« Enquêter à partir des traces textuelles du web »

Présentation

Jean-Philippe Cointet, Sylvain Parasie

L'essor du web et des réseaux sociaux offre aux chercheurs en sciences sociales un volume considérable de textes : des livres et des articles numérisés ou « nativement numériques », mais aussi des discussions sur des forums, des échanges sur les réseaux sociaux ou des propos tenus sur diverses plateformes en ligne. D'une grande hétérogénéité, ces matériaux textuels sont le support d'un registre étendu d'activités sociales : des citoyens débattent sur une plateforme délibérative des réformes souhaitables en matière de fiscalité et de services publics ; un travailleur présente son parcours, ses compétences et ses contacts professionnels sur LinkedIn ; des mères de famille échangent des conseils de puériculture sur un forum, etc. Combinés au développement d'outils informatiques, ces matériaux sont à l'origine d'une promesse qui a été formulée à l'intérieur et à l'extérieur des sciences sociales : notre compréhension du monde social serait plus profonde et plus précise si nous exploitions quantitativement les « traces » textuelles de tant d'activités sociales (Lazer et al., 2009).

Nombreux sont les chercheurs que cette promesse ne laisse pas insensibles. D'abord parce que le traitement quantitatif d'un matériau textuel volumineux laisse espérer que l'enquête pourrait embrasser un plus grand nombre d'objets qui les intéressent depuis longtemps : des interactions, des opinions, des façons de se présenter, etc. Ensuite parce qu'à la différence des questionnaires traditionnels, les matériaux textuels ainsi collectés n'ont pas été sollicités par le chercheur, ce qui leur donne un caractère moins artificiel et plus spontané. Enfin parce que ces textes sont souvent associés à des informations de nature relationnelle – un tweet cite un autre tweet, un individu affiche ses liens avec d'autres individus, etc. – qui s'intègrent naturellement au raisonnement en sciences sociales.

Cet élargissement des sources de l'enquête en sciences sociales se heurte toutefois à plusieurs obstacles. Ces matériaux textuels sont produits par des plateformes numériques qui non seulement sollicitent, encadrent et mettent en forme les expressions et les échanges, mais contrôlent aussi la façon dont les chercheurs y ont accès (ou non). Au point où la métaphore de la « trace » paraît bien illusoire, les chercheurs se retrouvant davantage à enquêter sur des plateformes que sur les activités sociales qui s'y déploient (Marres, 2017 ; Beuscart, 2017). Autre obstacle à l'enquête, l'analyse quantitative de corpus textuels reste une pratique confidentielle en sciences sociales, et tout particulièrement en sociologie, en dépit de son ancienneté (Demazière et al., 2006). L'essor de nouvelles techniques quantitatives issues des mondes informatiques ajoute un trouble supplémentaire lié à l'opacité des algorithmes et leur difficile appropriation par les chercheurs en sciences sociales. À ces deux obstacles s'en ajoute un autre plus fondamental encore : l'ancrage social des personnes qui s'expriment sur le web et les réseaux sociaux demeure en grande partie inconnu. Font souvent défaut des informations

aussi cruciales pour l'enquête que le niveau de revenus ou de diplôme, la catégorie sociale, et même l'âge ou le genre de ceux et celles qui prennent la parole en ligne.

Devant ces obstacles, certains chercheurs préfèrent renoncer à exploiter de tels matériaux, privilégiant les méthodes plus établies. À l'opposé, d'autres collègues jugent que ceux-ci imposent de renouveler en profondeur l'épistémologie des sciences sociales. C'est l'argument défendu par Dominique Boullier, selon qui les sciences sociales doivent prendre comme objet des « traces techniquement collectées », en cessant de les rattacher à des individus, à des groupes, à la société ou l'opinion (Boullier, 2015). Entre ces deux perspectives, certains chercheurs ambitionnent néanmoins de produire des connaissances à partir de ces matériaux textuels issus du web, sans pour autant remettre en cause l'épistémologie des sciences sociales (cf. Baya-Laffitte et Benbouzid, 2017). En se confrontant aux difficultés pratiques de l'enquête, ils doivent alors inventer de nouvelles façons d'enquêter en intégrant les traces du web dans les questionnements de sciences sociales.

