés

Critères	Contexte	Composante	Classe	Spéc.
frais de scolarité	Thématique sur l'augmentation des frais.	Thématique	GOUV	10
que nous	Énonciateur proche, prise en charge, collectif.	Dialogique	ETUD	10
Mais il	En majuscule, le « Mais » apparaît dans un contexte où un argument est présenté puis rejeté dans la phrase subséquente.	Dialectique	GOUV	6
associations étudiantes	Associations étudiantes impliquées dans le conflit, les actions et les décisions prises par les associations.	Thématique	GOUV	5
carrés rouges	Référence au symbole des étudiantes (carrés rouges) ou aux étudiants eux-mêmes.	Thématique	GOUV	5
centrales syndicales	Référence aux centrales ou organisations syndicales et à leur appui financier au mouvement de grève.	Thématique	GOUV	5
Mme Marois	Entité nommée. « Mme » est utilisée de façon ironique.	Thématique	GOUV	5
étudiants qui	Énonciateur distancé, se dirige à la 3e personne.	Dialogique	GOUV	5
de demain	Organisation temporelle : référence au temps futur, à l'avenir.	Dialectique	ETUD	5
ils n	Énonciateur distancé, se dirige à la 3e personne.	Dialogique	GOUV	5
juste part	Contestation sur la formule de la « juste part » dans le discours du gouvernement, plus souvent cité entre guillemets. Aussi, l'utilisation de l'adjectif « juste ». La formule « juste part » n'est pas très présente dans GOUV.	Thématique	ETUD	5
Martine Desjardins	Référence à Martine Desjardins, leader étudiant de la FÉUQ, Fédération étudiante universitaire du Québec. Référence à ses déclarations et actions.	Thématique	ETUD	5
Pauline Marois	Entité nommée, Pauline Marois, chef du Parti Québécois.	Thématique	GOUV	5
Réjean Parent	Réjean Parent, président de la Centrale des syndicats du Québec (CSQ), est souvent cité par son support aux étudiants en grève.	Thématique	ETUD	5
Mais elle	En majuscule, le « Mais » apparaît dans un contexte où un argument est présenté puis rejeté dans la phrase subséquente.	Dialectique	GOUV	4

à nous	Énonciateur proche, prise en charge, collectif.	Dialogique	ETUD	4
carré rouge	Référence au symbole des étudiantes (carrés rouges) et sur le port du symbole.	Thématique	GOUV	4
de la rue	Façon de se référer au mouvement étudiant, à la rue, en opposition aux institutions démocratiques parfois.	Thématique	GOUV	4
de notre	Énonciateur proche, prise en charge, collectif.	Dialogique	ETUD	4
de nouveaux	Structuration cumulative du récit, notion de saturation.	Dialectique	ETUD	4
et ceux	Énonciateur distancé, se dirige à la 3e personne.	Dialogique	GOUV	4
et plus	Structuration cumulative du récit, notion de saturation.	Dialectique	ETUD	4
gel des frais	Thématique au sujet de gel de frais, critique aux demandes des étudiants.	Thématique	GOUV	4
hausse des frais	À propos de l'augmentation des frais de scolarité.	Thématique	GOUV	4
Il y a	Modalisation exprimant de la certitude.	Dialogique	GOUV	4
leurs cours	Référence à 3e personne (les étudiants), acteur distancé.	Dialogique	GOUV	4
leurs membres	Référence à 3e personne (les étudiants), acteur distancé.	Dialogique	GOUV	4
même si	Coordination exprimant l'hypothèse, fonction argumentative.	Dialectique	GOUV	4
nos jeunes	Énonciateur proche, prise en charge, collectif.	Dialogique	ETUD	4
notre jeunesse	Énonciateur proche, prise en charge, collectif.	Dialogique	ETUD	4
nous les	Énonciateur proche, prise en charge, collectif.	Dialogique	ETUD	4
Parce que	Fonction argumentative, explications. Succède les questions posées par l'auteur même.	Dialectique	GOUV	4
pendant la	Structuration durative du récit.	Dialectique	GOUV	4
présidente de la FEUQ	Référence aux déclarations et actions de la présidente de la FEUQ, Martine Desjardins.	Thématique	ETUD	4
prêts et bourses	Référence au programme de prêts et bourses du gouvernement.	Thématique	GOUV	4
boycott des cours	Suremploi de « boycott » par opposition à « grève » dans ETUD.	Thématique	GOUV	4
qui ne sont	Modalisation exprimant de la certitude.	Dialogique	GOUV	4
du côté	Du côté des étudiants, du gouvernement — opposition de perspectives, introduction de l'opinion des parties du conflit.	Thématique	GOUV	4
Michelle Courchesne	Entité nommée. Ministre de l'Éducation qui a remplacé la précédente et qui poursuit les négociations avec les associations étudiantes.	Thématique	ETUD	4
Mme Beauchamp	Entité nommée. Ministre de l'Éducation du Québec qui a participé au début des négociations avec les étudiants.	Thématique	ETUD	4
Mme Courchesne	Entité nommée. Ministre de l'Éducation qui a remplacé la ministre précédente et qui poursuit les négociations avec les associations étudiantes.	Thématique	ETUD	4
salaire minimum	Proposition pour indexer les frais de scolarité au salaire minimum.	Thématique	GOUV	4
votre gouvernement	Positionnement énonciatif distancé par rapport au gouvernement. Dénote la non-reconnaissance de sa légitimité (votre versus notre). Présence de l'interlocuteur dans le texte.	Dialogique	ETUD	4
système économique	Système économique capitaliste	Thématique	ETUD	4
Madame Beauchamp	Médiateur impliqué dans le conflit, ironie avec l'utilisation de Madame	Thématique	ETUD	4
On le	Énonciateur distancé, impersonnel	Dialogique	GOUV	4
% en	Référence à l'augmentation des frais de scolarité.	Thématique	ETUD	3

accessibilité aux études	Préoccupation des auteurs avec le thème de l'accessibilité. Parle des concessions proposées par le gouvernement pour favoriser ladite accessibilité.	Thématique	GOUV	3
alors qu'	Contrapositions d'arguments, dénote de la contradiction.	Dialectique	ETUD	3
Assemblée nationale	Entité nommée. Assemblée nationale du Québec, décisions administratives et politiques.	Thématique	GOUV	3
au cours des	Structuration durative du récit.	Dialectique	GOUV	3
au terme	Structuration terminative du récit.	Dialectique	GOUV	3
aux jeunes	Indice d'interlocution. Représentation de l'énonciation, aux jeunes étudiants.	Dialogique	ETUD	3
avaient pas	Conjugaison à la troisième personne, sujet parlant distancé de son discours. Modalisation de la certitude.	Dialogique	GOUV	3
campagne électorale	Au sujet de la campagne électorale des libéraux.	Thématique	GOUV	3
car il	Explications fourni dans une même phrase	Dialectique	ETUD	3
cela ne	Opposition à une idée, négation à une proposition	Dialectique	GOUV	3
ces derniers	Référence à la 3e personne du pluriel (ils).	Dialogique	GOUV	3
cette année	Structuration temporelle durative.	Dialectique	GOUV	3
chefs syndicaux	Chefs de centrales syndicales qui appuient la grève.	Thématique	ETUD	3
classe moyenne	Critique au discours du gouvernement : les auteurs d'ETUD contestent l'argument selon lequel les coûts de l'éducation vont peser plus sur la classe moyenne après la hausse.	Thématique	ETUD	3
communauté universitaire	Ensemble d'étudiants universitaires, associations et professeurs	Thématique	ETUD	3
comme ça	Démonstratif reliant des propositions (c'est comme ça qui ça marche, des gens comme ça).	Dialectique	GOUV	3
crise étudiante	Manière de se référer à la grève.	Thématique	GOUV	3
dans ce dossier	Signifie au sujet de quelque chose. Plutôt informel que dans ETUD.	Dialectique	GOUV	3
dans le cas	Pour introduire une supposition, une hypothèse	Dialogique	GOUV	3
dans leur	Énonciateur distancé, se dirige à la 3e personne.	Dialogique	GOUV	3
de plus en plus	Structuration cumulative du récit.	Dialectique	ETUD	3
de votre	Réponse à des textes particuliers et à des personnes, par exemple : « De votre gouvernement. »	Dialogique	ETUD	3
des gens	Énonciateur distancé, se dirige à la 3e personne.	Dialogique	GOUV	3
Depuis des	Marqueur temporel, inchoativité.	Dialectique	ETUD	3
derniers jours	Structuration temporelle durative.	Dialectique	GOUV	3
des études supérieures	Manière de se référer aux études universitaires. Dans ETUD les auteurs utilisent plutôt « éducation » et « savoir ».	Thématique	GOUV	3
dès le	Marqueur temporel, inchoativité.	Dialectique	ETUD	3
dire que les	Contrapositions. Façon de présenter un argument contraire pour ensuite le réfuter.	Dialectique	ETUD	3
droit de grève	Se refaire au droit de grève qui était contesté par ceux qui se positionnaient contre les décisions adoptées par les assemblées étudiantes.	Thématique	ETUD	3

droits de scolarité	Spécificité de « droits de scolarité » dans le sous-corpus ETUD par opposition à « frais de scolarité » du côté GOUV.	Thématique	ETUD	3
en cinq ans	Temps de l'augmentation de frais de scolarité proposés par le gouvernement Charest.	Thématique	ETUD	3
En effet	Fonction argumentative, explications.	Dialectique	ETUD	3
est encore	Structuration cumulative du récit.	Dialectique	ETUD	3
est là	Utilisé pour montrer la raison, donner l'explication.	Dialectique	ETUD	3
est pas parce	Fonction argumentative, s'oppose à un argument prononcé, en le réfutant.	Dialectique	GOUV	3
et surtout	Structuration cumulative du récit.	Dialectique	ETUD	3
et on	Énonciateur distancé, impersonnel	Dialogique	GOUV	3
études universitaires	Manière de se référer aux études universitaires. Dans ETUD les auteurs utilisent plutôt les mots « éducation » et « savoir ».	Thématique	GOUV	3
gel des droits	Thématique sur le gel des droits de scolarité.	Thématique	GOUV	3
gouvernement Charest	Désignation. Différemment d'ÉTUD, où on retrouve « votre gouvernement », ici cette désignation a plus de déférence.	Thématique	GOUV	3
gouvernement du Québec	Entité nommée. Manière de se référer au gouvernement.	Thématique	ETUD	3
gouvernement québécois	Entité nommée. Manière de se référer au gouvernement.	Thématique	ETUD	3
il devrait	Modalisation exprimant l'incertitude.	Dialogique	GOUV	3
il est	Modalisation exprimant de la certitude.	Dialogique	GOUV	3
Il ne faut	Impératif, exprime un ordre.	Dialogique	GOUV	3
ils ont	Énonciateur distancé, 3e personne (beaucoup d'occurrences pour se référer aux étudiants).	Dialogique	GOUV	3
ils sont	Énonciateur distancé, 3e personne (beaucoup d'occurrences pour se référer aux étudiants).	Dialogique	GOUV	3
je pense	Modalité épistémique, exprime l'incertitude, l'opinion.	Dialogique	GOUV	3
je vois	Énonciateur proche, prise en charge, modalisation de la certitude, factuel, réel.	Dialogique	ETUD	3
joindre notre	Énonciateur proche, propose un contact avec son interlocuteur	Dialogique	ETUD	3
le 22	Date d'organisation des manifestations le 22 de chaque mois.	Thématique	ETUD	3
les étudiants n'	Mise à distance énonciatif. Modalisation de la certitude.	Dialogique	ETUD	3
Les étudiants ont	Mise à distance énonciatif. Modalisation de la certitude.	Dialogique	ETUD	3
ligne dure	Terme pour qualifier l'intransigeance du gouvernement du Québec, le manque de négociations raisonnables.	Thématique	ETUD	3
lutte contre	Référence à la lutte des étudiants (contre le gouvernement.).	Thématique	GOUV	3
médias sociaux	Référence aux médias utilisés par les étudiants comme moyen de lutte.	Thématique	GOUV	3
modèle québécois	Référence au modèle de social-démocratie, état de bien social, gestion publique.	Thématique	GOUV	3
Mais la	En majuscule, le « Mais » apparaît dans un contexte où un argument est présenté puis rejeté dans la phrase subséquente.	Dialectique	GOUV	3
Mais les	En majuscule, le « Mais » apparaît dans un contexte où un argument est présenté puis rejeté dans la phrase subséquente.	Dialectique	GOUV	3
ne paient	Terminaison 3e personne du pluriel (ils).	Dialogique	GOUV	3

niveau de vie	Référence sur la baisse du niveau de vie dans les dernières années ou du niveau de vie des futurs étudiants par rapport à leurs parents.	Thématique	ETUD	3
non pas	Explication qui suit la réfutation d'un argument.	Dialectique	ETUD	3
non seulement	Structuration cumulative du récit.	Dialectique	ETUD	3
nous avons	Énonciateur proche, prise en charge, identité collective.	Dialogique	ETUD	3
nous nous	Énonciateur proche, prise en charge, identité collective.	Dialogique	ETUD	3
nous pouvons	Énonciateur proche, prise en charge, identité collective.	Dialogique	ETUD	3
nous sommes	Énonciateur proche, prise en charge, identité collective.	Dialogique	ETUD	3
nous voulons	Énonciateur proche, prise en charge, identité collective.	Dialogique	ETUD	3
On a	Énonciateur distancé, impersonnel	Dialogique	GOUV	3
on a	Énonciateur distancé, impersonnel	Dialogique	GOUV	3
on est	Énonciateur distancé, impersonnel	Dialogique	GOUV	3
on fait	Énonciateur distancé, s'adresse à la 3e personne.	Dialogique	GOUV	3
On n	Énonciateur distancé, impersonnel	Dialogique	GOUV	3
ont pas	Terminaison verbale à la 3e personne du pluriel. Effacement énonciatif.	Dialogique	GOUV	3
par ailleurs	Fonction argumentative.	Dialectique	GOUV	3
par contre	Fonction argumentative, opposition.	Dialectique	GOUV	3
par la suite	Fonction argumentative, conséquence, aspect terminatif.	Dialectique	GOUV	3
pas besoin	Phrases négatives, réfutation d'une proposition, d'une idée, d'un argument.	Dialectique	GOUV	3
plus de	Structuration cumulative du récit.	Dialectique	ETUD	3
plus grand	Structuration cumulative du récit.	Dialectique	ETUD	3
point de presse	Se refaire à l'organisation ou la tenue de points de presse pour expliquer les demandes des étudiants.	Thématique	ETUD	3
pour tous	Thème sur l'accessibilité, l'éducation est un droit de tous.	Thématique	ETUD	3
premier ministre	Manière de se référer à Jean Charest.	Thématique	ETUD	3
première ministre	Se référant à Pauline Marois comme future première ministre	Thématique	GOUV	3
près de	Précède la présentation d'un chiffre particulier, structuration cumulative du récit.	Dialectique	ETUD	3
semble que	Avoir l'impression de quelque chose, ressentir.	Dialogique	GOUV	3
autres provinces	Utilisé dans des contextes qui comparent la situation économique du Québec à d'autres provinces.	Thématique	GOUV	3
proportionnel au revenu	Au sujet de la proposition d'une augmentation proportionnelle au revenu de la part du gouvernement.	Thématique	GOUV	3
régime des prêts	Régime de prêts et bourses du gouvernement du Québec	Thématique	GOUV	3
Quand on	Structuration temporelle durative.	Dialectique	GOUV	3
question de	Manière d'expliquer un enjeu particulier : « c'est une question de ».	Dialectique	ETUD	3
majorité de Québécois	La plupart de contextes montrent qu'il s'agit de la majorité d'individus favorables aux arguments du gouvernement ou contre les étudiants. Moins présent dans ETUD.	Thématique	GOUV	3
si les	Contidditionnel, Possibilité.	Dialogique	GOUV	3
ministre Line Beauchamp	Entité nommée. Ministre de l'Éducation du Québec qui a participé au début des négociations avec les étudiants. Se diffère de « Madame Beauchamp », plus communément employé dans ETUD.	Thématique	GOUV	3

