La Philologie computationnelle à l'École des chartes: premier bilan et perspectives

Jean-Baptiste Camps

Bibliothèque de l'École des chartes, t. 176, 2021, p. 000-000.

LA PHILOLOGIE COMPUTATIONNELLE À L'ÉCOLE DES CHARTES

Premier bilan et perspectives [1]

par

Jean-Baptiste CAMPS

Comment se porte, à l'École des chartes, cette discipline au nom peu familier qu'est la philologie computationnelle [2] ? Plus qu'une réflexion théorique ou épistémologique, qui a déjà été entreprise par ailleurs [3], la présente contribution entend revenir, de manière pragmatique, sur ces quelques dernières années qui ont vu la mise en place ou le renforcement d'enseignements et de projets de recherche dans ce domaine, avant de tracer des voies possibles de maturation pour l'avenir. Pour ce faire, l'exposé suivra un plan qui reprend la distinction souvent faite entre l'ecdotique proprement dite, d'une part, avec ses étapes canoniques qui vont de la *recensio* à l'*emendatio*, et ce que l'on appelle parfois la « haute critique » – notamment la critique attributive – et l'histoire des textes. Le propos s'efforcera enfin de tracer quelques perspectives d'avenir pour la philologie, en tant que science se préoccupant d'étudier des évolutions culturelles, dans un contexte où la massification des données disponibles permet d'envisager des questions fondamentales à une échelle jusque-là difficile : centaines, milliers ou centaines de milliers

---

1. Au seuil de cet article, je tiens à remercier très chaleureusement pour leurs relectures et conseils Olivier Canteaut, Thibault Clérice, Simon Gabay, Vincent Jolivet, Chahan Vidal-Gorène, ainsi qu'Olivier Poncet et David Feutry.

2. Associant la science des textes et les méthodes computationnelles, c'est-à-dire fondées sur le calcul, la philologie computationnelle est une discipline qui a recours aux mathématiques, aux statistiques, à l'informatique et à l'intelligence artificielle pour la production de corpus de textes et leur analyse. Se rattachant au courant de la science des données, elle est la prolongation de la philologie par les méthodes modernes.

3. Je me permets en renvoyer à Jean-Baptiste Camps, « Où va la philologie numérique ? », dans *Fabula-LHT*, t. 20, 2018, https://www.fabula.org/lht/20/camps.html.

Jean-Baptiste Camps, maître de conférences, École nationale des chartes-PSL, Centre Jean-Mabillon (EA 3628), 65 rue de Richelieu, 75002 Paris. <jean-baptiste.camps@chartes.psl.eu>.

de textes et de documents, parfois sur plusieurs siècles ou à l'échelle de continents, traités à la fois en série et avec un très grand degré de finesse, jusqu'au niveau de l'occurrence mot ou même, en deçà, jusqu'au signe graphique [4].

I. Préambule : déluge de données et expertise philologique.

Nous vivons, depuis le début des années 2000, une explosion de la quantité d'information disponible qui aurait fait pâlir d'envie les générations précédentes d'historiens, notamment ceux qui, dans les années 1950 et 1960, devaient, pour la collecte de données numériques, se reposer sur le lent travail des perforatrices de cartes [5].

Ce « déluge de données » touche tous les domaines de la connaissance et a conduit à l'émergence d'un nouveau paradigme scientifique, qui place les données au centre, ce que Jim Gray qualifie de *« data intensive*

4. Un exemple de très grand corpus est celui des *Cartae Europae Medii Aevi* (CEMA), qui recueillent environ 250 000 éditions de chartes européennes collectées par Nicolas Perreaux, pour un total de 75 millions de mots ; Nicolas Perreaux, « Possibilities, challenges and limits of a European charters corpus (*Cartae Europae Medii Aevi* – CEMA) », 2021, *preprint* : https://hal.archives-ouvertes.fr/hal-03203029. Le *Corpus of French Literary Fictions (1050-1920)*, en cours de constitution, vise quant à lui à collecter les formes longues de la littérature narrative française depuis la chanson de geste jusqu'aux romans populaires du XIX^e^ siècle, en se rapprochant autant que possible de l'exhaustivité pour la production conservée (il contient déjà notamment 80 % des romans connus par le catalogue de la Bibliothèque nationale de France pour la période 1470-1600 et 70 % pour la période 1600-1700) ; Pierre-Carl Langlais *et al.*, « From Roland to Conan : First results on the corpus of French literary fictions (1050-1920) », *DH 2022 Tokyo*, à paraître.

5. Un exemple pionnier et souvent cité de ce type d'entreprises est celle menée par Roberto Busa pour la réalisation de l'*Index Thomisticus*, mais les historiens et philologues français n'ont pas été en reste. On se reportera notamment avec intérêt aux travaux de Michel Allard *et al.*, *Analyse conceptuelle du Coran sur cartes perforées*, Paris, 1963, ou bien à ceux de Dom Jacques Froger, « Emploi de la machine électronique dans les études médiévales », dans *Bulletin de philosophie médiévale*, t. 3, 1961, p. 177-188, et au panorama qu'il dresse des usages de l'informatique en études médiévales : édition critique, lexicographie, stylistique, traduction automatique et analyse grammaticale, profil d'auteurs et même, au titre de perspective d'avenir, « lecture automatique » ou bibliothèques de données numériques. Dans ce domaine, il faut souligner le rôle joué par les opératrices et programmeuses, qui a déjà été noté dans le cadre des projets de R. Busa par Melissa Terras, « Blog : For Ada Lovelace Day – Father Busa's female punch card operatives », 15 octobre 2013, http://melissaterras.blogspot.com/2013/10/for-ada-lovelace-day-father-busas.html (consulté le 18 novembre 2021). Cela suscite d'ailleurs dès l'époque des réflexions sur la conservation de ces données et la mutualisation de briques de code chez les archivistes : François Burckard, « Les archives et l'informatique en France, perspectives et directions de recherches », dans *La gazette des archives*, n^o^ 75, 1971, p. 159-177. On retrouve ainsi en germe trois étapes de la vie des données : ingénierie et production des données, exploitation scientifique et pérennisation.

scientific discovery », plus couramment nommée *data science* ou *science des données* (en réalité, plutôt, science *par* les données ou *intensive en données*) [6].

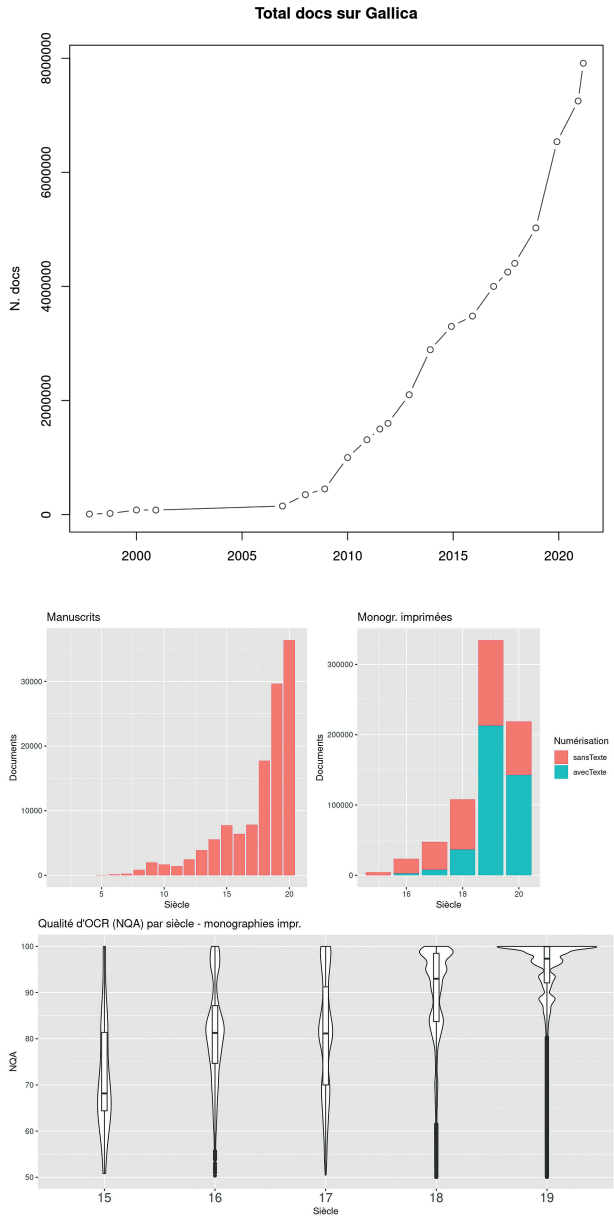
Ce « déluge » nous met ainsi à portée de quantités jusque-là non maniables de sources et pourrait bien renouveler l'ensemble des pratiques philologiques et historiques. Un cas particulièrement pertinent est celui des bibliothèques numériques. Pour celle de la Bibliothèque nationale de France (Gallica), cela s'est manifesté par une croissance exponentielle du nombre de documents disponibles, passant de 2 500 à sa création en 1997 à huit millions aujourd'hui (fig. 1).

Une quantité de données imposante, mais de quelles données s'agit-il ? En premier lieu, ces données sont encore essentiellement des fac-similés numériques, c'est-à-dire des photographies numériques des manuscrits, en d'autres termes, des images. L'intégration d'une couche de texte, généralement produite automatiquement par reconnaissance optique des caractères ou des écritures, est un enjeu, tout comme la quantité d'erreurs contenues dans ces transcriptions automatiques. Ce n'est pas le seul : la reconnaissance de la mise en page, du statut des éléments textuels ou non textuels distribués sur les pages des sources en question, la variété graphique (abréviations, graphies, etc.) appellent une série de traitements appuyés par l'intelligence artificielle et d'enrichissements pour pouvoir aller au-delà de la simple interrogation en texte brut.

