Le numéro 12 de la *Revue Roumaine d'Études Francophones* invite à une réflexion autour de l'hybridité et des métamorphoses qui lui sont associées, concepts qui ont refait surface ces derniers temps. Regroupés selon les deux axes traditionnels de la revue – littérature et linguistique –, les articles réunis dans ce recueil interrogent ces concepts sous différents aspects, laissant voir continuités, discontinuités, permanences, ruptures, renouveau dans leurs approches.

Si le couple conceptuel hybridité-métamorphoses traverse époques, disciplines, thématiques, approches épistémologiques, les articles recueillis rendent compte de la manière dont le questionnement autour de cette problématique complexe permet d'établir des corrélations, de proposer des constantes, de fournir des instruments d'investigation pour mieux appréhender phénomènes littéraires, manifestations culturelles et pratiques langagières.

**Cristina PETRAŞ**

ISSN 2065-8087

HYBRIDITÉ ET MÉTAMORPHOSES

Revue Roumaine d'Études Francophones No. 12/2020

Revue Roumaine d'Études Francophones

No.12/2020

Publication annuelle de l'Association Roumaine des Départements Universitaires Francophones (ARDUF)

# HYBRIDITÉ ET MÉTAMORPHOSES

JUNIMEA

# Un couple indivisible : travail digital et travail manuel dans les études littéraires

Ioana GALLERON[1]

## 1. Introduction

Bien connu et documenté, le « tournant digital »[2] des études de sciences humaines et sociales, et en particulier la digitalisation de la recherche en lettres – qui avait peu intégré, avant 1970, et même 1990[3], des méthodes statistiques dans la panoplie d'approches possibles des textes et des systèmes culturels[4] – est une réalité à

[1] Université Sorbonne-Nouvelle, UMR 8094 LATTICE, France.

[2] Quoiqu'ils ne soient pas, à strictement parler, à l'origine de l'expression, les auteurs du manifeste pour les humanités numériques en proposent une analyse ludique et complète : Jeffrey Schnapp, Todd Presner, Johanna Drucker and Peter Lunenfeld, « Digital Humanities Manifesto 2.0. », https://humanitiesblast.com/publications/ (consulté le 20 mai 2021).

[3] Immédiatement après la fin de la seconde guerre mondiale, l'utilisation de l'ordinateur pour des études de sciences humaines et sociales est envisagée par Roberto Busa, considéré comme un des « pères fondateurs » des humanités numériques. C'est avec la production, dans les années 1970, de son Index Thomisticus sur CDRom que celles-ci deviennent visibles. Il faut attendre les années 1990, ainsi que la démocratisation des ordinateurs personnels, pour qu'elles prennent leur essor. En Roumanie, les travaux de Solomon Marcus dans le domaine de la linguistique, mais aussi pour la modélisation du théâtre, s'inscrivent dans cette tendance, même s'ils sont restés sans beaucoup d'écho parmi les spécialistes des lettres et des langues.

[4] Les études littéraires se distinguent ainsi de l'histoire, qui n'a pas attendu la révolution numérique pour apporter des arguments fondés sur des (dé)comptes, ainsi que de la linguistique, qui, en mettant à distance l'intuition du chercheur comme point de départ pour la description de la langue, s'est engagée très tôt dans la construction de très larges corpus (plusieurs millions de mots), se donnant ainsi la possibilité d'étudier non seulement des

présent incontournable. À de multiples reprises, il a été démontré que le numérique apporte des réponses inédites et permet de résoudre des problèmes des plus complexes, comme la possibilité de représenter de façon enfin conviviale et plus lisible les multiples variantes d'un ouvrage dans le cadre d'une édition critique[5], l'attribution auctoriale de passages contestés et d'œuvres anonymes[6], l'analyse du style[7], l'observation de différentes évolutions sur le temps long, la réévaluation du rapport entre canon et production littéraire[8], et bien d'autres. Si les travaux qui n'intègrent pas de dimension numérique ne sont nullement menacés dans le contexte actuel, puisque l'effet esthétique ne peut avoir lieu ailleurs que dans une conscience humaine, il semble peu envisageable que le spécialiste des lettres de

---

de mots), se donnant ainsi la possibilité d'étudier non seulement des phénomènes attestés, mais également statistiquement significatifs. Soulignons toutefois que des travaux similaires ne sont pas entièrement absents dans les études de lettres : les tables d'Alexandre Joannidès permettent, dès 1901, d'avoir une perspective chiffrée sur le phénomène du théâtre classique en France (v. *La Comédie-Française de 1680 à 1900*, Paris, Plon-Nourrit et cie, 1901). Cependant, force est de constater qu'il s'agit d'approches ayant joui de peu d'échos et de suites, à la différence de ce qui s'est passé dans les autres disciplines citées dans cette note.

[5] Voir en ce sens Matthew James Driscoll et Elena Pierazzo (éd.), *Digital Scholarly Editing: Theories and Practices*, Cambridge, Open Book Publishers, 2016, http://dx.doi.org/10.11647/OBP.0095 (consulté le 20 mai 2021).