si on	Relation logique entre deux assertions.	Dialectique	GOUV	3
son gouvernement	Se réfère au gouvernement actuel, au gouvernement Charest. Contraste avec « votre gouvernement » du côté des étudiants. Indice de l'interlocution.	Dialogique	GOUV	3
suis sûr	Modalité épistémique, exprime la certitude.	Dialogique	GOUV	3
sur le dos	Thématique sur les coûts de l'éducation et sur le poids de ces coûts sur les étudiants.	Thématique	ETUD	3
Sur le	Manière d'aborder une question, plutôt informel comparativement à ETUD	Dialectique	GOUV	3
Parti québécois	Entité nommée. Critiques au Parti Québécois.	Thématique	GOUV	3
plus bas	Plus présent dans GOUV. Les frais de scolarité sont qualifiés comme bas, surtout par rapport à d'autres provinces canadiennes. Dans ETUD, bas se refaire aussi à des frais, mais dans ce contexte les auteurs reprennent l'argument du groupe contraire pour le rebattre.	Thématique	GOUV	3
tout comme	Structuration cumulative du récit.	Dialectique	ETUD	3
à cette	Démonstratif, utilisé pour parler des évenemments courants	Dialectique	ETUD	2
à tous	Dans la majorité des contextes se refère aux étudiants ou aux citoyens, et aborde le sujet de l'accessibilité aux études.	Thématique	ETUD	2
Au lieu	Opposition entre deux assertions	Dialectique	ETUD	2
arriver à	Terminatif. Arriver à une solution, entente, à ses fins.	Dialectique	ETUD	2
au détriment	Au détriment d'un groupe ou d'une idée (par exemple : majorité, étudiants, pauvres).	Dialectique	ETUD	2
au profit	Au profit d'un groupe (par exemple : riches, minorité, loi 78, quelques-uns).	Dialectique	ETUD	2
au sein	Au sein d'un groupe.	Dialectique	ETUD	2
autre côté	L'autre côté de la dispute : met en évidence de la polarisation existante entre les deux groupes.	Thématique	GOUV	2
avec eux	Référence aux étudiants, dénote de la proximité (par exemple : « je marche avec eux »).	Dialogique	ETUD	2
ce temps	Référence à un temps passé situé dans l'histoire	Dialectique	ETUD	2
ces étudiants	Indice d'interlocution. Référence aux étudiants, proximale, ces étudiants en grève.	Dialogique	ETUD	2
ceux et celles qui	Indice d'interlocution. Façon de l'énonciateur de se référer aux partisans ou opposants de la grève. Inclusion du féminin et masculin (différent dans GOUV).	Dialogique	ETUD	2
de leurs	Énonciateur distancé, se dirige à la 3e personne.	Dialogique	GOUV	2
deux parties	Les deux parties qui se confrontent, le gouvernement et les étudiants, thématique sur leur opposition.	Thématique	GOUV	2
droit de manifester	Thématique sur la nécessité d'encadrer juridiquement les manifestations, afin d'éviter des abus ou entraîner le désordre.	Thématique	GOUV	2
en matière	En s'agissant de, à propos de. Utilisés vis-à-vis des sujets abordés par les étudiants comme éducation, financement de l'enseignement, contribution étudiante, sécurité	Dialectique	ETUD	2
enseignement supérieur	Exprime la préoccupation avec l'accès à l'enseignement supérieur, ainsi que sa qualité.	Thématique	ETUD	2

est que	Modalisation exprimant de la certitude.	Dialogique	GOUV	2
et donc	Explications donnés dans une même phrase.	Dialectique	GOUV	2
étudiants universitaires	Référence aux étudiants qui seront affectés par la hausse.	Thématique	ETUD	2
les étudiants eux	Référence aux étudiants, mise à distance énonciatif.	Dialogique	GOUV	2
les gens	Énonciateur distancé, se dirige à la 3e personne, informel	Dialogique	GOUV	2
loi 78	Thématique autour de la loi 78 qui décrète l'augmentation des frais.	Thématique	ETUD	2
mais aussi	Souvent suivi de « non seulement… », exprime l'idée de somme, de saturation.	Dialectique	ETUD	2
majorité des étudiants	À propos des étudiants qui ne se sont pas engagés aux mouvements de grève. L'argument veut déqualifier le mouvement, en le caractérisant comme une réaction d'un groupe minoritaire.	Thématique	GOUV	2
milliards de dollars	Au sujet des dépenses gouvernementales et de gaspillage d'argent, mais aussi des dettes.	Thématique	ETUD	2
ne sont	(Ils ne sont) Négation, certitude.	Dialogique	GOUV	2
par les étudiants	Mise à distance énonciatif.	Dialogique	GOUV	2
partie de la	Faire partie d'un groupe, d'une société, d'une population.	Thématique	ETUD	2
se sont	Référence aux étudiants, mise à distance énonciatif.	Dialogique	GOUV	2
si elle	Relation logique entre deux assertions.	Dialectique	GOUV	2
si je	Relation logique entre deux assertions.	Dialectique	GOUV	2
si vous	Relation logique entre deux assertions.	Dialectique	GOUV	2
temps plein	Études à temps plein, thématique reliée à la charge de travail et impossibilité à concilier le travail et les études.	Thématique	ETUD	2
y a quelques	Trait temporel, passé proche.	Dialectique	GOUV	2

Coocurrents

Critères	Contexte	Composante	Classe	Spéc.
scolarité-dégel	Dégel dans le gouvernement de Bourassa.	Thématique	ETUD	10
hausse-proposer	Hausse proposée.	Thématique	GOUV	10
droit-augmenter	Augumenter les droits.	Thématique	ETUD	5
hausse-maintenir	Maintenir la hausse.	Thématique	GOUV	5
étudiant-collégial	Étudiants du niveau collégial, fédération d'étudiants.	Thématique	GOUV	5
hausse-raisonnable	La hausse est raisonnable.	Thématique	GOUV	5
droit-renoncer	Renoncer à la hausse des droits.	Thématique	ETUD	5
gouvernement-fédéral	Gouvernement fédéral.	Thématique	ETUD	5
droit-hausser	Hausser les droits de scolarité.	Thématique	GOUV	5
étudiant-détermination	Détermination des étudiants de poursuivre le mouvement.	Thématique	ETUD	5
québec-milliard	Relatif au budget de la province, investissement, dividende, dette, etc.	Thématique	ETUD	5
hausse-moratoire	Moratoire sur la hausse.	Thématique	ETUD	4
scolarité--annoncer	La hausse a été annoncée par le gouvernement.	Thématique	GOUV	4
hausse-approuver	Le gouvernement approuve la hausse. Difficulté à approuver la hausse.	Thématique	ETUD	4
étudiant-porte-parole	Leader des étudiants, porte-parole.	Thématique	GOUV	4

hausse-favorable	Personnes qui sont favorables à la hausse.	Thématique	ETUD	4
scolarité-imposer	La hausse a été imposée par le gouvernement Charest.	Thématique	ETUD	4
étudiant-opposer	Les étudiants s'opposent à la hausse.	Thématique	ETUD	4
hausse-décret	Annuler la hausse par décret.	Thématique	GOUV	4
gouvernement-accepter	Acceptation des étudiants à l'égard des propositions du gouvernement.	Thématique	ETUD	4
scolarité-afférent	Frais afférents en surplus des frais de scolarité.	Thématique	ETUD	4
droit-privilège	L'éducation est un droit et pas un privilège.	Thématique	ETUD	4
québec-diriger	Diriger le Québec, dans le sens de gouverner.	Thématique	ETUD	4
droit-pacte	Pacte international relatif aux droits économiques.	Thématique	ETUD	4
droit-international	Droits humains, pacte international des droits économiques.	Thématique	ETUD	4
hausse-affecter	L'accessibilité aux études est affectée par la hausse.	Thématique	ETUD	4
étudiant-entente	Entente entre les étudiants et le gouvernement.	Thématique	ETUD	4
droit-brimer	Brimer les droits des étudiants qui ne participent pas à la grève.	Thématique	GOUV	4
hausse-objet	La hausse des frais est l'objet du conflit.	Thématique	GOUV	4
scolarité-réclamer	Les étudiants réclament le gel, l'abolition des droits.	Thématique	GOUV	4
gouvernement-décréter	La hausse décrétée par le gouvernement.	Thématique	GOUV	4
gouvernement-argument	Arguments en faveur de la hausse.	Thématique	ETUD	4
gouvernement-succéder	Succession des gouvernements du Québec. Thème sur l'élection.	Thématique	GOUV	4
droit-expression	Droit de manifester, de s'exprimer, liberté d'expression.	Thématique	ETUD	4
droit-contractuel	Droits contractuels.	Thématique	ETUD	4
québec-révélateur	La crise est révélatrice des problèmes sociaux au Québec.	Thématique	GOUV	4
québec-conservateur	Conservateurs au gouvernement, mesures conservatrices.	Thématique	ETUD	4
gouvernement-décider	Décisions prises par le gouvernement.	Thématique	ETUD	4
droit-garantir	Droits (juridiques) garantis.	Thématique	ETUD	4
étudiant-cause	Cause étudiante.	Thématique	GOUV	3
scolarité-concerner	Le débat sur la hausse concerne toute la société.	Thématique	ETUD	3
québec-progressiste	Nécessité d'un gouvernement plus progressiste.	Thématique	ETUD	3
scolarité-parler	Parler des droits, de la hausse, de ce qu'il faut parler.	Thématique	ETUD	3
droit-abolition	Question de la gratuité. Abolition des droits.	Thématique	GOUV	3
québec-comparer	Le Québec comparé avec d'autres provinces sur l'aspect de l'éducation et du budget.	Thématique	GOUV	3
étudiant-déclencher	Grève, mouvement ou crise déclenchée par les étudiants.	Thématique	GOUV	3
gouvernement-impopularité	Impopularité du gouvernement	Thématique	GOUV	3
gouvernement-actuel	Gouvernement actuel.	Thématique	ETUD	3
gouvernement-scandale	Scandales relatifs au gouvernement.	Thématique	ETUD	3
hausse-brutal	La hausse est brutale.	Thématique	ETUD	3
scolarité-graduel	Hausse graduelle et modérée.	Thématique	GOUV	3
étudiant-injonction	Injonctions demandées par certains étudiants pour reprendre leurs cours.	Thématique	GOUV	3
québec-député	Députés de l'opposition.	Thématique	ETUD	3
gouvernement-ouverture	Ouverture du gouvernement.	Thématique	GOUV	3

gouvernement-sérieux	Le gouvernement ne prend pas au sérieux le mouvement ni les allégations contre lui.	Thématique	ETUD	3
québec-francophone	Québec francophone.	Thématique	ETUD	3
gouvernement-gérer	Le gouvernement a mal géré la crise.	Thématique	GOUV	3
gouvernement-reprocher	Reprocher le gouvernement.	Thématique	GOUV	3
québec-généreux	Les programmes sociaux du Québec sont généreux.	Thématique	GOUV	3
étudiant-échouer	Échec de négociations avec les étudiants.	Thématique	ETUD	3
étudiant-recruter	Recruter des étudiants dans les universités et dans les associations.	Thématique	ETUD	3
gouvernement-signifier	Explications sur les actions du gouvernement — «cela signifie».	Thématique	GOUV	3
gouvernement-preuve	Locution : faire preuve.	Thématique	ETUD	3
hausse-débat	Débat sur la hausse.	Thématique	ETUD	3
étudiant-empêcher	Empêcher d'autres étudiants à aller contre la grève ou suivre leurs cours.	Thématique	GOUV	3
étudiant-millier	Nombre d'étudiants organisés, dans les associations, dans les manifestations.	Thématique	GOUV	3
hausse-accompagner	La hausse devrait être accompagnée d'une bonification, d'un avantage.	Thématique	GOUV	3
scolarité-lutter	Les étudiants luttent contre la hausse ou contre d'autres choses?	Thématique	GOUV	3
étudiant-cégep	Étudiants des Cégeps.	Thématique	GOUV	3
droit-global	Les droits de scolarité ne représentent qu'une partie du coût global de l'éducation.	Thématique	ETUD	3
étudiant-exiger	Exigences des étudiants, fardeau exigé des étudiants.	Thématique	GOUV	3
gouvernement-souhaiter	Les choses que la population souhaite du gouvernement.	Thématique	GOUV	3
hausse-prêcher	Personnes qui prêchent en faveur de la hausse.	Thématique	ETUD	3
hausse-scander	Les étudiants scandent contre la hausse.	Thématique	ETUD	3
étudiant-contribution	Contribution des étudiants pour le financement de leurs études.	Thématique	GOUV	3
étudiant-profiter	Les étudiants devraient profiter de l'ouverture du gouvernement.	Thématique	GOUV	3
étudiant-début	Début du conflit.	Thématique	ETUD	3
gouvernement-dénouement	Possibilité de dénouement de la crise.	Thématique	GOUV	3
scolarité-boycotteur	Étudiants en grève.	Thématique	GOUV	3
droit-indexer	Indexation des droits au coût de la vie.	Thématique	ETUD	3
droit-primauté	Primauté du droit (juridique).	Thématique	ETUD	3
gouvernement-dupe	La population n'est pas dupe par rapport au discours du gouvernement.	Thématique	ETUD	3
québec-problème	Problèmes qui touchent le Québec.	Thématique	GOUV	3
étudiant-doctorat	Étudiants du doctorat.	Thématique	ETUD	3
étudiant-postsecondaire	Étudiants du postsecondaire.	Thématique	ETUD	3
étudiant-blâmer	Les étudiants sont blâmés.	Thématique	ETUD	3
étudiant-minimiser	Miniser le débat, minimiser les gains des étudiants.	Thématique	ETUD	3
scolarité-garderie	Manque d'argent pour financer l'éducation et les garderies.	Thématique	GOUV	3
droit-obligation	Relation entre les droits juridiques et les obligations.	Thématique	ETUD	3
québec-entier	Québec entier, idée de totalité.	Thématique	ETUD	3