Que l'on ne s'y trompe pas : ces vastes entrepôts numériques constituent de véritables mines d'informations, encore à peine effleurées, offrant peut-être la possibilité de réviser notre appréhension des productions écrites du passé, en dépassant les limites de ce qu'il est humainement possible de traiter pour aller au-delà de l'analyse de cas isolés, de l'étude d'un petit canon de textes et de documents, vers des modélisations de nos objets d'études plus globales et moins attachées à l'irréductibilité des cas particuliers. Et pourtant les méthodes traditionnelles de la critique philologique seraient bien en peine de manier ces vastes ensembles qui appellent, pour qu'il soit possible d'en tirer pleinement profit, à repenser nos procédés et à trouver la meilleure répartition et les meilleures synergies entre intelligence humaine et artificielle.

Pour ce faire, il est essentiel de concevoir des approches méthodologiques qui associent pleinement l'expertise philologique sur les sources, leur édition et leur critique, avec les apports des approches mathématiques, qu'il s'agisse d'analyse statistique, d'intelligence artificielle ou de modélisation, le tout étant servi par les capacités de traitement d'information et de calcul apportées par l'ordinateur. Si cela est sans doute assez nouveau pour les sciences du texte, il ne s'agit de rien d'autre que

6. Jim Gray, « Jim Gray on eScience : a transformed scientific method », dans *The fourth paradigm. Data-intensive Scientific Discovery*, éd. Tony Hey, Stewart Tansley et Kristin Tolle, Washington, 2009, p. XVII-XXXI.



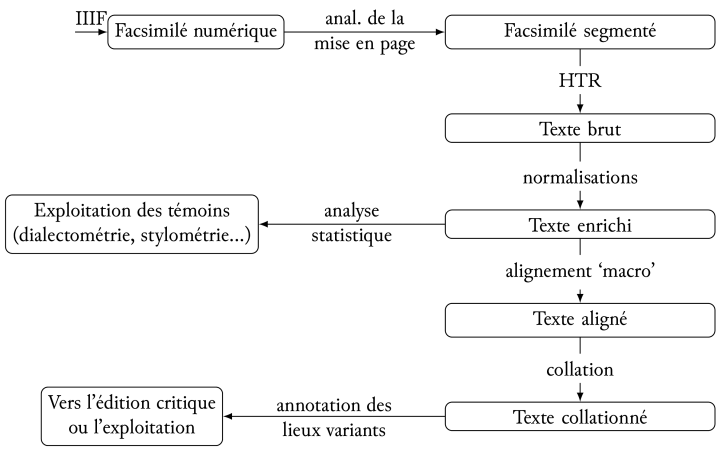
Ill. 1. Évolution du nombre total de documents sur Gallica (haut) ; répartition par type de source (manuscrit ou imprimé), date et existence ou non d'une couche de texte pour les numérisations de livres disponibles sur Gallica (centre) ; qualité moyenne desdites couches de texte, comme pourcentage de mots ne contenant pas d'erreur (bas).

de pratiques de sciences des données devenues désormais assez courantes et ordinaires dans de nombreuses autres disciplines scientifiques – la biologie par exemple.

Dans cette approche plaçant les données au centre [7], la transformation méthodologique se situe à la fois en amont, par la modification des modes de production des données, et en aval, par la révision de leurs méthodes d'analyse.

II. Ecdotique et intelligence artificielle : des chaînes de traitement du fac-similé numérique à la collation.

Au-delà de l'identification des sources et de leur description, la question centrale de l'ecdotique computationnelle est la conception de chaînes de traitement allant du fac-similé numérique à l'établissement et à l'enrichissement critique d'un texte, à l'instar de la démarche ecdotique « pré-numérique ». Différentes entreprises collectives engagées au sein de l'École des chartes témoignent de la mise en place d'une chaîne de traitement modulaire qui combine au mieux l'apprentissage machine et l'expertise humaine (fig. 2) [8]. Progressivement conçue et mise à jour depuis 2015, cette chaîne devrait aussi permettre l'inclusion d'une couche de texte à des manuscrits et incunables en français de la plate-forme Gallica [9].



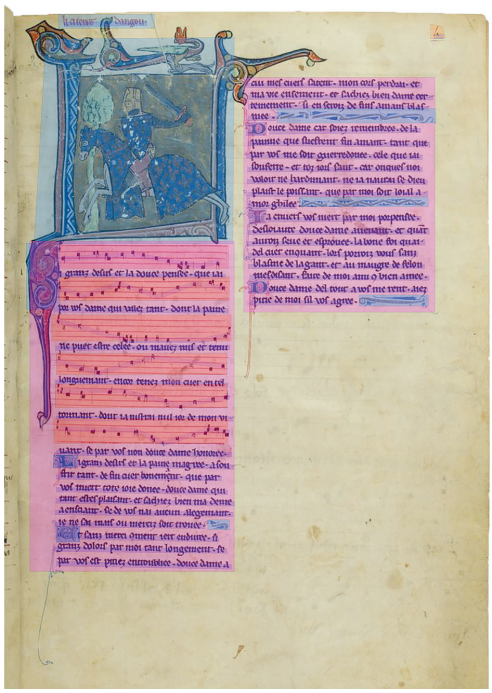
Ill. 2. Représentation des étapes conceptuelles d'une chaîne de traitement complète (J.-B. Camps, L. Ing et E. Spadini, « Collating medieval vernacular texts… »).

7. Voir J.-B. Camps, « Où va la philologie numérique ? »…

8. Jean-Baptiste Camps, Lucence Ing et Elena Spadini, « Collating medieval vernacular texts : aligning witnesses, classifying variants », dans *Digital Humanities Conference 2019, Complexities, Utrecht*, 2019, https://hal.archives-ouvertes.fr/hal-02268348/.

9. Dans le cadre du projet *Gallic(orpor)a*, liant l'Institut national de recherche en sciences et technologies du numérique (INRIA), l'université de Genève et l'École des chartes (Benoît Sagot, Simon Gabay, Ariane Pinche et Jean-Baptiste Camps).

1. Analyse de la mise en page. – La première étape de l'analyse des images est celle de la reconnaissance de la mise en page. *A minima*, il s'agit d'isoler les zones contenant le texte dont il faudra faire l'acquisition en les distinguant de toutes les autres, qu'elles portent par exemple des éléments de décor ou qu'elles soient relatives à l'organisation matérielle du volume (réclames, signatures, foliotation, etc.). Mais cette étape d'analyse de la mise en page peut aller plus loin et permettre une identification plus fine des différentes zones de la page. Pour permettre cette annotation, nous avons ainsi conçu une première version d'un vocabulaire contrôlé, SegmOnto, qui se fonde sur une distinction entre zones et lignes, et propose une série de types (contrôlés) et de sous-types (suggérés) pour chacune de ces classes (fig. 3) [10]. Avec suffisamment de données annotées, il devient possible d'entraîner des modèles d'analyse de la mise en page, circonscrivant et annotant automatiquement les zones [11].



Ill. 3. Bibl. nat. Fr., fr. 844 (chansonnier du roi), fol. 4, annoté par V. Mariotti et A. Pinche. Différents types de zones tirés du vocabulaire SegmOnto sont visualisés ici sous forme de rectangles de couleur (*Title* pour la rubrique, *Main* pour le corps de texte, *Decoration*, *DropCapital* pour les initiales, *Music*, *Numbering* pour le numéro de feuillet).

10. Simon Gabay *et al.*, « SegmOnto : common vocabulary and practices for analysing the layout of manuscripts (and more) », dans *16th International Conference on Document Analysis and Recognition (ICDAR 2021)*, 2021, https://hal.archives-ouvertes.fr/hal-03336528/.

11. Le logiciel Kraken dispose d'un moteur entraînable de reconnaissance de la mise en page : Benjamin Kiessling, « A modular region and text line layout analysis system », dans *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2020, p. 313-318, doi : 10.1109/ICFHR2020.2020.00064.

2. Reconnaissance automatique des écritures. – Si la reconnaissance optique des caractères (*optical character recognition* ou OCR) est considérée, pour les documents contemporains, comme un problème résolu depuis longtemps, il n'en va pas de même du traitement des documents imprimés anciens ou, à plus forte raison, des documents manuscrits. Dans ce domaine, toutefois, les progrès se sont révélés spectaculaires ces dernières années, grâce au renouvellement des technologies de réseaux de neurones et à l'émergence du paradigme de l'apprentissage profond (*deep learning*). Sous réserve de disposer de données d'entraînement en quantité suffisante, il devient ainsi possible d'entraîner des modèles capables de transcrire des manuscrits anciens avec un certain succès, voire un succès certain. En la matière, les progrès sont rapides. Lorsque nous – Thibault Clérice, Lucence Ing, Ariane Pinche et moi-même – avons entrepris en 2015-2016 l'entraînement de premiers modèles de transcription automatique appliqués à des manuscrits, tels que ceux de la chanson d'*Otinel*, des vies de saints en prose française ou du *Lancelot* en prose, les applications étaient encore rares [12] et la pratique la plus courante, avec le logiciel OCRopy, était d'entraîner un modèle par manuscrit, voire par main, à l'instar du travail sur les incunables des collègues du Center for Information and Language Processing de Munich [13]. Ces premières expériences en ont depuis lors stimulé d'autres, dans un cadre collaboratif, notamment sur l'occitan médiéval et prémoderne ou sur les imprimés français du XVII^e^ siècle [14].