[6] Pour une synthèse, lire Patrick Juola, « Authorship Attribution and the Digital Humanities Curriculum », in *Literary Education and Digital Learning: Methods and Technologies for Humanities Studies*, W. van Peer, S. Zynglier, V. Viana (éd.), s. l., IGI Global, 2010, p. 1-21.

[7] Multiples références possibles, lire par exemple David L. Hoover, « Quantitative Analysis and Literary Studies », in *A Companion to Digital Literary Studies*, Susan Schreibman and Ray Siemens (éd.), Oxford, Blackwell, 2008, http://www.digitalhumanities.org/companionDLS/.

[8] Pour les deux, renvoyons à Franco Moretti, *Graphs, Maps, Trees : Abstract Models for a Literary History*, London, New York, Verso, 2005, ainsi que son atlas du roman : *The Novel : History, geography and culture*. 1 (trad. de l'italien), Princeton, N.J., Princeton University Press, 2006.

demain ne dispose pas de compétences digitales, et le rapide développement des masters en « Humanités numériques »[9] (HN), ainsi que, plus largement, l'intégration de plus en plus ample de cours dits « d'informatique »[10] dans les études de lettres, langues et cultures européennes et non européennes est un bon indicateur quant à ce changement de perspective.

Ce n'est toutefois pas des formes et de l'intérêt de cette rapide pénétration du numérique dans les études littéraires qu'il sera question dans cet article, mais plutôt de la mise en avant d'une réalité moins connue, qui est celle de la dépendance du numérique de travaux manuels, impliquant des connaissances poussées de lettres et de langues. Non seulement les résultats des différents systèmes automatiques ou semi-automatiques impliqués dans les études rapidement énumérées plus haut ont-ils besoin de l'intelligence humaine pour devenir compréhensibles et stimulants, mais, en outre, leur fonctionnement même n'est possible sans un investissement considérable de temps humain, ainsi que de théories et de connaissances, explicites ou tacites, que l'on n'acquiert autrement que par la fréquentation des formations littéraires. Dès lors, il ne s'agit pas de donner, une fois de plus, des garanties quant au fait que le « distant reading » ne tue pas la lecture tout court, ou la lecture dite « de près »[11]. Avec une perspective un peu différente, l'objectif est de souligner à quel point, au moins dans l'état actuel des recherches, le numérique a besoin du manuel pour produire des résultats. Une telle affirmation a, sans doute, de quoi décourager plus d'un néophyte intéressé par la révolution numérique, et donnera de l'eau au moulin de ceux qui voient dans cette tendance un feu de paille ou une errance,

---

[9] Une liste à jour, mais sans doute pas complète, se trouve à l'adresse https://dhcr.clarin-dariah.eu/.

[10] Ils ne le sont pas, bien entendu, car les étudiants n'apprennent ni à coder, ni même à comprendre les principes de base du fonctionnement de l'ordinateur dans le cadre de ces enseignements. Selon les universités, les *curricula* sont plus ou moins ambitieux, s'articulant *a minima* autour de la présentation des « espaces numériques de travail » et d'exercices de bureautique.

[11] Pour une confrontation des deux perspectives, voir Barbara Herrstein Smith, « What Was 'Close Reading' ? A Century of Method in Literary Studies », *Minnesota Review*, Duke University Press, issue 87, 2016, p. 57-75.

éventuellement préjudiciable aux sciences humaines. Il n'en reste pas moins que cette dépendance est un fait, et que le refus de ses conséquences par les spécialistes des lettres leur est nocif, les privant de la possibilité d'intégrer, de façon étendue, critique et diversifiée, leurs cadres épistémiques dans la préparation des outils de la recherche de demain. L'imaginaire de la transformation des méthodes comme un processus venant de l'extérieur, que l'on embrasse de façon enthousiaste ou auquel on résiste, est celui d'une « colonisation », et non pas celui d'une hybridation disciplinaire, porteuse, certes, de nombre d'inconvénients mais, surtout, de promesses.

Dans une première partie, après un rapide survol des ressources actuellement existant en ligne, je montrerai à quel point leur utilisation de façon intégrée, et à la véritable hauteur de leurs possibilités, demande la création d'instances de négociation formées de spécialistes purement littéraires qui s'engagent dans la construction d'outils et du « web de données ». Dans une seconde partie, sur la base de plusieurs projets de recherche actuellement en cours, il sera question de différents objets sur lesquels ce qu'on appelle à présent « intelligence artificielle » présente un criant besoin d'intervention humaine, fondée sur la connaissance fine des théories et méthodes littéraires. Je finirai en esquissant un programme d'engagement dans le numérique, ou avec le numérique, à partir de ces différentes observations.

## 2. Dans la jungle des ressources numériques

En schématisant de façon rapide et sans doute discutable, sept types de ressources numériques sont actuellement disponibles aux spécialistes des sciences humaines et sociales.

