Applied Semiotics / Sémiotique appliquée, nº 26 (2018)

VERS UNE SÉMIOTIQUE COMPUTATIONNELLE :

ÉTUDE DE CAS ET PREMIÈRES EXPLORATIONS

Davide Pulizzotto, Jean-François Chartier, Université de Montréal
Jean-Guy Meunier, Louis Chartand, Francis Lareau, Université du Québec à Montréal
Louis Hébert, Université du Québec à Rimouski
davide.pulizzotto@gmail.com

Résumé
À l'ère des big data et de l'informatique omniprésente, quels sont les outils informatiques utilisables dans un contexte d'analyse de texte ? Comment peuvent être utilisées les théories et les méthodes de la sémiotique dans une telle analyse ? Le présent article propose une chaîne de traitement pour l'exploration d'un grand corpus journalistique. En particulier, cette chaîne permet l'identification d'un certain nombre de macrostructures qui caractérisent le corpus analysé. La contribution principale de ce travail consiste dans le transfert de connaissances de l'intelligence artificielle aux sciences humaines.

Introduction

Au printemps 2012, plus de 120 000 étudiantes et étudiants québécois ont commencé une grève qui a duré plusieurs mois et qui a initié un débat public sur le statut de l'éducation au Québec. Le *printemps érable* a fait l'objet d'une couverture médiatique complète et ce, dans plusieurs types de médias. La presse écrite a été l'un des principaux moyens d'expression pour les acteurs impliqués dans le débat, constituant ainsi une composante importante pour la construction de l'*opinion publique*.

Trois grands journaux québécois, *La presse*, *Le devoir* et *Le soleil*, constituent l'échantillon à partir duquel est menée notre analyse de l'opinion publique sur le *printemps érable*. Nous construisons un corpus de très grande dimension afin d'établir une base très exhaustive et nous essayons de répondre à deux types de questions : cognitives et méthodologiques. Les premières conduisent à des informations sur le phénomène sociopolitique et médiatique, à savoir les événements les plus importants ou les thématiques les plus récurrentes. Les deuxièmes mènent à une réflexion sur l'utilisation des outils de l'intelligence artificielle (AI) dans une analyse sémiotique.

Pour répondre à ce genre de questions, nous utilisons une méthode d'analyse de texte assistée par ordinateur et des outils d'analyse sémiotique du courant structuraliste. Nous définissons d'abord les concepts de *macrostructure* et de *schéma récurrent* comme étant des

éléments de révélation du contenu d'un texte. Nous abordons ces concepts à partir d'une théorie narrative de génération du sens. Par la suite, nous identifions une combinaison de modèles venant de l'AI pour la découverte des schémas récurrents, lesquels seront analysés au moyen d'un modèle actanciel inspiré de Greimas.

L'article est divisé en quatre sections. La première illustre la question de recherche et l'approche utilisée. La deuxième expose en détail la méthode d'analyse de texte assistée par ordinateur qui a été employée. La troisième précise la manière par laquelle les résultats ont été annotés et interprétés. Enfin, la quatrième aborde les interactions possibles entre sémiotique et AI.

1. Question de recherche

L'objet de cette étude est le traitement journalistique de la crise étudiante de 2012. Son objectif principal est d'identifier, en explorant un grand corpus, les éléments les plus généraux du contenu qui ont caractérisé ce traitement de l'information et de faire ceci avec l'aide d'une *méthode hybride* (entre sémiotique et AI). L'étude de la couverture médiatique du *printemps érable* au moyen d'une approche sémiotique peut se réaliser de plusieurs façons. Par exemple, la découverte des éléments thématiques composant la couverture médiatique de ce phénomène sociopolitique constitue l'une des approches possibles. La présente recherche se base plutôt sur l'étude du *niveau macrostructurel* des textes, en excluant l'analyse des microstructures de l'organisation du sens, sans s'arrêter à l'analyse thématique. Ainsi, en explorant un corpus journalistique comme celui du *printemps érable*, nous souhaitons connaître avant tout qui en sont les principaux acteurs et quels sont leurs enjeux majeurs. Une fois obtenue une description générale du corpus, nous pourrons détailler l'un ou l'autre des aspects déjà identifiés par cette première exploration. Généralement, les corpus utilisés pour ce genre de tâches sont très grands, ce qui cause plusieurs problèmes de traitement et d'analyse aux chercheurs en sciences humaines. Pour cette raison, l'assistance informatique se révèle une nécessité.

Ainsi, deux typologies de questions surgissent. D'une part, une typologie concerne la *dimension cognitive* de la recherche : quelles sont les caractéristiques principales des articles de presse qui ont traité du printemps érable ? D'autre part, cette recherche considère des questions d'ordre *méthodologique* : comment pouvons-nous découvrir de manière empirique et par assistance informatique ces caractéristiques ? Ces questionnements sont complémentaires aux objectifs de la recherche. L'exploration du corpus et son analyse seront effectuées en tentant de *regrouper de manière empirique et informatisée les articles de presse sur la base d'éléments macrostructurels qui caractérisent le niveau global de la sémantique du corpus*.

1.1. Hypothèses

L'élaboration des conditions initiales de notre raisonnement suit le double parcours décrit précédemment. D'une part, les hypothèses sémiotiques préparent le terrain pour l'*analyse des informations extraites du corpus*. D'autre part, les hypothèses méthodologiques nous permettent de *transférer une procédure de découverte particulière vers un modèle formel et computable* et d'identifier une série d'outils informatiques utilisables dans le cadre d'une analyse de texte assistée par ordinateur.