---

12. On notera qu'en France, 2015 est l'année de lancement du projet HIMANIS, dirigé par Dominique Stutzmann à l'Institut de recherche et d'histoire des textes (IRHT), qui a permis l'indexation plein texte de plus de 75 000 pages de manuscrits médiévaux ; Dominique Stutzmann, Jean-François Moufflet et Sébastien Hamel, « La recherche en plein texte dans les sources manuscrites médiévales : enjeux et perspectives du projet HIMANIS pour l'édition électronique », dans *Médiévales. Langues, Textes, Histoire*, t. 73, 2017, p. 67-96, doi : 10.4000/medievales.8198.

13. Le logiciel OCRopy, développé par Thomas Breuel, a été pionnier dans l'utilisation de réseaux de neurones de type *long-short-term-memory* (LSTM) pour la prédiction du texte ; Thomas M. Breuel *et al.*, « High-performance OCR for printed English and Fraktur using LSTM networks », dans *2013 12th International Conference on Document Analysis and Recognition (ICDAR)*, 2013, p. 683-687, doi : 10.1109/ICDAR.2013.140. Il est l'ancêtre direct du logiciel Kraken que nous utilisons aujourd'hui, voir Benjamin Kiessling, « Kraken – an universal text recognizer for the humanities », 2019, https://dh-abstracts.library.cmu.edu/works/9912 (consulté le 13 décembre 2021). Les usages sur des données historiques d'OCRopy ont été, à ma connaissance, initiés à Munich par Uwe Springmann et David Kaumanns, *Ocrocis : a High Accuracy OCR Method to Convert Early Printings into Digital Text*, 2015, http://cistern.cis.lmu.de/ocrocis/tutorial.pdf. On se reportera aussi, pour la présentation de premiers résultats, à Jean-Baptiste Camps, *La « Chanson d'Otinel » : édition complète du corpus manuscrit et prolégomènes à l'édition critique*, thèse de doctorat, dir. Dominique Boutet, univ. Paris-Sorbonne, 2016, https://halshs.archives-ouvertes.fr/tel-01664932, annexe B, p. 363, ainsi qu'au tutoriel : *id.*, « Homemade manuscript OCR (1) : OCRopy », *Sacré Gr@@l*, 2017, http://graal.hypotheses.org/786 (consulté le 6 février 2018).

14. Jean-Baptiste Camps et Gilles Guilhem Couffignal, « La production de corpus d'occitan médiéval et prémoderne : problèmes et perspectives de travail », dans *Fidelitats*

Car la problématique de la reconnaissance des écritures n'est pas qu'une problématique technique. Elle se pose d'abord comme une question de disponibilité de données d'entraînement, c'est-à-dire de données alignant fac-similé et transcription *a minima* au niveau de la ligne. En la matière, sans que cela fût l'objectif initial, il est permis de dire que l'École des chartes a été pionnière avec le projet de « fac-similés interactifs » lancé par Gautier Poupeau sur le portail Thélème [15], et qui trouve sa prolongation la plus récente dans le projet d'*Album de diplomatique européenne en ligne* (ADELE) [16]. Plus généralement, la tension se place entre la réutilisation – et l'alignement avec le fac-similé – de données au volume substantiel, comme celles qui résultent de la numérisation d'éditions critiques, mais qui ne sont pas conçues pour l'entraînement, et la production patiente de données *ad hoc* par transcription ou relecture humaine, selon des critères précisément définis, le plus souvent pour maintenir, pour chaque signe présent sur la source, une transcription unique. La première approche permet de tirer profit de deux siècles de travail philologique, *via* la réutilisation relativement rapide d'éditions préexistantes. Ce fut par exemple le choix réalisé par le projet HIMANIS, qui a ainsi pu exploiter la numérisation initiée par Olivier Canteaut à l'École des chartes de l'édition faite par Paul Guérin des actes royaux concernant le Poitou [17], ou bien plus récemment par le projet eNotre-Dame-de-Paris dirigé par Julie Claustre et Darwin Smith, auquel l'École est associée. Ce dernier projet a en effet opté pour une approche hybride, dans laquelle un modèle entraîné sur des données préexistantes est ensuite parfait et affiné (opération que l'on désigne généralement par l'expression anglaise de *fine tuning*) sur des données

---

e dissidencias/Fidélités et Dissidences, Actes du XII^e Congrès international de l'Association internationale d'études occitanes, Albi 2017, https://halshs.archives-ouvertes.fr/halshs-02050089/document ; Claire Jahan et Simon Gabay, *OCR17+ – Layout Analysis and Text Recognition for 17th c. French Prints*, 2021, https://github.com/e-ditiones/OCR17plus (consulté le 13 décembre 2021).

15. Pour une présentation contemporaine des premières initiatives de l'École en la matière, voir Olivier Guyotjeannin et Gautier Poupeau, « Le projet d'édition électronique du cartulaire blanc de l'abbaye de Saint-Denis et les projets électroniques de l'École nationale des chartes », dans *Le médiéviste et l'ordinateur*, t. 42, 2003, p. 93-99, https://www.persee.fr/doc/medio_0223-3843_2003_num_42_1_1598. Pour la période suivante, on se reportera à Florence Clavaud, « Les éditions électroniques de l'École nationale des chartes : objectifs, principes, outils et perspectives », dans *Bulletin de la Commission royale d'Histoire*, t. 176, 2010, p. 107-120, https://www.persee.fr/doc/bcrh_0001-415x_2010_num_176_1_1079.

16. *Album de diplomatique européenne en ligne*, éd. Olivier Guyotjeannin, Paris, 2022, https://adele.chartes.psl.eu/.

17. Paul Guérin et Louis Celier, *Recueil des documents concernant le Poitou contenus dans les registres de la chancellerie de France (Poitiers, 1881)*, éd. num. Olivier Canteaut, Olivier Guyotjeannin et Vincent Jolivet, Paris, 2011, http://corpus.enc.sorbonne.fr/actesroyauxdupoitou/ ; voir D. Stutzmann, J.-F. Moufflet et S. Hamel, « La recherche en plein texte… ».

produites à nouveaux frais. Ici, l'héritage de pratiques de transcription bien ancrées (et de modèles préexistants) facilite la production nouvelle d'un corpus homogène par différents scripteurs.

La seconde approche, elle, permet un contrôle plus fin des principes de transcription, une meilleure correspondance signe à signe à la source et une plus grande marge de manœuvre dans la définition des principes de transcription. Elle permet en outre généralement de conserver la trace des différentes étapes du travail éditorial : abréviations et leur résolution ultérieure, segmentation des mots, normalisations éventuelles, ponctuation, etc. Adoptée par de nombreux projets, elle pose la question du niveau auquel situer la transcription et d'éventuels traitements ultérieurs de normalisation.

La question de l'efficacité respective des deux approches n'est, elle, pas entièrement tranchée et n'a pas de réponse évidente. Sur un premier niveau, il est certes possible d'en comparer l'efficacité, sous conditions contrôlées et toutes choses égales par ailleurs, pour obtenir un texte final normalisé ; en cela, nos premières expériences ne montrent pas de gain net de la seconde approche, pour le latin du moins, la question pouvant être sensiblement différente pour l'ancien français et demandant des recherches supplémentaires qui sont en cours [18]. Sur un second niveau, les paramètres conditionnant le choix de l'une ou l'autre approche dépassent ce cadre strict de la performance mesurée par le score obtenu sur le ou les indicateurs choisis, pour intégrer des éléments de transparence et de vérifiabilité de la démarche ecdotique.

Reste la question des choix de transcription à opérer, en particulier dans le cadre de la seconde approche, et de la décomposition en différentes étapes du traitement ecdotique. S'il n'est pas lieu ici de disserter longuement sur la définition de principes de transcription [19], on notera que dans la typologie générale des possibles niveaux de transcription, telle que définie par Peter Robinson et Elizabeth Solopova et précisée par Dominique Stutzmann [20], seuls les niveaux graphématique

18. Jean-Baptiste Camps, Chahan Vidal-Gorène et Marguerite Vernet, « Handling heavily abbreviated manuscripts : HTR engines vs text normalisation approaches », dans *International Conference on Document Analysis and Recognition*, 2021, p. 306-316, doi : 10.1007/978-3-030-86159-9_21 ; une entreprise collective est en cours, tirant profit à la fois de corpus réalisés à l'École et à l'IRHT.

19. On se reportera à ce sujet à l'article de Frédéric Duval, « Transcrire le français médiéval : de l'"instruction" de Paul Meyer à la description linguistique contemporaine », dans *Bibliothèque de l'École des chartes*, t. 170, 2012, p. 321-342, doi : 10.3406/bec.2012.464252. Voir également J.-B. Camps, *La « Chanson d'Otinel »*…, p. CXCV-CCXVI.