1° Des bases de données (BDD) signalant ou pointant vers des ressources, plus généralistes (comme ceux des bibliothèques nationales ou d'Europeana), ou plus spécialisées[12], et destinées à

[12] Pour les textes du Moyen Âge (et plus largement de l'Antiquité au XVIII^e^ siècle), on peut consulter le portail Biblissima : https://portail.biblissima.fr/. La philosophie clandestine peut être tracée à partir des relevés effectués sur la plateforme http://philosophie-clandestine.huma-num.fr/index.html.

permettre l'identification de textes pertinents pour l'étude d'une question ou d'un groupe de questions données. Réalisées par de grandes institutions, ou dans le cadre de projets plus modestes, elles fournissent – et depuis plus de 30 ans déjà – une contrepartie numérique au traditionnel travail bibliographique mené à l'aide de fiches et de compilations.

2° Des bibliothèques numériques, que l'on peut de nouveau classer en bibliothèques à spectre large (Wikisource, Project Gutenberg, Gallica, Hathi Trust book collections), ou plus spécialisées (Women Writers Online, Montaigne à l'œuvre). Un critère supplémentaire de différenciation est le type d'offre que l'on trouve dans ces bibliothèques : des numérisations en mode image ou bien des textes « machine readable »[13]. Il est rare que l'on n'y trouve que l'un ou l'autre type, même si les pdf ou jpg (plus récemment, des images IIIF) restent indiscutablement majoritaires[14]. Le mode d'utilisation principal envisagé est, comme dans le cas des BDD mentionnées plus haut, analogue à celui d'une « consommation » *in situ*, à savoir la lecture, intégrale ou par extraits, effectuée par un être humain.

3° Des outils de traitement de texte brut, offrant des services plus ou moins sophistiqués : tokénisation, lemmatisation et étiquetage morpho-syntaxique[15], analyse stylistique et thématique[16],

[13] Guy Meunier propose d'utiliser l'appellation de « texte numérique » pour les premiers, obtenus par transduction et encodage, et de « texte dynamique » pour les seconds (J. G. Meunier, « Le Texte numérique : enjeux herméneutiques », *Digital Humanities Quaterly*, vol. 12, no. 1, 2018, http://www.digitalhumanities.org/dhq/vol/12/1/000362/000362.html, consulté le 20 mai 2021). Toutefois, la seconde appellation peine à s'implanter en français.

[14] Plus récemment, Gallica offre un service d'océrisation « à la volée » pour une partie des textes qu'elle met à disposition. Il n'est pas toutefois pas très clair quels types de textes sont concernés, et il va de soi que les résultats comportent des taux d'erreurs plus ou moins importants.

[15] Parmi de multiples systèmes disponibles en ligne, une mention spéciale mérite UDPipe, à la fois pour la qualité des résultats et pour la multitude de langues couvertes : http://lindat.mff.cuni.cz/services/udpipe/ (consulté le 20 mai 2021).

*topic modelling*[17], et quelques autres. Bon nombre de ces ressources offrent des interfaces en ligne, tout en existant dans des versions que l'on peut installer sur son ordinateur.

4° Des outils d'analyse de textes structurés et annotés, en format linéaire (XML-TEI notamment), ou tabulaire (BIO, CoNLL ou similaire). Certains de ces outils présentent une interface graphique et des applications en ligne (TXM, SketchEngine, Geobrowser, OpenRefine), d'autres fonctionnent uniquement sous la forme d'un logiciel installé sur la machine du chercheur (eXist, BaseX), d'autres, enfin, demandent de savoir écrire quelques lignes de code (R et Python, avec leurs « bibliothèques ») ou à tout le moins de pouvoir utiliser un « notebook » (comme Jupyter).

5° Des outils d'annotation, capables de prendre en charge des formats des plus différents (comme les logiciels de la famille CAQDAS[18]), ou bien spécialisés dans l'annotation de texte (BRAT, WebAnno), éventuellement à des fins ultérieures d'entraînement d'un système d'intelligence artificielle.

6° Des ressources permettant la création, le stockage, la publication et parfois l'archivage à long terme des données et des travaux de recherche. On peut citer dans cette catégorie les entrepôts comme Zenodo, des plateformes collaboratives comme Github, des services comme Nakala, etc. Mentionnées pour mémoire, il n'en sera pas question dans cet article.

7° Des ressources pour l'enseignement, comme des plateformes collaboratives (Moodle étant une des plus connues), souvent couplées avec des modules permettant des évaluations formatives ou sommatives. Ces ressources ne seront pas non plus abordées dans le développement qui suit.

---

[16] Par exemple, avec le populaire Voyant Tools (https://voyant-tools.org/, consulté le 20 mai 2021).

[17] Pour une présentation de cette technique et des outils qu'elle sollicite, voir Lisa M. Rhody, « Topic Modeling and Figurative Language », *Journal of Digital Humanities* 2, no. 1, 2012. (http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/, consulté le 20 mai 2021).

[18] *Computer-assisted qualitative data analysis software*, soit des logiciels comme Atlas.ti ou NVivo.