1.1.1 Hypothèses sémiotiques

Pour identifier les caractéristiques générales d'un grand corpus de presse, nous postulons d'abord l'existence, dans n'importe quel corpus, de *macrostructures* qui soutiennent les textes. Ce postulat se base sur les travaux du linguiste van Dijk, lequel définit la macrostructure comme une « structure de signification globale d'un texte » (Kintsch et Van Dijk, 1975: 101). Une macrostructure se forme dans un texte à partir de sa *cohérence sémantique*, ce qui détermine ses conditions d'existence. Ceci implique qu'une analyse sémantique d'un texte mène à l'identification d'une ou plusieurs macrostructures. Ensuite, nous émettons l'hypothèse qu'*il est possible de construire des groupes d'articles distincts qui dépendent de macrostructures différentes. Ceci est possible par la détection des schémas récurrents*. Ces derniers mettent en forme des *régularités* qui sont communes aux textes du corpus et qui sont déterminantes pour la stabilisation du message transmis. Ce sont les schémas récurrents qui véhiculent des macrostructures et qui permettent de réunir les articles similaires dans un même groupe.

Les macrostructures se situent à un niveau plutôt général de l'organisation du sens. Or, une des propriétés fondamentales de l'organisation du sens est la *narrativité*. La compétence narrative n'est pas seulement liée à la capacité de raconter des histoires, mais aussi à la création du monde et de la réalité (Bruner, 1991 ; Herman, 2009). Elle détermine donc une partie de la vie cognitive humaine (Young et Saver, 2001). À partir de ce postulat sémiocognitif, nous émettons la sous-hypothèse suivante : *les macrostructures qui déterminent le plan du contenu à un niveau plus global et général s'identifient par des schémas récurrents de type narratif*. Ceci implique que les schémas récurrents qui véhiculent les macrostructures ont des propriétés narratives et qu'ils peuvent donc être identifiés par des outils sémiotiques classiques. Les relations entre macrostructures et narrativité ont déjà été formalisées par van Dijk (Van Dijk, 1977).

Le *schéma actanciel* est le modèle choisi pour l'extraction des informations des schémas récurrents. Il se définit comme « un dispositif permettant, en principe, d'analyser toute action réelle ou thématisée » (Hébert, 2016). Ce modèle formel et général permet surtout d'extraire des éléments sémantiques à partir du rôle qu'ils jouent dans l'économie générale du texte analysé. L'analyse au moyen de ce dispositif consiste donc à classer les éléments de l'action dans les catégories déterminées par le modèle. Les composantes du modèle sont au nombre de six et elles se déploient sur trois axes : l'*axe du vouloir*, l'*axe du*

pouvoir et l'*axe de la transmission*. Dans notre cas, nous recourrons à un modèle légèrement plus précis (Hébert, 2016: 133) qui tient compte d'autres sous-catégories actantielles : *actant*/*négactant*, *actant possible*/*factuel*, *actant actif*/*passif*.

1.1.2 Hypothèses méthodologiques

Des méthodes d'assistance par ordinateur sont requises dans le cadre de cette recherche en raison des grandes dimensions du corpus analysé. L'utilisation de ces outils dépend de la possibilité d'exécuter certaines fonctions cognitives impliquées dans l'analyse de texte au moyen d'un dispositif informatique. Ceci est réalisable lorsqu'une fonction cognitive est formalisée dans un modèle computationnel (Meunier, 2017a). Donc, si une tâche peut être représentée formellement, il peut alors exister une fonction computable qui, une fois implémentée dans une machine, l'exécute (Meunier, 2017b). Ceci nous mène à la définition de l'hypothèse méthodologique générale : *la tâche de découverte des schémas récurrents peut être modélisée* dans un contexte computable et, plus particulièrement, *au moyen d'une combinaison de modèles* issus de la *sémantique vectorielle* et de l'*apprentissage automatique*.

Plus spécifiquement, nous devons *transposer* le corpus dans un format sur lequel il sera possible d'exécuter des calculs et des procédés algorithmiques (Weiss et coll., 2010). À cet égard, la *représentation vectorielle* (Salton et coll., 1975) de chacun de textes contenus dans le corpus permet d'obtenir une structure de données textuelles sur laquelle il est possible d'opérer la détection des schémas récurrents. L'utilisation d'un *modèle syntagmatique* et d'une *mesure de pondération* spécifique (Tf-Idf) accroit la possibilité de découvrir d'éléments sémantiques *distinctifs* et *spécifiques* à des groupes d'articles.

Par la suite, une fois obtenu un modèle vectoriel syntagmatique du corpus, une technique de *clustering* est utilisée pour identifier des groupes d'articles et pour découvrir des schémas récurrents. En effet, ce genre d'outils se base sur le calcul de similarité entre vecteurs et est surtout utilisé pour la reconnaissance des *formes spécifiques* à certains groupes de vecteurs (Aggarwal et Han, 2014). Le calcul de similarité ainsi effectué nous semble suffisant pour la détection des schémas récurrents spécifiques d'un corpus et pour l'association des articles à chacun d'eux.

2. Méthodologie

Les théories et les méthodes relatives à l'étude du discours journalistique sont abondantes. En général, les médias et l'article de presse, plus particulièrement, soulèvent des problèmes de différentes natures, ce qui exige une approche multidisciplinaire. La sémiotique est ainsi appelée à combler l'absence d'une science unique et articulée pour l'étude des médias. Celle-ci étant considérée comme une « hyperscience globale et intégrante » (Dobre, 2013 : 1), se prête bien à l'analyse de tout objet sémiotique au moyen de théories et méthodes

multidisciplinaires. Notre recherche propose une méthode hybride pour l'analyse du texte journalistique, composée d'outils issus de la sémiotique et de l'AI, d'une manière qui, à notre connaissance, n'a pas encore été considérée.