20. Peter Robinson et Elizabeth Solopova, « Guidelines for transcription of the manuscripts of the wife of Bath's prologue », dans *The Canterbury Tales Project Occasional Papers*, t. I, éd. Norman Blake et Peter Robinson, Oxford, 1993, p. 19-52, http://server30087.uk2net.com/canterburytalesproject.com/pubs/op1-transguide.pdf (consulté le 1er avril 2015) ; Dominique Stutzmann, « Paléographie statistique pour décrire, identifier, dater… Normaliser pour coopérer et aller plus loin ? », dans

(conservant les graphèmes de la source, sans leurs variantes formelles) et normalisé sont régulièrement utilisés. Pour le niveau qu'on appellera le mieux allographétique, qui enregistre les variantes de forme de chaque graphème, la variété même d'appellations rencontrées (transcription imitative, hyperdiplomatique, fac-similaire, etc.) témoigne de la relative rareté et de l'absence de systématisation de sa mise en œuvre, qui intéresse pourtant l'analyse paléographique computationnelle et la construction de profils de scribes. Quant au niveau « graphique », qui enregistrerait « chaque trait du manuscrit, chaque espace […] jusqu'à la décomposition des formes de lettres en traits distincts » [21], il semble bien peu utilisé. Du côté de l'École, un travail est en cours – mené par Ariane Pinche, Frédéric Duval et moi-même – dans le cadre du projet CREMMA LAB pour définir un ensemble de critères de transcription aussi homogènes que possible pour les transcriptions graphématiques (au moins) des manuscrits français, en termes de choix de représentation notamment du système abréviatif, de la segmentation des mots, de la ponctuation et des diacritiques anciens, des modes de correction ou bien encore de l'allographie. Des premiers entraînements de modèle ont permis de vérifier la pertinence de disposer d'échantillons variés par leur nature mais harmonisés par leur transcription pour l'entraînement de modèles plus généralistes [22].

Pour ce qui est de la dissociation de l'étape de reconnaissance des écritures proprement dite, et celles, ultérieures, qui sont nécessaires à l'obtention d'une transcription normalisée, les recherches nous ont d'abord menés vers l'entraînement de modèles de segmentation en mots et de modèles de résolution des abréviations. L'espacement dans les manuscrits est en effet une question complexe, qui ne peut totalement être réduite à une opposition binaire entre présence ou absence d'espace, et dont la retranscription pose des difficultés autant aux humains qu'aux

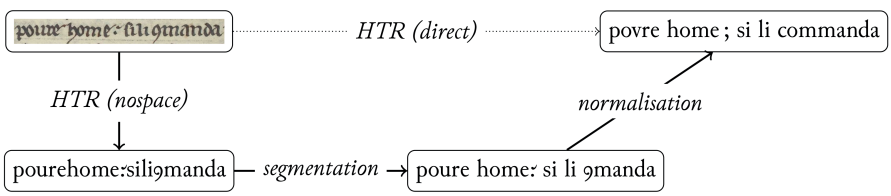
---

Kodikologie und Paläographie im digitalen Zeitalter 2 / Codicology and Palaeography in the Digital Age 2, éd. Fischer Franz, Christiane Fritz et Georg Vogeler, Norderstedt, 2011 (Schriften des Instituts für Dokumentologie und Editorik, 3), p. 247-277, https://halshs.archives-ouvertes.fr/halshs-00596970/.

21. « Every mark in the manuscript, every space […] even to the point of decomposition of letter forms into discrete marks », P. Robinson et E. Solopova, « Guidelines… ».

22. Avec un échantillon de 17 431 lignes tirées de 11 manuscrits français, le modèle « Bicerin » comptabilise un taux d'erreur caractère inférieur à 5 % ; Ariane Pinche, *CREMMA Medieval, an Old French Dataset for HTR and Segmentation*, 2021, https://github.com/HTR-United/cremma-medieval (consulté le 13 décembre 2021). Plus généralement, l'initiative HTR United, lancée par Thibault Clérice et Alix Chagué a permis de commencer à mutualiser des corpus d'entraînement en quantité intéressante ; Alix Chagué, Thibault Clérice et Laurent Romary, « HTR-United : mutualisons la vérité de terrain ! », 2021, https://hal.archives-ouvertes.fr/hal-03398740. On y trouve notamment des données produites dans le cadre du projet LECTAUREP (LEcture Automatique de REpertoires), qui associe Minutier central des notaires de Paris des Archives nationales, INRIA et EPHE, et auquel ont participé des diplômés de master de l'École.

modèles automatiques [23]. Quant à la résolution des abréviations et aux autres normalisations telles que l'insertion de lettres ramistes (*j* distinct de *i* et *v* distinct de *u* selon leur valeur phonologique), elles ont parfois été traitées comme des tâches de reconnaissance des écritures ; mais il est aussi possible de décomposer ce travail en une série de traitements spécialisés qui permettent de garder trace de chaque étape intermédiaire (fig. 4). Si des premières tentatives réalisées avec cette seconde approche se sont révélées prometteuses pour l'ancien français et ont permis de réaliser des analyses stylométriques sur un corpus relativement « brut » de pages manuscrites reconnues, des recherches complémentaires (en cours) s'avèrent nécessaires pour comparer les avantages, inconvénients et performances des deux approches, à la fois sur le court terme (qualité des résultats obtenus et effort nécessaire) et sur le long terme (construction ou non de corpus de transcriptions enregistrant à la fois les abréviations et leur résolution, une segmentation proche de la source et sa normalisation) [24].



Ill. 4. Approche directe (pointillés) *vs* approche modulaire (flèches pleines). L'image est segmentée en zones et lignes par un algorithme d'analyse de la mise en page, puis, soit le texte de la ligne est reconnu directement sous forme normalisée, soit il l'est sous forme graphématique, avant de passer par des étapes de segmentation et de résolution des abréviations.

3. Normalisations et lemmatisation. – D'autres normalisations ou enrichissements disposent d'une pertinence assez générale, en dépit du caractère technique que l'on peut parfois leur attribuer. Il en va ainsi de la lemmatisation, opération qui consiste, pour toutes les formes (fléchies,

23. F. Duval, « Transcrire le français médiéval… », p. 337 ; Maria Careri *et al.*, *Album de manuscrits français du XIII^e siècle*, Rome, 2001 ; A. Pinche, *CREMMA Medieval…*

24. Pour le travail initial sur le ms. fr. 411, on se reportera à Jean-Baptiste Camps, Thibault Clérice et Ariane Pinche, « Noisy medieval data, from digitized manuscript to stylometric analysis : Evaluating Paul Meyer's hagiographic hypothesis », dans *Digital Scholarship in the Humanities*, t. 36, 2021, p. ii49-ii71, https://doi.org/10.1093/llc/fqab033 ; ainsi qu'à la présentation des outils, Boudams (Thibault Clérice, « Evaluating deep learning methods for word segmentation of scripta continua texts in Old French and Latin », dans *Journal of Data Mining & Digital Humanities*, 2020, doi : 10.46298/jdmdh.5581) et Pie (Enrique Manjavacas, Akos Kádár et Mike Kestemont, « Improving lemmatization of non-standard languages with joint learning », 2019, *arXiv:1903.06939*). Pour le travail sur le latin, on consultera J.-B. Camps, C. Vidal-Gorène et M. Vernet, « Handling heavily abbreviated manuscripts… ».

porteuses de variantes graphiques, etc.) d'un lexème donné, à les rattacher à une forme canonique unique, le *lemme*, généralement l'infinitif pour un verbe, le masculin singulier pour les noms et adjectifs : ainsi, *aimerons*, *aimé*, *aymée*, etc., seront rattachés au lemme *aimer*. Cette opération, presque indispensable aux enquêtes linguistiques et lexicographiques, facilite en outre bon nombre d'interrogations thématiques ou lexicales sur les textes, leur indexation, leur comparaison et une variété d'opérations computationnelles utiles à l'historien [25].

Dans ce domaine, l'École a porté de longue date des initiatives telles que le projet OMNIA, mis en place en 2009, qui a permis la création de paramètres de lemmatisation pour le latin médiéval [26]. Depuis le lancement du projet LAKME en 2015 (puis, par la suite, du projet OMÉLIE), un travail chronophage de construction de corpus annotés (mot par mot) et d'entraînements de modèles a été l'objet d'un vaste effort collectif permettant la création de modèles performants pour l'ancien français, le français classique et moderne, le latin et bientôt le moyen français et peut-être le franco-italien [27].

Pour la production de modèles de lemmatisation, comme pour les autres tâches impliquant de l'apprentissage machine, la qualité des données produites ou revues par les humains est centrale. Dans ce contexte, sous l'impulsion de Thibaut Clérice, un écosystème d'ingénierie complet dédié à la lemmatisation a été mis en place, incluant une application dédiée à la post-correction et permettant le travail collaboratif, l'usage de référentiels (lexique de lemmes permis, liste d'étiquettes), le suivi dans le temps des modifications, le traitement des corrections par lots, etc. [28].

---

25. Renaud Alexandre, Bruno Bon et Anita Guerreau-Jalabert, « Variations graphiques, variations morphologiques et lemmatisation du latin médiéval », dans *Amicorum societas. Mélanges offerts à François Dolbeau pour son 65e anniversaire*, éd. Jacques Elfassi, Cécile Lanéry et Anne-Marie Turcan-Verkerk, Florence, 2013, p. 3-18 ; Gard Jenset et Barbara McGillivray, *Quantitative Historical Linguistics : a Corpus Framework*, Oxford, 2017, part. chap. 4, « Historical corpus annotation » et 5, « (Re)using resources for historical languages », p. 98-152.

26. Bruno Bon, « OMNIA : outils et méthodes numériques pour l'interrogation et l'analyse des textes médiolatins (3) », dans *Bulletin du centre d'études médiévales d'Auxerre BUCEMA*, t. 15, 2011, https://journals.openedition.org/cem/12015.

