Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

Réinterroger l'édition numérique et la consultation d'œuvres anciennes : traçabilité, accessibilité, interprétabilité

Emmanuel Giguet[1] Julia Roger[2]
(1) Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France
(2) MRSH, Université de Caen, 14000 Caen, France
emmanuel.giguet@cnrs.fr, julia.roger@unicaen.fr

RÉSUMÉ

Dans le domaine des humanités numériques et de l'édition d'œuvres anciennes, l'influence de la Text Encoding Initiative (TEI) a porté ses fruits et n'est plus à démontrer. Le contexte technologique est cependant propice à l'émergence de nouveaux modes de consultation et de diffusion. Nous nous appuierons sur la création d'une nouvelle interface de consultation des œuvres de Descartes pour traiter des questions de traçabilité des opérations, d'interopérabilité des ressources de TAL, et d'interprétabilité.

ABSTRACT

Culture Heritage and Digital Publishing : Traceability, Accessibility, Challenges.

In the field of Culture Heritage and Digital Publishing, the influence of the Text Encoding Initiative (TEI) is no longer to be demonstrated. However, new ways to access and share information emergence. Using Descartes Digital Corpora, we will address the challenges raised by new interfaces : traceability, interoperability of NLP resources, and interpretability.

MOTS-CLÉS : Text Encoding Initiative, interopérabilité, Numérisation, OCR.

KEYWORDS: Cultural Heritage, Document Indexing, Word Spotting, OCR.

1 Introduction

Dans le domaine des humanités numériques, l'influence de la Text Encoding Initiative (TEI) sur de très nombreux projets d'édition numérique a porté ses fruits et n'est plus à démontrer. La TEI a tout particulièrement favorisé la structuration des contenus, leur diffusion, leur annotation, leur analyse (Galleron *et al.*, 2018). À Caen notamment, et plus particulièrement à la Maison de la Recherche en Sciences Humaines, une édition en ligne des œuvres et de la correspondance de Descartes a été réalisée en suivant les prescription de la TEI (ANR ProDescartes, 2009-2014) (Carraud, 2009). Il s'agit du *Corpus Descartes* [1]. Stockée sous forme de documents au format XML TEI, l'œuvre peut être consultée par le public sous une forme esthétique, moderne, permettant la recherche de mots ou d'expressions en français classique grâce à la mise en oeuvre de technologies relevant de l'indexation automatique et du Traitement Automatique des Langues (TAL).

Si l'on peut se ravir de la mise à disposition au plus grand nombre de telles œuvres par l'intermédiaire de ces éditions numériques, si l'on peut se réjouir des outils de recherche qui peuvent leur être associés

1. https://www.unicaen.fr/puc/sources/prodescartes/

pour faciliter leur étude, il n'en reste pas moins que la forme numérique contemporaine prive le lecteur d'un certain rapport à l'ouvrage, certes physique, certes tactile, mais pas uniquement. Les éditions originales comportent des spécificités dont certaines sont abandonnées lors de la réalisation de l'édition numérique. L'on pense bien entendu à certaines particularités de la composition typographique de l'époque : l'usage de l'espace autour des signes de ponctuation qui diffère de l'usage contemporain et peut être normalisé, les césures propres à la mise en ligne du texte dans l'édition papier et qui doivent être abandonnées, le s court remplaçant sa forme longue, ſ, qui peut être systématisé, les ligatures esthétiques à abandonner, les variations graphiques d'un terme à harmoniser sur l'ensemble de l'œuvre. Ces choix éditoriaux profitent bien entendu tant au confort de lecture des utilisateurs, à la recherche de graphies particulières, qu'à la facilité de mise en œuvre de traitement automatique ; il n'en reste pas moins qu'une succession de *réductions* de ce genre, par rapport à la source, éloignent d'autant plus le lecteur de la version originelle du texte.