Notons enfin que des superpositions existent entre les différentes ressources, de nombreux catalogues donnant un accès direct à des bibliothèques, et certaines de ces dernières proposant à leur tour des formulaires de recherche qui s'appuient, en arrière-plan et souvent à l'insu du chercheur, sur des outils des catégories 3 et 4.

Même esquissé ainsi à grands traits, l'écosystème des ressources numériques pour la recherche en Sciences Humaines et Sociales (SHS) est impressionnant, et en constante augmentation. En comparaison, les études d'envergure produites à partir de ces ressources restent moins nombreuses, surtout si l'on regarde les cultures non-anglo-saxonnes. Une large place dans les colloques et conférences du domaine des humanités numériques reste réservée à la présentation de collections plus ou moins récemment numérisées, ou à la démonstration des fonctionnalités de différents outils, au détriment des résultats obtenus grâce à de telles collections et outils et à tel point que l'on parle parfois d'un déséquilibre des HN en faveur du numérique. Des justifications existent pour un tel état des faits, la communauté HN étant par définition un espace de réflexion sur les techniques et les méthodes, alors que les résultats sont à discuter dans les communautés disciplinaires d'origine des chercheurs. Mais, là aussi, le développement des travaux sur les objets (textes, collections, corpus) laisse à désirer : sans minimiser l'intérêt et l'importance de toute une série d'études, remarquons par exemple que sur 151.607 enregistrements depuis 2003 dans la BLF (*Bibliographie de la littérature française* : équivalent numérique de la bibliographie de la *RHLF*[19]), seuls 176 affichent le mot-clé « numérique », soit à peine un peu plus de 1%. En y regardant de plus près, on y trouve une proportion non négligeable de présentations de projets et d'outils, au détriment de la présentation des résultats obtenus grâce à ces initiatives numériques[20].

[19] *Revue d'Histoire littéraire de la France*.

[20] Le mot-clé « numérique » a été recherché dans l'interface en ligne de la BLF, consultable à l'adresse https://www.classiques-garnier.com/numerique-bases/blf. Le plus ancien enregistrement ainsi obtenu remontant à 2003, le nombre total d'entrées dans la base de 2003 à nos jours a été recherché. À noter que la BLF offre un dépouillement systématique, mais pas exhaustif : des travaux

Les raisons de cet état de fait sont multiples, et tiennent en premier lieu à la relative méconnaissance des types de ressources mentionnées plus haut. Des efforts de formation des chercheurs, plus jeunes ou moins jeunes, aux différentes possibilités restent nécessaires. Mais une partie de la réponse est également liée à une certaine inadaptation de ces outils par rapport aux questions de la recherche en lettres. En se focalisant uniquement sur les deux premières catégories (il sera question des deux autres dans la partie suivante), observons que, dans l'état actuel des choses, non seulement n'a-t-on pas un plein accès aux textes, qui restent majoritairement en format image, comme indiqué plus haut, mais que même la découverte et, encore plus, la mise en réseau de ces derniers s'avère difficile, voire impossible. Conçus pour permettre l'accès à un objet dont on connaît déjà l'existence et la description, catalogues et bibliothèques en ligne s'avèrent bien moins efficaces pour soutenir une recherche exploratoire, alors même que c'est de cette dernière qu'il s'agit, le plus souvent, dans le *distant reading* et plus largement dans les analyses littéraires assistées par ordinateur.

Un premier exemple soutenant cette affirmation est la difficulté de dégager de grandes lignes de force à partir des archives ou éditions numériques[21]. Au cours de la dernière dizaine d'années, de nombreux manuscrits ou textes rares[22], jadis négligés car hors canon ou réservés

---

parus dans la *Revue d'Histoire du Théâtre numérique*, par exemple, n'ont pas pu y être identifiés, même quand ils étaient en lien avec des textes littéraires français. Une sous-représentation des études numériques est donc à prendre en compte dans l'interprétation du pourcentage indiqué plus haut, même s'il est difficile d'en évaluer l'impact.

[21] Pour une différence entre archive et édition numérique, ainsi que pour une description des différents types d'éditions numériques, voir Ioana Galleron, Fatiha Idmhand, Marie-Luce Demonet, Cécile Meynard, Elena Pierazzo, et al. *Les Publications numériques de corpus d'auteurs – Guide de travail, grille d'analyse et recommandations* (V1-Novembre 2018). [Rapport de recherche], halshs-01932519.

[22] Brouillons préparatoires et autres éléments du dossier génétique d'œuvres plus ou moins connues (archives de Proust, de Claude Louis-Combet, de Paul Bourget, etc.), lettres des soldats de la première guerre mondiale, lettres d'intellectuels immigrés, etc.

