Livre des résumés: dix ans avec CAHIER : des corpus d'auteurs pour les humanités à leur exploitation numérique

Ioana Galleron, Idmhand Fatiha

Dix ans avec CAHIER : des corpus d'auteurs pour les humanités à leur exploitation numérique

Conférence internationale

Livre des résumés
Réunis et présentés par

Ioana GALLERON et Fatiha IDMHAND

7-10 juin 2021
Bordeaux
France

SOMMAIRE

Introduction

Fatiha Idmhand
ITEM, UMR 8132
Université de Poitiers, France

Ioana Galleron
LATTICE, UMR 8094
Université Sorbonne-Nouvelle, France

Pour sa dernière année d'existence, le consortium « Corpus d'auteurs pour les humanités : informatisation, édition, recherche » (CAHIER) a souhaité réunir l'ensemble des projets adhérents, mais aussi des représentants d'autres projets et initiatives numériques, pour une rencontre sur les nouveaux savoirs produits dans le domaine des sciences du texte grâce aux corpus numériques et aux bases de données.

Plutôt qu'un bilan du consortium, ce colloque s'est voulu un espace de dialogue entre les différents spécialistes des corpus d'auteurs sur les résultats de l'exploration de ces ressources après leur constitution, gestion, publication et pérennisation, ou après la création de nouvelles applications ou plateformes. Il s'agissait ainsi non pas de présenter un projet, une ressource, ou les facilités de consultation et d'interrogation de celle-ci (pas seulement et pas exclusivement en tout cas), mais de se concentrer sur les questions scientifiques que l'on peut résoudre de la sorte, ou sur la constitution de nouveaux champs d'interrogation dans le sillage de la production de corpus numériques. Le colloque attendait tout particulièrement des propositions et des communications permettant de mesurer la distance entre les présupposés théoriques de départ, qui ont mené à la constitution du ou des corpus, et les résultats concrets ou les perspectives épistémologiques dégagées à partir de ceux-ci.

Deux axes de réflexion ont été proposés :

1. Nouveaux regards sur l'histoire littéraire ou l'histoire des idées

Le « distant reading » de Franco Moretti, les « macro-lectures » de Matthew Jockers et autres études portant sur l'analyse des données du net ont ouvert des pistes et donné des idées sur la façon de penser les phénomènes culturels observables grâce à de grands volumes de données. Toutefois, nous sommes loin d'avoir épuisé la matière et une nouvelle histoire de la littérature, nationale ou mondiale, comme de nouvelles histoires des idées et des arts, restent à écrire. Des critiques commencent d'ailleurs à émerger à propos des approches citées plus haut, appelant par exemple à la constitution d'une « data-rich literary history » (K. Bode) qui prenne mieux en compte les complexités de l'articulation entre recensement numérique de sources, numérisation des textes, création des données et contextualisation historique.

Dès lors, dans le cadre de cet axe le colloque souhaitait accueillir des communications portant, par exemple, sur :

- l'exploration des données des catalogues de bibliothèques ;
- l'analyse en masse de mots clés liés à des projets de numérisation ;
- les thésaurus et les ontologies pour l'étude de la littérature, de l'histoire des idées, de la philosophie, et d'autres disciplines des SHS ;
- les régularités stylistiques et thématiques à l'échelle d'une génération d'écrivains ;

- les ruptures et les innovations nationales et transnationales détectées par ordinateur ;
  etc.

2. Linguistique, poétique et génétique numériques

Depuis plusieurs années, l'étude des discours, des genres et des styles a été renouvelée grâce à l'apport de l'informatique. Des ouvrages et numéros de revue récents montrent la fécondité de l'exploration textométrique et stylométrique de grands corpus littéraires, de l'analyse des chaînes de co-référence, de l'interrogation des motifs, de l'identification des entités nommées, des tâches de classification et autres techniques du TAL appliquées aux œuvres littéraires (voir, entre autres, Legallois et al., 2018 ; Frontini et Ganascia, 2019 ; Lebart, Pincemin et Poudat, 2019). Ces travaux avaient toute leur place dans le colloque qui espèrait, en outre, susciter un intérêt similaire auprès des spécialistes de la littérature qui se sont moins emparés de ces techniques et de ces résultats. Comment nos perspectives changent-elles sur le fonctionnement narratif, dramatique ou lyrique des textes à la lumière de ces apports ? Comment l'ordinateur permet-il d'observer sur de nouvelles bases la co-construction du sens grâce à la lecture ? Plus largement, quel est l'apport de l'ordinateur à la question insoluble de l'essence de la littérature ? Telles étaient quelques-unes des questions auxquelles le colloque espérait répondre.

Par ailleurs, d'intéressantes perspectives ont été ouvertes par le numérique pour étudier, sous des angles nouveaux, le processus de création des œuvres et les avants-textes. Alors que le traitement informatique permet une plus grande précision de la démarche processuelle de la génétique grâce aux images en haute résolution à partir desquelles le généticien peut travailler depuis n'importe quel pays du monde (De Biasi, 2010), l'édition numérique permet de mieux croiser les approches philologiques et génétiques (D'Iorio, 2010), tandis que le balisage des textes (en format XML/TEI) a enfin permis de visualiser la superposition de phases génétiques et de visualiser les strates et traces du processus de création (André, Pierazzo, 2013 ; Van Hulle, 2018). Même si la plupart des propositions (outils, langages ou méthodes) sont encore exploratoires, la génétique a vu émerger de nouveaux enjeux pour son champ avec le document nativement numérique, des « brouillons » et des traces sous forme de disques durs ou de cloud. Il semblait donc intéressant de s'interroger sur la place de l'herméneutique génétique dans ces travaux et projets, sur les progrès possibles dans le domaine de la visualisation des processus de création et sur les perspectives scientifiques de la génétique en tant que « science des processus » (De Biasi, 2017).

Enfin, au-delà de ces deux axes, le colloque se proposait d'accueillir toutes les propositions, venant d'horizons disciplinaires variés, qui concernaient l'exploitation des corpus d'auteurs en milieu numérique. Étaient également bienvenues les communications réfléchissant sur les limites (matérielles ou conceptuelles) des approches s'inscrivant dans le spectre des humanités numériques. Les participants étaient encouragés à mettre en évidence non seulement les résultats obtenus grâce aux corpus qu'ils ont construits ou interrogés, mais aussi les « angles morts » liés à la quantité, la nature et la structure des données utilisées. De même, étaient les bienvenues les réflexions portant sur la réutilisation de corpus produits ailleurs et sur la façon de gérer l'hétérogénéité des données qu'une telle réutilisation implique.

Or, le constat est sans appel : quoique tous les participants aient joué le jeu et essayé d'inscrire leur communication dans un des trois axes du colloque, nombreux sont les débordements par rapport aux cadres proposés : histoire littéraire, poétique et génétique, expérimentations. Telle communication dont l'objet s'inscrit dans le domaine de l'histoire littéraire implique naturellement des observations sur la poétique des textes, telle qu'elle peut être explorée grâce à l'ordinateur, et comporte une dimension expérimentale de ce même fait. À l'inverse, expérimentations et travaux

de textométrie ou de génétique assistée par le numérique s'inscrivent dans une histoire – de la littérature, du livre ou plus largement des idées.

Il a fallu dès lors renoncer à partager les textes selon ces catégories initiales, et essayer de les redistribuer selon d'autres affinités. Un partage possible s'est avéré celui entre des travaux portant sur un auteur, un groupe de textes issu d'une même plume, et des études fondées sur des corpus plus divers, construits ou non par les communicants. Dans le premier cas, le défi dont parlent tous les textes ici présentés est celui de rendre compte de la diversité d'un archipel textuel qu'unifie trompeusement la figure de l'« auteur ». Dans le second, les outils, les vocabulaires, les méthodes, les enjeux épistémologiques de la recherche des similitudes sont à l'honneur.

Nombreux sont les échos, les liens que l'on peut percevoir entre les communications, par-delà ces nouvelles catégories, de sorte que ces nouvelles rubriques peuvent à leur tout être mises en question. Mais elles ont au moins le mérite de correspondre plus étroitement aux propositions suscitées par notre appel à communications. Peut-être est-il permis de considérer qu'elles constituent une forme de bilan des dix années de CAHIER – non pas des corpus produits, des formations, ateliers et réunions de travail dans le cadre de ce consortium, mais du changement des méthodes qu'il aura provoqué et des perspectives qu'il aura ouvertes dans la communauté des chercheurs en sciences du texte.

Références

André, Julie ; Pierazzo, Elena, « Le codage en TEI des brouillons de Proust : vers l'édition numérique », *Genesis* [Online], 36|2013, Online since 09 July 2015, connection on 27 September 2020. URL: http://journals.openedition.org/genesis/1159 ; DOI : https://doi.org/10.4000/genesis.1159

Bode, Katherine, *A World of Fiction: Digital Collections and the Future of Literary History*, University of Michigan Press, 2018.

D'Iorio, Paolo, « Qu'est-ce qu'une édition génétique numérique? », in *Genesis*, 30, 2010, 49-53.

De Biasi, Pierre-Marc, Herschberg Pierrot, Anne, *L'œuvre comme processus*, CNRS Edition, Paris, 2017.

Galleron, Ioana ; Idmhand, Fatiha, « 'Réutilisabilité' : L'utilisateur dans l'édition électronique », *Humanistica*, numéro 1, 2019. https://revues.univ-lyon3.fr/humanites-numeriques/

Galleron, Ioana ; Idmhand, Fatiha ; Meynard, Cécile, « Que mille lectures s'épanouissent… Modélisation du personnage et expérience de 'crowdreading' » in *Digital Humanities Quaterly*, 1938-4122, 2018, http://www.digitalhumanities.org/dhq/vol/12/1/000363/000363.html

Galleron, Ioana ; Fatiha Idmhand ; Marie-Luce Demonet; Cécile Meynard; Elena Pierazzo, et al. *LES PUBLICATIONS NUMERIQUES DE CORPUS D'AUTEURS - Guide de travail, grille d'analyse et recommandations* (V1-Novembre 2018). [Rapport de recherche] Huma-Num ; identifiant : halshs-01932519

Ganascia, Jean-Pierre, Frontini, Francesca, *TAL et humanités numériques*, numéro spécial de la *Revue TAL*, vol. 60, no. 3, 2019.

Jockers, Matthew, *Macroanalysis. Digital Methods and Literary History*, University of Illinois Press, 2013.

Lebart, Ludovic, Pincemin, Bénédicte et Poudat, Céline, *Analyse des données textuelles*, Presses de l'université du Québec, 2019.

Legallois, Dominique, Thierry Charnois, and Meri Larjavaara, *The Grammar of Genres and Styles. From Discrete to Non-Discrete Units*, De Gruyter Mouton, 2018.

Moretti, Franco, *Distant reading*. New York, Verso, 2013.

Van Hulle, Dirk, « Intégrer la bibliothèque d'écrivain aux éditions génétiques numériques : le cas Beckett », *in* Belin, Olivier, Mayaux, Catherine, Verdure-Mary, Anne, *Bibliothèques d'écrivains : Lecture et création, histoire et transmission*, Rosenberg & Sellier, Torino, 2018. https://doi.org/10.4000/books.res.1856.

Première partie. Du corpus d'auteur à son exploitation

L'édition critique multimodale de sources anciennes. Une recherche collaborative pour la création de nouveaux outils

Marie-Agnès Lucas-Avenel
Centre Michel de Boüard
CRAHAM, UMR6273
Université de Caen Normandie, France

Marie Bisson
Maison de la recherche en sciences humaines
USR3486
Université de Caen Normandie, France

La situation des études visant à l'édition scientifique numérique des sources anciennes est paradoxale : l'accès en ligne aux textes anciens grâce à la numérisation des collections et des témoins manuscrits ainsi que leur insertion dans des bases de données permettent de démultiplier les modes d'interrogations et les recherches pluridisciplinaires sur ces sources, dont la lecture est indispensable à la connaissance des époques anciennes et médiévales. En outre, les spécialistes de ces périodes, qu'ils soient historiens ou littéraires, ont développé depuis plusieurs années des outils informatiques performants pour leurs recherches (en France, par exemple, les sites Ménestrel et celui de l'Equipex Biblissima, ainsi que ceux créés par l'IRHT ou le département des manuscrits de la BNF). Pourtant, très rares encore sont les projets qui aboutissent à de nouvelles éditions critiques numériques de sources anciennes et médiévales[1]. Parmi les diverses raisons, on peut invoquer l'aspect chronophage de ce type de recherches[2], qui s'accommode mal d'une programmation de la recherche sur des contrats courts. Néanmoins, la collaboration qui a été mise en place à Caen entre le CRAHAM (UMR 6273), le Pôle Document numérique (PDN) de la Maison de la Recherche en Sciences de l'Homme (USR 3486) de l'Université de Caen Normandie -- en lien avec des consortiums nationaux et internationaux (Huma-Num, Equipex Biblissima, GDRI Zoomathia) -- et les Presses universitaires de Caen (PUC) a permis de mener une réflexion sur le choix et l'utilisation des outils et de proposer de nouvelles pratiques. Suivant une démarche à la fois empirique et heuristique, cette collaboration a abouti à la création d'environnements numériques en XML-TEI outillant la recherche pour la réalisation d'éditions de sources, avec des retombées importantes aussi bien pour l'éditeur scientifique que pour les lecteurs (chercheurs et autres). Certaines des préoccupations de l'éditeur scientifique rejoignent en outre celles de l'éditeur matériel, qui l'un et l'autre veulent permettre aux lecteurs d'aujourd'hui d'accéder au texte ancien. Cela suppose, en partie, d'extraire le texte, dans ses différents états, de tout ce qui le rend abscons aux non spécialistes, en particulier la multiplicité de ses réalisations sur des supports variés, pour le restituer sous une autre forme et sur un autre support, papier ou numérique, tout en rendant compte de l'histoire du texte, de ses spécificités et des choix qui ont présidé à la conception de l'édition.

La communication présentera d'abord le contexte scientifique qui a permis de mettre en place ce partenariat interdisciplinaire (recherche philologique, ingénierie de recherche en humanités numériques et ingénierie en édition matérielle multimodale), puis exposera certains des résultats obtenus, grâce à ce partenariat, dans le domaine de l'ecdotique outillée par ordinateur : on montrera

[1] Pour les éditions de textes latins, l'un des précurseurs a été Bruno Bureau et le projet « Hyperdonat ». Ajoutons aussi l'édition du *De natura deorum* de Cicéron, PUC, par Clara Auvray-Assayas (https://www.unicaen.fr/puc/sources/ciceron/accueil).

[2] Cet aspect a été dénoncé, par exemple, par E. Pierazzo, « Digital Documentary Editions and the others », *Scholarly Editing : The Annual of the Association for Documentary Editing*, 2014, 35, p. 1-23 (PDF consultable en ligne).

en particulier combien la création d'un environnement XML dédié facilite et rend plus sûr le travail qui consiste à établir le texte et multiplie les possibilités de recherche sur la source et ses témoins à la fois pendant le travail de préparation à la publication, ainsi que pour et après la publication.

On s'appuiera sur l'édition critique de *l'Histoire du Grand Comte Roger* par Geoffroi Malaterra, composée en latin vers 1100. L'œuvre est la source littéraire la plus riche en informations sur les membres de la famille de Hauteville et sur les exploits accomplis par chacun d'eux, en particulier par Roger en Sicile. Le texte latin, qui est divisé en quatre livres et rédigé à la fois en prose et en vers, a été établi grâce à la collation exhaustive des manuscrits primaires (A, B - ou Be, copie de B, quand ce dernier fait défaut -, C, Z, conservés dans diverses bibliothèques de Sicile, Besançon et Barcelone), et des deux éditions qui ont marqué son histoire (la princeps de Geronimo Zurita [1578] et celle d'Ernesto Pontieri [1927-28]). Le texte latin est enrichi de trois types de notes : notes d'apparat critique, notes de sources, notes philologiques. Il est accompagné de la traduction française pourvue de notes explicatives, et l'ensemble est indexé (toponymes et anthroponymes). Suite à la redécouverte du manuscrit Z, Gianvito Resta avait montré les manques de l'édition de Pontieri et la nécessité de réaliser à nouveaux frais une édition scientifique[3].

Le travail de recherche pour cette édition, mené en même temps que la découverte des nouveaux outils, nous a incités à choisir le texte de Malaterra comme terrain expérimental pour l'encodage fin des variantes textuelles. En effet, d'une part, l'un des apports essentiels de la recherche sur l'œuvre de Malaterra était d'offrir un nouveau texte scientifiquement édité et pourvu d'un apparat critique complet et justifié ; d'autre part, le nombre de témoins retenus était raisonnable (sept en tout), avec des mains supplémentaires pour les témoins primaires, et les types de variantes textuelles (omissions et lacunes plus ou moins longues, additions, transpositions, substitutions, etc.) étaient suffisamment divers pour tester les différents cas de figure susceptibles d'intervenir dans d'autres travaux d'ecdotique. On a voulu vérifier si l'utilisation du module « Critical Apparatus » de la TEI, en suivant la méthode de la segmentation parallèle, permettait d'encoder les variantes en respectant les recommandations des Belles Lettres et de l'École des Chartes pour l'édition critique de sources et, le cas échéant, quel pouvait en être l'intérêt pour la recherche en ecdotique, mais aussi dans d'autres domaines : paléographie latine, histoire des textes, linguistique latine, historiographie et poésie latines.

Cette méthode permet d'éviter d'avoir à choisir un témoin de base et offre la possibilité d'attribuer à chacun le même statut. Elle permet ensuite d'obtenir une restitution de chacun en plus du texte édité, dans laquelle le lieu variant est marqué par un jeu de couleurs opposant le lemme à la leçon rejetée. Ce travail exige, d'une part, de définir précisément ce qu'il semble intéressant de marquer, ce qu'on cherche à obtenir et, éventuellement, d'opérer des choix, dont certains résultent du tâtonnement de la recherche. D'autre part, ce travail impose rigueur, précision et exhaustivité : tout manque, par exemple, fera croire, sans ambiguïté, à une lacune dans le témoin et toute erreur dans l'association du témoin à sa variante produira une incohérence dans sa restitution. Par ailleurs, tout le travail d'ecdotique sera aisément vérifiable par le lecteur grâce à la publication en ligne des fac-similés. Il s'agit là, sans doute, d'un des aspects à la fois les plus séduisants et les plus terrifiants de l'encodage en TEI de l'apparat critique selon la méthode de la segmentation parallèle : en offrant au lecteur la possibilité de disposer des restitutions de chaque témoin à partir de l'encodage des variantes, il impose une très grande discipline, mais il offre en même temps un formidable outil de relecture, de vérification et de comparaison des témoins.

En procédant ainsi, on a constaté que le module de la TEI donnait la possibilité d'encoder tous les types de variantes, à condition toutefois de renoncer à certaines habitudes des rédacteurs d'apparat critique, en particulier quand elles risquaient de provoquer des enchâssements ou chevauchements incompatibles avec le schéma XML[4]. Cette concession pourrait paraître

---

3 Resta G. (1964), « Per il testo di Malaterra e di altre cronache meridionali », in *Studi per il CL anno del Liceo ginnasio Tomasso Campanella di Reggio Calabria*, Reggio de Calabria, De Franco, p. 3-60.

4 Exemple de transposition qui affecterait hostes (I, 16, 1) : ut timidus hostes devitando. Comment faire en sorte que l'encodage ne marque pas hostes comme lieu variant ?

secondaire, sinon pour une publication papier, du moins pour une publication en ligne, dans laquelle les restitutions des témoins se substitueraient à des notes traditionnelles d'apparat critique ; cependant, il nous semble au contraire que l'apparat critique associé au texte édité doit rester l'outil d'accès privilégié à la compréhension de l'histoire textuelle qu'il a vocation à être et que l'édition numérique peut jouer un rôle dans la redécouverte de cet outil scientifique, qui est bien souvent délaissé dès qu'il est achevé. La méthode appliquée à l'encodage a évolué grâce à la création d'outils de plus en plus sophistiqués au sein de l'environnement dédié. Dans un premier temps, afin d'obtenir un apparat rédigé selon les normes traditionnelles pour la publication papier, on a créé un type de notes « apparat » et rédigé chaque unité critique telle qu'elle doit apparaître sur le papier en plus de l'encodage des variantes en XML TEI (version P5[5]). Ce choix entraînait une double opération, à la fois chronophage et source d'erreurs d'écriture, si bien qu'il n'était pas sûr que les apports post-publication suffisent à séduire de nouveaux chercheurs. Dès lors, considérant les difficultés et les enjeux, des améliorations ont été apportées par le PDN pour outiller la recherche : le nouvel environnement[6] conçu pour XMLMind XML Editor (version 8.3) permet désormais d'encoder une source ancienne, avec des fonctionnalités génériques et plus performantes, qui réduisent les risques d'erreurs et permettent d'obtenir semi-automatiquement la rédaction de la quasi-totalité des unités critiques. Ainsi, on peut d'autant mieux se concentrer sur le travail critique que le numérique offre de formidables solutions pratiques pour la comparaison des témoins : la vue synoptique des restitutions permet de considérer les variantes dans leur contexte phrastique et non plus de manière extrêmement fragmentée ; elle lève en outre les ambiguïtés que l'apparat critique seul peut parfois laisser par une forme de discours implicite[7]. Grâce à cet environnement de travail -- outil d'encodage et prototype de sites --, dont on fera une présentation, l'éditeur dispose d'un nouveau laboratoire de recherche qui modifie de manière conséquente les méthodes de réalisation de son travail et offre de nouvelles perspectives.

Cette communication vise ainsi à encourager les jeunes chercheurs à réaliser des éditions de sources multimodales à partir des outils accessibles et gratuits qui seront présentés, car ils permettent de répondre aux exigences des éditions traditionnelles, tout en améliorant les conditions de la comparaison textuelle ; en outre, ils facilitent l'encodage des données et leur interopérabilité. En effet, considérant la situation dans laquelle se trouvent les humanités classiques, l'utilisation des corpus de sources mis à disposition grâce au numérique restera l'apanage d'une minorité d'érudits, si ceux-ci ne sont pas accompagnés de l'enrichissement scientifique nécessaire à leur compréhension.

Références complémentaires

Andrews T., « The Third Way : Philology and Critical Edition in the Digital Age », *Variants*, 10, 2013, p. 61-76.

Apollon, Régnier, Bélisle (dir.), *L'Édition critique à l'ère du numérique*, Paris, L'Harmattan, 2017 (trad. fr. de *Digital Critical Editions*, University of Illinois Press, 2014).

Duval F., « Pour des éditions numériques critiques. L'exemple des textes français », *Médiévales*, 73, 2017, p. 13-30.

5 Ce travail a donné lieu à la publication papier et numérique de Geoffroi Malaterra, *Histoire du Grand Comte Roger et de son frère Robert Guiscard*, livres I & II, vol. 1, Marie-Agnès Lucas-Avenel (éd.), Caen, Presses universitaires de Caen (Fontes & Paginae), 2016, consultable en ligne : https://www.unicaen.fr/puc/sources/malaterra/accueil.

6 Cet environnement de travail pour l'édition de sources avec apparat critique en XML et la méthodologie d'encodage sont consultables et téléchargeables en ligne :
http://www.unicaen.fr/recherche/mrsh/document_numerique/outils/apparat.

7 Cette nouvelle méthode a été utilisée pour l'édition critique multimodale des livres III et IV, dont l'ensemble a constitué l'inédit de l'HDR, «Éditer des œuvres latines à l'ère du numérique. Édition critique multisupport de Geoffroi Malaterra, *Histoire du grand comte Roger et de son frère Robert Guiscard*(Livres III & IV) » (garant C. Jacquemard), 6 décembre 2019, Université de Caen Normandie.

Bisson Marie, Cannet Edith, Lucas-Avenel Marie-Agnès, *Environnement pour réaliser des éditions avec apparat critique en XML TEI P5*, documentation du Pôle document numérique, 2020, consultable en ligne : http://www.unicaen.fr/recherche/mrsh/sites/default/files/public/document_numerique/manuel_apparat.pdf.

Queste del saint Graal, Christiane Marchello-Nizia et Alexei Lavrentiev (éd.), Lyon, ENS de Lyon, 2019. Publié en ligne par la Base de français médiéval. Dernière révision le 2018-11-30.

Mathieu d'Edesse, *Chronique*, Tara Andrews (éd.), 2012, consultable en ligne : https://byzantini.st/ChronicleME/

Louis Meigret et la réutilisabilité des données

Cendrine Pagani-Naudet et **Laurent Vanni**
Bases, Corpus, Langage, UMR 7320
Université Nice Sophia Antipolis, France

La question de la réutilisation est au cœur du projet de la « Base Louis Meigret[1] ». Au-delà des gestes techniques que suppose la mise à disposition de données réutilisables, c'est un principe adéquat à la singularité de l'œuvre de Louis Meigret.

Le projet est né à l'occasion du colloque consacré à cet auteur en 2018 à Nice. Il s'agissait au départ de créer un lieu de référence, destiné à favoriser les échanges entre chercheurs. Les personnes qui travaillent sur Meigret ne se connaissent pas toujours, et pour peu qu'elles évoluent dans des cercles différents, ignorent leurs activités respectives. Au moment du colloque une base venait d'être mise en ligne (Bettens 2017), parallèlement, et à peu près à la même époque, *Le Tretté* faisait l'objet d'une transcription par A. Pelfrêne et B. Colombat (mise en ligne sur le site du CTLF).

Ce doublon, outre la fragmentation qu'il manifeste, rappelait un déséquilibre : la surexposition de certaines œuvres du grammairien, notamment *Le Tretté de la grammere françoeze*, et le relatif oubli de l'œuvre traduite. En outre, les œuvres linguistiques - et particulièrement *Le Tretté* - ont elles-mêmes été enfermées dans des modes de lecture qui en occultent bien des dimensions.

Compte tenu de la grande diversité des travaux suscités par Louis Meigret (histoire de la langue, histoire des idées, histoire des techniques, histoire du livre), ce que devait offrir cette base à l'usager était à définir (ou peut-être à ne pas définir). D'emblée donc s'est posée la question de la réutilisation : réutilisation des données existantes (la base conçue par O. Bettens), réutilisation de celles qu'on envisageait de mettre à disposition des futurs usagers (l'intégralité de l'œuvre de Louis Meigret, textes personnels et traductions).

Le projet était *a minima* de rassembler l'ensemble des textes, de faciliter la circulation d'un texte à l'autre, et de faire jaillir la cohérence quasiment organique de l'œuvre de Louis Meigret. Il s'agissait de permettre une appréhension globale de l'œuvre, tout en conservant la possibilité de revenir à la spécificité de chaque texte, sans l'assigner a priori à un genre (traité sur la langue / traduction), sans conditionner sa lecture par un outillage envahissant qui le rendrait « illisible » sous d'autres approches. Toutes les recommandations favorisant la réutilisation des données rejoignent donc la conviction que pour mieux connaître l'œuvre de Louis Meigret, l'essentiel réside dans la liberté laissée à l'usager (qu'il soit chercheur ou lecteur non expert) de choisir son mode de lecture (choix du support, choix des outils de visualisation et d'exploration), de réinventer le texte en modulant les points de vue.

La base est à l'heure actuelle dans une phase intermédiaire : elle est utilisable mais en train de se faire. Utilisable parce que les traductions ont été numérisées (pour celles qui ne l'étaient pas) transcrites et mises en ligne sur le site d'Hyperbase[2], avec toutes les fonctionnalités que permet cet outil. La matière est disponible mais encore partiellement fragmentée, et pas forcément réutilisable. Il convient de réfléchir à la manière de faire fusionner les deux bases existantes : celle qui rassemble les traductions et celle qui concerne les textes en graphie rénovée. Les deux ensembles ont leur cohérence interne mais cela ne justifie pas une dualité que dément la démarche originale de Meigret. Démarche qui se manifeste de manière exemplaire dans une œuvre comme *Le Menteur*. *Le Menteur* est une œuvre bicéphale. Par sa préface, c'est un traité sur l'orthographe. Intégralement composé en graphie rénovée, il appartient à la base conçue par O. Bettens qui rassemble les autres textes de Meigret relatifs à la langue française publiés chez C. Wechel. Conserver les particularités graphiques

[1] https://meigret.j2p.fr/
[2] http://hyperbase.unice.fr/hyperbase/

et ménager un accès au support original est donc crucial. Mais Le Menteur intéresse aussi l'histoire de la traduction : le dialogue de Lucien engage vers d'autres formes de lecture et d'exploration qui peuvent justifier son intégration à la base des traductions, et à d'autres corpus (alignement avec les textes sources et les traductions ultérieures).

Enfin reste à définir l'interface qui permettra de répondre aux questions qui ont motivé la création de l'outil, tout en restant disponible pour des investigations nouvelles. Un projet d'édition collective sur les traductions de Louis Meigret doit créer les conditions d'un dialogue entre chercheurs d'horizons divers, et permettre à l'usager d'expérimenter l'outil et d'en suggérer des améliorations. On voudrait en somme que l'usager crée la base et puisse la recréer à chaque moment.

Références

Bettens, Olivier, *Louis Meigret. Corpus phonétique*, édition électronique lemmatisée par Olivier Bettens, 2017, http://virga.org/phon16/.

Colombat, Bernard, Pelfrêne, Arnaud, 2017, *Le Tretté de la grammere françoeze, Le traité de la grammaire française*, Texte originel numérisé avec l'image de page correspondante, http://ctlf.ens-lyon.fr/i_tdm.asp?v=164.

Colombat, Bernard, Fournier, Jean-Marie, *Corpus des grammaires françaises de la Renaissance*, Garnier numérique, 2011.

Galleron, Ioana, Idmhand, Fatiha, « 'Réutilisabilité' : L'utilisateur dans l'édition électronique », *Humanistica*, numéro 1, 2019. https://revues.univ-lyon3.fr/humanites-numeriques/

Meigret, Louis, *Le Mẹnteur, ou l'incredule de Lucian, traduit de Grẹc en Frãçoẹs par Lou'is Meigrẹt* Paris, Chrestien Wechel, 1548.

Meigret, Louis, *Le Tretté de la grammere françoeze*, Paris, Chrestien Wechel, 1550.

Montagne, Véronique, Pagani-Naudet, Cendrine, (dir.), *Actualités de Louis Meigret, humaniste et linguiste*, Paris, Garnier, 2021.

Vanni, Laurent, 2017, « Hyperbase Web : Outil d'analyse statistique des données textuelles », ECLAVIT 2017, Paris, France. https://hal.archives-ouvertes.fr/hal-01804331

Éditer pour donner prise, éditer pour rendre visible. Le projet *Mythologia* et l'étude des processus de constitution du savoir dans une mythographie de la Renaissance

Céline Bohnert
Institut Universitaire de France
CRIMEL
Université de Reims Champagne-Ardennes, France

« [L]es penseurs de la Renaissance nous ont appris à mettre notre culture en perspective, à confronter nos coutumes et nos croyances avec celles d'autres temps et d'autres lieux. En un mot, ils ont créé les outils de ce qu'on pourrait appeler une technique du dépaysement. » (Lévi-Strauss, 2011)

Claude Lévi-Strauss définit ici un premier âge de l'humanisme, caractérisé par le recours aux textes et aux monuments, à défaut d'un lien direct avec l'altérité en rapport avec laquelle il s'élabore, l'Antiquité gréco-romaine. Aujourd'hui, au moment où l'humanisme de l'ère numérique cherche à se penser (Citton, 2010 ; Doueihi, 2011) et où il tend parfois à se définir par ses outils, les « techniques du dépaysement » élaborées par les savants de la Renaissance trouvent une actualité renouvelée. Pour autant, le lecteur contemporain s'y confronte aussi à une étrangeté accrue, tant leurs fondements épistémologiques diffèrent des nôtres.