27. Jean-Baptiste Camps *et al.*, « Corpus and models for lemmatisation and POS-tagging of Classical French theatre », dans *Journal of Data Mining and Digital Humanities*, 2020, doi : 10.46298/jdmdh.6485 ; J.-B. Camps *et al.*, « Corpus and models for lemmatisation and POS-tagging of Old French », 2021, *arXiv:2109.11442*. En ce qui concerne le franco-italien, on se reportera à l'article de Floriana Ceresato, « L'analisi lessicale dell'Entrée d'Espagne : bilancio di una prima sperimentazione », à paraître dans *Francigena*. Ces modèles ont été entraînés avec le logiciel Pie (E. Manjavacas, A. Kádár et M. Kestemont, « Improving lemmatization… »). Pour ce qui est de l'API, il s'agit de Thibault Clérice, *Pie Extended, an extension for Pie with pre-processing and post-processing*, 2020, https://doi.org/10.5281/zenodo.3883589.

28. Thibault Clérice et Julien Pilla, *Pyrrha*, 2021, https://github.com/hipster-philology/pyrrha (consulté le 13 décembre 2021). Pour une présentation générale du

En outre, des outils permettant le contrôle de la construction des jeux de données d'entraînement et en garantissant la reproductibilité, de même qu'une API (web et Python) gérant les différentes étapes entourant la lemmatisation, ont été mis en place [29].

4. Collation et autres enrichissements. – Parmi les usages possibles d'un texte enrichi d'informations linguistiques (lemme, catégories grammaticales, flexion), la collation constitue une étape ultérieure caractéristique d'une démarche ecdotique tendant vers une édition critique, ou au moins la mise en regard de différents témoins. Pour ce faire, des algorithmes dédiés à l'alignement automatique de textes existent, notamment empruntés à l'alignement de séquences en biologie computationnelle [30]. Toutefois, la variation observée dans les textes anciens, notamment les textes vernaculaires médiévaux, n'a que peu à voir avec celle des séquences génomiques. Elle conjoint en réalité différents types de variations, du plus (paléo)graphique au plus substantiel, qu'il importe généralement de pouvoir traiter de manière différenciée, en les neutralisant ou non. Pour ce faire, nous avons cherché à développer des outils qui fassent fonctionner ensemble un alignement macrostructurel, une lemmatisation visant à effacer la variation graphique, un algorithme d'alignement proprement dit (CollateX) et un algorithme dédié à l'annotation semi-automatique des variantes, pour distinguer variantes graphématiques, flexionnelles, morphosyntaxiques ou lexicales (dérivationnelles, paradigmatiques ou plus nettement séman-tiques) [31].

---

fonctionnement utilisateur, on se reportera à Ariane Pinche, « Annoter facilement un corpus complexe », dans *Actes des Rencontres lyonnaises des jeunes chercheurs en linguistique historique*, éd. Timothée Prémat et Ariane Pinche, Lyon, 2019 (Diachronies contemporaines), p. 48-58, https://halshs.archives-ouvertes.fr/halshs-02330147.

29. Thibault Clérice, Vincent Jolivet et Julien Pilla, « Building infrastructure for annotating medieval, classical and pre-orthographic languages : the Pyrrha ecosystem », *DH 2022 Tokyo*, hal-03606756. Une API (*application programming interface* ou interface de programmation d'applications) est « un ensemble normalisé de classes, de méthodes, de fonctions et de constantes qui sert de façade par laquelle un logiciel offre des services à d'autres logiciels » (*Wikipedia francophone*, art. « Interface de programmation », 13 mars 2022).

30. Un des logiciels parmi les plus utilisés est CollateX. Voir Ronald Haentjens Dekker, *et al.*, « Computer-supported collation of modern manuscripts : CollateX and the Beckett Digital Manuscript Project », dans *Digital Scholarship in the Humanities*, t. 30, 2015, p. 452-470, http://dsh.oxfordjournals.org/content/30/3/452.abstract.

31. J.-B. Camps, L. Ing et E. Spadini, « Collating medieval vernacular texts… ». L'outil en question est disponible sous la forme d'un prototype. Lucence Ing a également poursuivi le travail sur la collation assistée par ordinateur dans le cadre de sa thèse de doctorat : Lucence Ing, *Disparitions lexicales en diachronie : traitements automatiques sur le Lancelot en prose*, thèse de doctorat, École nationale des chartes, http://www.theses.fr/s221114.

D'autres enrichissements sont également possibles, tirant profit de l'annotation linguistique. Il en va ainsi de la reconnaissance d'entités nommées : personnes, lieux, organisation, etc. [32].

5. Du numérique pour quelles éditions ? – Ces chaînes de production, qui vont permettre de plus en plus facilement de faire l'acquisition du texte de différents témoins vont probablement soulever à nouveau la question de la signification et du rôle de l'acte d'édition des textes, de ses formes et des résultats que l'on attend d'elle. Comme F. Duval l'a souligné il y a quelques années, jusqu'ici l'édition numérique est restée en partie captive d'un modèle centré sur le document (unique) qui a laissé peu de place aux démarches critiques et aux éditions reconstructionnistes ou tournées vers la tradition [33]. Outre la prégnance chez les premiers défenseurs de l'édition électronique de modèles hérités de la tradition de la « nouvelle philologie » marqués par les thèses de Bernard Cerquiglini et son éloge de la variante [34], cet état de fait peut s'expliquer également par la lourdeur et le caractère chronophage du surcroît de

32. De premières expérimentations concernant l'annotation des entités nommées des cartulaires latins ont été entreprises à l'École, sous l'impulsion de Vincent Jolivet. Elles s'appuient sur les éditions numériques des *Cartulaires latins d'Île-de-France*, éd. Olivier Guyotjeannin *et al.*, Paris, 2009 (Éditions en ligne de l'École des chartes, 11), http://elec.enc.sorbonne.fr/cartulaires/. L'abondante annotation des personnes et des lieux de ces fichiers TEI ont récemment été revus et corrigés par Elena Ghiringhelli et Marguerite Vernet pour fournir une première vérité de terrain afin d'entraîner des modèles de reconnaissance.

33. F. Duval, « Pour des éditions numériques critiques : l'exemple des textes français », dans *Médiévales*, t. 73, 2017, p. 13-29, doi : 10.4000/medievales.8165. Cela vaut en tout cas pour l'édition des textes littéraires ; la question se pose de manière sensiblement différente dans le domaine des éditions de textes documentaires, dont les originaux sont plus souvent conservés ou les copies moins éloignées (dans l'espace, si ce n'est pas dans le temps) de l'original et les intermédiaires perdus moins nombreux, etc. Dans ce domaine, des premières expérimentations ont été menées à l'École des chartes, et on se référera notamment à Camille Desenclos et Vincent Jolivet, « Diple, propositions pour la convergence de schémas XML/TEI dédiés à l'édition de sources diplomatiques », dans *Digital Diplomatics : the Computer as a Tool for the Diplomatist ?*, éd. Antonella Ambrosio, Sébastien Barret et Georg Vogeler, Cologne/Weimar/Vienne, 2014, p. 23-30, https://hal.archives-ouvertes.fr/hal-01317540, ainsi qu'au prototype suivant, *L'édit de Nantes et ses antécédents (1562-1598)*, dir. Bernard Barbiche, Paris, 2005 ; 2e éd. 2009 (Éditions en ligne de l'École des chartes, 5), http://elec.enc.sorbonne.fr/editsdepacification/. Toutefois, même à l'École des chartes, l'édition numérique a fourni l'occasion d'expérimentations tournées vers les documents, telle que l'*Édition critique des carnets de prison et de la correspondance privée d'Henri Delescluze à Belle-Île (1851-1853)*, éd. Christine Nougaret et Florence Clavaud, Paris, 2015 (Éditions en ligne de l'École des chartes, 26). Un cas intermédiaire est fourni par l'édition des *Chroniques latines de Saint-Denis*, éd. Pascale Bourgain, Paris, 2005 et 2010 (Éditions en ligne de l'École des chartes, 13), qui édite un témoin spécifique, corrigé ponctuellement grâce à un manuscrit de contrôle.

34. Bernard Cerquiglini, « Éloge de la variante », dans *Langages*, t. 17, 1983, p. 25-35, doi : 10.3406/lgge.1983.1140.

travail généré par la production d'une édition électronique – à tel point que l'on peut se demander si cet éloge du document ne relève pas parfois plutôt d'une justification *a posteriori*.

Si la rareté du travail critique sur la tradition au sein des éditions numériques est regrettable, il n'en reste pas moins que la production de données réutilisables et cumulables est un aspect absolument central et fondamental (au sens propre), sans lequel aucune autre approche computationnelle n'est possible. Si la tâche peut parfois sembler fastidieuse et ingrate – mais n'est-ce pas la gloire des chartistes que de ne pas reculer devant ce type d'entreprise ? – et peu payante sur le court terme, elle conditionne toutes les possibilités d'automatisation ultérieures et la capacité à tester les hypothèses au niveau le plus macro.

