Les sciences Humaines aux prises avec les algorithmes

Jean-Claude Soulages, Julien Velcin, Solange Kurpiel, Luis Otávio Dias, Myrian del Vecchio, Frédéric Aubrun

Les sciences Humaines aux prises avec les algorithmes ou lorsque le verbe se fait chiffre.

The human sciences battling against the algorithms or when the verb is made number.

Las ciencias humanas a las tomas con los algoritmos o cuando el verbo se hace cifra.

Article inédit.

Jean-Claude Soulages

Jean-Claude SOULAGES est Professeur en Sciences de l'information et de la communication à l'Université Lumière Lyon 2 et chercheur au Centre Max Weber UMR 5283. Il est spécialiste de l'analyse des discours médiatiques et des productions culturelles. Ces recherches portent sur la construction des identités (nationales, sociales, de genre, etc.) à travers les médias d'information, la télévision et la publicité en relation avec l'imaginaire social.

Jean-claude.soulages@univ-lyon2.fr

Julien Velcin

Julien Velcin est enseignant-chercheur en Informatique à l'Université Lyon 2. Spécialiste en fouille de données, il travaille au sein du laboratoire ERIC et cherche à développer des nouveaux outils pouvant être employés pour analyser des données complexes issues du Web. Il collabore régulièrement avec des chercheurs en Sciences Humaines et Sociales dans le cadre de projets pluridisciplinaires

Julien.velcin@univ-lyon2.fr

Résumé

Notre recherche interdisciplinaire s'efforce d'élaborer des méthodologies qui permettent de documenter la cartographie sémantique et l'orientation discursive de quatre éditions du *Huffington Post* sur de très longues durées. Dans cet aller retour entre l'analyse supervisée par l'humain et le traitement algorithmique, la démarche retenue se caractérise par la mobilisation de deux paliers d'analyse ; le niveau du *texte,* sous traité à des algorithmes définissant un *domaine scénique* rattaché à un *thème-événement* et celui du *discours* supervisé par des annotations de l'analyste, dénotant un point de vue ou une posture énonciative (témoignage, expertise, dénonciation, etc.). A moyen terme, le travail des algorithmes sera d'apprendre (*machine learning*) à reproduire ces orientations sémantiques et discursives.

Mots clés

Big data, journalisme, discours, thématique

Abstract

Our interdisciplinary research tries to develop methodologies that allow to document the semantic cartography and the discursive orientation of four editions of Huffington Post on very long-lasting. In

this round trip between the analysis overseen by the human being and the treatment algorithmic, the reserved approach is characterized by the mobilization of two steps of analysis. The level of the text, under treaty in algorithms defining a scenic domain connected with a theme-event and that of the speech overseen by notes of the analyst, denoting a point of view or an enunciative posture (testimony, expertise, denunciation), etc.). In the medium term, the work of the algorithms will be to learn (learning machine) to reproduce these semantic and discursive orientations.

Keywords

Big data, journalism, discourse, topic, machine learning

Resumen

Nuestra investigación interdisciplinaria se esfuerza por elaborar metodologías que permiten documentar la cartografía semántica y la orientación discursiva de cuatro ediciones de Huffington Post sobre duraciones muy largas. En esta ida y vuelta entre el análisis supervisado por el humano y el tratamiento algorítmica, el paso retenido se caracteriza por la movilización de dos descansillos de análisis; El nivel del texto, bajo tratado a algoritmos que definen un campo escénico relacionado con un tema-acontecimiento y el del discurso supervisado por anotaciones del analista, denotan un punto de vista o una postura enunciativa (testimonio, peritaje, denuncia, etc.). A medio plazo, el trabajo de los algoritmos será aprender (machine learning) a reproducir estas orientaciones semánticas y discursivas.

Palabras clave

Big data, peridismo, discurso, tema, machine learning

Plan de l'article

Introduction

Notre contribution vise à interroger les mécanismes de « production de la société » introduits par la « data-ification » de notre environnement, non pas tant dans les pratiques et les objets générés par nos économies mais à travers le traitement et l'instrumentation de cette question par la recherche scientifique en sciences humaines et sociales. Quelles nouvelles pertinences devons-nous attendre de la prise en compte de ce phénomène et quelles retombées le recours à certains de ces outils sont-ils susceptibles d'introduire dans nos méthodologies ? En dépit de l'aura symbolique et du bonus de scientificité supposés du fait de l'appariement avec un tel appareillage, il convient de rappeler, pour l'exploration de cet immatériel qu'est l'étude du sens social, la nature des objets qui sont exposés au travail des algorithmes. Comme nous le rappelle François Rastier, tout corpus de données relève d'une pratique sociale, d'un type de discours ou d'un genre culturel encadrés par des ressources interprétatives mobilisées tant en production qu'en réception par des collectifs (Rastier, 2011). En bref, les SHS ne traitent pas de choses mais d'objets socialement construits. Ce postulat découle du fait que l'environnement de l'homme doit être considéré comme un environnement culturel, c'est-à-dire extrait de la nature et élaboré en exploitant et s'appropriant les ressources de cette dernière mais aussi en produisant et thésaurisant des savoirs informels, des connaissances d'arrière plan, devenus autant d'artefacts d'informations sur le monde. C'est grâce à eux que l'homo sapiens a pu interagir avec différents milieux de façon raisonnée et programmatique et donc en utilisant au mieux ces ressources.

