
Proposition Traitement automatique des données textuelles et pétitionnement en ligne : l'apport de l'analyse de discours en sciences de l'information et de la communication

Christine Barats*¹

¹CEDITEC – Université Paris-Est Créteil Val-de-Marne - Paris 12 – France

Résumé

Le déploiement du web s'est accompagné d'une profusion de données et est caractérisé par la prégnance de logiques algorithmiques (Cardon, 2015), ce qui pose la question de la place des SHS dans le contexte des big data (Boullier, 2015). Si les SHS disposent de méthodes quantitatives, en particulier dans le cas de données textuelles, le risque demeure d'un retour d'une forme de positivisme, non exempt d'une délégation aveugle à la machine (Née et al., 2017). Dès les années 1970, Michel Pécheux rappelait qu'en analyse de discours le traitement automatique de données textuelles ne signifie en aucun cas une interprétation automatique (Pécheux, 1969). Nous nous proposons de montrer l'apport de l'analyse de discours en Sciences de l'Information et de la Communication (SIC) pour examiner des volumes importants de données dans le cas de corpus numériques. Le volume des données et la multiplication d'outils de traitement automatique incitent en effet à privilégier une approche quantitative. Or, il s'agit de tenir compte des présupposés de ces outils qui orientent les résultats obtenus, invitant à renouveler les réflexions épistémologiques et méthodologiques sur le choix des données numériques et des méthodes (Venturini et al., 2014). Le recours de plus en plus fréquent à des outils de traitement automatique, comme dans le cas par exemple des outils d'opinion mining ou d'analyse de sentiments (Boullier, Lohard, 2012), réduit parfois les données textuelles (leur diversité et hétérogénéité s'avèrent partiellement gommées) et standardise les résultats ainsi que leur interprétation. Les corpus textuels issus du web reposent en effet sur une diversité et une hétérogénéité de données : données inscrites par les internautes et également données générées par le dispositif et relatives aux traces d'activité des internautes (horodatage, inscriptions liées à l'activité, mesures...). Cette diversité et cette hétérogénéité des données implique de renouveler les modalités de constitution et d'analyse de corpus web afin de tenir compte du dispositif étudié, de ses fonctionnalités et de son design (Bottini, Julliard, 2017 ; Longhi, 2017 ; Contamin et al., 2017 ; Paveau, 2017 ; Dagiral, Pailler, 2018), tout autant que des graphies et des inscriptions laissées par les internautes (Halté, 2016 ; Paveau, 2017 ; Mercier, Pignard-Cheynel, 2018). Nous montrerons à partir des résultats d'une recherche financée par l'ANR sur le pétitionnement en ligne les apports de l'analyse de discours en SIC, et en particulier en quoi l'articulation d'approches quantitatives et qualitatives s'avère féconde. L'accès aux données de la plateforme a constitué un potentiel sans précédent car les données collectées constituent à la fois des données textuelles alphanumériques (rubrique de dépôt de la pétition, cible, titre et texte de la pétition, date de lancement et de fin, nom du porteur de la pétition, commentaires,

*Intervenant

signataires) et également des traces de la dynamique de pétitionnement (horodatage, nombre de signataires, lieu d'habitation...). Nous avons été confrontés à de multiples difficultés pour la constitution des corpus (masse de données à traiter, anonymisation...). Cependant, la collecte massive a permis la constitution de différents corpus qui ont tenu compte des spécificités du dispositif, sans privilégier uniquement les pétitions les plus signées ou certaines thématiques. A l'opposé d'une vision idéalisée du web comme espace de liberté, nous avons mis l'accent sur le pouvoir de " faire faire " qui s'exerce au sein des environnements numériques (Badouard et al., 2016). Nous avons appréhendé la plateforme comme un dispositif qui incite, contraint et encadre l'acte de pétitionnement, en particulier au regard de l'architexte, c'est-à-dire " les objets informatiques qui sont en position de régir l'écriture, de lui donner ses formats et ses ressources " (Tardy, Jeanneret, 2007 : 24). Le site propose neuf rubriques pour le dépôt d'une pétition. Il est possible d'examiner l'incidence du choix des rubriques sur les caractéristiques textuelles des pétitions et des commentaires, les liens entre ces caractéristiques et le nombre de signataires, etc., mettant au jour le poids du dispositif, les caractéristiques des textes de pétition et également la diversité des raisons pour lesquelles les signataires signent une pétition. Nous montrerons l'intérêt de croiser les données afin de constituer des corpus *ad hoc* et examiner le poids du dispositif et de certaines variables. Nous montrerons également la nécessité d'articuler des approches quantitatives de la statistique textuelle, compte tenu du volume des données, et qualitatives (entretiens avec le concepteur de la base, analyse " à la main ", retour au texte, etc.).

Références

Barats C., Dister A., Gambette P., Leblanc J-M., Leblanc-Peres M., " Appeler à signer une pétition en ligne : caractéristiques linguistiques des appels ", *JADT 2018*, Rome, <http://lexicometrica.univ-paris3.fr/jadt/JADT2018/actes-jadt18.pdf>.

Barats C., Dister A., Gambette P., Leblanc J-M., Leblanc-Peres M., " Analyser des pétitions en ligne : potentialités et limites d'un dispositif d'étude pluridisciplinaire ", *JADT 2016 : 13èmes Journées internationales d'Analyse statistique des Données Textuelles*, Nice, juin 2016, <https://jadt2016.sciencesconf.org/83043>.

Barats Christine (Dir.), Manuel d'analyse du web, Collection U, Armand Colin, 2016.

Boullier, Dominique ; Lohard, Audrey. *Opinion mining et Sentiment analysis : Méthodes et outils*. Nouvelle édition [en ligne]. Marseille : OpenEdition Press, 2012

Boullier Dominique, " Les sciences sociales face aux traces du *big data*. Société, opinion ou vibrations ? ", *Revue française de science politique*, 2015/5 (Vol. 65), p. 805-828. DOI : 10.3917/rfsp.655.0805. URL : <https://www.cairn.info/revue-francaise-de-science-politique-2015-5-page-805.htm>

Dominique Cardon, *À quoi rêvent les algorithmes. Nos vies à l'heure des big data*, Paris, Seuil, La République des idées, 2015.

Née Emilie (dir.), Barats C., Fleury S., Leblanc J-M., Sitri F., Veniard M., *Méthodes et outils informatiques pour l'analyse des discours*, Presses Universitaires de Rennes, août 2017.

Paveau M.-A., *L'analyse du discours numérique*, Paris, Hermann, 2017.

Pêcheux Michel, *Analyse automatique du discours*, Paris, Dunod, 1969.

Reinert M., " Les "mondes lexicaux" et leur logique ", *Langage et société*, 66, pp. 5-39, 1993.

Reinert M., " Quelques interrogations à propos de l'"objet" d'une analyse de discours de type statistique et de la réponse "Alceste" ", *Langage et société*, 90, p. 57-70, 1998.

Tardy C., Jeanneret Y. (dir.), 2007, *L'écriture des médias informatisés – espaces de pra-*

tiques, Paris, Lavoisier.

Venturini T., Cardon D., Cointet J-P., 2014, " Présentation ", *Réseaux*, 188, pp. 9-21