Il existe en réalité sans doute une tension entre la nature d'édition (qui fixe un texte pour la lecture directe humaine) et celle de données (qui fournit des informations structurées pour des traitements sériels et informatiques). La possibilité de générer rapidement et de manière homogène de vastes corpus, grâce à l'intelligence artificielle, et de les interroger par des méthodes computationnelles pour chercher à observer des tendances ou des variations massives, sur une diachronie longue ou une diatopie large, repose quelque peu la question de la pertinence de l'effort consacré à la lente édition individuelle d'un texte unique, par rapport, par exemple, à la constitution de jeux de données d'entraînement échantillonnés. Si l'édition veut garder sa pertinence, elle doit être résolument critique et constituer un outil de connaissance problématisé d'une tradition.

III. La critique des textes face aux méthodes computationnelles.

Mais que faire de ces données et pourquoi se donner la peine de les collecter ? Et surtout, comment traiter des corpus de taille sans cesse croissante ? Un certain nombre de méthodes d'analyse justifient le travail de collecte et d'annotation décrit plus haut et en même temps rendent exploitables les jeux de données qui en résultent pour la production de connaissances nouvelles.

1. Données visuelles et paléographie. – Outre la reconnaissance automatique de la mise en page et du texte, déjà évoquée, le traitement de données visuelles *via* l'apprentissage profond est riche d'applications aux documents anciens. Le projet Filigranes pour tous utilise ainsi une architecture fondée sur des réseaux de neurones convolutifs pour réaliser des rapprochements entre un filigrane à identifier et une

base de données contributive de plus de 17 000 entrées [35]. Ailleurs, des compétitions visent au classement typologique des écritures ou à l'identification automatique des scripteurs *via* cette même catégorie de méthodes. Ainsi, en 2019, une compétition a montré qu'il était possible, à partir d'un ensemble de 20 000 images représentant environ 10 000 scripteurs de livres manuscrits, lettres et chartes ou documents légaux, provenant de bibliothèques numériques patrimoniales, de reconnaître le scripteur dans plus de 97 % des cas [36]. En outre, des expérimentations récentes menées par Chahan Vidal-Gorène sur les écritures arméniennes, dont la classification et la datation restent plus problématiques encore que pour les écritures latines, ont montré qu'il était possible d'affiner la typologie des écritures arméniennes et d'en améliorer ainsi la classification automatique grâce à un jeu de données représentatif constitué de seulement 2 300 images [37].

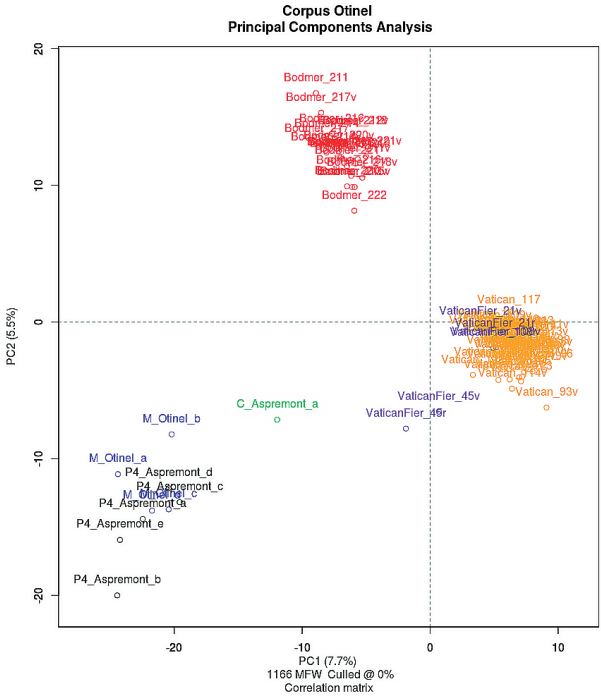
Sortant du domaine de l'analyse d'image pour entrer dans celui de l'analyse de données textuelles, il est également possible de construire, par analyse statistique, des profils de scribes qui tiennent comptent des allographes et signes abréviatifs particuliers utilisés par chacun – si tant est que des données allographétiques soient disponibles – et de produire des analyses multivariées reposant sur ceux-ci (fig. 5) [38]. Ce type d'analyse, par la nature des données qu'il envisage, se situe également au seuil d'autres formes d'analyse des données textuelles qui tiennent également compte de la variation régionale et individuelle dans les codes graphiques.

35. Oumayma Bounou *et al.*, « A web application for watermark recognition », dans *Journal of Data Mining and Digital Humanities*, t. 24, 2020, doi : 10.46298/jdmdh.6220.

36. Vincent Christlein *et al.*, « ICDAR 2019 competition on image retrieval for historical handwritten documents », dans *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019, p. 1505-1509, doi : 10.1109/ICDAR.2019.00242.

37. Chahan Vidal-Gorène et Aliénor Decours-Perez, « A computational approach of Armenian paleography », dans *International Conference on Document Analysis and Recognition – Workshops*, Lausanne, 2021, p. 295-305, doi : 10.1007/978-3-030-86159-9_20.

38. Pour l'analyse présentée en figure, voir J.-B. Camps, *La « Chanson d'Otinel »,* p. LIV.



	1	2
_ɑ	0,58	0,57
ɑl	0,55	0,57
pɑ	0,57	0,53
de	0,61	0,46
ɑr	0,52	0,55
vɑ	0,59	0,46
dɪ	0,59	0,45
_d	0,58	0,45
_k	0,66	0,30
_F	0,20	0,66
_ò	0,41	0,41
ſ_	0,72	0,11

Figure 5 : Analyse en composantes principales [39] de feuillets transcrits d'un corpus de manuscrits épiques (bigrammes de caractère), laissant à voir une disposition chronologique des différents manuscrits (à gauche) et sélection de quelques bigrammes de caractères (séquences de deux caractères consécutifs, par exemple, « sé », « éq », « qu », « ue », « en », « nc », « ce », « e_ », etc.) avec une forte contribution aux axes, parmi les quarante premiers (à droite).

2. Analyse de données textuelles : scriptométrie et stylométrie. – L'analyse de données textuelles est un champ qui a connu des développements assez anciens, notamment du côté de la stylométrie (la mesure du style), depuis le XIXe siècle [40]. Deux champs de ce vaste domaine sont peut-être

39. L'analyse en composantes principales est une analyse par réduction de la dimensionnalité, qui cherche, dans un espace de dimension *n* (défini par *n* variables, ici les bigrammes de caractères), à réduire ce nombre de dimensions corrélées entre elles en un plus petit nombre, décorrélées les unes des autres (appelées des « composantes principales »). On en visualise généralement les quelques dimensions les plus importantes (les deux ou trois premières, qui, dans notre cas, totalisent 13,2 % de l'information initiale). Il faut ensuite chercher à identifier les structures sous-jacentes résumées par ces axes.

40. Sur ce sujet, Jean-Baptiste Camps, « Computation », dans *En quête de sources. Dictionnaire critique*, dir. Frédéric Duval, Paris, 2021, p. 110-114 ; Florian Cafiero et Jean-Baptiste Camps, *Affaires de style*, Paris, 2022. Une synthèse désormais un peu ancienne est fournie par David I. Holmes, « The evolution of stylometry in humanities

d'un intérêt particulier pour les études philologiques, la dialectométrie ou scriptométrie, d'une part, qui s'intéresse aux variations linguistiques dans le temps et l'espace en tant qu'elles permettent notamment de localiser des textes et documents, et la stylométrie, d'autre part, qui permet en particulier de démêler des débats d'attribution des textes.

Qu'on la nomme dialectométrie (de dialecte, variante régionale parlée) ou plus proprement scriptométrie (de *scripta*, variante régionale écrite), l'étude quantitative de la variation diatopique dans des corpus de documents est pratiquée au moins depuis les années 1970 et 1980 et appliquée aux documents en ancien français depuis les travaux d'Anthonij Dees [41]. Pouvant avoir une vocation descriptive – qu'est-ce qui caractérise telle ou telle *scripta* régionale, ou de telle période –, elle peut également servir à fournir des hypothèses de datation et de localisation pour des documents. Il est ainsi possible, par exemple, de chercher à regrouper de manière non supervisée un corpus de copies de chansons de geste, ou bien, sur la base d'un corpus d'entraînement composé, idéalement, de documents datés et localisés (des chartes par exemple), d'entraîner des modèles d'apprentissage machine pour leur permettre d'attribuer des documents à l'une des classes vues à l'entraînement [42].

La stylométrie, elle, s'intéresse à la variation individuelle dans la langue, et étudie ce que l'on nomme des *idiolectes*, c'est-à-dire des variétés de langue propres à un individu, notamment dans une perspective d'attribution des textes à l'autorité disputée et anonyme. Ses fondements sont similaires à ceux des autres champs de l'analyse de données textuelles – quantification de certaines propriétés des textes *via* le compte d'occurrences, analyse statistique multivariée et apprentissage machine – mais c'est la nature des caractères étudiés qui s'en différencie. La stylométrie s'intéresse en effet aux propriétés (formelles) les moins conscientes

---

scholarship », dans *Literary and Linguistic Computing*, t. 13, 1998, p. 111-117, doi : 10.1093/llc/13.3.111.

41. Jean Séguy, « La dialectométrie dans l'*Atlas linguistique de la Gascogne* », dans *Revue de Linguistique romane*, t. 37, 1973, p. 1-24, pour ce qui est peut-être la première occurrence du terme de dialectométrie, et, pour celui de scriptométrie, peut-être Charles Théodore Gossen, « Méditations scriptologiques », dans *Cahiers de civilisation médiévale*, t. 22, 1979, p. 263-283, cité par Paul Videsott, « Introduzione », dans *Padania scrittologica*, Berlin, 2009, p. 7-268, https://doi.org/10.1515/9783484970458.7 (consulté le 18 janvier 2022).