à l'accès d'un groupe restreint de chercheurs, sont devenus accessibles grâce à leur publication en format image sur des sites OMEKA[23]. Or, alors que la plus rapide visite dans cette galaxie numérique permet d'identifier « à l'œil nu » des échos et des similitudes, leur relevé systématique s'avère bien moins aisé lorsqu'on essaie de le faire au moyen du dispositif précisément conçu pour ce type de travail, à savoir une requête sur les mots-clés. Des thématiques similaires (l'exil, l'impuissance créatrice, le questionnement sur les relations humaines, etc.) sont rendues avec des termes différents, et selon des pratiques tellement disparates que les tentatives de regroupement et de conciliation s'avèrent pratiquement impossibles. Plus largement, les représentations quant à ce qu'est un mot-clé divergent, comme tout coordonnateur d'un volume collectif ou d'un numéro spécial de revue aura pu s'en rendre compte dans un autre cadre.

Un autre manque susceptible d'expliquer le sous-développement de la recherche digitale en lettres est la présence disons fragmentaire du genre littéraire dans les catalogues des bibliothèques, ou plus largement dans les métadonnées des collections numérisées. Des champs de recherche par « sujet », « genre » ou « forme » existent bien, mais les résultats qu'ils proposent sont à la fois incomplets et d'une pertinence parfois discutable[24]. Même les collections les plus récentes, comme celles qui ont été numérisées dans le cadre du consortium français CAHIER, donnent des résultats mitigés en la matière, en raison de l'absence d'un protocole descriptif clair des textes du point de vue de leur typologie. Plus encore, au-delà du protocole, ce qui manque ce sont des notions de référence, organisées et définies de façon brève et non-ambiguë – autrement dit, des ontologies, *thesauri* ou vocabulaires contrôlés. Briques fondamentales du « web de données », assurant la communication entre des ressources conçues avec une grande variété de techniques et intentions, ils s'avèrent peu

[23] Voir par exemple les sites accueillis sur la plateforme eMan : https://eman-archives.org/EMAN/ (consultée le 20 mai 2021).

[24] Pour une plus ample discussion sur ce point, voir I. Galleron, M. L. Demonnet, F. Idmhand, A. Lavrentiev, A. Reanch-Ngô, « Décrire un corpus d'auteurs : un thésaurus pour les types de textes », à paraître.

connus et encore moins pratiqués par les littéraires, qui se reposent plutôt, en la matière, sur le travail des bibliothécaires. Possesseurs d'une connaissance intime des textes, des courants littéraires et de leurs complexités, les spécialistes du texte se détournent ainsi d'une des synthèses exigeant les compétences les plus élevées en la matière. Il est vrai qu'un tel travail est peu valorisé académiquement, qu'il prête le flanc à de multiples critiques et qu'il exige un engagement sur la durée, en raison des nombreuses évolutions nécessaires après la première publication : à la différence d'un livre ou d'un article, un vocabulaire contrôlé ou une édition numérique demandent d'assurer un « service après-vente ». Mais le numérique a besoin d'une telle hybridation par les connaissances et compétences dites « traditionnelles », dont le développement se trouverait, en retour, conforté dans le sillage d'un tel engagement.

### 3. Travaux de bénédictin et Turcs mécaniques

Au cours des dernières années, la transformation numérique, accélérée par endroits par le contexte pandémique, a créé le sentiment que, de nos jours, « tout » est en ligne. Cependant, laissant de côté ce qui est en ligne derrière des barrières diverses (paiement, technologies propriétaires, ressources réservées à un groupe qui les garde jalousement, pour des raisons pas toujours liées aux limitations imposées par le droit d'auteur), revenons brièvement sur cette masse de textes numérisés en mode image, dont nous sommes, le plus souvent, censés nous satisfaire. Oui, on peut lire commodément, de nos jours, des textes qui auraient demandé jadis de longs et coûteux voyages, et même une négociation serrée avec des institutions qui en ont la garde[25]. Mais on ne peut ni les manipuler, ni les éditer, et encore moins les annoter de façon satisfaisante, c'est-à-dire de façon à pouvoir retrouver rapidement et de façon fiable les annotations similaires. C'est du texte intégral que l'on a besoin pour pouvoir réaliser de telles opérations. Qu'à cela ne tienne ! La reconnaissance

[25] Que l'on pense par exemple à certaines ressources existant à la bibliothèque de l'Institut, à Paris, où l'entrée est impossible sans montrer patte blanche.

optique des caractères (OCR) n'a-t-elle pas fait des progrès extraordinaires, affichant de nos jours des taux de correction qui avoisinent le 100% ? C'est oublier – outre la façon dont ce taux est calculé[26], ainsi que le fait que les systèmes les plus performants sont relativement peu accessibles, car payants[27] – que l'OCR est toujours mis en échec par les typographies non-standardisées, antérieures à la première moitié du XIX$^{e}$ siècle, que les documents les plus intéressants sont parfois manuscrits, et que la reconnaissance des attributs de caractère (italiques, petites capitales, gras, etc.) n'est pas tout à fait performante, même pour les lettrages plus récents. Ainsi, dans la réalité des faits, la conversion des images en texte reste difficile, soit parce que le résultat est entaché de nombreuses erreurs dont la correction ne peut se faire que manuellement, soit parce que son amélioration dépend de la capacité du chercheur à entraîner (on reviendra sur le terme) un modèle OCR *ad hoc*, soit, enfin, en raison d'une combinaison des deux causes.

