Nouvelles perspectives en sciences sociales
Revue internationale de systémique complexe et d'études relationnelles

Avant-propos

Claude Vautier

Volume 11, Number 1, November 2015

Sur le thème de l'analyse de données textuelles informatisée

URI: https://id.erudit.org/iderudit/1035931ar
DOI: https://doi.org/10.7202/1035931ar

See table of contents

Publisher(s)

Prise de parole

ISSN

1712-8307 (print)
1918-7475 (digital)

Explore this journal

Cite this document

Vautier, C. (2015). Avant-propos. *Nouvelles perspectives en sciences sociales*, *11*(1), 15–23. https://doi.org/10.7202/1035931ar

Avant-propos

Claude Vautier
LEREPS, Université Toulouse 1 – Capitole

La livraison de ce mois de novembre 2015 de la revue *Nouvelles perspectives en sciences sociales* propose une plongée dans un monde qui, sans être récent (les premiers travaux datent des années 1960-1970), connaît depuis une quinzaine d'années une évolution considérable. Les progrès de l'informatique en sont, certes, l'une des causes premières. Mais le développement technologique a généré un profond mouvement de réflexions épistémologiques et méthodologiques (produit-on ainsi des connaissances valables, selon une méthode robuste, c'est-à-dire possédant une fiabilité…?).

Au-delà de ces questions très générales, ces approches nouvelles affrontent des défis multiformes : permettent-elles aux SHS de faire apparaître les caractères construits, historiques, contingents des phénomènes et d'éviter les biais liés à de possibles herméneutiques fragiles, du fait de l'insuffisance des matériaux ou de leur surabondance, projetant l'interprétation en première ligne de la recherche des résultats, ce qui n'est pas en soi scandaleux mais augmente le risque d'abus d'inférences fautives, de trop grande subjectivité dans le traitement du qualitatif comme du quantitatif?

Les multiples logiciels contemporains d'analyse *partiellement*[1] automatisée des textes forment une jungle dans laquelle le pro-

[1] C'est l'un des points d'achoppement des débats et il est important d'insister sur le fait qu'aucun logiciel de « textométrie » (ou de quelque autre nom qu'on

meneur, amateur non aguerri, a bien des chances de se perdre. Mal guidé, il risque de tomber dans de multiples pièges, ce que les détracteurs (de moins en moins nombreux) de la méthode ne manquent pas de souligner : perte de vue de la question, incompréhension des modes de calcul utilisés par le(s) logiciel(s), donc risque de conclusions hors de propos et d'utilisation aléatoire des logiciels (« au petit bonheur », en fonction de la disponibilité de l'outil), risque de leurre lié à l'interprétation des catégories construites par le logiciel, des représentations (graphes multiples et suggestifs)… Bien d'autres risques existent, qui sont pointés par les familiers de ces méthodes d'investigation[2].

Cependant, n'est-ce pas le lot de tout instrument de recherche? La construction de l'objet par le physicien se fait à travers une instrumentation qui, en fait, « produit » le phénomène observé (sa trace, sa forme observable) sur lequel le chercheur veut travailler : « On ne peut connaître de façon COMMUNICABLE, intersubjective, QUE DES DESCRIPTIONS, jamais une entité factuelle physique "elle-même", ni des phénomènes non décrits[3] ». En « triturant » le texte, le logiciel (ou les logiciels utilisés pour des croisements de traitements) fait émerger des « descriptions », met en évidence des catégories, des « clusters », suggère des rapprochements potentiellement porteurs de sens… Mais c'est le chercheur qui décide de ce sens, les « descriptions » du logiciel étant le matériau « brut » (pas tout à fait brut, cependant, puisque c'est d'un traitement du texte que ces descriptions sont nées). Physicienne, spécialiste des phénomènes quantiques, Miora Mugur-Schächter, ne dit pas autre chose :

veuille lui donner) ne peut fournir la moindre conclusion à partir des computations qu'il réalise.