2.1. Le corpus

Le corpus de travail que nous avons choisi d'utiliser est composé de 2 897 articles de presse provenant des quotidiens *La Presse*, *Le Soleil* et *Le Devoir*. La sélection des articles a été réalisée à l'aide d'une procédure de recherche par mots-clés sur la base de données *Biblio branchée* et le service *Eureka*. Les termes « étudiant, étudiants, étudiante et étudiantes » ont été retenus comme mots-clés pour restreindre la recherche des articles. Ce choix est fondé sur le *critère de saillance* (Flament et Rouquette, 2003), selon lequel des indicateurs lexicaux peuvent servir d'« ancrage empirique » au phénomène exploré. De plus, la recherche a été limitée à une période précise, soit du 15 février 2012 au 9 juin 2012, ce qui couvre le début du traitement médiatique, avec les premiers votes de grève, jusqu'à une semaine après la rupture des négociations entre les associations étudiantes et le gouvernement. Ceci a permis de cibler la période la plus importante de la grève, car la rupture des négociations a constitué une impasse qui s'est poursuivie jusqu'au retour en classe. Tous les articles des trois journaux respectant ces critères ont été inclus dans le corpus de travail.

Les éléments retenus de chaque article sont les suivants : *date de publication*, *journal de publication*, *auteur*, *titre*, *corps de l'article*. Les images ne sont pas présentes, ni la typologie d'article ou la position de l'article dans l'ensemble du quotidien. Les données réellement traitées lors de l'analyse sont le titre et le corps de l'article. Les articles se répartissent ainsi : 1 159 pour *La presse*, 906 pour *Le soleil* et 832 pour *Le devoir*. La très grande majorité d'entre eux est signée par un ou plusieurs journalistes (662 signataires différents) ; environ 8,5 % des articles ne portent pas de signature.

2.2 La représentation vectorielle du corpus

Cette étape est constituée par la transformation du corpus dans une matrice *U* et elle est effectuée après une étape de prétraitement des textes (Vijayarani et coll., 2015), nécessaire pour l'extraction des caractéristiques de chaque article de journal. Une des étapes les plus importantes est la racinisation des mots. Le modèle créé est représenté par le tableau 1. Chaque ligne de cette matrice modélise un article du corpus sous la forme d'un vecteur $\vec{s}_i = (v_{i1}, \dots, v_{ij})$ où v_{ij} correspond à la *valeur de pondération* de la $j^{ème}$ racine dans le $i^{ème}$ segment.

Tableau 1 : Matrice article-racine

		$racine_1$	$racine_2$	…	$racine_j$
$U =$	$article_1$	V_{11}	V_{12}	…	V_{ij}
	$article_2$	V_{21}	V_{22}	…	V_{ij}
	⋮	⋮	⋮	⋱	⋮
	$article_i$	V_{i1}	V_{i2}	…	V_{ij}

La représentation vectorielle du contenu sémantique des textes dépend surtout des valeurs qui remplissent la matrice, c'est-à-dire de la *pondération*, qui assigne un poids différent aux mots de chaque article. Il existe différentes méthodes de pondération (Salton, 1971). Nous avons choisi la pondération *Tf-Idf* (*Term Frequency-Inverse Document Frequency*). Cette fonction modélise le concept de discrimination et de spécificité sémantique et lexicale. La fonction du *Tf-Idf* augmente la valeur de manière proportionnelle au nombre de fois qu'un mot apparaît dans un document, mais de manière inversement proportionnelle à la fréquence du mot à l'intérieur du corpus. Ceci implique que, plus un mot est utilisé dans un petit nombre d'articles, plus sa valeur *Tf-Idf* sera élevée. Au contraire, plus un mot est utilisé de manière diffuse et relativement homogène dans les textes, plus le *Tf-Idf* sera faible. Pour ces raisons, le *Tf-Idf* est considéré comme une *fonction pour la découverte d'éléments de type lexicosémantique qui ont une valeur discriminatoire et distinctive*. À la fin de ce traitement, le corpus textuel est converti dans une matrice *Documents-Mots*, contenant 2 987 articles pour 18 576 racines des différents mots.

Les hypothèses et postulats linguistiques en cause lors de la transformation d'un corpus dans une matrice ainsi construite concordent avec ceux énoncés par Saussure lors de la définition du concept de *valeur* : « dans la langue, il n'y a que des différences » (Saussure, 1995). Dans ce cadre, le signe linguistique se met « en condition de signifier » seulement à l'intérieur d'un *système discret*, et non analogique, composé d'éléments qui s'opposent entre eux. Cette conception de la signification recoupe les hypothèses de la sémantique vectorielle (Sahlgren, 2008 : 6) et des modèles qui utilisent le *Tf-Idf* comme fonction de pondération, car ce qui se révèle important est la *distinction*, c'est-à-dire les éléments spécifiques et discriminatoires qui se manifestent par des dynamiques d'opposition.