Il est cependant incontestable qu'Internet offre au chercheur intéressé un grand nombre de textes « propres », dans des formats bruts ou plus ou moins structurés, que l'on peut manipuler au moyen d'un des outils décrits dans les catégories 3, 4 ou 5. L'établissement d'une concordance pour toutes les formes lexicales utilisées dans les *Rougon-Macquart* ou dans *la Comédie humaine* devient dès lors un jeu d'enfant, ainsi que l'identification des mots-clés de l'un et de l'autre ensemble. Au relevé manuel, fastidieux et sujet à l'erreur, des occurrences, le numérique apporte ainsi des réponses rapides et complètes. Toutefois, dans la plupart des études littéraires, ce type d'aide, quoique appréciée, n'est pas exactement ce dont on a le plus besoin.

C'est, en effet, moins aux mots et même aux constructions que les spécialistes de la discipline s'intéressent, qu'à des objets qui, à ce jour, restent difficiles voire impossible à saisir avec des outils

[26] Pour mémoire, le calcul s'effectue au niveau du caractère, espaces compris. Un taux de réussite de 98% signifie qu'il y aura deux erreurs par centaine de caractères, soit environ six mots dans une note comme celle-ci. Suffisamment pour que la lecture soit freinée et que la correction devienne chronophage.

[27] Le logiciel de référence reste ABBY Finereader, https://pdf.abbyy.com/fr/.

numériques : simili-objets comme les personnages, ancrages temporels et spatiaux, structures de profondeur (« mouvements » de l'argumentation ou parties de l'intrigue, par exemple), ou de surface (organisation du texte en parties et chapitres), etc. Cependant, l'aide numérique pour le travail sur ce type d'objets tarde à venir, comme on peut le constater, de nouveau, à partir de toute une série d'exemples. En laissant de côté les cas les plus difficiles, comme l'identification des métaphores[28], on peut dire que la raison se trouve, d'une part et de nouveau, dans le sous-investissement du numérique par les spécialistes de la littérature, mais aussi, d'autre part, dans une conception erronée quant à ce que peut le numérique, comme on le verra à partir des exemples suivants.

L'idée que les personnages d'une œuvre littéraire ne sont pas à étudier un par un, mais en tant que système organisé par des lignes de force et des tensions, n'est pas nouvelle ; sa mise en œuvre en contexte analogique est, en revanche, longue et complexe. Avec quelques exceptions, c'est toujours à partir de figures proéminentes ou grâce à des textes (surtout théâtraux) qui présentent un nombre limité de tels objets qu'elle a été mise en œuvre. À ces limitations, le numérique semble pouvoir apporter une réponse plus efficace, à la fois pour le relevé des instances (combien de personnages dans une œuvre littéraire), pour l'identification des occurrences (où et quand apparaissent-ils dans le texte) et pour la manipulation des extractions ainsi réalisées (regroupements, recoupements, calculs statistiques et visualisations). Les présentations de différents systèmes de reconnaissance d'entités nommées (NER) et, plus particulièrement, de chaînes de traitement de textes littéraires comme celle développée à Stanford[29], donnent l'espoir d'un accès rapide à de telles données.

---

[28] Voir en ce sens Suzanne Mpouli, « Chronique d'un échec : identification des métaphores dans les écrits des géographes », *Revue TAL*, ATALA (Association pour le Traitement Automatique des Langues), numéro spécial « TAL et humanités numériques », 60 (3), 2019.

[29] Lire par exemple David Bamman, Sejal Popat and Sheng Shen, « An Annotated Dataset of Literary Entities », NAACL, 2019, ou plus récemment, David Bamman, « LitBank : Born-Literary Natural Language Processing », *Debates in Digital Humanities*, 2020.

En réalité, des tests réalisés avec différents outils et sur des langues autres que l'anglais montrent que les résultats laissent grandement à désirer[30]. Entraînés sur des corpus de presse, les systèmes de NER sont mis en difficulté quand il s'agit de reconnaître des entités nommées dans les textes littéraires. Par ailleurs, les entités reconnues ne sont pas toutes, et de loin, des personnages : même en laissant de côté les noms de lieux et d'institutions que l'on retrouve inévitablement dans la liste, faut-il considérer que Molière, sous le buste duquel on discute dans un passage de *Nouma Roumestan*, est un personnage ? Enfin, même une fois le partage fait entre entités « personnage » et « non-personnage », reste le problème de la désambiguïsation ou, formulé autrement, de l'identification des entités similaires désignées différemment : si le lecteur de *La Chartreuse de Parme* sait que Gina del Dongo, la comtesse Pietranera, la duchesse Sanseverina, la comtesse Mosca et la tante de Fabrice sont une seule et même personne, tel n'est pas le cas des ordinateurs, pour lesquels la simple variation d'une majuscule ou d'une minuscule suffit à créer des entités différentes. Une amélioration des résultats peut venir de l'entraînement des systèmes, mais ceci exige l'annotation manuelle, préalable, d'un corpus suffisamment étendu[31] pour que l'ordinateur puisse en apprendre, grâce à l'analyse des contextes et par des méthodes d'essai et d'erreur, à quoi ressemble un personnage. En d'autres mots, les systèmes automatiques ou semi-automatiques ont besoin, pour