Le projet « Mythologia » entend observer ces « techniques du dépaysement » et les gestes intellectuels qui les fondent à travers l'édition d'une encyclopédie sur les croyances et les rites antiques, la *Mythologie* de Natale Conti. Empruntant aux autres mythographes renaissants, Conti expérimente simultanément des approches très différentes des corpus antiques, de l'élucidation philologique à une posture exégétique beaucoup plus marquée, voire à une forme de proto-anthropologie. Il déploie pour cela une forme de collectionnisme effréné propre à l'âge humaniste : en expérimentant différents modes d'organisation des textes et des images qui visaient à rendre compte de la totalité des connaissances, les érudits européens promouvaient le geste de la compilation comme une véritable méthode de production du sens.

Dans ce programme, l'édition de la *Mythologie* intervient à la fois comme préalable à l'étude, comme outil permettant la recherche et comme lieu de publication des résultats au fur et à mesure des étapes du projet – un lieu ouvert à de nouvelles manipulations pouvant confirmer, élargir ou tester à la fois les présupposés et les résultats présentés par l'équipe. C'est cette place spécifique de l'édition numérique dans un processus de recherche dont cette communication rendra compte : le geste éditorial est ici l'un des moments et l'un des moyens privilégiés de la recherche. Cette articulation entre exploration et édition est liée à la nature de l'objet de recherche : non un texte qu'il s'agirait de documenter, dans la tradition des éditions philologiques, mais des processus qu'il s'agit de rendre visibles. Ces processus sont de trois ordres.

- Il s'agit d'abord de variance : la *Mythologie* est un texte en mutation, qui, par là même, interroge les notions d'œuvre et de texte ainsi que les régimes et les frontières de l'autorité/auctoralité. L'édition rassemble donc quatre états significatifs du texte, deux états latins (Venise 1567 et Francfort 1581) et deux états français (Lyon 1612 et Paris 1627). Le pas de côté permis par l'édition numérique consiste dans l'étude de la constitution d'une constellation textuelle, plutôt que dans le recensement de variantes, notion qui suppose la prééminence d'un texte de référence sur d'autres versions considérées comme secondaires, préparatoires ou fautives.

- Le site *Mythologia* entend aussi faire voir les dynamiques qui président à la constitution du corpus. L'objectif de l'approche génétique est ici d'entrer dans l'atelier du mythographe afin de saisir sa façon d'agencer les savoirs et de dégager les modèles herméneutiques sous-jacents : entrer dans le laboratoire de Conti suppose de reconstituer, autant que possible, sa bibliothèque. Au-delà d'une étude des gestes intellectuels (Unsworth, 2000), l'enquête vise à expliciter les catégories de pensée qui président au regroupement des citations dans ces textes-centons.
- Enfin se pose la question des usages, de l'influence et de la circulation du texte. La démarche apparemment « bricoleuse » de Conti met en lumière le désir d'Antiquité de ses lecteurs ainsi que les mécanismes de légitimation du savoir. De même que les transformations du texte, l'étude de son succès, des stratégies de Conti et de ses éditeurs, traducteurs, illustrateurs et commentateurs, et de ses modes de diffusion doit contribuer à éclairer les processus instituants de l'Europe pré-moderne, qu'il faudra pister et intégrer sur le site.

Tenant compte conjointement de la vie propre du corpus et de ses variances, de son lien génétique avec un ensemble de corpus antérieurs et de ses usages, l'édition vise bien à donner prise sur l'objet et à faire voir une série de processus intellectuels et culturels qui président à une fabrique collective de l'Antiquité : elle est l'adjuvant et le moyen de la recherche en même temps qu'elle constitue un livrable en soi et une partie des résultats du projet.

Références

Citton, Yves, *L'avenir des humanités. Économie de la connaissance ou culture de l'interprétation ?,* Paris, La Découverte, 2010.

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

Lévi-Strauss, Claude, *L'anthropologie face aux problèmes du monde moderne*, Paris, Seuil, coll. La librairie du XXI^e^ siècle, 2011.

Unsworth, John, « Scholarly Primitives. What methods do humanities researchers have in common, and how might our tools reflect this », *Symposium on Humanities Computing : Formal Methods*, Experimental Practice. King's College, London, 2000.

Le *Parallèle des Anciens et des Modernes* de Charles Perrault, témoin d'une modernité conflictuelle

Delphine Reguig et **Emmanuelle Perrin**
IHRIM, UMR5317
Université Jean Monnet (Saint-Étienne), France

La proposition de communication porte sur le projet d'édition numérique du *Parallèle des Anciens et des Modernes* de Charles Perrault, placé sous la responsabilité de Delphine Reguig (IHRIM), professeure de langue et littérature françaises du XVII[e] siècle à l'université de Saint-Étienne. Paru en quatre tomes entre 1688 et 1697, ce texte central dans la Querelle des Anciens et des Modernes n'a jamais été édité, ni exploité dans son intégralité, en raison des difficultés posées par la richesse de l'intertextualité interdisciplinaire qu'il sous-entend. L'édition numérique s'est imposée pour offrir un cadre capable de traiter ces difficultés en rendant intelligible, grâce aux outils de navigation intra-textuelle et de traitement des données textuelles, la structuration logique du discours foisonnant produit par l'auteur.

Perrault met en scène un dialogue entre trois personnages : le Président, partisan des Anciens, l'Abbé, partisan des Modernes, et, entre les deux, le Chevalier, au tempérament enjoué et provocateur, prompt à mettre en perspective les deux aspects d'une modernité conflictuelle. La discussion se déroule à Versailles, cadre privilégié d'observation du « Siècle de Louis XIV ». Examinant les mérites respectifs des artistes, philosophes, savants antiques et contemporains dans l'architecture, la sculpture, la peinture, l'éloquence, la poésie, l'astronomie, la géographie, la navigation, la guerre, la philosophie, la médecine ou la musique, ce texte paraît emblématique d'une ambition, que nous appellerions aujourd'hui « interdisciplinaire », de refondation des savoirs. Pour couvrir l'ensemble des domaines traités par Perrault, l'équipe éditoriale est constituée de collègues spécialistes de littérature, histoire des idées, philosophie (en particulier des sciences et de la médecine), histoire de l'art, musicologie, histoire sociale. Le directeur du Centre de recherches du Château de Versailles nous accompagne en nous assurant la fiabilité et la richesse d'une iconographie conservée dans les archives du CRCV qui permet aux lecteurs de comprendre le cadre dans lequel se déroule le débat que retrace Perrault dans les quatre tomes de son œuvre.

Pour donner accès à ce texte fleuve voire labyrinthique, foisonnant de références, appelant largement l'iconographie et l'annotation, l'édition électronique, accessible à l'adresse [https://parallele-anciens-modernes.huma-num.fr], possède une véritable pertinence. La publication de ce texte se fonde sur un protocole éditorial destiné à en favoriser la lisibilité, avec la modernisation de la graphie et la régularisation de la ponctuation. Cette édition est également diffusée selon les principes FAIR : le texte est encodé en TEI, les métadonnées respectent les recommandations du consortium Cahier, l'indexation s'appuie sur un vocabulaire normalisé avec les notices d'autorités de la BNF. L'indexation fine reflète la densité de ce texte avec plus de 620 noms de personnes, 130 toponymes, 340 œuvres et 300 sujets. Elle ménage différents points d'accès et de recherche dans le texte (coréférences, recherche par personnage, par type de texte, etc.). Il s'agit de permettre à la fois une lecture continue et un parcours discontinu guidé par des repères balisés et mis en réseau. Dans la perspective du Linked Open Data, l'alignement sur les référentiels et les notices d'autorité (data.bnf.fr, IdREF, VIAF, ISNI) permet d'enrichir, d'uniformiser et de consolider les index en simplifiant les variantes des patronymes, toponymes et titres d'ouvrages. Par l'insertion de liens vers d'autres ressources numériques, l'enjeu est également d'accroître l'interopérabilité entre le texte édité et le réseau des références qu'il convoque. L'annotation est dans ce cadre un autre volet important de l'entreprise : elle comprend trois niveaux

d'intervention et propose des élucidations lexicales ou syntaxiques, des informations historiques ou contextuelles, des commentaires enfin permettant de situer et comprendre l'argumentation. Les quelque deux mille notes prévues témoignent encore une fois de la richesse de ce texte.

L'édition du *Parallèle* doit ainsi servir de base à une redécouverte de Charles Perrault, dont la lecture et l'étude sont aujourd'hui limitées aux Contes. Ce projet est donc une première étape vers la reconstruction de la figure de Perrault comme un polygraphe à l'écriture éminemment politique. Ancien bras droit de Colbert, Charles Perrault a vécu dans l'intimité du pouvoir et activement supervisé la construction du Château de Versailles comme lieu de rayonnement pour la monarchie absolue. Le choix de l'édition électronique est particulièrement pertinent dans ce cadre puisqu'il permet :

- de re-contextualiser les écrits de Perrault contemporains du *Parallèle* ; l'édition du *Parallèle* constitue la première ouverture vers l'élargissement du corpus patrimonial qui permettra de mieux connaître et comprendre cet auteur largement sous-investi par les études scientifiques.
- le texte du *Parallèle* permet d'approfondir la question ; du rapport entre littérature et pouvoir politique sous Louis XIV. Christian Jouhaud a bien montré la nature paradoxale de ce rapport dans son ouvrage Les Pouvoirs de la littérature mais sa réflexion porte essentiellement sur la première partie du XVIIe siècle : un prolongement vers les décennies de la fin du règne de Louis XIV paraît extrêmement pertinent car la Querelle est sans doute l'une des étapes majeures de l'évolution de ce rapport. Le texte de Perrault fournit des éléments de réflexion fondamentaux pour la saisie des enjeux politiques du développement des œuvres culturelles. Tout un pan de la réflexion concerne en particulier les enjeux de la traduction des œuvres de l'Antiquité en français, pensée comme légitime ou non, selon qu'il s'agit de donner accès à des textes perçus comme définitivement archaïques ou bien de faire rayonner un nouvel état de la culture française favorisé par la monarchie absolue. La nature politique du texte de Perrault engage par ailleurs une passionnante réflexion sur le temps à partir de l'idée d'historicité. L'interprétation axiologique de l'histoire développée par Perrault implique toute une vision de la diffusion du savoir et de la culture qui relativise la maîtrise scientifique des anciens. La nouvelle disponibilité du livre imprimé a en effet modifié pour Perrault le rapport à la connaissance : l'apprentissage par cœur cédant le pas à la réflexion et la méditation. Perrault semble décrire un processus de démocratisation de la République des lettres qui verrait l'érudition réservée à un petit nombre céder devant un plus grand égalitarisme affaiblissant le rapport d'admiration aux textes anciens pour lui substituer un rapport critique. Le paysage décrit par Perrault est un témoignage capital sur la constitution de cette nouvelle République des lettres dont il s'agit de dessiner les contours par une indexation fine et un réseau de liens hypertextuels dynamiques. À titre d'exemple, l'analyse du texte du *Parallèle* servira de base à la constitution d'un lexique numérique des temporalités classiques (siècle, époque, événement, modernité, tradition, héritage, antiquité, actualité, ancienneté, progrès, origine, modèle, répétition, relativité, éternité, rupture, cycle, providence, etc.), publié en ligne, fondement de la constitution rigoureuse du cadre notionnel d'époque et appelé à interagir avec d'autres bases de données.
- le texte du *Parallèle* nourrit enfin l'étude précise et approfondie de la séparation des champs du savoir, champs culturels, champs intellectuels dont Perrault montre à la fois la solidarité et la progressive distinction en fonction de méthodologies et perspectives divergentes. Et cela d'autant plus que Perrault s'attache à définir une démarche scientifique nouvelle, une méthodologie qui se construit contre le commentaire érudit, contre l'« appropriation » des auteurs, contre la « prévention » et pour « ses propres lumières » en valorisant l'exercice du jugement individuel. Perrault présente d'emblée son projet comme totalisant : « une entreprise au-dessus de [s]es forces » où il s'agit de

couvrir « tous les Beaux-Arts et toutes les Sciences » et d'en évaluer l'évolution depuis le « degré de perfection » atteint dans « les plus beaux jours de l'Antiquité » jusqu'au « Siècle où nous sommes » (préface). Affirmant préparer le terrain pour les successeurs qui achèveront son entreprise, Perrault entend écrire « une histoire exacte du progrès » qui permettrait notamment de comprendre le rôle du « raisonnement » et celui de l'« expérience » dans l'accès à la « modernité ». Les notions en jeu dans le texte sont constamment essentielles pour comprendre comment s'élabore un nouveau paradigme intellectuel : mémoire, autorité, patrimoine, preuve, imitation, jugement, expérience, méthode, autant de concepts qui font l'objet d'approfondissements et de questionnements à situer dans leur temps et à modéliser dans leur singularité.

L'enjeu est de mettre en évidence le caractère conflictuel de la définition de la modernité à un moment de reconfiguration intellectuelle majeure. Si le règne de Louis XIV a mis en place les conditions d'une institutionnalisation de la vie culturelle, en particulier avec la création des Académies, et s'il a conduit les auteurs et artistes à tirer de la période une forme de bilan, ils ne l'ont pas fait d'une manière consensuelle : la controverse pose clairement les enjeux de la définition d'une littérature et d'une pensée modernes, c'est-à-dire actuelles, à partir de la double expérience esthétique antique et contemporaine. En interrogeant les valeurs littéraires, le conflit qui secoue les milieux culturels et savants en France, au tournant des XVII^{e} et XVIII^{e} siècles, constitue une transition majeure vers une conception renouvelée de la création et de la diffusion du savoir. La médiation numérique est un outil essentiel pour donner à lire et à mesurer l'arborescence discursive que Perrault mobilise en fondant son manifeste polémique sur une énergie intellectuelle hors du commun. L'ambition de l'entreprise est finalement de donner accès à un texte « ancien » en le détachant des approches traditionnelles et des modélisations institutionnelles qui ont tendance à figer cet héritage. Il s'agit de montrer comment notre modernité peut explorer la complexe genèse de l'idée de modernité au cœur d'un texte qui en cristallise les enjeux philosophiques et politiques et dont l'édition numérique offre la possibilité d'éprouver simultanément toutes les dimensions.

Édition numérique, site dédié et réévaluation d'une œuvre. Les écrits de l'abbé Castel de Saint-Pierre

Julia Roger
Maison de la recherche en sciences humaines, USR3486
Université de Caen - Normandie, France

Carole Dornier
Histoire, Territoires & Mémoires, EA7455
Université de Caen - Normandie, France

Cette communication vise à expliquer les choix éditoriaux qui ont présidé à l'édition scientifique numérique des écrits de l'abbé de Saint-Pierre en fonction de la spécificité de ce corpus et des résultats obtenus au regard de ces choix.

On insistera donc sur la fonctionnalité des outils éditoriaux, en particulier à partir de deux niveaux de l'appareillage éditorial co-construit par le chercheur et l'ingénieur : 1) à l'échelle globale du corpus, la conception du site dédié et les outils de présentation des textes ; 2) à l'échelle d'un texte, les variantes et l'annotation XML/TEI.

1) Saint-Pierre, auteur prolixe, ayant rédigé une quantité impressionnante d'opuscules, dont il multipliait les versions, en a fait imprimer une grande partie dans des recueils sans grande cohérence, en fonction des opportunités. Cet ensemble foisonnant, évolutif, aux limites mal déterminées, intéressant avant tout un public savant, se prêtait mal à la notion d'oeuvres complètes et à une édition papier. Les fonds contenant les manuscrits sont dispersés et certains n'ont été ni décrits, ni exploités. Le corpus a surtout fait l'objet de recherches partielles, ce qui a empêché de mettre à jour le caractère cohérent et systématique de la pensée de Saint-Pierre.

À l'échelle globale du corpus, ont donc été privilégiés les outils de présentation suivants : a) la distribution des textes en modules thématiques ; b) l'inventaire des imprimés (ordre chronologique) et des manuscrits (par lieu de conservation) ; c) les ressources documentaires intéressant l'ensemble du corpus (note autobiographique, directives d'édition aux héritiers...) ; d) un index nominum pour mettre à jour en particulier des sources, et les raisons de l'abandon d'un index rerum) ; e) un moteur de recherche plein texte.

Ces différents outils permettent de confirmer le caractère systématique de la pensée de l'auteur appliquée à des objets divers : relations interétatiques, fiscalité, gouvernement des États, religion, politique culturelle, morale. Ils permettent aussi de souligner la récurrence d'influences, de thématiques, de notions et de termes, de les mettre en relation. Ils permettent de dépasser l'écueil de recherches jusqu'ici cloisonnées par discipline(s) universitaire(s), qui ne pouvaient éclairer les liens entre les différents thèmes abordés. La communication passera en revue, captures d'écran à l'appui, ces différents points.

2) À l'échelle d'un texte, on présentera les choix de structuration dans a) la présentation des variantes, b) l'appareil des notes, c) les tables des matières. Ces choix de structuration s'appuient sur certaines règles de l'édition des textes modernes, présentant des états imprimés et manuscrits -- l'autorité de ces derniers devant être évaluée en fonction de leur rapport aux premiers --, en tenant compte des particularités du corpus et des attentes de la communauté concernée. Tout en privilégiant donc l'autorité des imprimés en période moderne, une attention particulière est cependant accordée aux manuscrits corrigés par Saint-Pierre en vue d'une nouvelle édition, lorsque ceux-ci font état d'une évolution conceptuelle significative. Par ailleurs, les spécificités du corpus excluent de présenter de façon exhaustive toutes les variantes textuelles : un relevé exhaustif de

micro-variations orthographiques et stylistiques aurait en effet conduit à une inflation d'informations peu pertinentes, nuisant au soulignement des variations significatives qui mettent en lumière l'évolution du contexte de rédaction, celle de la pensée de l'auteur, son travail de réorganisation et de réutilisation des textes. Cette visée a ainsi dispensé de recourir au module « Critical Apparatus » de la TEI, plutôt adapté à la comparaison de témoins sur des corpus médiévaux constitués par des versions essentiellement manuscrites. Le modèle de données et le schéma TEI finalement retenus ont abouti à une identification des passages significatifs, modifiés par Saint-Pierre au fil du temps, par des éléments typés. Cet encodage, simple, suffit à confronter plusieurs versions desdits passages dans le moteur d'affichage XML du site de l'édition. Pour illustrer ce point, seront présentés trois exemples emblématiques d'empans de texte structurés en XML-TEI, dont on montrera la restitution sur l'interface du site, en soulignant leurs enjeux interprétatifs. Une rapide visualisation de certaines notes explicatives contenant des renvois mettra en évidence l'apport interprétatif des liens établis entre les textes édités, à l'intérieur d'un module thématique, mais aussi entre modules.

En conclusion, on insistera sur la réévaluation à laquelle conduit une édition qui s'efforce de mettre en évidence l'ensemble des relations qui unissent des textes jusqu'alors isolés par leurs modalités d'édition, de conservation ou d'exploitation scientifique, et de faire apparaître une pensée politique et morale très cohérente qui cherche à s'appliquer à des objets divers du fonctionnement social.

Les études zoliennes à l'ère du numérique. Bilan et perspective sur dix ans d'expérience

Olivier Lumbroso
DILTEC, EA 2288
Université Sorbonne-Nouvelle, France

Depuis le début des années 2010, le Centre d'étude sur Zola et le naturalisme, de l'ITEM-CNRS / ENS, dirigé par Alain Pagès et Olivier Lumbroso, a associé les humanités numériques à l'édition et l'interprétation des œuvres de Zola, qu'il s'agisse des corpus manuscrits (dossiers préparatoires de trois cycles), des corpus iconiques (dessins, photographies de Zola), des corpus romanesques (œuvres publiées), enfin des corpus épistolaires (les correspondances générales et les lettres internationales). C'est à travers plusieurs projets de recherche institutionnels que ce tournant numérique a été pris sur l'étendue des dix ans : Archiz (ANR, 2012), Lettres internationales (Labex TransferS, 2017), ScéNa (Translitterae, 2019), CorreZ (Item, 2018).

Le premier projet a abouti à la création d'un portail des archives manuscrites zoliennes, référence et ressources pour de nombreux chercheurs, étudiants et amateurs de l'œuvre de Zola. La finalité a été globalement patrimoniale et archivistique, en visant à rendre accessible, et dans des conditions satisfaisantes, une documentation souvent difficile à consulter en bibliothèque.

Le projet Labex TransferS a souhaité ouvrir la réflexion à une équipe internationale d'une trentaine de membres avec l'idée d'éditer, sur la plateforme EMAN de l'ITEM, l'ensemble des lettres internationales reçues par Émile Zola, notamment durant l'affaire Dreyfus. Ces lettres provenant du monde entier, il fallait une équipe internationale et locale en même temps, pour lire et comprendre, dans leur contexte historique, ces quelques 2000 lettres, aujourd'hui consultables sur un site ouvert. Le projet, dans les années à venir, est de construire l'édition annotée de l'ensemble de la correspondance de Zola (générale, internationale, intime).

Le projet ScéNa (« Scénarios naturalistes ») s'oriente vers la génétique scénarique et vers les études photographiques, toujours avec cette ambition d'articuler données empiriques, construction de livrables numériques et réflexion théorique, littéraire et interdisciplinaire. Ainsi, le projet scéNa mène actuellement deux volets de réflexion sur des corpus différents : les scénarios et canevas de Zola d'un côté, son œuvre photographique de l'autre. Dans les deux cas, un partenariat s'est noué avec la plateforme TACT pour le projet génétique et la MAP (Médiathèque de l'Architecture et du Patrimoine) pour le projet photographique.

La communication visera à développer la problématique transversale suivante : quelles relations les études zoliennes à l'ITEM ont-elles construit avec les instruments numériques sur l'étendue d'une décade ? Selon quelles méthodes et pour quels objectifs ? Une moitié de l'intervention sera consacrée à évoquer les choix, tournants et obstacles au cours du temps, l'autre moitié se focalisera sur le dernier projet : l'édition numérique annotée sur la plateforme TACT de l'université Grenoble Alpes des vingt ébauches des *Rougon-Macquart*, au moyen de la TEI. Pourquoi la TEI représente-t-elle une avancée pour l'interprétation des grands corpus manuscrits ?

Modélisation des textes littéraires. Entre temps limité et désir d'exhaustivité (retour sur deux projets d'éditions numériques)

Sarah Orsini
HiSoMA, UMR 5189
École normale supérieure de Lyon, France

Le tournant numérique a profondément reconfiguré les modalités d'analyse et d'édition des textes. Il a également transformé la façon de représenter le savoir et de le rendre accessible, ce qui induit une mutation importante des pratiques de lecture. Cela ouvre des perspectives d'une grande créativité : il nous appartient de réinventer nos méthodes de travail pour produire de nouvelles formes de savoirs. Cela a également induit la redéfinition de concepts-clés de la littérature tels que le texte, le livre ou le document, en phase avec le développement de la nouvelle philologie (Driscoll, 2010).

De fait, un apport précieux de la nouvelle philologie est l'attention portée à des « faits » sur le document qui dépassent le texte prêt-à-lire et qui sont regroupées en diverses « dimensions » variant en fonction de la lecture qu'on en fait : linguistique, sémantique, paléographique, littéraire, génétique, culturelle, etc. (Sperberg-McQueen, 2009 in Pierazzo, 2014). L'édition du document est donc une modélisation (McCarty, 2005) qui résulte de choix éditoriaux. Or, face à la grande créativité offerte par le tournant numérique, comment réaliser ces choix de façon satisfaisante ? Quelles dimensions privilégier et quel degré de précision viser ? Où s'arrêter pour que le projet soit terminable ?

Cette transformation des pratiques éditoriales, aussi créative soit-elle, n'est pas toujours confortable. En effet, nos méthodes héritées de la Renaissance et du XIX^{e} siècle continuent de faire école : nous sommes toujours formé.e.s aux méthodes des éditions critiques sur support papier, et ces méthodes permettent toujours de réaliser de fréquentes publications. De fait, comme le souligne Elena Pierazzo, si tous les livres suivent plus ou moins la même méthode, en raison des limites physiques de la page, les éditions numériques sont toutes différentes, car elles font toujours l'objet d'expérimentations méthodologiques et ne sont pas limitées en termes d'espace. L'espace presque illimité permet également de multiplier les lectures possibles du document, mettant chacune en lumière une de ses « dimensions ». Face à cette multitude d'opportunités, il faut alors choisir les intentions de l'édition, « ce qu'elle fait », pour aboutir à un résultat.

Aujourd'hui, des méthodes d'édition numérique ont été établies selon des centres d'intérêts principaux, souvent hérités des méthodes philologiques des siècles précédents (éditions philologiques avec apparat, reconnaissance des entités nommées, indexation, analyse linguistique automatisée, alignement de traduction), avec des guides de l'encodage à la visualisation. Mais d'une part chaque projet d'édition associe à sa manière ces méthodes, rendant nécessaire l'établissement d'outils de visualisation sur-mesure. D'autre part dans certains domaines il n'existe pas de protocole fixe, ce qui offre une grande liberté, mais induit aussi d'importants défis techniques. Par exemple, dans les éditions génétiques, malgré une volonté générale d'analyser et de donner à lire le texte dans sa chronologie, l'échelle de l'analyse et les éléments analysés varient en fonction des textes et des projets (Grésillon, 1994, Pierazzo, 2014).

Nous avons mené cette réflexion méthodologique dans le cadre du doctorat, dont une des missions consistait à réaliser un prototype d'édition génétique : un outil de lecture chronologique des brouillons du poème néolatin *Crepereia Tryphaena* de Giovanni Pascoli (1892). L'intérêt d'une telle édition est d'analyser le processus d'écriture, et c'était particulièrement intéressant dans le cas de la composition d'un poème néolatin qui supposait un travail de reconstitution de la langue latine, des contenus et des références intertextuelles. Notre démarche

était exploratoire : il s'agissait de tester le plus grand nombre d'axes de lecture possible à appliquer à un corpus réduit de brouillons :

- encodage en XML-TEI de la chronologie de l'écriture pour chaque brouillon, de l'échelle de la page à l'échelle de la lettre
- typologie des gestes d'écriture (analyse des types de modification et des causes)
- analyse métrique
- parallélisation des éléments sémantiquement proches d'un brouillon à l'autre (échelle du mot, du groupe de mots, de la strophe)
- encodage des références intertextuelles explicites et implicites.

L'encodage chronologique a abouti à un outil de lecture chronologique des brouillons au moyen de boutons permettant d'afficher le texte d'une campagne à l'autre (inspiré du Proust Protoype, André, Pierazzo, 2013, il ne repose toutefois pas sur une transcription intégrée au fac-similé). La typologie des gestes d'écriture a donné lieu à des tableaux statistiques, ainsi qu'à des commentaires au fil de la lecture chronologique. Cela a nécessité l'élaboration de visualisations sur-mesure, qui sont encore aujourd'hui à l'état de prototype et dont l'aspect peut sembler peu attirant puisqu'il a été réalisé sans aucune aide en développement et design. Or, l'apport de ces deux disciplines permettrait de proposer des solutions pour représenter le suivi d'un passage sélectionné d'un brouillon à l'autre (vision synoptique), ou représenter le travail métrique (lacunes comblées progressivement au moyen de tests plus ou moins satisfaisants), autant d'informations qui ne sont pour l'instant interrogeables qu'en XPath ou XQuery pour étayer le commentaire génétique.

Outre les difficultés techniques supposées par ces expérimentations, la principale limite rencontrée est celle du temps. En effet, un tel encodage n'est pas automatisable. Il est seulement possible de réaliser grossièrement la parallélisation des éléments proches avec un traitement automatique des langues (lemmatisation, co-occurrences et requêtes XQuery). Ainsi, pour traiter un corpus entier avec une telle précision, il faudrait être très nombreux ou y consacrer des dizaines d'années. Mais est-il souhaitable d'étendre une telle démarche à l'ensemble du corpus latin de l'auteur, voire du corpus italien ? Faut-il préférer une transcription de davantage de textes mais d'une moindre précision ? Est-ce que le lecteur en apprendra davantage sur l'écriture pascolienne si l'on traite tout le corpus de cette façon que si l'on choisit quelques dossiers représentatifs ? Et jusqu'où peut-on assister le lecteur dans sa confrontation avec le brouillon ? Pour comprendre profondément le geste d'écriture, une fréquentation assidue des brouillons reste nécessaire, et n'est que facilitée par l'édition. De fait, certains aspects du processus d'écriture restent inexplicables, soit parce que les raisons d'un geste ne sont pas explicites, soit parce que les mots nous manquent pour décrire ce qui relève d'une expérience sensible (c'est pourquoi il est bien plus facile d'expliciter les mises au point des derniers brouillons que les premières étapes, notamment la création de nouveaux contenus). L'autre limite rencontrée est celle de la subjectivité de telles analyses : même dans le cas où le déchiffrement serait aisé, la datation relative des éléments du brouillon relève de l'interprétation de l'éditeur. Tout en permettant au lecteur d'accéder à une transcription plus facile à lire et dynamique, elle déforme nécessairement le document en le modélisant.

À ces enjeux herméneutiques s'ajoutent les problématiques de la visibilité : quelle visibilité peut avoir un prototype (expérimentation d'une exhaustivité dans la précision de l'encodage) par rapport à l'édition génétique d'un corpus complet (exhaustivité dans la transcription des textes) ? Comment situer une telle production dans la foison des projets numériques ?

Nous pourrions penser que de telles problématiques sont caractéristiques de la génétique textuelle. Cependant, elles concernent également des projets d'édition de textes classiques. Le projet AgroCCol pour lequel nous travaillons désormais consiste à établir une édition numérique de type encyclopédique des textes agronomiques latins et grecs. Pour l'instant, pour des raisons de temps, nous avons sélectionné des textes sur la culture des céréales et des légumineuses, afin de pouvoir réaliser un encodage correspondant à une multiplicité d'axes de lecture :

- indexation de termes techniques pour constituer un dictionnaire sur l'agriculture antique
- constitution d'un thésaurus des noms de plantes et d'outils

- identification des entités nommées
- encodage thématique.

Certes, davantage de tâches sont automatisables entièrement (notamment le relevé des occurrences du dictionnaire technique) ou partiellement (analyse thématique), nous travaillons en équipe et bénéficions d'aides techniques importantes. Cependant, le choix d'une analyse qualitative (même si assistée par ordinateur) nous a poussé à restreindre notre corpus, et l'échéance du projet nous contraint parfois à arrêter la précision de nos analyses thématiques à ce qui a été établi. Cela pose à nouveau des problèmes de subjectivité (sur quels critères les textes ont-ils été choisis ? sur quels critères des thèmes sont-ils associés à des mots du texte ? l'encodage thématique est-il constant d'un texte à l'autre ?), de visibilité (quelle sera la place d'un tel corpus par rapport à des corpus anciens plus larges sans élucidation sémantique ou littéraire tels que Perseus ? En tant que lecteur, comment trouver une telle publication ?). Se pose également la question du rapport au lecteur : cette sélection satisfera-t-elle sa curiosité, même si elle ne traite pas de l'ensemble de l'agriculture ?

Ainsi, la facilitation de l'accès aux documents, aux textes ou aux données de la recherche augmente les possibilités de traiter des volumes plus grands et d'aspirer à une exhaustivité en matière de corpus (œuvre intégrale, mouvement, associations), tandis que la TEI nous pousse à viser une exhaustivité en termes de précision d'encodage. Mais le souhait de continuer à mener des analyses de précision correspond à un temps long de la recherche et ne peut pas être subordonné uniquement à la demande ou à des échéances courtes. Dans ce contexte, la démarche des données ouvertes et réutilisables est tout à fait stimulante. Cependant, elle ne peut aboutir que si l'on organise des synergies entre les publications numériques, soit en préférant travailler à la suite d'un projet préexistant, soit en construisant ensemble des répartitions entre divers projets, sans pour autant les uniformiser. Des projets de plateformes telles que Biblissima, ou Fonte Gaia pour les textes italiens pourraient constituer une première étape pour sortir de cette tendance à l'isolement des projets numériques les uns par rapport aux autres.

Références

TEI Consortium (ed.) *TEI P5 : Guidelines for Electronic Text Encoding and Interchange*. [Version 4.1.0] [Last updated on 19-08-2020]. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html [Consulté le 20/09/19].

