CAIRN.INFO
CHERCHER, REPÉRER, AVANCER

CE QUE LE BIG DATA *FAIT À L'ANALYSE SOCIOLOGIQUE DES TEXTES*

Un panorama critique des recherches contemporaines

Jean-Philippe Cointet, Sylvain Parasie

Presses de Sciences Po (P.F.N.S.P.) | « Revue française de sociologie »

2018/3 Vol. 59 | pages 533 à 557

ISSN 0035-2969
ISBN 9782724635669

Article disponible en ligne à l'adresse :

https://www.cairn.info/revue-francaise-de-sociologie-2018-3-page-533.htm

Pour citer cet article :

Jean-Philippe Cointet, Sylvain Parasie« *Ce que le* big data *fait à l'analyse sociologique des textes. Un panorama critique des recherches contemporaines* », *Revue française de sociologie* 2018/3 (Vol. 59), p. 533-557.
DOI 10.3917/rfs.593.0533

Distribution électronique Cairn.info pour Presses de Sciences Po (P.F.N.S.P.).
© Presses de Sciences Po (P.F.N.S.P.). Tous droits réservés pour tous pays.

La reproduction ou représentation de cet article, notamment par photocopie, n'est autorisée que dans les limites des conditions générales d'utilisation du site ou, le cas échéant, des conditions générales de la licence souscrite par votre établissement. Toute autre reproduction ou représentation, en tout ou partie, sous quelque forme et de quelque manière que ce soit, est interdite sauf accord préalable et écrit de l'éditeur, en dehors des cas prévus par la législation en vigueur en France. Il est précisé que son stockage dans une base de données est également interdit.

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)

[NOTE CRITIQUE]

Ce que le big data *fait à l'analyse sociologique des textes*
*Un panorama critique des recherches contemporaines**

Jean-Philippe Cointet
Sylvain Parasie

Résumé. Depuis les années 2000, de nouvelles techniques d'analyse textuelle font leur apparition au croisement des mondes informatiques, de l'intelligence artificielle et du traitement automatique de la langue. Bien qu'élaborées en dehors de toute préoccupation sociologique, ces techniques sont aujourd'hui mobilisées par des chercheurs – sociologues comme non-sociologues – dans le but de renouveler la connaissance du social en tirant parti du volume considérable de matériaux textuels aujourd'hui disponibles. En dressant un panorama des enquêtes sociologiques qui reposent sur la mise en données et le traitement quantitatif de corpus textuels, cet article identifie à quelles conditions ces approches peuvent constituer une ressource pour l'enquête sociologique. Les trois conditions qui émergent de notre analyse concernent : 1) la connaissance du contexte de production des inscriptions textuelles ; 2) l'intégration à l'enquête de données extérieures au texte lui-même ; 3) l'ajustement des algorithmes au raisonnement sociologique.

Mots-clés. *Big data* – Traces numériques – Corpus textuels – Lexicométrie – Réseaux sémantiques – Épistémologie – Intelligence artificielle – Traitement automatique de la langue

Depuis le milieu des années 2000, les sociologues ont accès à des matériaux textuels à la fois plus volumineux et plus hétérogènes. Ce sont d'abord des documents dont ils sont depuis longtemps familiers – articles de presse, publications scientifiques, documents administratifs ou juridiques, etc. –, mais qui sont désormais plus souvent disponibles sous une forme numérique. Ce sont également des traces textuelles qui sont produites par les individus dans le cadre de leurs activités en ligne – discussions sur les forums, échanges sur les réseaux sociaux, commentaires sur des sites de presse, évaluations de commerces, requêtes sur les moteurs de recherche, etc. Si bien que les sociologues, dont une partie était déjà habituée depuis plusieurs décennies à traiter quantitativement du texte grâce à l'informatique (Demazière *et al.*, 2006), accèdent aujourd'hui à des corpus textuels non seulement plus volumineux mais qui sont aussi liés à un registre plus large d'activités sociales. Ces corpus contiennent non seulement du texte, mais aussi des informations sur les relations entre locuteurs

* *Nous remercions les coordinateurs du numéro ainsi que Valérie Beaudouin, Marc Barbier, Bilel Benbouzid et David Demortain pour leurs remarques et suggestions. Les deux auteurs ont contribué de façon égale à la réalisation de l'article.*

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)

qui sont susceptibles d'intéresser les sociologues. C'est l'un des aspects du *big data*, souligné par de nombreux auteurs (Lazer *et al.*, 2009 ; Cardon, 2015 ; Marres, 2017 ; Beuscart, 2017).

Dans le même temps, se sont développés des dispositifs informatiques qui renouvellent en profondeur les façons de collecter, de traiter et d'analyser ces « données textuelles ». Ces dispositifs sont issus des mondes de l'informatique, de l'intelligence artificielle et du traitement automatique de la langue. Ils prennent la forme d'infrastructures de collecte et de stockage de données, mais aussi d'algorithmes ayant pour but de traiter et d'analyser ces données. À la différence des logiciels d'analyse textuelle dont l'usage s'est relativement routinisé – à l'image d'Iramuteq ou de Prospéro dans le contexte français –, ces algorithmes n'ont donc pas été conçus dans le monde de la recherche en sciences sociales. Plus encore, certains de ces algorithmes peuvent tout aussi bien être utilisés pour analyser des corpus textuels que pour catégoriser des images ou traiter des séquences d'ADN. Cette circulation des calculs dans des univers de connaissance très différents les uns des autres est un second aspect, tout aussi important, du *big data* (Cardon, 2015 ; Dagiral et Parasie, 2017).

Le fait qu'un volume croissant de textes puissent être collectés et traités de façon quantitative a nourri les ambitions des chercheurs venant de l'informatique et des sciences dures. Une partie d'entre eux ont prétendu renouveler la connaissance du social, mettant en question la juridiction de la sociologie (Abbott, 1988). Dès lors qu'on accède à des corpus massifs et visant l'exhaustivité, il serait alors possible, affirment ces chercheurs, de produire des connaissances d'une façon beaucoup plus inductive que ne le font la plupart des chercheurs en sciences sociales (Anderson, 2008). Ainsi, l'analyse statistique des mots contenus dans les millions de livres numérisés par Google permettrait de « reconstruire le squelette d'une nouvelle science » des faits culturels (Michel *et al.*, 2010). Par ailleurs, le traitement quantitatif d'inscriptions textuelles non sollicitées par le chercheur – à la différence des entretiens ou des questionnaires – offrirait un accès plus « naturel » aux opinions et aux interactions (Kitchin, 2014).

Les sociologues qui ont recours à ces dispositifs informatiques jouissent d'une visibilité croissante. Surtout en Amérique du Nord où ils sont aujourd'hui publiés par les revues de référence de la discipline (*American Journal of Sociology*, *American Sociological Review*), et non plus seulement dans des revues spécialisées ou interdisciplinaires. Certains d'entre eux sont des sociologues réputés, rompus aux méthodes traditionnelles, et qui voient dans ces nouvelles méthodes un moyen d'accroitre la portée de leurs analyses – à l'image de Paul DiMaggio ou Neil Fligstein. D'autres s'identifient au domaine des « sciences sociales computationnelles » (Lazer *et al.*, 2009), qui réunit à la fois des chercheurs formés en sociologie, en économie et en science politique, et des chercheurs venant de disciplines plus lointaines (physique, statistique, informatique, etc.). L'intégration de ces méthodes d'analyse textuelle à l'intérieur de la sociologie est encore limitée si l'on en juge par le nombre d'articles publiés – un peu plus d'une centaine au début de l'année 2018. Mais elles suscitent un écho plus important comme en témoigne la variété des objets sur lesquels des sociologues enquêtent avec ces dispositifs (culture, économie, science, politique, mouvements sociaux, etc.), et le nombre de conférences, de *workshops* et d'enseignements qui leur sont consacrés.

L'intérêt de la sociologie pour le matériau textuel, et son traitement quantitatif, est aujourd'hui établi. Voici plusieurs décennies que des sociologues constituent des corpus textuels et les traitent quantitativement au moyen de logiciels (Demazière

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)

et al., 2006 ; Chateauraynaud, 2003). Mais l'intégration de ces nouveaux dispositifs d'analyse textuelle au raisonnement sociologique n'a aujourd'hui rien d'évident. Comment des algorithmes et des infrastructures élaborées en dehors de toute préoccupation sociologique – et même parfois en dehors de toute perspective académique – peuvent-ils devenir une ressource pour l'enquête sociologique ? Doit-on y voir une menace pour le raisonnement sociologique ou, au contraire, une opportunité pour renouveler les méthodes de l'enquête sociologique (Savage et Burrows, 2007) ?

L'argument que nous défendons dans cet article est que, sous certaines conditions, ces méthodes émergentes peuvent être mises au service du raisonnement sociologique, dans un contexte où les mondes informatiques mettent en question la juridiction de la discipline. Ces conditions ont trait au contexte de production des inscriptions textuelles, à l'épaisseur sociale de ces inscriptions, et à l'articulation avec les catégories d'analyse propres à la discipline. C'est aussi pour contribuer à la définition de ces conditions que nous dressons un panorama critique des enquêtes sociologiques qui reposent sur la mise en données et le traitement quantitatif de corpus textuels. Ce faisant, nous prolongeons plusieurs revues de littérature récentes qui ont été réalisées en sociologie (Evans et Aceres, 2016), en science politique (Grimmer et Stewart, 2013) et en économie (Gentzkow *et al.*, 2017).

Pour réaliser ce panorama, nous avons d'abord réuni les articles de référence en informatique qui s'identifient à ces six familles de techniques d'analyse textuelle apparues à partir des années 2000 : nouvelles approches lexicométriques, modélisation thématique, analyse de réseaux sémantiques, analyse de sentiment, plongement de mots, analyse stylistique. Nous n'évoquerons donc pas ici les méthodes d'analyse textuelle quantitatives apparues antérieurement – notamment dans le contexte français à travers les logiciels Iramuteq ou Prospéro – dont l'intégration à l'enquête sociologique ne pose plus de difficultés majeures. Nous avons tout particulièrement prêté attention aux travaux qui, bien que réalisés par des informaticiens, portent sur des objets proches de ceux de la sociologie. Nous avons ensuite collecté tous les travaux de chercheurs se réclamant de la sociologie, ou ayant publié dans des revues de sociologie, qui mobilisent ces techniques dans leurs enquêtes. Tout en portant aussi attention à d'autres travaux remarquables qui s'inscrivent dans des disciplines voisines (science politique, histoire, économie), mais qui contribuent à la connaissance de l'action sociale.