Fervents défenseurs de la Text Encoding Initiative (Ide & Véronis, 1995), déployée dans nombre de projets d'édition menés à Caen depuis plus de 15 ans, nous souhaitons profiter de cet article pour requestionner la distanciation entre l'œuvre originale et l'œuvre numérique engendré par l'emploi de la TEI. L'on entend certes les arguments en faveur de l'étude et de la diffusion des œuvres, de parties d'œuvres, ou de citation d'un extrait (Buard, 2015), du confort de lecture attribué aux éditions numériques sur des dispositifs modernes, la facilité de mise en œuvre de techniques d'annotation et de technologies d'analyse automatique pour mener des études scientifiques.

On peut cependant regretter les pertes liées à la consultation d'un ouvrage paginé, où l'espace de la double page fait sens tant pour l'auteur que pour le lecteur (Tschichold & Paris, 1994), en plaçant par exemple une illustration, un schéma, ou une carte en regard d'un texte pour en faciliter l'interprétation, où l'épaisseur de l'ouvrage n'est plus perceptible, ne serait-ce que pour y glisser un marque-page, la composition originale et plus généralement la mise en forme porteuse de sens disparaît (Roger, 2015). C'est d'autant plus regrettable quand on sait à quel point la « mise en livre » influe directement sur la compréhension des textes (Chartier, 1997). Enfin, alors que la traçabilité des opérations de numérisation est souvent au cœur des processus de conversion et est un véritable gage de qualité, l'on peut regretter que le lien entre les dispositifs de l'édition numérique et l'édition physique dont elle est issue soit si ténu, les mots de l'édition numérique n'ayant par exemple plus d'ancrage, de lien avec l'ouvrage papier, si ce n'est au travers d'un numéro de page, voire d'un numéro de ligne sur lequel il apparaît : ce que Michel Melot a appelé « le pouvoir transcendantal du livre », pouvoir qui « est inscrit dans le pli » est ainsi perdu (Melot, 2006).

À la faveur du confinement, nous avons profité du calme qui nous était imposé pour penser une nouvelle manière de consulter en ligne les œuvres de Descartes, mais aussi de les étudier. Certains pourraient la qualifier de plus moderne, d'autres tout simplement d'alternative. Force est de constater qu'elle ne laisse personne indifférent. Présentée lors de la Fête de la science, l'événement national de diffusion de la culture scientifique, elle n'a laissée indifférent aucun public, enfants, élèves, étudiants, enseignants, les replongeant dans l'intimité du rapport au livre, réactivant un lien distendu, parfois perdu, avec le livre, objet autrefois pourtant si familier. Une fois retrouvé le plaisir de tourner les pages, de feuilleter, un geste ici simulé par le biais d'une interface tactile qui ne remplace cependant pas le plaisir du touché de la feuille, du ressenti de son épaisseur, de sa texture, de son odeur parfois même, il est alors temps d'entrer dans l'œuvre. Tout d'abord par la forme retrouvée, puis par le fond. Bien entendu, la médiation scientifique n'a pas été notre préoccupation première ; nous ne la négligeons cependant pas pour autant.

Dans notre culture interdisciplinaire, notre objectif est cependant autre. Pour un philosophe, un

historien, qu'apporte le retour à l'ouvrage dans ses recherches ? Le retour à une certaine forme de matérialité n'est-il qu'un artifice ? Une fois l'aspect ludique passé, quel est l'intérêt à poursuivre l'expérience de la matérialité, à la personnaliser ? Cette question prend davantage de sens, peut-être pourrait-on d'ailleurs en sourire, au regard du corpus support de l'expérimentation, celui de Descartes, et le discrédit qu'il avait pour l'objet-livre (Roger, 2015).

Dans cet article, nous allons tout d'abord repartir de l'état pré-crise sanitaire, à savoir les œuvres et la correspondance de Descartes codées dans le respect de la Text Encoding Initiative et présentées aux lecteurs à l'aide de transformations XSL (*Extensible Stylesheet Language Transformation*). Nous reviendrons sur les choix effectuées par l'équipe en charge de l'édition numérique. Dans un second temps, nous montrerons comment le contexte technologique s'est avéré propice à l'émergence d'un nouveau mode de consultation. Dans une troisième partie, nous évoquerons comment le Traitement automatique des langues a contribué au projet en travaillant l'interopérabilité des ressources. Nous terminerons par une discussion sur les perspectives d'une telle expérimentation.