C'est précisément dans cette démarche que s'inscrivent les chercheurs qui contribuent à ce numéro de *Réseaux*. Par le biais d'un appel, nous avons sollicité des recherches en sciences sociales qui intègrent un traitement quantitatif de matériaux textuels issus du web ou des réseaux sociaux – tweets, commentaires, interventions dans des forums, échanges sur Wikipédia, etc. Écartant les contributions exclusivement méthodologiques, nous avons retenu les articles de sociologues, de chercheurs en sciences de l'information et de la communication, de chercheurs en « sciences sociales computationnelle ». Portant sur des thèmes différents (politique, science, médecine, journalisme, mémoire collective, notamment), chacune de leurs enquêtes apporte une contribution substantielle à un domaine de recherche et porte un certain nombre d'innovations quant à la façon d'intégrer l'analyse du texte à une question de sciences sociales.

Dans cette introduction, nous soulignerons d'abord l'effervescence qu'on observe aujourd'hui autour de la « mise en données » du texte, et les enjeux que cela soulève pour les sciences sociales. Puis en nous appuyant sur plusieurs recherches notables et sur les travaux rassemblés dans ce numéro, nous identifierons plusieurs « chemins d'enquête » qui nous semblent offrir des solutions originales pour surmonter les obstacles associés au traitement quantitatif des matériaux textuels issus du web. Enfin, nous présenterons chaque article du numéro, en insistant sur les résultats originaux ainsi obtenus.

Effervescence autour de la « mise en données » des textes

Depuis une dizaine d'années, l'analyse quantitative de corpus textuels s'est considérablement renouvelée. Au croisement des mondes de l'ingénierie, de l'apprentissage automatique et du traitement automatique de la langue, des infrastructures numériques sont apparues qui visent à « mettre en données » des corpus souvent volumineux. Sous l'impulsion de recherches souvent financées par les industries du numérique, des algorithmes ont été conçus afin de traiter ces corpus textuels à grande échelle et de façon plus ou moins automatique. Ces techniques émergentes peuvent être classées en plusieurs familles (Cointet et Parasie, 2018) parmi lesquelles les approches *lexicométriques* (qui mesurent des fréquences de mots ou d'expression) ; l'*analyse de sentiment* (qui mesure le sentiment général ou les émotions qui sous-tendent un corpus à partir de marqueurs textuels) ; les *réseaux sémantiques* (qui analysent la structure sémantique d'un corpus en modélisant les textes comme un réseau de mots) ; les

modèles thématiques (qui produisent une représentation de la structure thématique d'un corpus de texte).

Le fait que des chercheurs en sciences sociales aient recours à des techniques quantitatives d'analyse textuelle ne constitue en rien une nouveauté. La pratique est ancienne non seulement aux États-Unis (Berelson et Lazarsfeld, 1948), mais aussi en France où le courant de l'analyse de données, autour de Benzécri, a très tôt diffusé ses méthodes en sociologie, en gestion, en science politique, etc. (Lebart et Salem, 1988 ; Beaudouin, 2016). La nouveauté réside plutôt dans le fait que les nouvelles techniques ont été élaborées dans des mondes et selon des perspectives qui sont très éloignées des sciences sociales. Dans le cas français, on constate en effet que les principaux logiciels d'analyse textuelle, qui sont encore utilisés aujourd'hui ont été conçus en étroite collaboration avec des chercheurs en sciences sociales (Demazière et al., 2006) – à l'instar d'Iramuteq ou de Prospéro (Chateauraynaud, 2003). Au contraire, les techniques les plus récentes circulent dans des mondes sociaux beaucoup plus variés (à la fois marketing, industriels et académiques) et se déploient dans des espaces disciplinaires très différents les uns des autres (biologie, physique, histoire, psychologie, linguistique, etc.).