hausse-tenant	Tenants de la hausse.	Thématique	ETUD	3
étudiant-établissement	Établissements d'enseignement.	Thématique	GOUV	3
gouvernement-attitude	Attitude du gouvernement.	Thématique	ETUD	3
étudiant-manif	Manière de se référer aux manifestations.	Thématique	GOUV	3
étudiant-rejeter	Les étudiants rejettent les offres proposées par le gouvernement.	Thématique	ETUD	3
droit-vote	Droit de vote.	Thématique	ETUD	3
québec-insulte	Insulte du gouvernement du Québec envers les étudiants.	Thématique	ETUD	3
droit-régressif	Le gel des droits est une mesure régressive.	Thématique	GOUV	3
scolarité-justifier	La hausse est justifiée. Arguments qui justifient la hausse.	Thématique	GOUV	3
étudiant-bonifier	Bonification du régime de prêts et bourses offertes par le gouvernement.	Thématique	ETUD	3
étudiant-réveiller	Réveil de la société, des étudiants.	Thématique	ETUD	3
droit-majorer	Majorer les droits de scolarité.	Thématique	GOUV	3
droit-veto	Droit de veto dans les assemblées.	Thématique	GOUV	3
hausse-baguette	Annuler la hausse par un coup de baguette magique.	Thématique	GOUV	3
hausse-objecter	Les étudiants objectent à toute hausse.	Thématique	GOUV	3
étudiant-dialogue	Dialogue entre étudiants et gouvernement.	Thématique	ETUD	3
scolarité-entêtement	Entêtement du gouvernement face à la hausse.	Thématique	ETUD	3
gouvernement-céder	Concessions du gouvernement.	Thématique	GOUV	3
scolarité-embûche	Obstacles présentés par la hausse.	Thématique	ETUD	3
scolarité-loi-cadre	Loi-cadre pour les frais de scolarité.	Thématique	ETUD	3
scolarité-tarif	Augmentation des tarifs.	Thématique	ETUD	3
étudiant-boursier	Étudiants qui se bénéficient d'une bourse du programme du gouvernement.	Thématique	GOUV	3
étudiant-éterniser	La crise étudiante s'éternise.	Thématique	GOUV	3
droit-bafouer	Bafouer les droits et les libertés, la démocratie.	Thématique	ETUD	3
étudiant-assujettir	Lois auxquelles les étudiants sont assujettis.	Thématique	ETUD	3
étudiant-échanger	Échange d'idées, débat.	Thématique	ETUD	3
droit-faveur	Faveur de la hausse.	Thématique	ETUD	3
droit-parole	Droit de parole.	Thématique	GOUV	3
scolarité-impôt	Discussion sur l'augmentation de crédits d'impôt pour ceux qui paient les droits de scolarité.	Thématique	ETUD	3
hausse-annonce	Annonce de la hausse.	Thématique	ETUD	3
hausse-privatisation	Privatisation des services publics.	Thématique	ETUD	3
hausse-contester	Les étudiants contestent la hausse.	Thématique	GOUV	3
hausse-reculer	Reculer sur la hausse. Le gouvernement ne devrait pas reculer.	Thématique	GOUV	3
étudiant-retourner	Retourner aux études, retourner en classe.	Thématique	GOUV	3
gouvernement-tort	Tort du gouvernement dans la crise.	Thématique	GOUV	2
gouvernement-réduire	Réduction de dépenses, réduire la pression, la hausse.	Thématique	GOUV	2
droit-inaliénable	Droit inaliénable.	Thématique	ETUD	2
québec-gouverne	Gouverne du Québec.	Thématique	GOUV	2
gouvernement-concession	Le gouvernement fait plusieurs concessions dans le conflit.	Thématique	GOUV	2
gouvernement-renverser	Renversement du gouvernement. Élections.	Thématique	GOUV	2
scolarité-accroître	La hausse permet d'accroitre l'efficacité, l'accessibilité et le	Thématique	GOUV	2

	financement des universités.			
québec-habiter	Dépense par habitant du Québec.	Thématique	GOUV	2
gouvernement-intransigeance	Intransigeance du gouvernement.	Thématique	ETUD	2
gouvernement-irresponsable	L'attitude du gouvernement est irresponsable.	Thématique	ETUD	2
gouvernement-suspendre	Suspension des trimestres, suspension de la loi.	Thématique	ETUD	2
étudiant-gréviste	Étudiants grévistes.	Thématique	GOUV	2
droit-défendre	Défendre les droits (juridiques).	Thématique	ETUD	2
droit-dépasser	Le conflit dépasse l'enjeu initial concernant les droits de scolarité.	Thématique	GOUV	2
québec-région	Région de Montréal, Québec, lieu géographique.	Thématique	GOUV	2
québec-aréna	Construction d'un aréna de hockey.	Thématique	ETUD	2
québec-dynamique	Manière de procéder : dynamique mal saine, dynamique politique.	Thématique	ETUD	2
québec-fouet	La hausse frappera «de plein fouet» les jeunes.	Thématique	ETUD	2
québec-surnombre	Les étudiants en surnombre se manifestent au Québec.	Thématique	ETUD	2
hausse-abolir	Pauline Marois propose l'abolition de la hausse.	Thématique	GOUV	2
hausse-équivaloir	Combien de journées de travail équivalent à l'argent de la hausse.	Thématique	GOUV	2
scolarité-dossier	Dossier de discussion, sujet.	Thématique	ETUD	2
scolarité-matériel	Matériel scolaire.	Thématique	ETUD	2
gouvernement-corrompre	Gouvernement corrompu.	Thématique	ETUD	2
gouvernement-incomber	Obligations du gouvernement.	Thématique	ETUD	2
gouvernement-argent	Argent du gouvernement, budget, utilisation de l'argent par le gouvernement.	Thématique	ETUD	2
hausse-mentalité	Mentalité de privatisation associée à la hausse.	Thématique	ETUD	2
hausse-modéré	Arguments contre la hausse modérée.	Thématique	ETUD	2
étudiant-saisir	Les étudiants doivent saisir l'offre du gouvernement.	Thématique	GOUV	2
gouvernement-consister	Explications sur les actions du gouvernement — «consiste à».	Thématique	GOUV	Spéc.
gouvernement-choisir	Choix du gouvernement.	Thématique	ETUD	10
étudiant-protester	Protester contre la hausse.	Thématique	GOUV	10
étudiant-allié	Alliés des étudiants.	Thématique	ETUD	5
étudiant-dénigrer	Dénigrer les étudiants.	Thématique	ETUD	5
gouvernement-subventionner	Subventions données par le gouvernement.	Thématique	GOUV	5
gouvernement-augmentation	Augmentation des frais, position du gouvernement.	Thématique	GOUV	5
québec-fiscal	Paradis fiscal, irresponsabilité fiscale.	Thématique	ETUD	5
étudiant-négocier	Négocier avec les étudiants.	Thématique	ETUD	5
québec-canadien	Québec dans le contexte canadien.	Thématique	GOUV	5
gouvernement-déplait	Les décisions du gouvernement déplait les étudiants.	Thématique	GOUV	5
gouvernement-artisan	Le gouvernement est l'artisan de leur propre malheur.	Thématique	ETUD	5
gouvernement-convier	Le gouvernement devrait convier les étudiants à la négociation.	Thématique	ETUD	4
gouvernement-mentir	Le gouvernement ment.	Thématique	ETUD	4
gouvernement-saboter	Le gouvernement veut saboter les discussions.	Thématique	ETUD	4
québec-décennie	Augmentation des frais au Québec dans la dernière décennie.	Thématique	GOUV	4
hausse-concret	Le gouvernement doit répondre aux problèmes concrets soulevés par les étudiants, faire de propositions concrètes sur la hausse.	Thématique	GOUV	4
étudiant-terrain	Terrain de négociation.	Thématique	ETUD	4

scolarité-payer	Ceux qui paient les frais de scolarité.	Thématique	GOUV	4
étudiant-collège	Étudiants des collèges.	Thématique	ETUD	4
gouvernement-décision	Décision du gouvernement.	Thématique	GOUV	4
hausse-prévoir	La hausse prévue.	Thématique	GOUV	4
scolarité-gratuité	Proposition des étudiants pour la gratuité des études.	Thématique	GOUV	4
droit-admissible	Avoir le droit d'être admissible au régime de prêts et bourses.	Thématique	GOUV	4
droit-démocrate	Démocratie et droits.	Thématique	GOUV	4
droit-efficacité	Droits de scolarité et efficacité des universités.	Thématique	GOUV	4
droit-explorer	Explorer des solutions pour la hausse des droits de scolarité.	Thématique	GOUV	4
droit-injecter	La hausse des droits va injecter de l'argent dans les universités.	Thématique	GOUV	4
hausse-étalement	Étalement de la hausse : discussion sur la proposition du gouvernement.	Thématique	GOUV	4
hausse-étirer	Proposition d'étirer la hausse dans le temps.	Thématique	GOUV	4
hausse-incontournable	La hausse est incontournable.	Thématique	GOUV	4
hausse-majoration	Défense de la majoration de la bourse pour compenser la hausse.	Thématique	GOUV	4
gouvernement-offre	offres du gouvernement aux associations étudiantes	Thématique	GOUV	4
scolarité-choix	La question de la hausse des frais pose un choix à la société et aux électeurs.	Thématique	GOUV	4
scolarité-applicable	Crédits d'impôt applicables à la hausse.	Thématique	ETUD	4
scolarité-controverse	La controverse sur la hausse.	Thématique	ETUD	4
scolarité-élevé	Provinces où les droits de scolarité sont plus élevés.	Thématique	ETUD	4
scolarité-foulée	Locution : dans la foulée.	Thématique	ETUD	4
scolarité-montée	La montée des droits de scolarité.	Thématique	ETUD	4
scolarité-récurrent	Notion de charge, récurrence des frais, de dépenses.	Thématique	ETUD	4
gouvernement-muscler	Conservatisme musclé du gouvernement.	Thématique	GOUV	3
gouvernement-user	Gouvernement usé par la crise.	Thématique	GOUV	3
gouvernement-fermer	Fermer la porte.	Thématique	GOUV	3
étudiant-réduction	Compensations offertes à la hausse : réduction de taxes, etc.	Thématique	GOUV	3
étudiant-affirmer	Affirmations à propos des étudiants.	Thématique	ETUD	3
gouvernement-minier	Référence aux investissements du gouvernement dans le secteur minier, sans recevoir des redevances.	Thématique	ETUD	3
gouvernement-adopter	Mesures adoptées par le gouvernement concernant différents sujets.	Thématique	GOUV	3
hausse-automne	Automne : entrée en vigueur de la hausse.	Thématique	ETUD	3
scolarité-amener	Réactions amenées par la hausse des frais.	Thématique	ETUD	3
hausse-baser	Arguments sur lesquels se base la hausse.	Thématique	ETUD	3
hausse-coalition	Coalition contre la hausse.	Thématique	ETUD	3

Annexe F. Articles du corpus

Identifiant	Titre	Source
news20120215LE20120215_a9_droits	Universités - Hausse des droits contre dérives immobilières...	Le Devoir
news20120215OP120215226595245	Laissez-les sécher	Le Journal de Montréal
news20120216LE20120216_a8_avant	Droits de scolarité - Avant-goût électoral	Le Devoir
news20120217LE20120217_a9_vivement	Vivement un printemps québécois	Le Devoir
news20120218LE20120218_b5_sens	Grève étudiante - Individualisme contre sens de la communauté	Le Devoir
news20120218QT0016	La hausse des frais concerne tout le monde	Le Quotidien
news20120218NV0030	Plein de gaspillage dans les universités	Le Nouvelliste
news20120218LA0039	Prêt à payer plus d'impôt	La Presse
news20120219OR120219227381050	Un citoyen inquiet	Le Journal de Québec
news20120221MO0045	À la jonction de deux groupes	Métro (Montréal)
news20120221NV0022	Pourquoi faire la grève ?	Le Nouvelliste
news20120221LS0043	Respecter les étudiants	Le Soleil
news20120222LS0037	Comparer, un jeu dangereux	Le Soleil
news20120222TB0019	Gardez le cap!	La Tribune
news20120222TB0023	Gâtés, nos étudiants?	La Tribune
news20120222LE20120222_a8_bonnes	Lettres - De bonnes études...	Le Devoir
news20120222LE20120222_a8_heureuse	Lettres - Heureuse d'étudier au Québec	Le Devoir
news20120222LS0035	Nous sommes choyés	Le Soleil
news20120222OR120222227611928	Un sale coup bas	Le Journal de Québec
news20120223NV0027	Le Nouvelliste	Bataille pour l'avenir
news20120223LE20120223_a8_gel	Lettres - Un gel plutôt que la gratuité	Le Devoir
news20120223LE20120223_a8_iphone	Libre opinion - J'ai un iPhone, donc je suis?	Le Devoir
news20120224LT0018	Entre l'idéal et la réalité, des choix...	Le Droit
news20120224TB0015	La démocratie élastique	La Tribune
news20120224QT0021	La force du nombre	Le Quotidien
news20120224LE20120224_a8_moral	Libre opinion - Grèves étudiantes : la fin d'un engagement moral?	Le Devoir
news20120224LE20120224_a8_calibre	Universités - Quel calibre?	Le Devoir
news20120224LA0041	Où est la véritable solidarité?	La Presse
news20120225LS0052	Des leaders avant tout	Le Soleil
news20120225LT0031	L'incompréhension du ministre	Le Droit
news20120225LE20120225_b5_marre	Marre!	Le Devoir
news20120225NV0050	Sortir pour les mauvaises raisons	Le Nouvelliste
news20120225VE0018	Une grève mal engagée avec ses dérapages	La Voix de l'Est
news20120226LS0036	Faites votre part	Le Soleil
news20120226LS0035	Le mouvement étudiant a raison	Le Soleil
news20120227NV0017	Jamais, jamais, jamais!	Le Nouvelliste
news20120227MO0031	Lettre d'une étudiante	Métro (Montréal)
news20120228LE20120228_a9_niveau	Pour la hausse du niveau de débat!	Le Devoir
news20120228LE20120228_a9_pays	Un pays de projets	Le Devoir