Mais ces informations sont longtemps restées des données inertes que l'homme utilisait de façon intuitive et ponctuelle, qui pouvaient lui servir de signaux (les nuages qui annoncent l'orage) ou s'incarner dans des outils (les armes). Si elles aboutissaient à une projection dans un futur immédiat ou à une certaine maîtrise de l'environnement naturel, elles ne permettaient pas un enrichissement cognitif et cumulatif de sa connaissance du monde. Depuis des millénaires, de l'invention du langage, de l'écriture en passant par le phonographe (Kittler, 1999), jusqu'aux machines intelligentes contemporaines, cet enrichissement a été rendu possible et ce stockage de l'information est devenu accessible à presque tous. Aujourd'hui, avec nos interfaces intelligentes, ces interactions, routines, savoirs informels sont stockés et traduits dans le langage des chiffres et circulent dans des réseaux. Pourquoi les stocker dans un cerveau, si une interface les gère ? Pourquoi s'encombrer de supports physiques si l'accès au *cloud* — la galaxie des chiffres— et grâce à un « métamédium » (Manovich, 2010, p. 62) nous permet d'accéder à tous les services et toutes les informations vingt quatre heures sur vingt quatre ? C'est bien un accroissement sans précédent des savoirs et de la « zone proximale de développement » de l'homme contemporain (Vygotski, [1934] 1997) qui s'est opéré avec ces formes d'externalisation et de circulation des informations. L'intelligence de l'homme s'est petit-à-petit déposée dans son décor quotidien. Ce à quoi nous assistons, c'est à une intégration d'une parcelle d'intelligence dans nos pratiques et nos objets quotidiens. S'agit-il d'un rétrécissement de notre monde ou plutôt de sa duplication donnant le jour à son décalque invasif dans notre vie quotidienne ? Mais comment maîtriser ces phénomènes ou comment les englober dans une nouvelle couche d'intelligibilité ?

Une opportunité mais aussi des écueils

Pour les sciences humaines et sociales, la multiplication de ces données pousse non seulement dans un premier temps à leur prise en compte mais surtout à la production d'outils adéquats. En effet, l'ordinateur n'a été jusqu'à présent conçu que comme un outil supplétif, pour certains décoratif et tout à fait subsidiaire. Il a représenté un assemblage disparate de fonctionnalités déjà existantes et dupliquées mais il n'en est pas moins devenu un terminal et un réceptacle incontournables pour nos pratiques contemporaines. Or, il est temps de l'utiliser non plus comme un simple marteau ou un instrument parmi une panoplie d'outils. Il s'agit de l'intégrer comme une véritable prothèse, géniteur

d'une nouvelle intelligence partagée, intelligence tout à la fois humaine et artificielle susceptible de générer une complémentation homme machine. Toutefois, il ne s'agit en aucun cas de céder aux logiques industrielles, technico-commerciales qui sont celles des outils miracles qui n'opèrent le plus souvent qu'un formatage des données. Il va de soi que cette démarche ne peut aboutir sans la mise en pratique d'une « interdisciplinarité focalisée » entre sciences dures (informatique) et sciences humaines (sociologie, SIC, sciences du langage, etc.) du fait qu'elles sont conduites à travailler de façon complémentaire sur le même objet (Charaudeau, 2009). Toutefois, si cette « data-ification » effective des données que les chercheurs rencontrent offre l'opportunité de travailler autrement sur de nouveaux corpus d'objets, elle présente néanmoins certains risques et de nombreux écueils.

Le premier écueil tient au traitement et la sélection de données qui tendent à n'objectiver que certains phénomènes, objectivation toujours partielle et arbitraire. Il faut bien émettre une hypothèse qui tient plutôt d'une sorte de postulat — et, à l'épreuve des faits, la tenir souvent pour une utopie — qui veut que le discours des hommes fait d'énoncés symboliques de toutes sortes, constitutifs de la discursivité sociale de nos collectivités, possède une logique systémique mais aussi une complexité certaine. Or les traitements numériques de données nous imposent de passer par des protocoles d'indexation et de codage des énoncés qui ont tendance à durcir cette logique ; c'est ce que tout un chacun peut constater en interrogeant par mots-clés Google, Twitter ou d'autres médias sociaux. Souvent, les résultats obtenus nous inciteraient à assimiler le traitement cognitif du cerveau humain à la configuration résultant du traitement des algorithmes. Il nous faut repousser la tentation d'une telle isotopie qui se voudrait pour beaucoup quasi-idéale. S'il peut exister certaines corrélations, il faut nous prévenir contre cette alternative ou conformité trop facile. Tout le monde le sait, la carte ne se confond jamais avec le territoire, sauf dans les récits fantastiques de Jorge Luis Borges. Les signes ou les pratiques que les sémiologues et les sociologues ont longtemps et toujours scrutés sont devenus des données traduites dans une autre langue, celle des algorithmes, or ces mêmes algorithmes sont susceptibles de dégager des structurations spécifiques tout à fait autres. De quelle formes d'intelligibilité déduite des data, s'agit-il? Quel est leur degré de pertinence, ne sont-elles finalement qu'un simple fantasme statistique et privé de toute pertinence cognitive et anthropique ? Ne sommes nous pas en train d'agréger statistiquement une fiction à des données ? Et quelle est l'identité de ce méta-locuteur qui se cache dans les coulisses du *deep learning* ? La société qui se parle à elle-même à travers le marc de café des data ? Ou bien n'est ce pas à nouveau, derrière les sentences du cloud, un jeu de ventriloquie et une instance transcendante qui reprendrait la parole ou le risque d'une « gouvernance par les nombres » (Supiot, 2015) et finalement le verbe qui se fait chiffre ?