42. Sur la première perspective, notablement maniée par Hans Goebl, on pourra consulter son article, « L'aménagement scripturaire du Domaine d'Oïl médiéval à la lumière des calculs de localisation d'Anthonij Dees effectués en 1983 : une étude d'inspiration scriptométrique », dans *Medioevo romanzo*, *Seminario 2011: Il problema della scripta*, Venise, 13-14 ottobre 2011, http://www.medioevoromanzo.it/modules/content/index.php?id=14. Pour le corpus épique, les résultats de ma première enquête se trouvent dans Jean-Baptiste Camps, « Manuscripts in time and space : experiments in scriptometrics on an old French corpus », dans *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities (CRH-2)*, éd. Andrew U. Frank *et al.*, Vienne, 2018, p. 55-64, https://arxiv.org/abs/1802.01429.

et les moins sujettes à la falsification comme les moins tributaires du contenu (du fond) des textes : mots-outils, ordre des parties du discours, fréquence des affixes, de la ponctuation, etc. Récemment, nous avons pu l'appliquer à la polémique récurrente portant sur l'attribution des œuvres de Molière [43], mais elle est également très riche en potentialité pour des périodes, comme la période médiévale, où l'anonymat est fréquent ou bien où les attributions sont nombreuses et contradictoires, comme c'est le cas par exemple pour la poésie lyrique des troubadours et des trouvères [44]. Au-delà du texte, ces méthodes peuvent également envisager les mélodies et tenter d'attribuer celles-ci en fonction, par exemple, des séquences de notes ou d'intervalles (tierce, quarte, quinte…) qu'on y retrouve [45].

Des tentatives d'appliquer ce type de méthodes à l'identification des faux diplomatiques existent également – telles que celle menée par Jeroen De Gussen (Anvers) sur la fausse *Donation de Charlemagne* (D Kar 286), que ses résultats attribuent à l'abbé Suger [46]. Elles posent bien sûr la difficulté de circonvenir le formulaire diplomatique pour extraire des éléments stylistiques, mais ouvrent un champ d'enquête encore peu exploré et riche de potentielles découvertes.

3. Entre outil ecdotique et instrument de connaissance de la tradition : le stemma. – Un autre domaine, présent de longue date au sein du champ de la philologie computationnelle, est celui de la stemmatologie. Selon la définition de Frédéric Duval, celle-ci est une « discipline cherchant à reconstruire la transmission des textes sur la base des relations entre les témoins survivants » [47].

De manière classique, la stemmatologie permet, à partir de l'enregistrement d'une base de données de variantes, de reconstruire de manière assistée par ordinateur une ou plusieurs hypothèses généalogiques

43. Florian Cafiero et Jean-Baptiste Camps, « Why Molière most likely did write his plays », dans *Science Advances*, t. 5, 2019, doi : 10.1126/sciadv.aax5489.

44. Pour des études récentes produites par des équipes impliquant des chercheurs de l'École, nous renvoyons à titre d'exemple, pour ce qui est de l'hagiographie française en prose, à J.-B. Camps, T. Clérice et A. Pinche, « Noisy medieval data… » ; pour les œuvres des trouvères, dans le cadre du projet MARITEM, à Jean-Baptiste Camps *et al.*, « Editing and attributing musical texts : the chansonnier du roi and the MARITEM project », dans *EADH2021: Interdisciplinary Perspectives on Data, 2nd International Conference of the European Association for Digital Humanities,* Krasnoyarsk, 2021, https://halshs.archives-ouvertes.fr/halshs-03260116/. Pour une présentation vulgarisée d'un certain nombre de dossiers stylométriques et une brève histoire de la discipline, F. Cafiero et J.-B. Camps, *Affaires de style…*

45. J.-B. Camps *et al.*, « Editing and attributing musical texts ».

46. Jeroen De Gussem, *Collaborative Authorship in Twelfth-Century Latin Literature : a Stylometric Approach to Gender, Synergy and Authority*, thèse de doctorat, université de Gand, 2019, part. chap. 7, « Forging ties : Suger and the donation of Charlemagne », p. 203-234.

47. Frédéric Duval, *Les mots de l'édition de textes*, Paris, 2015.

concernant la transmission d'un texte, c'est-à-dire l'enchaînement des témoins conservés et des étapes intermédiaires supposées depuis l'original ou l'archétype d'une tradition textuelle (la racine). Dans cette première perspective, très concrète, figure notamment le module logiciel que nous avons développé et qui implémente une méthode décrite par E. Poole se fondant sur l'étude des désaccords entre témoins [48].

Mais une autre perspective est possible. L'étude sérielle de généalogies de copies manuscrites ou d'éditions imprimées permet d'approcher les mécanismes de diffusion et de circulation des textes, comme de répondre aux débats qui ont agité la philologie durant tout le XX^e^ siècle – on songe notamment à celui qu'a ouvert Joseph Bédier et qui portait sur la prépondérance des *stemmata* bifides. Dans cette optique, nous avons commencé à constituer une collection en libre accès et collaborative de généalogies de manuscrits, OpenStemmata, en partenariat avec des collègues des universités de Heidelberg et de Genève, Gustavo Riva et Simon Gabay [49]. Cette base s'enrichit progressivement de contributions proposées par la communauté des utilisateurs. À terme, elle pourra fournir une matière utile à l'étude et à la modélisation des processus de transmission des textes et des idées, dans une perspective évolutionniste, telle que nous allons à présent l'aborder.

IV. Transmission des textes et évolution culturelle.

Pourquoi certaines productions culturelles, certaines œuvres, certains textes, certaines versions s'imposent-elles massivement au détriment de nombreuses autres qui ne connaissent aucune postérité ? Pourquoi certaines rares œuvres entrent-elles dans le mince canon littéraire et pourquoi ne conservons-nous qu'un seul manuscrit du *Roland* d'Oxford ? Quelle est la part du hasard, des mécanismes d'autorenforcement – l'argent qui va au riche, aussi théorisé sous le nom d'attachement préférentiel – ou du résultat de caractères internes, propres aux œuvres ?

En dépit de son apparent caractère très spécialisé, la stemmatologie permet en réalité d'appréhender des mécanismes fondamentaux de la

48. Jean-Baptiste Camps et Florian Cafiero, « Stemmatology : an R package for the computer-assisted analysis of textual traditions », dans *Proceedings of the second workshop on corpus-based research in the humanities (CRH-2)*, éd. Andrew U. Frank *et al.*, Vienne, 2018, p. 65-74, https://hal.archives-ouvertes.fr/hal-01695903/. La méthode elle-même est décrite dans Jean-Baptiste Camps et Florian Cafiero, « Genealogical variant locations and simplified stemma : a test case », dans *Analysis of Ancient and Medieval Texts and Manuscripts : Digital Approaches*, éd. Tara Andrews et Caroline Macé, Turnhout, 2014 (Lectio, 1), p. 69-93, doi : 10.1484/M.LECTIO-EB.5.102565.

49. Jean-Baptiste Camps, Simon Gabay et Gustavo Fernández Riva, « Open Stemmata : a digital collection of textual genealogies », dans *EADH2021: Interdisciplinary Perspectives on Data, 2nd International Conference of the European Association for Digital Humanities*, Krasnoyarsk, 2021, https://halshs.archives-ouvertes.fr/halshs-03260086.

transmission des idées et des productions culturelles. Elle constitue un cas particulier de l'étude évolutionniste de ceux-ci, des tessons de poterie au développement des théories scientifiques ou à la diffusion des « memes » sur la Toile [50]. Les grilles de lecture que l'on peut y appliquer rejoignent la formulation darwinienne de l'évolution, comme un processus tendu entre variation (apparition de variantes ou « mutations ») et fixation (conservation d'une seule d'entre elles, disparition des autres), reposant sur deux forces fondamentales que sont le hasard (la dérive génétique) et la sélection (due à un avantage dans un contexte particulier).

Dans le domaine culturel, la philologie et la linguistique partagent avec la biologie des racines conceptuelles qui remontent aux XVIII^e^ et XIX^e^ siècles, et à l'élaboration concomitante de perspectives généalogiques et de représentations arborées, par exemple chez Schlyter en philologie, Darwin en biologie ou Schleicher en linguistique [51]. Toutefois, l'émergence contemporaine d'un champ consacré plus généralement à l'étude de l'évolution culturelle remonte aux années 1980 et 1990 et trouve dans l'archéologie une partie de ses premières applications. Dans son étude de 1995 sur la variation décorative (*i.e.*, non fonctionnelle) dans les poteries de la période sylvicole moyenne et tardive en Amérique du Nord (entre 200 av. J.-C. et 1000 après), Fraser Neiman conclut ainsi que la diversité de motifs que l'on observe dans la décoration d'ensembles de poteries durant la période varie selon deux forces qui sont, pour l'une, le hasard, et pour l'autre, le rythme avec lequel les innovations sont transmises d'un groupe à un autre [52]. Ce faisant, il introduit la notion de dérive (*drift*) et montre qu'une des forces principales derrière le triomphe de certaines variantes stylistiques sur les autres est dû au hasard de tirages au sort successifs.