André Julie et Elena Pierazzo, « Le codage en TEI des brouillons de Proust : vers l'édition numérique », *Genesis. Manuscrits -- Recherche -- Invention*, no 36, Sigales, 2013, p. 155-161. DOI : https://doi-org.acces.bibliothequediderot.fr/10.4000/genesis.1159 [consulté le 16/11/2020]

Casenave, Joana, « Le positionnement éditorial dans l'édition critique numérique », *Digital Studies/Le champ numérique*, vol. 9, no 1, 2019. DOI : http://doi.org/10.16995/dscn.348 [consulté le 15 janvier 2020].

Driscoll Matthew J., « The Words on the Page : Thougts on Philology, Old and New », dans Judy Quinn et Emily Lethbridge (éd.), *Creating the Medieval Saga : Versions, Variability and Editorial Interpretations of Old Norse Saga Literature*, Denmark, University Press of Southern Denmark, 2010, p. 85-102.

Grésillon Almuth, *Éléments de critique génétique : lire les manuscrits modernes*, Paris, Presses Universitaires de France [1994], 2016.

McCarty Willard, *Humanities computing*, Basingstoke, Palgrave Macmillan, 2005.

Orsini Sarah, *Les Carmina de Giovanni Pascoli : édition traduite et commentée d'une sélection de poèmes latins et édition numérique d'une sélection de brouillons*, Université Lyon 2, Università Roma Tre, 2019.

Crepereia Tryphaena :
https://github.com/SarahOrsini/Edition_genetique_Pascoli_CrepereiaTryphaena/tree/master/Lecture%20chronologique, [consulté le 16/11/2020]
Pierazzo Elena, *Digital Scholarly Editing: Theories, Models and Methods*, sans lieu, 2014.
Sperberg-McQueen C. Michael, « How to Teach your Edition How to Swim », *Literary and Linguistic Computing*, vol. 24/1, 2009, p. 27-52.

Une ontologie pour Henri et ses amis

Pierre Willaime
Archives Henri-Poincaré, UMR7117
Université de Strasbourg, France

Cette communication se propose de présenter méthodologiquement l'édition numérique d'un corpus d'auteur, celui de la correspondance d'Henri Poincaré (1854-1912, approx. 2100 lettres[1]). Le traitement et l'analyse du corpus posent des questions plus générales, propres aux technologies utilisées (celles du Web sémantique) et à leur applicabilité en humanités numériques. Ce projet est porté collectivement par les Archives Henri-Poincaré (UMR7117 - CNRS, Universités de Lorraine et de Strasbourg).

Le Web sémantique peut être vu comme particulièrement adapté à la description d'un corpus dans le cadre d'un projet d'humanités numériques. En effet, cette extension du Web permet de modéliser les entités, relations, concepts, ou autres métadonnées pouvant décrire les documents d'archives, les acteurs historiques en question et leur contexte. Cette structuration, permise par les technologies standardisées (W3C) constituant le Web sémantique (Berners-Lee et al, 2001) telles que RDF (Resource Description Framework) et OWL (Web Ontology Language), évite l'écueil courant des bases de données relationnelles, adaptées à un projet mais peu interopérables. Plus encore, la « sémantisation » d'un corpus ouvre la voie à son traitement automatique par des méthodes computationnelles.

Travail sur le long cours (depuis 1999), l'édition de ce corpus a été portée sous Omeka S en 2018. Ce CMS a été privilégié à la fois pour sa facilité d'édition, qui permet à tout membre du projet d'avoir un pouvoir d'action, et pour ses fonctionnalités liées au web sémantique.

Omeka S permet facilement de décrire chaque contenu à l'aide de schémas de métadonnées standardisés (Dublin Core dans sa version « DCMI Metadata Terms », Bibliographical ontology, BIO, Relationship, etc.). Cet effort en vue de l'interopérabilité se heurte parfois aux objectifs de recherche, nécessitant souvent des propriétés ad hoc. Les ontologies existantes n'apportent pas l'expressivité d'un encodage en XML-TEI par exemple. Pour ne pas choisir entre d'un côté l'interopérabilité et le respect des standards, et de l'autre une description scientifique fine et adaptée au corpus, nous avons développé une ontologie spécialisée pour l'histoire des sciences et l'édition de correspondances scientifiques. Cette ontologie, nommé AHPo, s'aligne sur les schémas de métadonnées préexistants lorsque cela est possible.

Cette présentation présentera l'ontologie AHPo à l'aide de schémas permettant de comprendre sa structuration. Cette ontologie, et le choix du Web sémantique en général, permettent la mise en place d'outils pour la recherche. Nous en détaillerons deux en présentant les capacités d'interrogation permises par le langage SPARQL et la visualisation du réseau de personnes gravitant autour d'Henri Poincaré. Le langage SPARQL permet de formuler des requêtes complexes pour exploiter les liens entre les éléments d'un graphe de données. Cela ouvre de nouvelles possibilités pour le chercheur en termes de recoupement et de fouille fine du corpus. L'outil de visualisation du réseau de personnes permet de son côté une représentation dynamique et temporelle du contexte social d'écriture des lettres. Il est basé sur l'ontologie et sa description fine des relations entre entités.

De manière plus générale, cette communication souhaite mettre en avant les possibilités techniques d'exploitation de corpus permissent par le Web sémantique. Si elles ne permettent pas d'atteindre une granularité aussi fine qu'un balisage XML-TEI détaillé, les technologies sémantiques mettent cependant l'accent sur le corpus vu comme un ensemble en décrivant

[1] henripoincare.fr

précisément les relations entre entités et en allant plus loin qu'une simple édition numérique. Les deux technologies peuvent conduire à des visions complémentaires d'un même corpus. Il serait intéressant de coupler ces usages, présentés souvent comme deux approches incompatibles et comme deux directions pour un projet d'humanités numériques.

Corpus romanesque et lexicométrie

Ouafae Benzina
Faculté des Lettres et des sciences Humaines
Université Moulay Ismail-Meknès, Maroc

Nous nous proposons dans cette contribution d'étudier le vocabulaire du corpus romanesque de Guy de Maupassant : *Une vie*, *Bel-Ami*, *Mont-Oriol*, *Pierre et Jean*, *Notre cœur* et *Fort comme la mort*. Ce corpus, nous allons le décortiquer à travers des moyens scientifiques, à savoir la statistique et l'informatique. Comment l'ordinateur peut-il ouvrir de nouvelles perspectives à la littérature et aux études linguistiques ?

Parmi les outils informatiques qui traitent des corpus littéraires, nous avons opté pour le logiciel Hyperbase (version 7.1 pour Windows). Un logiciel de statistique lexicale, élaboré par Etienne Brunet en 1989 et développé environ chaque année. Il en est actuellement à la version 10, disponible depuis 2018.

Le traitement statistique des données avec le logiciel Hyperbase permet d'étudier la structure du vocabulaire d'un corpus, étude qui s'intéresse en général à la distribution des fréquences, à l'étude des hapax, à la richesse lexicale, ainsi qu'à l'accroissement lexical. En plus de l'étude de la structure d'un corpus, Hyperbase permet, en effet, de déduire des conclusions également au niveau du contenu lexical en étudiant la distance lexicale, le vocabulaire spécifique ainsi que l'évolution d'un même auteur au cours de la période pendant laquelle il a produit son œuvre.

À cet égard, à l'aide de ce logiciel, nous avons constitué une base que nous avons nommée MAUPASSA.EXE à partir des six romans de l'auteur normand.

Les textes ont été hébergés par le Laboratoire d'Informatique de Besançon, par la suite, ils étaient placés, gracieusement, dans une bibliothèque numérique « Athéna », par Thierry Selva, sur son site « Maupassant par les textes » (http://maupassant.free.fr/).

Le travail se fera donc sur deux dimensions : l'une compte N, nombre de mots du texte (occurrences) ; l'autre dénombre V, vocabulaire du texte (vocables). Dans notre corpus, nous avons relevé une étendue N (corpus) de 549056 occurrences et 22932 vocables.

À cet effet, notre travail d'analyse statistique s'articulera en deux volets. Dans le premier, nous aurons à démonter ce qui particularise les romans de Guy de Maupassant par l'observation de son vocabulaire spécifique. Ce dernier est déterminé avec le logiciel Hyperbase en relevant le vocabulaire en excédent et le vocabulaire déficitaire dans le corpus. Le logiciel fait en même temps une comparaison avec le Trésor de la langue française en calculant l'écart réduit de chaque forme dans chacune des parties du corpus. Si le vocable en question est significativement sur-employé par notre auteur, nous déduisons que l'écrivain éprouve une certaine attirance pour ce vocable. A l'inverse, s'il est significativement sous-utilisé, par rapport à l'usage qu'en font les autres, l'auteur éprouve une certaine répulsion pour ce mot qu'il évite ou qu'il oublie.

Dans Hyperbase, les données du TLF sont insérées en tant que norme et servent de base de calcul en indiquant la différence entre notre corpus et celle de Frantext. Il s'agit donc de voir les mots préférés de Maupassant et la signification qu'il leur prête.

Dans notre étude du vocabulaire spécifique de Maupassant, nous nous intéresserons au vocabulaire positif et négatif. Par vocabulaire positif/négatif, nous entendons les vocables particulièrement sur-employés et sous-employés qui vont nous donner une idée des thèmes traités ou négligés. Autrement dit, le vocabulaire positif représente le vocabulaire excédentaire qui, après une comparaison avec le TLF (XIX^{e} et XX^{e} siècle), a un écart réduit positif. Quant au vocabulaire négatif ou déficitaire, il désigne les formes qui n'apparaissent pas beaucoup dans le corpus et dont la présence par rapport au TLF est insignifiante. Ces formes ont un écart réduit négatif.

Dans ce sens, la lecture des listes des mots présentées, par le logiciel Hyperbase pour chaque texte, nous amène à constituer les champs associatifs qui nous permettrons d'interpréter ce vocabulaire.

Une autre façon d'approcher globalement le vocabulaire de notre corpus : c'est d'observer le vocabulaire le plus caractéristique dans chacun des textes qui le constituent, en prenant pour référence l'ensemble de ce corpus. Le vocabulaire spécifique est déterminé avec le logiciel Hyperbase par le calcul de l'écart réduit de chaque forme dans chacune des parties du corpus.

Dans le deuxième volet de notre étude, il sera question de l'étude de l'évolution du vocabulaire de notre corpus qui interroge l'ensemble du vocabulaire et mesure l'évolution de chaque terme, par le coefficient de corrélation de Bravais-Person. Avec ce calcul réalisé par le logiciel Hyperbase, nous saurons quels sont les mots que Guy de Maupassant emploie de plus en plus et ceux qu'il abandonne progressivement dans son œuvre romanesque.

Ce coefficient examine donc une éventuelle relation entre une distribution et la chronologie. Il évolue entre deux limites : +1 (pour une progression) et -1 (pour une régression), comme l'indique Brunet : « la valeur du coefficient est négative ou positive selon que le mot est en régression ou en progression » (Brunet : 1998).

Le calcul de corrélation nous permet donc de suivre, globalement, la trajectoire des mots dans les différents textes du corpus. Le tri des résultats permet de constituer la liste (décroissante) des mots en progression et la liste (croissante) des mots en régression. Cet examen du contenu lexical nous permettra donc d'enrichir notre connaissance de la spécificité stylistique des textes de notre corpus et de préciser l'évolution de l'écriture de Guy de Maupassant.

À cet égard, nous avons remarqué qu'il y a eu un renouvellement lexical chez Maupassant. L'auteur passe du lexique des parties du corps, des verbes d'action et de l'argent (termes concrets) au vocabulaire de la beauté, de l'art, des salons des sentiments et des préoccupations morales (termes abstraits).

En effet, le thème de l'argent, très dominant dans les quatre premiers romans de Maupassant, disparaît complètement dans ses deux derniers romans, *Fort comme la mort* et *Notre cœur*. À partir de *Pierre et Jean*, le romancier observateur devient romancier psychologue.

En supprimant tout problème matériel à ses personnages et en les faisant évoluer dans des milieux favorisés, Maupassant estime mieux s'attarder sur leur personnalité. Dans *Fort comme la mort*, à une seule exception près, on ne trouvera aucune somme d'argent numériquement définie ; il en va de même dans *Notre cœur*.

Enfin, nous verrons dans quelle mesure les résultats obtenus sont en adéquation avec les analyses stylistiques classiques sur l'œuvre romanesque de l'auteur normand.

Références

1- Œuvre de Guy de Maupassant

Bel-Ami, Paris, Pocket, 2006.
Fort comme la mort, Paris, Albin Michel, 1983.
Mont-Oriol, Paris, Gallimard, 2002.
Notre cœur, Paris, Gallimard, 1993.
Pierre et Jean, Paris, Albin Michel, 1999.
Une vie, Paris, Presses-Pocket, 1977.

2- Ouvrages et articles sur Maupassant et son œuvre

Amis de Flaubert et Maupassant, *Maupassant 2000, Bulletin Flaubert-Maupassant* no. 9, Dieppe, 2001.

Benamrhar, Abdelkrim, « Le miroir dans les romans de Guy de Maupassant », in *Langues et littératures*, vol. XI, Rabat, 1993, p. 125-138.

Besnard-Coursodon, Micheline, *Étude thématique et structurale de l'œuvre de Maupassant : le piège*, Paris, Nizet, 1973.
Bonnefis, Phillipe, *Comme Maupassant*, Lille, Presses Universitaires de Lille, 1981.
Bury, Marianne, *La Poétique de Maupassant*, Paris, SEDES, 1994.
Bury, Marianne, *Maupassant*, Paris, Nathan, 1991.
Bury, Marianne, *Une Vie de Guy de Maupassant*, Paris, Gallimard, 1995.
Cleret, Anne-Marie et Réauté, Brigitte, *Bel-Ami de Maupassant*, Paris, Hachette, 1999.
Delaisement, Gérard, *Maupassant Journaliste et Chroniqueur*, Paris, Albin Michel, 1956.
Dizol, Jean-Marie, *Guy de Maupassant*, Toulouse, Milan, 1997.
Fonyi, Antonia, *Maupassant 1993*, Paris, Kimé, 1993.
Frebourg, Olivier, *Maupassant, le clandestin*, Paris, Mercure de France, 2000.
Giacchetti, Claudine, « Les hauts et les bas : la conquête de l'espace dans *Bel-Ami* de Maupassant », in *Revue romane*, XXVI, 2, 1991, p. 219-229.
Giacchetti, Claudine, *Maupassant, espaces du roman*, Paris, Droz, 1993.
Malrieu, Joël, *Bel-Ami de Guy de Maupassant*, Paris, Gallimard, 2002.
Morand, Paul, *Vie de Guy de Maupassant*, Paris, Flammarion, 1942.
Rochefort-Guillouet, Sophie, *Étude sur Maupassant et le roman*, Paris, Ellipses, 1999.
Salem, Jean, « Le bestiaire imaginaire de Guy de Maupassant », in *Maupassant et l'écriture, Actes du colloque de Fécamp 21-22-23 mai 1993*, Paris, Nathan, 1993, pp.129-138.
Salem, Jean, *La Philosophie de Maupassant*, Paris, Ellipse, 2000.
Santelli, Claude, *Mon Ami Maupassant*, Paris, Éditions 1, 1998.
Satiat, Nadine, *Maupassant*, Paris, Flammarion, 2003.
Savinio, Alberto, *Maupassant et l'« Autre »*, Paris, Gallimard, 1977.
Tassart, François, *Nouveaux souvenirs intimes sur Guy de Maupassant (inédit)*, Paris, Nizet, 1962.
Tolstoï, Léon, *Guy de Maupassant*, Montpellier, L'ANABASE, 1995.
Trevor, A., Le V. Harris, *Maupassant et Fort comme la mort : Le roman contrefait*, Paris, Nizet, 1991.
Vial, André, *Guy de Maupassant et l'art du roman*, Paris, Nizet, 1954.

3- Ouvrages de statistique lexicale

Benzecri, Jean-Paul, *L'analyse des données 1 : La Taxinomie*, Paris, Dunod, 1976.
Benzecri, Jean-Paul, *L'analyse des données 2 : L'analyse des correspondances*, Paris, Dunod, 1976.
Bernet, Charles, *Le Vocabulaire des tragédies de Jean Racine. Analyse statistique*, Genève-Paris, Slatkine-Champion,1983.
Bernet, Charles, « Faits lexicaux. Richesse du vocabulaire. Résultats », in *Études sur la richesse et la structure lexicales*, Paris-Genève, Champion-Slatkine, 1988, pp. 1-11.
Bernard, Michel, *Introduction aux études littéraires assistées par ordinateur*, Paris, PUF, 1999.
Bernard, Michel, « Rêvons un peu... Essai de prospective sur les études littéraires assistées par ordinateur » in *Mesures et démesure dans les lettres françaises au XX^e siècle, Hommage à Henri Béhar*, Paris, Honoré Champion, 2007, p. 359-369.
Brunet, Étienne, *Le Vocabulaire français de 1789 à nos jours*, Genève-Paris, Slatkine-Champion, vol. I, 1981.
Brunet, Étienne, *Le Vocabulaire de Proust I, étude quantitative*, Genève, Slatkine, 1983.
Brunet, Étienne, « La structure lexicale dans l'œuvre de Hugo », in *Études sur la richesse et la structure lexicales*, Paris-Genève, Champion-Slatkine, 1988, p. 23-42.
Brunet, Étienne, *Le Vocabulaire de Victor Hugo*, Genève - Paris, Slatkine - Champion, vol. I, 1988.
Brunet, Étienne, *Compte d'auteurs*, Paris, Honoré Champion, 2009.
Évrard, Étienne et Mellet Sylvie, « Les Méthodes quantitatives en langues anciennes », in *Lalies 18, Actes des sessions linguistique et littérature*, Paris, Presses de l'École Normale Supérieure, 1998, p. 111-155.
Gicquel, Bernard, *Stylistique littéraire et Informatique*, (publ. par le) Centre d'études et de recherches sur les textes électroniques littéraires, Arras, Artois presses université, 1999.

Hefied, Ali, *Statistique linguistique : Aspects stylostatistiques du vocabulaire dans quinze voyages extraordinaires de Jules Verne*, Thèse de doctorat d'État, Fès, 1999.

Juilliard, Michel, « Du bon choix d'un corpus et de son bon usage », in *Mots chiffrés et déchiffrés. Mélanges offerts à Étienne Brunet*, textes rassemblés par Sylvie Mellet et Marcel Vuillaume, Paris, Honoré Champion, 1998, p. 139-116.

Kastberg Sjöblom, Margareta, *L'Écriture de J.M.G. Le Clézio : Des mots aux thèmes*, Paris, H. Champion. 2006.

Lebart, L. et Salem, A., *Statistique textuelle*, Paris, Dunod, 1994.

Lenoble, Michel, « Statistique lexicale et critique littéraire le mariage impossible ? » in *Méthodes quantitatives et informatique dans l'étude des textes*, Genève-Paris, Slatkine-Champion, 1986, p. 565-574.

Luong, Xuan Nhuam et Novi, Michel, « Représentations arborées de données textuelles », in *Méthodes quantitatives et informatique dans l'étude des textes*, Genève-Paris, Slatkine-Champion, 1986, p. 575-586.

Magri, Véronique, « Stylistique générique et statistique pour une poétique du récit de voyage », in *JADT' 06, Volume II, Actes des 8es journées internationales d'analyse statistique des données textuelles*, Besançon, Presses Universitaires de Franche-Comté, 2006, p. 655-666.

Magri, Véronique, *Le Voyage à pas comptés*, Paris, Honoré Champion, 2009.

Mayaffre, Damon, *Le Poids des mots*, Paris, Honoré Champion, 2000.

Valorisation du fonds Bourget de l'ICP. Une revalorisation d'un auteur grâce à son exposition numérique ?

Dominique Ancelet-Netter et **Guillaume Boyer**
UR « Religion, culture et société »
Institut Catholique de Paris (ICP), France

L'Institut catholique de Paris a reçu en don un fonds exceptionnel de la part des héritiers de l'académicien Paul Bourget avec la bibliothèque personnelle et les écrits du for privé de l'écrivain mais avec une interdiction de publication par l'auteur. Cette alliance de ces deux ensembles est rarissime et représente un corpus d'auteur unique et une masse considérable de données. Deux questions se posent alors. Comment intégrer la bibliothèque dans une édition génétique de son œuvre représentée par ses brouillons et journaux (en croisant les envois et les traces de lectures dans les ouvrages avec les notes de lectures dans les journaux, agendas et autres notes préparatoires tout en les associant aux citations présentes dans son œuvre...) ? Comment organiser et structurer au mieux le *data mining* (fouille de données) sur ce grand gisement que constituent les écrits intimes du fonds Bourget avec par exemple la constitution d'un index des noms propres de personnes, de lieux ou d'institutions et des titres d'œuvres pour une future édition numérique ?

La bibliothèque de Fels de l'Institut catholique de Paris conserve les journaux intimes de Paul Bourget et de son épouse, sa bibliothèque, une partie de sa correspondance passive, et quelques manuscrits d'œuvres romanesques, journalistiques et dramatiques : il s'agit du fonds sur Bourget le plus considérable en France, avec celui de la Bibliothèque nationale de France. Ce fonds comprend onze volumes de journal, vingt-huit agendas et huit carnets de notes, tenus par Bourget de 1870 à 1933[1], et onze volumes de journal et vingt-et-un agendas tenus par sa femme Minnie (née Julia Louise Amélie David) de 1890 à 1926[2]. L'état du fonds présente cependant un angle mort. Sans explication, les années en 8 sont manquantes pour les agendas de Bourget entre 1906 et 1933, sauf l'année 1928, retrouvée récemment. La bibliothèque de Fels ne conserve en outre aucun cahier intime entre 1871 et 1878, interdisant l'accès à la période des premiers dîners ou cercles littéraires de l'académicien et des informations précieuses sur cette période d'histoire littéraire[3].

Tout ce que la fin-de-siècle compte d'écrivains et d'œuvres se retrouve dans la bibliothèque de Bourget. Elle contient plus de quatre-mille-trois cents volumes avec plus de mille envois recensés. Un troisième ensemble est plus hétéroclite. Il intègre pêle-mêle les lettres envoyées à Bourget par Huysmans, Ferdinand Brunetière, Edith Wharton et Juliette Adam[4], des factures et livres de comptes et les papiers libres glanés dans les exemplaires truffés de sa bibliothèque comme les divers brouillons et notes préparatoires[5]. Un rare croquis d'audience de Jean-Louis Forain en

1 ICP, bibliothèque de Fels, Ms français 664/1 à 664/39.

2 ICP, bibliothèque de Fels, Ms français 665/1 à 664/32

3 Michel Mansuy signale dans sa biographie des documents qui ne figurent pas dans le fonds, notamment un cahier daté de juillet 1884 à mars 1887 et des notes de voyages des années 1886-1889. À noter aussi des caviardages et des découpes dans les journaux sans qu'il soit possible de déterminer s'ils sont le fait de Bourget ou de ses ayants droit.

4 ICP, bibliothèque de Fels, Ms français 664/40. Quelques lettres ont fait l'objet d'une publication : Bénédicte Coste, « Two Unpublished Letters from Walter Pater to Paul Bourget », *The Pater Newsletter*, n° 61/62, printemps/automne 2012, p. 4-20 ; Virginia Ricard, « An Unknown Letter from Edith Wharton to Minnie Bourget », *Edith Wharton Review*, vol. 33, n° 2, 2017, p. 351-360 ; Pierre Brunel, André Guyaux, *Huysmans, Cahier de l'Herne*, 1985.

5 S'y trouve notamment une copie manuscrite avec corrections et parties autographes de la pièce *Le Tribun* ainsi que des notes diverses ayant servi à la rédaction du roman *La Duchesse bleue* et des nouvelles « Un scrupule » et « L'Apostat ». https://bibliotheques.icp.fr/rechercher/collections-patrimoniales/fonds-particuliers/fonds-paul-bourget

juillet 1914 lors du procès d'Henriette Caillaux où Bourget fut cité en témoin, a été aussi trouvé lors du récolement réalisé de la bibliothèque de l'auteur[6].

La valorisation de ce fonds est menée conjointement par une enseignante-chercheuse de la Faculté des Lettres de l'ICP et par le responsable du fonds patrimonial et ancien de la bibliothèque de Fels. Il s'inscrit dans le cadre du projet Universitas de l'UR « Religion, culture et société » de l'ICP, dans l'axe prévoyant une collaboration étroite entre bibliothèques et chercheurs à l'aune des humanités numériques. Le premier objectif a été patrimonial : conservation et sécurisation du fonds des journaux intimes et des agendas de l'académicien par une campagne de numérisation au total de près de 7 000 pages. Le deuxième objectif aurait dû être celui d'une édition critique génétique avec une transcription de ces textes. Mais deux obstacles se sont levés, l'un financier, l'autre juridique et éthique. Dans son testament, en date du 25 janvier 1935, Bourget a refusé toute publication de ses écrits intimes et de sa correspondance après sa mort. L'écrivain a dénoncé la « littérature d'autobiographie » et a même employé l'expression, à propos du *Journal* d'Amiel, de « maladie du journal intime[7] ». Même s'il peut apparaître paradoxal pour Bourget d'avoir alors légué à la postérité ses journaux intimes, il a été décidé de respecter la volonté de l'auteur, même en l'absence d'ayants droit identifiés, et d'employer des stratégies d'évitement de l'interdiction de la publication des écrits du for privé. Un autre ensemble du fonds Bourget pouvait être rendu public : sa bibliothèque et ses envois. Le récolement confié entre novembre 2019 et février 2020 à une stagiaire de l'École des bibliothécaires documentalistes (école associée à l'ICP) a permis de recenser l'ensemble des envois de la bibliothèque de Bourget dans ce double objectif patrimonial et scientifique. Les notices bibliographiques présentes dans le SUDOC seront ainsi enrichies de la transcription de ces envois. Une exposition virtuelle présente depuis avril 2020 quelques-uns de ces envois à Bourget, accompagnées de notices permettant une meilleure compréhension des liens de sociabilité littéraire de l'auteur[8]. Cette exposition en ligne, appelée à s'enrichir, est la première étape dans l'élaboration d'un site consacré à l'auteur dans le cadre de la bibliothèque numérique de l'ICP. Elle sera suivie par la mise en ligne, grâce un logiciel de littérométrie, de la cartographie de la galaxie des relations littéraires de Bourget grâce à ces envois identifiés et transcrits. Financés exclusivement sur fonds privés issus du mécénat, le projet « Fonds Bourget » de l'ICP se poursuivra aussi en collaboration avec l'EBD, par une mission prochaine confiée à une stagiaire afin d'établir, dans un projet documentaire, la bibliographie de et sur Bourget la plus exhaustive possible qui sera mise en ligne (avec des liens hypertextes quand ils seront possibles) sur ce futur site consacré à l'auteur. Un stagiaire en Master humanités numériques est en cours de recrutement pour une co-construction de ce site dédié avec un prestataire extérieur. Si la publication intégrale des manuscrits intimes n'est pas envisageable, alors le contournement, pour éviter les angles morts, consistera en une première description matérielle et technique des journaux et agendas intimes pour un versement dans Calames (Catalogue en ligne des archives et des manuscrits de l'enseignement supérieur), puis dans le futur site Bourget, de plus, en indexant les manuscrits intimes par mots-clefs afin de proposer des matériaux-sources pour les chercheurs. Les champs sont vastes : l'histoire littéraire (Bourget est un homme de réseaux et de salons littéraires), histoire des droites françaises (proximité avec Barrès et l'Action française), histoire de l'éducation, de la psychologie et de la pédagogie, mais aussi de la géographie (nombreuses descriptions de voyages, Europe, USA, Liban dans les manuscrits intimes) ou l'histoire du genre avec les journaux croisés Minnie David / Paul Bourget. Une consultation encadrée des manuscrits numérisés sera permise par la base de collationnement descriptive mise en ligne. Face à la quantité, la nature et la

---

6 Il figure Berthe Gueydan, première épouse de Joseph Caillaux, appelée elle aussi à la barre lors du procès. L'écrivain a d'ailleurs gardé par-devers lui, comme une relique, le numéro du *Figaro* daté du 16 mars 1914, jour de l'assassinat du directeur Gaston Calmette. Cette identification est à mettre au crédit de Romain Larguier, chef magasinier à la bibliothèque de Fels.

7 Paul Bourget, « La maladie du journal intime », *Nouvelles pages de critique et de doctrine*, t. II, Paris, Plon, 1922, p. 15-26.

8 Cette exposition virtuelle est consultable dans la bibliothèque numérique de l'ICP https://bibliotheque-numerique.icp.fr/ via l'onglet « Expositions ».

structure des données utilisées, le défi va être dans la détection et la sélection des mots-clefs pour chacun des manuscrits afin de se situer ultérieurement dans une co-construction collaborative en éditions critiques et en partenariat notamment pour des doctorants de tous les champs disciplinaires des sciences humaines et sociales. L'édition électronique de la correspondance passive pourra intervenir ultérieurement pour parfaire la mise à disposition du fonds Bourget de la bibliothèque de Fels de l'Institut catholique de Paris.

Avec pragmatisme et rigueur, et en collaboration avec des stagiaires étudiants en humanités numériques, nous souhaitons, dans le double objectif d'une conservation patrimoniale et d'une valorisation scientifique, permettre l'accessibilité à un corpus rare et riche en le faisant passer de l'ombre des sous-sols d'une bibliothèque universitaire à la lumière de l'exposition numérique pour un auteur qui s'est raréfié dans le champ des études littéraires. L'académicien Paul Bourget, en tant que poète, essayiste, critique, romancier, journaliste, dramaturge au carrefour des XIXe et XXe siècles, mérite mieux que le mépris et l'oubli dans lesquels il a été si longtemps confiné.

Usages de la textométrie en histoire de la philosophie

Anatole Lucet
TRIANGLE, UMR 5206
École Normale Supérieure de Lyon, France

Un corpus de trois millions de mots : plusieurs centaines d'articles, quelques ouvrages théoriques, plusieurs romans et nouvelles, de nombreuses traductions, une imposante correspondance. C'est l'œuvre foisonnante laissée par le philosophe allemand Gustav Landauer (1870-1919) en trente années d'activité de plume. En quoi les outils numériques peuvent-ils nous aider à appréhender un tel corpus ?

Marqué par les mystiques médiévaux, les auteurs romantiques allemands et le scepticisme linguistique, le philosophe s'efforça de trouver une parole qui parvienne à dire le monde - et le transformer - sans l'enfermer dans un système de lois rigides, y compris linguistiques. Ce penseur de la communauté fut un adversaire résolu du socialisme scientifique prôné par les marxistes de son temps ; c'est à ce titre qu'il élabora une pensée qui, par sa forme comme par son contenu, se dérobait face aux tentatives de systématisation. Privilégiant une parole vivante, mouvante et non systématique, il refusa catégoriquement de donner une formalisation scientifique à sa pensée. La conséquence de ce refus : un manque criant de définitions, d'énoncés et de propositions stables, ainsi qu'un sentiment de confusion pour les lecteurs accoutumés à une exposition plus géométrique des thèses philosophiques.

En effet, l'auteur qui tournait en dérision les analyses scientifiques de la liberté humaine n'a pas laissé apparentes les chevilles qui structuraient sa pensée, préférant l'effectivité de la suggestion au formalisme de l'argumentation. Pour retrouver les définitions qui font la cohérence de sa pensée, il est donc nécessaire de s'immerger dans la profusion de ses petits écrits : près d'un millier d'articles, deux fois plus de lettres. Il s'agit en effet de retrouver, disséminés sur plus de trente ans d'écriture, le sens des concepts-clefs de la pensée d'un auteur. C'est à ce stade que le recours à l'analyse textométrique permet non seulement de retracer l'usage d'un concept à travers les différentes périodes de son activité rédactionnelle, mais également de synthétiser des définitions lorsque celles-ci font défaut.