L'article s'organise de la façon suivante. Dans une première partie, nous présentons ces nouvelles méthodes d'analyse textuelle issues des mondes de l'informatique et de l'intelligence artificielle. Nous montrons que ces méthodes sont utilisées par les chercheurs de ces domaines comme un moyen de remettre en cause la juridiction de la sociologie, c'est-à-dire le monopole que la discipline revendique sur l'étude de l'action sociale. Dans une deuxième partie, nous présenterons les principaux travaux sociologiques qui mobilisent ces méthodes, en montrant que celles-ci sont bel et bien mises au service de l'enquête sociologique. Enfin, la dernière partie discute des conditions permettant l'intégration de ces techniques émergentes à l'éventail des méthodes d'enquête en sociologie.

Méthodes émergentes et juridiction de la sociologie

Méconnus de la plupart des sociologues français, les nouveaux dispositifs d'analyse textuelle qui sont au cœur de cet article bousculent l'écologie des disciplines en général, et de la sociologie en particulier. Apparus au croisement de mondes souvent

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)

très éloignés des sciences sociales (Cointet, 2017), ils sont utilisés par certains chercheurs en informatique dans le but de produire des connaissances sur des objets qui sont traditionnellement ceux de la sociologie – l'imitation entre individus, les relations de pouvoir, les opinions ou les représentations, notamment. Dans les pages qui suivent, nous présentons chacun de ces dispositifs, en expliquant dans quel contexte ils ont été conçus et dans quelle mesure leur prétention s'étend à l'analyse de l'action sociale.

Le renouveau de la lexicométrie

Voici plusieurs décennies que la lexicométrie a sa place dans les sciences sociales, que ce soit aux États-Unis (Berelson et Lazarsfeld, 1948) ou en France (Lebart et Salem, 1988). Elle consiste à analyser les textes en comptant les mots qu'ils contiennent, c'est-à-dire en identifiant les motifs que dessine la répétition des éléments du lexique d'un corpus[1]. Critiquée par certains sociologues qui y voient une façon superficielle de traiter les textes (Chateauraynaud, 2003), la lexicométrie a néanmoins été régulièrement mobilisée en sciences sociales pour mettre au jour des représentations ou des idéologies. Depuis dix ans, l'approche lexicométrique a connu un renouvellement alors que s'élaboraient les premières bases de données textuelles massives issues du *web*.

En calculant des fréquences de mots ou d'expressions, des chercheurs en informatique – souvent rejoints par des chercheurs en sciences naturelles – ont profité de ces corpus massifs pour prétendre renouveler la connaissance de certains faits sociaux ou culturels. Un des projets les plus marquants remonte à la fin des années 2000, quand des ingénieurs, physiciens et biologistes collaborent avec Google pour construire un corpus composé de 4 % de l'ensemble de la littérature mondiale jamais imprimée. Par-delà la performance technologique que représente l'indexation d'une telle masse de données (plus de 5 millions de livres depuis 1800), les auteurs publient dans *Science* un article qui annonce la fondation d'une nouvelle science des faits culturels qu'ils appellent « *culturomics* » (Michel *et al.*, 2010). Ils annoncent « étendre les frontières de l'enquête quantitative rigoureuse à un large ensemble de phénomènes nouveaux couvrant les sciences sociales et les humanités ». Ne se limitant pas aux promesses, ils mesurent les fréquences de plusieurs termes afin de saisir des phénomènes aussi variés que l'évolution lexicale ou grammaticale des langues, la diffusion des technologies ou les effets de censure. À titre d'exemple, les auteurs montrent, en construisant des cohortes des personnes les plus « célèbres » nées à différentes périodes, que si l'accession à la « gloire » intervient de façon de plus en plus prématurée au fil de l'histoire, les célébrités retombent également dans l'oubli de plus en plus rapidement. On constate ici qu'un objet classiquement étudié par les sociologues (Chenu, 2008) est accaparé par des chercheurs qui mettent en avant le volume de leurs données et le caractère systématique de leur méthode.

D'autres chercheurs en informatique ont collaboré avec des sociologues ou des historiens, toujours dans le but de faire accéder les sciences humaines et sociales à une plus grande scientificité. C'est le cas de cette informaticienne qui, en compagnie d'un physicien et d'un historien, a analysé un corpus de 112 000 décisions de justice

1. La méthode de l'analyse factorielle développée par Jean-Paul Benzécri occupe, encore aujourd'hui, une place majeure en France. Elle se distingue néanmoins des approches lexicométriques plus simples par sa prise en compte des relations qui s'établissent entre les termes au sein d'un texte.

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)

rendues par la cour d'assises de Londres sur la période 1760-1913. Mesurant la fréquence des termes employés par cette cour, les chercheurs identifient un tournant au début du XIXe siècle, quand les crimes violents sont réprimés de façon plus forte que les crimes non violents (Klingenstein *et al.*, 2014). Ils mettent en évidence la divergence croissante, jusqu'au début du XXe siècle, des lexiques associés aux procès pour crimes violents et aux procès pour crimes non violents. Mobilisant Norbert Elias, ils identifient un « processus de civilisation » au terme duquel la puissance publique contrôle de plus en plus la violence.

Une partie de la recherche en sciences humaines et sociales a aussi contribué au renouveau des approches lexicométriques – autour du mouvement des « humanités numériques ». Ainsi, l'historien de la littérature Franco Moretti est un fervent défenseur de cette approche qu'il incorpore à sa pratique de « lecture distante » (Moretti, 2013). En collaboration avec l'historien des sciences Dominique Pestre, il a ainsi récemment publié un article dans lequel l'ensemble des rapports annuels de la Banque mondiale depuis les années 1960 sont analysés en suivant le même principe de comptage longitudinal (Moretti et Pestre, 2015). Ils identifient une rupture majeure dans le langage de l'institution à la fin des années 1970, quand la Banque impose une langue plus codée et plus éloignée des acteurs économiques réels.

Mais de nombreuses critiques ont été adressées à ces approches lexicométriques, en histoire et en sociologie tout particulièrement. Plusieurs chercheurs ont pointé les problèmes liés à la sélection des livres numérisés par Google, et interrogé plus fondamentalement la démarche qui consiste à attribuer un sens aux occurrences d'un terme sans interroger ses contextes d'apparition (Chateauraynaud et Debaz, 2010 ; Guerrini, 2011).

L'analyse de sentiment

Un autre type d'analyse de contenu textuel a émergé au début des années 2000 dans les laboratoires d'informatique, souvent en collaboration avec des acteurs industriels (IBM, Yahoo). Il s'agit de l'analyse de sentiment (*sentiment analysis*), qui désigne un ensemble de techniques visant à mesurer le « sentiment général » d'un texte à partir de marqueurs textuels produisant une mesure du « sentiment positif » ou du « sentiment négatif » associé à ce texte (Pang *et al.*, 2002). Reposant sur des algorithmes d'apprentissage automatique, ces techniques ont été développées à partir de corpus textuels volumineux issus du *web* (constitués par exemple de critiques de films publiées par des internautes). Les promesses liées à l'exploitation des traces sociales du *web* à des fins industrielles et commerciales ont soutenu une recherche importante dans le domaine (Pang et Lee, 2008).

Les ambitions des chercheurs se sont étendues aux objets traditionnels des sciences sociales – tout particulièrement les sondages d'opinion. Profitant des volumes de textes publiés par des personnes ordinaires sur les réseaux sociaux, un courant de recherche s'est formé autour de l'idée que la mesure automatique du sentiment de ces inscriptions textuelles permettrait de concurrencer les sondages politiques traditionnels. Parmi d'autres, Brendan O'Connor et ses collègues ont ainsi montré que les sondages portant sur l'approbation du Président Obama pendant l'année 2009 étaient fortement corrélés à la mesure du sentiment des *tweets* mentionnant le nom du président (O'Connor *et al.*, 2010). La promesse est alors que l'analyse de sentiment offrirait la possibilité de mesurer les mouvements de l'opinion, d'une façon plus efficace et moins coûteuse que les sondages.

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)

Ces techniques ont été très discutées, d'abord à l'intérieur des communautés informatiques et linguistiques. Il est apparu que la mesure des sentiments exigeait une modélisation de la sémantique et des aspects figuratifs du langage – et notamment de l'ironie (Bosco *et al.*, 2013). Le domaine de recherche est néanmoins resté actif, en incorporant des modèles syntaxiques plus complexes (Duric et Song, 2012), en capturant des échelles d'émotions plus riches que le modèle dichotomique classique distinguant sentiment positif/négatif (Kim *et al.*, 2013), et en saisissant les multiples cibles du jugement dans un même énoncé (Ruiz et Poibeau, 2015).

Si plusieurs sociologues et politistes ont mobilisé ces algorithmes dans leurs enquêtes, leur usage est aujourd'hui limité[2]. La plupart des chercheurs en sciences sociales sont restés critiques, contestant l'idée qu'on puisse dire quelque chose de l'opinion publique à partir de la mesure du sentiment des *tweets* (Mitchell et Hitlin, 2013). D'autres ont formulé une critique plus fondamentale de l'analyse de sentiment, expliquant qu'il est très difficile de déléguer à la machine la mesure d'opinions et de sentiments qui dépendent de dynamiques sociales que la seule analyse du texte ne peut entièrement saisir (Boullier et Lohard, 2012).

L'analyse stylistique

Élaborées au croisement de la linguistique et de l'informatique, les techniques d'analyse stylistique sont encore embryonnaires, mais elles se distinguent par leur intérêt pour des objets proches de la sociologie (relations de pouvoir, discriminations, etc.). Leur ambition est de saisir les caractéristiques stylistiques d'un texte en procédant à l'examen quantitatif systématique d'un certain nombre de critères inspirés de la sociolinguistique. L'objectif est de dire quelque chose des relations entre les locuteurs, en analysant quantitativement leurs interactions langagières, médiatisées ou de face-à-face. Leur hypothèse est que les variations stylistiques des échanges discursifs, identifiées quantitativement, permettent de rendre compte de la nature des relations entre les locuteurs.