[30] Une comparaison des performances de différents systèmes a été menée dans le cadre du projet européen « Distant reading » (CA 16204). Pour la présentation de la méthodologie et quelques résultats, voir Francesca Frontini, Carmen Brando, Joanna Byszuk, Ioana Galleron, Diana Santos and Ranka Stanković, « Named Entity Recognition for Distant Reading in ELTeC », *Proceedings of CLARIN Annual Conference 2020*, C. Navarretta and M. Eskevich (eds.), Virtual Edition, 2020. p. 37-41.

[31] L'intelligence artificielle à réseaux de neurones est particulièrement gourmande, exigeant plusieurs gigas de données pour s'entraîner. Les systèmes plus traditionnels se contentent de volumes plus modestes (échantillons de 1000 à 2000 mots), mais, pour être robuste, l'annotation doit être négociée, menée en parallèle, puis confrontée et arbitrée, ce qui implique plusieurs mois de travail d'équipe.

fonctionner, de littéraires qui fassent au préalable les opérations qu'on leur demande d'accomplir. Considérés sous cet aspect, les programmes actuellement les plus avancés de traitement de textes littéraires ressemblent pas mal à cet automate de la fin du XVIII[e] siècle qui gagnait aux échecs grâce à un joueur caché dans son sein, ayant donné son nom à un célèbre service d'Amazon.

En revenant aux entités nommées et à l'entraînement des machines, observons – outre le peu d'investissement des littéraires dans de tels travaux, très similaire à leur désintérêt pour les ontologies signalé plus haut – que les quelques campagnes de ce genre, qui ont eu lieu ou qui se déroulent actuellement, mettent en lumière plusieurs limites de l'exercice. D'une part, il n'existe pas de véritable norme, dans le champ des études de lettres, quant à ce qu'est un personnage[32] ; d'autre part, même quand une définition de travail est adoptée de façon pragmatique dans le cadre d'un projet, l'annotation reste hautement subjective et, plus d'une fois, imprécise. C'est du moins ce qui ressort d'un projet comme DEMOCRAT, ayant mené un travail minutieux sur les chaînes de co-référence[33], probablement le plus avancé à ce jour, mais dont les outils numériques continuent à être affectés par les quatre chevaliers de l'Apocalypse de tout langage humain, à savoir la polysémie, l'ambiguïté, l'ironie et l'implicite. À moyen terme, le travail numérique sur les personnages va sans doute progresser, mais uniquement au prix d'une double hybridation – celle des lettres par le

[32] À titre d'exemple, un test effectué avec une centaine d'étudiants sur « Continuité des parcs » (un récit en trois paragraphes, de José Luis Borgès) montre que les lecteurs y reconnaissent entre 3 et 8 personnages. Pour une présentation complète de la méthodologie et des conclusions, voir Galleron, Ioana ; Idmhand, Fatiha ; Meynard, Cécile, « Que mille lectures s'épanouissent… Modélisation du personnage et expérience de 'crowdreading' », *Digital Humanities Quaterly*, volume 12, no. 1, 2018, (http://www.digitalhumanities.org/dhq/vol/12/1/000363/000363.html, consulté le 20 mai 2021).

[33] Frédéric Landragin, *Rapport final du projet ANR Democrat,* « *Description et modélisation des chaînes de référence : outils pour l'annotation de corpus et le traitement automatique* ». [Rapport de recherche], Agence Nationale de la Recherche – France, 2020, hal-02533314.

numérique, qui lui imposera ses exigences de rigueur et d'exhaustivité, mais aussi celles du numérique par les lettres, qui lui enseignera à travailler en régime d'incertitude et à se contenter, par endroits, d'approximations. Dans tous les cas, des travaux manuels d'affinement et de ré-annotation resteront nécessaires, avant de pouvoir utiliser de tels résultats.

C'est à des conclusions similaires qu'est parvenue l'équipe du projet BASNUM, consacré à la numérisation et à l'analyse du *Dictionnaire universel* d'Antoine Furetière, dans la version revue, corrigée et augmentée par Basnage de Beauval, publiée en 1701. Prenant le relai d'un travail pilote de digitalisation de la lettre C par saisie et encodage manuel, le module GROBID-Dictionaries[34] a permis des avancées rapides et significatives : à l'heure actuelle, l'intégralité du dictionnaire (plus de 3000 pages) existe dans une version finement structurée, qui permet de faire des recherches à la fois sur la nomenclature, les informations grammaticales, les termes, les étymologies et, bien entendu, les sens[35]. À y regarder de plus près, on constate toutefois que de nombreux points restent encore à corriger, en dépit du temps considérable passé à annoter des extraits pour l'entraînement. Confronté à la manière non-systématique dont Furetière, puis Basnage et ses collaborateurs, livrent des informations, en plaçant prononciation, étymologie, sens et phraséologie dans différents endroits et en les combinant de façon souvent peu prévisible, GROBID-Dictionaries s'est plus d'une fois « emmêlé les pinceaux », comme l'on dit familièrement, agglutinant des informations distinctes ou, au contraire, classant sous des rubriques différentes des éléments qui auraient dû aller ensemble. Un aspect qui lui a particulièrement posé problème est la distinction entre les exemples forgés et les citations, qui restent à ce jour à identifier manuellement. Une nouvelle génération de GROBID, actuellement en préparation et fondée sur une technologie différente,