2 Une liste convaincante en est donnée par Guillaume Ollivier, *Panorama critique des analyses textuelles informatisées en SHS*, http://www.academia.edu/2854000/Panorama_critique_des_analyses_textuelles_informatis%C3%A9es_en_SHS, site consulté le 21 septembre 2015. L'auteur, loin d'être un adversaire des analyses textuelles informatisées, pointe les risques pour que le chercheur puisse y échapper le mieux possible.

3 Miora Mugur-Schächter, « Objectivité, relativités, relativisme », dans Grasce, *Entre systémique et complexité, chemin faisant… Mélanges en l'honneur du professeur Jean-Louis Le Moigne*, Paris, Presses universitaires de France, 1999, p. 175. Les capitales sont de l'auteur.

> Une fois accomplie cette première étape de génération d'une entité-objet, on passe à la deuxième étape d'une certaine connaissance de l'entité-objet générée. Cette entité, telle qu'elle émerge de l'opération de sa génération, n'atteint pas le niveau de l'observable par l'homme. Il s'agit donc de l'amener à y produire des manifestations observables[4].

Cependant, le traitement informatique ne réduit-il pas trop fortement la richesse, la complexité du texte? C'est sans doute l'un des enjeux du débat (et des articles ici publiés) que de répondre à cette question. Mais on ne voit pas pourquoi ce traitement serait plus radical et destructeur que l'application de méthodes statistiques classiques au réel observable en SHS. L'intervention du qualitatif (même si ce sont des éléments quantitatifs qui sont en cause : des comptages et des opérations sur des nombres) offre la possibilité de ré-enrichir la réflexion. Simon Laflamme milite depuis des années pour que soit reconnue l'idée selon laquelle « la statistique […] fournit bon nombre d'informations dont l'interprétation la dépasse toujours parce qu'elles appellent la théorisation[5] » et le fait qu'« en elle-même, la statistique est pauvre, puisqu'elle n'est qu'une manière d'observer le monde que dessinent les modélisations. Elle est aussi pauvre que toute analyse à laquelle fait défaut l'interprétation, fût-elle qualitative[6] ».

Dans leur ouvrage paru en 1994, Ludovic Lebart et André Salem écrivaient :

> Le choix d'une stratégie de recherche ne peut être opéré qu'en fonction d'objectifs bien définis. Quel type de texte analyse-t-on? Pour tenter de répondre à quelles questions? Désire-t-on étudier le vocabulaire d'un texte en vue d'en faire un commentaire stylistique? […]
> Bien entendu, aucune méthode d'analyse figée une fois pour toutes ne saurait répondre entièrement à des objectifs aussi diversifiés[7].

Les textes présentés dans ce volume répondent à cette proposition. Tous, d'une manière différente, insistent sur l'intérêt, voire

4 *Ibid.*, p. 177.

5 Simon Laflamme, « Analyse statistique linéaire et interprétation systémique », *Nouvelles perspectives en sciences sociales*, vol. 4, n° 1, 2008, p. 145.

6 *Ibid.*, p. 159.

7 Ludovic Lebart et André Salem, *Statistiques textuelles*, 1994, http://lexicometrica.univ-paris3.fr/livre/st94/STU0INTRO.pdf, p. 7-8, site consulté le 5 octobre 2015.

la nécessité, de mêler les approches, de multiplier les analyses et les logiciels. Tous s'interrogent sur la valeur des outils, leur capacité à orienter le chercheur sans le contraindre de façon inacceptable. Tous proposent cette réflexion à travers une application empirique et la plupart montrent que l'utilisation des méthodes d'analyse textuelle informatisée leur permet de mettre à jour des caractéristiques qui, sans doute, seraient passées inaperçues sans ces méthodes. Nombreux également sont ceux qui donnent à voir que l'application des logiciels d'analyse textuelle, en permettant de traiter des corpus très importants, fondent plus fortement les interprétations de ces corpus.