Parmi les recherches qui saisissent des objets communément étudiés par des sociologues, citons l'enquête des chercheurs de l'université de Cornell (Danescu-Niculescu-Mizil *et al.*, 2012). Empruntant à la sociolinguistique l'hypothèse selon laquelle la coordination langagière perceptible dans les matériaux textuels peut servir à mesurer des différences de pouvoir au sein d'un groupe, ils analysent un corpus de 240 000 conversations entre contributeurs sur Wikipédia. Ils construisent alors une mesure de la « coordination linguistique » entre les individus qui discutent de l'édition des articles, en les distinguant selon le statut qu'ils ont sur Wikipédia. Les chercheurs montrent ainsi que les individus se coordonnent linguistiquement davantage avec les administrateurs de la plateforme qu'avec ceux qui ne le sont pas. Ils montrent également que lorsqu'un individu accède au statut d'administrateur, il se coordonne moins avec ceux qui n'ont pas le même statut que lui. Bien que conduite par des informaticiens, on voit qu'une telle recherche utilise des variables qui sont d'une nature sociale, et met en œuvre un questionnement qui est familier pour les sociologues.

Une autre enquête est allée plus loin en étudiant un phénomène classique de la sociologie américaine : les pratiques discriminatoires de la police aux États-Unis. Des

2. Une cinquantaine d'articles ayant recours à l'analyse de sentiment ont été publiés en science politique, communication, anthropologie, études culturelles et psychologie sociale, comparés à quelques centaines en gestion (*source* : *Web of Science*).

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)

chercheurs en informatique et en linguistique, de l'université de Stanford, analysent les propos que tiennent les policiers américains aux automobilistes qu'ils arrêtent sur la route (Voigt *et al.*, 2017). Mobilisant des travaux sociolinguistiques (relatifs aux « actes de langage qui menacent la face »), ils construisent un score qui évalue le respect dont les policiers font preuve dans ces situations. Ils appliquent ensuite cette mesure à un corpus constitué de 36 700 retranscriptions d'échanges provenant des caméras que portent les policiers d'Oakland. Ils montrent ainsi que les policiers parlent systématiquement de façon moins respectueuse aux citoyens noirs, et ce quels que soient la couleur de peau du policier, la gravité de l'infraction ou le lieu de l'arrestation. On voit ici à quel point ce type d'enquête identifie un problème social, en mesure la portée et remet en cause les explications avancées par les acteurs du débat.

Les réseaux sémantiques

Depuis la fin des années 2000, des chercheurs spécialisés dans l'analyse de réseaux se sont mis à traiter quantitativement des corpus textuels. Ils s'inscrivent moins dans la tradition sociologique de l'analyse des réseaux sociaux – apparue dans les années 1960-1970 autour des figures d'Harrison White, Ronald Burt ou Mark Granovetter – que dans la communauté dite des « réseaux complexes » apparue à la fin des années 1990 dans le sillage des physiciens Albert-László Barabási et Duncan Watts. À partir des outils de la physique statistique, ces chercheurs ont voulu mettre au jour les principes universaux qui gouvernent la topologie des réseaux, qu'ils soient naturels, géographiques ou sociaux. Certains de ces physiciens sont devenus sociologues, et conduisent leurs recherches au sein des grandes entreprises du numérique, à l'image de Duncan Watts chez Microsoft.

Bien que quelques sociologues aient précocement traité des corpus textuels comme des réseaux[3], la démarche est devenue plus fréquente avec l'essor du *web* social. Le fait de pouvoir accéder à des volumes importants d'inscriptions textuelles et d'informations relationnelles sur les locuteurs (sous la forme de liens ou de citations) a donné à ces chercheurs issus des sciences dures l'opportunité de s'emparer des objets de la sociologie – les phénomènes d'imitation, la constitution de collectifs en ligne, etc. Et ce d'autant plus que la disponibilité de certaines mesures issues de la linguistique computationnelle leur a permis d'adapter leurs métriques à des corpus textuels.

Parmi les travaux notables, citons celui de Jure Leskovec et Jon Kleinberg qui a porté sur un objet classique de la sociologie des médias : le cycle de l'information (Leskovec *et al.*, 2009). Ces informaticiens ont mis en place une méthode algorithmique fondée sur la théorie des graphes, qui vise à identifier automatiquement, et à une large échelle, les phrases similaires qui se diffusent dans les médias sur une période de temps donnée. Exploitant un corpus de 90 millions de documents publiés par l'ensemble des médias américains pendant la campagne présidentielle de 2008, ils mettent au jour plusieurs phénomènes que les sociologues ne parvenaient alors pas à saisir de façon quantitative – en particulier la dynamique de diffusion des sujets entre les médias traditionnels et les *blogs*. Plus récemment, Eytan Bakshy, chercheur en analyse de réseaux chez Facebook, a voulu savoir dans quelle mesure les

3. Au début des années 1980, Michel Callon et ses collègues du Centre de sociologie de l'innovation ont développé la méthode des « mots associés » destinée à identifier des fronts de recherche à partir de publications scientifiques (Callon *et al.*, 1983).

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)

utilisateurs de Facebook étaient soumis à des informations contraires à leurs préférences politiques (Bakshy *et al.*, 2015). À partir d'un corpus de dix millions d'inscrits, il montre que, plus que l'algorithme de Facebook, c'est le réseau d'amis de l'utilisateur et ses propres choix qui expliquent l'homogénéité des informations auxquelles il est soumis.

Les modèles thématiques

Dans cet éventail des techniques émergentes, la modélisation thématique (*topic models*) est celle dont l'usage se développe le plus rapidement parmi les sociologues[4]. Ces modèles rassemblent un ensemble de méthodes d'apprentissage qui visent à révéler la structure thématique d'un corpus de textes[5]. Ils s'inscrivent dans une tradition informatique plus ancienne – la recherche d'information – qui permettait déjà, depuis les années 1990, de décrire les documents dans un espace vectoriel de bien plus faible dimension que la taille du vocabulaire. Les modèles thématiques s'appuient sur une double hypothèse. Premièrement, ils postulent qu'il existe un nombre fini de thèmes ou sujets à l'intérieur d'un corpus textuel : à chaque thème correspond une certaine distribution des mots du vocabulaire. Deuxièmement, ils postulent que les documents sont eux-mêmes composés d'un mélange de plusieurs thèmes. Ces modèles thématiques visent justement, à partir des données empiriques, à inférer conjointement les mots ou expressions qui composent les thèmes, et la façon dont chaque sujet se distribue dans l'ensemble des documents.

Le principal créateur de la méthode est David Blei, chercheur en informatique de l'université de Princeton et maintenant à Columbia. Selon lui, les modèles thématiques ont vocation à être utilisés dans de nombreuses disciplines : « En plus des applications dans des sciences comme la génétique ou les neurosciences, on peut imaginer que les modèles thématiques viennent au service de l'histoire, de la sociologie, de la linguistique, des sciences politiques, du droit, de la littérature comparée, et tout autre domaine où les textes constituent un objet d'étude principal. » (Blei, 2012, p. 84). De fait, comme pour de nombreux algorithmes d'apprentissage, ces modèles thématiques ont fait l'objet d'applications dans des domaines aussi variés que l'analyse du génome (La Rosa *et al.*, 2015), la reconnaissance d'images (Cao et Fei-Fei, 2007) ou l'étude des *tweets* (Ramage *et al.*, 2010). Depuis la fin des années 2000, cette méthode est de plus en plus utilisée par les chercheurs en gestion, sociologie, science politique, économie et histoire.

L'objectif des chercheurs impliqués dans le développement de ces algorithmes – à commencer par D. Blei lui-même – n'a jamais été de concurrencer les sciences sociales. L'ambition affirmée est plutôt d'offrir aux chercheurs un outil permettant d'identifier rapidement la structure thématique d'un corpus de textes dont le volume rend impossible toute lecture humaine. Bien qu'il ait lui-même appliqué son modèle[6] à plusieurs corpus intéressants pour les sociologues – l'ensemble des articles publiés dans le *Yale Law Journal* ou dans la revue *Science* –, sa démarche visait plutôt à

4. Une centaine d'articles ayant recours à la modélisation thématique ont été publiés dans la sociologie au sens large (science politique, communication, anthropologie, études culturelles, psychologie sociale) (*source* : *Web of Science*).

5. La librairie Mallet, disponible sur Java, R et Python, est la plus utilisée aujourd'hui.

6. Celui-ci se nomme *LDA* pour « Allocation de Dirichlet Latente ».

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)

montrer l'intérêt documentaire que cela pouvait représenter pour des chercheurs en sciences sociales (Blei, 2012).

Les méthodes de plongement de mots

Les algorithmes de « plongement de mots » ou de « plongement lexical » (*word embedding*) forment la classe la plus récente parmi les méthodes caractéristiques des *big data*. Née dans les laboratoires de Google, cette famille d'algorithmes s'appuie sur des techniques d'apprentissage automatique appliquées à de très larges corpus de données souvent issus du *web*. À partir de ces corpus, les algorithmes apprennent la « position » de chaque mot du vocabulaire dans un espace vectoriel limité à quelques centaines de dimensions. Le processus d'apprentissage garantit que deux entités lexicales partageant les mêmes contextes se retrouveront à proximité l'une de l'autre dans l'espace euclidien dans lequel elles sont plongées.

Dans ce domaine de recherche récent, des informaticiens ont conçu plusieurs applications, dont certaines font écho aux objets de la sociologie. Ils utilisent en effet ces algorithmes – et notamment le plus connu, appelé *word2vec* (Mikolov *et al.*, 2013) – pour identifier automatiquement des relations d'analogie dans de vastes corpus textuels. Ces analogies peuvent être de nature syntaxique (singulier/pluriel, masculin/féminin) ou de nature sémantique (les capitales peuvent être automatiquement identifiées à partir des noms de pays). Certains chercheurs en informatique s'en servent pour identifier les stéréotypes racistes ou sexistes dans de gros corpus textuels, et les mesurer systématiquement (Caliskan-Islam *et al.*, 2017).