2.3 Clustering

La *classification non supervisée* (*clustering*) de données non structurées et la reconnaissance de leurs formes récurrentes sont utilisées dans des domaines de connaissance différents, sur plusieurs types de données et pour de multiples champs d'application. Les modèles de *clustering* peuvent donc varier en fonction du domaine, de l'application ou du type de données. Énormément de travaux ont aussi été consacrés au traitement de données

textuelles (Steinbach et coll., 2000 ; Aggarwal et Zhai, 2012). Dans cette recherche, un algorithme de type *hard clustering* est utilisé, car il est le plus adapté pour nos travaux. Cet algorithme ne génère pas de groupes ayant des relations hiérarchiques entre eux ou qui se chevauchent. L'hypothèse classificatoire est la suivante : la meilleure partition possible peut être obtenue en classant un article dans un seul et unique *cluster*, et ce, afin d'éviter des classes floues. Les avantages de ce genre d'algorithme sont principalement liés à une plus grande interprétabilité des résultats et à une facilité de traitement des données, ce qui est généralement recherché pour une première exploration d'un corpus textuel.

Notre recherche utilise l'algorithme *K-moyennes*. D'un point de vue mathématique, cet algorithme génère le nombre de *clusters k* choisi par l'utilisateur, en minimisant la moyenne de la distance euclidienne au carré entre les documents et le centre de chaque *cluster k*. De fait, cet algorithme produit plusieurs groupes d'articles qui sont proches d'un *centroïde* c_k, ce qui correspond au centre géométrique du *cluster*. À la fin de cette étape, une série de groupes d'articles (*clusters*) sont identifiés en fonction de leur similarité sémantique. Ceux-ci sont ensuite analysés afin de détecter les motifs récurrents du *cluster* et, en fonction de notre hypothèse, de leur associer des structures narratives.

L'opération de *clustering* a donc permis, par voie algorithmique et mathématique, la construction de 26 groupes d'articles sur la base de leurs similarités. Les 26 groupes ont été identifiés par une évaluation de la qualité de plus de 50 partitionnements potentiels. Dans le cadre de cette recherche, les groupes (*cluster*) de documents sont analysés afin de détecter la ou les macrostructures permettant de les rassembler et qui, selon notre sous-hypothèse, correspondent à une structure de type narratif. En d'autres termes, la régularité lexicale sur laquelle le *clustering* se base pour la détection des groupes d'articles similaires est le reflet d'une régularité sémantique qui agit à un niveau profond de l'organisation du *sens*. L'organisation de ce niveau profond est assujettie à une grammaire de type narratif. Le protocole d'analyse qui suit a comme objectif de vérifier la possibilité d'interpréter les résultats du *clustering* en suivant cette sous-hypothèse.

2.4 Le protocole d'analyse des clusters

L'analyse des *clusters* est effectuée à partir d'un protocole expérimental qui vise à reconstruire les schémas récurrents de chaque groupe d'articles. Deux étapes sont réalisées, soit la sélection d'un échantillon représentatif pour chaque groupe et l'annotation manuelle de cet échantillon selon un schéma actanciel.

2.4.1 La sélection d'un échantillon représentatif

La chaîne de traitement développée permet de sélectionner un échantillon représentatif d'articles à analyser. Ce processus de sélection automatisée demeure un des éléments les plus importants de cette recherche, car il soutient l'analyse sémiotique par une méthode automatisée et guidée de construction d'un échantillon pour l'analyse à partir d'un grand corpus.

Un avantage majeur de l'utilisation des algorithmes de *clustering* est la possibilité de considérer chaque groupe d'articles dans un espace géométrique, ce qui permet d'établir un modèle heuristique simple et efficace pour la sélection automatisée des échantillons représentatifs du corpus. En effet, l'algorithme utilisé se base sur un nombre *k* de centroïdes qui sont itérativement déplacés dans l'espace géométrique dans le but d'obtenir une partition optimale des articles situés sur ce même espace. Lorsque les centroïdes ne peuvent plus être déplacés, l'algorithme s'arrête, car un équilibre s'est établi dans le système et une partition optimale de données a été obtenue. Chaque article du groupe sera alors sélectionné en fonction de sa distance avec le centroïde du *cluster*. Les documents les plus proches du centroïde forment ainsi un *centre de gravité* sur la base duquel est formé un *cluster*.

Pour sélectionner cet agglomérat d'articles, il est requis de calculer la similarité (métrique cosinus) de chaque article par rapport à chaque centroïde. Cette façon de procéder constitue une des méthodes les plus utilisées en recherche d'information (*information retrieval*). À la fin du processus, les groupes d'articles sont triés selon leur proximité avec les centroïdes. Les 20 premiers articles les plus proches du centroïde sont ainsi retenus et constituent l'échantillon représentatif du *cluster*. L'ensemble des échantillons ainsi formé représente globalement le corpus en fonction des composantes sémantiques retrouvées (*clusters*).

2.4.2 L'annotation des échantillons

L'objectif principal de cette étape est de trouver un nombre minimal de schémas actantiels représentant un nombre maximal d'articles pour chaque échantillon. Le modèle d'annotation retenu utilise des outils issus de la sémiotique greimassienne et se base sur l'hypothèse sémiotique déjà énoncée : le niveau profond du sens dans le discours journalistique s'organise de manière narrative. Sur cette base, un des modèles les plus simples de la théorie greimassienne sera employé, soit le modèle actantiel, dans sa version modifiée par Hébert. Ce modèle est donc utilisé comme ancrage théorique du protocole d'annotation et il balise l'interprétation des échantillons des *clusters*. De manière pratique, chaque article de l'échantillon est lu et annoté sur la base du modèle actantiel dans le but de révéler et de souligner les similarités structurelles avec les autres.

