Computation

Jean-Baptiste Camps

Computation: Compter les mots, compter les choses… Au delà de la simple quantification, les méthodes d'analyse fondées sur le calcul, au sens large, aidées des machines à calculer que sont les ordinateurs, permettent une nouvelle appréhension et de nouvelles interrogations sur les sources, textuelles notamment, mais pas uniquement.

Sans remonter jusqu'au Massorètes et à leurs décompte des mots de la Bible, on relève dès le milieu du XIXe siècle des tentatives de mesure du style et de datation ou d'attribution d'œuvres disputées, qui mènent à l'apparition du terme de stylométrie (Tannery, 1899; Hockey, 2004; Delcourt, 2002), et même à l'élaboration d'une première `machine à compter', servant, via le travail de deux opératrices, à recueillir plus rapidement les données statistiques d'un texte (Mendenhall, 1901). Du côté de la critique textuelle, l'étude des généalogies textuelles (stemmatique ou stemmatologie), de par son caractère algorithmique, s'est prêtée assez tôt à des réflexions de nature mathématique (Greg, 1927; Quentin, 1926; Maas, 1937). En matière de computation philologique, deux branches à l'histoire longue méritent d'être distinguée, l'une se préoccupant de «basse critique», c'est-à-dire de démarche ecdotique, et de collation comme de stemmatologie, et l'autre de « haute critique », c'est-à-dire notamment d'attribution des textes.

En termes d'ecdotique, les développements les plus récents de la science des données, de l'intelligence artificielle et du traitement automatique des langues, rendent possible de nouveaux modes de constitution de corpus ou d'interrogation des sources, qu'il s'agisse de reconnaissance automatique des écritures manuscrites, pouvant servir la recherche plein-texte dans un vaste fonds comme les registres de la chancellerie royale française des années 1302-1483 (Stutzmann, Moufflet et Hamel, 2017), de l'ajout d'informations telles que des annotations d'ordre linguistique (lemmes, parties du discours, …; Manjavacas et al., 2019), de la reconnaissance des entités nommées (Torres Aguilar, 2017), voire de la conception de chaînes de traitement complètes allant de l'image numérisée de la source à l'édition critique ou à l'analyse scientifique du contenu, à l'échelle d'un manuscrit unique (Camps, Pinche et Clérice, 2019), voire peut-être de très vastes collections (Abbott, 2017).

En dépit de son caractère de « niche », la stemmatologie dispose d'une histoire ancienne, relativement au domaine envisagé. Le besoin de traiter des traditions très touffues a amené assez tôt des tentatives de transposition calculable de l'algorithmie des variantes, chez Dom Quentin (1922) travaillant sur la Vulgate ou chez Dom Froger (1968), peut-être le premier à avoir eu recours aux ordinateurs en la matière. Les méthodes algorithmiques pour l'établissement de la généalogie de traditions textuelles sont un domaine actif jusqu'à nos jours, et qui a connu ses écoles, notamment à Amsterdam, parmi les élèves d'Anthonij Dees (Van Reenen et al., 1996; 2004). Au sein du champ, des méthodes fondées sur la mathématisation de principes traditionnels de la critique textuelle (Poole, 1979) voisinent avec des méthodes empruntées à la biologie (plus exactement à la phylogénétique et à la cladistique) ou à la recherche en informatique, par exemple aux algorithmes de compression (Roos et al., 2009; voir aussi Hoenen, 2018). Des recherches plus théoriques examinent, quant à elles, les différentes configurations arborescentes possibles d'un point

de vue mathématique (Hoenen et al., 2017), ou ont recours à la modélisation et à la simulation pour confronter les formes observées des arbres à la variation de paramètres tels que les taux de copie ou de destruction des manuscrits (Weitzman, 1987).

