HAL
open science

L'analisi testuale

Claire Lemercier

To cite this version:

Claire Lemercier. L'analisi testuale. Deborah Paci. La storia in digitale. Teorie e metodologie, Edizioni Unicopli, pp.291 - 292, 2019, 9788840020914. hal-02455457

HAL Id: hal-02455457

https://sciencespo.hal.science/hal-02455457

Submitted on 29 Jul 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Version originale française du texte publié comme : Claire Lemercier, « L'analisi testuale », in Deborah Paci (a cura di), *La storia in digitale: teorie e metodologie*, Milan, Unicopli, 2019, p. 291-292 (trad. Jacopo Bassi).

L'analyse de textes

Claire Lemercier

L'analyse de textes assistée par ordinateur a été proposée dès l'époque des ordinateurs à cartes perforées, en particulier en histoire littéraire et en histoire politique. Mais elle n'a jamais été adoptée par un grand nombre d'historien.ne.s. Il n'existe même pas de vocabulaire stabilisé pour l'évoquer. On parle d'analyse de discours, de contenu, lexicale, textuelle, d'exploration de corpus, de lexicométrie, logométrie, stylométrie, textométrie, plus récemment de « lecture à distance », de *text mining* ou de *topic modeling*, pour évoquer des pratiques qui ont pourtant beaucoup en commun. En mesurant les fréquences d'usages de mots (et souvent les proximités dans les phrases entre ces mots) elles caractérisent les ressemblances et les différences entre des textes (et souvent entre des auteurs, des publications ou encore des époques).

Du fait de ce rendez-vous régulièrement manqué entre l'histoire et l'analyse (quantitative) de textes, ceux qui comptent de toute façon les mots le font de plus en plus sans les historien.ne.s. Par exemple, les *cultoromics* supposent que Google Books représente bien l'ensemble des publications et que les publications « reflètent » toute l'histoire de l'humanité. Ce désintérêt pour la critique des sources et la bibliographie historienne entretient celui des historien.ne.s pour les outils numériques d'analyse de textes, et vice versa.

S'intéresser à ces outils confronte vite à des choix peu agréables. Les plus simples portent sur un corpus biaisé et mal défini (les Google Ngrams) ou proposent des visualisations à la fois simplistes et peu lisibles (les nuages de mots). Les plus avancés, comme le *topic modeling*, pas encore traduits dans la langue de l'histoire, demandent un long apprentissage ou font figures de boîtes noires. Même des logiciels développés pour et avec les sciences humaines et sociales, comme IraMuTeQ ou TXM, manquent de tutoriels adaptés à l'histoire et de collections de belles applications.

L'analyse de textes va-t-elle enfin entrer dans la boîte à outils de l'historien.ne ? La quantité grandissante de sources numérisées la rend plus rapide (le temps de numérisation était souvent un obstacle infranchissable auparavant). Elle rend toutefois encore plus nécessaire la réflexion critique sur ce qui est en ligne et ce qui n'y est pas (la plupart des textes documentant les dominé.e.s, notamment). Le choix de *ce que l'on regarde de loin*, plus encore que les décisions en matière de méthodes, oriente en effet les résultats.

Les proclamations excessivement objectivistes ont beaucoup nui à l'analyse de textes. On n'explicitera jamais assez tout ce qu'elle laisse à l'interprétation qualitative : le choix du corpus, donc ; certains paramétrages, selon les méthodes ; le sens donné aux résultats. Il existe bien des manières différentes de compter dans des textes. Beaucoup de ces manières n'impliquent d'ailleurs pas de logiciels spécialisés : compter des figures de style, par exemple, se fait bien mieux avec un tableur (ou une indexation en TEI, etc.). C'est une autre manière de quantifier : après une première lecture, une interprétation et une catégorisation humaines, plutôt que par induction directe à partir de tous les mots du texte. Considérer ainsi le choix d'utiliser l'ordinateur ou non, de quantifier ou non, et selon quel principe, à chaque étape permet de sortir de la fausse alternative entre le tout humaniste et le tout informatique.

Bibliographie

S. Graham, I. Milligan & S. B. Weingart, *Exploring big historical data: the historian's macroscope*, London, Imperial College Press, 2016.

C. Lemercier & C. Zalc, *Quantitative Methods in the Humanities. An Introduction*, Charlottesville, University of Virginia Press, 2019.

D. Mayaffre, *Vers une herméneutique matérielle numérique. Corpus textuels, Logométrie et Langage politique*, manuscrit d'habilitation à diriger des recherches, Université Nice Sophia Antipolis, 2010.

F. Moretti, *Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850), Critical Inquiry*, 36 (2009), pp. 134-158.