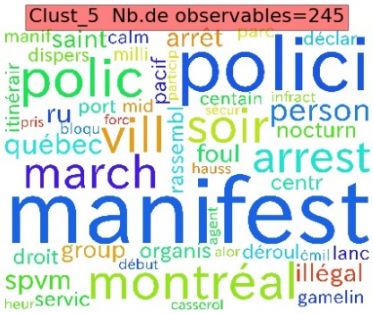
3. Résultats et interprétation

Les 26 *clusters* ont permis de déterminer une série de macrostructures. Nous présenterons une description d'un seul de ces *clusters*, ce qui constitue un exemple de l'analyse qui peut être conduite avec notre méthodologie. Les autres groupes d'articles obtenus peuvent être consultés au moyen d'une application web construite spécialement pour ce faire (https://davidepulizzotto2.shinyapps.io/printemps_erable/).

3.1 Le cluster nº 5

Le *cluster* nº 5 contient 245 articles. Les termes les plus importants sont : « manifestation », « police », « policier », « marche », « ville », « Montréal », « soir », « arrestation » et, naturellement, le mot pôle « étudiant » (figure 1). Des vingt articles composant l'échantillon, neuf ont été publiés par *Le Soleil*, six par *Le Devoir*; cinq par *La Presse*. Les journalistes les plus présents sont Matthieu Boivin (4 articles) et David Santerre (3 articles) et la période de couverture est distribuée de manière relativement homogène entre mars et juin, avec une concentration majeure pour le mois de mai (8 articles).

Figure 1 : Nuage des mots du cluster nº 5



L'axe du vouloir est représenté par un schéma récurrent dans lequel l'objet de valeur A est poursuivi par les sujets B qui s'opposent aux anti-sujets C (tableau 2). Cet axe est le plus souvent sous-entendu, mais quelquefois il est exprimé explicitement par le journaliste. Par exemple : « La police a eu besoin de plusieurs heures, jeudi soir, pour ramener le calme » (article 263). Cet axe du vouloir est le plus représentatif du groupe d'articles analysés et constitue le programme narratif α. Il s'impose aussi comme épicentre de chaque article, à partir duquel différents discours ou points de vue sont exprimés.

Tableau 2 : Axe du vouloir (schéma actantiel)

PN : α	**A**	**B**	**C**
ACTEURS OBSERVÉS	Ramener le calme, s'opposer aux actes de violence	Police et maire de Montréal	Manifestants, casseurs et militants radicaux
ACTANT	*Objet de valeur*	*Sujet*	*Anti-sujet*

PN : β	**A^{I}**	**B^{I}**	**C^{I}**
ACTEURS OBSERVÉS	Liberté de manifester	Manifestants	Police et politiciens
ACTANT	*Objet de valeur*	*Sujet*	*Anti-sujet*

Plus rarement, l'axe du vouloir est inversé, formant ainsi le plan narratif β (trois articles sur vingt). Ceci fait de l'anti-sujet du premier axe son sujet principal et du droit à manifester son objet de valeur. Cet axe est à maintes reprises exprimé par l'appel des étudiants qui dénoncent « l'attitude provocatrice de la police qui […] leur [dénie] le droit de manifester » (article 263). Cependant, dans le programme narratif β, le sujet n'est jamais représenté à travers les acteurs « casseurs » ou « militants radicaux », ce qui démontre deux visions opposées sur les manifestants entre les deux programmes narratifs. Les deux expressions « casseurs » et « militants radicaux » comportent des jugements de valeur qui ne sont pas envisagés par le programme narratif β.

Le programme narratif α, qui est le plus important et le plus récurrent de l'échantillon analysé, est constitué de plusieurs axes du pouvoir (tableau 3). L'actant D diffère des autres actants par son statut d'actant *passif*, ce qui qualifie davantage le type de dynamique en cause. *Les manifestants s'opposent aux dispositions de la police et du maire, qui veulent contrôler les trajets des manifestations, en ne communiquant pas l'itinéraire*. Pour les manifestants et certains leaders étudiants, fournir le trajet signifie « céder à la peur et à la répression » (article 1 814), ce qui brime le droit d'expression et la liberté à manifester.

Tableau 3 : Axe du pouvoir (schéma actantiel)

PN : α	**D**	**E**	**F**
ACTEURS OBSERVÉS	Manifestants ne fournissent pas le trajet	Manifestants lancent des projectiles	Casseurs fracassent des vitrines
ACTANT	*Opposant*	*Opposant*	*Opposant*
POSSIBLE/FACTUEL	*Factuel*	*Factuel*	*Factuel*
ACTANT/NÉGACTANT	*Opposant*	*Opposant*	*Opposant*
PASSIF/ACTIF	*Opposant passif*	*Opposant actif*	*Opposant actif*

PN : α	**G**	**H**	**I**
ACTEURS OBSERVÉS	Déclaration de manifestation illégale et avis de dispersion	Règlement municipal « anti-masques »	Arrestations de masse ou ciblées
ACTANT	*Adjuvant*	*Adjuvant*	*Adjuvant*
POSSIBLE/FACTUEL	*Factuel*	*Factuel*	*Factuel*
ACTANT/NÉGACTANT	*Adjuvant/Non-opposant*	*Adjuvant*	*Non-opposant*
PASSIF/ACTIF	*Adjuvant actif*	*Adjuvant actif*	*Adjuvant actif*

Tableau 4 : Axe de la transmission (schéma actantiel)

PN : α	**L**	**M**
ACTEURS OBSERVÉS	État de droit, société, politique	Société, état de droit, respect de la loi
ACTANT	*Destinateur*	*Destinataire*

PN : β	**L^I**	**M^I**
ACTEURS OBSERVÉS	Droits de l'homme	Société
ACTANT	*Destinateur*	*Destinataire*

Les actants G, H et I sont des adjuvants et ils ont une position ambigüe à l'intérieur de α, car ce sont des moyens qui mènent à des actions violentes, comme les arrestations de masse ou musclées et l'utilisation des gaz urticants et de lacrymogènes, ce qui ne correspond pas à l'objet de valeur du programme narratif. Cependant, ils constituent aussi des moyens par lesquels la police atteint son objectif immédiat : *mettre fin à une manifestation déclarée illégale et ramener le calme en ville*. Les actants E, F, G et H sont aussi les éléments qui rendent possible l'intervention de la police.