Cette étude permet de reconstituer la définition de l'une des notions nodales de l'œuvre de Gustav Landauer : le terme « esprit » (Geist), accompagné de ses multiples dérivés. Cette notion, dont l'histoire est très dense au sein de la tradition philosophique allemande, fait ici l'objet d'une analyse essentiellement internaliste à l'aide d'outils d'analyse textométrique. En effet, si la singularité et l'importance - quantitative, mais également conceptuelle - de cette notion chez Landauer a bien été relevée par les commentateurs jusqu'ici, aucune définition complète n'a encore été proposée. Cet exposé propose de retracer les étapes de cette élaboration conceptuelle, de manière à présenter les clefs - notamment philologiques - qui ont permis de restituer la cohérence d'un système de pensée construit à l'encontre de toute systématisation. L'étude statistique des occurrences de la notion de Geist, de son environnement sémantique et des évolutions structurant son usage permettront d'aboutir à une compréhension précise, fondée sur le texte dans sa complexité et sa globalité, d'un terme à l'interprétation délicate. Ce travail offre une occasion de discuter de la pertinence et des limites épistémologiques d'une analyse textométrique en histoire de la philosophie.

Élaborer, numériser, mettre en ligne et exploiter un corpus d'auteur. Exemples de deux cas pratiques en littérature

Marianne Froye et **France Marchal Ninosque**
ELLIADD, EA 4661
Université de Bourgogne Franche-Comté, France

À travers un dialogue entre deux expériences de chercheurs sur des corpus d'auteurs (un romancier, Louis-Combet et un poète, Frénaud), France Marchal-Ninosque et Marianne Froye vont tenter de conceptualiser l'approche scientifique de la constitution et de l'exploitation d'un corpus littéraire, selon la méthode inductive mise en place par ces deux enseignantes-chercheuses à l'Université de Bourgogne Franche-Comté (EA Elliadd).

Un retour expérimental sera porté sur les trois paradigmes du travail du chercheur : l'expertise (pour l'établissement d'un corpus), le conseil (pour la plateforme numérique conçue par un ingénieur informaticien), l'interprétation (pour l'exploitation du corpus). Un regard sera porté sur la plateforme numérique Fanum qui accueille des corpus littéraires et artistiques.

Puis la réflexion portera sur la nature du corpus littéraire devenu numérique, à la fois signifiant et matière. Une fois numérisé, un corpus implique qu'il soit pensé comme tel : sous son format numérique, ce corpus implique des choix qui lient l'iconicité et la plasticité d'un nouveau document, devenu numérique. Les choix graphiques, les liens hypertextes peuvent-ils transformer le discours qu'on peut faire à partir de l'objet originel, c'est-à-dire le manuscrit ? L'interface homme-machine qu'est une plateforme numérique de corpus littéraires (l'exemple de la plateforme Fanum de l'UFC sera développé) peut-elle influencer l'exploitation traditionnelle des archives littéraires (génétique, éditions de texte...) ?

I. Présentation des projets : de la diversité des corpus vers une plateforme commune

À partir d'un bref historique de deux expériences, les enseignantes-chercheuses tenteront de montrer les principaux enjeux des projets en humanités numériques sur des corpus d'auteurs. Les deux expériences montrent une variété certaine : un romancier et un poète, des pratiques diverses dans le processus créatif : des brouillons relativement propres pour Claude Louis-Combet et des manuscrits très raturés et fournis pour Frénaud, mais un but commun : exploiter et interpréter les états intermédiaires de la création littéraire et concevoir une plateforme générique de visualisation et d'exploitation.

I. 1. Claude Louis-Combet (CLC) : expérience de France Marchal-Ninosque

Le travail effectué s'est composé de six principales étapes : il a fallu dans un premier temps collecter les manuscrits et les brouillons de l'auteur, photocopiés. Le projet de conservation a été initié dès les années 1990. S'en est suivie l'entreprise de numérisation d'un fonds in vivo, que l'auteur vivant alimentait régulièrement. Différents personnels ont numérisé le fonds et ont informé les fiches, tels des archivistes. L'ingénieur informaticien a ensuite conçu l'interface pour mettre à disposition de la communauté scientifique l'ensemble des données, tout en réfléchissant avec l'institution aux enjeux juridiques. Depuis lors, les chercheurs novices et confirmés exploitent le fonds dans le cadre de thèses, de colloques monographiques ou thématiques.

I.2. Frénaud numérique : expérience de Marianne Froye

L'expérience sur l'œuvre de Frénaud est différente dans son processus ; il s'est agi dans un premier de reconstituer le fonds qui avait été désorganisé après le legs opéré par la veuve du poète. À cette étape, il fallait essentiellement inventorier le fonds pour connaître le contenu de toutes les pochettes et boites léguées. La reconnaissance du projet par le consortium Cahier a permis des avancées très importantes dans le traitement du fonds. La numérisation exigeait des subventions peu conséquentes que les institutions ou les appels à projet peinaient à financer, car la rentabilité scientifique à court terme ne leur semblait pas patente. La confiance de Cahier en ce projet a permis d'avancer rapidement une fois la numérisation effectuée. Le processus créatif de Frénaud mêle écriture de poèmes et exégèse. Le corpus qui semblait alors le plus approprié pour affiner la connaissance de son geste d'écriture était composé de deux œuvres : *La Sorcière de Rome* et *Gloses à la Sorcière*. Le choix a mêlé arguments scientifiques, pratiques, pragmatiques et génétiques. Notre volonté était d'étudier le processus de création de Frénaud. Nous souhaitions donc avoir en regard l'œuvre poétique et son exégèse rédigée par le poète lui-même, pour mettre en évidence sa pratique d'écriture. L'ensemble le plus achevé et le plus accessible physiquement correspond à ces deux ensembles de brouillons.

I.3 FANUM[19] / FANA[20] : Multiarch

L'ingénieur informaticien de l'EA ELLIADD a conçu deux interfaces de visualisation l'une FANA[21] pour les arts du spectacle vivant, l'autre FANUM pour les manuscrits de corpus littéraires. L'évolution souhaitée est la réunion des deux plateformes en une seule, pour visualiser et exploiter l'ensemble des fonds sur MultiArch. Ces plateformes ont permis de rendre accessible la totalité de ces données à l'ensemble de la communauté scientifique. La création d'un outil générique serait un apport décisif. L'intégration récente du projet Frénaud à cette structure de recherche permet de faire évoluer l'offre de visualisation offerte par FANUM et vise à appliquer le TAL à ces œuvres littéraires pour en renouveler l'approche et en approfondir l'interprétation génétique.

II. En amont et en aval : le corpus comme colonne vertébrale

Le premier bilan tiré de ces différentes expériences en humanités numériques consacre l'importance du corpus. Elle se justifie pour différentes raisons : sans la constitution d'un corpus raisonné et réfléchi, aucune exploitation n'est possible. Des compétences nécessaires sont multiples : elles sont celles des chercheurs, des ingénieurs informaticiens, des juristes et des archivistes.

II.1 La constitution du corpus. Le chercheur expert : l'alpha

Le chercheur endosse plusieurs casquettes lorsqu'il entreprend de traiter numériquement un corpus littéraire. Il est, en amont du projet, celui qui possède l'expertise par sa connaissance du fonds à traiter, par sa capacité à déchiffrer l'écriture parfois illisible de l'auteur étudié. Son expertise sur l'histoire de sa discipline, de la critique, les études génétiques et sur l'histoire littéraire est également essentielle à la constitution optimale d'un corpus d'étude. Enquêteur à la recherche d'indices dans les manuscrits, il reconstruit le cheminement de la création, il devient de ce fait l'épicentre de multiples ressources qui sont éparses. Il est donc celui qui, en construisant un corpus fiable et solide, légitime la recherche. Étape et rôle essentiels finalement peu visibles dans l'expertise scientifique

[19] Fonds d'Archives Numériques.
[20] Fonds d'Archives Numériques Audiovisuelles.
[21] https://fanum.univ-fcomte.fr//fana/?f=1

II.2 L'exploitation et la valorisation du corpus. Le chercheur interprète : l'oméga

L'autre rôle essentiel du chercheur se situe en aval du projet, au moment de l'interprétation des données recueillies. En conjuguant ses pratiques de recherche à celles d'autres disciplines, comme la linguistique et le TAL par exemple, le chercheur interprète des occurrences lexicales, des modifications opérées par l'auteur, ou encore des correspondances textométriques.

L'autre domaine d'expertise du chercheur est l'interprétation historique qu'il mène sur le corpus ainsi constitué. La numérisation facilite l'étude des influences d'un auteur, met en évidence les phénomènes d'intra- et d'intertextualité et permet de constituer plus facilement le réseau intellectuel explicite ou implicite de l'auteur étudié.

Ces données, dont l'accès est grandement facilité, permettent une interprétation stylistique de plus grande ampleur, grâce au traitement numérique du fonds. L'outil informatique permet une visualisation optimisée de différents manuscrits, de différentes versions d'un même extrait.

Finalement, le chercheur peut interpréter la genèse de l'œuvre. Or, la numérisation et son exploitation linguistique opèrent au moins un déplacement, voire une évolution de cette critique. Elle est possible in vivo, elle impose de réfléchir désormais aux statuts des différents états physiques du manuscrits : le papier, la version numérisée, le fichier sur la plateforme...

II.3 L'étape médiane. Le traitement du fonds : le chercheur consultant

Mais le chercheur et le projet scientifique ne seraient rien sans l'aide d'autres ressources humaines primordiales au traitement du fonds. Le chercheur se met en retrait pour être davantage un consultant qu'un acteur lors de cette étape médiane. L'ingénieur informaticien est la cheville ouvrière de la conception de la plateforme de visualisation. L'apport novateur du numérique est sans conteste l'accessibilité d'un fonds à l'ensemble de la communauté au-delà des limites de temps et d'espace. Or, sans sa visualisation optimale et optimisée, le fonds numérique resterait tout aussi inaccessible que ne l'est le fonds papier. L'informaticien permet donc un véritable accès au fonds constitué. Les projets en Humanités numériques seraient donc des projets Janus, à la fois tournés vers les sciences humaines et vers les sciences informatiques. L'expertise de l'ingénieur comme des chercheurs en informatique pour développer de nouvelles fonctionnalités est essentielle. Malgré les formations que les littéraires pourraient suivre, leur connaissance du code restera toujours en-deçà de celle d'un informaticien.

Les besoins humains dépassent ces deux disciplines et appellent d'autres ressources de personnel universitaire, notamment les juristes. L'accessibilité rendue possible par le support numérique pose la question des droits. Consulter un manuscrit en bibliothèque nécessitait jusqu'à présent une autorisation des ayant-droit, qui est à repenser lorsqu'une vue numérisée du même document est publiée sur un site internet. Les juristes permettent donc d'encadrer l'accessibilité à la plateforme et traitent également avec les maisons d'édition pour les auteurs qui ne sont pas tombés dans le domaine public.

Enfin, les besoins humains sont également au sein de la communauté scientifique : les projets en humanités numériques mobilisent des compétences pluridisciplinaires : les littéraires ont besoin des linguistes, des informaticiens, des historiens... Cette pluridisciplinarité favorisée par ce renouvellement épistémologique tend à une transdisciplinarité qui reste pour le moment utopique.

III. Une terminologie pour une méthodologie

L'apport du numérique pour traiter les fonds d'archives d'auteurs est indéniable, mais il implique aussi des changements de paradigme importants, que nous tenterons de conceptualiser dans un dernier temps. Les conséquences épistémologiques sont profondes et demandent à être éclaircies. Le numérique a notamment comme enjeu une redéfinition de la critique génétique. Il impose donc de réfléchir à une terminologie qui permet d'ordonner la pensée et d'organiser la

méthodologie façonnée de manière intuitive, inductive, empirique, au gré des expériences en génétique. Elle permet de poser un regard rétrospectif pour mettre du sens sur le travail du chercheur généticien et essayer d'avancer plus rapidement et plus efficacement sur les projets suivants.

III.1 Une terminologie pour donner à voir les archives

La constitution d'un corpus, comme nous l'avons évoqué à partir de deux exemples précis, impose des choix. Le chercheur qui constitue son corpus en vue de le numériser donne une orientation scientifique à son projet. Le fruit de la numérisation est une lecture du chercheur de l'ensemble du fonds. En agissant ainsi, le chercheur donne à voir à l'ensemble de la communauté scientifique des vues numérisées qui ne correspondent plus finalement à un fonds vierge de toute entreprise scientifique comme les chercheurs pourraient y avoir accès lorsqu'ils se déplacent en bibliothèques pour découvrir des manuscrits. Le chercheur à l'origine d'un projet influe la suite des recherches génétiques sur un auteur ou sur un ensemble de manuscrits. La mise à disposition n'est pas totalement neutre. Il convient donc de réfléchir aux statuts de ces différentes archives physiques et / ou dématérialisées pour comprendre la perspective scientifique de la démarche du chercheur. En ce sens, la mise à disposition sur Nakala de données fairisées modifie en profondeur les rapports aux archives. Leur accessibilité et leur réutilisabilité sont deux apports indéniables qui invitent à ce que nous réfléchissions aux statuts des différents états ou statuts des archives.

III.2 Une terminologie pour découper / pour décrire le texte et pour l'encoder

Travailler sur des fonds littéraires de genres différents et sur des archives audiovisuelles montre toute la difficulté à trouver le vocabulaire adéquat pour décrire la réalité du document. Or, ce n'est pas uniquement une question lexicale. Le vocabulaire employé est consubstantiel à la réalité décrite. Espérer concevoir une plateforme générique impose une grammaire commune. Plusieurs raisons en sont la cause, la première est sans aucun doute l'étape de l'encodage. La seconde, celle de la visualisation. Les choix des balises de l'encodage dépendent de cette grammaire ; l'arborescence du site et sa granularité également. Que définir comme « texte » ? Que signifie « document » ?

III.3 Une terminologie pour exploiter le texte

Finalement, le changement de paradigme apporté par le numérique entraine la conception de nouveaux outils d'analyse, notamment ceux développés avec les chercheurs en informatique ou les ingénieurs de recherche. Le numérique automatise et fiabilise certains résultats, comment le chercheur en SHS s'approprie-t-il ces données ? Le décompte d'occurrences est certainement plus fiable lorsqu'une machine l'effectue que lorsque c'est un humain. Les outils conceptuels à disposition gardent-ils toute leur efficacité ? Comment peut-on les faire évoluer ?

J. M. G Le Clézio et les genres littéraires. Étude textométrique d'un corpus littéraire

Margareta Kastberg Sjöblom
ELLIAD, EA4661
Université de Franche-Comté, France

La notion de genre reste encore aujourd'hui l'institution première du code littéraire, bien qu'elle ait souvent été discutée et remise en question. Certains théoriciens la considèrent avec réserve, affirmant que chaque genre littéraire en englobe plusieurs, et les hésitations terminologiques manifestent ce caractère « d'appartenance multiple et emboîtante » de tout écrit littéraire. En effet, la codification des genres n'est pas chose aisée et les catégories ne sont pas stabilisées. Le système traditionnel propose - ou nous impose - selon le code générique institutionnel, certaines classifications reconnues : romans, nouvelles, essais, etc.

Pourtant des études ont montré que les genres existent, et qu'il serait inconcevable sur le plan purement linguistique de nier l'existence de différentes typologies de textes. L'analyse textométrique valide cette idée, l'opposition générique est extrêmement claire et permet de définir des caractéristiques génériques en s'appuyant, non sur des valeurs anthropologiques ou sociales, mais sur les propriétés mêmes des textes.

Plusieurs écrivains français remettent en question ou refusent même le cloisonnement en genres, parlant d'une seule et unique écriture. Parmi ces auteurs certains ont une large production qui se décline en plusieurs genres littéraires. C'est le cas de l'œuvre de J. M. G. Le Clézio, un de plus grands écrivains contemporains à laquelle nous nous intéressons dans cet exposé.

Le présent exposé propose d'étudier les variations et les oppositions génériques, en s'appuyant sur un corpus informatisé et lemmatisé, et en exploitant les techniques quantitatives. L'œuvre de Le Clézio présente en effet une riche variété de textes qui se déclinent en différents genres. Bien qu'il évoque souvent une « écriture unique » affranchie des genres, déclare n'appartenir à aucun groupe et tente même de transgresser un système social établi, les différentes typologies de textes permettent d'observer dans ses écrits des variations à tous les niveaux.

L'opposition entre différentes typologies est en effet toujours présente et souvent même prépondérante dans les analyses statistiques. Les spécificités génériques dans les analyses de statistique lexicale est si forte qu'elle empêcherait même de fonder de grands espoirs sur les méthodes quantitatives pour attribuer un texte à un écrivain plutôt qu'à un autre. Un excellent exemple de la force de ce clivage générique est celui de certaines tragédies de Molière que l'on a attribuées à Corneille (Brunet, 2000).

Ces variations déjà bien documentées, sont-elles observables également à l'intérieur d'un corpus ou dans l'œuvre d'un seul écrivain ? Comment évoluent-elles ?

Le Clézio s'est lui-même intéressé à tout le processus de la création littéraire et il affirme un refus de certaines normes en érigeant ce refus en contestation sociale. Accepter les conventions du roman, ou de tout autre type d'écriture présentait, pour l'écrivain, surtout au début de sa création, le risque de s'enfermer dans un système sociopolitique, dans un cloisonnement conventionnel des genres qui le dérangeait profondément. Tout au long de sa production littéraire, Le Clézio a en effet tenté des expériences en transgressant les catégories et les genres, notamment celui du roman.

« Tout ce qu'a écrit Le Clézio », remarque Michelle Labbé (Labbé : 1999), « du moins jusqu'à *Désert*, contient le roman de sa lutte contre le roman, sa quête de l'écriture, la grande histoire d'amour de l'œuvre ». Il ne propose pas de théorie structurée sur la création romanesque ni de critique de forme sur le roman dit « traditionnel », comme ont pu le faire ses contemporains N.

Sarraute, A. Robbe-Grillet, J. Ricardou ou Ph. Sollers, mais plutôt des réflexions fréquentes, récurrentes et dispersées.

On trouve en effet ces réflexions sur la littérature dans toute l'œuvre leclézienne, aussi bien dans les articles et les essais que dans les préfaces, les nouvelles, les romans et même les épigraphes aux chapitres de romans.

L'œuvre de JMG Le Clézio est riche et s'étend sur presque soixante ans. Notre corpus informatisé et numérisé est constitué tout d'abord des six premières œuvres, classées, par leur style particulier et innovant, comme appartenant à l'École du « nouveau roman » : *Le Procès-verbal*, *La Fièvre*, *Le Déluge*, *Le Livre des fuites*, *La Guerre* et *Voyages de l'autre côté*. Les neuf romans qui suivent cette période, considérés par les critiques comme plus « traditionnels », sont les suivants : *Désert*, *Le Chercheur d'or*, *Voyage à Rodrigues* (écrit sous forme de journal personnel), *Angoli Mala*, *Onitsha*, *Étoile errante*, *La Quarantaine*, *Poisson d'or* et *Hasard*. *Mydriase* et *Vers les icebergs* sont difficiles à classer dans un genre précis, ce sont plutôt des récits poétiques. Le corpus inclut ensuite les recueils de nouvelles : *Mondo et autres histoires*, *La Ronde et autres faits divers* et *Printemps et autres saisons*. Les essais littéraires sont de différentes époques. *L'Extase matérielle* et *L'Inconnu sur la terre* traitent de thèmes généraux tandis que *Trois villes saintes* et *Le Rêve mexicain ou la pensée interrompue* s'intéressent exclusivement à la culture amérindienne. Celle-ci constitue également le principal sujet des ouvrages à vocation ethnologique, *Les Prophéties du Chilam Balam* et *La Fête chantée*, tandis que *Sirandanes* s'intéresse à la culture de l'île Maurice. Sont inclus en outre dans le corpus deux livres pour enfants : *Voyage au pays des arbres* et *Pawana* ; la biographie *Diego et Frida*, et le récit de voyage *Gens des nuages*.

Notre corpus contient plus de deux millions occurrences et plus de cinquante mille lemmes pour trente-et-une œuvres. Il a été numérisé et traité par le logiciel Hyperbase, version 10. Le traitement textométrique automatisé ouvre la voie à des interprétations et à des études différentes du corpus, basées sur des données statistiques qui permettent une analyse contrôlée et systématisée.

L'analyse du corpus montre ici que le lexique, la morphosyntaxe, la structure et le rythme de récit varient avec les genres. Les oppositions génériques sont tout d'abord observables dans la structure du vocabulaire et dans son évolution ; c'est l'étude de la richesse lexicale, de la diversité du vocabulaire, de l'accroissement lexical ainsi que des hapax qui permet de tirer des conclusions sur ce phénomène.

L'analyse de la structure lexicale du corpus permet en effet de constater le rôle primordial du genre littéraire. Les essais, les ouvrages ethnologiques et la biographie présentent une richesse lexicale avec une grande spécialisation du vocabulaire ainsi que des apports lexicaux importants dans notre corpus. La bipolarité de la structure confirmée par l'analyse statistique, avec un vocabulaire qui tend soit vers l'abondance soit vers le dépouillement, est le fidèle témoin du paradoxe de l'écriture leclézienne et oppose ainsi le sous-genre « nouveau roman » au genre « roman traditionnel ».

L'étude des parties du discours et de la syntaxe à travers une analyse « stylométrique », possible grâce aux versions lemmatisées et étiquetées du corpus, permet de relever aussi certains aspects morphologiques et syntaxiques qui différencient les types de textes.

Dans notre corpus, ce deuxième critère, morphologique, montre que la première période « nouveau roman » se démarque grammaticalement toujours du reste par son usage important du substantif et de l'adjectif, mais aussi par l'emploi de l'impératif et, paradoxalement pour une écriture expérimentale, par l'usage de formes temporelles très traditionnelles comme le passé simple. L'étude de temps verbaux et de l'usage très personnel qu'en fait Le Clézio permet de mieux cerner une technique qui consiste à donner au récit cette valeur universelle tant appréciée par ses lecteurs.

Une écriture qui change est une des caractéristiques fondamentales de notre corpus. En effet, il n'y a pas de « stabilisation » du style mais, au contraire, des écarts grandissants chez le Clézio. Toutefois, bien que les procédés morphosyntaxiques ne soient pas statiques, que les techniques d'expression changent, qu'elles évoluent et qu'elles soient constamment mises en question, c'est l'opposition générique qui reste prépondérante.

L'opposition générique opère aussi au niveau thématique. L'étude de la distance lexicale entre les différents livres du corpus ainsi que celle des spécificités lexicales, mettent en exergue les variations isotopiques, récurrentes dans ce corpus « multigénérique ». De la même façon que dans les analyses structurales et morphosyntaxiques, l'analyse des corrélats sémantiques et thématiques révèle aussi des caractéristiques de chaque typologie présente dans ce corpus et l'analyse factorielle montre que les mêmes orientations des textes se retrouvent aussi bien au niveau thématique.

En effet, le refus de genres est souvent une position idéaliste ou sociopolitique. Aussi, bien que Le Clézio refuse toute appartenance à un genre littéraire et que les critiques aient souvent souligné le mélange des genres dans un même ouvrage, nos analyses montrent que l'appartenance à un genre précis de chacun de ses livres est bien réelle.

Chaque genre littéraire a en fait son anatomie, sa physiologie et son fonctionnement, et cela transparaît très clairement dans les différents textes qui forment l'œuvre leclézienne.

Références

Adam J.-M., *Les textes types et prototypes : Récit, description, argumentation, explication et dialogue*, Paris, Armand Colin, collection Fac. linguistique, 2005.

Étienne Brunet, « Peut-on mesurer la distance entre deux textes ? », in Rastier F. (éd.) *Corpus littéraires - Recueil et numérisation, analyses assistées, didactique*, Paris, 2000.

Kastberg Sjöblom M., *L'écriture de J.M.G. Le Clézio -- Des mots aux thèmes*, Paris, Honoré Champion, 2006.

Kastberg Sjöblom M., Leblanc J.-M., « Extraction des isotopies d'un corpus textuel - analyse systématique des structures sémantiques et des cooccurrences, à travers différents logiciels textométriques », Ch. Cusimano (éd.), *Texto! Textes & Cultures*, Numéro XVII-3, 2012, http://www.revue-texto.net/docannexe/file/3059/texto_kastberg_leblanc.pdf.

Kastberg Sjöblom M., « La textométrie au service de l'analyse comparative de discours de présidents africains », in M. Kastberg Sjöblom A. Barry, A. Chauvin-Vileno, (éds.) *Actes du 6ème colloque du réseau discours d'Afrique : Nouvelles voix/voies des discours politiques en Afrique francophone*, Besançon, Cahiers de la MSHE -- Université de Franche-Comté, 2021(en cours de publication).

Labbé M., *Le Clézio, l'écart romanesque*, Paris, L'Harmattan, 1999.

Malrieu D. & Rastier F. (2002) « Genres et variations morphosyntaxiques », in Angel Martin Municio (éd.), *Actas del segundo seminario de la escuela interlatina de altos estudios en linguistica aplicada, Matemàticas y tratamiento de corpus*, Logroño, Fundaciòn San Millàn de la Cogolla, 2002, p. 61-84.

Muller Ch., *Principes et méthodes de statistique lexicale*, Paris, Hachette, 1977.

Rastier F., *La mesure et le grain. Sémantique de corpus*, Paris, Champion, 2011.

Lectures zoliennes.
Des manuscrits aux adaptations quelles données numériques en jeu ?

Table ronde animée par
Alina Iuliana Nastase Gonzalez
FoReLLiS, UMR 7285
ITEM, UMR 8132
INSPE CAEN
Université de Caen Normandie, France

Participants

Laurélenn Brière (éditrice scolaire et conceptrice pédagogique)
Alina Nastase Gonzalez (enseignante-chercheuse en didactique de la littérature)
Olivier Dobremel - Dobbs (auteur scénariste)
Jean-Sébastien Macke (ingénieur d'études)

Cette table ronde offre aux différents participants l'occasion d'engendrer une collaboration ponctuelle et future prometteuse, autour d'une question centrale dans le champ des études de réceptions des œuvres littéraires classiques à l'ère de l'hypersphère (Debray, 2002): la création transmédiale comme fabrication, prenant en compte la nature d'un médium différent (numérisation d'un corpus d'auteur) pour penser la diffusion du roman naturaliste d'Émile Zola (*La Bête humaine* et *L'Œuvre* plus précisément). Compétence du XXI^{e} siècle, un des « 4C » (Lamri, 2018), la collaboration, envisagée aujourd'hui entre ces différents acteurs, « passeurs » de Zola, permettra d'articuler autour des apports informatiques de l'exploitation des données, des projets artistiques, pédagogiques et éditoriaux, des projets de recherche en didactique de la littérature, génétique et esthétique romanesque, déjà existants.

C'est tout d'abord une connexion humaine et réflexive que nous visons à assurer le temps de cette table ronde réunissant quatre intervenants de domaines artistiques et de recherche différents. Mais aussi une synergie entre les usages multiples des bases de données textuelles relatives à un auteur : tous ces projets ont en ligne de mire la visualisation et la diffusion de la matière littéraire sous des aspects numériques hétérogènes, sur le plan de la genèse et/ou de la lecture (scolaire ou grand public).

Les projets des participants

En 2017, Alina Nastase Gonzalez, ATER à l'INSP'E de Caen et chercheuse en didactique de la littérature, lance, dans l'Académie de Poitiers, le projet d'une adaptation de La Bête humaine en jeu vidéo 2D à usage pédagogique. Quelques années plus tard, le dispositif didactique de cette recherche-action atteint sa forme provisoire de prototype test sur Unity, suite au travail de conception et de programmation réalisé par un étudiant de l'IUT Métiers du jeu vidéo de Bobigny, entre mai et août 2020. Né d'une passion pour l'œuvre de Zola, ce jeu vidéo pédagogique est mis au service de la recherche en didactique comme outil de visualisation de données textuelles et génétiques et moyen ludique d'incitation à la lecture des oeuvres patrimoniales.

En 2018, voit le jour la bande dessinée *La Bête humaine*, d'après le roman d'Émile Zola, par l'auteur scénariste Dobbs, de son vrai nom, Olivier Dobremel (avec Germano Giorgiani au dessin). Ce sociologue de formation, enseignant d'histoire du cinéma et de la bande dessinée à l'université Montpellier III et dans d'autres écoles supérieures, dit avoir répondu à une commande de son éditeur Hachette, intéressé de proposer plusieurs adaptations littéraires. Le scénariste se

penche alors sur *La Bête humaine* comme premier roman noir, mettant en avant plan le point de vue du criminel, choix subjectifs, influencés par son passé en criminologie et son attrait pour le réalisme.

Depuis 2017, Laurélenn Brière, actuellement éditrice dans le milieu scolaire (réalisation de manuels, cahiers d'exercices et d'ouvrages parascolaires LSH pour le collège et le lycée) et conceptrice pédagogique pour la LaPsyDÉ Labschool de l'universitaire (Université de Paris), s'intéresse à la pédagogie numérique et aux possibilités qu'offrent les langages XML et HTML. Elle envisage alors de concevoir une édition enrichie du roman *L'Œuvre*, pas comme les autres, sous la forme d'un site internet.

Jean-Sébastien Macke, ingénieur d'études à l'Institut des Textes et Manuscrits Modernes (ITEM/CNRS-ENS), depuis 2016, participe aux activités de l'équipe « Supports et tracés » qui travaille à l'analyse codicologique (papiers, encres, filigranes) des manuscrits du 18e au 21e siècle. Il poursuit également, au sein de l'ITEM, sa collaboration avec l'équipe Zola, en développant le projet « ArchiZ » et en mettant en œuvre un nouveau projet d'édition numérique de la correspondance générale d'Émile Zola.

Objectifs

Nous espérons ainsi consolider un ensemble systémique de pensées et d'actions convergentes, pivoté depuis son épicentre, par ses forces centrifuges et concentriques créant ainsi une tension positive dont nous analyserons les effets sur comportement lectorial « du public », dans un contexte de transmission numérique et multimédiale. Une tension crée par les technologies numériques à l'œuvre dans la conception de ces artefact zoliens présentés ici : les archives numériques de Zola (ITEM), le roman *La Bête humaine* en jeu vidéo 2D pédagogique et en bande dessinée, le roman *L'Œuvre* en version site internet.

Prenons alors l'objet littéraire, l'œuvre zolienne, et faisons-la tourner comme cette « étrange toupie » (Sartre, 1948 : 16) qui n'existe qu'en mouvement lors de la lecture et qui ne dure que le temps de celle-ci. Cette force centrifuge du mouvement serait l'impulsion donnée par le chercheur en didactique, exprimant l'essence littéraire de l'objet qui l'occupe. Imaginons ensuite des bouts de matière projetés vers l'extérieur et attrapés dans la trajectoire de ces multiples cercles concentriques qui définissent le champ d'action des différents domaines invités à collaborer autour de cette table ronde. Un mouvement oscillatoire d'adaptation commence alors, pour harmoniser la rotation en faisant converger leurs objets spécifiques préalablement modifiés, à nouveau vers le centre. Qu'advient-il à la fin de ce tournoiement du texte littéraire dans les diverses orbites ? En modulant sa forme de représentation, dans quelle mesure arrive-t-on à préserver sa substance ?

Les échanges montreront peut-être, par le biais des particularités de chaque milieu, à quel point il demeure essentiel pour la didacticienne - dans une approche transversale - d'adopter la posture du « didacticien gyrovague » (Chevallard, 2016 : 40), sans affiliation disciplinaire connue, dispensé de « montrer ses papiers » (Biagioli, 2014), qui sort de son univers spécifique pour aller vers des domaines dont il ne maîtrise pas les contenus mais dont il se servira naturellement pour atteindre les objectifs fixés. Cette méta-approche, anthropologique, des différents champs d'analyse trouve sa légitimité dans l'amplitude de l'espace de circulation et de transmission du patrimoine littéraire à l'ère de la numérosphère, ou de l'hypersphère (Louise Merzeau, 1998 : 19). En définitive, ce sera une belle opportunité pour chacun.e d'entre nous d'adopter, face à l'œuvre colossale et moderne de Zola, ce point de vue élargi, participant à la redéfinition de la lecture littéraire patrimoniale et des ses pratiques dans le champ scolaire et extra-scolaire, grâce à l'utilisation d'un corpus de données numériques.