Un deuxième écueil est à éviter qui caractérise les postures de la plupart des démarches d'analyse automatique de la langue, ce récif lexical qui débouche souvent sur un comptage uniforme des mots, sur lequel viennent se fracasser beaucoup de recherches. Face à la langue et sa nébuleuse de signifiants, l'analyse peut échouer et le plus souvent se contenter, par défaut, de ce cabotage. Le risque que l'on court alors avec ces approches comptables, c'est d'obérer le fait que le discours observé est inscrit dans un genre issu d'une pratique sociale qui détermine en grande partie le sens des énoncés. Faire l'impasse sur ces contraintes situationnelles, c'est risquer de dissoudre la valeur de signifiés au profit de l'agglutination mécanique de signifiants, ce que les informaticiens redoutent et dénomment du vocable de « gap sémantique ». Car il convient de garder à l'esprit que ce sont un cadre institutionnel et des normes communicationnelles qui structurent la sémiosis sociale et contribuent à générer l'intercompréhension de ce sens social. Certes on ne peut reprocher à des linguistes cette pratique qui n'est qu'une extension de leur corpus disciplinaire ou à certains chercheurs d'utiliser la langue comme seul instrument d'analyse mais d'autres approches privilégient un autre seuil de configuration et d'interprétation, qui ne relève pas exclusivement de la langue mais du discours. Or, dans beaucoup de démarches, ce biais lexical est naturalisé et surpondéré, on découvre des corrélations de signifiants, des agrégations de désignations, des corrélations générant des nuages de mots qui ne renvoient le plus souvent qu'à un dictionnaire des idées reçues. Il faut

donc redoubler ces premières indexations purement lexicales par un autre seuil d'analyse, celui de la mise au jour des configurations discursives qui structurent les énoncés, et qui sont introduites dans un premier temps par la supervision externe du chercheur en opérant pour paraphraser Kant une synthèse transcendantale des données.

Dernier écueil qui tient plus de la déploration. Il faut cesser de céder à cette hypnose thaumaturgique de l'intelligence artificielle avec son escorte de périls et de miracles et dissiper le rideau de fumée de soumission et d'impuissance présent chez de nombreux chercheurs en SHS. Est-il utile de rappeler que les machines intelligentes comme toutes les machines sont des créatures engendrées et conçues par l'homme ? Que, sans lui, elles ne seraient rien, sinon un amas de matériaux et un gargouillis de signaux. Il faut à tout prix mettre un terme à ce complexe d'incomplétude et de faiblesse. Depuis des millénaires, tout notre environnement est fait d'intelligence humaine incarnée et distribuée dans des objets et des services. Si nous avons affaire de plus en plus à un comportement intelligent ou autonome de ces derniers, c'est en fait d'une intelligence humaine augmentée et instrumentée dont il s'agit. C'est bien ce partage et les bénéfices attendus du recours à l'emploi de ces nouveaux types d'instrumentation qui nous ont poussé à entreprendre une recherche interdisciplinaire portant sur les *big data* émanant aujourd'hui de l'univers médiatique[1].

Une data-ification des messages médiatiques

En effet, les médias, tout comme le langage, ont toujours constitué des ressources symboliques qui s'interposent entre le monde et l'être humain. Ils élaborent pour ce dernier un patrimoine et un monde commun et représentent aujourd'hui avec internet des supports innombrables de connaissance et de réfraction de notre monde, accessibles désormais à tous grâce à la nébuleuse des *big data*. En conséquence, pour l'analyse des discours médiatiques, le recours au traitement algorithmique et à la prothèse que représente l'outil informatique est non seulement possible mais quasiment incontournable. Pour les chercheurs désireux d'explorer la galaxie médiatique cette démarche comporte deux enjeux.

Le premier est de palier l'incommensurabilité de l'approche de certains objets qui confronte le chercheur à une nébuleuse d'occurrences qui outrepasse la compétence analytique de toute analyse qualitative assumée par un cerveau humain mais à laquelle il s'agit de ne pas renoncer. Or, depuis la quasi data-ification de tout l'univers médiatique il devient possible de sous-traiter certaines opérations d'analyse à des algorithmes et donc de partager des tâches et surtout d'étayer certains résultats au moyen de protocoles statistiques de validation. Par ailleurs, si les occurrences médiatiques signifiantes sont trop nombreuses, trop multiformes pour être explorées par un seul observateur, elles n'ont aucunement le privilège ou la prétention de dépendre d'une seule corporation et d'une seule discipline et leur analyse impose le recours à cette « interdisciplinarité focalisée » avec les sciences du chiffre déjà évoquée.

Le second enjeu tient à l'examen et l'indexation automatisée, non supervisée de vastes corpus d'énoncés qui sont susceptibles de mettre au jour des formes de corrélations, des redondances, des configurations non-visibles, en un mot la structure insue de certains phénomènes de discours. En effet, dans un premier temps, il s'agit de sous-traiter des opérations à des algorithmes —le traitement et le regroupement de données—, et donc de partager des tâches, et en prenant appui sur des

. . . . . .