L'axe de la transmission du programme α est caractérisé par des actants généralement implicites, à savoir : *l'état de droit et la société* qui confère à la police sa mission. Le destinataire des actions de la police ne peut qu'être similaire au destinateur. Certaines déclarations des politiciens rapportées dans les articles analysés expriment plus explicitement le destinataire du programme narratif α, et ce, en prenant position pour justifier les actions de la police. Par exemple, le ministre Bachand a déclaré : « les gens ont le droit de manifester

et d'exprimer leurs droits, mais ils n'ont pas le droit d'empêcher les autres de travailler. Tout le monde a des droits dans la société et ça doit se faire avec respect. Ils ont le droit de manifester, mais de bloquer le centre-ville, ça, c'est interdit et c'est inacceptable dans une société » (article 20).

Le *cluster* nº 5 regroupe les articles sur la base d'une macrostructure principale, celle de la *lutte aux manifestations illégales et violentes*. Les actants impliqués forment une structure narrative récurrente dans laquelle un sujet principal, la police, contrarie les actes illégaux des manifestants au moyen des avis de dispersion et des arrestations. Il s'agit de l'épicentre à partir duquel différents articles traitent des *avantages d'une marche pacifique* (article 1 020), de la possibilité *d'interdire les masques* pendant les manifestations (article 263), des *conséquences économiques des manifestations violentes* (article 983), de la *légitimité des actions violentes de la police* (article 152), du risque de *perturbation des événements liés au Grand Prix de F1* (article 2 365), de la *légitimité de la loi spéciale 78* (article 1 950) ou de la simple description des événements (articles 1 913 et 2 187), etc. En dépit de la diversité des points de vue et des sujets qui sont abordés dans les différents articles, le programme narratif α, de même que le β, même si dans une moindre mesure et de manière inversée, véhiculent la macrostructure sur laquelle tous les articles de ce *cluster* sont fondés.

4. Champs de recherche pour une *sémiotique computationelle*

Le domaine de la *sémiotique computationnelle* peut être élaboré à l'intérieur de l'espace qui lie la sémiotique et l'AI. Il existe déjà des travaux qui traitent de manière plus spécifique la relation entre ces deux disciplines. En suivant la distinction faite par Tanaka-Ishii (Tanaka-Ishii, 2015), nous pouvons identifier trois grandes zones d'interaction qui mènent vers des champs de recherche différents :

1) l'étude du fonctionnement de l'ordinateur comme machine manipulant des symboles ;
2) l'étude de l'utilisation et du contrôle de l'ordinateur par l'humain ;
3) l'étude des objets sémiotiques au moyen de l'ordinateur.

4.1 Ordinateur comme machine sémiotique

Le premier champ de recherche exploite différentes théories des sciences humaines pour l'amélioration des performances d'une machine. On va de l'exploration de l'architecture computationnelle (Etxeberria et Ibáñez, 1999) à la comparaison, très critiquée, entre computation et pensée (Queiroz et Merrell, 2008 ; Fetzer, 2011 ; Rapaport, 2012). Dans ce dernier cas, les travaux de Peirce ont été repris plusieurs fois (Ketner, 1988), en raison de la relative facilité à aborder sa théorie du point de vue mathématique (Nadin, 1977). De plus,

l'interprétation sémiotique du phénomène computationnel (Meunier, 1989, 2014) s'affirme de plus en plus comme un thème central dans le domaine des humanités numériques.

4.2 Relations homme-machine

Les projets faisant partie du deuxième champ de recherche sont plus nombreux et le rôle qu'y joue la sémiotique est aussi plus clair (Nadin, 2011). Il existe trois sous-champs de recherche : 1) les interfaces homme-machine (IHM) ; 2) les systèmes d'information ; 3) les langages de programmation. Ces trois sous-champs partagent le même objectif, qui est d'améliorer le design et la conception des systèmes informatiques pour faciliter l'interaction entre l'humain et l'ordinateur.

Pour le premier sous-champ de recherche, la sémiotique offre un cadre théorique à partir duquel l'élaboration des logiciels et de chaque modèle d'interaction homme-machine peut être conçue et améliorée. Ce genre de travaux a donné naissance à l'*ingénierie sémiotique* (De Souza, 2005), qui utilise les modèles sémiotiques pour le design informatique et l'organisation fonctionnelle des IHM. Le deuxième sous-champ est rattaché à la *sémiotique organisationnelle*, qui a été inaugurée dans les années 1970 (Stamper, 1973) et qui est surtout enracinée dans les théories de Morris (*semiotic ladder*) (Liu, 2000). Le troisième domaine est strictement lié à l'étude sémiotique des langages de programmation, un projet qui remonte aux années 1960 (Zemanek, 1966) et pour lequel les travaux de Peirce et Morris représentent un point de référence important. L'intérêt envers ce genre de recherche est revenu à la surface récemment avec des études plus complètes (Tanaka-Ishii, 2010), dans lesquelles une catégorisation des modèles, des types et des systèmes de signes des langages de programmation a été proposée sur la base des deux grands courants de la sémiotique (Peirce et Saussure).