Ainsi, une autre approche, médiologique (Dubray, 2000), de la transmission culturelle, techniquement déterminée par les moyens de transports adaptés à chaque médiasphère - chapeaute les dimensions spécifiques de chacun de ces projets autour de plusieurs axes.

En premier lieu, afin d'interroger et d'éclairer le processus de réception sous l'aspect de sa diffusion / valorisation et de réception lectoriale (novice ou experte), nous nous proposons de scruter ici l'horizon de ces formes transmédiales actuelles (BD, jeu vidéo, site internet, corpus

numériques) - objets sémiotiques secondaires (Louichon, 2015) - et de mettre en orbite les données numériques inhérentes à leur création.

Tout comme les différentes médiasphères s'étayent, s'imbriquent et se restructurent sous la domination de la dernière venue, la moins coûteuse ou la plus efficace, les approches d'analyse, de création et de transmission des romans, des avant-textes et des documents inédits de Zola cherchent à faire corps commun à l'aune de cette rencontre. Dévié depuis longtemps de son orbite graphosphérique (Debray, 2010) par des pratiques d'usage et de consommation culturelles renouvelées, le texte et les avants-textes de Zola, subissent et profitent d'un double phénomène d'actualisation: ses adaptations transmédiales et sa diffusion numérique à grande échelle. À partir de là, l'enjeu des différés acteurs réunis par cette contribution, peut apparaître comme étant l'harmonisation d'un regard éduqué et informé (en amont) sur la morphogenèse des données (Moretti, etc.) réfléchissant à la facilitation des regards lambda (utilisateurs lecteurs et public scolaire) sur le texte zolien ainsi modifié.

En deuxième lieu, notre rôle commun pourrait alors devenir celui de fixer un point de convergence de nos actions respectives sur le texte original, à savoir, l'interface comme espace de visualisation de ces données mobilisées, à des degrés différents d'opacité. Il nous revient donc d'embrasser les interfaces polymorphes de nos réalisations d'un même regard afin de pouvoir interroger d'une voix plurielle la lecture de l'œuvre de Zola que nous entendons inscrire dans la circulation des usages actuels, en direction des publics visés, plus ou moins experts.

Ceci nous contraint donc à prêter une attention particulière aux avant-textes, aux textes documentaires de nos recherches, mobilisés dans nos entreprises d'adaptation. Nous pourrons ainsi mettre en exergue l'espace génétique de nos productions transmédiales dans ses aspects plus techniques de structuration et d'agencement des données, recueillies, crées et réfléchies pour les nouveaux besoins des usagers.

Questionnements

Dans le cadre large du questionnement offert par l'appel à communication, il serait donc intéressant de s'interroger sur la place de l'herméneutique génétique dans ces travaux et projets, sur les progrès possibles dans le domaine de la visualisation des processus de création et sur les perspectives scientifiques de la génétique en tant que « science des processus » (De Biasi, 2017), d'autres problématiques, plus spécifiques à nos objets de recherche semblent intéressantes à formuler :

- dans quelle mesure le jeu pourrait être envisagé comme solution technologique de visualisation, de scénarisation, d'exploitation et de production de connaissances sur les corpus numérisés? L'univers du jeu élargit en amont, le champ d'action du lecteur - programmeur en l'incitant à avoir la parole efficace pour aménager des effets lors du codage informatique du texte ; en aval, celui du lecteur - joueur en le poussant vers le hors texte, plus précisément vers les avant-textes. L'expérimentation pédagogique du prototype test favorisera la manipulation des ressources numériques implémentées dans l'infrastructure du jeu et conduira le chercheur en didactique à repérer l'impact de l'artefact numérique sur les stratégies de lecture. Enfin, le développement du jeu La bête humaine correspond au désir d'intégrer l'artefact numérique dans la conceptualisation des parcours de lecture vecteurs de communication scientifique. L'usage spécialiste et spécialisé des outils et méthodes dans la fouille de données en amont, lors de la phase de conception du jeu, prépare le terrain d'exploitation sous une forme simplifiée, à destination d'usages moins experts.
- l'adaptation du même roman en bande dessinée, peut être perçue elle aussi, comme un acte éditorial et créatif de sauvegarde d'une œuvre patrimoniale nous définissant tous et générant, face à l'oubli, ce sentiment de perte collective. Il tiendra à cœur à l'artiste

scénariste de mettre en avant le travail de recherche déployé, en lien étroit avec un spécialiste des chemins de fer, donnant lieu à la constitution d'une base importante de données numériques. Ce n'est qu'un aspect du large spectre de données numériques qu'implique la genèse de ce projet. Par ailleurs, sera ici posée la question sensible de la fidélité au texte original et des intersections possibles avec d'autres adaptations existantes, l'écran par exemple, que l'artiste a su tenir à l'écart afin de conférer à sa création un regard personnel, vierge de toute contamination. Enfin, pourra être évoquée la problématique plus technique mais passionnante du *game design* en rapport avec la création numérique de bandes dessinées et par le même biais, la question de la fragilité des données dans un environnement créatif numérique.

- le projet d'une édition numérique enrichie d'un autre roman de Zola, *L'Œuvre*, sous forme de site internet, répond à deux besoins principalement: rendre accessible le texte pour un public non initié (collégien, lycéen, FLE ou lecteur du dimanche) et réaliser un travail de balisage XML du texte et favoriser l'extraction de données pour un public initié (étudiant ou chercheur). Cette entreprise soulève de nombreuses questions, du côté de l'intérêt pédagogique comme de la production technique d'un site internet comme interface de l'œuvre. Quel rôle joue l'esthétique du site, objet d'un véritable travail de réflexion du concepteur en amont, dans l'aide à l'apprentissage du lecteur non initié ? L'ergonomie du site assure une utilisation simple et efficace des diverses ressources, le choix des ressources textuelles numérisées, à des fins pédagogiques, se veut le plus varié possible : paragraphe explicatif, carte interactive, frise chronologique, extraits de films ou représentation de tableaux, etc. Enfin, la question de la production d'un site en XML puis en HTML relève d'un double enjeu : permettre, lors de la création du site, de coordonner les différentes pages avec une grande maniabilité, et, lors de l'exploitation du site, de fournir une base de données existante et accessible à tous.
- depuis 10 ans, le Centre d'étude sur Émile Zola et le naturalisme a entrepris un vaste projet de numérisation et de mise en ligne des archives zoliennes (manuscrits, textes imprimés, iconographie, correspondance, etc.). Dans le même temps, les technologies ont constamment évolué et un des enjeux de l'équipe a été de s'approprier ces outils numériques afin d'offrir à la communauté des chercheurs un corpus large dont la question serait de savoir en quoi cela a pu modifier l'approche scientifique de l'œuvre de Zola. Plusieurs questions émergent de ce projet : comment le numérique a-t-il ainsi modifié les pratiques de recherche dans l'étude d'un auteur classique ? Est-ce que ces pratiques ont favorisé ou non la recherche collaborative ? Quels en sont les éventuels écueils ? Mais on pourrait également réfléchir aux apports du numérique dans le cadre des études génétiques et notamment de la génétique éditoriale avec l'édition, du roman *Nana* sur « Variance » (Université de Lausanne), outil de comparaison de textes imprimés permettant de comparer différentes éditions afin de mettre en évidence et d'analyser les variantes, entre la parution en feuilleton et la première édition.

Quelques questions ouvertes à la salle

Comment les différents acteurs cartographient-ils le domaine des données textuelles (para-, méta-, hyper-, inter-) de l'univers romanesque zolien, dans la perspective plus large d'une ingénierie pédagogique (Musial, Tricot, 2020) ? Comment des projets de transposition et de numérisation distincts arrivent-ils à faire corps commun, en commençant, tout d'abord par faire connaissance ?

Quelles motivations communes inervent ces projets, décidés à manifester une action de type magique sur le livre, via les instruments spécifiques à leurs disciplines respectives ?

Comment ces adaptations et numérisations parviennent-elles à dynamiser l'acte de lecture littéraire dans un double sens: la motiver et motiver à travers d'elle : « is considered motivated someone who is energized or activated toward an end » (Ryan et Deci, 2000 : 54) ?

Déstructurant et émiettant l'original, l'infrastructure des nouveaux médias choisis, décontextualisent le plus souvent les données génétiques et textuelles. Qu'en est-il donc de la fidélité à l'original et du respect pour son intégralité dans la perspective d'une narratologie transmédiale (Baroni, 2017) ? Dans quelle mesure sommes-nous en droit de proposer ces (dé)lires aux actuelles et futures générations de lecteurs qui se pencheront sur notre ouvrage ?

Quelles lectures de Zola apparaissent aujourd'hui comme légitimes aux yeux du public, des institutions ? Peut-on parler de lectures plus ergonomiques, axées sur la visualisation des données numériques grâce à de nouvelles médialités et interfaces ? Pour quels usages, en direction de quel public ? Dans une perspective de l'esthétique de la réception située du côté du lecteur (Baroni, 2007), comment la scénarisation des données numériques génère-t-elle le pouvoir de la fiction via les trois émotions narratives fondamentales : la surprise, le suspense et la curiosité (Jouve, 2019 : 23) ?

Références

Baroni, R., « Pour une narratologie transmédiale », *Poétique*, 2017/2 (no. 182), 2017, p. 155-175. https://www.cairn-int.info/revue-poetique-2017-2-page-155.htm

Baroni, R., *La Tension narrative – suspense, curiosité et surprise*, Paris, Seuil, 2007.

Biagioli, N., « Didactique(s) : un singulier-pluriel. Réaction aux points de vue développés », in *Éducation et didactique*, 8-1, 2014. https://journals.openedition.org/educationdidactique/1870

Chevallard, Y. « Des didactiques des disciplines scolaires à la didactique comme science anthropologique », in *Éducation et didactique*, 8-1, 2014. https://journals.openedition.org/educationdidactique/1863

De Biasi, P-M., Herschberg Pierrot, A. *L'Œuvre comme processus*, Paris, CNRS Edition, 2017.

Débray, R., *Introduction à la médiologie*, Paris, PUF, 2002.

Galleron, I., Idmhand, F., « 'Réutilisabilité' : L'utilisateur dans l'édition électronique », in *Humanistica*, numéro 1, 2019, https://revues.univ-lyon3.fr/humanites-numeriques/

Jouve, V., *Les pouvoirs de la fiction. Pourquoi aime-t-on tant les livres ?*, Paris, Armand Colin, 2019.

Lamri, J., *Les Compétences du 21e siècle. Comment faire la différence ? Créativité, Communication, Esprit Critique, Coopération*, Malakoff: Dunod, 2018.

Louichon, B., « Le patrimoine littéraire : un enjeu de formation », in *Tréma*, 43. https://journals.openedition.org/trema/3285

Lebrun, M., « Lire des textes littéraires à l'ère des humanités numériques », *Revue de recherches en littératie médiatique multimodale*, 1, 2015. https://doi.org/10.7202/1047789ar

Merzeau Louise, « Ceci ne tuera pas cela », *Les cahiers de médiologie*, 1998/2 (N° 6), p. 27-39. DOI : 10.3917/cdm.006.0027.

Merzeau, L., *Introduction à la médiologie*, Paris, Presses Universitaires de France, 2000.

Moretti, F. « Introduction. La littérature, à sa mesure », in *La Littérature au laboratoire*, édité par Franco Moretti, 7-18, Paris, Les éditions d'Ithaque, 2016.

Musial, M. et Tricot, A. *Précis d'ingénierie pédagogique*, Louvain, De Boeck Supérieur, 2020.

Ryan, R., Deci, E., « Intrinsic and Extrinsic Motivations : Classic Definitions and New Directions », Contemporary Educational Psychology, 25, 2000, p. 54-67. http://selfdeterminationtheory.org/SDT/documents/2000_RyanDeci_IntExtDefs.pdf .

Sartre, J.-P., *Qu'est-ce que la littérature* ? Paris, Gallimard, 1948.

Schuwey, C., *Interfaces. L'apport des humanités numériques à la littérature*, Suisse, Livreo-Alphil, 2019.

Deuxième partie. À vol d'oiseau : traitements, motifs, ontologies et visualisations de corpus complexes

Diffuser et expliciter les traités d'agriculture de l'Antiquité. Les humanités numériques dans le projet AgroCCol

Sarah Orsini
HiSoMA, UMR 5189
École normale supérieure de Lyon, France

Le projet AgroCCol est piloté à l'ENS de Lyon par Maëlys Blandenet et est financé par l'ANR. Une des principales missions de ce projet[1] consiste en la réalisation d'une édition numérique d'un recueil des textes agronomiques autour de la culture des céréales et des légumineuses (Hésiode, Théophraste, Caton, Varron, Columelle, Virgile, Pline l'Ancien, Galien, Palladius, les Géoponiques).

En effet, l'établissement et la traduction de ces textes doit être mise à jour en raison de leur technicité. L'enjeu principal est de faire le point sur les termes techniques employés par les auteurs (noms de plantes, d'animaux, d'outils) et sur les pratiques agricoles (moissonner, amender le sol, gérer l'eau, stocker ...) en mettant en relation les textes avec les réalités archéologiques ainsi qu'avec les usages actuels. C'est pourquoi l'équipe réalisant ce projet associe des chercheurs en littérature ancienne[2], en littérature moderne[3] et des archéo-botanistes[4].

Cette édition repose sur un encodage XML-TEI des textes (à partir d'éditions critiques revues) et de leur traduction (à partir de traductions libres de droit revues et adaptées). Nous réalisons ensuite deux types d'indexation sur ces textes. Tout d'abord nous réalisons une indexation lexicale, en constituant un dictionnaire technique (noms de plantes, outils, aménagements, gestes et travaux agricoles). Ce dictionnaire est associé à un thésaurus que nous avons constitué et qui fait correspondre les noms de plantes et d'outils entre le latin, le grec et le français. Ce thésaurus est enrichi de fiches sur les données littéraires et archéologiques sur les concepts. Ensuite, une indexation thématique met en rapport l'ensemble des textes en fonction du thème choisi (par exemple les outils ou les animaux nuisibles). L'édition propose aussi un index des entités nommées. Enfin, des notices scientifiques, reliées aux textes et aux index par des liens hypertextes, permettent de se documenter sur les pratiques agricoles et sur les termes techniques sous la forme d'articles.

Cette édition offre donc les avantages suivants :

- Les textes et leur traduction ainsi que les articles scientifiques sont en libre accès.
- La traduction des termes techniques est harmonisée entre latin, grec et français.
- La recherche d'occurrences concerne trois langues.
- L'indexation thématique permet de systématiser l'analyse et de circuler entre les textes de façon comparative.
- La navigation entre le texte et les outils d'élucidation est facilitée et fait gagner du temps au lecteur. Cela permet de diversifier les lectures et le lectorat : un public de chercheurs et un public plus large (enseignants, étudiants)

---

1 Nous avons également réalisé une exposition virtuelle « Le ménage des champs » sur la transmission de ces textes de l'Antiquité à la Renaissance et une édition scientifique de Columelle, *De re rustica*, II ainsi que des manifestations scientifiques sur ce sujet.

2 Maëlys Blandenet et Michel Jourde (ENS de Lyon - HiSoMA), Pascal Luccioni et Marine Bretin-Chabrol (Lyon 3 - HiSoMA).

3 Michel Jourde (ENS de Lyon - IHRIM).

4 Marie-Pierre Ruas (CNRS) et Michel Chauvet (INRA - UMR AMAP de Montpellier).

Encyclopédies médiévales en milieu numérique. Les nouveaux enjeux de SourcEncyMe pour le traitement des *auctoritates*

Isabelle Draelants
IRHT, UPR841
CNRS, France

Cette communication propose un regard rétrospectif et réflexif sur les humanités numériques dans le domaine de l'histoire des textes ou des sciences de l'érudition, à partir de du cas du programme SourcEncyMe dédié à l'étude des sources des encyclopédies médiévales.

Les humanités numériques ne sont plus une discipline récente et le web est un cimetière à projets. L'obsolescence des outils et technologies mises en œuvre en est une des causes, l'arrêt des financements obtenus par dépôt de projet en est une autre. Ce temps brévissime des financements et de la technologie va à l'encontre de la pérennisation requise par la recherche, et surtout de la nature même de l'érudition, qui nécessite expertise et temps long. L'investissement considérable en travail et en expérience se heurte au vieillissement rapide des outils, et le chercheur en « humanités » doit sortir de son domaine d'expertise pour devenir à la fois entrepreneur et ouvrier de l'édition électronique. Nombre de chercheurs en sciences de l'érudition ont le sentiment de perdre leur âme, car on dispose de moins en moins de temps pour approfondir sa discipline, alors qu'il est devenu inimaginable de ne pas recourir aux outils numériques. Dans le même temps, les exigences ont augmenté : il ne s'agit plus seulement de créer chacun son outil pour mettre en ligne des résultats vérifiés, mais d'utiliser des langages (ex. html), des balisages (ex. TEI) et des référentiels toujours plus nombreux. Il faut désormais aussi offrir en open source le codage utilisé et la description des données (DMP), sans parler de la transformation du vocabulaire qui doit emprunter au jargon informatique : les contenus érudits de toute nature perdent leur nom pour être étiquetés en « données » ou « métadonnées ». Enfin, il faut tendre à une interopérabilité croissante entre les réalisations en humanités numériques.

Tels sont les défis que doivent affronter les projets fondés sur un patrimoine textuel, cherchant à rassembler un corpus de textes vérifiés et à en enrichir le contenu scientifique au fil des ans. SourcEncyMe (Sources des Encyclopédies Médiévales) est un de ces programmes qui doit conjoindre croissance des contenus spécialisés et évolution des outils techniques. Après une dizaine d'années d'existence, un déménagement institutionnel en 2014 et une mise à disposition du public en 2016, l'outillage du projet SourcEncyMe nécessite une complète rénovation, en cours, et des moyens humains pour introduire des contenus de la qualité attendue par les publications d'érudition traditionnelles.

SourcEncyMe rassemble les encyclopédies médiévales latines pour étudier leurs sources, c'est-à-dire tous les auteurs et les œuvres utilisées par les encyclopédistes, en particulier au « siècle d'or des encyclopédies médiévales », le XIII^{e} siècle. Tissées de 75 à 95 % de citations de textes antérieurs ou contemporains, elles représentent un objet emblématique de l'intertextualité des œuvres médiévales. L'enjeu de leur étude est de donner accès virtuellement à la bibliothèque fréquentée par les « compilateurs » auteurs des encyclopédies, autrement dit à un large héritage littéraire, scientifique et philosophique.

La communication vise à montrer, avec un retour d'expérience sur SourcEncyMe, comment les défis ci-dessus peuvent être affrontés et quelles solutions mises en œuvre. Les objectifs scientifiques du projet et ses réalisations en matière de corpus d'encyclopédies médiévales seront exposés ; les questions relatives à la pseudépigraphie (attribution des *auctoritates*), à l'identification des citations, et aux annotations sur la tradition textuelle, seront abordées à partir d'exemples précis. On explorera également le volet technique du passage de bases de données

reliées en PhP-MySQL à une base de données unifiée XML native (BaseX) et d'une interface d'administration collaborative à un outil de balisage « chercheur-friendly » en mode « stand-off embarqué » au format XML-TEI, ainsi que les liens à créer avec les référentiels d'autorités des institutions patrimoniales nationales et internationales, la potentialité de l'identification automatique des sources, et l'interopérabilité possible avec d'autres réalisations.

L'édition de textes fragmentaires en TEI xml : stratégies d'encodage

Estelle Debouy
ArScAn – AnTET UMR 7041
Université de Poitiers, France

Un certain nombre de textes ne sont connus que par la tradition indirecte, c'est-à-dire qu'ils ont été transmis uniquement sous la forme de citations chez d'autres auteurs. Ces fragments peuvent se présenter sous la forme de citations précises ou bien sous la forme de simples allusions, ou encore de paraphrases plus ou moins éloignées du texte original. En les insérant dans son texte, un auteur les préserve de l'oubli tout en leur conférant d'autres significations[1]. C'est précisément cette double caractéristique que doit restituer une édition numérique de textes fragmentaires : l'éditeur doit s'attacher à représenter les citations des fragments afin d'être en mesure de les repérer et de les extraire, et de fournir par la même occasion un inventaire précis des auteurs et des œuvres citées ; il doit également éditer le contexte où ces citations apparaissent car c'est le contexte qui lui permettra de proposer une lecture interprétative et scientifique du texte. L'un des enjeux d'une édition numérique de fragments est donc de dé- passer les limites des corpus relevant de la culture de l'imprimé où les citations sont reproduites sous la forme d'extraits décontextualisés.

L'objet de cette présentation est de réfléchir à une méthodologie pour mener à bien une telle édition numérique. Cette réflexion a notamment été entamée dans le cadre du projet d'édition de textes latins « The Digital Latin Library[2] ». Elle a aussi été menée plus précisément dans le cadre de l'édition de corpus de textes anciens fragmentaires, comme « The Leipzig Open Fragmentary Texts Series » (LOFTS), qui s'insère dans l'« Open Philology Project[3] » et dont l'ambition est de numériser des éditions imprimées d'oeuvres fragmentaires et faire des liens vers les sources, mais également de produire directement des éditions numériques d'oeuvres fragmentaires[4]. À titre d'exemple, citons au sein du LOFTS le projet « Digital Athenaeus[5] » ou encore l'édition des fragments des historiens grecs dont la version en ligne, réalisée sous la direction de Monica Berti[6], reproduit l'édition imprimée de Müller (1841-1872).

En m'appuyant sur ces travaux, je souhaite réfléchir à la façon de présenter en TEI les références aux sources citées (*apparatus fontium*), les variantes des fragments eux-mêmes (apparat critique) et le contexte dans lequel ils apparaissent. Je m'appuierai pour cela sur un exemple concret, celui des fragments d'atellanes : il ne reste que quelque trois cents vers de ces comédies latines écrites au Ier s. avant notre ère et transmises sous la forme de citations par le grammairien latin Nonius (IV^{e} s. apr. J.-C.) qui, pour composer son *Dictionnaire* (*De compendiosa doctrina*), a choisi des extraits d'auteurs servant à illustrer l'emploi d'un mot ou d'une locution rare. Comme l'œuvre

1 Sur la réalité textuelle qu'est le fragment, voir Françoise Daviet-Taylor et Laurent Gourmelen (dir.), *Fragments : Entre brisure et création*, Angers, Presses universitaires de Rennes, 2016.

2 Samuel J. Huskey et Hugh Cayless, « Guidelines for Encoding Critical Editions for the Library of Digital Latin Texts », 2020, https://digitallatin.github.io/guidelines/LDLT-Guidelines.html.

3 Cf. http://sites.tufts.edu/perseusupdates/2013/04/04/the-open-philology-project-and-hum boldt-chair-of-digital-humanities-at-leipzig/.

4 Monica Berti, « Annotating Text Reuse within the Context : The Leipzig Open Fragmentary Texts Series (LOFTS) », dans *Text, kontext, kontextualisierung. Moderne kontextkonzepte und antike literatur*, Zürich : Hildesheim, 2018, p. 223-234.

5 L'œuvre d'Athénée, qui se présente comme une compilation en quinze livres de propos de table de savants grecs et romains qui foisonnent de citations d'auteurs anciens, s'avère être une bibliothèque inestimable d'œuvres perdues. Cf. http://digitalathenaeus.org/.

6 Monica Berti, « Historical Fragmentary Texts in the Digital Age », dans *Digital classical philology*, Berlin, De Gruyter, p. 255-276, http://www.dfhg-project.org/.

d'Athénée, le traité de Nonius est donc une source précieuse de textes perdus qu'il faut pouvoir encoder de façon à les identifier, les analyser et les extraire. Je montrerai, pour finir, que le modèle TEI adopté peut être produit automatiquement à partir d'une saisie en LuaLaTeX grâce au logiciel ekdosis[7].

[7] Robert Alessi, *ekdosis - Typesetting TEI xml Compliant Critical Editions*, 2020. https://mirrors.ctan.org/macros/luatex/latex/ekdosis/ekdosis.pdf.

XML, TEI et graphes dans le corpus Ichtya. Le traitement, la visualisation et l'analyse des noms de poissons et créatures aquatiques

Marie Bisson
Maison de la recherche en sciences humaines, USR3486
Université de Caen - Normandie, France

Brigitte Gauvin
Centre Michel de Boüard, CRAHAM, UMR6273
Université de Caen – Normandie, France

Pierre-Yves Buard et **Barbara Jacob**
Maison de la recherche en sciences humaines, USR3486
Université de Caen - Normandie, France

Présentation du programme de recherche

Le corpus *Ichtya* (B. Gauvin et T. Buquet resp.) est intégré au corpus CAHIER depuis sa création. Il est articulé en 3 volets[1] :

- des éditions de textes multisupports avec traduction[2] ;
- la bibliothèque numérique *Ichtya*[3] ;
- un thesaurus numérique des noms de poissons et créatures aquatiques[4].

Le groupe *Ichtya* rassemble des chercheurs et ingénieurs du Centre Michel de Bouärd[5] et des ingénieurs du pôle « Document numérique[6] ». Les réalisations de ce groupe de travail sont le résultat d'un travail étroit entre les deux équipes.

Problématique de la communication

Notre communication se propose d'insister plus particulièrement sur la méthodologie appliquée au traitement de l'ichtyonymie dans le corpus. Il s'agit de montrer les liens étroits entre les technologies choisies, la matière scientifique et les problématiques de recherche qui ont conduit le groupe *Ichtya* à construire un outil de recherche particulier : le thesaurus des noms de poissons et créatures aquatiques.

Présentation de la bibliothèque

La bibliothèque *Ichtya* est une bibliothèque numérique qui rassemble des textes latins consacrés à l'ichtyologie qui furent publiés dans l'Antiquité, au Moyen Âge et à la Renaissance. Elle

---

1 Ceux-ci sont complétés par une bibliographie sur Zotero.org : https://www.zotero.org/groups/ichtya/items.

2 La première édition a été publiée en 2013 : B. Gauvin, C. Jacquemard et M.-A. Lucas-Avenel (éd.), *Hortus sanitatis : Livre IV, Les Poissons*, Caen, Presses universitaires de Caen (Fontes et paginae), 2013. Consultable en ligne : https://www.unicaen.fr/puc/sources/depiscibus/. Devraient suivre les éditions du livre 24 du *De animalibus* d'Albert le Grand et les livres 6 et 7 du *De natura rerum* de Thomas de Cantimpré.

3 La bibliothèque *Ichtya* est accessible à l'adresse suivante : https://www.unicaen.fr/ichtyalab/bibliotheque/accueil. Elle est partiellement publique.

4 Le thesaurus des poissons et créatures aquatiques est accessible en intégralité à l'adresse suivante : https://www.unicaen.fr/ichtyalab/thesaurus/accueil. Il est alimenté régulièrement au fur et à mesure de l'indexation du corpus de la bibliothèque *Ichtya*. Il comprend à la date de cette proposition de communication 2 290 entrées.

5 https://www.craham.cnrs.fr/

6 http://www.unicaen.fr/recherche/mrsh/document_numerique

s'inspire de la *Bibliotheca Ichtyologica* de Peter Artedi (1705-1735) et a pour vocation de mettre en ligne et à disposition des lecteurs un corpus latin consacré au savoir ichtyologique.

Le corpus, entièrement encodé en XML-TEI, est composé actuellement de 21 textes (8 en ligne publique) et 5 traductions (2 en ligne publique). Ces textes, attribués ou anonymes, sont de nature hétérogène (textes sacrés, encyclopédies, dialogues, poèmes) et de longueur très variable (fragments, chapitres, livres entiers).

La bibliothèque a été conçue d'emblée comme un corpus que les outils numériques permettraient de valoriser et de dynamiser, d'où le choix du langage XML pour permettre une interopérabilité entre les différents volets. Deux objectifs présidaient à l'entreprise : le souhait d'étudier le lexique des noms de poissons latins à travers les siècles ; et la volonté de cerner avec précision comment se fait la transmission des savoirs de l'Antiquité jusqu'à la Renaissance dans un domaine précis. Les outils permettant ce type de valorisation ont donc été choisis ou élaborés dans cette perspective. Le second objectif a été rempli par l'indexation systématique des sources : pour chaque œuvre, les sources ont été identifiées segment par segment et le lecteur, au fil de sa lecture, voit inscrite en marge l'origine du passage concerné et peut afficher la totalité du texte source pour se livrer à une comparaison. Le premier objectif a été atteint grâce à la constitution d'un thesaurus indépendant des citations[7]. En adoptant les usages du numérique, on arrive ainsi à valoriser le corpus et offrir de nouveaux outils pour la recherche.

Une méthodologie particulière pour l'indexation

Le corpus de la bibliothèque *Ichtya* ayant été traité de manière à permettre l'étude des noms de poissons latins à travers les âges, un index était donc indispensable pour permettre aux visiteurs de faire une recherche par nom de poisson ou de créature aquatique et d'accéder immédiatement à tous les passages du corpus *Ichtya* dans lequel ce nom apparaît.

Pour la première édition de texte (*De piscibus* de *l'Hortus sanitatis*[8]), la méthodologie d'indexation retenue était traditionnelle : un marqueur normalisé précédait le terme rencontré[9]. Pour la bibliothèque *Ichtya*, au lieu d'être pensée à plat, texte après texte, terme après terme, la méthodologie d'indexation a été conçue sous forme de thesaurus indépendant[10].

Le thesaurus

Le thesaurus est construit en XML-TEI. Il est composé d'autant de fichiers XML (notices) que de formes latines (au nominatif) ou vernaculaires rencontrées dans le corpus de la bibliothèque *Ichtya*. Chaque notice présente toujours la référence précise à la source dans laquelle le terme apparaît. Cette indication de source s'accompagne, autant que possible, d'une ou plusieurs identifications et, pour les appellations latines et grecques, de la référence scientifique qui valide ces identifications. Ces identifications peuvent être accompagnées d'une note de commentaire. Les notices peuvent aussi présenter deux sortes de renvois sous forme de liens : d'une part à la forme principale en cas de paronymie, de variante orthographique ou de forme vernaculaire, indication qui figure en tête de la fiche, à la place de l'identification ; de l'autre aux autres termes désignant le même animal sous un autre nom. L'indexation par le biais du format XML permet de faire des liens directement d'une forme à l'autre. Ce thesaurus fournit un outil de première utilité pour l'étude des synonymies et polyonymies entre les noms de poissons dans les traités ichtyologiques.

---

7 Le répertoire constitué pour l'édition du *De piscibus* est consultable à cette adresse : https://www.unicaen.fr/puc/sources/depiscibus/citations.

8 B. Gauvin, C. Jacquemard et M.-A. Avenel (éd.), *Hortus sanitatis : Livre IV, Les Poissons*, Caen, Presses universitaires de Caen (Fontes et paginae), 2013. Consultable en ligne : https://www.unicaen.fr/puc/sources/depiscibus/

9 Le template renseignant la forme normalisée avait été inséré devant chaque terme qu'on voulait retrouver dans l'index.

10 Après sa publication aux Presses universitaires de Caen, l'indexation XML du texte du *De piscibus* a donc dû être mise à jour pour intégrer la bibliothèque *Ichtya*.

Chaque forme de nom de poisson ou créature aquatique rencontré dans le corpus de la bibliothèque *Ichtya* fait donc l'objet d'une notice XML-TEI et chaque occurrence est liée à une notice du thesaurus[11].