2 Le corpus ProDescartes et les premières éditions numériques

L'objectif premier de ProDescartes (ANR 2009-2014) était de publier, à l'heure du numérique, toutes les œuvres du grand philosophe – œuvres publiées et inédites ainsi que la correspondance –, dans leurs éditions originales.

Ce retour aux sources était jugé nécessaire par les spécialistes au regard de l'histoire éditoriale des textes cartésiens, ces derniers ayant subi une succession de déformations, scories, ajouts inauthentiques qui ont progressivement dénaturé l'intention de l'auteur. Qui sait, par exemple, que le *Discours de la méthode* qu'on lit aujourd'hui n'est que la préface d'une œuvre originelle beaucoup plus ample, illustrée, constituée de trois Essais de physique et de mathématiques ?

Le second objectif était d'outiller le corpus ainsi produit en français classique et en latin avec des méthodes et techniques actuelles de recherche et d'ingénierie en informatique : moteur de recherche lemmatisé et multilingue, outil d'annotation collaboratif ainsi que différents outils issus du TAL ; le tout à destination des philosophes, pour favoriser la recherche sur les études cartésiennes et enregistrer en temps réel les résultats du cartésianisme.

Mais un des obstacles majeurs au choix des éditions originales était linguistique : le français classique et le latin étaient peu dotés, à l'époque, de ressources pour produire le moteur de recherche espéré.

Malgré une grande fidélité de principe à la source du XVII^{e} siècle, un certain nombre de propriétés des éditions originales, jugées secondaires par rapport à l'interprétation des textes, ne sont donc pas restituées dans la version TEI : nous nous en sommes défendus dans les textes de présentation. Qu'on songe aux particularités typographiques ou à celles de la composition. Naturellement, l'accès au fac-similé des éditions originales reste doublement possible : des liens qui correspondent à l'élément *pagebreak*, noté <pb>, de la TEI affichent chaque page des éditions papier conservées à la BNF. Un autre lien, au bas de chaque fenêtre de la visionneuse renvoit encore à la même page, mais en haute résolution sur le site de Gallica. Mais ce renvoi au mode image de l'imprimé n'est qu'une option de lecture : on peut masquer ces liens et ne lire le texte que sous sa forme numérique. Au sens propre comme au sens figuré, le lien avec la forme du texte validée par Descartes est perdu. C'est un défaut qu'il nous faut combler, au nom même de l'idée que Descartes se faisait de la « mise en livre » de ses propres ouvrages à l'édition desquels il a activement travaillé. N'oublions pas que Descartes

n'enseignait pas et que ses ouvrages étaient les seuls véhicules de sa pensée. Certes, en philosophie cartésienne, la lecture des bons livres n'est pas suffisante à la découverte de la vérité car ceux-ci ne nous apprennent pas à la chercher. Mais pratiquer une lecture ordonnée de ses propres livres, constitue une étape vers la vérité, dans la mesure ou le fil du texte est lui-même méthodique. Devenant sensible à l'ordre même dans lequel se présente les objets décrits par Descartes, le lecteur progresse sur la voie de l'appropriation de la méthode.

Des solutions plus souples de consultation des imprimés originaux sont donc à la fois attendues, non seulement par le grand public, les historiens du livre mais aussi par les spécialistes de Descartes qui savent que la nature démonstrative de ses énoncés est chez lui soutenue par un ensemble de dispositifs typographiques et graphiques qu'il a choisis ou élaborés (Carraud, 2011).

FIGURE 1 – Le site Internet de l'édition en ligne *Corpus Descartes*

3 Un contexte technologique favorable à la traçabilité

Les réflexions que nous développons ici et le prototype qui en est issu résultent en fait de plusieurs avancées technologiques majeures en matière de reconnaissance automatique des caractères (OCR) sur ouvrages anciens. Ceci nous a permis de rétablir le lien entre l'ouvrage numérisé et l'édition numérique au format TEI que nous avions à disposition.