Les chercheurs en sciences sociales sont donc confrontés à une tension. D'un côté, ils sont davantage en mesure d'accéder à un grand nombre de textes qui sont plus étroitement liés aux mondes sociaux dans lesquels évoluent leurs contemporains ; d'un autre côté, s'ils veulent analyser ce matériau, ils sont contraints de mobiliser des algorithmes qui sont étrangers au raisonnement des sciences sociales, et qui – comme souvent avec les algorithmes de l'apprentissage automatique – présentent une opacité considérable (Cardon et al., 2017). Les algorithmes apprenants sont au cœur d'une question qui peut être posée en ces termes : comment le chercheur peut-il se donner les moyens de renouveler la connaissance de sciences sociales, sans déléguer l'analyse à la machine ? Depuis quelques années, plusieurs initiatives visent à identifier les apports et les limites de ces méthodes d'enquête en sociologie (Evans et Aceves, 2016 ; Cointet et Parasie, 2018), en science politique (Grimmer et Stewart, 2013) et en économie (Gentzkow, Kelly et Taddy, 2017).

Toutefois, l'enjeu pour les sciences sociales ne se réduit pas à savoir s'il est opportun ou non d'importer des techniques issues d'espaces de recherche éloignés. Des informaticiens, des physiciens, des linguistes et des psychologues utilisent déjà ces techniques pour s'emparer d'objets traditionnellement étudiés par les sciences sociales. En étudiant des mouvements sociaux (Budak et Watts, 2015), des interactions sociales (Danescu-Niculescu-Mizil, ce numéro), des discriminations (Voight et al., 2017) ou le cycle de l'information (Leskovec et al., 2009) via le traitement à grande échelle de corpus textuels, ces chercheurs viennent bousculer la juridiction des sciences sociales.

Chemins d'enquête

Mais arrêtons-là l'évocation des enjeux, pourtant majeurs, liés à l'analyse quantitative des traces textuelles. Le parti pris de ce numéro a été de rassembler des enquêtes qui n'abordent pas de façon frontale des questions de méthodes, mais posent des questions de recherche substantielles – comment se construit la mémoire collective sur le web ? Comment les militants se mobilisent-ils contre leurs adversaires lors d'une campagne électorale ? Quels sont les ressorts de l'engagement des participants aux projets de science citoyenne ? Comment se transforme dans le temps la structure des échanges sur un forum ? Les personnes qui consultent des informations factuelles sur des plateformes en ligne peuvent-elles former des publics ?

En répondant à ces questions à partir du traitement quantitatif de matériaux textuels issus du web, les auteurs ont été confrontés aux obstacles que nous avons présentés – en premier lieu le manque d'informations sur l'ancrage social des locuteurs. Ils ont ainsi été conduits à explorer ce que nous appelons des « chemins d'enquête », c'est-à-dire un ensemble de « ficelles » d'un métier de chercheur en sciences sociales qui se transforme. Ce sont trois chemins que nous allons maintenant évoquer à la lumière des contributions de ce numéro mais qu'empruntent également d'autres travaux récents qui nous semblent dessiner une même topographie : (1) la combinaison des méthodes qualitatives et quantitatives ; (2) l'identification de groupes sociaux ; (3) le traitement des énoncés textuels comme supports de relations entre acteurs.

Combiner les méthodes

Un premier chemin consiste à associer l'analyse quantitative de corpus à d'autres méthodes d'enquête plus conventionnelles, notamment qualitatives. C'est là une façon de donner une plus grande épaisseur sociale aux acteurs qui s'expriment en ligne, en accédant à leurs trajectoires et à leurs interprétations. L'article que Valérie Beaudouin consacre dans ce numéro à la construction de la mémoire collective de la Grande Guerre sur le web est emblématique d'une approche « quali-quanti » (Venturini, Cardon et Cointet, 2014). La sociologue double le traitement quantitatif d'un corpus de sites (graphe de réseaux ; identification des classes lexicales d'un forum via Iramuteq) d'une dizaine d'entretiens auprès d'acteurs impliqués dans la mémoire de la Guerre 14-18. Comme elle l'explique dans son article, cette combinaison des méthodes lui permet d'accéder au sens que les acteurs attribuent à leurs pratiques d'écriture ; de reconstituer des échanges interpersonnels qui n'apparaissent pas dans le corpus en ligne ; mais aussi de tester auprès des acteurs la pertinence des analyses statistiques.