Pour le moment, quelques chercheurs en sciences sociales les utilisent pour analyser des corpus textuels[7]. À titre d'exemple, un historien de l'Antiquité a collaboré avec un linguiste computationnel pour explorer un corpus composé de 11 261 textes en latin couvrant deux millénaires (Bjerva et Praet, 2015). Une fois calculé un plongement de mots pour l'ensemble du vocabulaire, ils ont mesuré la connexion entre concepts (modernité, liberté, romanité, etc.) et personnages historiques (Cassiodorus, Agapetus, Theodora, etc.). En interprétant les distances relatives des mots et des concepts, les chercheurs mettent en évidence la façon dont Théodoric est parvenu à se détacher des origines gothiques de son pouvoir pour s'imposer comme le légataire de l'empire romain en Italie.

La possibilité d'avoir accès à des corpus textuels non seulement plus volumineux, mais aussi plus profondément liés à un registre plus large d'activités sociales a donc correspondu à l'émergence de nouvelles techniques d'analyse textuelle. Au-delà de leur hétérogénéité, ces techniques quantitatives ont en commun d'avoir été conçues dans des mondes académiques ou industriels éloignés des sciences sociales. Si elles en viennent à partager des objets et des ambitions communes avec la sociologie, c'est soit qu'elles adoptent une démarche de connaissance proche de la sociologie (à l'image de certains chercheurs en linguistique computationnelle), soit que leur ambition industrielle les conduise à concevoir des dispositifs informatiques destinés à qualifier des activités sociales. Nous allons voir maintenant comment des sociologues en font aujourd'hui une méthode d'enquête, au même titre que les méthodes plus traditionnelles.

7. Le nombre de publications en sociologie, même entendu dans un sens large, se limite à quelques unités.

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)

Des techniques mises au service de l'enquête sociologique

Depuis la fin des années 2000, un nombre croissant de sociologues ont ajouté certaines de ces techniques émergentes à leur répertoire de méthodes. Il s'agit aussi bien de sociologues qui occupent des positions importantes dans la discipline (P. DiMaggio, N. Fligstein), dont certains sont déjà familiers de l'analyse quantitative des textes (Peter S. Bearman, Francis Chateauraynaud), que de jeunes chercheurs formés en sociologie, ou encore de sociologues qui viennent de disciplines plus lointaines (physique, statistique, informatique, etc.). Soit ces sociologues effectuent un travail d'importation de méthodes conçues en dehors des sciences sociales, soit ils mettent au point des méthodes originales en collaborant étroitement avec des chercheurs en informatique. Les corpus textuels qu'ils analysent sont très variés, à la fois par leur taille (de quelques milliers à plusieurs milliards de caractères) et par leur nature (articles de presse, documents administratifs, *tweets*, etc.).

En parcourant leurs travaux, nous allons voir que ces chercheurs parviennent à mettre ces techniques au service de l'enquête sociologique, soit en étudiant de nouveaux objets, soit en étendant le domaine de validité de leurs analyses, soit encore en mettant au jour de nouvelles régularités sur des objets déjà connus. Pour ces chercheurs, ces nouvelles méthodes d'analyse textuelle apparaissent comme un moyen de poursuivre trois grandes perspectives : 1) étudier des cadres d'interprétation ; 2) reconstituer des stratégies d'acteurs ; 3) analyser l'énonciation comme un acte social.

Étudier des cadres d'interprétation

Parmi l'ensemble des nouvelles méthodes que nous avons évoquées, ce sont les algorithmes de « modélisation thématique » (*topic modeling*) qui ont d'abord été importés en sociologie (Mohr et Bogdanov, 2013). Un certain nombre de sociologues, parmi les plus reconnus, y ont vu un moyen d'identifier des « cadres d'interprétation » dans des corpus textuels volumineux. S'inscrivant dans la filiation de David Snow, ils veulent repérer des ensembles d'éléments discursifs qui suggèrent une interprétation particulière d'une personne, d'un évènement, d'une pratique ou d'une situation (Snow, 2004).

En collaboration avec D. Blei – que nous évoquions plus haut –, P. DiMaggio et son collègue Manish Nag ont ainsi réalisé une enquête sur le financement des arts par le gouvernement fédéral américain (DiMaggio *et al.*, 2013). Le point de départ de leur enquête concerne le changement de perception publique du financement fédéral des arts : alors que celui-ci faisait l'objet d'un large consensus depuis les années 1960, il est contesté dans les années 1990, ce qui conduit à la diminution des aides fédérales aux artistes. À partir d'un algorithme de recherche, les chercheurs identifient les articles de presse publiés sur la période 1986-1997 dans les grands journaux américains, et qui comportent une mention du programme fédéral de soutien aux arts. Ils constituent ainsi un corpus de 8 000 articles, soit 3 millions de mots. L'application du modèle de D. Blei génère douze listes de termes qui correspondent à autant d'ensembles thématiques distincts (des « *topics* »), que les sociologues rattachent à plusieurs cadres d'interprétation : les controverses autour de l'attribution des bourses par le gouvernement (topic n° 2) ; les délibérations du Congrès américain (topic n° 5) ; le rôle de l'art dans le développement des villes (topic n° 1), etc. Ils mesurent et représentent l'évolution de l'importance relative de chaque cadre dans le temps. Ce qui leur permet d'identifier le moment où certains cadres sont plus ou

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)

moins mobilisés dans le débat public. Ils révèlent ainsi que c'est au printemps 1989, peu de temps après l'élection de Georges Bush père, que le financement fédéral des arts cesse d'être célébré pour être vigoureusement contesté. En construisant des coefficients statistiques, les auteurs montrent également que certains journaux mobilisent prioritairement certains cadres lorsqu'ils abordent le sujet du financement des arts.

D'autres sociologues ont eu recours aux modèles thématiques afin d'identifier des cadres d'interprétation dans des corpus moins volumineux, de façon à donner plus de force à leur argument. C'est le cas de N. Fligstein et ses collègues lorsqu'ils enquêtent sur la crise financière de 2008 (Fligstein *et al.*, 2017). Ils veulent comprendre pourquoi la Federal Reserve – organisme en charge de la politique monétaire aux États-Unis – n'a pas perçu l'importance de la crise alors que se multipliaient les signes d'effondrement du système financier. Le corpus qu'ils construisent est assez classique et de taille modeste : il s'agit des comptes rendus de 72 réunions du *Federal Open Market Committee* qui se sont tenues sur la période 2000-2008. En s'appuyant sur la théorie de la « construction du sens » dans les organisations (Weick, 1995), ils utilisent la méthode des modèles thématiques pour identifier une dizaine de cadres interprétatifs de la crise financière qui est alors en train d'apparaitre, et saisir leur évolution dans le temps. Leur hypothèse est que la discussion des experts est dominée par les concepts macroéconomiques, ce qui les aurait rendus aveugles aux liens entre le marché immobilier et les marchés financiers. Or la distribution statistique des différents cadres montre précisément que la bulle immobilière est totalement absente des discussions avant l'été 2005. La modélisation est ici utilisée pour confirmer une hypothèse générale, N. Fligstein et ses collègues se lançant ensuite dans une lecture minutieuse du corpus. C'est ainsi qu'ils montrent que les experts de la Federal Reserve réfutent longtemps l'existence d'une bulle immobilière, jugeant que l'augmentation des prix serait justifiée par les fondements macroéconomiques.

L'analyse de sentiment commence également à être utilisée par quelques sociologues dans une perspective proche. La contribution de René Flores est ici particulièrement remarquable. Dans un article récent, il étudie l'effet de l'adoption d'une loi anti-immigration en Arizona sur l'opinion publique, à partir d'un corpus de 250 000 *tweets* (Flores, 2017). Il ne s'agit pas ici de mettre au jour des cadres d'interprétation, mais plutôt de mesurer la polarité d'un grand nombre de *tweets*, selon qu'ils sont plus ou moins opposés ou favorables à l'immigration. Le sociologue mesure, avant et après l'adoption de la loi, de quelle façon évolue la position des auteurs de *tweets* à l'égard de l'immigration. Il montre ainsi que la loi a eu pour effet d'augmenter le nombre de *tweets* abordant le sujet de l'immigration, et que les prises de parole opposées à l'immigration y sont de plus en plus importantes. Mais ce n'est pas que les auteurs de *tweets* ont changé d'avis sur le sujet : ce sont les personnes déjà mobilisées contre l'immigration qui prennent de plus en plus la parole. L'analyse de sentiment est ici utilisée pour rendre compte d'une mobilisation en ligne, difficilement appréhendable par les méthodes sociologiques traditionnelles.

Reconstituer des stratégies d'acteurs

S'identifiant au domaine émergent des « sciences sociales computationnelles », d'autres sociologues s'inscrivent dans une démarche très différente. Ils collaborent étroitement avec des informaticiens pour mettre au point des méthodes d'enquête à partir de techniques existantes mais peu utilisées en sociologie (traitement du langage naturel, clusterisation, analyses de réseaux, simulation multi-agents, etc.). Pour ces

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)

chercheurs, ces méthodes offrent l'opportunité d'identifier de façon plus profonde, dans des corpus textuels, les stratégies des individus et des organisations.

James Evans incarne cette approche de façon particulièrement stimulante. Ce sociologue à l'université de Chicago collabore étroitement avec des mathématiciens, des informaticiens, des biologistes, des statisticiens et d'autres sociologues pour mettre au point des méthodes innovantes d'analyse textuelle. En 2015, lui et ses collègues publient un article dans l'*American Sociological Review* qui porte sur les stratégies de recherche des scientifiques (Foster *et al.*, 2015). Leur corpus est particulièrement massif, puisqu'il contient l'ensemble des résumés de chimie biomédicale parus sur la base MEDLINE entre 1934 et 2008 – soit plus de 6,4 millions de résumés scientifiques. La question des chercheurs est tout à fait classique : quelles stratégies de recherche les scientifiques choisissent-ils ? La méthode l'est beaucoup moins, puisqu'elle consiste à identifier de façon automatique les entités chimiques mentionnées dans les résumés ainsi que les relations entre ces entités. L'état de la connaissance est ainsi modélisé sous la forme d'un réseau de co-occurrences entre composés chimiques déjà liés au sein d'un même résumé. De cette façon, les auteurs entendent reconstituer « l'espace des chemins de recherche » empruntés par les scientifiques depuis soixante-dix ans. Ils distinguent plusieurs stratégies de recherche, selon que le chercheur propose une relation qui est nouvelle ou qui est déjà connue ; selon que les entités chimiques ainsi reliées sont elles-mêmes déjà connues ou qu'elles n'ont pas été explorées jusque-là ; et selon que ces entités appartiennent à des domaines de recherche communs ou différents. Ils développent ensuite un modèle probabiliste qui suit une logique de choix rationnel. Ils postulent en effet que chaque scientifique choisit une stratégie de recherche en considérant l'ensemble des liens qui sont possibles à un temps *t*, ainsi que l'ensemble des liens qui ont déjà été explorés par les chercheurs du domaine considéré. À partir de ce modèle, J. Evans et ses collègues mesurent la distribution des différentes stratégies des chercheurs dans le temps. Ils montrent ainsi que les stratégies les plus risquées (où le chercheur propose une relation entre des entités inconnues et distantes) sont beaucoup moins fréquemment observées que les stratégies moins risquées (où des relations entre entités déjà testées), et que ce rapport est resté stable depuis trente ans. Ils montrent également que les stratégies les plus risquées sont récompensées par un plus grand nombre de citations et de prix, mais qu'une stratégie moins risquée permet de maximiser le nombre de citations.