Alors qu'il y a encore dix ans, l'océrisation d'ouvrages anciens était d'une qualité qui nécessitait des révisions majeures, voire une resaisie intégrale du texte de référence, les progrès rendus possibles par des technologies relevant de l'apprentissage automatique permettent aujourd'hui d'avoir une localisation des graphies et des taux de reconnaissance tout à fait acceptables, en particulier lorsque l'on utilise des outils spécialement entraînés sur les corpus contemporains de l'ouvrage à traiter. En l'occurrence, nous avons eu recours au logiciel de reconnaissance de caractères $Kraken$, paramétré par un modèle de langue correspondant au français du XVII^{e} siècle (Kiessling, 2019; Tanguy, 2020). La qualité des résultats obtenus tient tant aux performances du modèle qu'aux images de documents sur lesquelles ils sont appliqués, provenant de Gallica.

Gallica, la bibliothèque numérique de la Bibliothèque nationale de France (Bermes, 2020; Bertrand

& Girard, 2016), met en effet à disposition du public et des chercheurs, aux fins de consultation et téléchargement, de très nombreux ouvrages. Dans cette collection, se trouvent les œuvres de Descartes telles que le *Discours de la méthode* (Identifiant : ark :/12148/btv1b86069594). La numérisation des ouvrages qui a été réalisée permet d'accéder à des images de haute résolution, compatibles avec l'emploi d'un logiciel de reconnaissance de caractères. La reconnaissance de caractères accessible via le site en ligne de Gallica n'est pas parfaite mais des initiatives participatives laissent entrevoir une amélioration constante des ressources disponibles (Andro & Saleh, 2015).

Dans notre expérimentation, la reconnaissance de caractères proposée par Gallica n'a pas été utilisée. Nous avons utilisé celle issue du logiciel *Kraken*. Si tant est que la qualité de la reconnaissance soit au rendez-vous, nous ne nous sommes pas lancés dans une correction manuelle qui aurait été fastidieuse et contre-productive pour ce qui n'était alors qu'une expérimentation, une preuve de concept. La version TEI réalisée dans l'ANR ProDescartes constitue en effet une référence qu'il aurait été préjudiciable de ne pas considérer comme telle.

Dans Gallica aussi bien que dans *Kraken*, l'unité manipulée, traitée, est la page numérisée. À notre disposition, la sortie de *Kraken*, avec pour chaque page, les lignes reconnues, et pour chaque ligne, les *segments* ou tokens océrisés. L'unité manipulée dans l'édition numérique TEI est quant à elle l'ouvrage. Aussi, notre stratégie a-t-elle été de réaliser un alignement automatique de la version océrisée et de l'édition TEI. L'alignement automatique fait l'objet d'études depuis de très nombreuses années dans le domaine de la traduction automatique (Brown *et al.*, 1993; Och & Ney, 2003). Cette technologie peut cependant être utilisée dans des cadres autres que la traduction, à partir du moment où il s'agit de synchroniser deux flux textuels.

Les tokens produits par *Kraken* étant localisés dans la page au moyen d'une boîte englobante (i.e, *bounding box*) précisant ses coordonnées, sa hauteur et largeur, il est envisageable de retisser le lien de traçabibilité manquant dans la version TEI entre chacun des mots du texte et celui de la version numérisée originale. Pour ce faire, il nous a fallu réaliser sur le document TEI une segmentation en *mots* alignable avec celle de *Kraken*. Alors que la solution la plus satisfaisante aurait été de suivre les consignes de transcription du projet ANR, nous avons, par soucis d'efficacité, procédé à une rétro-ingénierie qui nous a permis d'accéder au plus vite aux règles les plus courantes : codage des ponctuations, substitution des s longs, réécriture des & en « et ». La sortie de *Kraken* a quant à elle été retouchée sur deux points affectant l'alignement : le tiret « - », reconnu comme non logique « ¬ », et la normalisation Unicode, pour les voyelles avec diacritique produites par composition.