Certains auteurs adoptent une orientation nettement méthodologique et s'intéressent aux conditions de mise en œuvre de logiciels dont les soubassements doivent être explorés et explicités, et donc aux méthodes possibles, mais pas équivalentes et non indifférentes, d'application de ces logiciels. D'autres proposent plutôt des résultats obtenus par la mise en œuvre d'une analyse textuelle, venant ainsi illustrer d'une certaine manière les propos précédents. La plupart des auteurs, cependant, mêle à l'envie ces deux approches.

Dans la première catégorie, Jean-Marc Leblanc s'interroge sur les protocoles possibles pour éviter que « la quantification de données textuelles produisant listes, tableaux et visuels [fasse que ces derniers ne soient] parfois commentés de manière aléatoire » et défend une « démarche expérimentale », « privilégiant le retour au texte », tout en proposant une « typologie des outils et des méthodes », « explicitant les postulats méthodologiques qui sont à la base de leur développement ». L'objectif de l'auteur est clairement de « valider la démarche lexicométrique sur de petits corpus » et de montrer que des croisements de logiciels et de méthodes permettent une meilleure analyse des corpus, l'une des questions qui émergent concernant « le rapport qu'il convient de faire aujourd'hui entre lexicométrie et sciences des textes, *data visualisation*, TAL (Traitement Automatique des Langues), etc. »

Dans un même mouvement, Martine Paindorge, Valérie Fontanieu et Jacques Kerneis réfléchissent sur les résultats obtenus

par leur analyse des contenus des « programmes et ressources pour l'enseignement » publiés par le Ministère de l'Éducation Nationale. Ce faisant, les auteurs en évaluent les avantages et les limites et plaident pour l'utilisation conjointe de plusieurs méthodes (analyse « manuelle », utilisation des logiciels *Tropes* et *Alceste*), qui permet de faire apparaître « des informations plus nombreuses et plus variées, lesquelles peuvent servir à valider les résultats d'une des méthodes, notamment la méthode "manuelle" pour laquelle la subjectivité du chercheur est perçue comme un écueil ». Cette « méthodologie articulée, qui reste à construire, offre également la possibilité de croiser les premiers résultats obtenus afin d'émettre de nouvelles hypothèses ».

Se basant sur l'analyse d'un corpus peu commun, celui de « quatre-vingt-trois rituels de Chevaliers Kadosh de la collation du fonds de l'atelier de recherches Sources », Bernard Pateyron, Maurice Weber et Pierre Germain veulent montrer que la lexicométrie contemporaine est à la fois performante, dispersée (entre auteurs, équipes, logiciels) et fondée sur des outils sous-jacents parfois obsolètes : « Il reste que la part proprement formelle d'axiomatisation est à la traîne de la théorie de l'information et reste à construire ». Ils veulent aussi « convaincre les chercheurs, tant les "non-historiens" que les "non-statisticiens", de leur aptitude à […] utiliser » ces méthodes, testées ici sur un corpus historique, qui permettent de faire apparaître les proximités et éloignements entre des sources dont la filiation postulée doit être fondée.

Pour Elisa Omodei, Yufan Guo, Jean-Philippe Cointet et Thierry Poibeau, c'est aussi la combinaison de deux approches qui permet d'extraire d'un corpus concernant des publications scientifiques des éléments saillants, en l'occurrence des catégories distinguant « le vocabulaire conceptuel des éléments d'ordre méthodologique ». Eux aussi plaident pour une combinaison des analyses (« recherche par nature pluridisciplinaire ») et un retour au texte, proposant, de façon originale et intéressante, de se retourner « vers des spécialistes d'histoire des sciences pour poursuivre ce travail en collaboration ».