news20120229MO0036	Cette fameuse responsabilité sociale	Métro (Montréal)
news20120229LS0033	Le iPhone et la solidarité	Le Soleil
news20120302OR120302228889283	Un flop	Le Journal de Québec
news20120309OP120309229806723	Les nouveaux curés	Le Journal de Montréal
news20120310OP120310229937769	Au-delà de l'argent	Le Journal de Montréal
news20120312LE20120312_b7_contestation	À l'école de la contestation	Le Devoir
news20120312LA0032	Un gros « pas dans ma cour »	La Presse
news20120314OP120314230493361	Le choeur des enfants de coeur	Le Journal de Montréal
news20120314LA0046	Les petits carrés rouges	La Presse
news20120320OP120320231607668	Rentabilité électorale	Le Journal de Montréal
news20120323OP120323232002931	Monique nous manque	Le Journal de Montréal
news20120323LA0040	Négocier? Non. Écouter? Oui.	La Presse
news20120324LE20120324_b3_pas	Les premiers pas	Le Devoir
news20120326OP120326232492997	Lettre ouverte aux recteurs	Le Journal de Montréal
news20120326OP120326232493039	Questions d'argent	Le Journal de Montréal
news20120327OP120327232623063	Lettre à Éric Duhaime	Le Journal de Montréal
news20120328OP120328232883104	Des propos scandaleux !	Le Journal de Montréal
news20120401OR120401233968795	Courriers	Le Journal de Québec
news20120401OR120401233968816	On a besoin de jeunes gagnants	Le Journal de Québec
news20120402QT0016	L'accessibilité n'est pas menacée	Le Quotidien
news20120402OP120402234098871	Le carré jaune	Le Journal de Montréal
news20120402OP120402234098902	Pourquoi négocier ?	Le Journal de Montréal
news20120403TB0020	Entendez nos voix, Monsieurs Charest	La Tribune
news20120403LT0023	Plein les bras	Le Droit
news20120403VE0023	Pourquoi hausser les droits de scolarité?	La Voix de l'Est
news20120403OP120403234261950	Diplôme de boîte de céréales	Le Journal de Montréal
news20120403OP120403234261946	Un peuple allergique aux chiffres	Le Journal de Montréal
news20120404OR120404234719678	Aveos, Rio Tinto, étudiants...	Le Journal de Québec
news20120404QT0019	Du sable dans l'engrenage?	Le Quotidien
news20120404LE20120404_a9_sous	Grève étudiante : au-delà des sous	Le Devoir
news20120404VE0024	Un conflit à régler ou un règlement de comptes entre les générations?	La Voix de l'Est
news20120404LA0007	La belle vie	La Presse
news20120404CY4512188	Grève étudiante : la bataille de l'image	La Presse
news20120405LT0026	Conflit étudiant	Le Droit
news20120405MO0057	Debout, les 55 ans et plus!	Métro (Montréal)
news20120405LE20120405_a8_nombrils	Grève étudiante - Les nombrils	Le Devoir
news20120405LT0025	La grève tourne à la guérilla	Le Droit
news20120405LE20120405_a9_arguments	La réplique Droits de scolarité - Des arguments en demi-teintes	Le Devoir
news20120405LS0040	Profiter de l'ouverture	Le Soleil
news20120405NV0024	Sortie de crise	Le Nouvelliste
news20120406LE20120406_a8_etincelle	Grève étudiante - L'étincelle	Le Devoir
news20120406LT0026	L'argent au mauvais endroit	Le Droit
news20120406LE20120406_a9_mouille	La ministre se mouille	Le Devoir

news20120406QT0017	Le dialogue au lieu de la réprimande	Le Quotidien
news20120406LE20120406_a8_comment	Lettres - Comment comprendre ?	Le Devoir
news20120406LE20120406_a8_proposition	Lettres - Une proposition	Le Devoir
news20120406LT0027	Pas la bonne manière	Le Droit
news20120406LT0025	Près de la date-butoir	Le Droit
news20120406QT0014	Respecter la démocratie	Le Quotidien
news20120406MO0028	Trop financées, mal gérées	Métro (Montréal)
news20120406LT0028	Une grève inutile et injuste	Le Droit
news20120406OP120406235537319	Étudiants confondus	Le Journal de Montréal - Final
news20120407TB0027	Des mesures pour accroître l'accessibilité universitaire	La Tribune
news20120407LT0024	Droits de scolarité : Québec doit reculer	Le Droit
news20120407NV0025	Droits de scolarité : quatre concepts à évaluer	Le Nouvelliste
news20120407VE0013	En réponse à Monsieur C.A.	La Voix de l'Est
news20120407OR120407235833172	Grèves étudiantes : saviez-vous que?	Le Journal de Québec
news20120407TB0116	L'intégration, c'est être sur le même chemin	La Tribune
news20120407OR120407235833135	La crise du bacon rouge	Le Journal de Québec
news20120407NV0026	Le Québec n'a plus les valeurs à la bonne place	Le Nouvelliste
news20120407LA0008	Des grenailles	La Presse
news20120407OP120407235832252	Le non-débat	Le Journal de Montréal
news20120408OR120408235967586	Morosité printanière	Le Journal de Québec
news20120408OP120408235967354	Déprime pascale	Le Journal de Montréal
news20120409OR120409236130735	Chialer ou choisir	Le Journal de Québec
news20120409OR120409236130731	La travailleuse contre l'assisté social	Le Journal de Québec
news20120409NV0019	Philosophie distincte	Le Nouvelliste
news20120409LA0033	Étudiants : le début de la fin	La Presse
news20120410OR120410236229344	Le chat sort du sac...	Le Journal de Québec
news20120410LE20120410_a6_recherche	Libre opinion - Des oubliées : les études supérieures et la recherche	Le Devoir
news20120410VE0012	Ne cédez pas à la pression	La Voix de l'Est
news20120410MO0042	Scolarité et Plan Nord	Métro (Montréal)
news20120410OP120410236228032	Jouer à la révolution	Le Journal de Montréal
news20120411LS0049	La hausse nuit-elle réellement?	Le Soleil
news20120411LT0027	Les leçons apprises sur le piquet de grève	Le Droit
news20120411LE20120411_a8_cartons	Lettres - Des cartons rouges	Le Devoir
news20120411VE0017	Oui à l'accessibilité aux études	La Voix de l'Est
news20120411LS0051	Un vent de fraîcheur	Le Soleil
news20120411OP120411236553612	Directions aplatventristes	Le Journal de Montréal
news20120411LA0055	Quelle logique marchande?	La Presse
news20120412LT0029	Conflit étudiant au Québec	Le Droit
news20120412NV0019	Continuez!	Le Nouvelliste
news20120412MO0045	En réalité, un gel de frais signifie une baisse	Métro (Montréal)
news20120412LT0028	La gratuité a un prix	Le Droit
news20120412TB0015	La négociation a toujours sa place	La Tribune

news20120412LA0039	La valeur des diplômes	La Presse
news20120412LE20120412_a8_irresponsable	Libre opinion - L'attitude du gouvernement est irresponsable	Le Devoir
news20120412TB0016	Retournez en classe	La Tribune
news20120412VE0012	Une offre trompeuse	La Voix de l'Est
news20120412LT0031	Une vision plus humaine	Le Droit
news20120412OP120412236753446	Le party	Le Journal de Montréal
news20120413LE20120413_a8_enlisement	Grève étudiante - Enlisement	Le Devoir
news20120413LE20120413_a9_ministre	La ministre fait fausse route	Le Devoir
news20120413MO0036	Le véritable courage 2	Métro (Montréal)
news20120413MO0035	Merci de ne pas lâcher	Métro (Montréal)
news20120413NV0031	Oui à la grève	Le Nouvelliste
news20120413LA0009	La provocation	La Presse
news20120413LA0047	Le syndrome du chaton	La Presse
news20120414NV0028	C'est carrément de l'intimidation	Le Nouvelliste
news20120414TB0078	Carré rouge, monsieur Charest!	La Tribune
news20120414LT0024	Démagogie ignorante	Le Droit
news20120414TB0073	Des élections pour sortir de la crise?	La Tribune
news20120414LE20120414_b5_appel	Grève étudiante - Appel au dialogue	Le Devoir
news20120414LE20120414_b5_envers	L'envers de la médaille	Le Devoir
news20120414LE20120414_b4_greve	Lettres - Il n'y a pas de grève étudiante	Le Devoir
news20120414TB0077	Parlez-vous!	La Tribune
news20120414VE0025	SVP, M. Charest, ne reculez pas!	La Voix de l'Est
news20120414QT0020	Vieilles ornières	Le Quotidien
news20120415LS0037	La société doit gagner	Le Soleil
news20120415OP120415237572853	Le party (2)	Le Journal de Montréal
news20120416LE20120416_a7_peuples	Autochtones, droits de scolarité et Plan Nord - Peuples invisibles... à l'université	Le Devoir
news20120416LT0026	Comme un sot!	Le Devoir
news20120416LE20120416_a6_vlan	Garderies à 7$ - Et vlan!	Le Devoir
news20120416MO0032	Injustice sociale	Métro (Montréal)
news20120416LT0025	Investir dans l'éducation	Le Droit
news20120416LT0027	Le modèle universitaire ontarien	Le Droit
news20120416LS0039	Les absents ont tort	Le Soleil
news20120416LT0028	Propos infantilisants	Le Droit
news20120416LE20120416_b7_manifester	Médias - Manifester au temps du numérique	Le Devoir
news20120417NV0018	Avoir de la CLASSE...	Le Nouvelliste
news20120417LT0024	Charest et la majorité	Le Droit
news20120417LT0023	Démocratie tolérante	Le Droit
news20120417NV0016	Éducation et mode de vie	Le Nouvelliste
news20120417LE20120417_a8_carton	Grève étudiante - Carton d'invitation	Le Devoir
news20120417LT0022	Laissez-nous étudier!	Le Droit
news20120417LE20120417_a8_antidialogue	Libre opinion - La stratégie antidialogique	Le Devoir
news20120417LS0039	Main tendue et soufflet	Le Soleil
news20120417NV0019	Notre lutte est altruiste	Le Nouvelliste

news20120417LT0025	Ordis, iPad et tutti quanti	Le Droit
news20120417LT0026	Si pauvres (?) et si branchés	Le Droit
news20120417LA0003	Twilight zone	La Presse
news20120417OP120417237833619	Deux classes d'étudiants	Le Journal de Montréal
news20120417LA0039	Notes sur un boycott	La Presse
news20120418VE0019	Est-il possible de changer d'idée sur la question de la hausse des frais de scolarité?	La Voix de l'Est
news20120418NV0023	Le Nouvelliste	Le Nouvelliste
news20120418VE0018	Pourquoi est-ce que le conflit perdure?	La Voix de l'Est
news20120418LS0028	Soyons logiques	Le Soleil
news20120419MO0043	C'est un boycott	Métro (Montréal)
news20120419TB0020	Des enseignants interpellent la rectrice	La Tribune
news20120419LE20120419_a9_ideologique	Droits de scolarité - Une infamie idéologique	Le Devoir
news20120419LT0027	L'injonction n'a rien réglé	Le Droit
news20120419LE20120419_a8_strategie	Libre opinion - La stratégie du cul-de-sac	Le Devoir
news20120419TB0018	Pas de son ressort?	La Tribune
news20120419VE0014	Tirer dans tous les sens	La Voix de l'Est
news20120419TB0017	Une violence inacceptable	La Tribune
news20120419LT0032	Violence policière	Le Droit
news20120419OP120419238371143	Que se passe-t-il au Québec?	Le Journal de Montréal
news20120420LT0022	Discuter ou démissionner	Le Droit
news20120420LE20120420_a8_greve	Grève étudiante - Faire ses CLASSEs	Le Devoir
news20120420LE20120420_a9_etat	Intimidation et conflit étudiant - L'université en état de siège	Le Devoir
news20120420LT0017	Journalisme et sécurité	Le Droit
news20120420LE20120420_a8_libre	Libre opinion - L'art de la guerre	Le Devoir
news20120420VE0018	Pathétique !	La Voix de l'Est
news20120420TB0017	Un appel à la raison	La Tribune
news20120420TB0016	Vous avez tort, M. Charest	La Tribune
news20120420LA0007	Je te tiens, tu me tiens par la barbichette	La Presse
news20120421LS0051	Appel au pragmatisme	Le Soleil
news20120421LS0058	Assumez-vous	Le Soleil
news20120421LS0056	Autres temps, autres moeurs...	Le Soleil
news20120421LS0062	Croient-ils en leur avenir?	Le Soleil
news20120421VE0019	Démocratie	La Voix de l'Est
news20120421LS0061	La loi de la CLASSE	Le Soleil
news20120421LS0059	Le cégep responsable	Le Soleil
news20120421LE20120421_b4_ecoles	Lettres - Et les écoles privées...	Le Devoir
news20120421LE20120421_b4_faites	Lettres - Faites votre travail	Le Devoir
news20120421LE20120421_b5_mere	Nous, mères indignées	Le Devoir
news20120421LT0025	Une lettre au premier ministre	Le Droit
news20120421LS0063	Violence aveugle	Le Soleil
news20120422OP120422238817667	Philo 101	Le Journal de Montréal
news20120423LE348211	« School as a business ? »	Le Devoir
news20120423NV0021	Des solutions, ça presse!	Le Nouvelliste