Résultats textuels obtenus

Cette méthodologie nous a permis :

- de générer l'index de la bibliothèque *Ichtya* : l'index peut être mis à jour au fur et à mesure de son alimentation publique et permettre ainsi la consultation transversale de la bibliothèque *Ichtya*[12] ;
- de donner accès dynamiquement à une notice de thesaurus lors de la lecture d'un texte (chaque terme indexé est signalé par la couleur et la notice est accessible au clic) ;
- de proposer un site *Thesaurus* indépendant permettant d'accéder à l'ensemble des notices créées pour le corpus *Ichtya* dans son intégralité, via le sommaire, le moteur de recherche ou encore les liens internes entre chaque notice.

L'utilisation du langage XML à la fois pour les éditions, la bibliothèque *Ichtya* et le thesaurus de noms de poissons et créatures aquatiques, au moyen des recommandations de la Text Encoding Initiative (TEI), permet une interopérabilité totale. Chaque élément indexé dans une édition de texte, dans le corpus de la bibliothèque *Ichtya* ou dans l'index des zoonymies se trouve ainsi immédiatement traité et actif dans les données des deux autres supports : un nom de poisson indexé dans un texte de la bibliothèque se trouve immédiatement relié à une fiche dans le thesaurus.

Exploration graphique

La constitution indépendante du thesaurus nous a également permis de proposer une lecture du thesaurus sous forme de graphes dynamiques. À partir de l'encodage en arbre XML-TEI, des graphes ont pu être générés pour chaque notice, permettant de mettre en évidence les liens établis par les chercheurs : identification ; variantes graphiques ; notices en relation.

La chaîne de traitement développée s'appuie donc sur les informations scientifiques contenues dans les notices du thesaurus et en particulier sur les liens construits pendant les phases d'annotation. Le langage RDF permet d'exprimer explicitement ces liens sous la forme de relation entre les deux poissons (entre la notice de départ du lien et sa destination). Une fois le graphe RDF produit, un générateur de diagramme est appliqué pour produire une forme graphique intégrée au site du thesaurus. Cette intégration permet d'examiner un réseau à différentes échelles en proposant un système de zoom, de mettre en lumière un poisson et ceux qui lui sont directement liés ou encore de consulter une fiche depuis un diagramme.

La constitution de réseaux de notices permet aussi de modifier l'unité, ou le grain, de consultation. En effet, le lecteur peut examiner l'ensemble des poissons entretenant des relations de quelque nature que ce soit (traduction, variante, identification, etc.). Les diagrammes permettent en définitive de consulter l'ensemble d'un réseau de poissons liés les uns aux autres. Cette modification de l'unité de consultation à travers la mise à disposition de visualisations de réseaux de notices permet de faciliter l'étude de la circulation des savoirs : le fait de regrouper dans un même graphe « synoptique » toutes les notices connectées les unes aux autres peut permettre aux chercheurs d'identifier de nouveaux liens qui pourraient leur avoir échappé pendant les phases d'annotation.

D'un point de vue informatique et documentaire, il s'agit aussi d'explorer les limites du modèle de données arborescent (XML) à travers une expérimentation directe sur le terrain

[11] Le nom du poisson dans le texte est encodé au moyen de l'élément term. L'élément est qualifié d'une valeur de langue (attribut @xml:lang) et d'une référence à sa notice (identifiant de la notice en valeur d'attribut @ref).

[12] Voir ainsi l'index généré pour l'état actuel public de la bibliothèque *Ichtya* : https://www.unicaen.fr/ichtyalab/bibliotheque/index_poissons.

scientifique. On pourra alors évaluer les solutions de passage d'une organisation des données en arbre à un modèle plus expressif en tirant parti de la finesse de l'annotation mise en place par les spécialistes pendant les phases d'étude. En effet, le modèle de graphe RDF permet d'expliciter, en plus des parties de textes habituellement annotées en XML, les relations entretenues par ces différents textes. Là où le XML permet d'exprimer la seule relation contient/est contenu par, le RDF n'offre aucune limitation dans la caractérisation des relations entre les éléments qui compose un graphe. Il s'agit pour nous d'exploiter non seulement les parties de textes annotées pendant le travail de recherche, mais aussi, et peut-être surtout, les relations tracées entre ces parties de textes en particulier à des fins de visualisation.

Par ailleurs, en articulant les visualisations produites à partir des graphes RDF et les interfaces textuelles exploitant des textes encodés en XML-TEI, nous présentons une solution tirant parti des deux modèles de données sans surcoût d'encodage manuel.

Enfin, cet enrichissement sémantique des données permettra, à terme, de proposer l'interrogation du corpus sur les relations existant entre les différents poissons constitutifs du corpus. L'exploitation des types de relations pourra, par exemple, permettre la création d'un sous-corpus composé uniquement des noms de poissons, latins et vernaculaires, se référant à un même animal, à travers les siècles et les langues.

Nous proposons ici une méthode d'exploitation pour l'enrichissement des interfaces de consultation. Il s'agit, à partir des instances XML, de reprendre l'analyse réalisée par les chercheurs pour en extraire le réseau sous-jacent et le rendre visualisable et manipulable en explicitant les relations implicites existant entre les notices de poissons. Notre méthode permet d'exploiter ce réseau en tant que tel du point de vue informatique et documentaire, ce qui est, en définitive, rarement le cas dans ce type de projet ; or, les intérêts pour la recherche et l'étude des textes anciens sont nombreux.

Notre communication se propose de présenter plus en détail :

- le corpus et les objectifs de recherche du groupe *Ichtya* en termes d'ichtyonymie ;
- les notices du thesaurus en TEI et l'établissement du lien fait entre elles et les textes de la bibliothèque *Ichtya* : on montrera l'apport de cette méthodologie pour encoder et analyser le corpus ;
- la génération des graphes dans l'objectif de proposer une visualisation non textuelle et ce que cela apporte d'un point de vue documentaire ;
- les principes d'articulation des modèles de données en fonction des besoins et sans annotation manuelle supplémentaire.

Références

Bisson Marie, Gauvin Brigitte et Jacob Barbara, *Environnement d'édition scientifique en XML-TEI utilisé dans le cadre du programme Ichtya pour encoder les compilations médiévales*, Documentation du Pôle Document numérique, 2020. Consultable en ligne : http://www.unicaen.fr/recherche/mrsh/sites/default/files/public/document_numerique/manuel_ichtya.pdf.

Bisson Marie, Gauvin Brigitte, Jacquemard Catherine, *Rédiger une notice pour le thesaurus des créatures aquatiques du corpus Ichtya*, Documentation du Pôle document numérique, 2018. Consultable en ligne : http://www.unicaen.fr/recherche/mrsh/sites/default/files/public/document_numerique/manuel_ichtyonymie.pdf.

Bisson Marie, Goloubkoff Anne, « Les notices d'autorité en XML-TEI : un outil pour l'accroissement collaboratif de connaissances et l'indexation d'éditions de sources », *Tabularia*, 2020, 'Les sources des mondes normands à l'heure du numérique'. Consultable en ligne : https://doi.org/10.4000/tabularia.4176.

Buard, Pierre-Yves, « Le réseau de la baleine ou la visualisation de l'histoire d'un texte », dans *Inter litteras & scientias. Recueil d'études en hommage à Catherine Jacquemard*, éd. Brigitte Gauvin et Marie-Agnès Lucas-Avenel, Caen, Presses Universitaires de Caen (Miscellanea), 2019, p. 185-198.

Jacquemard Catherine, Gauvin Brigitte, Lucas-Avenel Marie-Agnès (ed.), avec la collaboration de C. Février et F. Lecocq, *HORTVS SANITATIS, Livre IV, Les poissons*, Caen, Presses universitaires de Caen (Fontes & Paginae), 2013. Consultable en ligne : http://www.unicaen.fr/puc/sources/depiscibus/accueil.

Kummer, Robert, « Semantic Technologies for Manuscript Descriptions - Concepts and Visions », dans *Codicology and Palaeography in the Digital Age*, 2 éd. Franz Fischer, Christiane Fritze, et Georg Vogeler, Books on Demand, Norderstedt, 2010, p. 133-154.

La Base de français médiéval et le consortium CAHIER. Dix ans d'échanges et de collaborations

Alexei Lavrentiev
IHRIM, UMR5317
CNRS, France

Céline Guillot-Barbance
IHRIM, UMR5317
École Normale Supérieure de Lyon, France

Le projet de la Base de français médiéval (BFM, http://txm.bfm-corpus.org) fait partie des membres fondateurs du Consortium CAHIER. Les origines du projet remontent à la fin des années 1980 et son évolution a suivi, et parfois anticipé, les grandes tendances du développement des humanités numériques. L'expérience de la BFM a permis de contribuer à plusieurs chantiers du consortium CAHIER : l'accès libre aux données, les normes d'encodage (et notamment l'usage de la TEI), la typologie textuelle, la mise en place de chaînes éditoriales ouvertes. Les échanges qui se sont produits dans les groupes de travail et lors des ateliers du consortium permettent à leur tour d'améliorer les pratiques d'encodage et les outils proposés aux utilisateurs de la BFM et d'assurer une plus grande interopérabilité et pérennité des données.

La constitution de la BFM a commencé par la numérisation de l'édition de la *Queste del saint Graal* d'A. Pauphilet (1923) et la base a été enrichie au fil des ans grâce à des vacations, aux contributions de doctorants de Ch. Marchello-Nizia, aux échanges avec des collègues et, plus récemment, grâce à des financements ANR. A ce jour, la BFM comprend 170 textes composés entre le 9e et le 15e siècle, soit près de 4,7 millions de mots. Pour la *Queste del saint Graal* l'édition de Pauphilet a d'ailleurs été remplacée dans le corpus par une édition numérique originale (Marchello-Nizia et Lavrentiev 2019). Une augmentation importante du corpus est prévue en 2021. Les textes de la BFM sont étiquetés en morphosyntaxe et lemmatisés (avec ou sans vérification) et bénéficient du balisage XML-TEI enrichi. En particulier, le discours direct est balisé dans l'ensemble du corpus, ce qui permet de mener des recherches sur l'oral représenté (Guillot-Barbance et al. 2018). De nombreuses thèses et travaux de recherche ont été réalisés grâce aux données de la BFM. Notamment, la partie médiévale du corpus de la *Grande grammaire historique du français* (Marchello-Nizia et al. 2020) est entièrement issue de la Base de français médiéval. La BFM est accessible en ligne grâce au logiciel « portail TXM » (http://textometrie.org), les textes peuvent être consultés librement et l'accès au moteur de recherche et d'analyse est donné gratuitement sur simple inscription (Guillot-Barbance et al. 2017).

Quand la TGIR Corpus (prédécesseur d'HumaNum) lance en 2010 l'appel à la création de consortiums de corpus, la BFM faisait déjà, depuis 2004, partie du Consortium international pour les corpus de français médiéval (CCFM, http://ccfm.ens-lyon.fr). Cette organisation informelle, n'ayant jamais bénéficié d'un financement spécifique et dont l'activité s'est estompée après 2008, a néanmoins permis d'entamer la réflexion et de publier des documents de travail sur les normes communes d'encodage et de description des textes, ainsi que sur les conditions d'accès aux corpus et d'échange de données. L'expérience du CCFM a sans doute inspiré l'initiative de la TGIR Corpus et a servi de point de départ pour certaines activités de CAHIER (notamment pour l'organisation de groupes de travail et pour la rédaction de guides de bonnes pratiques).

L'un des premiers groupes de travail de CAHIER visait à traiter les questions juridiques liées aux droits d'auteurs (et d'éditeurs) et à la mise à disposition de corpus. La BFM avait commencé comme un ensemble de concordanciers échangés dans un cadre privé entre chercheurs sous la forme de CD-ROM ou de tirages papier, puis elle avait progressivement ouvert l'accès à

l'interrogation et au téléchargement du corpus sur Internet. Elle avait une longue histoire de relations complexes avec les éditeurs commerciaux. Denise Pierrot, qui s'est occupée des questions juridiques pour la BFM, a également joué un rôle important dans le groupe de travail correspondant de CAHIER et a contribué à la rédaction du Guide des bonnes pratiques. La situation juridique a évolué suite aux décisions de justice dans le procès Droz contre Garnier numérique (2014 et 2017), ce qui a rendu possible la mise à disposition libre de textes historiques (hors apparat critique).

La pérennisation des données, grâce notamment à l'usage de l'encodage XML-TEI pour les textes et les annotations, a été la priorité pour la BFM depuis le début des années 2000 (Guillot et Heiden 2002). Le consortium CAHIER a pu bénéficier de la documentation de la BFM relative à l'encodage TEI du corps du texte et de l'entête (teiHeader) des documents. La BFM a été l'un des premiers projets à se conformer au modèle élaboré pour l'outil Web-OAI de CAHIER permettant le moissonnage des métadonnées (http://weboai.cahier.huma-num.fr). Toujours dans le domaine des métadonnées, les descripteurs typologiques de BFM, tels que le genre ou le domaine du texte, ont servi de base au thésaurus élaboré par le groupe de travail « Typologie textuelle " de CAHIER. Ce thésaurus, très riche et soigneusement structuré, permettra à son tour de préciser les métadonnées des futurs corpus de la BFM et de faciliter l'analyse de données de corpus agrégés à partir de plusieurs sources.

La pérennisation des données de recherche conformément aux principes FAIR (Findability, Accessibility, Interoperability and Reusability) est l'activité du Consortium CAHIER fortement encouragée par la TGIR Huma-Num. Même si la vision qui semble se dégager des recommandations du conseil scientifique d'Huma-Num et qui consiste à exiger avant tout le dépôt des données dans l'outil Nakala nous paraît un peu réductrice, il est certain que l'archivage pérenne et l'accessibilité des données sont extrêmement importants. Le soutien que CAHIER assure pour le dépôt des textes et des images dans NAKALA est très précieux pour la BFM.

La formation aux outils d'édition, d'analyse et de publication de corpus ouverts, ainsi que l'échange de bonnes pratiques éditoriales, a toujours été l'une des principales activités de CAHIER. La BFM, qui se développe en étroite collaboration avec la plateforme de préparation, d'analyse et de préparation de corpus TXM (Heiden et al. 2010) a pu partager son expérience, et des membres de l'équipe BFM ont animé de nombreuses séances de formation lors des ateliers CAHIER.

Le consortium CAHIER a également favorisé les échanges entre l'équipe TXM et le Pôle document numérique de la MSH de Caen qui développe la chaîne éditoriale Métopes (grâce notamment au financement d'un stage en 2017). La BFM bénéficie actuellement de certains éléments de Métopes pour la mise en page de ses éditions au format PDF et des scripts de traitement automatique permettant d'importer dans TXM des documents XML-TEI créés avec Métopes ont été élaborés.

La BFM est un projet qui a commencé bien avant la création du Consortium CAHIER et qui va sans doute continuer à se développer après la disparition de CAHIER dans sa forme actuelle de consortium de la TGIR Huma-Num. Quel que soit l'avenir du Consortium, nous sommes convaincus que les ressources numériques, les méthodes et les outils de travail élaborés grâce ou avec le soutien de CAHIER ainsi que les relations humaines et les partenariats de recherche qui se sont tissés au cours des dix ans de ses activités continueront à jouer un rôle important dans la communauté des humanités numériques.

Références

Guillot-Barbance, Céline, Heiden, Serge et Lavrentiev, Alexei, « Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique », *Diachroniques*, 7, 2017, p. 167-184.

Guillot-Barbance, Céline, Alexei Lavrentiev, Serge Heiden et Bénédicte Pincemin. « Diachronie de l'oral représenté : délimitation et segmentation interne du dialogue (IX^{e}-XV^{e} siècle) », in Wendy Ayres-Benett, Anne Carlier, Julie Glikman, Thomas Raindford, Gilles Siouffi et

Carine Skupien Dekens (éds.) *Nouvelles voies d'accès au changement linguistique. Actes du colloque de la SIDF*, Paris, Classiques Garnier, 2018, p. 279-296.

Marchello-Nizia, Christiane, Combettes, Bernard, Prévost, Sophie et Scheer, Tobias (éds.), *Grande grammaire historique du français*, Berlin, De Gruyter, 2020.

Heiden, Serge, Jean-Philippe Magué et Bénédicte Pincemin, « TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement », in S. Bolasco, I Chiari et L. Giuliano (éds.) *Proceedings of the 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*, Rome, Edizioni Universitarie di Lettere Economia Diritto, 2010, p. 1021-1032.

Heiden, Serge et Barbance-Guillot, Céline, « Capitalisation des savoirs par le web : une application de la TEI pour l'encodage et l'exploitation des textes de la Base de Français Médiéval », in P. Kunstmann, F. Martineau, et D. Forget (éds.) *Ancien et moyen français sur le Web : enjeux méthodologiques et analyse du discours [Actes du colloque d'Ottawa, 4-5 oct. 2002]*, Ottawa, Éditions David, 2003, p. 77-92.

Marchello-Nizia Christiane et Lavrentiev, Alexei (éds.), *Queste del saint Graal*, Lyon, ENS de Lyon, 2019.

Pauphilet, Albert (éd.), *La Queste del Saint Graal. Roman du XIII^e^ siècle,* Paris, Champion, 1923.

Des données au corpus.
L'exploitation numérique des mazarinades

Karine Abiven et **Gael Lejeune**
STIH, EA 4509
Université Paris-Sorbonne, France

Le projet « Antonomaz[1] » (ANalyse auTOmatique et NumérisatiOn des MAZarinades) se propose d'explorer un ensemble de brefs imprimés parus en France lors de la Fronde (1648-1653) - entre 5000 et 5500 unités, environ 68 000 pages. Utilisés par des communautés multiples (linguistiques, historiens, littéraires), ces écrits posent pourtant des problèmes d'accès spécifiques : comme ces petits livrets étaient peu couteux, et produits en masse, ils ne sont pas rares, et les exemplaires sont nombreux, y compris numériques (d'Europeana à Google Livre en passant par les bibliothèque numériques Gallica ou Mazarinum, que le présent projet contribue à abonder en fac-similés numériques de haute qualité). Mais pour que ces textes soient exploitables, les tâches demeurent nombreuses, à des niveaux disciplinaires divers :

- au plan de la linguistique de corpus (il reste à les constituer en corpus cohérents) ;
- de la bibliographie matérielle et de la philologie numérique (documentation de métadonnées floues pour des pamphlets souvent anonymes, et imprimés à la va-vite) ;
- du TAL (élaboration de méthodes efficaces d'océrisation appropriées pour ces documents anciens ; classification, supervisée ou non, qui permet par exemple d'obtenir des cluster de textes ainsi mieux compréhensibles par de nouveaux contextes) ;
- des humanités numériques (par exemple, imaginer des visualisations originales qui puissent permettre de mettre en réseaux des textes d'actualité, et donc intrinsèquement dépendants de leur contexte).

Nous espérons obtenir différents résultats :

- abonder en métadonnées nouvelles la Base Bibliographique de la Bibliothèque Mazarine (datation des documents, atelier d'imprimerie d'origine, parti politique d'origine) ;
- construire et partager des chaines de traitements pour diverses tâches (océrisation de PDF anciens, annotation automatique, repérage d'entités nommées, etc.) ;
- fournir aux diverses communautés intéressées une base de données requêtable et téléchargeable.

[1] https://cahier.hypotheses.org/antonomaz

La Bibliothèque dramatique. L'édition numérique d'un corpus de pièces de théâtre du XVII^{e} siècle

Amélie Canu
Centre « Léon Robin », UMR8061
CNRS, France

Claire Carpentier
CELLF, UMR8599
CNRS, France

Ce poster présentera les étapes d'un projet d'édition numérique d'un corpus de pièces de théâtre du XVII^{e} siècle, à travers le prisme de l'expérience d'un ingénieur dans un laboratoire de recherche. Débuté en 2013 à l'initiative de Georges Forester, professeur à Paris IV, le projet Bibliothèque dramatique a été lancé en 2014 ; la base a été étoffée au fil des années jusqu'à compter aujourd'hui 125 pièces. Huit ans après le début du projet, elle vit toujours et constitue un objet éditorial dynamique, utilisé par les étudiants et les chercheurs.

Première partie : la genèse du projet

Un ingénieur. Un chercheur. Une discussion. De cette discussion naîtra une série de questions et de doutes pour l'ingénieur isolé dans son laboratoire de recherche : de quelles ressources dispose-t-il ? À qui s'adresser ? Comment assurer la pérennité du projet ?

Sortir de cet enclavement permet à l'ingénieur de dissiper ses doutes. Il se tourne alors vers la communauté de l'édition numérique (Médici, Cahier...), commence à poser des questions, à chercher des appuis (Labex voisin, universités traitant des corpus similaires...). Le projet se développe alors comme un échange de connaissances où chacun trouvera une place bien spécifique, en fonction de ses possibilités.

Deuxième partie : la réalisation technique

Pour encoder, une évidence s'impose alors à l'ingénieur : il va lui falloir traverser un océan de balises. La documentation de la TEI est colossale ; les possibilités de lecture vastes ; les lacunes techniques présentes. Et là une révélation pour l'ingénieur : il doit mieux comprendre les attentes du chercheur pour mener sa barque dans cet océan. C'est uniquement grâce à cela qu'il pourra repérer les balises qui lui seront utiles et réaliser des choix importants. La relation chercheur-ingénieur se noue et grandit autour des choix d'encodage. Au début, il faut tâtonner : certains choix d'encodage ne seront pas retenus, d'autres, une fois testés, s'avèreront réellement importants et efficaces.

Troisième partie : cycle de vie d'une édition numérique

Une fois les choix arrêtés, l'encodage de chaque pièce est de plus en plus facile et la base peut être lancée et révélée au public, ce qui est une source réelle de satisfaction pour tous les participants. Huit ans plus tard, elle vit encore aujourd'hui grâce à ceux qui ont repris le flambeau, s'assurant que le site puisse continuer à être visible au fil des ans. Cela nécessite de réels moyens, à ne pas négliger au moment de se lancer.

Marc Michel Rey ou l'invention d'un corpus

Fabienne Vial-Bonacci
IHRIM, UMR5317
CNRS, France

Christelle Bahier-Porte
IHRIM, UMR5317
Université Jean Monnet - Saint-Etienne, France

Lorsque Jeroom Vercruysse, professeur à l'université de Bruxelles, a légué au laboratoire IHPC (désormais IHRIM) les documents qu'il a réunis pendant plusieurs décennies autour de la correspondance et des archives du libraire Marc Michel Rey, s'est d'emblée posée une question de méthode : de traitement (des milliers de pages photocopiées, de notes manuscrites), d'exploitation scientifique, et de diffusion au public. Sous ce titre un peu provocateur, nous voudrions faire le point avec cinq ans de recul, sur ce qui est devenu le projet Marc Michel Rey (http://rey.huma-num.fr/presentation). Ce sont en effet les possibilités offertes par l'édition numérique qui ont permis que cet ensemble de papiers disparates puisse constituer un corpus, certes ouvert et évolutif, mais structuré, exploitable et donc - c'était la première mission confiée par J. Vercruysse -, diffusable au public de chercheurs.

Il ne s'agira pas seulement de retracer brièvement les différentes étapes de ce projet (établissement d'un inventaire critique, protocole d'édition, indexation, alignement, fairisation, visualisation, diffusion...) mais surtout de montrer comment chaque étape « technique » a fait évoluer la notion même de corpus telle qu'elle avait pu être imaginée en amont du projet. Le projet n'est pas achevé et nous tenons à lui conserver sa nature expérimentale. Celle-ci se caractérise par le partage de méthodologies au sein du laboratoire, avec l'appui du pôle Humanités numériques. En outre, le projet d'édition numérique des archives et de la correspondance de ce libraire a fait naître un projet corollaire de base de données et d'analyse par les outils de la vision par ordinateur des ornements utilisés par le libraire dans une optique d'authentification des ouvrages publiés par ce libraire[1]. Si ce projet a sa propre autonomie, il a pour vocation de s'articuler étroitement à l'édition numérique de la correspondance, techniquement et intellectuellement. Cette extension récente du projet ouvre encore la notion de corpus qui prend, par exemple, un tout autre sens pour des chercheurs dans le domaine de l'image et de la vision par ordinateur. La dimension expérimentale, que nous voudrions mettre en avant dans cette présentation, se décline alors dans le champ disciplinaire dont relève cette recherche. Convoquant l'histoire du livre (bibliographie matérielle, censure, commerce...), l'histoire des idées, l'étude des réseaux, la littérature, le projet Marc Michel Rey entend contribuer au renouvellement des méthodes de l'histoire des idées et de l'histoire du livre dont l'édition numérique a profondément bouleversé les pratiques ces dernières années.

Le consortium Cahier a accompagné l'évolution de ce projet par le partage des compétences qu'il permet[2] et plus spécifiquement à deux moments clés de sa jeune histoire : le début de l'encodage en XML-TEI et au moment de la prospective pour le développement du volet « Ornements » du projet. Ce sera pour nous l'occasion de témoigner de la pertinence et de l'efficacité de tels lieux d'échanges et d'accompagnement des projets d'édition numérique.

---

1 Projet ROIi (Rey's Ornament Image investigation) financé par l'ANR (2020-2024) en partenariat avec le Laboratoire Hubert Curien (UMR 5516)

2 Et notamment au moment de la rédaction du guide de l'édition numérique des correspondances.

Projet MeThAL : ressources numériques pour une relecture du théâtre en alsacien

Pablo Ruiz Fabo
Carole Werner
Delphine Bernhard
Pascale Erhart
Dominique Huck
LiLPa, EA 1339
Université de Strasbourg, France

Le projet MeThAL[1] part du constat que le manque de ressources numériques appropriées rend difficiles les études quantitatives en analyse dramatique du théâtre dialectal et en sociolinguistique historique de l'Alsace. Nous développons un corpus encodé en TEI de pièces en alsacien (de 1870 à 1914), disponible publiquement[2] et visant des pratiques FAIR (Wilkinson et al., 2016). Notre source principale est la collection de 150 pièces numérisée en mode image par la BNU[3]. La valorisation du corpus à travers une interface d'exploration est aussi visée.

Plusieurs travaux et projets témoignent de l'intérêt des méthodes assistées par ordinateur pour l'analyse théâtrale, fondées sur de larges corpus numériques. Des numéros spéciaux de la *Revue d'Historiographie du Théâtre* (2017) et du *Theatre Journal* (2016) en fournissent des exemples ; cf. leurs introductions respectives (Galleron, 2017a ; Tompkins, 2016). D'autres projets proches de nos intérêts sont DraCor (Fischer & Börner, 2019), QuaDramA[4] ou les ressources de Dramacode et le site « Théâtre classique[5] ». Notre projet veut constituer un premier pas pour rendre possibles de telles ressources et analyses pour le théâtre en alsacien. Nous présenterons le travail effectué vers ces objectifs et les défis existants.

Nous avons commencé l'océrisation et l'encodage TEI du corpus (chaîne éditoriale décrite dans Ruiz Fabo et al., 2020) ; les premières pièces sont en ligne sur notre dépôt git, sous licence ouverte [2]. La FAIRisation du corpus sera complétée par son dépôt sur une plate-forme d'exposition de données comme Nakala (cf. recommandations du consortium CAHIER, Idmhand & Galleron, 2020). Les pièces encodées sont également disponibles sur la plate-forme DraCor[6].

Afin d'obtenir un premier aperçu du corpus, nous avons transcrit le *dramatis personae* de 109 pièces (1.091 personnages), annoté le sexe, âge et profession des personnages si disponibles, et donné une estimation de leur classe sociale[7]. Ces métadonnées aident à extraire des interactions entre des personnages définis par différentes variables sociales, ce qui permet d'examiner les sujets abordés et le langage utilisé en fonction de ces variables. Les métadonnées sociales aident aussi à observer la représentation de divers groupes socio-politiques dans les pièces et son évolution temporelle. Le corpus pourrait ainsi aider à quantifier des tendances sur ces aspects identifiées par des études précédentes effectuées sans le support d'un corpus numérique (Gall, 1974 ; Huck, 1998, 2005 ; von Hülsen, 2003).

Un de nos défis consiste à créer une ressource représentative : le corpus couvre prioritairement la production strasbourgeoise et la numérisation et l'encodage de pièces d'autres

1 https://methal.pages.unistra.fr/
2 Le dépôt (https://git.unistra.fr/methal/methal-sources) est mis à jour graduellement.
3 Bibliothèque nationale et universitaire de Strasbourg (portail Numistral) : https://numistral.fr/.
4 Quantitative Drama Analysis : https://quadrama.github.io/.
5 Dramacode : https://github.com/dramacode; Théâtre Classique : https://www.theatre-classique.fr/
6 https://dracor.org/als
7 Ces métadonnées sont dans une base de données et leur formalisation dans le corpus en TEI (cf. Galleron, 2017b) est prévue.

origines (cf. Gall, 1974) sont nécessaires. Un deuxième défi est posé par la variété dans la scripturalisation de l'alsacien, qui demande une identification automatique de variantes orthographiques comme pré-requis pour la comparaison du contenu des pièces via le *topic modeling* ou la textométrie. Ayant conscience de ces défis, le projet a la volonté de poser une première pierre pour l'analyse du théâtre en alsacien sous un angle inédit pour cette tradition.

Références

Fischer, F., Börner, I., « Programmable Corpora : Introducing DraCor, an Infrastructure for the Research on European Drama », *Digital Humanities* 2019, 5, https://dev.clariah.nl/files/dh2019/boa/0268.html

Gall, J.-M., *Le Théâtre populaire alsacien au XIX^e^ siècle*, Istra, 1974.

Galleron, I., « Études théâtrales et humanités numériques : Une introduction », *Revue d'Historiographie du Théâtre : Études théâtrales et humanités numériques*, 4, 2017.

Galleron, I., « Conceptualisation of Theatrical Characters in the Digital Paradigm : Needs, Problems and Foreseen Solutions », *Human and Social Studies*, 6(1), 2017, p. 88-108. https://doi.org/10.1515/hssr-2017-0007

Huck, D., *D'r Herr Maire (1898) de Gustave Stoskopf. Entre ethnologie et littérature : Les Alsaciens en auto-représentation*, *Recherches Germaniques*, 28, 1998, p. 163-190.

Huck, D., « Le 'Théâtre Alsacien de Strasbourg' et la production dramaturgique de ses fondateurs (1898-1914) », in J.-M. Leveratto, J. Benay, O. Thomas, & S. Wuttke, *Culture et histoire des spectacles en Alsace et en Lorraine : De l'annexion à la décentralisation (1871-1946)*, Peter Lang, 2005, p. 198-222.

Idmhand, F., Galleron, I., *Guide pour la FAIRisation des données des corpus d'auteurs. Groupe de travail [Data_Cahier]. [Research Report]*, Huma-Num, 2020. https://halshs.archives-ouvertes.fr/halshs-02889777

Ruiz Fabo, P., Bernhard, D., & Werner, C., « Création d'un corpus FAIR de théâtre en alsacien et normalisation de variétés non-contemporaines », in *Deuxièmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT)*, 2020, p. 34-43. https://hal.archives-ouvertes.fr/hal-03047152

Tompkins, J., « Editorial Comment : Theatre, the Digital, and the Analysis and Documentation of Performance », *Theatre Journal*, 68(4), 2016, xi-xiv. https://doi.org/10.1353/tj.2016.0103

von Hülsen, B., *Szenenwechsel im Elsass : Theater und Gesellschaft in Strassburg zwischen Deutschland und Frankreich 1890-1944*, Leipziger Universitätsverlag, 2003.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, et al., « The FAIR Guiding Principles for scientific data management and stewardship », *Scientific Data*, 3(1), 2016, 160018. https://doi.org/10.1038/sdata.2016.18

La routine et le style.
Exploration outillée des formules d'ouverture et de clôture dans des correspondances peu-lettrées de la Première Guerre mondiale

Agnès Steuckardt
PRAXILING, UMR5267
Université Paul Valéry - Montpellier III, France

Sybille Große
Université de Heidelberg, Allemagne

Beatrice Dal Bo
LPL, UMR 7309
Aix-Marseille Université, France

Lena Sowada
Université de Heidelberg, Allemagne

L'exploration outillée des corpus, par l'extraction de concordances, de segments répétés, de cooccurrences, met en évidence les régularités textuelles. Comme en témoignent par exemple les analyses des rapports professionnels (Née, Sitri, Veniard, 2014) ou des discours de vœux (Leblanc, 2016), cette approche se révèle particulièrement pertinente pour l'étude de genres fortement codifiés : plus le corpus étudié présente de passages obligés, plus leur mise en série éclaire à la fois sur les régularités caractéristiques du genre et sur les spécificités stylistiques de chaque auteur. Qu'en est-il du genre épistolaire ? En tension entre convention et liberté d'écriture, la lettre est un poste d'observation de la dialectique entre le formulaire et le singulier.