Des recherches récentes empruntent un chemin similaire, combinant l'analyse statistique de corpus textuels issus du web avec l'administration de questionnaires. L'enquête s'enrichit ainsi de nouvelles variables qui ne peuvent pas être collectées à partir des seules plateformes numériques. Analysant des corpus issus de Facebook, Irène Bastard et ses collègues ont par exemple conçu une application qui offre aux utilisateurs une carte de leur réseau d'amis en échange d'informations sur leurs activités ou leurs propriétés sociales (Bastard et al., 2017). C'est aussi la démarche de Christopher Bail et ses collègues, qui ont voulu identifier les raisons pour lesquelles certaines associations parviennent à utiliser Facebook pour stimuler la conversation du public (Bail, Taylor et Mann, 2017). Les trois sociologues ont collecté les posts et commentaires publiés sur les pages Facebook de plus de 200 associations dans le domaine de l'autisme et du don d'organes. À partir de plusieurs techniques automatiques d'analyse du langage, ils qualifient ces publications selon qu'elles relèvent d'un registre cognitif ou émotionnel. Mais pour démontrer que la mobilisation du public est véritablement fonction du registre de la publication, ils doivent vérifier si une variable cachée n'intervient pas, qui serait liée à la taille, aux ressources ou à d'autres caractéristiques des associations. À partir d'un questionnaire envoyé aux associations, ils concluent que le choix du bon registre de publication est, parmi tous les facteurs, celui qui explique le plus l'attention que rencontre l'association sur le réseau social.

Certains chercheurs doublent l'analyse quantitative d'une étude qualitative du même corpus textuel. C'est ce que font Neil Fligstein et ses collègues lorsqu'ils enquêtent sur l'attitude de l'instance américaine en charge de la politique monétaire lors de la crise financière de 2008 (Fligstein, Brundage et Schultz, 2017). Pour comprendre pourquoi la *Federal Reserve* n'a pas perçu l'importance de la crise alors que se multipliaient les signes d'effondrement du système

financier, ils analysent la structure thématique de comptes rendus de réunions internes. Ils prolongent leurs analyses statistiques d'une étude minutieuse du contenu des discussions qui se sont tenues à l'intérieur de l'organisme. Ce qui permet aux chercheurs de reconstituer la manière dont les acteurs ont ajusté leurs interprétations aux changements du contexte macro-économique.

Identifier des groupes sociaux

Un deuxième chemin consiste à regrouper les individus qui produisent les inscriptions textuelles. Les locuteurs peuvent être rassemblées de bien des façons, selon qu'ils partagent des traits socio-démographiques, des opinions, des pratiques, des ressources, un lieu de résidence, un problème, etc. Ces groupes ne se donnent souvent pas à voir directement à partir des données issues des plateformes du web, même si la plupart des individus qui s'y expriment catégorisent les autres participants (Velkovska, 2002). C'est que les plateformes prêtent peu d'attention aux catégories sociales en privilégiant les individus. On voit donc des chercheurs faire preuve d'une grande ingéniosité pour identifier des groupes au-delà de la myriade des contributions individuelles. L'article que Madeleine Akrich publie dans ce numéro est emblématique de cette démarche. Enquêtant sur la dynamique des échanges sur un forum consacré au dépistage prénatal, elle collecte près de 170 000 messages postés entre 2005 et 2018. Disposant de peu d'informations sur les contributrices – en dehors de leur genre que leur pseudonyme laisse paraître sans ambiguïté –, la sociologue élabore plusieurs métriques de la participation qui lui permettent d'isoler plusieurs groupes de participantes. À partir du volume et de la distribution temporelle de la participation dans ce forum particulier et dans les autres forums de la plateforme, Madeleine Akrich distingue un groupe minoritaire de femmes qui viennent sur ce forum alors qu'elles rencontrent un problème spécifique lié au dépistage, et un groupe majoritaire de femmes qui contribuent régulièrement à la plateforme et fréquentent régulièrement ce forum particulier afin de soutenir les femmes du premier groupe. L'usage fin de ces métriques de participation lui permet de reconstituer l'histoire de ce forum et les différentes modalités d'engagement dans la discussion.