news20120423QT0016	Le privé doit faire sa part	Le Quotidien
news20120423LE0002	Lettre à mes étudiants	Le Devoir
news20120424NV0046	Agitation politique	Le Nouvelliste
news20120424NV0044	Ceux qui décident	Le Nouvelliste
news20120424VE0019	Des étudiants ignorés, discrédités, ridiculisés	La Voix de l'Est
news20120424NV0047	Des propos irresponsables	Le Nouvelliste
news20120424LT0029	Élus pour gouverner	Le Droit
news20120424LS0034	Indigne d'un gouvernant	Le Soleil
news20120424LE0005	L'urgence d'agir	Le Devoir
news20120424LT0027	La CLASSE, au-dessus des lois	Le Droit
news20120424LS0036	La leçon de démocratie	Le Soleil
news20120424VE0018	S'excuser pour se remettre au-dessus de la mêlée	La Voix de l'Est
news20120424LT0031	Un sondage sans effet sur la crise	Le Droit
news20120424OP120424239078735	La CLASSE déclassée	Le Journal de Montréal - Final
news20120424LA0048	Six pistes de solution	La Presse
news20120425TB0024	Assez c'est assez !	La Tribune
news20120425QT0026	La hausse aura plus d'impacts en région	Le Quotidien
news20120425TB0026	Les droits des uns et des autres	La Tribune
news20120425LE20120425_a8_souviens	Lettres - Je me souviens	Le Devoir
news20120425LE20120425_a8_violent	Libre opinion - Violent débat... avec moi-même	Le Devoir
news20120425LT0023	Pour une solidarité intergénérationnelle	Le Droit
news20120425OP120425239243694	Grèves en culottes courtes	Le Journal de Montréal
news20120425OP120425239243689	Qui est égoïste?	Le Journal de Montréal
news20120426LT0034	Choux gras et chienlit	Le Droit
news20120426LE20120426_a8_machiavel	Grève étudiante - Machiavel à Québec	Le Devoir
news20120426VE0012	L'accès à l'université doit être un droit, pas un privilège	La Voix de l'Est
news20120426LE20120426_a8_face	Libre opinion - La face cachée de la grève étudiante	Le Devoir
news20120426VE0013	Rassembleur, un projet de société étude-travail-famille?	La Voix de l'Est
news20120426LT0035	« Terrorisme »? Déplorable!	Le Droit
news20120427LT0028	Crise de société?	Le Droit
news20120427LE348533	Et la paix sociale ?	Le Devoir
news20120427MO0026	La table est mise pour des élections	Métro (Montréal)
news20120427LE348560	Le droit à mon opinion	Le Devoir
news20120427NV0025	Les étudiants ont raison	Le Nouvelliste
news20120427LS0049	Mon opinion change	Le Soleil
news20120427TB0021	Profs du secondaire, concernés et solidaires!	La Tribune
news20120427VE0017	Solution régressive	La Voix de l'Est
news20120427VE0016	Un système à nos frais	La Voix de l'Est
news20120427LA0011	Le mur	La Presse
news20120427OR120427239601750	Élections envisagées	Le Journal de Québec
news20120427LA0051	Tasser la CLASSE	La Presse
news20120428LT0027	Coup de poker	Le Droit
news20120428TB0029	En faveur d'un médiateur	La Tribune

news20120428VE0014	Forte pression sur le monde étudiant	La Voix de l'Est
news20120428TB0025	Indignés et fiers	La Tribune
news20120428LA0224	La pauvreté, grande responsable	La Presse
news20120428LS0051	Le devoir des étudiants	Le Soleil
news20120428TB0028	Les ACEF contre les hausses	La Tribune
news20120428LS0056	Opposés à la réélection de Denis Brière	Le Soleil
news20120428LS0054	Retournez aux études	Le Soleil
news20120428TB0026	Revenir à l'essentiel	La Tribune
news20120428VE0015	Un débat aux proportions nationales	La Voix de l'Est
news20120428OR120428239768866	Élu pour gouverner	Le Journal de Québec
news20120430LT0018	Crédibilité perdue	Le Droit
news20120430LS0039	Dérive autoritaire	Le Soleil
news20120430LS0036	Forcé à l'abandon	Le Soleil
news20120430NV0018	Humour et politique	Le Nouvelliste
news20120430LE2012-04-30_348806	Injonctions et grève étudiante - La primauté du droit en péril?	Le Devoir
news20120430LS0037	S'ouvrir au compromis	Le Soleil
news20120430OP120430239998728	Au royaume de la poutine	Le Journal de Montréal
news20120430OP120430239998697	Le chat sort du sac	Le Journal de Montréal
news20120430LA0033	Petit rappel sur la démocratie	La Presse
news20120501LE2012-05-01_348908	L'exemple de Lesage	Le Devoir
news20120501LA0010	Le temps file	La Presse
news20120502LA0049	Au tour des étudiants	La Presse
news20120502LA0047	Un braquage surréaliste	La Presse
news20120502OP120502240291873	Une bouillabaisse	Le Journal de Montréal
news20120503LA0013	La brèche	La Presse
news20120503OP120503240456673	Trois partis, trois visions	Le Journal de Montréal
news20120504LE2012-05-04_349172	Le cœur rouge et la tête verte	Le Devoir
news20120504OP120504240587045	Chèque de paie ou de BS	Le Journal de Montréal
news20120504OP120504240587056	En bon père de famille	Le Journal de Montréal
news20120504LA0038	L'erreur libérale	La Presse
news20120507LE2012-05-07_349436	La faute des autres	Le Devoir
news20120507OR120507240980335	Il est prêt	Le Journal de Québec
news20120507OP120507240979438	Le mors aux dents	Le Journal de Montréal
news20120507OP120507240979526	Une amère déception	Le Journal de Montréal
news20120508LE2012-05-08_349509	Le drap rouge	Le Devoir
news20120508LA0006	Pari (presque) impossible	La Presse
news20120509LA0048	Le mythe des 189 millions	La Presse
news20120510LA0013	L'Université Laval s'installe au cœur de Montréal	La Presse
news20120510LE2012-05-10_349717	Le catalyseur	Le Devoir
news20120511OP120511241504553	Un extrémisme destructeur	Le Journal de Montréal
news20120512LA0002	La nuit des longues négos	La Presse
news20120512LE2012-05-12_349932	Le nouvel ennemi public	Le Devoir
news20120514LE2012-05-14_349983	Médias - De guerre d'image lasse	Le Devoir
news20120514LA0004	Pour en finir avec la grève	La Presse

news20120514OP120514241926820	Ah, nos « valeurs »	Le Journal de Montréal
news20120515LE2012-05-15_350111	Place aux muscles	Le Devoir
news20120515LA0012	Une décision lucide	La Presse
news20120515OP120515242254453	Coup de tonnerre	Le Journal de Montréal
news20120516LA0051	Au pays des « momos »	La Presse
news20120516QVHM120516242328378	Le désir de censure	24 heures Montréal
news20120516OP120516242327176	Le SYSTÈME réplique	Le Journal de Montréal
news20120517LE2012-05-17_350262	Le choc des perspectives	Le Devoir
news20120517LA0015	Un pari risqué	La Presse
news20120517LA0059	Le clivage franco-anglo	La Presse
news20120518OP120518242585513	Une matraque jouet	Le Journal de Montréal
news20120519LE2012-05-19_350476	Le pyromane	Le Devoir
news20120519LE2012-05-19_350472	Les spectateurs enragés	Le Devoir
news20120519OP120519242983703	Deux dans une	Le Journal de Montréal
news20120520OP120520243108961	Le vendeur	Le Journal de Montréal
news20120522OP120522243336805	Dérapages et démesure	Le Journal de Montréal
news20120523OP120523243436542	Que faire ?	Le Journal de Montréal
news20120523LA0060	Sauver la social-démocratie	La Presse
news20120524LE2012-05-24_350764	On fait quoi ?	Le Devoir
news20120524OP120524243634030	Qui brime la liberté d'expression ?	Le Journal de Montréal
news20120525OP120525243797872	Journal d'une salope	Le Journal de Montréal
news20120525OR120525243799189	Négocier à la TV	Le Journal de Québec
news20120526LA0008	Comme au référendum	La Presse
news20120526LE2012-05-26_350962	Loi 78 - L'audace retrouvée	Le Devoir
news20120527OP120527244059190	Avec Aron	Le Journal de Montréal
news20120528LA0029	Cherchez l'erreur	La Presse
news20120528OP120528244190540	La peur rouge	Le Journal de Montréal
news20120528OP120528244190511	Les deux côtés de la bouche	Le Journal de Montréal
news20120529LE2012-05-29_351133	Le conflit étudiant selon Wikipédia	Le Devoir
news20120529LE2012-05-29_351131	Les marrons	Le Devoir
news20120531LE2012-05-31_351313	L'œuf de Colomb	Le Devoir
news20120531OP120531244616759	Chacun son camp	Le Journal de Montréal
news20120531OP120531244616754	Le salaud, l'idiot, et le larbin	Le Journal de Montréal
news20120601LT0019	Aux urnes!	Le Droit
news20120601LE2012-06-01_351421	Conflit étudiant - Se comparer avec Mai 68 ?	Le Devoir
news20120601TB0017	Guerre d'image	La Tribune
news20120601LS0039	Intransigeance étudiante	Le Soleil
news20120601LS0040	L'envie de les répudier avant terme	Le Soleil
news20120601LS0042	Le conflit étudiant : l'utilisateur-payeur en question	Le Soleil
news20120601LS0041	Le point de bascule?	Le Soleil
news20120601NV0020	Un printemps universel?	Le Nouvelliste
news20120601MO0036	Une autre raison de s'indigner	Métro (Montréal)
news20120601LE2012-06-01_351442	La politique du pire	Le Devoir
news20120601LE2012-06-01_351422	Les miettes tombées de la table	Le Devoir

news20120601LA0049	Peut-on sortir de l'impasse?	La Presse
news20120602LS0064	Cacophonie québécoise	Le Soleil
news20120602LS0066	Comme du bacon	Le Soleil
news20120602LE2012-06-02_351469	Conflit étudiant - La poésie de la police	Le Devoir
news20120602LT0036	Éliane Laberge entre dans la danse	Le Droit
news20120602LE2012-06-02_351471	Il faut en sortir, non ?	Le Devoir
news20120602LS0068	La CSN à la défense du travail journalistique	Le Soleil
news20120602NV0027	La solution temporaire	Le Nouvelliste
news20120602LT0035	Le syndrome des casseroles	Le Droit
news20120602TB0026	Révolution tarifaire	La Tribune
news20120602VE0008	Une responsabilité partagée	La Voix de l'Est
news20120602LA0012	La crise, prise deux	La Presse
news20120602LE2012-06-02_351538	Monarcho-libéraux contre républicains	Le Devoir
news20120602LE2012-06-02_351474	Une bénédiction	Le Devoir
news20120603LS0027	Le fardeau de renflouer les universités québécoises	Le Soleil
news20120603OP120603244976060	Après nous le déluge	Le Journal de Montréal
news20120604LE2012-06-04_351571	Conflit étudiant - Des sondages contradictoires ?	Le Devoir
news20120604MO0041	L'union fait la force	Métro (Montréal)
news20120604MO0040	Léo Bureau-Blouin	Métro (Montréal)
news20120604TB0025	Québec doit opter pour les logiciels libres	La Tribune
news20120604NV0012	Retourner en arrière pour mieux repartir	Le Nouvelliste
news20120604LA0034	La grave erreur de Mme Marois	La Presse
news20120605LE2012-06-05_351628	Conflit étudiant - Du carré rouge au drapeau du Québec	Le Devoir
news20120605NV0020	Crise ou crisette?	Le Nouvelliste
news20120605MO0031	Deux cents de l'heure	Métro (Montréal)
news20120605LS0040	Je sais que je ne sais pas	Le Soleil
news20120605NV0021	Pour quelques dollars de plus	Le Nouvelliste
news20120605LE2012-06-05_351652	Radio - Le printemps d'ici vu de Là-bas si j'y suis	Le Devoir
news20120605OP120605245241621	Les oubliés	Le Journal de Montréal
news20120606NV0021	Arrêtez de nous dénigrer!	Le Nouvelliste
news20120606LE2012-06-06_351726	Conflit étudiant - L'art de gouverner, et celui de la résistance	Le Devoir
news20120606LE2012-06-06_351727	Crise sociale - Le long souffle du printemps érable	Le Devoir
news20120606NV0025	De meilleurs citoyens	Le Nouvelliste
news20120606LT0028	La population réagit au conflit étudiant	Le Droit
news20120606TB0017	Les fantômes du passé de M. Charest	La Tribune
news20120606LA0044	Les victimes de la démographie	La Presse
news20120606OP120606245340465	Twitter et nous	Le Journal de Montréal
news20120607LS0043	Bel exemple	Le Soleil
news20120607LT0031	Ça va nous coûter!	Le Droit
news20120607TB0014	Jusqu'où les étudiants iront-ils?	La Tribune
news20120607VE0023	Le rêve d'une éducation libre de toute contrainte	La Voix de l'Est
news20120607NV0025	Un martyr nommé Khadir	Le Nouvelliste
news20120607OP120607245502122	Les banquiers	Le Journal de Montréal
news20120608LS0034	Jouer le jeu du PLQ	Le Soleil

news20120608NV0023	Nous sommes avec vous	Le Nouvelliste
news20120608LT0021	Quand un député désobéit	Le Droit
news20120608LE2012-06-08_351949	Une nouvelle forme d'activisme politique	Le Devoir
news20120609LE2012-06-09_352030	Conflit étudiant - La diversité dans la rue	Le Devoir
news20120609TB0078	Diviser pour régner	La Tribune
news20120609LS0127	Excès de zèle	Le Soleil
news20120609LS0129	Exercer son droit	Le Soleil
news20120609LA0047	Savoir dire non	La Presse
news20120609LE2012-06-09_352083	Crise étudiante - Mai 68, en gros	Le Devoir
news20120609LA0005	Le dérapage verbal	La Presse
news20120609OP120609245730778	Lettre à Jacques Villeneuve	Le Journal de Montréal
news20120611QT0013	Autre son de cloche	Le Quotidien
news20120611LI79124a2e-b30b-11e1-8897-22ba906e545b	Le nouveau monde et la crise des valeurs	Libération
news20120611LI78476534-b30b-11e1-8897-22ba906e545b	Manifestations étudiantes au Québec : de « l'enfant-roi » au porteur du rêve	Libération
news20120611LE2012-06-11_352137	Politique - Un simulacre de démocratie au Québec?	Le Devoir
news20120611VE0014	Pourquoi un tel mépris, sinon pour masquer une injustice?	La Voix de l'Est
news20120611LS0029	Solidarité bidirectionnelle	Le Soleil
news20120611LS0027	Un événement fondateur	Le Soleil
news20120611LS0028	Vous résonnez, nous raisonnons	Le Soleil
news20120611OP120611245992973	Nos artistes sur Twitter	Le Journal de Montréal
news20120612LE2012-06-12_352179	Au-delà de la crise : retrouver les voies du Québec	Le Devoir
news20120612NV0020	Chialeux, les Québécois	Le Nouvelliste
news20120612MO0044	L'échec de Laurent Proulx	Métro (Montréal)
news20120612LS0025	Une solution fiscale	Le Soleil
news20120612OP120612246156538	Le péché originel	Le Journal de Montréal
news20120612LA0045	Vent de folie	La Presse
news20120613LS0038	Discipline et démocratie	Le Soleil
news20120613MO0039	Diviser pour mieux régner	Métro (Montréal)
news20120613LE0030	L'illusion de l'élection	Le Devoir
news20120613LS0039	Les chaudrons ont des noms...	Le Soleil
news20120613NV0025	Manifester pour un avenir meilleur!	Le Nouvelliste
news20120613LT0030	Perquisition chez les Khadir	Le Droit
news20120613OP120613246255079	Défendons-nous!	Le Journal de Montréal
news20120613OP120613246255078	Les fachos	Le Journal de Montréal
news20120614LE2012-06-14_352379	La réplique des casseroles aux urnes - Le désir profond d'un Québec progressiste	Le Devoir
news20120614NV0019	Le fond du baril	Le Nouvelliste
news20120614MO0049	Un désaveu politique	Métro (Montréal)
news20120614LE2012-06-14_352381	Une distinction qui s'impose	Le Devoir
news20120614OP120614246387989	J'ai des questions, comme ça...	Le Journal de Montréal
news20120614OP120614246387960	Le Québec nazi ?	Le Journal de Montréal
news20120615LE2012-06-15_352504	Conflit étudiant - La part de l'étudiant	Le Devoir