En analysant à l'aide du logiciel *SPAD* plus de 11 000 articles de la presse canadienne et française, Roger Gervais tente de montrer que le processus de mondialisation ne génère pas forcément de l'homogénéisation des sociétés et que les réactions à ce phénomène supposé n'amènent pas non plus un morcellement, une forte différenciation du monde, mais, qu'en fait, les deux mouvements se produisent de façon concomitante. Les réflexions méthodologiques de l'auteur sur l'analyse textuelle font état du fait que la perte de sens que craignent certains chercheurs dans la mise en œuvre du logiciel ne se produit pas. Bien au contraire, nous dit Gervais, la possibilité de mobiliser des corpus d'une grande importance quantitative permet de faire ressortir les éléments fondamentaux qui traversent ce corpus, sans perdre les éléments qui permettent de revenir en permanence au texte et « on ne perd pas de vue la relation entre les mots et les idées ». Mais, affirme l'auteur, il n'est pas besoin, dans une analyse textuelle informatisée, de revenir aux divers contextes de l'écrit : qui est l'auteur, où vivait-il, dans quelle situation a-t-il écrit?... et, citant Bachelard, il nous demande de penser au fait que « [l]'empirisme est alors en quelque manière déchargé; il n'a plus à rendre compte à la fois de tous les caractères sensibles des substances mises en expérience[8] ».

Fabienne Baider s'intéresse au discours politique de madame Le Pen et montre que ce discours transporte des éléments de brouillage permettant à la fois de rompre apparemment avec le discours que tenait son père, d'inscrire ce nouveau discours dans le « bon sens », la « raison » et la « cohérence », de l'insérer dans un genre (le genre féminin) et de générer la proximité et l'empathie avec les électeurs, les sympathisants et certains autres citoyens susceptibles de se rapprocher du parti. L'idée selon laquelle ce serait « une tradition familiale » que de « renverser le stigmate » et, dans un discours populiste, d'affirmer qu'« on compatit avec les immigrés, mais [qu']on refuse les immigrés » ou celle qui permet à la députée européenne de s'affirmer « comme la protec-

8 Gaston Bachelard, *La formation de l'esprit scientifique*, Paris, Vrin, 2004 [1938], p. 127.

trice des plus vulnérables – immigrés inclus –, contre les élites corrompues », tout cela apparaît avec une certaine clarté dans l'analyse menée à l'aide, notamment, de *TermoStat*, permettant une herméneutique suffisamment distanciée.

Maud Hidalgo, Isabelle Ragot-Court et Chloé Eyssartier, quant à elles, ont comparé, à l'aide d'*Alceste*, les discours des automobilistes et des conducteurs de deux-roues sur la circulation inter-files pratiquée par les deux roues dans les embouteillages. Les cinq classes lexicales qui ressortent de cette analyse permettent d'informer le sujet d'étude en proposant des catégories appelées « conscience technique » et « conscience sociale », en montrant l'importance, dans les phénomènes d'accidents de la route, des représentations et comportements induits des usagers, de même que de mécanismes identitaires. L'analyse a également permis de faire émerger une problématique de contextes de circulation, qui montre que les automobilistes parisiens et marseillais, par exemple, ont des comportements différents, ce qui conduit les auteures à suggérer que « en matière de sécurité routière, la prise en compte de spécificités qui renvoient aux réalités du terrain est utile pour l'efficacité et la pertinence des actions de prévention et plus globalement des politiques publiques ».