Notre proposition centrera l'attention sur les moments les plus codifiés de la lettre : les formules d'ouverture et de clôture, que nous étudierons dans des lettres de familles peu-lettrées, écrites pendant la Première Guerre mondiale. Notre corpus (690.662 tokens) est composé de 2378 lettres et cartes postales dont les auteurs, issus de différentes régions françaises, se caractérisent par un niveau d'instruction élémentaire. Transcrit selon les standards de la TEI, il est interrogeable par le logiciel TXM. L'encodage des « opener » et « closer » en tant que séquences textuelles est prévu par la TEI, ce qui en facilite l'extraction ; l'utilisation d'expressions régulières dans les requêtes permet d'en identifier les patrons lexico-syntaxiques.

Que nous apprend l'exploration outillée des débuts et fins de lettre sur les formules épistolaires et sur le rapport qu'entretient avec elles le sujet d'écriture ? Peut-on confirmer l'hypothèse selon laquelle il y aurait « une grande variation dans les formules de début et plutôt une grande stabilité dans les formules de fin » (Walter, 2018 : 13) ? Laissent-elles s'exprimer un « style », comme l'avance Chantal Wionet, notant que « l'idée de littérature peut s'en trouver troublée » (2015 : 191) ? Après avoir situé notre approche des phénomènes de récurrence discursive, nous décrirons les routines épistolaires attestées par notre corpus et proposerons d'aborder les variations qu'introduisent certains épistoliers comme des modulations stylistiques.

1. Appréhender la récurrence discursive

Les disciplines qui s'occupent de langage s'intéressent depuis longtemps aux phénomènes de récurrence discursive et en proposent différentes catégorisations : topos des rhétoriciens, leitmotiv des stylisticiens, devenu motif, et repris par la linguistique de corpus (Ganascia, 2001 ; Quiniou et al., 2012), rituel et routine en analyse conversationnelle (Goffman, 1967 ; Coulmas,

1981) repris par l'analyse du discours (Gülich, 1997 ; Née, Sitri, Veniard, 2014), traditions discursives dans une perspective diachronique (Schlieben-Lange, 1983), patrons discursifs en grammaire des constructions (Östman 2005 ; Meier, 2020), ou encore phraséologie des lexicologues (Mel'cuk, 1998). Chacune de ces catégorisations, ancrée dans une approche spécifique, saisit des objets linguistiques sensiblement différents, du schéma argumentatif à la locution figée.

Nous plaçant dans une perspective d'analyse du discours, nous envisageons les récurrences des débuts et fins de lettre comme des unités discursives qui accomplissent des actes de langage et comme des séquences linguistiques qui présentent des similarités lexicales et syntaxiques. Les études sur les correspondances peu-lettrées les ont abordées par le terme de formule, dans une approche ethnologique à la recherche de modèles « préécrits » (Bruneton-Governatori, Moreux, 1997 : 82) ou textuelle, les décrivant comme des formules structurantes du texte (text-structure formulae) (Rutten, van der Wal, 2014 : 82). Le terme de rituel permet de prendre en considération la dimension interactionnelle et sociale de la lettre ; celui de routine, tout en conservant l'ancrage dans une praxis sociale, met en avant la récurrence linguistique : « Une routine discursive consiste en la mise en relation de séquences linguistiques récurrentes, partiellement figées (i. e. les patrons), avec des déterminations discursives et des fonctions textuelles propres à un genre ou une sphère d'activité » (Née, Sitri, Veniard, 2014 : 2119). À la croisée des approches interactionnelle, discursive, outillée, le terme de routine nous permettra ici d'appréhender les débuts et fins de lettre, à la fois en tant que réalisation linguistique à décrire, mais aussi - parce que routine renvoie aux processus à l'œuvre dans l'acte d'écriture - en tant que préconstruit linguistique que le scripteur s'approprie (Große, Sowada, 2020).

2. Routines épistolaires

On a identifié dans deux études antérieures les combinatoires récurrentes des ouvertures (Große et al., 2016 : 3) et des clôtures (Steuckardt et al., 2020 : 113).

2.1 Ouverture

Les formules d'ouverture s'inscrivent dans le schéma :

lieu / date / adresse / formule d'ouverture subjective.

Plus de la moitié des lettres comportent le lieu de la rédaction, et la quasi-totalité la date et l'adresse. Afin de produire la formule d'adresse, les scripteurs disposent d'éléments linguistiques qu'ils agencent de façon variée : le déterminant possessif de 1ère personne, des adverbes comme très, bien ; des adjectifs tels que cher, petit, aimé, adoré ; des noms communs qui explicitent la relation des deux épistoliers (mari/femme, époux/épouse, etc.) ou le nom propre du destinataire. Citons quelques patrons syntaxiques attestés :

- déterminant possessif + adv. + adj. + nom de relation + adj. :

(1) Ma chére Pette Fémme cherie (Paul, 13/02/1916)

(2) Ma chère petite femme aimèe (Félicien, 15/07/1917)

- adv. + adj. + nom de relation (+ nom propre) :

(3) Bien chère Epouse (Pierre, 05/09/1914)

(4) Bien chèr soeur Marie (Joseph Antoine, 27/03/1915)

- déterminant possessif + adv. + adj. + nom propre :

(5) Ma tres chère Marie (Pierre, 13/10/1914)

(6) Mon tres chèr Piérriliou (Marie, 28/11/1914)

Parfois, un effet de cumul peut être observé : Mon tres chér bien aimé petit mari (Marie, 25/12/1914).

La formule subjective qui suit l'adresse est moins figée, même si l'on trouve aussi certaines structures récurrentes, comme par exemple C'est avec [dét./adv./adj.] plaisir que... qui sert à accuser réception de correspondance ou se réfère à l'acte de répondre. La structure se retrouve dans les textes de différents scripteurs de nos corpus (194 occurrences), par exemple :

(7) Mes Chers Parents.

C'est avec le plus grand plaisir que je fais réponse à votre honorée du 3 écoulé (Joseph, 25/03/1917)

(8) Biens chers Parents,

C'est avec grand plaisir que j'ai reçu votre lettre du 16 Cnt qui (Paul, 19/02/1915)

2.2 Clôture

La partie de la clôture se caractérise essentiellement par une forme de salutation et de signature. Presqu'un tiers des lettres présentent le patron syntaxique :

Reçois + SN apostrophe + SN cod + SP + signature.

qui comporte souvent une subordonnée relative avant la signature (10) ou qui est introduit par une autre structure formulaire En attendant le plaisir de (11) :

(9) Reçoit ma bien chère femme les plus tendre amitiés de ton cher mari Arcis. F. (Félicien, 22/08/1914)

(10) Recois mon bien aimé mari de bons baisers de plain de tendresse da ta petite femme qui t'aime et pense sans cesse à toi pour la vie V. Arcis (Victoria, 08/12/1915).

(11) En attendant le plaisir de vous voir Reçevez Biens chers Parents un grand et tendre Baiser de votre fils qui vous aime pour la vie (Paul) (Paul, 21/09/1914)

On relève une certaine variété lexicale dans les éléments qui constituent les syntagmes apparaissant dans ce patron : par exemple, comme SN en fonction d'objet direct, on trouve les plus douces caresses, mille baisers, ma sincère affection, mes plus gros baisers, etc. La subordonnée relative permet une appropriation :

(12) reçoit ma petite femme les plus douces carresses et les plus tendres amitiés de ton mari qui met deux gros baiser sur la lettre. Arcis. F (Félicien, 04/10/1914)

(13) Reçois mon bien cher mari mes baisers les plus affectueux et mes amitié bien sinceres ta petite femme qui t'envoie un 420 de gros mimis et au petit Félicien pour la vie V. Arcis (Victoria, 18/09/1916)

La clôture de la lettre semble également être un lieu privilégié pour l'expression des sentiments. Juste avant ou après la signature, les scripteurs peuvent ajouter une dernière pensée, un dernier message d'affection pour leur conjoint :

(14) Recois mon bien aime mari de gros baisers de ta petite femme qui t'aime et t'envoie un 75 de bons mimis un gros mimi au petit Félicien sa petite Victoria voudrai bien le voir. pour la vie V. Arcis (Victoria, 30/09/1916).

(15) Recois ma bien aimèe petite femme de bons baisers affectueux ton petit mari qui t'aime et t'embrasse de tout coeur pour la vie. F Arcis le petit Félicien envoit ses meilleurs mimis à la petite Victoria et qu'il n'ai pas du souci il est toujours bien sage (Félicien, 26/05/1916)

Malgré la prégnance de l'amorce par reçois, la créativité individuelle des scripteurs trouve ainsi, dans cette partie finale, une place qui paraît plus grande qu'au début de la lettre.

3. Modulations stylistiques

Peut-on parler de « style » pour caractériser les formules des scripteurs peu-lettrés ? Par définition, les peu-lettrés ne maîtrisent pas complètement les normes linguistiques de l'écrit. Pour autant, leur écriture est-elle dépourvue de choix, manifestant une singularité, une sensibilité, voire une recherche d'expressivité ? On propose de ne pas considérer a priori la non-conformité aux standards de l'écrit normé comme un obstacle dirimant au style. Cependant le caractère stéréotypé des formules n'est-il pas en contraction avec la notion même de style ? Ce ne sera pas ici dans la perspective d'une description du genre, ni d'une évaluation d'après les canons de la tradition littéraire ou discursive que nous appréhendons le style, mais en tant que manière de s'exprimer propre à un individu. L'extraction des concordances, si elle met en évidence les routines relevées en deuxième partie, permet aussi le repérage de modulations, et la « récurrence de l'apparition de ces écarts » (Philippe, 2005 : 77).

On en prendra illustration dans une modulation apportée à la formule d'ouverture, pourtant fortement routinisée. Un des patrons de formule subjective, qui suit l'adresse, se présente sous la forme : Me/Nous + voici/voilà (54 occurrences). Son plus grand utilisateur, Jules Ramier, le mobilise à deux moments de sa correspondance. Il écrit d'abord :

(16) Ma chère Bien aimee Léonie Me voila de nouveau à toi pour te causer un peu sur ma santée laquelle est toujours parfaite (18/01/1915).

Trois lettres de janvier et février reprennent à l'identique le patron :

Me voilà de nouveau à toi pour + te + infinitif,

où « à toi » fonctionne comme un attribut, apportant une prédication sur le pronom me. Le propos est de signifier l'entière disponibilité du destinateur au destinataire. Deux ans plus tard, Jules reprend l'amorce « Me voilà », mais dans une construction plus surprenante :

(17) Bien Chère Epouse Me voila un petit moment de silence pour te tracer quelle que ligne (20/10/1917).

L'attribut à toi a disparu, comme si Jules choisissait d'en faire l'ellipse, la deuxième personne étant peut-être suffisamment instanciée par le pronom te. Il en résulte une incertitude sur l'interprétation syntaxique de l'énoncé. Il semble que me voilà forme un bloc présentatif équivalent à voilà. La prédication devient : voilà un petit moment de silence. Au tour figé signifiant la disponibilité au destinataire se substitue une notation plus personnelle, motivée par la situation du soldat, pris dans le vacarme de la guerre. Comme l'enquête textométrique permet de le montrer, l'évocation du bruit est récurrente chez Jules, qui décrit par exemple :

> (18) on se dirait toujours au 14 juillet les obus qui tonnent, les fusées qui éclairent le terrain les bombes qui éclatent (26/02/1915).

La modulation qu'il donne à la formule apparaît ainsi révélatrice de sa sensibilité propre.

En dégageant, grâce à une exploration outillée du corpus, les récurrences textuelles, on a pu identifier les routines épistolaires à l'œuvre chez les scripteurs peu-lettrés et en démontrer la prégnance, particulièrement au moment de commencer la lettre. Si l'analyse outillée permet ainsi de démontrer l'existence de préconstruits discursifs, caractéristiques du genre épistolaire, elle met cependant aussi en évidence, par les variations qu'elle dévoile, la capacité des scripteurs à sortir de la routine et à tracer leur cheminement singulier dans l'écriture.

Références

Bruneton-Governatori, Ariane, Moreux, Bernard, « Un modèle épistolaire populaire », Daniel Fabre (dir.), *Par écrit. Ethnologie des pratiques d'écriture quotidiennes*, Paris, Éditions de la Maison des Sciences de l'Homme, 1997, 79-103.

Coulmas, Florian (ed.), *Conversational routine : Explorations in standardized communication situations and prepatterned speech*, The Hague, Mouton, 1981.

Ganascia, Jean-Gabriel, « Extraction automatique de motifs syntaxiques », *Actes de TALN 2001*, Tours, 2-5 juillet 2001.

Goffman, Erving, *Interaction Ritual*, Londres, Cox and Wyman, 1967.

Große, Sybille, Sowada, Lena, « Socialisation écrite et rédaction épistolaire de scripteurs moins expérimentés - lettres des soldats de la Grande Guerre », *Romanistisches Jahrbuch* 71, 2020, p. 82-129. DOI 10.1515/roja-2020-0003.

Große, Sybille, Steuckardt, Agnès, Sowada, Lena, Dal Bo, Beatrice, « Du rituel à l'individuel dans les correspondances peu lettrées de la Grande Guerre », F. Neveu et al. (éds.), *Actes du 4e Congrès mondial de linguistique française*, EPD Sciences, 2016, p. 1-15. DOI 10.1051/shsconf/20162706008.

Gülich, Elisabeth, « Routineformeln und Formulierungsroutinen. Ein Beitrag zur Beschreibung formelhafter Texte », in R. Wimmer (éd.), *Wortbildung und Phraseologie. Studien zur deutschen Sprache*, Vol 9, Tübingen, Narr, 1997, p. 131-176.

Leblanc, Jean-Marc, *Analyses lexicométriques des vœux présidentiels*, Londres, ISTE éditions, 2016.

Meier, Kerstin, *Semantische und diskurstraditionelle Komplexität. Linguistische Interpretationen zur französischen Kurzprosa*, Berlin/Boston, de Gruyter, 2020.

Mel'cuk, Igor, « Collocations and lexical functions », A. Cowie (éd.), *Phraseology. Theory, Analyses, and Applications*, Oxford, Oxford University Press, 1998, p. 23-53.

Née, Émilie, Sitri, Frédérique, Vienard, Marie, « Pour une approche des routines discursives dans les écrits professionnels », F. Neveu et al. (éds.), *Actes du CMLF 2014 -- 4e Congrès mondial de linguistique française EDP Sciences*, 2014, p. 2113-2124.

Östman, Jan-Ola, « Construction discourse: a prolegomenon », J.-O. Östman, M. Fried, (éds.), *Construction Grammars. Cognitive Grounding and Theoretical Extensions*, Amsterdam, Benjamins, 2005, p. 121-144.

Philippe, Gilles, « Traitement stylistique et traitement idiolectal des singularités langagières », *Cahiers de praxématique*, 44, 2006, p. 77-92.

Quiniou, Solen, Cellier, Peguy, Charnois, Thierry, Legallois, Dominique, « Fouille de données pour la stylistique : cas des motifs séquentiels émergents », *Proceedings of the 11th International Conference on the Statistical Analysis of Textual Data*, Liege, 2012, p. 821-833.

Rutten, Gijsbert, van der Wal, Marijke J., *Letters as Loot. A sociolinguistic approach to seventeenth- and eighteenth- century Dutch*, Amsterdam / Philadelphia, Benjamins, 2014.

Schlieben-Lange, Brigitte, *Traditionen des Sprechens*, Stuttgart, Kohlhammer, 1983.

Steuckardt, Agnès, Große, Sybille, Dal Bo, Beatrice, Sowada, Lena, « Le rituel et l'individuel dans les pratiques d'écriture : l'exemple de la clôture dans des correspondances peu lettrées de la Grande Guerre », W. Remyssen, S. Tailleur (éds.), *L'individu et sa langue*, Laval, Presses de l'Université de Laval, 2020, p. 103-126.

Walter, Richard, « L'édition numérique de correspondances. Guide méthodologique », 2018, https://cahier.hypotheses.org/guide-correspondance.

Wionet, Chantal, « Styles de l'écrit intime », in A. Steuckardt (dir.), *Entre village et tranchées. L'écriture de Poilus ordinaires*, Uzès, Inclinaison, 2015, p. 181-191.

Les données des catalogues de bibliothèques pour créer, sélectionner et évaluer des collections de textes littéraires

Nanette Rißler-Pipka
José Calvo Tello
Andreas Lüschow
Goettingen University, Allemagne

Face au grand nombre de textes littéraires non étudiés, jamais discutés, survolés puis oubliés par l'histoire littéraire, on a toujours espéré que la numérisation de livres en grande quantité serait une solution qui nous rapproche de l'idéal d'un corpus équilibré. Cependant, les raisons et les mécanismes de sélection de textes littéraires évalués de haute qualité sont assez ostensibles et présents dans la discussion critique : dans de nombreux cas, ces textes n'appartiennent pas au canon littéraire parce qu'ils sont écrits dans des langues minoritaires ou à des endroits périphériques par des personnes marginalisées.

Aujourd'hui, nous savons que le système d'exclusion et d'élitisme est tout aussi efficace lorsqu'il s'agit de textes littéraires numériques (Calvo Tello 2016, Robinson 2019).

Cependant, les catalogues de bibliothèque et les bibliographies gardent toutes les informations sur les textes imprimés, ou mêmes sur des manuscrits non publiés. Une révision de l'histoire littéraire est-elle possible à l'aide de catalogues de bibliothèques et de métadonnées bibliographiques ?

Le processus de sélection lors de la constitution de collections de textes est normalement structuré par des critères transparents (Schöch 2017, Gius et al. 2019). Les critères sont plus ou moins équivalents pour les textes imprimés ou numériques. Mais si les chercheurs souhaitent consulter une liste ou un catalogue de tous les romans écrits en français entre 1820 et 1850 pour préparer une étude sur Balzac, ils n'ont pratiquement aucune chance de créer une telle liste. La méthode courante utilisée est de se référer à une liste d'auteurs qui ont influencé Balzac ou qui ont été influencés par ses œuvres (par exemple George Sand) ; on y ajoutera l'inspiration à partir des études d'autres chercheurs. On finit par comparer deux ou trois romans, plus des données contextuelles comme des incidents biographiques. Ne serait-il pas préférable et nécessaire de dresser une liste des romans publiés pour la période étudiée et pour la langue ou la zone culturelle concernée ?

Afin de créer une telle liste, nous proposons d'utiliser les catalogues des bibliothèques. Il faut un catalogue qui permet aux chercheurs de trouver les données bibliographiques qui correspondent aux questions de recherche et aux catégories littéraires (genre, époque, œuvre, auteur). C'est valable aussi pour les collections de textes imprimés que numériques. Une liste des critères et des catégories garantira la transparence de la sélection des textes d'une collection. Les défis concernant certains aspects de la sélection des données et les données FAIR (Wilkinson et al. 2016) peuvent être testés pour une période, un genre et un domaine culturel spécifiques (par exemple, les œuvres littéraires en langues romanes publiées à une période spécifique), appliqués aux données du K10plus (https://wiki.k10plus.de/ catalogue des bibliothèques allemandes).

- Facile à (re)trouver : les catalogues des bibliothèques se concentrent sur des éditions uniques d'une œuvre ou même sur des copies individuelles ; cependant, les corpus littéraires numériques enregistrent principalement des métadonnées sur l'œuvre littéraire elle-même.
- Accessible : même si le catalogue de la bibliothèque peut fournir des informations sur la disponibilité et l'accessibilité d'un texte imprimé, ce n'est pas nécessairement le cas pour la copie numérique.

- Interopérable : pour dresser une liste complète de toutes les œuvres publiées au cours d'une période, d'un genre et dans une langue donnée, il faut disposer de plusieurs sources de données. Bien que les catalogues de bibliothèques utilisent des normes de métadonnées, la mise en correspondance et la fusion sont des défis communs.
- Réutilisable : les catalogues ne contiennent pas toujours les catégories utilisées comme critères pour la constitution de collections de textes (par exemple, les informations sur le genre) ou nécessaires pour l'application de méthodes numériques (par exemple, l'indication du format et de la qualité des données)

Si nous savons comment utiliser les données des catalogues de bibliothèque pour la recherche on réussit aussi à offrir un catalogue avec une fonction de chercher-trouver flexible en offrant aussi une vue globale à propos la relation quantitative entre textes disponible en version numérique ou imprimé.

Références

Calvo Tello, José, « Estado de la digitalizaciòn de la Edad de Plata : un anàlisis cuantitativo », *Revista de Humanidades Digitales*, no. 1, 2016.

Gius, Evelyn, Katharina Krüger, and Carla Sökefeld, « Korpuserstellung Als Literaturwissenschaftliche Aufgabe », in *DHd 2019 Digital Humanities : Multimedial & Multimodal. Konferenzabstracts*, Frankfurt am Main, 2019. https://doi.org/10.5281/zenodo.2600812.

Gantert, Klaus, *Bibliothekarisches Grundwissen*, De Gruyter Saur, 2016. https://www.degruyter.com/view/title/302969.

Robinson, Peter, « Gender, Feminism, Textual Scholarship, and Digital Humanities », in *Intersectionality in Digital Humanities*, edited by Barbara Bordalejo and Roopika Risam, 89-108. Collection Development, Cultural Heritage, and Digital Humanities. Leeds, Arc Humanities Press, 2019.

Schöch, Christof, « Aufbau von Datensammlungen », in *Digital Humanities*, edited by Fotis Jannidis, Hubertus Kohle, and Malte Rehbein, 223-233. Stuttgart, J. B. Metzler, 2017. https://doi.org/10.1007/978-3-476-05446-3_16.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al., « The FAIR Guiding Principles for Scientific Data Management and Stewardship » *Scientific Data* 3 (March), 2016. 160018. https://doi.org/10.1038/sdata.2016.18.

De Confucius à Djebar, de Dante à Lispector. Ce que la criticométrie nous apprend sur la réception des écrivains et de leurs œuvres

Carolina Ferrer
Université du Québec à Montréal, Canada

L'année 2017, en nous basant, d'une façon générale, sur l'analyse systémique introduite dans les sciences sociales par Niklas Luhmann (2000) et Immanuel Wallerstein (2004) et, plus particulièrement, sur le concept de champ littéraire développé par Pierre Bourdieu (1992 ; 1997), ainsi que sur la théorie des polysystèmes proposée par Itamar Even-Zohar (1990), nous avons inauguré un chantier de recherche dans le but de cartographier la littérature mondiale.

Selon Pierre Bourdieu, « le monde social moderne se décompose en une multitude de microcosmes, les *champs*, dont chacun possède des enjeux, des objets et des intérêts spécifiques (champ littéraire, scientifique, politique, universitaire, juridique, des entreprises, religieux, journalistique) » (Bourdieu, 1997, p. 119). De plus, il est pertinent de rappeler, toujours selon le sociologue français, qu'il existe des liens qui se forment, non seulement entre les auteurs, mais aussi par rapport aux agents et aux institutions qui constituent le champ (Bourdieu et Wacquant, 1997, p. 72). En ce sens, la configuration des champs se caractérise par des relations complexes entre ces différentes composantes.

Par ailleurs, les interactions entre les différentes littératures nationales constituent aussi des phénomènes complexes. Afin de mieux les comprendre, nous utilisons les études polysystémiques développées par Even-Zohar. En suivant le sémiologue israélien, nous retiendrons que le terme « système » renvoie ici à des relations fonctionnelles ayant pour but l'élaboration d'hypothèses sur les phénomènes à l'étude. Ainsi, le système littéraire est défini comme « [t]he network of relations that is hypothesized to obtain a number of activities called 'literary,' and consequently these activities themselves observed via that network » (Even-Zohar, 1990, p. 28). Dans cette ligne de pensée, s'éloignant de la plupart des études de littérature comparée, Even-Zohar déclare sa surprise, voire son indignation, concernant l'état lacunaire des études sur les interférences littéraires. Il affirme :

> This reluctance to deal with certain basic questions is indeed incomprehensible when compared to any other field of knowledge. Such basic questions are, for instance : what is interference for, why does it emerge, what are its main features, how does it work, when and under what conditions may it emerge, function for some longer time, and decline? It is inconceivable that such questions should be deliberately ignored just because people are skeptical about the accessibility of adequate answers. (Even-Zohar, 1990, p. 53).

Dans l'élaboration de sa théorie, un aspect central est donc la formulation d'hypothèses dans le but de dégager des lois d'interférence littéraire. En particulier, nous nous pencherons sur les lois d'interférence littéraire en examinant les dimensions géopolitique et linguistique des systèmes et sous-systèmes littéraires.

Comme annoncé, nous inscrivons cette étude dans le cadre des recherches empiriques. À ce sujet, concernant la théorie systémique, nous remarquons que, comme le souligne Kees van Rees :

> System theoreticians usually focus on theoretical and philosophical discussions involving the exegesis of what the founders, such as Bourdieu and Luhmann, have meant by their concepts and propositions on the one hand and the adaptation of certain propositions on the other. Despite system theoreticians' claims to understand the meaning of "model" as

a set of hypothesis, they keep aloof from emprirical research aimed at testing specific hypothesis (van Rees, 1997, p. 91).

Justement, un de nos objectifs sera d'articuler des concepts théoriques avec la recherche empirique. Pour y parvenir, nous employons l'approche méthodologique de la criticométrie (Ferrer, 2011).

La criticométrie

Initialement développée par Price (1963), la scientométrie doit son existence aux instruments mis au point par Eugene Garfield (2005), fondateur de l'Institute for Scientific Information, connu sous le nom de Thomson ISI pendant plusieurs années et récemment devenu Clarivate. L'objectif de la scientométrie est de mesurer et d'analyser l'activité dans les sciences et la technologie. Par analogie, nous avons appelé notre approche la criticométrie, puisqu'elle a pour objectif de mesurer et d'analyser l'activité critique en arts, et plus particulièrement en littérature. La méthode ici proposée correspond à l'adéquation des indicateurs scientométriques à la réalité des bases de données utilisées en sciences humaines et en arts.

Comme présenté au Tableau 1, il existe deux catégories d'indicateurs. Ceux de la première catégorie reçoivent le nom d'indicateurs descriptifs ou d'indicateurs d'activité. Ceux de la seconde catégorie sont les indicateurs relationnels. « Les premiers fournissent des données sur le volume et l'impact des activités de recherche, tandis que les seconds recherchent les liens et les interactions entre chercheurs et domaines, de manière à décrire les contenus des activités et leur évolution » (Callon et al, 1993, p.39). Les indicateurs peuvent être mesurés à plusieurs niveaux d'agrégation : chercheurs, groupe de recherche, institut, pays, discipline.

Tableau 1. Catégories d'indicateurs

Catégories d'indicateurs	Indicateurs	Fonctions
Descriptifs	Dénombrement des articles Dénombrement des citations Dénombrement des brevets	Mesurer le volume et le dynamisme de la recherche à différents niveaux d'agrégation.
Relationnels	Cosignatures Cocitations Mots associés Facteur d'impact	Mettre en lumière les interactions entre les acteurs des systèmes nationaux et internationaux de science et technologie.

Le plus simple des indicateurs descriptifs est le dénombrement des publications ou des brevets. Un autre indicateur descriptif est le dénombrement des citations. Celui-ci correspond au nombre de fois qu'un texte est cité dans une autre publication. Supposément, cet indicateur signale la qualité d'une publication. Cependant, cet argument a été longuement débattu et il en résulte qu'il est préférable de le considérer comme un indicateur de visibilité. De plus : « Des investigations encore plus fines peuvent être tentées en retournant aux contextes de citation de manière à faire ressortir la transformation ou la confirmation des énoncés, les transferts et les emprunts de techniques expérimentales » (Callon et al, 1993, p.60). À son tour, le plus simple des indicateurs relationnels est la cosignature de publications. Un des problèmes que pose cet indicateur est la différence d'habitudes de publication dans les différentes disciplines.

Afin de pallier les défaillances de l'indicateur des citations, Henri Small (1973) a créé celui des cocitations. Il s'agit de comptabiliser le nombre de fois où deux citations apparaissent simultanément dans les articles. Cette coïncidence de références signalerait un lien plus étroit entre

les documents qui les contiennent. Une autre méthode pour remédier les problèmes de l'indicateur des citations a été créée par Michel Callon. Plutôt que s'intéresser aux références, il propose la comptabilisation des co-occurrences de mots contenus dans les documents. « Plus les mots co-occurrent fréquemment dans des textes différents et plus les problèmes de recherche et les connexions entre ces problèmes se renforcent » (Callon et al, 1993, p.81).

Alors que l'indicateur des cocitations permet l'obtention d'un réseau d'auteurs, celui des co-occurrences se traduit en un réseau de mots associés dont la force dépend du nombre de co-occurrences dans les différents documents. Cependant, comme le souligne Loett Leydesdorff (1998), l'utilisation des mots-clés est de nature très variée, ce qui, selon l'auteur, rend préférable l'utilisation de l'analyse des citations pour étudier l'évolution de la science.

Le facteur d'impact est un indicateur qui porte sur les périodiques (Vinkler, 2004). Développé par le *ISI* dans le *Journal of Citation Reports*, le facteur d'impact correspond au « nombre de citations de tout type d'articles, sur 2 ans, concernant le premier auteur, divisé par le nombre d'articles édités à l'exception des éditoriaux, lettres, résumés de congrès » (Devaux, 2002).

Dans le cas de la criticométrie, nous retiendrons le point de vue de Kees van Rees, selon qui « a reliable indicator of the quality attributed to a work of art is permanent and intensive attention –in the form of (spoken or written) discourses » (van Rees, 1997, p. 93).

Dans le cadre d'une publication critique, Tableau 2, l'écrivain et l'œuvre littéraire analysés sont des objets d'étude pour l'auteur du document en question. En suivant van Rees, la référence critique constitue une reconnaissance de l'écrivain de la part du critique, alors que son texte établit une relation discursive avec l'œuvre analysée.

Tableau 2. Relations entre la publication critique et l'œuvre littéraire

	Auteur citant	**Publication critique**
Écrivain cité	Objet d'étude	Reconnaissance ou attention
Œuvre citée	Objet d'étude	Relation discursive

Par rapport aux écrivains qui sont cocités dans une publication critique, Tableau 3, ceux-ci peuvent entretenir une relation de prédécesseur/successeur ou contemporains. Dans tous les cas, ce sont des écrivains que le critique veut comparer en fonction d'une certaine thématique ou approche. Il en va de même pour les ouvrages cocités : ces textes sont l'objet d'une analyse comparée. De plus, les ouvrages littéraires cocités peuvent entretenir des relations que, en suivant Genette (1982), nous appellerons transtextuelles.

Tableau 3. Relations entre deux œuvres cocitées dans une publication critique

	Écrivain cité 2	**Œuvre citée 2**
Écrivain cité 1	Prédécesseur, contemporain, successeur	Source d'inspiration ou d'influence
Œuvre citée 1	Source d'inspiration ou d'influence	Relation inter/transtextuelle

Ainsi, grâce à la criticométrie, nous mettrons au profit l'essor du numérique et des données massives ou *big data* (Boyd and Crawford 2012; Mayer-Schönberger and Cukier 2013) et nous montrerons que cette disponibilité d'information, auparavant inimaginable, peut contribuer à l'avancement de la connaissance du champ littéraire. Dans ce cas précis, nous analyserons la réception critique des écrivains et de leurs œuvres en exploitant plus de deux millions de références, qui sont le résultat d'une communauté de chercheurs du monde entier. Par conséquent, cette étude s'appuie sur la loi des grands nombres.