news20120615LE2012-06-15_352503	Le centre droit, la voie des jeunes Québécois ?	Le Devoir
news20120615VE0014	Voilà pourquoi j'irai manifester	La Voix de l'Est
news20120615LE2012-06-15_352506	Un premier ministre assis	Le Devoir
news20120615LA0047	Le secret de la mayonnaise	La Presse
news20120616AN0024	Qui viole actuellement la démocratie au Québec?	L'Acadie Nouvelle
news20120616LE2012-06-16_352606	Une ambition pour le Québec : une éducation accessible pour tous	Le Devoir
news20120616OP120616246681427	Le méchant, méchant chroniqueur	Le Journal de Montréal
news20120618NV0023	Le droit à tous de s'instruire	Le Nouvelliste
news20120618NV0025	Le droit d'exercer le pouvoir à sa convenance...	Le Nouvelliste
news20120618SR3249416	Nous sommes solidaires du « Printemps érable »	Le Soir - 1ER
news20120618LA0036	L'OCDE est contre le gel	La Presse
news20120618OP120618246944033	Svp, enlevez-le, M me Marois	Le Journal de Montréal
news20120619LE2012-06-19_352769	Violence - Malhonnêteté éhontée	Le Devoir
news20120620TB0034	Des enfants gâtés?	La Tribune
news20120620LS0048	Est-ce à moi de payer?	Le Soleil
news20120620LS0049	Les subventionnés	Le Soleil
news20120620TB0031	Ma proposition pour sortir de la crise	La Tribune
news20120621NV0030	Casserole et utopie	Le Nouvelliste
news20120621NV0028	Cessez ce tintamarre	Le Nouvelliste
news20120621NV0029	Grève étudiante : la dérive	Le Nouvelliste
news20120621MO0049	Le Christ ressuscité !	Métro (Montréal)
news20120621TB0017	Oui à la démarche des étudiants	La Tribune
news20120621QT0013	Un référendum pour trancher	Le Quotidien
news20120622NV0016	Carré jaune	Le Nouvelliste
news20120622LS0032	La passion de l'intérêt public	Le Soleil
news20120622LE0003	Le 24 juin, on marche POUR le Québec !	Le Devoir
news20120623LS0052	Ingratitude	Le Soleil
news20120623NV0038	Les impacts de la hausse en Mauricie	Le Nouvelliste
news20120623LS0053	Printemps érable à la grecque	Le Soleil
news20120623VE0014	Quelle St-Jean aurons-nous ?	La Voix de l'Est
news20120623OP120623247664276	Les humoristes indignés s'indigneront-ils ?	Le Journal de Montréal
news20120626LS0024	Sens des mots et bon sens	Le Soleil
news20120627LS0033	La fête de quels Québécois?	Le Soleil
news20120627NV0024	Un éditorial au contenu abusif	Le Nouvelliste
news20120628LE2012-06-28_353420	Faux coup d'cochon	Le Devoir
news20120628LT0023	Quand les élites sont dépassées	Le Droit
news20120629TB0016	Vivement les élections!	La Tribune
news20120630VE0012	Merci M. Cohen...	La Voix de l'Est
news20120630LS0051	Pensons aux étudiants	Le Soleil
news20120702OP120702248941880	Le Québec coupé en deux	Le Journal de Montréal
news20120706OP120706249433051	La langue de chez nous	Le Journal de Montréal
news20120708OP120708249728045	Pencher du même bord	Le Journal de Montréal
news20120713OP120713250481656	GNB au Noël du campeur	Le Journal de Montréal

news20120718OP120718251169635	Élections, piège à cons?	Le Journal de Montréal
news20120801VE0013	Il y a les indignés et les indignes	La Voix de l'Est
news20120802LA0041	À la recherche du moins pire	La Presse
news20120804OR120804242286878	Ne cassez pas les étudiants !	Le Journal de Québec
news20120807OP120807253659839	Dans la cour des grands	Le Journal de Montréal
news20120808TB0018	Des faits à rétablir? En effet...	La Tribune
news20120810LE2012-08-10_356453	Lettre à une campeuse en colère	Le Devoir
news20120811VE0017	Le bruit de l'orage politique électoral	La Voix de l'Est
news20120811TB0017	Merci d'être là, Gabriel Nadeau-Dubois!	La Tribune
news20120811OP120811254218382	Quelle démocratie ?	Le Journal de Montréal
news20120815VE0010	Quel pays voulons-nous ?	La Voix de l'Est
news20120820OP120820255463715	La guerre des nourrices	Le Journal de Montréal
news20120824VE0011	L'éducation : un droit pour tous	La Voix de l'Est
news20120829VE0015	But des élections... détourné	La Voix de l'Est
news20120905LA0066	La boîte à surprise	La Presse
news20120908LA0055	Les deux mains liées	La Presse
news20120908OP121021264344120	Les deux visages du harcèlement	Le Journal de Montréal
news20120910OP120910258347128	Un coup de baguette magique	Le Journal de Montréal
news20120912LA0041	Vont-ils se « casser » ici aussi?	La Presse
news20120914OP120914258805232	Respectez le Parlement !	Le Journal de Montréal
news20120922OP120922259985298	De la rue au Parlement	Le Journal de Montréal
news20120924LA0043	Gouvernance rétroactive	La Presse
news20120927LA0052	La leçon des carrés rouges	La Presse
news20121016OP121016263684835	La poule aux oeufs d'or	Le Journal de Montréal
news20121022OP121022264508138	La pure vérité	Le Journal de Montréal
news20121025OP121025265031984	On ne peut pas descendre plus bas	Le Journal de Montréal

Annexe G. Critères textuels de la matrice M4_Globaux

Terme	Spécificité	Composante	Type	Classe
éducation	+16	Thématique	Lemme	ETUD
société	+11	Thématique	Lemme	ETUD
autochtone	+8	Thématique	Lemme	ETUD
mépris	+8	Thématique	Mots	ETUD
Courchesne	+6	Thématique	Lemme	ETUD
jeunesse	+6	Thématique	Lemme	ETUD
ministre	+6	Thématique	Mots	ETUD
@card@	+5	Thématique	Lemme	ETUD
président	+5	Thématique	Lemme	ETUD
grève	+5	Thématique	Lemme	ETUD
juste part	+5	Thématique	Segment répété	ETUD
Martine Desjardins	+5	Thématique	Segment répété	ETUD
Réjean Parent	+5	Thématique	Segment répété	ETUD
génération	+5	Thématique	Lemme	ETUD
savoir	+5	Thématique	Mots	ETUD
communauté	+5	Thématique	Lemme	ETUD
pacifique	+5	Thématique	Lemme	ETUD
jeune	+5	Thématique	Lemme	ETUD
Desjardins	+4	Thématique	Lemme	ETUD
accessible	+4	Thématique	Lemme	ETUD
inégalité	+4	Thématique	Lemme	ETUD
économie	+4	Thématique	Mots	ETUD
fondamental	+4	Thématique	Lemme	ETUD
Madame Beauchamp	+4	Thématique	Segment répété	ETUD
Michelle Courchesne	+4	Thématique	Segment répété	ETUD
Mme Beauchamp	+4	Thématique	Segment répété	ETUD
Mme Courchesne	+4	Thématique	Segment répété	ETUD
présidente de la FEUQ	+4	Thématique	Segment répété	ETUD
système économique	+4	Thématique	Segment répété	ETUD
Beauchamp	+4	Thématique	Lemme	ETUD
développement	+4	Thématique	Lemme	ETUD
juste	+4	Thématique	Mots	ETUD
accès	+4	Thématique	Lemme	ETUD
université	+4	Thématique	Lemme	ETUD
construction	+4	Thématique	Lemme	ETUD
humain	+4	Thématique	Lemme	ETUD
population	+3	Thématique	Lemme	ETUD
commission	+3	Thématique	Mots	ETUD
violence	+3	Thématique	Lemme	ETUD
art	+3	Thématique	Mots	ETUD
enseignement	+3	Thématique	Lemme	ETUD
instruire	+3	Thématique	Lemme	ETUD
instance	+3	Thématique	Lemme	ETUD
résistance	+3	Thématique	Mots	ETUD
entreprise	+3	Thématique	Lemme	ETUD
reconnaître	+3	Thématique	Mots	ETUD
souvenir	+3	Thématique	Lemme	ETUD
% en	+3	Thématique	Segment répété	ETUD

chefs syndicaux	+3	Thématique	Segment répété	ETUD
classe moyenne	+3	Thématique	Segment répété	ETUD
communauté universitaire	+3	Thématique	Segment répété	ETUD
droit de grève	+3	Thématique	Segment répété	ETUD
droits de scolarité	+3	Thématique	Segment répété	ETUD
en cinq ans	+3	Thématique	Segment répété	ETUD
gouvernement du Québec	+3	Thématique	Segment répété	ETUD
gouvernement québécois	+3	Thématique	Segment répété	ETUD
le 22	+3	Thématique	Segment répété	ETUD
ligne dure	+3	Thématique	Segment répété	ETUD
niveau de vie	+3	Thématique	Segment répété	ETUD
point de presse	+3	Thématique	Segment répété	ETUD
pour tous	+3	Thématique	Segment répété	ETUD
premier ministre	+3	Thématique	Segment répété	ETUD
sur le dos	+3	Thématique	Segment répété	ETUD
commun	+3	Thématique	Lemme	ETUD
avenir	+3	Thématique	Lemme	ETUD
immobilier	+3	Thématique	Lemme	ETUD
refuser	+3	Thématique	Lemme	ETUD
administrations	+3	Thématique	Mots	ETUD
adoption	+3	Thématique	Lemme	ETUD
matraque	+3	Thématique	Lemme	ETUD
famille	+3	Thématique	Mots	ETUD
stratégie	+3	Thématique	Lemme	ETUD
homme	+3	Thématique	Lemme	ETUD
quart	+3	Thématique	Lemme	ETUD
diviser	+3	Thématique	Mots	ETUD
solidarité	+3	Thématique	Lemme	ETUD
guerre	+3	Thématique	Mots	ETUD
asseoir	+3	Thématique	Mots	ETUD
valeurs	+3	Thématique	Mots	ETUD
dérive	+3	Thématique	Lemme	ETUD
images	+2	Thématique	Mots	ETUD
revendication	+2	Thématique	Lemme	ETUD
arrogance	+2	Thématique	Lemme	ETUD
citoyen	+2	Thématique	Lemme	ETUD
mobilisation	+2	Thématique	Lemme	ETUD
Bourassa	+2	Thématique	Lemme	ETUD
indigne	+2	Thématique	Lemme	ETUD
réussite	+2	Thématique	Lemme	ETUD
endettés	+2	Thématique	Mots	ETUD
pourrir	+2	Thématique	Mots	ETUD
campus	+2	Thématique	Lemme	ETUD
supérieur	+2	Thématique	Mots	ETUD
corruption	+2	Thématique	Lemme	ETUD
ministère	+2	Thématique	Lemme	ETUD
contribuable	+2	Thématique	Mots	ETUD
mobiliser	+2	Thématique	Lemme	ETUD
critique	+2	Thématique	Mots	ETUD
iPhone	+2	Thématique	Lemme	ETUD
répressif	+2	Thématique	Lemme	ETUD
UdeM	+2	Thématique	Lemme	ETUD
unir	+2	Thématique	Lemme	ETUD
imposé	+2	Thématique	Mots	ETUD
travailler	+2	Thématique	Mots	ETUD