D'autres sociologues ont mis en œuvre des démarches similaires, en collectant des matériaux issus des réseaux sociaux. C'est le cas de Christopher Bail et ses collègues, de l'université de Duke, qui étudient les stratégies de communication des organisations militantes sur Facebook (Bail *et al.*, 2017). Leur objectif est de montrer que certaines façons d'utiliser les réseaux sociaux ont pour effet de stimuler la conversation du public. Les auteurs collectent l'ensemble des *posts* et commentaires publiés sur les pages Facebook de plus de 200 associations spécialisées dans le domaine de l'autisme et du don d'organes. À partir de plusieurs techniques automatiques d'analyse du langage, C. Bail et ses collègues qualifient ces publications selon qu'elles relèvent d'un registre cognitif ou émotionnel. Ils montrent d'abord, à une échelle globale, que les registres cognitifs et émotionnels dominent alternativement les espaces conversationnels de ces associations sur Facebook. Leur analyse montre ensuite qu'une association suscite un plus grand intérêt du public en mobilisant davantage un registre cognitif quand c'est le registre émotionnel qui domine (et inversement). Au moyen d'un questionnaire envoyé aux associations, les chercheurs collectent un grand nombre de variables sociologiques (taille de l'association, etc.),

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)

et en 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.

Des chercheurs en économie publique ont également mobilisé ces méthodes pour étudier les stratégies des organisations. C'est le cas d'une recherche particulièrement innovante conduite par l'économiste Julia Cagé, en collaboration avec des informaticiens de l'Institut national de l'audiovisuel (Cagé *et al.*, 2017). Avec l'appui de l'INA, les chercheurs ont collecté l'ensemble des contenus produits et mis en ligne par les médias français d'information politique et générale pendant l'année 2013. Soit un corpus de plus de 2,5 millions de documents issus de sites *web*, de journaux locaux et nationaux, de stations de radio, de chaines de télévision et d'agences de presse. Les chercheurs utilisent alors plusieurs algorithmes de traitement du langage naturel. À l'aide d'un algorithme de détection d'évènements, qui repère des évènements communs à partir d'un ensemble de mots similaires, les chercheurs identifient près de 25 000 évènements couverts par les médias français en 2013. Ils utilisent ensuite un algorithme de détection de plagiat, à partir duquel ils montrent que 64 % du contenu publié au sein de ces évènements est copié d'un autre média. Tout ceci permet à J. Cagé et ses collègues de mettre au jour les variables qui expliquent pourquoi une organisation médiatique est susceptible de financer la production d'informations originales plutôt que de plagier ses concurrents. À l'aide d'une série de régressions, les chercheurs montrent ainsi que si l'embauche de journalistes supplémentaires affecte positivement la production d'informations originales, les conséquences en termes d'audience sont minimes.

Pour un bon nombre de sociologues, ou de chercheurs proches de la sociologie, les méthodes issues de l'intelligence artificielle aident donc à mieux identifier les stratégies des acteurs et des organisations.

Analyser l'énonciation comme un acte social

Un dernier ensemble de sociologues privilégient certaines de ces méthodes pour saisir l'énonciation comme un acte social. Le point commun de leur démarche est de prendre appui sur des modélisations élaborées de l'énonciation, qui empruntent à des approches sémiotiques. Autrement dit, ils identifient plusieurs catégories de termes – selon qu'ils désignent des agents, des actes, des scènes ou des motifs par exemple – entre lesquelles ils reconstituent les relations. Cela leur permet de mettre au jour non pas des ensembles thématiques ou des stratégies, mais un ensemble d'actions accomplies par le locuteur lorsqu'il prend la parole. Une telle perspective s'inscrit dans la continuité de travaux sociologiques qui avaient importé, il y a plusieurs dizaines d'années, les approches en termes de linguistique de l'énonciation – au sens de Mikhaïl Bakhtine en France (Boltanski *et al.*, 1984) ou de Kenneth Burke aux États-Unis (Franzosi, 1989).

Comparés aux travaux évoqués précédemment, les corpus analysés sont ici plutôt traditionnels – il s'agit de récits biographiques, de rapports publics, plus rarement de discussions en ligne – et de taille assez modeste. L'originalité vient plutôt du traitement effectué sur le matériau textuel. Le travail de Peter Bearman et Katherine Stovel, sociologues à l'université de Columbia et à l'université de Washington, est particulièrement intéressant de ce point de vue (Bearman et Stovel, 2000). Ils ont collecté 600 récits rédigés en 1934 dans le cadre d'un concours organisé par le parti nazi, qui avait demandé à certains de ses compatriotes d'expliquer comment ils étaient devenus nazis. Les deux sociologues veulent expliquer comment des citoyens ordinaires en

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)

sont venus à se déclarer nazis avant même l'arrivée au pouvoir d'Hitler, alors même qu'ils rencontraient l'hostilité de leur entourage. Les sociologues divisent d'abord le corpus en deux parties (la partie des récits qui concerne le fait de « devenir nazi » ; celle qui a trait au fait d'« être nazi »). Ils identifient ensuite tous les « éléments » présents dans le récit – des évènements historiques, des évènements personnels, des opinions ou des émotions – dont ils repèrent automatiquement de quelle façon ils sont reliés les uns aux autres. Chaque récit est ainsi analysé comme un réseau, dont P. Bearman et K. Stovel identifient les régularités structurelles – autrement dit, la façon dont certains éléments sont systématiquement associés les uns aux autres ou dissociés les uns des autres. C'est ainsi qu'ils parviennent à des résultats très stimulants. Ils montrent d'abord que, dans la partie du corpus « devenir un nazi », les éléments cognitifs font le lien entre les éléments d'action ; mais que dans la partie du corpus « être un nazi », les éléments cognitifs ne relient plus les actions. Ce qu'ils interprètent ainsi : le fait d'être un nazi induit l'absence de réflexivité sur soi-même. Ils montrent ensuite que les réseaux narratifs font de moins en moins de place aux institutions traditionnelles (l'église, l'école, le travail, etc.) et aux relations familiales et amicales – ce qui signifie que la trajectoire nazie implique la séparation avec les relations sociales habituelles. Les deux sociologues montrent enfin qu'en devenant nazis les individus adoptent une interprétation binaire de la réalité – les nœuds « chaos » et « ordres » devenant plus centraux dans les récits. On voit donc comment l'analyse du matériau textuel permet ici, via les réseaux narratifs, de caractériser des transformations identitaires.

Les travaux de John Mohr, sociologue des réseaux à l'université de Californie, s'inscrivent dans la même approche. En collaboration avec des sociologues et un informaticien, il propose une méthode pour identifier l'action du gouvernement américain en matière de défense à travers des textes définissant la stratégie de sécurité nationale des États-Unis depuis 2010 (Mohr *et al.*, 2013). L'originalité de la démarche ne réside pas ici dans la taille du corpus – celui-ci comporte onze textes d'une dizaine de pages chacun. Prenant appui sur la linguistique de l'énonciation, les chercheurs proposent une modélisation des textes suivant le schéma de Kenneth Burke donnant à voir des actes, des scènes, des agents, des formes d'action et des motifs. Ils utilisent ainsi des algorithmes d'extraction d'entités nommées pour identifier les agents (des pays, des organisations, etc.), des parseurs sémantiques pour leur associer des actes (soutenir, améliorer, partager, etc.) et une modélisation thématique pour dégager les scènes (conflit, énergie, etc.) qui offrent autant d'éléments de contexte à ces actes. Ils identifient ensuite de façon automatique les réseaux qui lient, pour chaque scène, les agents et les actes. Ce qui leur permet notamment de montrer à quel point le 11 septembre 2001 a profondément modifié la façon dont le gouvernement américain construit sa stratégie de défense. Les questions des « armes de destruction massive » et de la « sécurité intérieure » occupent dès lors une place majeure quelles que soient les scènes concernées, et aussi bien sous la présidence de Georges W. Bush que celle de Barack Obama.

De l'autre côté de l'Atlantique, plusieurs sociologues cherchent également à étudier l'énonciation comme action sociale, en mobilisant les techniques algorithmiques. C'est le cas de Francis Chateauraynaud (2003), dont les travaux précurseurs sur le logiciel Prospéro reposaient déjà sur une modélisation du langage inspiré de la linguistique pragmatique. Dans ses travaux plus récents, le sociologue analyse les controverses portant sur les nanotechnologies (Chateauraynaud, 2014). Analysant un corpus de 5 800 textes en partie issus du *web*, le sociologue produit plusieurs cartes de réseaux afin d'identifier des trajectoires argumentatives spécifiques de la controverse. C'est

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)

ainsi qu'il identifie, pour chaque acteur de la controverse, la façon dont leur argumentation relie les différentes entités (laboratoires, concepts, critiques, etc.). La démarche de Sylvain Parasie et Jean-Philippe Cointet (2012) est assez similaire, lorsqu'ils étudient de quelle manière les participants à des forums politiques municipaux en ligne construisent leurs argumentations. En s'appuyant sur un modèle actantiel de la prise de parole en public, ils montrent que l'expression des citoyens diffère profondément selon la taille de la ville dans laquelle ils résident. Ainsi, dans les petites communes, les locuteurs s'adressent de façon individuelle les uns aux autres, expriment un lien personnel avec le maire, mais sans dévoiler leur identité. Alors que, dans les communes moyennes, les locuteurs s'adressent à l'opinion et non à des individus, empruntant le registre de la dénonciation et de la révélation. Ici encore, les textes sont envisagés comme des réseaux, et le recours aux techniques algorithmiques repose sur une modélisation de l'énonciation.