Cette approche peut être vue comme une application au champ culturel de la théorie neutraliste de l'évolution qui postule que l'essentiel de la variété génétique que l'on observe résulte du hasard, dans un contexte où la plupart des mutations sont neutres et où seul un tout petit nombre d'entre elles est doté d'un avantage en termes de sélection naturelle. Cette théorie permet de fournir un scénario de référence, dont on cherchera

---

50. Fraser D. Neiman, « Stylistic variation in evolutionary perspective : inferences from decorative diversity and interassemblage distance in Illinois woodland ceramic assemblages », dans *American Antiquity*, t. 60, 1995, p. 7-36, doi : 10.2307/282074 ; David Chavalarias et Jean-Philippe Cointet, « Phylomemetic patterns in science evolution. The rise and fall of scientific fields », dans *PLos ONE*, t. 8, 2013, doi : 10.1371/journal.pone.0054847. Sur la philologie et la stemmatologie comme disciplines évolutionnistes, voir en dernier lieu *Handbook of Stemmatology*, éd. Philipp Roelli, Berlin/Boston, 2020, doi : 10.1515/9783110684384.

51. D. C. J. Schlyter et D. H. S. Collins, *Corpus iuris Sueo-Gothorum Aaniqui...*, Stockholm, 1827 ; Charles Darwin, *On the Origin of Species by Means of Natural Selection*, Londres, 1859 ; August Schleicher, *Die Darwinsche Theorie und Sprachwissenschaft : offenes Sendschreiben an Herrn Dr. Ernst Häckel*, Stockholm, 1873.

52. F. D. Neiman, « Stylistic variation in evolutionary perspective… ».

à voir à quel point les données que l'on observe s'en écartent ou non. À partir de là, tout un champ s'ouvre qui est celui de la mesure, dans la transmission des productions culturelles, des éventuels écarts au hasard que l'on peut tenter d'observer en comparant les résultats de modèles mathématiques et de simulations aléatoires d'une part et les données observées de l'autre.

Dans cette perspective, nous avons entrepris, avec Julien Randon-Furling, de modéliser les processus de transmission des textes par copie manuscrite. Certaines particularités étranges observées de longue date par les philologues – telles que la fameuse « bifidité » relevée par J. Bédier –, pourraient ainsi s'expliquer par l'effet mécanique du hasard. Si ce n'était pas le cas, il s'avérerait dès lors intéressant de mesurer et de tenter d'expliquer les écarts à ce scénario de référence. Pour ce faire, nous avons décidé d'avoir recours à l'arsenal méthodologique des systèmes complexes et de la physique statistique, et avons entrepris d'élaborer un modèle multi-agents avec des paramètres simples tels que le taux de copie et le taux de destruction, de réaliser des milliers de simulations, pour pouvoir ensuite en comparer les résultats avec les données observables dans les traditions réelles étudiées par les philologues [53]. Si ce long programme de travail tout juste entamé s'avère fructueux, il pourrait jeter de nouveaux éclairages sur les processus fondamentaux de transmission des textes.

Si cette première approche demeure externe à la substance des textes en eux-mêmes, une perspective complémentaire est celle qui se préoccupe de leur contenu et de l'avantage comparatif que des propriétés stylistiques ou thématiques peuvent fournir à certaines productions culturelles vis-à-vis des autres. Les travaux très récents de Nicolas Baumard et de ses collègues montrent ainsi que la place accordée aux motifs narratifs liés au sentiment amoureux va de pair avec des facteurs économiques tels que le niveau de confort matériel, suggérant que la place accordée à l'amour, plutôt qu'un simple trait hérité de certaines traditions littéraires (comme celle de l'amour courtois), pouvait être un trait donnant un avantage comparatif à certaines œuvres sur les autres [54]. Des études nombreuses sur le canon littéraire ont cherché dans les textes des propriétés stylistiques qui différencieraient dans leur substance même les textes qui ont intégré ce cercle très restreint de tous ceux qui en ont été exclus, cela toutefois avec un succès qu'il convient sans doute de relativiser [55].

---

53. Jean-Baptiste Camps et Julien Randon-Furling, « A dynamic model of manuscript transmission », dans *Workshop on Computational Methods in the Humanities (COMHUM 2018)*, Lausanne, 2018.

54. Nicolas Baumard *et al.,* « The cultural evolution of love in literary history », dans *Nature Human Behavior*, 2022, doi : 10.1038/s41562-022-01292-z.

55. Un des travaux notables dans ce domaine est celui produit par l'équipe de Stanford : Mark Algee-Hewitt *et al.*, *Canon/Archive : Large-scale Dynamics in the Literary Field*, Stanford, 2016 (Pamphlets of the Stanford Literary Lab, pamphlet 11),

Dans cette perspective, nous avons entrepris, avec Nicolas Baumard, Pierre-Carl Langlais et Olivier Morin, d'étudier les fortunes d'un genre littéraire, le roman de chevalerie, sur la très longue durée, depuis ses origines dans l'épopée médiévale et les romans courtois jusqu'à ses possibles métamorphoses dans la *fantasy* et la littérature populaire contemporaine [56]. Pour ce faire, un vaste corpus est en cours de constitution qui, pour la période de 1470 à 1700, s'appuie sur le système de cotation mis en place par Nicolas Clément à la Bibliothèque du roi en 1684-1688, et surtout sur la révision de 1730 qui a intégré la cote Y2 regroupant les romans. Grâce aux données disponibles sur Gallica et sur Google books, aux corpus préexistants [57] et aux résultats de projets tels que Gallic(orpor)a [58], il devient possible de constituer un corpus représentant une part très substantielle de la production, du moins telle qu'elle est reflétée par la cote Y2, soit entre 70 et 80 % des titres.

Outre la dérive (due au hasard) et la sélection (due à une meilleure adaptation à un contexte), d'autres mécanismes peuvent être à l'œuvre dans le succès et la pérennité de certaines productions culturelles, et particulièrement des mécanismes d'autorenforcement évoqués plus haut. Dans le domaine par exemple de l'étude de l'évolution du lexique, certains résultats suggèrent ainsi que, plus un mot est utilisé, plus il tire de ce capital une stabilité qui assure son utilisation future, ce qui expliquerait l'étonnante stabilité dans le temps, depuis l'indo-européen, de certaines racines d'emploi très fréquent, telle que celle de *deux*, *zwei*, *two*, ou *do* en hindi [59].

V. En conclusion : massification et renouvellement des perspectives.

La production de corpus massifs en diachronie ou diatopie large, de même que les avancées de la recherche dans le domaine de la modélisation mathématique, des systèmes complexes et de l'évolution culturelle,

---

http://litlab.stanford.edu/LiteraryLabPamphlet11.pdf (consulté le 21 janvier 2022). L'un des étudiants du master Humanités numériques de l'École, Jean Barré, a entrepris de répliquer ces analyses sur un corpus de 3 000 romans français des XIXe et XXe siècles pour son mémoire (dir. T. Poibeau et J.-B. Camps), dont les résultats permettront peut-être de jeter plus de jour sur cette problématique.

56. Pierre-Carl Langlais *et al.*, « From Roland to Conan. First results on the corpus of French literary fictions (1050-1920) », *DH 2022 Tokyo*, à paraître.

57. On songe tout particulièrement au corpus des fictions littéraires de Gallica, qui contient déjà 19 240 documents entre 1600 et 1996 ; Pierre-Carl Langlais, « Fictions littéraires de Gallica / Literary fictions of Gallica », avril 2021, https://zenodo.org/record/4751204.

58. Sur ce projet, voir ci-dessus, note 6.

59. Mark Pagel, Quentin D. Atkinson et Andrew Meade, « Frequency of word-use predicts rates of lexical evolution throughout Indo-European history », dans *Nature*, t. 449, 2007, p. 717-720, doi : 10.1038/nature06176.

permettent ainsi d'envisager à nouveaux frais des questions historiques et philologiques qui ont de longue date occupé la critique.

Il y a encore une dizaine d'années, Franco Moretti, l'inventeur du concept de *distant reading*, voulant étudier l'évolution des romans anglais de 1740 à 1850, justifiait le choix de se limiter à l'analyse des titres en notant ceci :

> D'ici quelques années, nous disposerons d'archives numériques contenant le texte intégral de (presque) tous les romans jamais publiés ; mais pour l'instant, les titres restent le meilleur moyen d'aller au-delà du 1 % de romans qui constituent le canon et d'avoir un aperçu du champ littéraire dans son ensemble [60].

Nous y sommes désormais (ou presque). Mais qu'allons-nous faire de toutes ces données ? Allons-nous continuer à tenter de leur poser les mêmes questions – jadis restreintes par les capacités techniques limitées qui étaient les nôtres – ou bien est-ce qu'un changement complet de paradigme de recherche va se réaliser, tirant profit de ces données et d'un arsenal méthodologique sans cesse croissant pour revisiter le cadre théorique même de nos disciplines et la nature des questions auxquelles nous cherchons à répondre ? La transformation des concepts et la maturation de nos interprétations, de nos cadres théoriques, sont plus lentes que l'évolution de la technique et des outils à notre disposition. C'est pourtant ce vers quoi nous devons diriger nos efforts, pour dépasser le stade de l'expérimentation et de la prouesse techniques, au profit d'une compréhension renouvelée des faits historiques et des dynamiques des sociétés humaines.

Jean-Baptiste Camps.

---

60. « In a few years, we will have a digital archive with the full texts of (almost) all novels ever published ; but for now, titles are still the best way to go beyond the 1 percent of novels that make up the canon, and catch a glimpse of the literary field as a whole » (je traduis) ; Franco Moretti, « Style, Inc. reflections on seven thousand titles (British novels, 1740-1850) », dans *Critical Inquiry*, t. 36, 2009, p. 134-158.