[1] La recherche en cours JADN (journalisme à l'heure du numérique) regroupe différentes universités (Curitiba, Lyon, Bordeaux, Beyrouth,) partenaires autour de chercheurs du Centre Max Weber (sociologie) et du laboratoire ERIC (informatique) de Lyon 2. Elle est financée par l'Institut des systèmes complexes Rhône-Alpes et vise à étudier 4 éditions du *Huffington Post* (France, Brésil, USA, Liban) dans leur contexte régional.

protocoles statistiques d'émettre un certain nombre d'hypothèses interprétatives. Une démarche d'hybridation s'impose donc dans ce partage des tâches entre l'homme et la machine. Car c'est bien l'appariement avec l'intervention supervisée de l'analyste qui seule peut permettre en dernière instance de dégager le sens des énoncés. C'est à celle-ci qu'il incombe la tâche de validation ou non de ce premier socle de traitement automatique afin d'opérer la mise en œuvre d'un second processus, herméneutique cette fois-ci, la complémentation et l'interprétation des résultats. On renouerait ainsi avec une vieille dichotomie sur laquelle reposait la première sémiologie structurale, une analyse immanente des données référentielles redoublée par une phase interprétative et analytique reposant sur une série d'interprétants externes (des inférences sociologiques, culturelles, politiques, etc.) apportés par la supervision de l'analyste (Houdebine, 2015). Car l'agencement des signifiants ne peut à lui seul rendre compte et épuiser le sens d'un discours. Comme les pragmaticiens l'ont démontré, l'analyse en immanence ne peut, à elle seule, réduire le sens des énoncés. Certes il est toujours possible d'isoler et d'inventorier des significations éparses, mais il est chimérique et souvent biaisé de vouloir résoudre le sens cohésif produit par un énoncé situé, renvoyant à une instance et un contexte énonciatif prédéterminé sans recourir à des interprétants externes. On peut prolonger cette analogie entre le traitement informatique automatisé du matériau linguistique et un traitement supervisé par l'intermédiaire d'annotations humaines. Le premier ordonnancement du matériau langagier porte sur les signifiants, leur organisation, leur récurrence, leur éloignement ou leur proximité sémantique à travers des procédés d'agglutination, reposant implicitement sur la similarité des référents mobilisés dans différents énoncés. Or le plus souvent le signifié tend à outrepasser la stricte nomenclature exprimée par une simple référence, l'usage social du signe peut opacifier la relation au référent ou bien la débrayer. Il peut aussi, dans un jeu d'ambivalence délibérée, en postuler plusieurs. Et c'est bien la supervision apportée par les inférences émises par le sujet interprétant qui va rendre compte du sens d'un énoncé dans tel ou tel contexte. Dans un premier temps la référence, dans un second l'inférence. Et à moyen terme, le travail des algorithmes sera d'apprendre (*machine learning*) à reproduire les orientations sémantiques et discursives produites par l'homme à partir du premier traitement — en observant les pratiques de validation, de qualification et de distribution des énoncés par le superviseur.

Un terrain d'application

La recherche en cours JADN, déjà évoquée, à travers le déroulement de ses différentes étapes peut nous servir de guide. Il s'agit en l'occurrence de l'élaboration de méthodologies destinées à la navigation dans un flux de données qui permettent de documenter une cartographie sémantique concernant quatre éditions du Huffington Post sur de très longues durées et, à plus long terme, — il est toujours permis de rêver—, en temps réel. Dans cet aller retour entre l'analyse supervisée par l'humain et le traitement algorithmique, une collaboration dialectique prend place qui permet de mettre au jour les procédés de construction des énoncés mais aussi les mécanismes sur lesquels repose leur interprétation par l'homme, assisté par la machine. La question demeure bien sûr de pouvoir définir quel est le degré et quel est le type de supervision assurés par l'homme. Sur le plan des contenus, le recours au *data mining* et à une cartographie sémantique des thématiques traitées nous permet de dégager une sorte de "*newsscape*" ou "d'*eventscape*" de chacune des éditions régionales du *Post* et de révéler ainsi les phénomènes de focalisation, de hiérarchie, de pondération ou de *gate keeping* de tel ou tel « thème-événement » (Soulages, 2002) dans ses éditions locales et simultanément les interactions possibles entre le local et le global[2].

. . . . . .

[2] Le choix du *Huffington Post* tient à la flexibilité et l'hybridité de sa logique rédactionnelle en rupture par rapport aux normes en vigueur jusque là. La réussite de ce pure player incarne la mutation en cours dans l'univers du journalisme. Présenté à la presse et au public en 2005 aux USA, comme un média indépendant qui vise à révolutionner le journalisme grâce à une offre numérique "alternative" tant du point de vue technologique que politique, le *Huffington Post* se décline aujourd'hui, dans plus de vingt pays, et connaît au niveau mondial plus de 200 millions d'usagers. « Cette conversation