L'ensemble des travaux que nous venons d'évoquer témoigne d'une véritable intégration de ces méthodes algorithmiques à l'enquête sociologique. Loin de jouer un rôle superficiel dans l'enquête, ces dispositifs d'analyse textuelle sont ici mobilisés de plusieurs façons différentes. Soit il s'agit d'étudier de nouveaux objets qui mettent à mal les méthodes d'enquête traditionnelles (des mobilisations en ligne) ; soit il s'agit d'étendre la portée empirique de l'enquête en permettant de couvrir un très grand nombre d'occurrences sans faire appel à un quelconque échantillonnage (l'ensemble des recherches dans un domaine scientifique donné) ; soit encore il s'agit de jeter un regard neuf sur de petits corpus, en procédant à des analyses originales. Prenant appui sur ce panorama, nous allons maintenant discuter des conditions auxquelles l'utilisation de ces techniques émergentes doit se plier pour contribuer au raisonnement sociologique.

À quelles conditions peut-on élargir l'éventail des méthodes ?

L'accès à de nouvelles sources de données textuelles, associé à l'émergence de nouvelles techniques algorithmiques, modifie l'écologie des disciplines. Des informaticiens, des linguistes et des chercheurs en sciences sociales s'intéressent aux mêmes sources, se retrouvent à partager les mêmes dispositifs. Devant la multiplication des recherches, une question se pose : à quelles conditions le recours à ces techniques algorithmiques, pour analyser des corpus textuels traditionnels ou émergents, contribue-t-il véritablement à l'enquête sociologique ? Le mot « sociologique » ne doit pas ici être pris dans le sens restreint des frontières disciplinaires, mais plutôt pour désigner l'ensemble des recherches qui prétendent rendre compte de l'action sociale (Passeron, 1991 ; Lemieux, 2014). Il s'agit donc de nous demander ce que ces enquêtes récentes – conduites par des sociologues comme par des non-sociologues – nous disent des conditions auxquelles ce type d'enquête doit se plier pour contribuer à une meilleure connaissance de l'action sociale. Ces conditions sont aujourd'hui l'objet d'une exploration collective, qui est loin d'être achevée. Pour autant, trois conditions nous semblent apparaitre, qui portent sur le contexte de production des traces textuelles ; l'intégration à l'enquête de données extérieures au texte lui-même ; et l'ajustement des algorithmes au raisonnement sociologique.

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)

Connaitre le contexte de production des textes

Pour contribuer à la connaissance de l'action sociale, le sociologue doit d'abord connaitre le contexte dans lequel ont été produites les inscriptions textuelles qu'il entend traiter quantitativement. Toutes les recherches que nous avons présentées mettent en données des textes qui n'ont pas été produits dans un objectif de recherche en sciences sociales. On sort donc du cadre classique dans lequel un sociologue récupère les données produites par des institutions statistiques. Comme ont commencé à le faire certains chercheurs (Bastin et Francony, 2017), il est donc indispensable d'objectiver les cadrages opérés par les individus, les organisations et les plateformes numériques qui coproduisent ces inscriptions textuelles. Cette nécessité est particulièrement grande lorsque ces textes sont organisés, stockés et mis en forme par des plateformes numériques (Google, Facebook, Twitter, etc.). Il faut alors comprendre dans quel type de relation s'engagent les usagers de ces plateformes, et de quelle façon ces plateformes orientent les formes de participation, limitent les informations auxquelles le chercheur a accès. Ainsi, le chercheur qui traite un corpus de discussions sur Wikipédia devra connaitre les procédures complexes qui régulent les conflits d'édition, ainsi que les différences de statuts entre les contributeurs. Dans le cas contraire, il court le risque de ne pas voir qu'un élément structurel des relations entre les usagers s'explique par les procédures mises en place par la plateforme.

Cette condition reste vraie, même lorsqu'aucune plateforme numérique n'est impliquée. Souvenons-nous qu'Harold Garfinkel jugeait parfaitement normal que les dossiers médicaux d'une clinique reflètent très mal l'activité de son personnel. C'est que, expliquait-il, le sens des inscriptions ne peut être déterminé par un lecteur si celui-ci ne connait rien de l'organisation : il faut connaitre l'histoire des transactions entre le médecin et son patient, entre celui qui écrit et celui qui lira (Garfinkel, 1967, p. 315-323). De la même façon, il est souvent crucial pour le chercheur de prendre en compte l'horizon interactionnel et organisationnel dans lequel les inscriptions textuelles ont été produites. Ce n'est pas un hasard si les enquêtes les plus intéressantes mobilisent souvent des corpus de textes qui présentent une grande cohérence du point de vue des acteurs eux-mêmes, et qui sont destinés à coordonner les actions d'individus ou d'organisations hétérogènes. C'est le cas de N. Fligstein et ses collègues, qui ont analysé les transcriptions des réunions de l'organisme chargé de la définition de la politique monétaire américaine (Fligstein *et al.*, 2017). C'est également le cas de Nicolas Baya-Laffite et J.-P. Cointet (2014), qui étudient l'évolution thématique des conférences climatiques, à partir d'un corpus composé des *Bulletins de négociation de la terre*. Parce que ces bulletins sont utilisés par les acteurs des négociations climatiques, ils présentent une forte standardisation qui facilite leur traitement algorithmique, et simplifie l'identification de la dynamique thématique des négociations.

Cette première condition incite donc à la plus grande prudence vis-à-vis des approches algorithmiques (notamment lexicométriques) qui mettent en équivalence un grand nombre d'inscriptions textuelles sans que l'on sache quoi que ce soit du contexte dans lequel chaque inscription a été produite.

Ajouter de l'extériorité aux textes

Une deuxième condition stipule que, pour contribuer à la connaissance de l'action sociale, le chercheur ne doit pas se limiter à une analyse strictement internaliste du texte. D'une façon ou d'une autre, il doit mettre en rapport certaines propriétés internes

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)

au texte (thèmes, tournures, styles, entités nommées, etc.) avec d'autres éléments qui sont extérieurs au texte : des collectifs, des ressources, des organisations, des professions, des lieux, etc. De ce point de vue, le panorama des recherches que nous dressons dans cet article appelle deux remarques. D'une part, il apparait que les sociologues ne sont pas les seuls chercheurs à intégrer des informations extérieures au texte. C'est ce que font des chercheurs en informatique ou en linguistique computationnelle (Danescu-Niculescu-Mizil *et al.*, 2012 ; Voigt *et al.*, 2017), ce qui renforce leur capacité à rendre compte d'actions sociales. *A contrario*, certaines enquêtes conduites par des sociologues accordent une place réduite à ces éléments extérieurs au texte – à l'image de James Evans et ses collègues analysant les stratégies de recherche à partir d'un corpus d'articles scientifiques (Foster *et al.*, 2015).

D'autre part, on constate que la combinaison des méthodes d'enquête constitue le moyen le plus sûr pour donner une épaisseur sociale aux inscriptions textuelles traitées quantitativement. Les chercheurs ont recours à des procédés variés. Certains collectent des données socio-démographiques ou économiques auprès des institutions statistiques, et les associent au matériau textuel. C'est le cas de S. Parasie et J.-P. Cointet (2012), lorsqu'ils analysent les échanges sur les forums des communes du Nord-Pas-de-Calais à l'occasion des élections municipales de 2008. À partir de données provenant de l'Insee, ils relient l'expression des internautes à la morphologie de la commune concernée (nombre d'habitants, densité de la commune, etc.). D'autres chercheurs élaborent des questionnaires à destination des acteurs qui prennent la parole sur les plateformes en ligne. Plusieurs sociologues qui analysent des corpus textuels issus de Facebook conçoivent des applications qui fournissent un service aux acteurs – en leur offrant une carte de leur réseau ou une évaluation de leur communication en ligne – en échange de données sur leurs activités ou leurs propriétés sociales (âge, genre, niveau de diplôme ou lieu de résidence) (Bastard *et al.*, 2017 ; Bail *et al.*, 2017). D'autres chercheurs encore tirent parti des métadonnées fournies par les plateformes pour donner de l'épaisseur sociale aux locuteurs – à l'image de R. Flores qui identifie les auteurs de *tweets* qui résident en Arizona et militent contre l'immigration (Flores, 2017), ou de Valérie Beaudouin et Dominique Pasquier (2017) qui isolent des profils de rédacteurs amateurs de critiques de films à partir d'informations issues de la plateforme. On voit ici l'inventivité dont les chercheurs font aujourd'hui preuve pour donner une épaisseur sociale aux inscriptions textuelles. C'est là une étape nécessaire pour intégrer ces méthodes et ces corpus à l'enquête sociologique.

Ajuster les algorithmes à l'enquête sociologique

Une dernière condition, tout aussi importante, concerne la façon dont les algorithmes sont ajustés à l'enquête sociologique. Cet ajustement, nous l'avons dit, est d'autant plus problématique que ces algorithmes sont issus de mondes éloignés des sciences sociales, et qu'ils sont utilisés comme des « briques »[8] dont l'articulation avec le raisonnement sociologique peut s'avérer problématique. La récente enquête de C. Bail et ses collègues (2017) est à ce titre emblématique. Enquêtant sur la capacité des associations à stimuler la conversation sur Facebook, ils ont recours à un

8. Pratiquement, ces briques prennent la forme de librairies – sous R et Python – qui effectuent un traitement spécifique et limité sur le matériau textuel. Ces librairies sont produites par des chercheurs, accessibles en ligne, et combinables les unes avec les autres dans une suite de traitements.

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)

algorithme conçu par des psycholinguistes[9] qui vise à identifier les états mentaux des individus à partir du vocabulaire qu'ils utilisent. Les résultats auxquels ils parviennent reposent intégralement sur cet algorithme, qu'ils utilisent pour identifier automatiquement les styles « cognitifs » ou « émotionnels » des *posts* des associations, et ainsi déterminer la stratégie qui leur permet d'augmenter leur nombre de commentaires sur Facebook.