Le vaste champ de l'analyse computationnelle des données textuelles est uni par un certain nombre de principes méthodologiques, tels que les méthodes statistiques employées et le recours aux calculs de fréquence (de mots ou de séquences), en dépit d'une variété d'appellations qui dépendent plutôt de l'objet étudié que du choix de méthode. Dans ce domaine, la stylométrie a probablement été première (dès le XIXe siècle, comme on l'a vu plus haut), et reste jusqu'à aujourd'hui un domaine très actif. À partir de la quantification de traits essentiellement grammaticaux (usage des mots-outils, affixes, et autres morphèmes grammaticaux, séquences de natures grammaticales des mots, etc.), elle vise à établir ou vérifier l'autorité de textes anonymes ou disputés, ou à déceler des collaborations, comme dans l'œuvre de Hildegarde de Bingen (Kestemont et al., 2015).

Malgré une unité méthodologique assez large sur plan statistique, on distingue notamment, surtout en France, en sus de la stylométrie qui les précède de plus d'un siècle, et de l'analyse de données textuelles au sens large, une série de champs, pourtant assez largement marqués par un héritage commun, celui de l'analyse du discours et de la statistique lexicale, et par des figures pionnières comme Charles Muller ou Étienne Brunet, selon l'objet de leur étude, qu'il s'agisse du lexique (lexicométrie), des textes (textométrie, voir Pincemin, 2011), du discours (logométrie, voir Mayaffre, 2005), de la variation linguistique (dialectométrie et scriptométrie), ou, nouvelle arrivée, la littérature (littérométrie), chacun avec ses écoles et ses centres: à l'ENS Saint-Cloud, à Nice, ou, pour la dialectométrie, à Amsterdam, où furent réalisés sous la direction d'A. Dees les atlas des formes des textes en ancien français (Dees et al., 1980 et 1987) et à Salzbourg. En outre, pour l'extraction des thèmes et sujets évoqués dans un corpus de sources, le philologue peut bénéficier des développements en recherche d'information: un outil tel que la modélisation de sujet (*topic modelling*) permet d'interroger par exemple la définition des genres du théâtre classique à partir de l'extraction automatique des sujets et des thèmes (Schöch, 2017).

En cheminant vers la matérialité du texte, notons aussi les riches possibilités offertes par l'analyse computationnelle à la paléographie, où par exemple la vision par ordinateur permet l'identification des mains ou la classification des écritures (Kestemont, Christlein et Stutzmann, 2017). Le livre manuscrit fait également l'objet d'études de codicologie quantitative, depuis les travaux fondateurs de Bozzolo et Ornato (1980), et rejoint méthodologiquement l'analyse de données non textuelles, même si potentiellement extraites des sources.

Si l'on quitte à présent ainsi l'analyse des données textuelles, chez les historiens l'histoire quantitative avant l'ordinateur est avant tout histoire économique et sociale, et trouve ses pionniers à la fin du XIXe et au début du XXe siècle, avec notamment l'établissement des premières séries de données sur l'évolution des prix, même si ce serait surtout la crise de 1929 qui jouerait le rôle d'élément déclencheur dans la constitution de ce champ (Chaunu,

1972), avec notamment dans les années 1930 les travaux de François Simiand et Ernest Labrousse (Bourin et Zadora-Rio, 2013). De ce point de vue, cette approche est héritière de la statistique comme science servant à l'étude de leurs populations par les États. La gamme méthodologique disponible pour les historiens est d'ailleurs toujours assez largement commune aux sciences sociales computationnelles dans leur ensemble, qu'il s'agisse des statistiques descriptives, de l'analyse causale, ou encore de la modélisation et de l'analyse des réseaux, ou plus largement de l'outillage propre aux systèmes complexes.

Un des premiers enjeux concerne la mise en place de séries chiffrées longues, permettant d'observer des évolutions de long terme, des tendances, avec les difficultés que cela pose en termes de collecte des données dans les sources, de formalisation et de manipulation de celles-ci. L'école des Annales fournit de nombreux exemples de ce type d'approche, à l'instar des travaux de Micheline Baulant sur le salaire des ouvriers du bâtiment à Paris, de 1400 à 1726 (Baulant, 1971). Une fois les données collectées, c'est au tour des méthodes de la statistique descriptive d'entrer en jeu pour résumer l'information sous forme graphique (courbes, histogrammes, graphe de dispersion…) ou numériques (indicateurs sur la distribution, moyenne, médiane, écart-type, etc.).