Pour Audrey Arnoult, il s'agit de percevoir quelles représentations de l'anorexie mentale portent les discours médiatiques et, à travers cela, de voir comment les logiciels d'analyse textuelle peuvent aider à « saisir le sens que les discours médiatiques donnent à ce sujet ». Elle se demande, par ailleurs, si l'utilisation de logiciels ne contraint pas trop le chercheur en lui imposant, sans qu'il en soit toujours conscient, ses propres procédures d'analyse statistique. L'objet de l'auteure n'est donc pas tellement l'anorexie mentale que les aides à l'étude de ce sujet qu'apportent les deux logiciels choisis et combinés, *Modalisa* et *Iramuteq*. L'auteure montre que la création d'un corpus restreint grâce à *Modalisa* a permis, avec *Iramuteq*, de mettre en lumière des univers lexicaux, divers, évolutifs, hétérogènes, voire peu marqués, avec des formes peu significatives. D'où l'interrogation de l'auteure : « quels logiciels pour quel usage? » qui résume sa

conclusion : « Le chercheur ne peut faire l'économie d'une réflexion sur les postulats théoriques qui les sous-tendent [les procédés statistiques, NDA] afin de comprendre ce que "produit" le logiciel ».

Enfin, Maria Zimina et Serge Fleury traitent, de leur côté, d'une question importante qui s'est développée dans le sillage des améliorations des logiciels d'analyse textuelle : l'analyse automatique des corpus multilingues parallèles et comparables. Ce secteur de la recherche a permis le développement des logiciels de plus en plus performants de traduction en ligne, même si ce n'est là qu'une des retombées des recherches en ce domaine. D'une façon générale, leur sujet est l'exploration des méthodes disponibles pour pratiquer l'alignement d'ensembles de textes qui se ressemblent. Présentant plusieurs méthodes de traitement des textes implémentées dans le logiciel *Trameur*, les auteurs nous montrent la possibilité de divers parcours interprétatifs des données textuelles, grâce à la souplesse de l'architecture de stockage Trame/Cadre.

Depuis 1992, les Journées d'analyse de données textuelles (JADT)[9] réunissent tous les deux ans les chercheurs adeptes des travaux d'analyse textuelle et des quinze à vingt logiciels que l'on trouve aujourd'hui sur le marché (mais certains sont gratuits). Les auteurs du présent volume ont travaillé, quant à eux, avec et/ou sur *Modalisa*, *Iramuteq*, *Termostat*, *SPAD*, *Alceste*, *Prospero*, *Sphinx*, *Leximappe*, *TMX*, *Tropes*, *Lexico*, *Hyperbase*, *Coocs*, *Dtmvic*, *Le Trameur*, soit un éventail large de ce que l'on peut aujourd'hui utiliser en analyse textuelle informatisée. Pour cette raison qui s'ajoute à la qualité des textes, cette nouvelle livraison de *Nouvelles perspectives en sciences sociales* devrait devenir une référence dans le domaine des analyses textuelles informatisées[10]. Par sa diversité, sa multitude des angles d'approche, mêlant aspects théoriques et méthodologiques, ses contributions

9 On trouve les actes de ces journées sur le site : http://lexicometrica.univ-paris3.fr//.

10 On se reportera avec profit au texte de Jean-Marc Leblanc (dans ce volume) pour un aperçu des débats autour et de l'évolution de la terminologie et des typologies possibles dans ce domaine.

constamment situées entre la réflexion sur l'outil, ses capacités et insuffisances possibles, et sa mise en œuvre, il peut devenir un instrument de réflexion comme un guide de choix et d'application des multiples logiciels existants.

Si cette réflexion possède un intérêt de plus en plus admis dans les sciences humaines et sociales, elle permet aussi d'appuyer les efforts que la revue *NPSS* encourage, à savoir la volonté de modéliser relationellement des systèmes complexes. L'apport des analyses textuelles informatisées est, en effet, de permettre de se donner des corpus importants de discours publics et privés qui recèlent en eux de nombreuses informations sur les individus, les systèmes et l'histoire des éléments réels sous-jacents à ces concepts. Grâce à la puissance de l'outil, nous pouvons imaginer de repérer, sans aucune hypothèse sur la rationalité et la psyché des acteurs, leurs histoires de vie insérées dans l'histoire collective, leurs modes d'engagement dans la cité comme dans leur propre microcosme, les caractéristiques et les évolutions des systèmes dans lesquels nos analyses les font évoluer.