La machine comme véhicule autonome

A l'intersection de l'informatique et de la statistique, des travaux récents relatifs à la fouille de textes traitent de l'extraction, de la visualisation et du suivi automatique de thème-événements, en particulier à partir de documents que l'on trouve à profusion sur Internet (*news*, twitts, articles de presse...). La fouille de textes (*text mining* en anglais) est un domaine dont l'objectif consiste à développer des modèles et des algorithmes pour l'analyse automatique de vastes volumes de données textuelles. Contrairement aux méthodes développées en traitement automatique des langues, elle favorise des approches issues de la statistique et de l'analyse des données. Parmi les nombreuses applications abordées, la fouille de textes permet, entre autres, d'extraire automatiquement et sans *a priori* les thématiques abordées dans un ensemble de textes (*topic modeling ou topic learning* en anglais). De nombreux algorithmes ont été proposés, principalement basés sur des outils mathématiques de décomposition de matrices (Paatero, Tapper, 1994) ou utilisant des modèles graphiques probabilistes (Blei et al., 2003). Certains algorithmes permettent aussi de suivre les thématiques dans le temps, et ainsi de détecter les thématiques déclinantes ou émergentes. Afin d'étudier ces données, nous avons mis en place un aspirateur de flux RSS sur quatre versions du Huffington Post (américaine, française, brésilienne, arabe) à partir de juin 2016, à la fois sur les articles de presse que sur les blogues. Chaque article est associé à un titre, au corps de l'article, à une date et à un auteur.

La première étape portait sur un corpus de trois mois de collecte (plus précisément du 20 juin au 8 septembre 2016) et se concentraient uniquement sur les trois premières versions et sur les articles de presse (12 067 articles pour les USA, 4133 pour la France, 2355 pour le Brésil). Concernant la partie automatique portant sur l'extraction des thématiques, nous avons utilisé le modèle LDA (Blei et al., 2003) dans son implémentation parallèle disponible via la librairie MALLET[3]. Les hyper-paramètres du modèle correspondant respectivement à l'a priori sur les thématiques et sur les documents sont symétriques et ont été estimés automatiquement à partir des données (Mccallum et al., 2009). Après plusieurs tentatives manuelles, nous avons choisi de fixer le nombre k de thématiques à 100. Cette distribution automatique a été soumise aux chercheurs en SHS. A l'issue de leur analyse une majeure partie des thématiques a été conservée car jugées suffisamment pertinentes par les sociologues. Plus précisément : 15 ont été écartées en français, 16 en anglais et 16 en portugais. Beaucoup d'entre elles ont été fusionnées manuellement en une quinzaine de macro-thématiques (politique étrangère, climat, femme, show business, etc.). Les algorithmes sont parvenus ainsi à distribuer des flux d'énoncés en sous-ensembles que les chercheurs en SHS peuvent dispatcher en rubriques correspondant aux orientations thématiques d'une pratique sociale précise : le journalisme. Il restera donc aux algorithmes à apprendre de l'homme et à le seconder et c'est à cette deuxième étape qu'a été consacrée l'évolution de l'interface logicielle (Velcin, Soulages alii, 2017).

Vers une interface homme machine

Dans une deuxième étape, pour répondre aux centres d'intérêt des différentes équipes de chercheurs en SHS et afin d'affiner la démarche, deux thématiques ont été retenues ; la « condition féminine » et le « réchauffement climatique ». Parallèlement, à la suite du premier traitement non supervisé, l'élaboration d'un outil de navigation à l'intérieur de ces bases de données, s'est imposé. Le développement du logiciel Newsbrowsers, a permis l'extraction de corpus d'articles susceptibles

[suite de la note]

mondiale » (dixit l'un de ses fondateurs, Ariana Huffington) fait cohabiter la production d'articles d'actualité nationale et globale avec l'expression d'opinions très souvent locales sous forme de blogues signés par des experts, des personnalités reconnues ou encore de témoignages hors normes (50%).

[3] http://mallet.cs.umass.edu/

de mettre au jour l'actualité des deux thèmes événements. Du premier traitement algorithmique a été retenu un thésaurus de descripteurs (*tags*) apparus en relation avec les deux thématiques permettant l'extraction de corpus d'articles traitant de ces deux macro-thèmes. Le traitement a débouché sur la distribution de ces derniers en différents *domaines scéniques* (Soulages, 2002) relatifs au différents tags attribués — comme /harcèlement sexuel/ viol/ sexualité/ famille/mariage, etc. ou bien pollution /ou /biodiversité/, etc. Dans un premier temps, les articles comportant les descripteurs ou tags attribués par le traitement automatique, sont validés (ou invalidés) par les chercheurs qui peuvent toujours introduire de nouveaux descripteurs appliqués au corpus retenu mais qui n'avaient pas été retenus lors de la première étape d'indexation (par exemple : femme noire pour le corpus brésilien ou sport pour les corpus français et étasunien). A moyen terme, à partir de ces premières expérimentation semi-supervisées, on peut envisager que le développement de l'interface logicielle autorisera le suivi de la distribution temporelle des articles et des blogues dans chacune des éditions permettant une analyse des relations (reprises, dénégations, droit de suite, etc.) qui se tissent entre les références du discours journalistique et les pratiques amateures qui l'accompagnent en mettant au jour d'éventuelles corrélations. Il deviendra ainsi possible d'analyser la réactivité des échanges entre ces deux univers, la traçabilité et l'impact d'une nouvelle, d'un événement, leurs trajectoires et leurs reprises, leurs répercussions, leurs zones de partage. L'indexation de ces éditions du *Post* autorise également le couplage de l'analyse macro du flux à une analyse de discours plus qualitative qui s'attache non seulement au suivi, à une période précise, mais aussi à l'apparition puis la disparition d'un thème-événement. Cette description sémantique indexée serait à même de scruter les variations dans le traitement d'un même thème-événement ou d'un domaine scénique particulier, et autoriserait le prélèvement d'échantillons significatifs pour une analyse de discours plus poussée de certains corpus d'articles ou de blogues.