Mais au-delà de la synthèse et de la première approche des tendances que fournissent les statistiques descriptives, le chercheur sera bientôt tenté de chercher à identifier, à partir de ses sources, des mécanismes à l'œuvre, en commençant par des évolutions communes ou contraires à plusieurs facteurs. L'étude du rapport entre deux variables, telles que hauteur et largeur des manuscrits (Bozzolo et Ornato, 1980), et des corrélations, qui n'impliquent pas nécessairement causalité, peut ensuite mener à la formulation d'hypothèses et d'un modèle explicatif. Tout particulièrement, pour les historiens, elles peuvent aussi être une manière de traiter les données manquantes ou lacunaires,

L'étude sur le climat à Genève aux XVI-XVIIIe siècle d'Anne-Marie Piuz (1974), qui s'inscrit dans une dynamique d'études sur le climat au sein des Annales, initiée notamment par Emmanuel Le Roy Ladurie, étudie les variations communes aux températures et aux jours d'avance ou de retard aux vendanges, et entre ceux-ci et le prix du blé. Une fois ces corrélations détectées (i.e., été pourri, vendanges tardives, prix du blé élevé, ou, à l'inverse, températures élevées, vendanges précoces, prix du blé bas) et précisément coefficientées, il devient possible d'inférer certaines valeurs manquantes (par exemple, la température, lorsque l'on connaît la date des vendanges et le prix du blé), et de constituer une série plus complète permettant l'étude des variations climatiques sur la période. Cherchant à étudier la variation de la consommation d'alcool durant la Prohibition aux États-Unis, mais ne disposant pas – et pour cause – de statistiques officielles en la matière, Jeffrey A. Miron (1999) modélise la relation, décalée dans le temps, entre cirrhoses du foie et consommation d'alcool pendant les périodes précédant et suivant la Prohibition, pour pouvoir inférer la variation de la consommation d'alcool.

L'emploi de systèmes d'équations, pour modéliser des comportements historiques, éventuellement appuyé sur des simulations par ordinateur, a également des applications du

côté des systèmes complexes, c'est-à-dire des systèmes dans lesquels de l'interaction des comportements individuels d'un grand nombre d'agents peuvent naître des effets collectifs difficiles à prédire, éventuellement marqués par des points critiques pouvant engendrer un changement soudain d'état (« transitions de phase »). Parvenue en histoire depuis les sciences sociales, et dans celles-ci depuis la physique et les sciences du vivant ou du climat, la recherche sur les systèmes dynamiques et multi-agents est peut-être pour le moment plus présente en archéologie, en anthropologie historique ou dans l'étude de la pré- et proto-histoire (Barceló et Del Castillo, 2016). On relève néanmoins des applications portant notamment sur les mouvements de foule et les batailles telles qu'Azincourt (Clements et Hughes, 2004), ainsi que des modèles de population et des processus stochastiques de naissance et mort appliqués à la transmission manuscrite des textes (Weitzman, 1987; Cisne, 2005), ou bien encore des modèles épidémiologiques mis au service de la reconstitution de la diffusion des épidémies du passé, au premier chef desquelles la peste noire (Christakos et al., 2005).

Ce type de modélisation, notamment lorsqu'elle porte sur les transports, les échanges ou le commerce, peut également se coupler à l'étude des réseaux et au traitement d'information spatialisée, comme c'est le cas notamment dans le projet ORBIS, qui étudie les déplacements dans le monde antique (Meeks et Grossner, 2012), ou bien encore dans l'étude de Schmid et al. (2015), qui croise circuits d'échanges, variations climatiques et épisodes pesteux pour confronter l'hypothèse de réservoirs de peste en Occident à celle de réintroductions successives depuis l'Orient.