[34] Mohamed Khemakhem, Luca Foppiano, Laurent Romary, « Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields », in *electronic lexicography,* Leiden, eLex Conference, 2017. hal-01508868v2 (consulté le 20 mai 2021).

[35] L'ensemble est consultable à l'adresse https://nakala.fr/collection/11280/542f1501.

pourra probablement alléger en partie ce travail de bénédictin, mais il est difficile d'imaginer un horizon temporel où l'on n'y aura plus recours : 6000 fois plus acérée que l'intelligence artificielle[36], celle des humains, que les études littéraires continuent à aiguiser et à démultiplier, reste indispensable pour départager et faire sens à partir de telles données.

## Conclusion

Conscientes de l'importance croissante du numérique dans la société moderne, de nombreuses institutions d'enseignement supérieur et de recherche cherchent à s'engager dans le mouvement en investissant dans les outils et les formations. Dans les pays les plus riches, des programmes réguliers de numérisation sont lancés ou se poursuivent, et les infrastructures se complètent et se démocratisent. Ce que cet article aura essayé de mettre en évidence, c'est que le succès du « tournant numérique » dépend aussi d'un troisième facteur, qui est celui de l'engagement des humanistes dans la conception et l'amélioration de l'écosystème. La révolution numérique n'est pas seulement une question d'apprentissage d'un savoir-faire que l'on transporte et adapte à son propre champ : la compréhension des limites de ce savoir et la conscience du travail ingrat qu'il s'agit de mener pour son expansion sont partie intégrante de la transformation. Nul besoin d'être un crack en JavaScript ou Python pour prendre part aux chantiers rapidement décrits plus haut, et le succès des campagnes de *crowd-sourcing* montre que nombreux sont les membres du large public qui s'en sont rendu compte. Il est d'autant plus dommage, dès lors, que les spécialistes des lettres ne soient pas plus nombreux à s'embarquer dans l'entreprise, paralysés par de fausses représentations d'un travail qui a besoin, à plus d'un titre, de leurs lumières. Les associations savantes, déjà existant dans nos disciplines, constituent des espaces de discussion et d'engagement, à partir desquelles des plans d'action peuvent être

[36] Ordre de grandeur proposé par Yan Le Cun dans *Quand la machine apprend. La révolution des neurones artificiels et de l'apprentissage profond*, Paris, Odile Jacob, 2019.

tracés : programmation des numérisations en fonction des besoins et en tenant compte des exigences scientifiques d'équilibre et de représentativité ; constitution de groupes *ad hoc* de réflexion sur des vocabulaires contrôlés ; mise en place d'interfaces de discussion avec les informaticiens, afin de demander l'ouverture (ou la réouverture !) de chantiers sur des objets d'études qui nous sont propres ; réflexion sur les modalités d'évaluation du travail digital, qui ne donne pas lieu aux mêmes types de produits que la recherche « traditionnelle » – voici autant de directions dans lesquelles il est urgent d'agir. Les humanités numériques restent humaines non seulement parce que leurs objets sont ceux qui intéressent les spécialistes des disciplines dites « non scientifiques », non seulement parce que leurs résultats augmentent notre compréhension de l'humain et de la société, mais aussi parce que le travail humain a encore de beaux jours devant lui dans la sphère numérique.

**Bibliographie**


Douehi, Milad, *Pour un humanisme numérique*, Paris, Éditions du Seuil, 2011.

Edmond, Jennifer (sous la direction de), *Digital Technology and the Practices of Humanities Research*, Cambridge, Open Book Publishers, 2020.

Klein, Laura F. et Gold, Matthew K., « Digital Humanities: the Expanded Field », in *Debates in the Digital Humanities 2016*, sous la direction de Matthew K. Gold et Lauren F. Klein, Minneapolis et Londres, University of Minnesota Press, 2016 [s. p., en ligne: https://doi.org/10.5749/9781452963761]

McCarty, Willard, *Humanities Computing*, Basingstoke, Palgrave MacMillan, 2005.

Rockwell, Geoffrey, « What Is Text Analysis, Really ? », *Literary and Linguistic Computing*, no. 18.2, 2003, p. 209-219.

Terras, Melissa, « Peering inside the Big Tent », in *Defining Digital Humanities: A Reader*, sous la direction de Melissa Terras, Julianne Nyhan et Edward Vanhoutte, Farnham, Ashgate, 2013, p. 263-270.