délégué	+2	Thématique	Lemme	ETUD
instruction	+2	Thématique	Lemme	ETUD
ONU	+2	Thématique	Lemme	ETUD
réveil	+2	Thématique	Lemme	ETUD
Rocher	+2	Thématique	Lemme	ETUD
salarial	+2	Thématique	Lemme	ETUD
policier	+2	Thématique	Lemme	ETUD
étudiant	+2	Thématique	Mots	ETUD
procéder	+2	Thématique	Lemme	ETUD
à tous	+2	Thématique	Segment répété	ETUD
étudiants universitaires	+2	Thématique	Segment répété	ETUD
loi 78	+2	Thématique	Segment répété	ETUD
milliards de dollars	+2	Thématique	Segment répété	ETUD
partie de la	+2	Thématique	Segment répété	ETUD
temps plein	+2	Thématique	Segment répété	ETUD
nous	+12	Dialogique	Lemme	ETUD
que nous	+10	Dialogique	Segment répété	ETUD
«	+9	Dialogique	Lemme	ETUD
»	+9	Dialogique	Lemme	ETUD
notre	+7	Dialogique	Mots	ETUD
Madame	+4	Dialogique	Lemme	ETUD
à nous	+4	Dialogique	Segment répété	ETUD
de notre	+4	Dialogique	Segment répété	ETUD
nos jeunes	+4	Dialogique	Segment répété	ETUD
nous les	+4	Dialogique	Segment répété	ETUD
votre gouvernement	+4	Dialogique	Segment répété	ETUD
sentiment	+3	Dialogique	Mots	ETUD
aux jeunes	+3	Dialogique	Segment répété	ETUD
de votre	+3	Dialogique	Segment répété	ETUD
je vois	+3	Dialogique	Segment répété	ETUD
les étudiants n	+3	Dialogique	Segment répété	ETUD
Les étudiants ont	+3	Dialogique	Segment répété	ETUD
nous avons	+3	Dialogique	Segment répété	ETUD
nous nous	+3	Dialogique	Segment répété	ETUD
nous pouvons	+3	Dialogique	Segment répété	ETUD
nous sommes	+3	Dialogique	Segment répété	ETUD
nous voulons	+3	Dialogique	Segment répété	ETUD
voulons	+3	Dialogique	Mots	ETUD
M	+3	Dialogique	Lemme	ETUD
je	+3	Dialogique	Mots	ETUD
michele	+3	Dialogique	Lemme	ETUD
ouimet	+3	Dialogique	Lemme	ETUD
J	+3	Dialogique	Mots	ETUD
votre	+2	Dialogique	Mots	ETUD
lapresse	+2	Dialogique	Lemme	ETUD
joindre	+2	Dialogique	Mots	ETUD
Professeur	+2	Dialogique	Mots	ETUD
avec eux	+2	Dialogique	Segment répété	ETUD
ces étudiants	+2	Dialogique	Segment répété	ETUD
ceux et celles qui	+2	Dialogique	Segment répété	ETUD
de demain	+5	Dialectique	Segment répété	ETUD
et	+4	Dialectique	Mots	ETUD
-il	+4	Dialectique	Mots	ETUD
de nouveaux	+4	Dialectique	Segment répété	ETUD
et plus	+4	Dialectique	Segment répété	ETUD

demain	+4	Dialectique	Lemme	ETUD
-nous	+4	Dialectique	Mots	ETUD
,	+4	Dialectique	Lemme	ETUD
alors qu	+3	Dialectique	Segment répété	ETUD
car il	+3	Dialectique	Segment répété	ETUD
de plus en plus	+3	Dialectique	Segment répété	ETUD
Depuis des	+3	Dialectique	Segment répété	ETUD
dès le	+3	Dialectique	Segment répété	ETUD
dire que les	+3	Dialectique	Segment répété	ETUD
En effet	+3	Dialectique	Segment répété	ETUD
est encore	+3	Dialectique	Segment répété	ETUD
est là	+3	Dialectique	Segment répété	ETUD
et surtout	+3	Dialectique	Segment répété	ETUD
non pas	+3	Dialectique	Segment répété	ETUD
non seulement	+3	Dialectique	Segment répété	ETUD
plus de	+3	Dialectique	Segment répété	ETUD
plus grand	+3	Dialectique	Segment répété	ETUD
près de	+3	Dialectique	Segment répété	ETUD
question de	+3	Dialectique	Segment répété	ETUD
tout comme	+3	Dialectique	Segment répété	ETUD
moment	+3	Dialectique	Mots	ETUD
véritable	+3	Dialectique	Mots	ETUD
aujourd hui	+3	Dialectique	Mots	ETUD
manifestement	+3	Dialectique	Lemme	ETUD
quelle	+3	Dialectique	Mots	ETUD
pourtant	+2	Dialectique	Lemme	ETUD
Depuis	+2	Dialectique	Mots	ETUD
Lors	+2	Dialectique	Mots	ETUD
Toutefois	+2	Dialectique	Mots	ETUD
moments	+2	Dialectique	Mots	ETUD
voilà	+2	Dialectique	Lemme	ETUD
notamment	+2	Dialectique	Mots	ETUD
encore	+2	Dialectique	Mots	ETUD
à cette	+2	Dialectique	Segment répété	ETUD
arriver à	+2	Dialectique	Segment répété	ETUD
au détriment	+2	Dialectique	Segment répété	ETUD
Au lieu	+2	Dialectique	Segment répété	ETUD
au profit	+2	Dialectique	Segment répété	ETUD
au sein	+2	Dialectique	Segment répété	ETUD
ce temps	+2	Dialectique	Segment répété	ETUD
en matière	+2	Dialectique	Segment répété	ETUD
mais aussi	+2	Dialectique	Segment répété	ETUD
artistes	+14	Thématique	Mots	GOUV
frais de scolarité	+10	Thématique	Segment répété	GOUV
gauche	+8	Thématique	Lemme	GOUV
Marois	+8	Thématique	Mots	GOUV
remboursement	+7	Thématique	Lemme	GOUV
frais	+7	Thématique	Mots	GOUV
carré	+7	Thématique	Lemme	GOUV
boycott	+6	Thématique	Mots	GOUV
beurre	+5	Thématique	Mots	GOUV
province	+5	Thématique	Lemme	GOUV
majorité	+5	Thématique	Lemme	GOUV
humoriste	+5	Thématique	Lemme	GOUV
gel	+5	Thématique	Lemme	GOUV
idéologique	+5	Thématique	Lemme	GOUV

associations étudiantes	+5	Thématique	Segment répété	GOUV
Mme Marois	+5	Thématique	Segment répété	GOUV
Pauline Marois	+5	Thématique	Segment répété	GOUV
boycotter	+5	Thématique	Lemme	GOUV
Twitter	+5	Thématique	Mots	GOUV
boycottage	+5	Thématique	Lemme	GOUV
bas	+5	Thématique	Lemme	GOUV
côté	+4	Thématique	Lemme	GOUV
cours	+4	Thématique	Mots	GOUV
fasciste	+4	Thématique	Lemme	GOUV
Québécois	+4	Thématique	Mots	GOUV
boycott des cours	+4	Thématique	Segment répété	GOUV
de la rue	+4	Thématique	Segment répété	GOUV
du côté	+4	Thématique	Segment répété	GOUV
gel des frais	+4	Thématique	Segment répété	GOUV
hausse des frais	+4	Thématique	Segment répété	GOUV
prêts et bourses	+4	Thématique	Segment répété	GOUV
rouge	+4	Thématique	Lemme	GOUV
bonification	+4	Thématique	Lemme	GOUV
session	+4	Thématique	Lemme	GOUV
hausse	+4	Thématique	Lemme	GOUV
inflation	+4	Thématique	Mots	GOUV
votent	+4	Thématique	Mots	GOUV
péquiste	+4	Thématique	Lemme	GOUV
promettre	+4	Thématique	Lemme	GOUV
syndical	+4	Thématique	Lemme	GOUV
association	+4	Thématique	Lemme	GOUV
intimider	+4	Thématique	Lemme	GOUV
geler	+3	Thématique	Lemme	GOUV
pont	+3	Thématique	Lemme	GOUV
Barbe	+3	Thématique	Mots	GOUV
Dubois	+3	Thématique	Mots	GOUV
ajuster	+3	Thématique	Lemme	GOUV
méfait	+3	Thématique	Lemme	GOUV
minimum	+3	Thématique	Mots	GOUV
responsabilité	+3	Thématique	Lemme	GOUV
PQ	+3	Thématique	Mots	GOUV
annuler	+3	Thématique	Mots	GOUV
écrire	+3	Thématique	Lemme	GOUV
social-démocratie	+3	Thématique	Mots	GOUV
impasse	+3	Thématique	Mots	GOUV
groupe	+3	Thématique	Lemme	GOUV
petit	+3	Thématique	Lemme	GOUV
budget	+3	Thématique	Mots	GOUV
annoncées	+3	Thématique	Mots	GOUV
dénouer	+3	Thématique	Lemme	GOUV
centrale	+3	Thématique	Lemme	GOUV
revenu	+3	Thématique	Lemme	GOUV
accessibilité aux études	+3	Thématique	Segment répété	GOUV
Assemblée nationale	+3	Thématique	Segment répété	GOUV
autres provinces	+3	Thématique	Segment répété	GOUV
campagne électorale	+3	Thématique	Segment répété	GOUV
crise étudiante	+3	Thématique	Segment répété	GOUV
des études supérieures	+3	Thématique	Segment répété	GOUV
études universitaires	+3	Thématique	Segment répété	GOUV

gel des droits	+3	Thématique	Segment répété	GOUV
gouvernement Charest	+3	Thématique	Segment répété	GOUV
lutte contre	+3	Thématique	Segment répété	GOUV
majorité de Québécois	+3	Thématique	Segment répété	GOUV
médias sociaux	+3	Thématique	Segment répété	GOUV
ministre Line Beauchamp	+3	Thématique	Segment répété	GOUV
modèle québécois	+3	Thématique	Segment répété	GOUV
Parti québécois	+3	Thématique	Segment répété	GOUV
plus bas	+3	Thématique	Segment répété	GOUV
première ministre	+3	Thématique	Segment répété	GOUV
proportionnel au revenu	+3	Thématique	Segment répété	GOUV
régime des prêts	+3	Thématique	Segment répété	GOUV
manifestant	+3	Thématique	Lemme	GOUV
mécanisme	+3	Thématique	Lemme	GOUV
menacer	+3	Thématique	Lemme	GOUV
micro	+3	Thématique	Lemme	GOUV
bourse	+3	Thématique	Lemme	GOUV
journaliste	+3	Thématique	Lemme	GOUV
parental	+3	Thématique	Lemme	GOUV
appui	+3	Thématique	Lemme	GOUV
militant	+3	Thématique	Lemme	GOUV
péréquation	+3	Thématique	Mots	GOUV
piste	+3	Thématique	Lemme	GOUV
affiche	+3	Thématique	Mots	GOUV
Pellerin	+3	Thématique	Mots	GOUV
salope	+3	Thématique	Mots	GOUV
camper	+3	Thématique	Lemme	GOUV
libéraux	+3	Thématique	Mots	GOUV
chiffre	+3	Thématique	Lemme	GOUV
annulation	+3	Thématique	Lemme	GOUV
chronique	+3	Thématique	Mots	GOUV
fermeté	+3	Thématique	Mots	GOUV
candidats	+3	Thématique	Mots	GOUV
atténuer	+3	Thématique	Lemme	GOUV
propositions	+3	Thématique	Mots	GOUV
compromis	+3	Thématique	Mots	GOUV
accessibilité	+2	Thématique	Lemme	GOUV
démocratie	+2	Thématique	Lemme	GOUV
leader	+2	Thématique	Lemme	GOUV
bord	+2	Thématique	Lemme	GOUV
bacon	+2	Thématique	Mots	GOUV
boutonnière	+2	Thématique	Mots	GOUV
défendu	+2	Thématique	Mots	GOUV
proposées	+2	Thématique	Mots	GOUV
test	+2	Thématique	Mots	GOUV
nazi	+2	Thématique	Lemme	GOUV
photographe	+2	Thématique	Lemme	GOUV
sexiste	+2	Thématique	Lemme	GOUV
perturbation	+2	Thématique	Lemme	GOUV
camarades	+2	Thématique	Mots	GOUV
Hitler	+2	Thématique	Mots	GOUV
Ouest	+2	Thématique	Mots	GOUV
appuyer	+2	Thématique	Lemme	GOUV
salaire	+2	Thématique	Mots	GOUV
position	+2	Thématique	Lemme	GOUV

entrée	+2	Thématique	Lemme	GOUV
familial	+2	Thématique	Mots	GOUV
sympathie	+2	Thématique	Lemme	GOUV
paient	+2	Thématique	Mots	GOUV
indexation	+2	Thématique	Mots	GOUV
rattrapage	+2	Thématique	Lemme	GOUV
$	+2	Thématique	Mots	GOUV
contribuables	+2	Thématique	Mots	GOUV
pression	+2	Thématique	Mots	GOUV
CAQ	+2	Thématique	Mots	GOUV
répéter	+2	Thématique	Lemme	GOUV
fardeau	+2	Thématique	Mots	GOUV
camp	+2	Thématique	Lemme	GOUV
anarchie	+2	Thématique	Mots	GOUV
descendre	+2	Thématique	Mots	GOUV
Brin	+2	Thématique	Mots	GOUV
concrets	+2	Thématique	Mots	GOUV
faction	+2	Thématique	Mots	GOUV
fascisme	+2	Thématique	Mots	GOUV
bêtise	+2	Thématique	Lemme	GOUV
comédien	+2	Thématique	Lemme	GOUV
disproportionner	+2	Thématique	Lemme	GOUV
dividende	+2	Thématique	Lemme	GOUV
empresser	+2	Thématique	Lemme	GOUV
maman	+2	Thématique	Lemme	GOUV
partisan	+2	Thématique	Lemme	GOUV
autre côté	+2	Thématique	Segment répété	GOUV
deux parties	+2	Thématique	Segment répété	GOUV
droit de manifester	+2	Thématique	Segment répété	GOUV
majorité des étudiants	+2	Thématique	Segment répété	GOUV
ont	+5	Dialogique	Mots	GOUV
étudiants qui	+5	Dialogique	Segment répété	GOUV
ils n	+5	Dialogique	Segment répété	GOUV
on	+5	Dialogique	Lemme	GOUV
Les	+5	Dialogique	Mots	GOUV
leur	+4	Dialogique	Lemme	GOUV
et ceux	+4	Dialogique	Segment répété	GOUV
Il y a	+4	Dialogique	Segment répété	GOUV
leurs cours	+4	Dialogique	Segment répété	GOUV
leurs membres	+4	Dialogique	Segment répété	GOUV
On le	+4	Dialogique	Segment répété	GOUV
qui ne sont	+4	Dialogique	Segment répété	GOUV
ils	+4	Dialogique	Mots	GOUV
veulent	+3	Dialogique	Mots	GOUV
quelqu un	+3	Dialogique	Lemme	GOUV
avaient pas	+3	Dialogique	Segment répété	GOUV
ces derniers	+3	Dialogique	Segment répété	GOUV
dans le cas	+3	Dialogique	Segment répété	GOUV
dans leur	+3	Dialogique	Segment répété	GOUV
et on	+3	Dialogique	Segment répété	GOUV
il devrait	+3	Dialogique	Segment répété	GOUV
il est	+3	Dialogique	Segment répété	GOUV
Il ne faut	+3	Dialogique	Segment répété	GOUV
ils ont	+3	Dialogique	Segment répété	GOUV
ils sont	+3	Dialogique	Segment répété	GOUV