Au croisement de la théorie mathématique des graphes et de l'étude des interactions sociales, l'analyse de réseaux est un incontournable des méthodes computationnelles en histoire comme en sociologie. Dès les années 1960, Gardin & Garelli (1961) ont l'idée de s'appuyer sur la théorie des graphes pour étudier les réseaux commerciaux en Cappadoce au XIXe siècle avant J.-C, à partir de la saisie de données extraites de tablettes cunéiformes, fournissant peut-être la première application de la théorie des graphes à des documents historiques (Plutniak, 2018). À partir de tablettes décrivant des transactions marchandes et financières, souvent complexes et impliquant de multiples acteurs, il est possible d'extraire un ensemble de relations deux à deux; une fois en possession de quelques milliers de liens de ce type, une représentation sous forme de graphe de nœuds (les individus) et d'arêtes (les liens) fournit une manière élégante de synthétiser l'information, pour la rendre à nouveau maniable, et y déceler des structures. Les applications de l'analyse de réseaux aux sources historiques sont trop nombreuses pour être toutes évoquées ici (voir notamment Lemercier, 2005; Bourin et Zadora-Rio, 2013), mais on retiendra notamment l'étude de Padgett et Ansell (1993), qui analyse l'ascension des Médicis à Florence en fonction de la position intermédiaire de cette famille dans un réseau d'élites autrement marqué par des disjonctions.

Terminons en évoquant les outils dédiés au traitement de l'information spatialisée, utilisés au premier chef par les archéologues, tels que les systèmes d'information géographique.

Ceux-ci permettent de stocker, visualiser et analyser des données spatiales, qu'il s'agisse de géolocaliser les lieux évoqués dans des sources textuelles comme de reconstituer l'implantation de communautés à partir de données archéologiques. La spatialisation de lieux extraits de textes pose la question de leur éventuelle reconnaissance automatique, via par exemple un système de reconnaissance des entités nommées, ainsi que du maintien d'un référentiel de lieux: un outil collaboratif en ligne tel que *Recogito* permet ainsi la reconnaissance semi-automatique des noms de lieux, le lien avec un référentiel et la géolocalisation, avec une focale sur les sources narratives de l'Antiquité classique et du Moyen Âge européen (Simon et al., 2015). Des applications existent également au croisement de la spatialisation et de la diachronie, voire de l'analyse des réseaux, qu'il s'agisse d'étudier l'espace urbain parisien (Noizet, Bove et Costa, 2013) ou le parcellaire de terrois villageois du Moyen Âge au cadastre napoléonien, en s'appuyant sur les sources fiscales et notamment les terriers et compoix (Rodier et al., 2013).

Jean-Baptiste Camps

Bibliographie

En pied de notice

Bourin (Monique) et Zadora-Rio (Élisabeth), «La mesure au Moyen Âge et les mesures des médiévistes: remarques en forme de conclusion», dans *Mesure et histoire médiévale: XLIIIe Congrès de la SHMESP (Tours, 31 mai-2 juin 2012)*, Paris, 2013, p. 374-396.

Guerreau (Alain), *Statistique pour historiens*, Paris, 2004, http://elec.enc.sorbonne.fr/statistiques/stat2004.pdf.

Jockers (Matthew), *Text Analysis with R for Students of Literature*, Cham, 2014 (Quantitative methods in the Humanities and Social Sciences), http://www.springer.com/statistics/computational+statistics/book/978-3-319-03163-7.

Juola (Patrick) et Ramsay (Stephen), *Six Septembers: Mathematics for the Humanist*, 2017, http://digitalcommons.unl.edu/zeabook/55.

Lemercier (Claire) et Zalc (Claire), *Méthodes quantitatives pour l'historien*, Paris, 2007 (Repères, 507).

O'Sullivan (James), *Digital Humanities for Literary Studies: Theories, Methods, and Practices*, University Park, à paraître (**depuis 2017**).

En bibliographie générale

Abbott (Alison), « The 'Time Machine'reconstructing ancient Venice's social networks », *Nature News*, 546-7658 (2017), p. 341.

Barceló (Juan A.) et Del Castillo (Florencia), *Simulating Prehistoric and Ancient Worlds*, Cham, 2016, doi: 10.1007/978-3-319-31481-5.

Baulant (Micheline), « Le salaire des ouvriers du bâtiment à Paris, de 1400 à 1726 », *Annales*, 26 (1971), p. 463-83, doi: 10.3406/ahess.1971.422372.