D'autres articles du numéro identifient également des groupes sous-jacents. C'est le cas de l'enquête d'Élise Tancoigne et Jérôme Baudry, qui porte sur les participants à un projet de science citoyenne dans le domaine de l'astronomie. À partir des classes lexicales qui se dégagent de l'analyse Iramuteq[1] d'un corpus de 40 000 profils postés sur la plateforme, les deux chercheurs identifient six groupes de participants qui se distinguent par des modes d'engagement spécifiques dans le projet. Au-delà de l'extrême variété des présentations individuelles, l'enquête aboutit à distinguer des groupes dont les membres partagent des pratiques et des interprétations. L'article de David Chavalarias, Noé Gaumont et Maziyar Panahi, tout comme celui de Pierre Ratinaud et ses collègues, identifie des « communautés politiques » sur Twitter pendant la campagne pour l'élection présidentielle française de 2017. Chaque communauté est définie par les auteurs comme un ensemble d'utilisateurs de Twitter qui font circuler des messages politiques sans les modifier ou de façon minimale.

L'identification de tels groupes apparaît souvent comme la seule manière de construire une interprétation sociologiquement valide à partir de corpus textuels issus du web. C'est ce dont témoigne l'enquête de René Flores, qui a étudié l'effet sur l'opinion publique de l'adoption d'une loi anti-immigration en Arizona (Flores, 2017). Traitant un corpus de 250 000 tweets à l'aide d'algorithmes d'analyse de sentiment, il constate que l'adoption de la loi correspond à

[1] Logiciel reposant sur la méthode de classification de Max Reinert, qui identifie des classes lexicales dans un corpus textuel.

une augmentation des prises de position anti-immigration chez les utilisateurs de Twitter résidant en Arizona. Mais il serait sociologiquement faux, montre-t-il, d'en conclure que cette loi a modifié l'opinion des citoyens. L'étude des trajectoires de contribution sur un temps plus long indique que ce sont les personnes déjà mobilisées contre l'immigration qui prennent de plus en plus la parole. On voit ici que c'est l'identification des groupes sociaux – ici, les militants opposés à l'immigration – qui permet de saisir sociologiquement comment se déplace le débat en ligne.

Saisir les actions qui traversent les énoncés

Le dernier chemin exploré par certains auteurs de ce numéro implique d'identifier des actions ou les relations entre acteurs qui s'expriment à travers les énoncés. L'analyse textuelle ne se limite pas ici à dégager la structure des thèmes contenus dans un corpus, mais plutôt à reconnaître dans le texte lui-même des relations entre locuteurs et plus largement entre des « actants ». Les énoncés sont ici modélisés par le chercheur, de façon à mettre au jour un ensemble d'actions accomplies par le locuteur lorsqu'il prend la parole. Ce chemin d'enquête est bien antérieur au web et aux réseaux sociaux, puisqu'il s'inscrit dans la continuité de recherches en linguistique de l'énonciation (Mikhaïl Bakhtine, Kenneth Burke), et de leur importation dans la sociologie française (Boltanski et al., 1984 ; Chateauraynaud, 2003) et américaine (Franzosi, 1989).

L'enquête de Sylvain Parasie et Jean-Philippe Cointet publiée dans ce numéro repose sur cette démarche. S'interrogeant sur les publics qui se forment autour d'une plateforme américaine d'information sur les homicides commis à Los Angeles, les sociologues analysent près de 29 000 commentaires postés par des internautes en lien avec un meurtre précis. Comme les deux chercheurs disposent ici encore d'un minimum d'informations sur les locuteurs, ils s'inspirent du modèle actantiel autrefois déployé par Luc Boltanski (1984), et qu'ils avaient déjà mobilisé (Parasie et Cointet, 2012). Ce modèle envisage toute prise de parole comme la construction de relations entre différents « actants », qui sont ici le locuteur qui prend la parole sur la plateforme, la victime du meurtre, les responsables du drame, et le public dont on sollicite la compassion ou que l'on prend à témoin d'une injustice. Utilisant des algorithmes d'apprentissage supervisé pour qualifier les relations entre actants, Sylvain Parasie et Jean-Philippe Cointet distinguent les locuteurs qui manifestent un lien personnel avec la victime, de ceux qui s'expriment à distance soit pour rendre hommage soit pour discuter du problème de la violence urbaine.