4.3 L'ordinateur comme outil pour l'analyse sémiotique

Ce dernier champ de recherche nous intéresse de plus près, car il est centré sur l'analyse et l'interprétation des objets sémiotiques. Cette interaction entre sémiotique et AI est la moins explorée (Tanaka-Ishii, 2015). Cependant, l'intérêt envers le traitement informatique de systèmes sémiotiques est un des projets de recherche les plus importants de l'AI. Ceci est démontré par la naissance de plusieurs disciplines pour lesquelles les sciences humaines commencent à développer un intérêt à la fois théorique et méthodologique : sémantique vectorielle (Salton et coll., 1975 ; Sahlgren, 2006), sémantique distributionnelle (Fabre et Lenci, 2015), apprentissage machine (Bishop, 2006), recherche d'information (Manning et coll., 2009), traitement automatique des langues (Manning et Schütze, 1999), fouille de texte (Aggarwal et Zhai, 2012) et exploration de données (Aggarwal, 2015).

Conclusion

Pour traiter une grande quantité d'information à des fins d'analyse, un support complémentaire aux outils et méthodologies sémiotiques doit être envisagé. Le domaine de l'AI fournit plusieurs outils qui peuvent être utilisés dans un contexte d'assistance à l'analyse de textes. En sémiotique, ces champs de recherche ont été toutefois très peu explorés. Ce travail a esquissé une des pistes de recherche possibles pour la sémiotique, en développant une méthode d'analyse des macrostructures d'un corpus journalistique par assistance informatique. La chaîne de traitement construite, bien que se limitant à utiliser des outils standard en informatique, permet une exploration efficace d'un grand corpus de textes, ce qui offre aussi une solution à des problèmes récurrents en sciences humaines. Ainsi, la contribution principale de cet article consiste dans le transfert de connaissances de l'AI aux sciences humaines.

Les hypothèses sur la base desquelles la présente recherche a été construite ont pu être partiellement vérifiées. Un certain nombre de groupes d'articles distincts ont pu être identifiés, et ceci en profitant des macrostructures partagées par les articles. L'échantillon d'articles de chacun de ces groupes a aussi révélé une structure narrative sous-jacente, qui a été identifiée au moyen d'un (ou plusieurs) schéma actanciel. La découverte de ces groupes et des schémas récurrents que les articles partagent a été possible grâce à une combinaison des modèles issus de la sémantique vectorielle (*modèle syntagmatique* et *pondération Tf-Idf)* et de l'apprentissage automatique (*clustering*). Cependant, il n'a pas encore été possible de calculer la marge d'erreur de la méthode, car les échantillons qui ont fait l'objet d'évaluation et le protocole d'analyse utilisé ne permettent pas d'atteindre ce genre de résultats.

De plus, l'article offre de multiples éléments de réflexion. Le changement méthodologique que l'assistance informatique comporte mène, par exemple, à des réflexions sur l'érection d'une *sémiotique computationnelle*. En particulier, les outils de *text mining* issus de l'AI et utilisés dans le cadre de cette recherche, comportent la construction d'une méthodologie *hybride*, dans laquelle une approche sémiotique d'analyse de texte rencontre des approches quantitative et informatique. L'utilisation de ce genre de méthodes implique des retombées théoriques menant à des réflexions sur la nature computationnelle des modèles sémiotiques classiques. Par exemple, le lien entre la sémantique vectorielle et la sémiotique saussurienne a été rapidement esquissé. Un parallélisme entre macrostructures et *clustering* a été plus largement discuté, ce qui pourrait être considéré comme un exemple à partir duquel relancer le projet sémiotique au sein des humanités numériques.

Notice biobliographique
Davide Pulizzotto est doctorant au programme de sémiotique de l'Université du Québec à Montréal (UQAM). Il est aussi chercheur auprès du Laboratoire d'ANalyse Cognitive de l'Information (LANCI), dans lequel il s'occupe d'analyse conceptuelle assistée par ordinateur. Il est aussi membre, avec Francis Lareau, Louis Chartrand et Jean-François Chartier, de l'équipe de recherche responsable de la conceptualisation et du développement d'une méthode de lecture

et d'analyse de textes assistées par ordinateur (LACTAO) sous la direction de Jean Guy Meunier. Ses recherches portent principalement sur la sémiotique et les humanités numériques. Il s'intéresse plus particulièrement aux croisements disciplinaires entre intelligence artificielle et sémiotique dans un contexte d'analyse de texte. Louis Hébert est professeur à l'Université du Québec à Rimouski (UQAR) et directeur de *Signo – Site Internet bilingue de théories sémiotiques* (www.signosemio.com).

Ouvrages cités

AGGARWAL, C. C. (2015), *Data Mining. The Textbook*, New York, Springer.

AGGARWAL, C. C. et HAN, J. (dir.) (2014), *Frequent Pattern Mining*, New York, Springer.

AGGARWAL, C. C. et C. ZHAI, (dir.) (2012), *Mining Text Data*, New York, Springer.

BISHOP, C. M. (2006), *Pattern Recognition and Machine Learning*, Singapour, Springer.

BRUNER, J. S. (1991), « Narrative Construction of Reality », *Critical Inquiry*, vol. 18, nº 1, p. 1-21.

DE SOUZA, C. S. (2005), *The Semiotic Engineering of Human-Computer Interaction*, Cambridge, MIT press.