Bourin (Monique) et Zadora-Rio (Élisabeth), «La mesure au Moyen Âge et les mesures des médiévistes: remarques en forme de conclusion», dans *Mesure et histoire médiévale: XLIIIe Congrès de la SHMESP (Tours, 31 mai-2 juin 2012)*, Paris, 2013, p. 374-396.

Bozzolo (Carla) et Ornato (Ezio), *Pour une histoire du livre manuscrit au Moyen Âge: trois essais de codicologie quantitative*, Paris, 1980.

Camps (Jean-Baptiste), Pinche (Ariane) et Clérice (Thibault), « Stylometry for Noisy Medieval Data: Evaluating Paul Meyer's Hagiographic Hypothesis », dans *Digital Humanities Conference 2019 - DH2019, Utrecht, Netherlands*, Utrecht, 2019, **https://arxiv.org/abs/-2012.03845.**

Christakos (George), Olea (Ricardo A.), Serre (Marc L.), Yu (Hwa-Lung) et Wang (Lin-Lin), *Interdisciplinary Public Health Reasoning and Epidemic Modelling: The Case of Black Death*, Cham, 2005, doi: 10.1007/3-540-28165-7.

Cisne (John L.), « How Science Survived: Medieval Manuscripts' *Demography* and Classic Texts' Extinction », *Science*, 307-5713 (2005), p. 1305-1307.

Clements (Richard R.) et Hughes (Roger L.), « Mathematical modelling of a mediaeval battle: the battle of Agincourt, 1415 », *Mathematics and computers in simulation*, 64–2 (2004), p. 259-269.

Dees (Anthonij), van Reenen (Pieter), et De Vries (Johan A), *Atlas des formes et des constructions des chartes françaises du XIIIe siècle*, Tübingen, 1980 (Beihefte zur Zeitschrift für romanische Philologie, 178), doi: 10.1515/9783111328980.

Dees (Anthonij), Dekker (Marcel), Huber (Onno), et van Reenen-Stein (Karin), *Atlas des formes linguistiques des textes littéraires de l'ancien français*, Tübingen, 1987 (Beihefte zur Zeitschrift für romanische Philologie, 212), doi: 10.1515/9783110935493.

Froger (Jacques), *La Critique des textes et son automatisation*, Paris, 1968 (Initiation Aux Nouveautés de La Science).

Hoenen (Armin), *Tools, Evaluation and Preprocessing for Stemmatology*, thèse de doct., Goethe University Frankfurt, 2018.

Hoenen (Armin), Eger (Steffen) et Gehrke (Ralf), « How Many Stemmata with Root Degree k? », *Proceedings of the 15th Meeting on the Mathematics of Language*, 2017.

Kestemont (Mike), Moens (Sara) et Deploige (Jeroen), « Collaborative authorship in the twelfth century: A stylometric study of Hildegard of Bingen and Guibert of Gembloux », *Literary and Linguistic Computing*, 30-2 (2015), p.199-224.

Kestemont (Mike), Christlein (Vincent) et Stutzmann (Dominique), « Artificial Paleography: Computational Approaches to Identifying Script Types in Medieval Manuscripts », *Speculum*, 92-1 (2017), p. 86-109, 10.1086/694112.

Lemercier (Claire), « Analyse de réseaux et histoire », *Revue d'histoire moderne & contemporaine*, 52-2 (2005), p. 88-112, doi: 10.3917/rhmc.522.0088.

Manjavacas (Enrique), Kádár (Akos) et Kestemont (Mike), « Improving lemmatization of non-standard languages with joint learning », *arXiv preprint* arXiv:1903.06939, 2019.

Mayaffre (Damon), « De la lexicométrie à la logométrie », *Astrolabe*, 2005, p. 1-11, https://hal.archives-ouvertes.fr/hal-00551921/document.

Meeks (Elijah) et Grossner (Karl), « ORBIS : An interactive scholarly work on the Roman world », *Journal of Digital Humanities*, 1–3 (2012), p. 1–3.

Miron (Jeffrey A.), « The Effect of Alcohol Prohibition on Alcohol Consumption », *NBER Working Paper*, 7130 (1999), http://www.nber.org/papers/w7130.