L'enquête de Christian Danescu-Niculescu-Mizil et ses collègues emprunte un chemin proche, même si ses auteurs ne sont pas chercheurs en sciences sociales mais en informatique. Partant d'hypothèses sociolinguistiques, ils étudient les relations de pouvoir entre les contributeurs à Wikipédia en mesurant la façon dont ceux-ci imitent les façons de parler de leurs interlocuteurs. Ils cherchent à montrer qu'un individu tend à réutiliser systématiquement certains mots employés par son interlocuteur si le statut social de ce dernier est plus élevé que le sien (typiquement lorsqu'un wikipédien ordinaire s'adresse à un « administrateur »). Ici encore, le contenu des énoncés textuels est analysé de façon à identifier des relations entre acteurs.

Si ces différents chemins d'enquête sont encore aujourd'hui en cours d'exploration, la cohérence des propositions réunies dans ce numéro montre combien le partage et l'échange autour de ces ficelles peut être utile pour la communauté.

Présentation des articles

En plus de leur intérêt méthodologique, les articles de ce numéro apportent une contribution substantielle à plusieurs domaines de recherche. Le texte de Madeleine Akrich étudie ainsi les discussions entre participants d'un forum sur le dépistage prénatal en se fixant une double exigence : rendre compte de l'histoire de ce forum sur le temps long et décrire dans un même mouvement ses structures sémantiques et sociales. Des dizaines de milliers de messages provenant de milliers de contributeurs sont ainsi analysés pour rendre compte des agencements collectifs qui se construisent au sein du forum. En attachant aux contributeurs des profils d'activités, l'auteure parvient à qualifier avec beaucoup de finesse le type d'engagement individuel et le profil global de chaque sous-forum. Elle offre ainsi une analyse originale de la dynamique temporelle des assemblages sociaux qui sous-tendent un forum.

David Chavalarias, Noé Gaumont et Maziyar Panahi s'intéressent eux à la structure des échanges entre commentateurs politiques sur Twitter pendant la dernière campagne présidentielle. À partir d'un corpus de 60 millions de tweets, ils reconstruisent de façon automatique des « communautés politiques » dont les membres se retweetent régulièrement les uns les autres. La simple qualification de la provenance des énoncés permet aux auteurs de mesurer et suivre avec beaucoup de finesse l'évolution jour après jour des stratégies partisanes de soutien à son ou sa candidate ou d'attaque à l'encontre de ses concurrents. Plusieurs échelles temporelles coexistent. On distingue ainsi l'évolution journalière du ton des partisans des candidats qui peut fluctuer au gré des événements politiques majeurs qui scandent la campagne, et la dynamique propre du scrutin, qui au lendemain du premier tour fait apparaître non pas des changements stratégiques mais des reconfigurations plus purement structurales de l'organisation des soutiens à un parti ou son leader.

Élise Tancoigne et Jérôme Baudry enquêtent sur l'engagement dans les sciences participatives à partir du cas de SETI@home, un projet d'astronomie qui existe depuis deux décennies. Ils combinent une analyse de l'activité des contributeurs avec l'étude quantitative d'un corpus de plusieurs dizaines de milliers de fiches dans lesquelles les participants se présentent et expliquent les raisons de leur engagement. Les deux chercheurs montrent ainsi que les participants correspondent peu à l'image que s'en font les concepteurs des projets de science citoyenne. Se dessine l'image d'individus qui sont moins intéressés par la « Science » que par le dispositif de mise en réseau des participants et de leurs ordinateurs.

Pierre Ratinaud, Nikos Smyrnaios, Julien Figeac, Guillaume Cabanac, Ophélie Fraisier, Gilles Hubert, Yoann Pitarch, Tristan Salord et Thibaut Thonet proposent eux aussi un traitement systématique d'un corpus massif de tweets publiés pendant l'élection présidentielle française de 2017. Sans faire l'objet d'hypothèses préalables, le contenu des tweets est analysé pour découvrir la structure du discours des militants préalablement rattachés à leur famille politique. L'analyse lexicale est ainsi dupliquée pour mieux saisir la façon dont certains motifs récurrents du discours politique se retrouvent (ou non) sur l'ensemble du spectre politique et se synchronisent (ou non) avec l'agenda médiatique.