La plupart des sociologues privilégient néanmoins les algorithmes qui leur laissent une certaine autonomie dans l'élaboration de leurs catégories d'analyse, ou qui permettent des allers et retours entre le calcul et l'analyse. C'est ce qui explique en partie la popularité actuelle des modèles thématiques chez les sociologues. À l'instar de logiciels plus connus en France (Iramuteq), cette famille d'algorithmes laisse en effet une grande place à l'interprétation du chercheur qui s'appuie sur sa propre expertise pour identifier les cadres à partir des ensembles de termes produits de façon automatique. Le succès actuel des modèles thématiques s'explique aussi par le travail de traduction réalisé par P. DiMaggio *et al.* (2013). Le chercheur américain défend l'argument selon lequel ces algorithmes s'ajustent bien au raisonnement sociologique, à la fois parce qu'ils s'appuient sur une conception relationnelle de la signification et qu'ils postulent que les textes ne reflètent pas toujours un point de vue unique, mais sont traversés par plusieurs « voix ».

L'ajustement implique aussi de pouvoir modifier l'algorithme pour le conformer à des hypothèses sociologiques. Ce qui exige de la part du chercheur des compétences informatiques et statistiques peu communes. C'est ce que fait R. Flores (2017) lorsqu'il modifie un algorithme d'analyse de sentiment, faisant en sorte que celui-ci puisse identifier plusieurs entités qui sont jugées par un même locuteur, et tienne compte de la distance entre les mots exprimant un jugement et la cible de ce jugement. À un niveau plus collectif, il est particulièrement important que les sociologues mettent en place des procédures de validation, tout particulièrement pour les algorithmes qui sont les plus fermés. Pour reprendre l'exemple de C. Bail que nous citions plus haut, il est intéressant que celui-ci ait entrepris de valider auprès d'un échantillon de chercheurs la classification des *posts* Facebook selon leur caractère cognitif ou émotionnel (Bail *et al.*, 2017).

Les trois conditions que nous venons de dessiner ne sont pas limitatives, et font aujourd'hui l'objet d'une exploration collective. Mais elles doivent permettre d'envisager l'intégration de ces techniques émergentes à l'éventail des méthodes sociologiques.

*

* *

S'ils n'ont pas attendu ces dernières années pour analyser des corpus textuels de façon quantitative (Chateauraynaud, 2003 ; Demazière *et al.*, 2006), les sociologues sont aujourd'hui confrontés à une situation nouvelle. D'une part, ils ont désormais accès, sous une forme numérique, à des inscriptions textuelles qui sont à la fois volumineuses et de plus en plus associées à des données de nature relationnelle. D'autre part, un grand nombre de techniques informatiques sont apparues dans des mondes très éloignés des sciences sociales, qui offrent à des informaticiens, statisticiens et linguistes l'opportunité de produire des connaissances sur les objets

9. Il s'agit du LIWC (*Linguistic Inquiry Word Count*).

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)

traditionnels de la sociologie. Dans les disciplines voisines de la sociologie, les chercheurs examinent actuellement avec attention de quelle manière ces techniques émergentes peuvent être intégrées à leur répertoire de méthodes (Grimmer et Stewart, 2013 ; Gentzkow *et al.*, 2017). C'est à une telle réflexion que nous avons voulu contribuer à travers cet article, en offrant un panorama critique des recherches sociologiques actuelles.

Notre argument est que, bien que provenant de mondes éloignés des sciences sociales, ces méthodes émergentes de mise en données et de traitement de corpus textuels peuvent être mises au service du raisonnement sociologique. Mais seulement pourvu que plusieurs conditions soient satisfaites, qui concernent le contexte de production des inscriptions textuelles, l'enrichissement des inscriptions textuelles par des informations extérieures, et l'ajustement des algorithmes au raisonnement sociologique. Ces conditions étaient plus facilement respectées dès lors que les sociologues utilisaient des logiciels tels que Prospéro ou Iramuteq, ceux-ci ayant été conçus et ajustés à l'enquête sociologique. Dans le contexte contemporain, le respect de ces conditions devient plus problématique alors que la nature des données, les familles d'algorithmes et les pratiques de recherche se diversifient. C'est pourquoi il est nécessaire de formaliser un ensemble de règles articulées à des pratiques de recherche, et de mettre au point des protocoles de validation partagés. Un chantier important est ouvert, dont dépend la capacité de la discipline à maintenir sa juridiction sur la connaissance de l'action sociale.

Jean-Philippe Cointet

médialab – Sciences Po
27, rue Saint Guillaume
75337 Paris cedex 07

jeanphilippe.cointet@sciencespo.fr

Sylvain Parasie

Laboratoire interdisciplinaire sciences innovations sociétés (LISIS)
Université Paris-Est Marne-la-Vallée
5, boulevard Descartes
77420 Champs-sur-Marne

sylvain.parasie@univ-paris-est.fr

Références bibliographiques

Abbott A., 1988, *The System of Professions. An Essay on the Division of Expert Labor*, Chicago (IL), The University of Chicago Press.

Anderson C., 2008, « The End of Theory: The Data Deluge Makes the Scientific Method Obsolete », *Wired magazine*, 16, 7: *https://www.wired.com/2008/06/pb-theory/* (consulté le 12 juin 2018).

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)

BAIL C. A., TAYLOR W. B., MANN M., 2017, « Channeling Hearts and Minds: Advocacy Organizations, Cognitive-Emotional Currents, and Public Conversation », *American Sociological Review*, *82*, 6, p. 1188-1213.

BAKSHY E., MESSING S., ADAMIC L. A., 2015, « Exposure to Ideologically Diverse News and Opinion on Facebook », *Science*, *348*, 6239, p. 1130-1132.

BASTARD I., CARDON D., CHARBEY R., COINTET J.-P., PRIEUR C., 2017, « Facebook, pour quoi faire ? Configurations d'activités et structures relationnelles », *Sociologie*, *8*, 1, p. 57-82.

BASTIN G., FRANCONY J.-M., 2017, « L'inscription, le masque et la donnée. *Datafication* du web et conflits d'interprétation autour des données dans un laboratoire invisible des sciences sociales », *Revue d'anthropologie des connaissances*, *10*, 4, p. 505-530.

BAYA-LAFFITE N., COINTET J.-P., 2014, « Cartographier la trajectoire de l'adaptation dans l'espace des négociations sur le climat. Changer d'échelle, red(u)ire la complexité », *Réseaux*, 188, p. 159-198.

BEARMAN P. S., STOVEL K., 2000, « Becoming a Nazi: A Model for Narrative Networks », *Poetics*, *27*, 2-3, p. 69-90.

BEAUDOUIN V., PASQUIER D., 2017, « Forms of Contribution and Contributors' Profiles: An Automated Textual Analysis of Amateur on Line Film Critics », *New Media & Society*, *19*, 11, p. 1810-1828.

BERELSON B., LAZARSFELD P. F., 1948, *The Analysis of Communication Content*, Chicago (IL), New York (NY), 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.

BJERVA J., PRAET R., 2015, « Word Embeddings Pointing the Way for Late Antiquity », *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Association for Computational Linguistics, p. 53-57.

BLEI D. M., 2012, « Probabilistic Topic Models », *Communications of the ACM*, *55*, 4, p. 77-84.

BOLTANSKI L., DARRÉ Y., SCHILTZ M.-A., 1984, « La dénonciation », *Actes de la recherche en sciences sociales*, 51, p. 3-40.

BOSCO C., PATTI V., BOLIOLI A., 2013, « Developing Corpora for Sentiment Analysis: The Case of Irony and Senti-TUT », *IEEE Intelligent Systems*, *28*, 2, p. 55-63.

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

BOULLIER D., LOHARD A., 2012, *Opinion mining et Sentiment analysis. Méthodes et outils*, Marseille, OpenEdition Press.

CAGÉ J., HERVÉ N., VIAUD M.-L., 2017, *L'information à tout prix*, Paris, INA.

CALISKAN A., BRYSON J. J., NARAYANAN A., 2017, « Semantics Derived Automatically from Language Corpora Contain Human-Like Biases », *Science*, *356*, 6334, p. 183-186.

CALLON M., COURTIAL J.-P., TURNER W. A., BAUIN S., 1983, « From Translations to Problematic Networks: An Introduction to Co-Word Analysis », *Social Science Information*, *22*, 2, p. 191-235.

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)

Cao L., Fei-Fei L., 2007, « Spatially Coherent Latent Topic Model for Concurrent Segmentation and Classification of Objects and Scenes », *Computer Vision*, iccv 2007, ieee 11th International Conference on ieee, p. 1-8.

Cardon D., 2015, *À quoi rêvent les algorithmes. Nos vies à l'heure des big data*, Paris, Le Seuil.

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

Chateauraynaud F., 2014, « Trajectoires argumentatives et constellations discursives. Exploration socio-informatique des futurs vus depuis le nanomonde », *Réseaux*, 188, p. 121-158.

Chateauraynaud F., Debaz J., 2010, « Prodiges et vertiges de la lexicométrie », *blog « Socio-informatique et argumentation »*, mis en ligne le 23 décembre : *http://socioargu.hypotheses.org/1963* (consulté le 12 juin 2018).

Chenu A., 2008, « Des sentiers de la gloire aux boulevards de la célébrité. Sociologie des couvertures de *Paris Match*, 1945-2005 », *Revue française de sociologie*, *49*, 1, p. 3-52.

Cointet J.-P, 2017, *La cartographie des traces textuelles comme méthodologie d'enquête en sciences sociales*, Habilitation à diriger des recherches, École normale supérieure, soutenue le 6 novembre 2017.

Dagiral É., Parasie S., 2017, « La "science des données" à la conquête des mondes sociaux : ce que le "Big Data" doit aux épistémologies locales » 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. 85-104.

Danescu-Niculescu-Mizil C., Lee L., Pang B., Kleinberg J., 2012, « Echoes of Power: Language Effects and Power Differences in Social Interaction » dans *Proceedings of the 21st International Conference on World Wide Web*, p. 699-708.

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

Desrosières A., 2008, « Analyse des données et sciences humaines : comment cartographier le monde social ? », *Journ@l électronique d'histoire des probabilités et de la statistique*, *4*, 2, p. 11-19.

DiMaggio P. J., Nag M., Blei D. M., 2013, « Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of u.s. Government Arts Funding », *Poetics*, *41*, 6, p. 570-606.

Duric A., Song F., 2012, « Feature Selection for Sentiment Analysis Based on Content and Syntax Models », *Decision Support Systems*, *53*, 4, p. 704-711.