DOBRE, D. (2013), *Analyse du discours de presse: projet sémiotique*, Bucarest, Editura Universităţii din Bucureşti.

ETXEBERRIA, A. et J. IBÁÑEZ (1999), « Semiotics of the Artificial: The "Self" of Self-Reproducing Systems in Cellular Automata », *Semiotica*, vol. 127, nº 1-4, p. 295–320.

FABRE, C. et A. LENCI, (2015), « Distributional Semantics Today », *Traitement automatique des langues*, vol. 56, nº 2, p. 7-20.

FETZER, J. H. (2011), « Minds and Machines : Limits to Simulations of Thought and Action », *International Journal of Signs and Semiotic Systems*, vol. 1, nº 1, p. 39-48.

FLAMENT, C. et M.-L. ROUQUETTE (2003), *Anatomie des idées ordinaires : comment étudier les représentations sociales*, Paris, A. Colin.

HÉBERT, L. (2016), *Dictionnaire de sémiotique générale*, *Signo*, http://www.signosemio.com/documents/dictionnaire-semiotique-generale.pdf, consulté le 17 mai 2016.

HERMAN, D. (2009), « Narrative ways of worldmaking », *Narratology in the age of cross-disciplinary narrative research*, nº 20, p. 71.

KETNER, K. L. (1988), « Peirce and Turing: Comparisons and conjectures », *Semiotica*, vol. 68, nº 1-2, p. 33–62.

KINTSCH, W. et T. A. VAN DIJK (1975), « Comment on se rappelle et on résume des histoires », *Langages*, nº 40, p. 98-116.

LIU, K. (2000), *Semiotics in information systems engineering*, Cambridge, UK, Cambridge University Press.

MANNING, C. D., P. RAGHAVAN et H SCHÜTZE. (2009), *Introduction to information retrieval* (Online edition), Cambridge, UK, Cambridge University Press.

MANNING, C. D. et H SCHÜTZE (1999), *Foundations of statistical natural language processing*, Cambridge, MIT press.
MEUNIER, J.-G. (1989), « Artificial intelligence and sign theory », *Semiotica*, vol. 77, nº 1-3, p. 43-64.
MEUNIER, J.-G. (2014), « Humanités numériques ou computationnelles : Enjeux herméneutiques », *Sens-Public*.
MEUNIER, J.-G. (2017a), « Humanités numériques et modélisation scientifique », *Questions de communication*, nº 31, p. XX.
MEUNIER, J.-G. (2017b), « Sémiotique et computation », *Applied Semiotics / sémiotique appliquée*.
NADIN, M. (1977), « Sign and fuzzy automata », *Semiosis*, vol. 1, nº 5, p. 19-26.
NADIN, M. (2011), « Information and Semiotic Processes. The Semiotics of Computation », *Cybernetics & Human Knowing*, vol. 18, nº 1-2, p. 153-175.
QUEIROZ, J. et F. MERRELL (2008), « On Peirce's Pragmatic Notion of Semiosis — A Contribution for the Design of Meaning Machines », *Minds and Machines*, vol. 19, nº 1, p. 129-143.
RAPAPORT, W. J. (2012), « Semiotic Systems, Computers, and the Mind: How Cognition Could Be Computing », *Int. J. Signs Semiot. Syst.*, vol. 2, nº 1, p. 32–71.
SAHLGREN, M. (2006), « The Word-space model » (thèse de doctorat), Stockholm University, Stockholm.
SAHLGREN, M. (2008), « The distributional hypothesis », *Italian Journal of Linguistics*, vol. 20, nº 1, p. 33-54.
SALTON, G. (1971), *The SMART retrieval system: Experiments in automatic document processing*, NJ, Prentice-Hall, Upper Saddle River.
SALTON, G., A. WONG et C.-S. YANG (1975), « A vector space model for automatic indexing », *Communications of the ACM*, vol. 18, nº 11, p. 613-620.
SAUSSURE, F. de (1995), « Principes généraux », dans *Cours de linguistique générale*, Paris, Payot, p. XX.
STAMPER, R. K. (1973), *Information in Business and Administrative Systems*, New York, John Wiley & Sons.
STEINBACH, M., G. KARYPIS et V. KUMAR (2000), « A Comparison of Document Clustering Techniques », *KDD workshop on text mining*, nº 400, p. 1-2.
TANAKA-ISHII, K. (2010), *Semiotics of programming*, New York, Cambridge University Press.
TANAKA-ISHII, K. (2015), « Semiotics of Computing: Filling the Gap Between Humanity and Mechanical Inhumanity », *International Handbook of* Semiotics, New York, Springer, p. 981-1002.
VAN DIJK, T. A. (1977), « Semantic macro-structures and knowledge frames in discourse comprehension », *Cognitive Processes in Comprehension*, p. 3-32.

VIJAYARANI, S., M. J. ILAMATHI et M. NITHYA (2015), « Preprocessing Techniques for Text Mining An Overview », *International Journal of Computer Science and Communication Networks*, vol. 5, nº 1, p. 7-16.

WEISS, S. M., N. INDURKHYA et T. ZHANG, T. (2010), « From Textual Information to Numerical Vectors », dans *Fundamentals of Predictive Text Mining*, New York, Springer, p. 13-38.

YOUNG, K. et J. SAVER (2001), « The Neurology of Narrative », *SubStance*, vol. 30, nº 1, p. 72-84.

ZEMANEK, H. (1966), « Semiotics and Programming Languages », *Communications of the ACM*, vol. 9, nº 3, p. 139–143.