Valérie Baudoin s'attelle à une tâche peu commune dans les études numériques puisqu'elle reconstitue avec forces détails et en multipliant les points de vue le travail de construction sur le web d'une mémoire collective de la Grande Guerre. Combinant cartographie du web, analyses textuelles et entretiens, elle montre la nature des contenus échangés entre amateurs au sein des divers fils de discussion, mais aussi le type de ressources qu'ils mobilisent et valorisent. Dans le même temps, son enquête quantitative et qualitative lui permet de souligner la

permanence des frontières entre historiens amateurs et professionnels dans la construction de la mémoire collective.

Jean-Philippe Cointet et Sylvain Parasie enquêtent sur les publics qui se forment autour des occurrences diffusées par une plateforme appelée « The Homicide Report », laquelle fournit une information standardisée sur tous les homicides commis à Los Angeles. Dans quelle mesure, s'interrogent-ils, un « public » au sens fort du terme peut-il se former, autrement dit un être collectif qui partage des interprétations communes ? Pour répondre à cette question, ils mobilisent une méthode d'analyse textuelle très peu utilisée en sciences sociales, qui repose sur des algorithmes d'apprentissage supervisé, pour étudier un corpus de 28 000 commentaires. Ils montrent d'abord que les participants élaborent des interprétations communes à partir des occurrences qui leur sont adressées, en combinant trois façons de « faire public ». Ils soulignent ensuite que l'exploitation sociologique des inscriptions textuelles permet de réduire le fossé entre les enquêtes quantitatives sur les audiences et les études plus qualitatives.

Nous publions également la traduction d'un article de chercheurs en informatique, qui s'inscrit dans une veine de recherches très méconnue en France. Articulant des techniques de traitement automatique de la langue avec des théories sociolinguistiques, Christian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang et Jon Kleinberg étudient les relations de pouvoir à partir de corpus de conversations issues de Wikipédia et des échanges entre avocats et juges de la Cour suprême des États-Unis. Ils montrent ainsi que les participants à une interaction émettent des signaux linguistiques qui expriment la relation de pouvoir qui s'établit entre eux-mêmes et leurs interlocuteurs. Un individu qui discute avec un autre individu dont le statut social est supérieur au sien tendra à réutiliser systématiquement certains des termes que son interlocuteur utilise. Cette contribution montre que le matériau textuel peut être exploité pour interroger de manière empirique et à grande échelle des relations de pouvoir qui sont classiquement étudiées par les sciences sociales.

Enfin, deux articles sont publiés en *varia*. Le premier porte sur les espaces de co-working, son auteur Basile Michel montrant comment l'organisation de l'espace, les interactions entre co-workers et la régulation des flux révèlent différents modèles du travail communautaire. Le second est la traduction, par Danier Burnier, d'un texte du sociologue américain Harvey Molotch originellement paru en 1994 sous le titre « Sortir de chez soi » (*Going out*). Ce texte offre un contrepoint précieux à un numéro qui pourrait laisser penser que les sociologues peuvent se contenter d'observer le monde derrière leur écran. Constatant que la sociologie occupe une place réduite dans le débat public, Molotch encourage ses confrères à s'interroger sur la manière dont ils se comportent, à découvrir d'autres mondes et à vivre des vies riches.

Références

Bail, C. A., W. B. Taylor, M. Mann (2017), « Channeling Hearts and Minds : Advocacy Organizations, Cognitive-Emotional Currents, and Public Conversation », *American Sociological Review*, vol.82, n°6, p.1188-1213.

Bastard, I., D. Cardon, R. Charbey, J.-P. Cointet, C. Prieur (2017), « Facebook, pour quoi faire ? Configurations d'activités et structures relationnelles », *Sociologie*, vol.8, n°1, p.57-82.

Baya-Laffitte, N., B. Benbouzid (2017), « Imaginer la sociologie numérique », *Sociologie et sociétés*, vol.49, n°2, p.5-32.