Noizet (Hélène), Bove (Boris) et Costa (Laurent), éds, *Paris de parcelles en pixels: analyse géomatique de l'espace parisien médiéval et moderne*, Paris, 2013.

Padgett (John F.) et Ansell (Christopher K.), « Robust Action and the Rise of the Medici, 1400-1434 », *American Journal of Sociology*, 98-6 (1993), p. 1259-1319.

Pincemin (Bénédicte), « Sémantique interprétative et textométrie », *Corpus*, 10 (2011), p. 259-269, http://journals.openedition.org/corpus/2121.

Piuz (Anne-Marie), « Climat, récoltes et vie des hommes à Genève, XVIe-XVIIIe siècle », *Annales*, 29 (1974), p. 599–618, doi: 10.3406/ahess.1974.293496.

Plutniak (Sébastien), « Aux prémices des humanités numériques ? La première analyse automatisée d'un réseau économique ancien (Gardin & Garelli, 1961). Réalisation, conceptualisation, réception », *ARCS: Analyse de réseaux pour les sciences sociales*, 2018, https://hal.archives-ouvertes.fr/hal-01870945.

Poole (Eric), « L'analyse stemmatique des textes documentaires », dans *La pratique des ordinateurs dans la critique des textes*, Paris, 1979, p.151-161.

Quentin (Henri), *Mémoire sur l'établissement du texte de la Vulgate*, Rome, 1922.

Rodier (Xavier), Hautefeuille (Florent), Le Couédic (Mélanie), Leturcq (Samuel), Jouve (Bertrand) et Fieux (Étienne), « De l'espace aux graphes: mesurer les dynamiques spatiales des terroirs villageois », dans *Mesure et histoire médiévale: XLIIIe Congrès de la SHMESP (Tours, 31 mai-2 juin 2012)*, Paris, 2013, p. 99-118.

Roos (Teemu), Heikkilä (Tuomas) et Myllymäki (Petri), « A compression-based method for stemmatic analysis », dans *Proceeding of the 2006 conference on ECAI 2006: 17th European Conference on Artificial Intelligence August 29– September 1, 2006, Riva del Garda*, 2006, p.805-806.

Schmid (Boris V.), Büntgen (Ulf), Easterday (W. Ryan), Ginzler (Christian), Walløe (Lars), Bramanti (Barbara) et Stenseth (Nils Chr.), « Climate-driven introduction of the Black Death and successive plague reintroductions into Europe », *PNAS*, 112-10 (2015), p. 3020-3025, doi: 10.1073/pnas.1412887112.

Schöch (Christoph), « Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama », *Digital Humanities Quarterly*, 11-2 (2017), http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html.

Simon (Rainer), Barker (Elton), Isaksen (Leif) et de Soto Cañamares (Pau), « Linking early geospatial documents, one place at a time: annotation of geographic documents with Recogito », *e-Perimetron*, 10-2 (2015), p. 49-59.

Stutzmann (Dominique), Moufflet (Jean-François) et Hamel (Sébastien), « La recherche en plein texte dans les sources manuscrites médiévales : enjeux et perspectives du projet HIMANIS pour l'édition électronique », *Médiévales*, 73 (2017), doi : https://dx.doi.org/10.4000/medievales.8198.

Torres Aguilar (Sergio), « La reconnaissance des entités nommées dans les bases numériques de chartes médiévales en latin : le cas du Corpus Burgundiae Medii Aevi (Xe-XIIIe siècle) », *Médiévales*, 73 (2017), doi: 10.4000/medievales.8182

Van Reenen (Pieter), Van Mulken (Margot) et Dyk (Janet), éds, *Studies in stemmatology*, Amsterdam, 1996.

Van Reenen (Pieter), Den Hollander (Aurelius Augustinus), Van Mulken (Margot) et Roeleveld (Annelies), éds, *Studies in stemmatology II*, Amsterdam, 2004.

Weitzman (Michael P.), « The Evolution of Manuscript Traditions », *Journal of the Royal Statistical Society, series A (General)*, 150-4 (1987), p. 287-308.