La réception critique des écrivains

Afin d'obtenir les références des 212 littératures nationales qui correspondent aux nations identifiées par les Nations Unies, nous avons interrogé la plus importante base littéraire, la *Modern Language Association International Bibliography*, que désormais nous appellerons MLAIB. Cette base contient plus de 2,8 millions de références et couvre plus de 160 ans de publications effectuées par la communauté académique internationale. Plusieurs types de documents y sont répertoriés : articles, monographies, livres de recueils, chapitres de livres, éditoriaux, thèses. Cependant, dans cette recherche, nous avons omis les thèses, car elles sont restreintes essentiellement à celles des États-Unis, ce qui biaiserait les données. Les références par littérature nationale ont été obtenues en utilisant les termes « littérature espagnole », « littérature française », « littérature belge », et ainsi de suite, dans le champ « littérature nationale » de MLAIB. La période couverte par les publications critiques est de 1850 à 2016. L'échantillon contient plus de 1,6 million de références.

Nos premières analyses des résultats obtenus nous ont permis d'observer que les études littéraires se concentrent sur un nombre restreint de littératures nationales : celle des États-Unis et celles d'un petit groupe de nations européennes. Ce constat est confirmé lorsque nous observons la place prépondérante qu'occupent certains écrivains et œuvres en tant qu'objets d'étude, eux aussi majoritairement européens (Ferrer, 2018, p. 94).

Par ailleurs, nous avons identifié 1.110 écrivains qui cumulent au moins 1% de la bibliographie critique sur leur littérature d'appartenance.

Dans la recherche ici proposée, nous étudierons la réception critique d'un sous-ensemble d'écrivains, que nous avons choisis de façon à atteindre une plus grande diversité continentale, linguistique et de genre. Le Tableau 4 contient la liste des 20 écrivains sélectionnés, qui appartiennent à 20 littératures nationales différentes.

Tableau 4 : Écrivains sélectionnés

Écrivain	Littérature nationale	Continent	Genre	Langue
Coetzee, J. M. (1940-)	Afrique du Sud	Afrique	Homme	Anglais
Djebar, Assia (1936-2015)	Algérie	Afrique	Femme	Français
Maḥfūẓ, Najīb (1912-2006)	Égypte	Afrique	Homme	Arabe
Senghor, Léopold (1906-2001)	Sénégal	Afrique	Homme	Français
Borges, Jorge Luis (1899-1986)	Argentine	Amériques	Homme	Espagnol
Hébert, Anne (1916-2000)	Canada	Amériques	Femme	Français
James, Henry (1843-1916)	États-Unis	Amériques	Homme	Anglais
Lispector, Clarice (1924-1977)	Brésil	Amériques	Femme	Portugais
Agnon, S. Y. (1888-1970)	Israël	Asie	Homme	Hébreu
Confucius (551 B.C.-479 B.C.)	Chine	Asie	Homme	Mandarin
Desai, Anita (1937-)	Inde	Asie	Femme	Anglais
Natsume Sōseki (1867-1916)	Japon	Asie	Homme	Japonais
Austen, Jane (1775-1817)	Royaume-Uni	Europe	Femme	Anglais
Balzac, Honoré de (1799-1850)	France	Europe	Homme	Français
Cervantes, Miguel de (1547-1616)	Espagne	Europe	Homme	Espagnol
Dante (1265-1321)	Italie	Europe	Homme	Italien
Dostoevskiĭ, Fedor (1821-1881)	Russie	Europe	Homme	Russe
Mann, Thomas (1875-1955)	Allemagne	Europe	Homme	Allemand

Frame, Janet (1924-2004)	Nouvelle-Zélande	Océanie	Femme	Anglais
White, Patrick (1912-1990)	Australie	Océanie	Homme	Anglais

Du point de vue continental, 6 écrivains s'inscrivent en Europe, 4 dans les Amériques, 4 en Asie, 4 en Afrique et 2 en Océanie. Concernant les langues d'expression des écrivains, il y en a 11. Par rapport au genre, l'échantillon est composé de 14 hommes et de 6 femmes. Finalement, les périodes de vie des écrivains s'étendent du 6e siècle av. J.-C. au 21e siècle.

Chaque écrivain sélectionné sera analysé, dans un premier temps, de façon individuelle afin de montrer la place qu'il occupe au niveau de sa littérature nationale d'appartenance. Nous allons aussi identifier le nombre d'études qui portent sur leurs œuvres. Par rapport à la réception internationale, nous déterminerons les langues et les lieux de diffusion de la critique. Dans un deuxième temps, nous analyserons les cocitations des écrivains sélectionnés, afin d'établir les réseaux dans lesquels ils s'inscrivent, notamment du point de vue national, continental et générique. Ainsi, pour chacun de ces différents niveaux d'analyse, nous allons élaborer des indicateurs descriptifs et relationnels.

À travers cette recherche, nous espérons montrer, d'une part, la pertinence d'étudier la littérature en nous basant sur une approche systémique capable de mettre en lumière les relations qui existent entre les écrivains, les œuvres, les langues et les littératures nationales. D'autre part, nous voulons faire ressortir l'importance d'étudier la littérature de façon empirique, notamment en exploitant les données et les métadonnées disponibles depuis l'avènement du numérique, et en introduisant des méthodes quantitatives, telles que la criticométrie, dans notre discipline.

Références

Bourdieu, Pierre, Wacquant, Loïc J. D. ,*Réponses. Pour une anthropologie réflexive*, Paris, Seuil, 1992.

Bourdieu, Pierre, *Méthodes pascaliennes*, Paris, Seuil, 1997.

Bourdieu, Pierre, *Les règles de l'art. Genèse et structure du champ littéraire*, Paris, Seuil, 1992.

Boyd, Danah, and Kate Crawford, « Critical Questions for Big Data Provocations for a Cultural, Technological, and Scholarly Phenomenon », *Information Communication & Society,* 15.5, 2012, p. 662-679.

Callon, Michel, Jean-Pierre Courtial et Hervé Penan, *La Scientométrie*, Paris, Presses universitaires de France, coll. « Que Sais-Je ? » no. 2727, 1993.

Clarivate, https://clarivate.com

Devaux, Valérie, « Veille. L'évaluation de la recherche », *Dossiers de synthèse documentaire*. 2002, CNRS-INIST2007. http://veille.inist.fr/.

Even-Zohar, Itamar, « Polysystem Theory », *Poetics Today* ,11, 1990, p. 1-268.

Ferrer, Carolina. « Les études littéraires à l'ère de la mondialisation : traces et trajets au prisme des nouveaux observables numériques », *Zizanie*, 2.1 (2018) : 76-101.

Ferrer, Carolina, « El boom hispanoamericano: del texto a la pantalla », in *Nuevas aproximaciones al cine hispánico : Migraciones temporales, textuales y étnicas en el bicentenario de las independencias iberoamericanas (1810-2010)*, Barcelona, Promociones y Publicaciones Universitarias, 2011, p. 79-101.

Genette, Gérard, *Palimpsestes. La littérature au second degré*, Paris, Seuil, 1982.

Leydesdorff, Loet, « Theories of Citation ? », *Scientometrics* 43.1, 1998, p. 5-25.

Luhmann, Niklas, *Art As A Social System*, Stanford, Stanford University Press, 2000.

Mayer-Schönberger, Viktor and Kenneth Cukier, *Big data. A revolution that will transform how we live, work, and think*, Boston and New York, Houghton Mifflin Harcourt, 2013.

Modern Language Association International Bibliography, www.mla.org

Price, Derek de Solla, *Little Science, Big Science*, New York, Columbia University Press, 1963.

Small, Henry, « Co-citation in the Scientific Literature : A New Measure of the Relationship Between Two Documents », *Journal of the American Society for Information Science,* 24, 1973, p 265-269.

Thomson Reuters Web of Knowledge. http://www.isiwebofknowledge.com/

United Nations. http://data.un.org/Default.aspx.

Van Rees, Kees, « Modelling the Literary Field: From System-Theoretical Speculation to Empirical Testings », *Canadian Review of Comparative Literature/Revue Canadienne de Littérature Comparée*, March/mars 1997, p. 91-101.

Vinkler, Peter, « Characterization of the Impact of Sets of Scientific Papers : The Garfield (Impact) Factor », *Journal of the American Society for Information Science and Technology* 55.5, 2004, p. 431-435.

Wallerstein, Immanuel. *World Systems Analysis : An Introduction*, Durham, Duke University Press, 2004.

RIRE : une base de données pour explorer vers et humour

Anne-Sophie Bories
Université de Bâle, Suisse

Quoi de commun entre les vers et les blagues ? À première vue, on les placerait plutôt aux pôles opposés d'une hiérarchie conventionnelle des formes littéraires. Et pourtant, on les rencontre souvent ensemble. Les poètes même les plus sérieux ont composé des pièces humoristiques et parfois obscènes, mêlé à leurs œuvres sérieuses des jeux de mots et des facéties. Hugo en est un bon exemple, qui y recourt régulièrement.

Poésie et humour violent volontiers les maximes de Grice, selon lesquelles une communication doit viser la concision et fuir l'ambiguïté, ils s'appuient au contraire sur les insuffisances de la langue pour produire avec précision le degré d'ambiguïté voulu, nous offrant le plaisir de significations superposées et trompeuses. Les procédés de la versification et de l'humour partagent certains traits : le recours à la monotonie, la construction d'un horizon d'attente, qui permettent la production de discordances, de contrepoints, la suggestion d'énoncés latents, ou la mise en scène d'une chute.

Dans le cadre du projet « Le Rire des vers » (SNSF), nous explorons systématiquement différents aspects de ces deux procédés, et construisons pour cela une vaste et riche base de données.

La base RIRE rassemble des textes numérisés couvrant trois larges corpus, et trois catégories de données. Nous recueillons d'abord des données fines sur la versification, résultat du traitement automatique des textes par le programme d'analyse Malhebe développé au CRISCO par Richard Renault et Éliane Delente. Ces données nous placent dans la filiation de Benoît de Cornulier, dont les travaux sur la métrique ont fourni aux spécialistes de versification des outils systématiques et opérants. S'ajoutent à ces données sur la versification des données linguistiques, parties du discours et lemmes notamment. Nous effectuons en outre un relevé méticuleux de procédés stylistiques assimilables à des plaisanteries. Pour cette partie de notre effort, nous nous basons sur les travaux de Victor Raskin et Salvatore Attardo sur l'humour langagier. Leurs grilles d'analyses des blagues (jokes), fondées sur la description de double scénarii et sur l'examen de divers éléments d'ancrage, nous ont servi à établir un protocole de description efficace non seulement des procédés humoristiques, mais aussi de nombreux autres phénomènes stylistiques plus ou moins apparentés à la syllepse. Notre base RIRE permet ainsi l'exploration multimodale de trois grands corpus : l'un consacré à la poésie des XIX^{e} et XX^{e} siècles, l'autre aux couplets de vaudeville de 1830 à 1835, le troisième à la chanson des XIX^{e} et XX^{e} siècles.

Au moyen de cet outil audacieux, nous voulons explorer les liens qui se tissent sur le plan stylistique et thématique au sein des textes en vers, entre la forme même des vers et la construction du sens, ou plutôt la construction des sens, puisque nous nous intéressons tout particulièrement aux moments des textes qui superposent des sens parfois incompatibles, suggèrent des énoncés latents capables d'enrichir ou de déstabiliser l'énoncé principal, freinent la lecture par leur incongruité, suscitent parfois le rire, plus souvent le sourire, ou simplement le plaisir d'une lecture multipliée, d'une concentration de la langue, de l'inattendu. Ces efforts visent d'abord à éclairer l'évolution et le fonctionnement de deux phénomènes hautement ancrés : les formes métriques, qui portent avec elles, selon les époques, les lieux et les cercles, leur propre lot de sens historique et stylistique, et l'humour au sein des textes versifiés, qui soulève des questions d'ordre historique, culturel et sociologique, notamment parce que l'humour, pour être opérant, nécessite un juste degré d'investissement de la part du public. Ensuite, notre travail vise à proposer une exploration fonctionnelle des figures stylistiques qui reposent sur des mécanismes proches de l'humour, et qui

peuvent être décrits avec la même série d'outils. Cette exploration systématique est à la fois multimodale, en ceci qu'elle combine divers outils d'analyse, et multifocale, en ceci qu'elle associe une lecture de loin à une lecture de près, combine différentes longueurs focales pour renouveler la lecture stylistique des textes.

MotiveR : un programme pour la stylistique

Dominique Legallois
LATTICE, UMR8094
Université Sorbonne-Nouvelle, France

Antoine Silvestre de Sacy
HumaNum, France

L'objectif de la communication est de présenter au public un outil informatique écrit sous la forme d'un script R : MotiveR[1]. MotiveR est un programme permettant d'identifier par méthode non supervisée des patrons syntaxiques (les motifs) sur-représentés dans un genre ou chez un auteur, par comparaison avec d'autres genres ou avec d'autres auteurs. Ces motifs sont des unités plus abstraites (ou schématiques) que des ngrammes de tokens, tout en étant plus spécifiques que des ngrammes de Part of Speech. L'équilibre entre schématicité et spécificité des unités est ainsi recherchée.

Certains motifs d'un texte peuvent être statistiquement sur-représentés (par rapport à d'autres textes) ; ils peuvent alors être considérés comme des unités stylistiques caractéristiques.

Nous donnons ici quelques exemples issus d'une analyse dans laquelle nous avons comparé cinq auteurs : Balzac, Dumas, Hugo, Sand et Zola. Le corpus était composé de 10 romans de chacun de ces auteurs (huit pour Hugo). Chez G. Sand, on peut relever, parmi des dizaines d'autres, les deux motifs sur-employés suivants :

1- ADJ et ADJ comme DETPOSS NC

> Jeannie était mince et petit comme sa mère, dont il avait toute la retirance. (*François le Champi*)
>
> Elle ne le haïssait point d'être calculateur et positif comme son siècle. (*Indiana*)

2- si ADJ et si ADJ que

> Néanmoins Mme Aldini était si gracieuse et si bienveillante, que mon brave homme de père, [...] ne sut que répondre à ses douces paroles et à ses généreuses promesses (*La Dernière Aldini*).
>
> j'avais pour elle un attachement si légitime et si profond, que je ne pensais pas faire un serment téméraire (*La Dernière Aldini*)

Chez Zola :

3- ADJ ainsi que un NC

> Un moulin, avec ses ailes, demeura seul, ainsi qu'une épave (*La Terre*)
>
> Par moments, des rues transversales qui dévalaient, des trouées brusques montraient l'immensité de Paris, profonde et large ainsi qu'une mer (*L'Œuvre*)

[1] MotiveR fera par la suite l'objet d'un package R.

4- ADJ, avec DETPOSS ADJ NC

> un instant, il était resté surpris et plein de gêne, devant cette fille déjà savante, avec ses grands yeux candides (*La Joie de vivre*)
>
> La jeune fille qui écoutait, souriante, avec son clair regard si froid et si décidé, eut une brusque affirmation du menton (*L'Argent*).

MotiveR est également programmé pour détecter des motifs sur des corpus en langue anglaise. Ainsi, en comparant les deux traductions anglaises du *Ventre de Paris* (Zola) de Kuransky et de Nelson (publiées toutes les deux en 2009), on peut mettre en évidence grâce à l'outil, ce motif (encore un fois, parmi des dizaines d'autres) spécifique chez Kuransky (c'est-à-dire sur-représentés par rapport à la traduction de Nelson) :

5- , slightly PASTPART by the NOUN

> The mat that covered the floor, the soft yellow wallpaper , the imitation oak oilcloth, all gave a coolness to the room, slightly softened by the shine of a brass lamp that hung from the ceiling and sprawled above the table with its large transparent porcelain shade.
>
> The tall brown-haired clerk, with flashing eyes in her calm face , slightly reddened by the cold, sat on a high wooden chair , peacefully writing, apparently undisturbed by the commotion of the hunchback, who seemed to ripple the edges of her skirts.

Outre l'identification et le calcul de motifs (par calcul des spécificités), MotiveR propose également : une représentation en WordCloud, un calcul Tf-idf, une analyse par AFC, une représentation de la densité des motifs dans une œuvre (permettant de voir les endroits dans le texte où un motif est particulièrement employé), un calcul de l'évolution temporelle (ou dans une œuvre) des motifs, un concordancier. Le concordancier permet à partir des calculs statistiques réalisés en amont, d'étudier les figures stylistiques caractéristiques d'un auteur en lecture proche. Chez Louis-Ferdinand Céline, par exemple, sa propension à la dislocation avec rappel par un pronom peut ainsi être modélisée :

6 -- le NC, il me VIMP

> Les passants, ils me remarquaient (*Mort à crédit*)
>
> La môme, elle me refaisait des gestes (*Mort à crédit*)
>
> La fatma, elle me fait signe de venir (*Mort à crédit*)

La force du programme réside dans sa capacité à fournir à l'utilisateur un ensemble de fonctions exécutables allant des textes bruts, aux analyses statistiques et à des visualisations permettant d'exposer ses résultats, tout en gardant une grande généralité lui permettant d'être utilisé dans des disciplines et problématiques très diverses (stylistique littéraire, attribution d'auteurs, linguistique des genres textuels, etc.). Le format Tidy utilisé favorise l'utilisation des résultats par d'autres outils R.

Le script est facilement utilisable par des utilisateurs non informaticiens. Lors du colloque, nous proposons donc de présenter les différentes fonctions de MotiveR illustrées par plusieurs cas d'analyse.

Références

Legallois, Charnois et Larjavaara, « The balance between quantitative and qualitative literary stylistics : how the method of 'motifs' can help » in Legallois, Charnois et Larjavaara, The Grammar of Genres and Styles : From Discrete to Non-Discrete Units, De Gruyter Mouton, 2018, p.168-193

Quiniou, Cellier, Charnois et Legallois, « What About Sequential Data Mining Techniques to Identify Linguistic Patterns for Stylistics ? » in *Lecture Notes in Computer Science*, Springer, 2012.

Méta-données.
De l'accessibilité des sources à l'élaboration des objets de la recherche (données et modèles)

Marie-Hélène Lay
FORELL, EA3816
Université de Poitiers, France

La communauté des sciences humaines, et tout particulièrement celle qui se retrouve au sein de CAHIER, se réjouit, à n'en pas douter, du développement des humanités numériques et de la disponibilité croissante de données, qu'elles soient textuelles, orales, ou multi-modales. Seuls les plus jeunes d'entre nous ne se souviennent pas d'une époque où nos objets de recherche n'existaient pas au format numérique, où les plus gros des laboratoires de recherche, dans les années 1990 disposaient de moins d'information textuelle au format numérique que tout un chacun sur son laptop en 2020. Beaucoup d'énergie est aujourd'hui encore mobilisée pour la mise à disposition de nos corpus (corpus d'auteurs ou autres), de façon plus ou moins coordonnée, ce qui montre bien la nécessité de lieux d'échanges comme CAHIER, dont l'extension constante au cours des 10 années écoulées a permis de sensibiliser et de fédérer une part toujours plus nombreuse de notre communauté : il nous est nécessaire de partager nos expériences et de consolider une communauté de pratiques, nécessaire de construire ensemble le paradigme numérique de nos pratiques scientifiques.

Car il faut bien le dire, l'enthousiasme éprouvé à l'idée de disposer de tant de sources et ressources se trouve tempéré par deux caractéristiques constitutives du paradigme numérique : d'une part la prise de conscience du fait qu'il ne suffit pas de disposer des textes « saisis », par exemple en format .txt, .odt ou .doc pour en faire des données de la recherche au format numérique disponibles pour la communauté, d'autre part la prise de conscience d'un phénomène parfois appelé « infobésité » : la masse de données est telle qu'on ne sait pas toujours comment en faire bon usage, bon usage scientifique s'entend, qui suppose que les données de la recherche aient été patiemment élaborées pour cerner au plus près, de la façon la plus honnête et la plus exhaustive possible, les éléments permettant de nourrir réflexion et argumentation autour d'une question posée. Ce qui est nouveau dans l'environnement numérique, c'est que l'on peut apporter une réponse unifiée à toutes ces problématiques : la pratique de l'annotation, c'est-à-dire l'adjonction de métadonnées, permet de structurer les fonds documentaires, de localiser les ressources et de documenter les perspectives scientifiques, de transformer des données dites « brutes » en données de la recherche.

Mais ce point doit sans doute être mieux élucidé et la pratique de l'annotation mieux exposée. En effet, il est certain que l'annotation, et plus particulièrement l'annotation des contenus, est extrêmement coûteuse en moyens humains : il peut donc sembler vain de s'y contraindre, comme il est vain de chercher à utiliser des annotations « standard » pour aborder un point délicat de nos recherches. Notre réponse-réflexe est alors : « vu le temps et les ressources nécessaires, mieux vaut se concentrer sur notre recherche elle-même que sur toutes ses étapes bien trop coûteuses par rapport à notre objet d'étude ».

Présentées ainsi, comment ne pas souscrire à ces réticences ? Mais comment ne pas remarquer par ailleurs que ces données appelées des vœux des chercheurs à la fin du XX^{e} et au début du XXI^{e} siècle ne semblent pas avoir massivement rénové la recherche de ceux qui sont d'abord des littéraires et des spécialistes de leurs auteurs. Il me semble raisonnable d'envisager que la raison en est assez simple : sauf à entrer dans des démarches d'analyse de type textométrique, le chercheur accédant aux ressources textuelles au format numérique reste dans une approche pré-numérique des textes.

Car de fait, le constat s'impose, les données patiemment élaborées et rendues disponibles sont largement sous exploitées. La simple disponibilité de ressources gigantesques, d'un tas de données en tas ne suffit pas à les rendre accessibles de façon pertinente, on risque de s'y « noyer », même si elles sont proprement cataloguées et indexées, d'un point de vue de « bibliothèque ». Certes, le phénomène n'est pas nouveau, ce dont on trouve des formulations plaisantes, comme « Vous croulez sous vos données ? C'était déjà le cas du temps de Voltaire[1] », phénomène dont nous avons tous fait l'expérience lorsque nous recherchions une citation précise dans un paquet de notes prises de façon trop peu organisée, sur des petites fiches cartonnées dispersées çà et là : un volume d'information assez restreint ne permet donc pas forcément d'échapper au sentiment de noyade. Aujourd'hui comme alors, il nécessaire d'organiser sources et ressources pour en permettre la localisation et le partage[2], mais aussi (et surtout ?) l'usage.

Parallèlement à la disponibilité de ressources se pose donc la question de leur exploitation. Leur volume est devenu tel, que la contrainte de l'outillage informatique s'impose d'elle-même : la consultation séquentielle, ou plus précisément, dans le cas présenté ici, la lecture linéaire par un humain n'est certainement pas la bonne méthode. Outre la transposition à l'environnement numérique des pratiques documentaires traditionnelles, permettant l'organisation des ressources et l'accès aux documents sont apparus des outils assurant l'accès direct au contenu des documents[3] et le traitement de l'information localisée : c'est le cas d'outils comme TXM, bien représenté au cours de cette journée.

C'est un point de vue un peu différent qui sera adopté ici (complémentaire des autres éléments évoqués ci-dessus). L'annotation de contenu ne sera pas d'abord présentée comme une pratique unifiée permettant la réutilisabilité des données, ou une étape utile avant le recours à des outils de traitement statistique. L'annotation manuelle de contenu sera réinscrite dans la lignée des pratiques pré-numériques de l'enrichissement des textes, enrichissement au sens ou une information nouvelle est associée à une information initiale : c'est ce que nous faisons communément en préparant une lecture critique, en produisant diverses versions d'un texte, en corrigeant des épreuves pour un éditeur, en corrigeant des travaux d'étudiants. Dans les pratiques d'annotation manuelle dont il est question ici, les métadonnées sont généralement insérées dans le document lui-même[4].

Annoter, c'est, formulé de la façon la plus simple, ajouter des « notes » à un document, ajouter une information B à une information initiale A. C'est enrichir des données (information A) par d'autres données (information B). Ces autres données sont appelées méta-données. En des termes plus techniques, on peut emprunter à Fort (2012, p. 17[5]) la définition suivante : « l'annotation recouvre à la fois le processus consistant à apposer (ad-) une note sur un support, l'ensemble des notes ou chaque note particulière qui en résulte et ce, sans préjuger a priori de la nature du support considéré (texte, vidéo, images, etc.), du contenu sémantique de la note (note chiffrée, valeur choisie dans un référentiel fermé ou texte libre), de son positionnement global ou local, ni de son objectif (visée évaluative ou caractérisante, simple commentaire discursif) ».

Notre objectif sera ici de montrer en quoi la pratique de l'annotation manuelle n'est pas une contrainte supplémentaire imposée par la mise à disposition des informations au format

---

1 https://rslnmag.fr/ « Regards sur le numérique », juin 2016.

2 Il y a nécessité à organiser les informations lorsqu'elles sont nombreuses, afin de permettre un accès pertinent à leur contenu. On n'imagine pas de bibliothèque qui poserait en vrac les livres sur les tables et les rayonnages et attendrait qu'on lise tout ou partie des livres les uns après les autres, au hasard. (On est bien sûr libre de se perdre entre les rayonnages comme on est libre de cliquer sur des hyperliens, au hasard, mais il arrive qu'on décide d'avoir un comportement plus *goal-driven*).

3 Cf. protocole OAI, ou le groupe de travail sur la typologie des textes de CAHIER.

4 On peut aussi faire de l'annotation dite « déportée », c'est-à-dire qu'elle donne lieu à un fichier séparé du corpus originel, mais synchronisé avec celui-ci. Entrer dans les détails de cette technique n'apporterait rien ici à notre propos.

5 Fort K. (2012), *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus*, thèse de Paris 13.

numérique, mais en quoi elle est un outil de façonnage des données de la recherche, un outil d'expérimentation des modèles que nous élaborons, un laboratoire de formulation et de test, un outil de pertinent venant à l'appui de nos heuristiques[6].

Références

André V., Canut E., « Mise à disposition de corpus oraux interactifs : le projet TCOF (Traitement de Corpus Oraux en Français) », *Pratiques*, no. 147-148, 2010, p. 35-51.

Baude O., Corpus oraux, guide des bonnes pratiques, 2006. http://www.dglf.culture.gouv.fr/recherche/corpus_parole/Corpus_Oraux_GBP%202006_version_imprimee.pdf

Baude O., « Aspects juridiques et éthiques de la conservation et de la diffusion des corpus oraux », *RFLA* XII-1, 2007, p. 85-97.

Bilger M., *Corpus - Méthodologie et applications linguistiques*, Paris, Champion, 2000.

Broudoux E., Scopsi C., « Métadonnées sur le web : les enjeux autour des techniques d'enrichissement des contenus », *Études de communication*, no. 36, 2011, p. 9-22.

Burnard L., « Text Encoding for Information Interchange. An Introduction to the Text Encoding Initiative », *Proceedings of the Second Language Engineering Conference*, TEI J31, 1995.

Burnard L., « Metadata for corpus work », in Wynne, M. (ed.) *Developing linguistic corpora : a guide to good practice*, Oxford, Oxbow Books, 2005, p. 30-46. http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter3.htm

Burnard L., « Qu'est-ce que l'annotation et pourquoi en parle-t-on de manière si inquiétante ? », *École thematique annotation de données langagières*, 2011. http://www.lattice.cnrs.fr/IMG/pdf/burnard-annotation.pdf

Burnard L., « Encoder l'oral en TEI : démarches, avantages, défis.... », Séminaire de l'institut du numérique, 2012. http://ecrin.u-paris10.fr/post/2012/05/11/Lou-burnard-%E2%80%9CEncoder-l%E2%80%99oral-en-TEI-%3A-d%C3%A9marches,-avantages,-d%C3%A9fis%E2%80%A6.%E2%80%9D

Cappeau P., Gadet F., « L'exploitation sociolinguistique des grands corpus. Maître-mot et pierre philosophale », *RFLA XII-1*, 2007, p. 99-110.

Charaudeau P., « Dis-moi quel est ton corpus, je te dirai quelle est ta problématique », *Corpus*, no. 8, 2009, p. 37-66.

Dalbera J.-P., « Le corpus entre données, analyse et théorie », *Corpus*, no. 1, 2002, p. 89-105.

Fort K., *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus*, thèse de Paris 13, 2012.

Habert B., Nazarenko A., Salem A., *Les Linguistiques de corpus*, Paris, Armand Colin, 1997.

Habert B., « Des corpus représentatifs : de quoi, pour quoi, comment? », *Linguistique sur corpus. Études et réflexions*, Mireille Bilger (resp.), Perpignan, Presses Universitaires de Perpignan, Collection Cahiers de l'université de Perpignan no. 31, 2000, p. 11–58.

Houdé O., *Catégorisation et développement cognitif*, Paris, PUF, 1992.

Hunston S., *Corpora in Applied Linguistics*, Cambridge, Cambridge University Press, coll. « Cambridge Applied Linguistics », 2002.

[6] Un raisonnement ou une méthode heuristique (ou une heuristique) est une méthode de résolution de problème qui ne s'appuie pas sur une analyse détaillée ou exhaustive du problème. Elle consiste à fonctionner par approches successives en s'appuyant, par exemple, sur des similitudes avec des problèmes déjà traités afin d'éliminer progressivement les alternatives et ne conserver qu'une série limitée de solutions pour tendre vers celle qui est optimale. Jean-Louis Le Moigne (1991), en donne cette définition : « Une heuristique est un raisonnement formalisé de résolution de problème (représentable par une computation connue) dont on tient pour plausible, mais non pour certain, qu'il conduira à la détermination d'une solution satisfaisante du problème. »

Kraif, O., « Les concordances pour l'observation des corpus : utilité, outillage, utilisabilité », J. Chuquet (dir.), *Le langage et ses niveaux d'analyse – Cognition, production de formes, production du sens*, Rennes, Presses Universitaires de Rennes, 2011, p. 67-80.

Lay, M.-H., Zaysser, L., « A Generic Model for Reusable Lexicons : The GENELEX Project », *Linguistics and Literary Computing*, vol. 9, n°1, 1994, p. 47-56.

Lay, M.-H., « Pour une exploration humaniste des textes : AnaLog », *Actes JADT'2010*, Roma, 2010. http://lexicometrica.univ-paris3.fr/jadt/jadt2010/allegati/JADT-2010-1045-1056_106-Lay.pdf

Lay M.-H., *Linguistique de/sur corpus et Humanités Numériques*, Synthèse d'HDR soutenue publiquement le 24 Novembre 2015, Université de Poitiers, 2015.

Lay, M.-H., P. Caron et R. Defiolle (dirs.), *L'Enjeu des méta-données dans les corpus textuels*, Rennes, PUR, 2019.

Leech G., « Adding Linguistic Annotation », *Developing Linguistic Corpora : a Guide to Good Practice*, ed. M. Wynne. Oxford, Oxbow Books, 2005, p.17-29. http://ahds.ac.uk/linguistic-corpora/.

Le Moigne, J.-L., *La Modélisation des systèmes complèxes*, Paris, Dunod, 1991.

Martinet A., *Grammaire fonctionnelle du français*, Paris, Didier, 1979.

Mellet S., « Corpus et recherches linguistiques : introduction », *Corpus*, no. 1, 2002, p. 5-12.

Pédauque R.T., *Le document à la lumière du numérique*, C&F éditions, 2006.

Rastier F., « Enjeux épistémologiques de la linguistique de corpus », *La Linguistique de corpus*, Rennes, Presses Universitaires de Rennes, 2005, p. 31-46.

Salaün, J-M., « La redocumentarisation, un défi pour les sciences de l'information », *Études de communication*, n° 30, 2007.

Sinclair J. M. H., « Corpus and Text : Basic Principles », *Developing Linguistic Corpora : A Guide to Good Practice*, 2005, p. 1-16.

Wynne M., « Archiving, Distribution and Preservation », *Developing Linguistic Corpora : a Guide to Good Practice*, ed. M. Wynne. Oxford, Oxbow Books, 2005, p. 71-78. http://ahds.ac.uk/linguistic-corpora/