Une analyse de contenu

La démarche retenue se caractérise par la mobilisation de deux paliers d'analyse ; le niveau du *texte* et celui du *discours* — même s'il faut bien convenir que cette partition demeure purement heuristique. Elle nous autorise toutefois à transformer des "données" en signes interprétables en dépit du fait que ces deux paliers d'indexation correspondent à deux tags ou plus attribués à chacun des articles ou blogues. En effet, d'un côté, nous lui associons un item sémantique, ce que nous dénommons un domaine scénique rattaché à la thématique climat, par exemple, / pollution /, de l'autre nous introduisons un descripteur-utilisateur dénotant un fait de discours (un point de vue ou une posture énonciative) qui peut prendre la forme d'un témoignage, d'une expertise ou d'une dénonciation. Mais d'un point de vue théorique, cette distinction a bien toute sa pertinence puisqu'elle atteste, au niveau macro, des enjeux sociétaux assumés par chacune des rédactions.

Le premier palier qui a été sous traité à des calculs algorithmiques, que nous pouvons définir comme le niveau du texte des articles, est révélateur d'une nomenclature lexicale et d'un système morphosyntaxique fermes (vocabulaire, syntagmes lexicaux, etc.), configuré et circonscrit par des contraintes de référenciation du propos et de stratégies génériques (la cohésion de tout article de presse autour d'une information et d'un angle, la personnalisation d'un blogue autour d'une thématique et d'un point de vue). Il s'agit dans un premier temps de repérer pour le logiciel la présence de certains termes ou expressions lexicales illustrant ou rattachés à une thématique (grâce à des descripteurs validés par les chercheurs) dans les quatre langues et dans de gros corpus d'articles (5 mois continus pour chaque édition du Huffington Post juillet /novembre 2016). A chacun des articles, Newsbrowsers attribue un ou plusieurs tags, que le superviseur va valider ou invalider. Ce premier traitement permet pour chaque édition de circonscrire le spectre sémantique et les types de référenciation concédés à chacune des thématiques et d'en dresser en quelque sorte la cartographie sémantique. Pour des raisons de concision nous ne présentons ici que certains des résultats obtenus

suite au traitement de la thématique /réchauffement climatique/ et dans 2 corpus seulement (Tableau n°1).

DOMAINES SCENIQUES	**FRANCE**	**USA**
CLIMATOSCEPTICISME	1,23%	16,89%
BIODIVERSITE	1,85%	12,67%
ALIMENTATION	2,06%	1,36%
AUTOMOBILE	2,26%	0,90%
POLLUTION	2,88%	11,16%
ENERGIES RENOUVELABLES	3,91%	4,68%
SANTE	6,17%	3,47%
DEVELOPPEMENT DURABLE	6,38%	5,88%
DECHETS RECYCLAGE	14,20%	4,07%
DEBAT POLITIQUE	17,28%	23,23%
CHANGEMEMENT CLIMATIQUE	41,77%	15,69%

Tableau 1 : les domaines scéniques du thème-événements réchauffement climatique en pourcentage dans les deux éditions ; Huffington Post US (631 articles tagués) et France (518 articles tagués)

Au terme de cette première étape, la pertinence de certains descripteurs retenus nous ait apparue tout à fait relative, soit du fait de leur sous-représentation dans un corpus donné (ex : déforestation dans les corpus français et nord-américain) soit du fait de leur proximité (automobile et particule fine fusionnés en automobile).

Une analyse de discours

A cette première indexation sémantique et distribution en domaines scéniques des articles, succède l'introduction d'un second palier cette fois-ci totalement supervisé par les chercheurs en SHS, celui de l'attribution de descripteurs-utilisateurs révélateurs, cette fois-ci, de faits de discours. Ces derniers témoignent de postures énonciatives (témoignage, dénonciation, mobilisation, etc.) ou de points de vue discursifs (principe de précaution, hypothèses, initiatives, etc.) des locuteurs. C'est donc bien le chercheur qui cette fois-ci seul va attribuer ce descripteur-utilisateur à chaque article. Cette deuxième étape part de l'hypothèse que tout énoncé comporte toujours une visée d'influence, voire une dimension argumentative enfouie et le plus souvent indirecte (Amossy, 2012) et, dans les deux cas, est porteur d'un point de vue orientant l'interprétation en réception, même si cela peut s'opérer à travers des stratégies d'effacement énonciatif typique du discours journalistique comme dans le cas des faits ou des propos rapportés (Rabatel, 2017).

	initiatives	témoignage	hypothèses	dénonciation	innovation	mobilisation	préconisation	principe de précaution
France	24%	1%	10 %	18%	4%	14%	12%	17%
USA	18%	1%	8 %	45%	1%	6%	3%	18%

Tableau 2 : les postures énonciatives et les points de vue en pourcentage dans les deux éditions

Faute de place, il va de soi que l'interprétation de ces résultats fera l'objet d'autres publications. Quelques remarques peuvent y suppléer. Les deux médias divergent dans les postures énonciatives mobilisées. L'édition américaine accorde un privilège manifeste à la posture de dénonciation (45 % contre 18 % pour l'édition française) confirmant la polarisation du traitement médiatique de la thématique aux USA. Par contre, le Huffington Post France atteste d'un privilège accordé aux initiatives (24 %), aux dénonciations (18 %) et aussi au principe de précaution (17 %).