Beaudouin, V. (2016), « Retour aux origines de la statistique textuelle : Benzécri et l'école française d'analyse des données », *13e Journées Internationales d'Analyse statistique des Données Textuelles*, p.17-27.

Berelson, B., P. F. Lazarsfeld (1948), *The Analysis of Communication Content*, Chicago et New York, University of Chicago Press.

Beuscart, J.-S. (2017), « Des données du Web pour faire de la sociologie... du Web ? » dans P.-M. Menger, S. Paye (dir.), *Big data et traçabilité numérique. Les sciences sociales face à la quantification massive des individus*, Paris, Collège de France, p.141-161.

Boltanski, L., Y. Darré et M.-A. Schiltz (1984), « La dénonciation », *Actes de la recherche en sciences sociales*, vol.51, p.3-40.

Boullier, D. (2015), « Les sciences sociales face aux traces du big data. Sociétés, opinion ou vibrations ? », Revue française de science politique, vol.65, n°5-6, p.805-828.

Budak, C., D. J. Watts (2015), « Dissecting the spirit of Gezi : Influence vs. selection in the Occupy Gezi movement », *Sociological Science*, vol.2, p.270-397.

Cardon, D., J.-P. Cointet, A. Mazières (2018), « La revanche des neurones », *Réseaux*, vol.5, n°211, p.173-220.

Chateauraynaud, F. (2003), *Prospéro. Une technologie littéraire pour les sciences humaines*, Paris, CNRS Éditions.

Cointet, J.-P., S. Parasie (2018), « Ce que le big data fait à l'analyse sociologique des textes. Un panorama critique des recherches contemporaines », *Revue française de sociologie*, vol.59, n°3, p.533-557.

Demazière, D., C. Brossaud, P. Trabal et K. Van Meter (dir.) (2006), *Analyses textuelles en sociologie. Logiciels, méthodes, usages*, Rennes, Presses Universitaires de Rennes.

Fligstein, N., J. S. Brundage, M. Schultz (2017), « Seeing like the Fed: Culture, cognition, and framing in the failure to anticipate the financial crisis of 2008 », *American Sociological Review*, vol.82, n°5, pp.879-909.

Flores, R. D. (2017), « Do Anti-Immigrant Laws Shape Public Sentiment ? A Study of Arizona's SB 1070 Using Twitter Data », *American Journal of Sociology*, vol.123, n°2, p.333-384.

Franzosi, R. (1989), « From words to numbers : A generalized and linguistics-based coding procedure for collecting textual data », *Sociological Methodology*, 19, p.263-298.

Gentzkow, L., B. T. Kelly, M. Taddy (2017), « Text as data », National Bureau of Economic Research, Cambridge (MA), *NBER Working Paper 23276*.

Grimmer, J., B. M. Stewart (2013), « Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts », *Political Analysis*, vol.21, n°3, p.267-297.

Lazer, D., A. Pentland, L. Adamic et al. (2009), « Computational social science », *Science*, vol.323, n°5915, p.721-723.

Lebart, L., A. Salem (1988), *Analyse statistique des données textuelles. Questions ouvertes et lexicométrie*, Paris, Dunod.

Leskovec, J., L. Backstrom, J. Kleinberg (2009), « Meme-Tracking and the Dynamics of the News Cycle », *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p.497-506.

Marres, N. (2017), *Digital Sociology : The Reinvention of Social Research*, Cambridge, Polity Press.

Parasie, S., J.-P. Cointet (2012), « La presse en ligne au service de la démocratie locale. Une analyse morphologique de forums politiques », *Revue française de science politique*, vol.62, n°1, p.45-70.

Velkovska, J. (2002), « L'intimité anonyme dans les conversations électroniques sur les *webchats* », *Sociologie du travail*, vol.44, p.193-213.

Venturini, T., D. Cardon et J.-P. Cointet (2014), « Présentation », *Réseaux*, vol.188, n°6, p.9-21.

Voight, R., N. P. Camp, V. Prabhakaran et al. (2017), « Language from Police Body Camera Footage Shows Racial Disparities in Officer Respect », *Proceedings of the National Academy of Sciences*, vol.114, n°25, p.6521-6526.