La procédure de segmentation en tokens utilisée pour la version TEI isole les élisions pour produire deux tokens là où l'OCR n'en produit qu'un seul. La segmentation de la version TEI est en quelque sorte plus fine sur cet aspect. À l'inverse, la version imprimée comporte de nombreuses césures à l'origine de deux tokens dans la version océrisée, césures non reproduites dans la version numérique. La procédure d'alignement gère ces différences de segmentation en proposant des correspondances multiples, le principe directeur étant d'utiliser la segmentation de la version TEI pour alimenter le moteur de recherche avec des unités faisant davantage sens, géolocalisées avec les informations issues de l'ocrisation. Des problèmes de segmentation se posent également en diachronie et auront des répercussions lors de l'indexation des documents pour une interrogation en forme classique ou contemporaine. Des formes agglutinées en français classique se sont désagglutinées au fil du temps, et inversement (Blumenthal *et al.*, 2017) : *pource* devenu *pour ce*, *au paravant* devenu *auparavant*.

Alors que l'équipe éditoriale de la version TEI s'est attachée à restituer le plus fidèlement possible le discours de Descartes, elle n'a pas restituée l'intégralité des éléments propres aux imprimés retenus – réclames, repères de l'imposition, titres courants ou encore césures de fin de ligne et de fin de page.

La prise en compte de l'unité documentaire constituée par les imprimés a notamment nécessité, au sein du flux TEI, la reconstruction de l'unité *page* et de l'unité *ouvrage*. Pour reprendre les catégories du modèle FRBR (IFLA, 1998) issu du monde des bibliothèques, il a en effet fallu concilier – au sein d'une même instance – la description de l'*expression* de la philosophie cartésienne [contenu intellectuel] et celle de la *manifestation* [publication].

L'opération d'alignement, à proprement parlé, nécessite de tenir compte des éléments de composition et d'imposition mentionnés plus haut et qui n'avaient pas lieu d'être repris dans l'édition numérique. L'alignement ne peut donc être total puisque certains éléments n'ont pas leur correspondant dans l'édition en ligne. Les caractéristiques de la forme imprimée, présentes dans la version océrisée mais sans correspondance dans la version numérique, sont prises en charge par l'algorithme d'alignement. Les inter-titres et notes de l'auteur, placés en manchette selon la volonté de Descartes, ont été reproduits et annotés en TEI. Leur sérialisation par l'OCR pertube le flux textuel, différemment selon qu'ils apparaissent en page paire ou impaire. Ces phénomènes locaux devront être pris en charge par l'alignement automatique et gérés comme une forme de *non-parallélisme*.

L'algorithme d'alignement que nous avons conçu utilise une stratégie de type diviser-pour-régner : des séquences de tokens identiques dans les deux versions à aligner permettent de contraindre les alignements potentiels restants à traiter. Ces séquences identiques sont comparables aux *cognats*, éléments invariants, utilisés dans l'alignement automatique (Simard *et al.*, 1992; Kraif, 1999).

L'alignement ainsi obtenu, bien que perfectible, est tout à fait acceptable et nous montre la voie vers une traçabilité retrouvée entre l'édition numérique et l'édition originale. Par la suite, et comme suggéré par les relecteurs, il conviendra d'évaluer la qualité de l'alignement obtenu au regard de solutions alternatives telles que CollateX (Haentjens Dekker *et al.*, 2015).

4 Des techniques de visualisation propices à l'immersion

La section précédente nous a montré qu'il était possible de maintenir un lien permanent entre l'édition originale et l'édition numérique sous forme TEI. D'une des versions, il est possible de passer à l'autre et réciproquement. Bien plus encore, d'un mot identifié sur un fac-similé de l'œuvre, il est possible de situer son occurrence dans l'édition numérique. Il convient cependant de préciser qu'alors que la localisation dans le fac-similé passe simplement par l'intermédiaire de la boîte englobante calculée lors de l'océrisation, la localisation dans l'édition numérique n'est pas aussi immédiate. En effet, le texte au format TEI ne permet pas d'identifier chacun des mots de manière individuelle, que ce soit par le biais d'un élément du langage, ou par le biais d'un identifiant. Il est donc nécessaire de réaliser une segmentation automatique en mots du contenu textuel du document TEI et de maintenir un lien permanent entre le résultat de cette segmentation et le flux textuel du document.