Le croisement de nos deux paliers d'indexation permet alors de scruter l'orientation argumentative et discursive de chacun des domaines scéniques à l'intérieur de chaque corpus (figure 1 et 2).

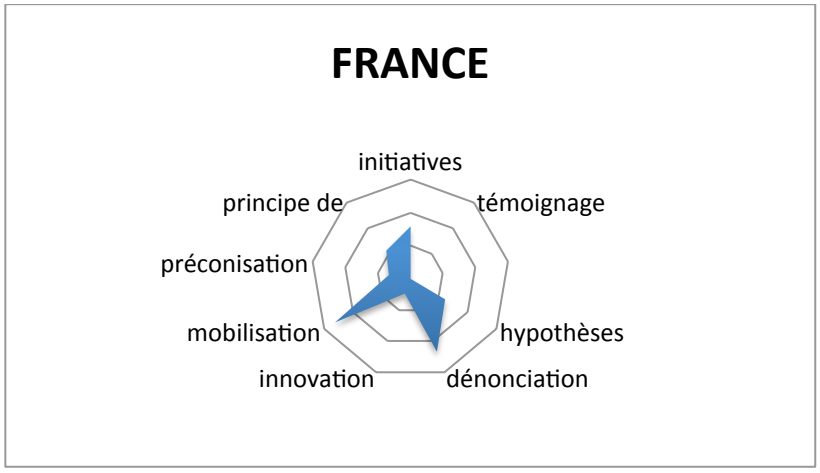

Figure 1 : les postures énonciatives et les points de vue en pourcentage du domaine scénique /débat politique/ dans l'édition française.

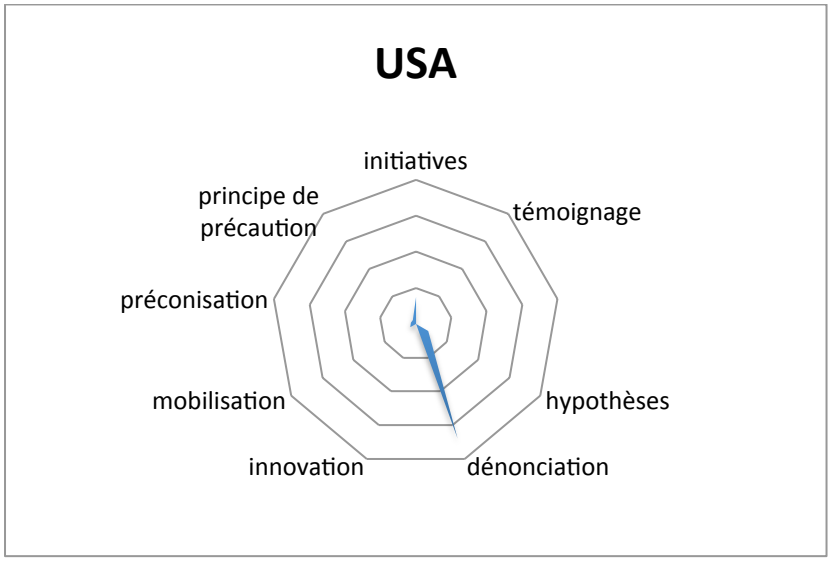

Figure 2 : les postures énonciatives et les points de vue en pourcentage du domaine scénique /débat politique/ dans l'édition nord-américaine.

En définitive, la thématique réchauffement climatique, si elle est bien envisagée dans l'édition française sous de nombreux aspects, tant du coté des domaines scéniques explorés que des postures énonciatives invoquées, n'est traitée dans l'édition américaine — et sans doute la période choisie joue un grand rôle—, que sous pratiquement un seul aspect, celui du conflit, de la polémique et de la dénonciation, et donc de la légitimité d'un tel questionnement (le poids des postures

climatosceptiques comme de leurs adversaires) alors que dans l'édition française le thème-événement est très peu remis en question et est tout à fait légitime dans le débat public. Du reste, en scrutant le domaine scénique /débat politique/ on retrouve la focalisation de l'édition américaine sur la dénonciation, alors que son traitement dans la version française propose un spectre plus large des points de vues et des postures. En définitive, nos analyses invalident, d'une part la vision homogénéisante très souvent décriée du traitement de médias transnationaux comme le Huffington Post, mais aussi les effets de contagion de certaines idéologies (le climatoscepticisme) et du même coup une certaine stabilité et pérennité des espaces publics nationaux.

Conclusion

Pour tenter de conclure sur une recherche toujours en cours, cette instrumentation du regard à travers le recours au *big data* et aux algorithmes, génère une nouvelle ontologie, cette ontologie digitale qui va petit à petit représenter le liquide amniotique d'un nouveau type de recherche. Recherche qui prendrait le contrepied de cette information de plus en plus désintermédiée qui court toujours le risque d'être éviscérée de son sens. Car la compilation des news, la saturation des sources et la montée d'un certain data-journalisme ne proposent parfois que l'agglutination de signifiants privés de signifiés.