je pense	+3	Dialogique	Segment répété	GOUV
ne paient	+3	Dialogique	Segment répété	GOUV
On a	+3	Dialogique	Segment répété	GOUV
on a	+3	Dialogique	Segment répété	GOUV
on est	+3	Dialogique	Segment répété	GOUV
on fait	+3	Dialogique	Segment répété	GOUV
On n	+3	Dialogique	Segment répété	GOUV
ont pas	+3	Dialogique	Segment répété	GOUV
semble que	+3	Dialogique	Segment répété	GOUV
si les	+3	Dialogique	Segment répété	GOUV
son gouvernement	+3	Dialogique	Segment répété	GOUV
suis sûr	+3	Dialogique	Segment répété	GOUV
ceux	+3	Dialogique	Mots	GOUV
gens	+3	Dialogique	Mots	GOUV
serait	+3	Dialogique	Mots	GOUV
sent	+3	Dialogique	Mots	GOUV
allez	+3	Dialogique	Mots	GOUV
tu	+2	Dialogique	Lemme	GOUV
normalement	+2	Dialogique	Lemme	GOUV
eux-mêmes	+2	Dialogique	Mots	GOUV
seraient	+2	Dialogique	Mots	GOUV
blogue	+2	Dialogique	Lemme	GOUV
de leurs	+2	Dialogique	Segment répété	GOUV
est que	+2	Dialogique	Segment répété	GOUV
les étudiants eux-mêmes	+2	Dialogique	Segment répété	GOUV
ne sont	+2	Dialogique	Segment répété	GOUV
par les étudiants	+2	Dialogique	Segment répété	GOUV
se sont	+2	Dialogique	Segment répété	GOUV
Mais	+11	Dialectique	Mots	GOUV
ne	+8	Dialectique	Lemme	GOUV
pas	+6	Dialectique	Lemme	GOUV
…	+5	Dialectique	Mots	GOUV
Mais elle	+4	Dialectique	Segment répété	GOUV
même si	+4	Dialectique	Segment répété	GOUV
Parce que	+4	Dialectique	Segment répété	GOUV
pendant la	+4	Dialectique	Segment répété	GOUV
au cours des	+3	Dialectique	Segment répété	GOUV
au terme	+3	Dialectique	Segment répété	GOUV
cela ne	+3	Dialectique	Segment répété	GOUV
cette année	+3	Dialectique	Segment répété	GOUV
comme ça	+3	Dialectique	Segment répété	GOUV
dans ce dossier	+3	Dialectique	Segment répété	GOUV
derniers jours	+3	Dialectique	Segment répété	GOUV
est pas parce	+3	Dialectique	Segment répété	GOUV
Mais la	+3	Dialectique	Segment répété	GOUV
Mais les	+3	Dialectique	Segment répété	GOUV
par ailleurs	+3	Dialectique	Segment répété	GOUV
par contre	+3	Dialectique	Segment répété	GOUV
par la suite	+3	Dialectique	Segment répété	GOUV
pas besoin	+3	Dialectique	Segment répété	GOUV
Quand on	+3	Dialectique	Segment répété	GOUV
si on	+3	Dialectique	Segment répété	GOUV
Sur le	+3	Dialectique	Segment répété	GOUV
parce	+3	Dialectique	Lemme	GOUV
TOUS	+3	Dialectique	Mots	GOUV

voulez	+2	Dialectique	Mots	GOUV
-ils	+2	Dialectique	Mots	GOUV
--	+2	Dialectique	Mots	GOUV
?	+2	Dialectique	Mots	GOUV
Bref	+2	Dialectique	Mots	GOUV
hier	+2	Dialectique	Mots	GOUV
là-dessus	+2	Dialectique	Mots	GOUV
très	+2	Dialectique	Lemme	GOUV
et donc	+2	Dialectique	Segment répété	GOUV
si elle	+2	Dialectique	Segment répété	GOUV
si je	+2	Dialectique	Segment répété	GOUV
si vous	+2	Dialectique	Segment répété	GOUV
y a quelques	+2	Dialectique	Segment répété	GOUV

Glossaire

Actualisation / actualiser : « opération interprétative permettant d'identifier ou de construire un sème en contexte » (Rastier, 1994, p. 221).

Classifieur : traduit de l'anglais *classifier*. Les classifieurs sont des modèles informatiques basés sur des algorithmes d'apprentissage supervisée et leur rôle est d'inférer la classe des objets (par exemple, des documents) à partir d'exemples fournis préalablement. Le classifieur effectue la classification automatique de textes.

Critère textuel : donnée textuelle sélectionnée par la méthode textométrique, le critère textuel constitue une variable de description des textes numériques. Dans la méthode de fouille d'opinions préconisée dans cette étude, les textes numériques sont représentés par un ensemble de critère textuel, c'est-à-dire, par un ensemble de données textuelles sélectionnées. Nous réservons le terme critère textuel pour parler de la donnée textuelle retenue après une analyse textométrique du corpus.

Donnée textuelle : donnée extraite de la chaîne textuelle aux fins d'analyse informatique. Nous utilisons « donnée textuelle » ou la forme plus courte « donnée » de manière interchangeable. Les formes graphiques isolés, les segments répétés et les cooccurrents sont considérés comme des données textuelles.

Fond sémantique : se dit d'un texte ou d'un corpus. Se réfère à l'ensemble d'isotopies génériques actualisées dans un texte ou dans un corpus et qui confère à ces derniers leur thème générique. Par exemple, l'isotopie /navigation/ dans « L'amiral ordonna de carguer les voiles » (Hébert, 2001) fonde aussi sont fond sémantique ou thème générique (le texte parle de navigation).

Forme graphique : suite de caractères bornée par des espaces ou des marques de ponctuation (Lebart et Salem ,1994 ; Polguère, 2008). Dans le cadre de ce travail, une donnée textuelle peut être formé par plus d'une forme graphique (par exemple, « par ailleurs »).

Forme sémantique : ensemble de molécules sémiques présentes dans un texte ou dans un corpus.

Isotopie sémantique : effet de la récurrence d'un même sème sur un empan donné, soit sur une phrase, un paragraphe, un texte ou un corpus. (Hébert, 2001). « Ex. : dans « Je bois

de l'eau », 'bois' et 'eau' contiennent le sème inhérent /liquide/. On représente les isotopies sémantiques, comme les sèmes qui les définissent, à l'aide de barres obliques (/isotopie/). (Hébert, 2001, p. 212). Il existe deux types d'isotopie selon le statut des sèmes qui les constituent : l'isotopie générique, formée par des sèmes génériques et l'isotopie spécifique (ou molécule sémique) formée par des sèmes spécifiques.

Lexique : ensemble d'unités significatives d'une langue. Dans la notion de lexique, nous excluons les mots grammaticaux : pronoms, déterminants, conjonctions, prépositions et verbes auxiliaires.

Locution : « (…) groupe de mots (nominal, verbal, adverbial) dont la syntaxe particulière donne à ces groupes le caractère d'expression figée et qui correspondent à des mots uniques. Ainsi, *faire grâce* est une loctuion verbale (ou verbe composé) correspondant à *gracier ; mettre le feu* est une locution verbale (ou verbe composé) équivalant à *allumer ; en vain* est une locution adverbiale (ou adverbe composé) correspondant à *vainement.* » (Dubois et coll., 2007)

Molécule sémique : récurrence groupée d'un ensemble de sèmes spécifiques. Défini comme un « groupement stable de sèmes, non nécessairement lexicalisé, ou dont la lexicalisation peut varier » (Hébert, 2001). Par exemple, dans *L'Assommoir* de Émile Zola, une molécule sémique qui regroupe les sèmes /jaune/, /chaud/, /visqueux/ et /néfaste/ est lexicalisée dans les mots 'alcool', 'sauce', 'morve', 'huile', 'pipi' (Hébert, 2001).

Morphème : « signe minimal, indécomposable dans un état synchronique donné » (Rastier et coll. 1994, p. 223). Par exemple, le mot 'rétropropulseurs' compte cinq morphèmes : rétro -, pro-, puls —, -eur, — s.

Mot : unité appartenant au lexique. Le terme mot est employé ici dans un sens plus général que celui de donnée textuelle ou forme graphique. Il est défini en termes saussurien, comme un signe linguistique constitué d'une forme acoustique qu'on appelle signifiant et d'un ensemble de propriétés sémantiques qu'on appelle « signifié ». Pour mentionner les mots dans notre texte, nous adoptons comme convention d'écriture les guillemets simples (par exemple : le mot 'pomme').

Mot grammatical : Mots liés à la grammaire en fonction de leur sens et leur comportement en langue (Polguère, 2008). Il s'agit de mots appartenant à une partie de discours de classe

lexicale fermée. Selon Polguère (2008, p. 101), « une partie de discours est une classe lexicale fermée (…) si l'ensemble des éléments qui la composent est stable ». Nous considérons les pronoms, les déterminants, les conjonctions et les prépositions.

Mot lexical : Mots liés au lexique d'une langue. Il s'agit de mots appartenant à une partie de discours de classe lexicale ouverte. Selon Polguère (2008, p. 100), « une partie de discours est une classe lexicale ouverte si l'ensemble des éléments qui la composent peut varier sans que cela entraîne une modification importante du comportement de la langue ». Ces parties du discours sont le verbe, le nom, l'adjectif et l'adverbe.

Mot subjectif : mot utilisé pour porter une appréciation ou une impression envers un objet. Un mot subjectif peut être mélioratif ou péjoratif. Les mots subjectifs mélioratifs accordent une appréciation positive et favorable, tels que 'beau', 'gracieux', 'bon', etc. Les mots subjectifs péjoratifs accordent une appréciation négative et défavorable par exemple, 'désagréable', 'triste', 'laid', etc.

Occurrence : se réfère à l'occurrence d'une donnée textuelle ou d'une forme dans le texte. Le nombre d'occurrences d'une donnée textuelle dans le corpus ou dans un texte est sa fréquence totale.

Sème : nous définissons les sèmes comme un trait sémantique ressortant de la confrontation de plusieurs mots dans un texte. En considérant le mot comme un signe linguistique doté d'un signifiant et d'un signifié, le sème serait la propriété sémantique du mot activée dans un contexte précis, en fonction de la présence d'autres mots qui actualisent le même sème. Les sèmes « sont d'ordre métalinguistique et résultent d'une analyse ou d'une validation humaine effectuée, par exemple, par un linguiste » (Valette, 2010, p. 26). Comme convention d'écriture, nous allons représenter les sèmes, ainsi que les isotopies que le forment, à l'intérieur de deux barres obliques (par exemple : /faiblesse/). À de fins de stylistique nous utilisons le terme 'trait sémantique' comme synonyme de 'sème'.

Sème générique : marque l'appartenance d'un sème à une classe sémantique plus générique. Par exemple, 'couteau', 'cuillère' et 'fourchette' ont en commun le sème générique /couvert/ (Hébert, 2001).

Sème spécifique : sème qui oppose un mot à d'autres au sein de la classe sémantique qu'il appartient. Par exemple, 'homme' et 'femme' ont en commun le sème générique /être

humain/, qui marquent leur appartenance dans cette classe, et se distinguent par le sème spécifique /sexe masculin/ et /sexe féminin/, qui les opposent à l'intérieur de la classe (Hébert, 2001).

Signe : signe linguistique, tel que défini par Ferdinand de Saussure, constitué de l'association entre un contenu, appelé signifié, et une forme ou image acoustique, appelée signifiant. Nous considérons que la séparation entre le signifié et le signifiant n'est qu'une abstraction puisqu'un n'a pas d'existence sans l'autre (Mounin, 1993 ; Polguère, 2008).

Signifiant : « dans la théorie saussurienne du signe, le signifiant est la forme concrète, perceptible à l'oreille (l'image acoustique) qui revoie à un concept, le signifié. C'est donc un son ou une suite de sons, qui peuvent être représentés secondairement par des symboles graphiques » (Mounin, 1993, p. 300)

Signifié : composante d'un signe saussurien, à laquelle renvoie le signifiant. « Il s'agit d'un concept résumé de l'intension (ou compréhension) de la classe d'objets évoquée par le signifiant » (Mounin, 1993, p.301).

Syntagme : suite de formes qui sont connectées directement ou indirectement par des relations syntaxiques (Polguère, 2008).

Sujet parlant : Se dit du locuteur. Distinct d'énonciateur (représentation que le sujet parlant se fait de soit même dans le texte)

Palier : « degré de complexité. Les principaux paliers sont le morphème, le syntagme, la période, et le texte » (Hébert, 2001, p. 223).

Parcours interprétatif : « suite d'opérations permettant d'assigner un ou plusieurs sens à une suite linguistique » (Hébert, 2001, p. 223). L'opération de base d'un parcours interprétatif, réalisé par un interprétant, consiste à d'identifier les relations d'identité et d'altérité qui sont produites par l'effet de la récurrence des sèmes dans les différents paliers de complexité. Les relations que l'interprétant identifie contraignent le sens de ce qu'il lit et interprète. Ce concept postule que personne ne peut pas tout dire à propos d'un texte. L'effet de l'interaction sémique impose des limites sur le sens et autorisant ce qu'il est plausible de dire (le sens ou le sens que l'on peut assigner à une suite linguistique).

SRAP : Systèmes de recommandation et d'articles de presse. Il s'agit d'un type particulier de système de recommandation, utilisés pour la recommandation d'articles de presse.

Texte : « suite linguistique empirique attestée, produite dans une pratique sociale déterminée, et fixée sur un support quelconque » (Rastier et coll., 1994, p. 168)

Textométrie : « L'analyse statistique des données textuelles (ADT), ou Textométrie, est un ensemble particulier de pratiques relevant du champ général de la linguistique de corpus. Elle comprend des traitements statistiques (analyse factorielle des correspondances, spécificités fondées sur le modèle hypergéométrique, etc.) et des outils de visualisation des corpus (nuages de mots, histogrammes, etc.) et documentaires (concordanciers) destinés à l'aide à l'interprétation des textes » (Eensoo et Valette, 2015, p.3).

Thème générique : « fond sémantique constitué par la récurrence d'un ou plusieurs sèmes génériques. Les thèmes génériques déterminent ainsi le « sujet » du texte (ou la topique en linguistique) en induisant par des faisceaux d'isotopies les impressions référentielles dominantes. (Gérard, 2004, p. 301).

Thème spécifique : « molécule sémique relevant de la composante thématique (vs dialogique, dialectique). » (Gérard, 2004, p. 301).

Unité de contexte : on appelle une unité de contexte le résultat du découpage opéré par les logiciels textométriques. Un texte peut être découpé par exemple dans un ensemble de phrase ou paragraphe. Le texte, en tant que contenu d'un document, constitue également une unité de contexte, même si aucun découpage n'a été effectué.

Unité morphosyntaxique : unités relevant de la syntaxe des textes. Il peut s'agir de mots simples (par exemple, pronoms et adverbes) ou de locutions verbales et adverbiales.

Vocabulaire : ensemble de mots simples du corpus.