Evans J. A., Aceves P., 2016, « Machine Translation: Mining Text for Social Theory », *Annual Review of Sociology*, *42*, 1, p. 21-50.

Fligstein N., Brundage J. S., Schultz M., 2017, « Seeing Like the Fed: Culture, Cognition, and Framing in the Failure to Anticipate the Financial Crisis of 2008 », *American Sociological Review*, *82*, 5, p. 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*, *123*, 2, p. 333-384.

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)

FOSTER J. G., RZHETSKY A., EVANS J. A., 2015, « Tradition and Innovation in Scientists' Research Strategies », *American Sociological Review*, *80*, 5, p. 875-908.

FRANZOSI R., 1989, « From Words to Numbers: A Generalized and Linguistics-Based Coding Procedure for Collecting Textual Data », *Sociological Methodology*, 19, p. 263-298.

GARFINKEL H., 2007, « De "bonnes" raisons organisationnelles pour de "mauvais" dossiers cliniques » dans H. GARFINKEL, *Recherches en ethnométhodologie*, intro. M. BARTHÉLÉMY, L. QUÉRÉ, Paris, Presses universitaires de France, p. 297-323 [1re éd. 1967].

GENTZKOW M., KELLY B. T., TADDY M., 2017, « Text as Data », Cambridge (MA), National Bureau of Economic Research, *NBER Working Paper 23276*.

GRIMMER J., STEWART B. M., 2013, « Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts », *Political Analysis*, *21*, 3, p. 267-297.

GUERRINI A., 2011, « Analyzing Culture with Google Books: Is It Social Science? », *PacificStandard*, August 7: *https://psmag.com/economics/culturomics-an-idea-whose-time-has-come-34742* (consulté le 12 juin 2018).

KIM S., LI F., LEBANON G., ESSA I., 2013, « Beyond Sentiment: The Manifold of Human Emotions », *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, p. 360-369.

KITCHIN R., 2014, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*, London, Sage.

KLINGENSTEIN S., HITCHCOCK T., DEDEO S., 2014, « The Civilizing Process in London's Old Bailey », *Proceedings of the National Academy of Sciences*, *111*, 26, p. 9419-9424.

LA ROSA M., FIANNACA A., RIZZO R., URSO A., 2015, « Probabilistic Topic Modeling for the Analysis and Classification of Genomic Sequences », *BMC Bioinformatics*, *16*, 6, p. 2-9.

LAZER D., PENTLAND A., ADAMIC L. *et al.*, 2009, « Computational Social Science », *Science*, *323*, 5915, p. 721-722.

LEBART L., SALEM A., 1988, *Analyse statistique des données textuelles. Questions ouvertes et lexicométrie*, Paris, Dunod.

LEMIEUX C., 2014, « Étudier la communication au XXIe siècle. De la théorie de l'action à l'analyse des sociétés », *Réseaux*, 184-185, p. 279-302.

LESKOVEC J., BACKSTROM L., KLEINBERG J., 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.

MICHEL J.-B., KUI SHEN Y., PRESSER AIDEN A. *et al.*, 2010, « Quantitative Analysis of Culture Using Millions of Digitized Books », *Science*, *331*, 6014, p. 176-182.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S., DEAN J., 2013, « Distributed Representations of Words and Phrases and their Compositionality », *Advances in Neural Information Processing Systems*, p. 3111-3119.

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)

MITCHELL A., HITLIN P., 2013, « Twitter Reaction to Events Often at Odds with Overall Public Opinion », Pew Research Center, March 4: *http://www.pewresearch.org/2013/03/04/twitter-reaction-to-events-often-at-odds-with-overall-public-opinion/* (consulté le 12 juin 2018).

MOHR J. W., BOGDANOV P., 2013, « Topic Models: What they Are and Why they Matter », *Poetics*, *41*, 6, p. 545-569.

MOHR J. W., WAGNER-PACIFICI R., BREIGER R. L., BODGANOV P., 2013, « Graphing the Grammar of Motives in National Security Strategies: Cultural Interpretation, Automated Text Analysis and the Drama of Global Politics », *Poetics*, *41*, 6, p. 670-700.

MORETTI F., 2013, *Distant Reading*, London, Verso.

MORETTI F., PESTRE D., 2015, « Bankspeak. The Language of World Bank Reports », *New Left Review*, 92, p. 75-99.

O'CONNOR B., BALASUBRAMANYAN R., ROUTLEDGE B. R., SMITH N. A., 2010, « From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series », *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, *11*, 122-129, p. 1-2.

PANG B., LEE L., 2008, « Opinion Mining and Sentiment Analysis », *Foundations and Trends in Information Retrieval*, *2*, 1-2, p. 1-135.

PANG B., LEE L., VAITHYANATHAN S., 2002, « Thumbs up? Sentiment Classification Using Machine Learning Techniques », *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, 10, p. 79-86.

PARASIE S., COINTET J.-P., 2012, « La presse en ligne au service de la démocratie locale. Une analyse morphologique de forums politiques », *Revue française de science politique*, *62*, 1, p. 45-70.

PASSERON J.-C., 1991, *Le raisonnement sociologique. L'espace non-poppérien du raisonnement naturel*, Paris, Nathan.

RAMAGE D., DUMAIS S. T., LIEBLING D. J., 2010, « Characterizing Microblogs with Topic Models », *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Mediap*, p. 130-137.

REINERT M., 1993, « Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars », *Langage et société*, 66, p. 5-39.

RUIZ P., POIBEAU T., 2015, « Combining Open Source Annotators for Entity Linking through Weighted Voting », *Joint Conference on Lexical and Computational Semantics*, June, Denver (USA), p. 211-215.

SAVAGE M., BURROWS R., 2007, « The Coming Crisis of Empirical Sociology », *Sociology*, *41*, 5, p. 885-899.

SNOW D. A., 2004, « Framing Processes, Ideology and Discursive Fields » dans D. A. SNOW, S. A. SOULE, H. KRIESI (eds.), *The Blackwell Companion to Social Movements*, Chichester, Blackwell Publishing, p. 380-412.

VOIGT R., CAMP N. P., PRABHAKARAN V. *et al.* 2017, « Language from Police Body Camera Footage Shows Racial Disparities in Officer Respect », *Proceedings of the National Academy of Sciences*, *114*, 25, p. 6521-6526.

WEICK K. E., 1995, *Sensemaking in Organizations*, Thousand Oaks (CA), Sage.

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)

ABSTRACT

What Big data does to the sociological analysis of texts?
A review of recent research

Since the 2000s, new techniques of text analysis have emerged at the crossroads of computer science, artificial intelligence and natural language processing. Although they were developed independently of any sociological theory, these methods are now being used by researchers—sociologists and non-sociologists alike—to produce new knowledge of the social domain exploiting the massive volume of textual materials now available. By providing an overview of recent sociological investigations that are based on quantitative analyses of textual corpora, this article identifies three conditions under which these approaches can be a resource for sociological inquiry. The three conditions that emerge from our analysis concern: 1) knowledge of the context of production of textual inscriptions; 2) integration of external data into the study itself; 3) the adaptation of algorithms for sociological reasoning.

Key words. BIG DATA – DIGITAL TRACES – TEXTUAL CORPORA – LEXICOMETRY – SEMANTIC NETWORKS – EPISTEMOLOGY – ARTIFICIAL INTELLIGENCE – AUTOMATIC LANGUAGE PROCESSING

ZUSAMMENFASSUNG

Was *Big data* der soziologischen Textanalyse antut
Eine kritische Überschau der zeitgenössischen Forschung

Seit Beginn der Jahre 2000 erscheinen neue Techniken der Textanalyse an der Kreuzung der Welt der Informatik, der künstlichen Intelligenz und der automatischen Sprachverarbeitung. Obwohl sie außerhalb aller soziologischen Absichten erarbeitet wurden, werden diese Techniken heute in der sowohl soziologischen als auch nicht soziologischen Forschung eingesetzt, um die Kenntnis des Sozialen mithilfe des jetzt verfügbaren umfangreichen Textmaterials zu erneuern. Der Artikel stellt eine Überschau der soziologischen Umfragen zusammen, die sich auf die Verdatung und die quantitative Behandlung der Textkorpora stützt und identifiziert so, unter welchen Bedingungen diese Ansätze eine Ressource für die Sozialforschung darstellen können. Drei Bedingungen zeichnen sich aus unserer Analyse ab: 1) die Kenntnis des Produktionskontexts der Texteinträge: 2) die Integrierung in die Forschung der Daten außerhalb des eigentlichen Textes; 3) die Anpassung der Algorithmen an die soziologische Argumentation.

Schlagwörter. *BIG DATA* – DIGITALE SPUREN – TEXTKORPORA – LEXIKOMETRIE – SEMANTISCHE NETZE – EPISTEMOLOGIE – KÜNSTLICHE INTELLIGENZ – AUTOMATISCHE SPRACHVERARBEITUNG

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)

Resumen

El efecto del *Big data* en el análisis sociológico de textos
Un panorama crítico de las investigaciones actuales

Desde los años 2000 han ido apareciendo nuevas técnicas de análisis textual, en la encrucijada entre la informática, la inteligencia artificial y el tratamiento automatizado de la lengua. Aunque han sido elaboradas fuera de cualquier preocupación sociológica, estas técnicas suelen ser utilizadas por investigadores – tanto sociólogos como no sociólogos – para renovar el conocimiento de lo social sacando partido del volumen considerable de materiales textuales del que se dispone hoy en día. A través de un panorama de las encuestas sociológicas basadas en la puesta en datos y en el tratamiento cuantitativo de corpus textuales, se identifica en este artículo en qué condiciones estos métodos pueden pasar a constituir un recurso para la encuesta sociológica. Las tres condiciones que emergen al final de nuestro análisis tienen que ver con: 1) el conocimiento del contexto de producción de las inscripciones textuales; 2) la integración en la encuesta de datos ajenos al propio texto; 3) la adecuación de los algoritmos al razonamiento sociológico.

Palabras claves. *Big data* – Huellas digitales – Corpus textuales – Lexicometría – Redes semánticas – Epistemología – Inteligencia artificial – Tratamiento automatizado de la lengua

Document téléchargé depuis www.cairn.info - Université de Strasbourg - - 93.19.63.141 - 05/11/2018 10h52. © Presses de Sciences Po (P.F.N.S.P.)