Autre conséquence, la notion académique de texte ou de document, bornée en production par le tunnel de l'actualité et le formatage d'un média — très souvent analysée à l'aulne du prisme d'un micro-horizon par le chercheur—, a implosé, confronté à un bain d'intertextualité, d'interdiscursivité et d'intermédialité. Certes, on peut accepter la métaphore de « la base de données », relationnelle et donc vivante, que Lev Manovich superpose au réseau des réseaux (Manovich, 2012), mais on doit y opposer le paradigme de l'encyclopédie de savoirs, de l'intelligence et de l'expérience humaine. Il ne faut en aucun cas céder à cet arraisonnement de notre environnement par ces machines qui enregistrent et indexent en continu le présent et renoncer à transpercer la nébuleuse des *big data*, pour y retrouver les matériaux d'un discours.

On peut aisément comprendre que ce qu'apporte ces démarches, c'est bien un élargissement de ces technologies de l'intellect dont nous parle Jack Goody qui ont pris naissance avec le langage et l'écriture et cette « littératie » (Goody, 2007) qui recouvre, selon le philosophe, l'inscription de la matérialité mais aussi de la symbolique du sens. C'est aujourd'hui, cet univers que délivre nos interfaces digitales, sources de paroles, d'écrits et d'images. C'est dans ce sens que notre équipe pluridisciplinaire s'est investie pour développer des méthodologies et des outils d'apprentissage supervisés, tout en limitant l'effort de supervision (on parle d'approche faiblement supervisée). Ceux-ci devraient permettre de mieux retrouver les thématiques après la curation du spécialiste, voire les registres du discours (ex. témoignage, dénonciation, etc.).

N'oublions pas que les sciences de l'esprit, la sémiologie, les sciences du langage et les théories structuralistes ont accompagné et ont souvent devancé la société de l'information digitale qui est aujourd'hui la nôtre. Si la physique, la chimie ont pu constitué la charpente de la technoscience et du capitalisme industriel des deux derniers siècles, ce sont les data et donc les signes qui composent aujourd'hui l'infrastructure de notre environnement cognitif mais aussi la texture de notre environnement social et de ce capitalisme cognitif qui émerge sous nos yeux (Moulier Boutang, 2007). Et, si ce sont bien les sciences cognitives et les sciences du chiffre et du signe qui apparaissent comme les instruments et les leviers de ces bouleversements c'est à elles de relever le défi de leur maîtrise et de leur compréhension.

Références bibliographiques

Amossy Ruth, (2012), L'argumentation dans le discours, Paris, Nathan Université.

Blei David. M., Ng Andrew. Y., Jordan Michael.I. and Lafferty John (2003), « Latent Dirichlet Allocation. » Journal of Machine Learning Research, vol. 3.

Charaudeau Patrick, (2010), « Pour une interdisciplinarité « focalisée » dans les sciences humaines et sociales », Questions de communication n°17, Nancy, Presses universitaires de Nancy, p. 195-22.

Goody Jack, (2007), Pouvoirs et savoirs de l'écrit, Paris, La dispute.

Houdebine, Anne-Marie, (2015) « De l'imaginaire linguistique à l'imaginaire culturel », La Linguistique, vol. 51, fasc. 1/2015, Paris, PUF, pp. 4-39.

Kittler Friedrich, (1999), Gramophone, film, typewriter. Stanford, Stanford University Press, [1986].

Mccallum Andrew, David. M. Mimno, et Hanna. M. Wallach (2009). Rethinking lda : « Why priors matter. » In Advances in Neural Information Processing Systems , pp. 1973–1981.

Manovich Lev, (2010), Le langage des nouveaux médias, Dijon, Les presses du Réel.

Manovich Lev, (2012), « Database as a symbolic form », en ligne consulté le 15 mars 2018 « http://manovich.net/index.php/projects/database-as-a-symbolic-form.

Moulier Boutang Yann, (2007), Le Capitalisme cognitif. La nouvelle grande transformation, Paris, collection Multitudes/Idées, Éditions Amsterdam.

Paatero Pentti., Tapper Unto. (1994), « Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. » Environmetrics, vol. 5 (2), pp. 111-126, Wiley Online Library.

Rabatel Alain, (2017) Pour une lecture linguistique et critique des médias. Empathie, éthique, point (s) de vue, Limoges, éditions Lambert Lucas.

Rastier François, (2011), « Langage et pensée : dualité sémiotique ou dualisme cognitif ? », Intellectica, n°56, 2011/2, pp. 29-79.

Soulages Jean-Claude,(2002), « Un thème événement : la guerre en ex-Yougoslavie (1990-1994) », Questions de Communication n°1, Nancy, Presses universitaires de Nancy, p. 69-79.

Supiot Alain, (2015), La gouvernance par les nombres, Paris, Fayard, coll. « Poids et mesures du monde ».

Velcin Julien, Soulages Jean-Claude, Kurpiel Solange, Dias Luis Otávio, Del Vecchio Myrian, Aubrun Frédéric, (2017), « Fouille de textes pour une analyse comparée de l'information diffusée par les médias en ligne : une étude sur trois éditions du Huffington Post », Extraction et Gestion des Connaissances (EGC), Grenoble, France, en ligne https://halshs.archives-ouvertes.fr/hal-01571265.

Vygotski Lev, (1997), Pensée et langage, Paris, La Dispute, [1934].