Dès lors, toute action de recherche, de navigation, effectuée sur la version numérique peut être réalisée sur le fac-similé : le fac-similé peut se substituer à la version TEI. Pour ce faire, et retrouver le feuilletage de l'édition originale, nous avons choisi de mettre en œuvre le module Turnjs de (Garcia, 2012) qui permet une consultation d'un ouvrage en mode livre, en simulant l'opération de feuilletage. Si ce système a trouvé des usages dans le monde des magazines, le principe du feuilleteur peut être observé dans un contexte scientifique avec la plateforme DocExplore (Tranouez *et al.*, 2012). Ce système de présentation, à base de feuilletage, peut paraître simplement ludique de prime abord. Il permet cependant de recréer l'espace de la double page en vis-à-vis, reproduisant un visuel

comparable à celui de l'ouvrage, validé par l'auteur. Ceci facilite l'accès au contenu et la localisation dans l'ouvrage, via l'épaisseur. Enrichi d'un sommaire permettant l'accès direct aux parties et chapitres, l'expérience utilisateur devient alors plus riche que celle obtenue sur un navigateur internet classique.

Le module Turnjs dispose en outre d'un système de gestion de couches (*layer*) permettant de gérer le surlignage et l'annotation. Nous l'illustrerons dans la section suivante, pour la mise en contexte des termes et expressions recherchés.

FIGURE 2 – Le module Turnjs offre une vue double page en vis-à-vis

5 L'interopérabilité des ressources pour la consultation des sources anciennes

Les méthodes de recherche et d'analyse de fonds anciens posent de nombreuses difficultés (Heiden & Prévost, 2002). Pour le lecteur qui ne maîtrise pas le français classique, il convient de gérer une orthographe instable au fil de l'ouvrage, une segmentation en mots qui a évolué, des formes verbales méconnues (Lefeuvre, 2014). Il peut alors parfois sembler plus aisé d'interroger le fonds en français contemporain. Pour les experts mêmes, il convient, lors des recherches, que la prise en compte des variantes se fassent de la manière la plus transparente possible, que le lemme puisse être convoqué et appelle ses différentes formes, quelles que soient d'ailleurs les variantes orthographiques de ces formes. Il s'agit là d'interopérabilité des ressources lexicales.

Dans le prototype sur lequel nous travaillons, nous avons choisi d'offrir la possibilité de rechercher des mots ou expressions aussi bien en français classique, qu'en français contemporain, à partir de formes ou de lemmes, tout en gérant des variantes ou instabilités orthographiques. Pour atteindre une qualité satisfaisante du moteur de recherche, il conviendrait de réaliser un étiquetage morpho-syntaxique robuste des ouvrages en français classiques, tel (Camps *et al.*, 2021) qui se concentre sur le genre théâtral en français classique ou (Blumenthal *et al.*, 2017) qui propose un corpus diachronique du XVI^{e} au XX^{e}. À ce stade, nous avons procédé sans étiqueteur. Afin de pouvoir effectuer les recherches sus-mentionnées via l'interface de consultation, nous sommes partis du dictionnaire de Wikisource qui nous permet d'obtenir des correspondances entre français classique et français contemporain. Nous avons ensuite fait correspondre les formes contemporaines de Wikisource avec celles du GLÀFF, un Gros Lexique À tout Faire du Français (Hathout *et al.*, 2014), de manière à obtenir les lemmes des

différentes formes. Ces ressources sont utilisées par le moteur de recherche intégré pour interroger le fonds d'une manière très naturelle, avec une mise en valeur en contexte des occurrences trouvées directement sur le fac-similé, en utilisant le système de couche d'enrichissement de Turnjs, et le mécanisme de changement automatique de page du module pour passer d'une occurrence à l'autre.

Forme classique	Forme contemporaine
au parauant	auparavant
abaissemens	abaissements
abaissoient	abaissaient
abaissoit	abaissait
abandonneroit	abandonnerait
abandonneroyent	abandonneraient
abandonnoient	abandonnaient
abandonnois	abandonnais

TABLE 1 – Wikisource fournit des correspondances français classique/contemporain

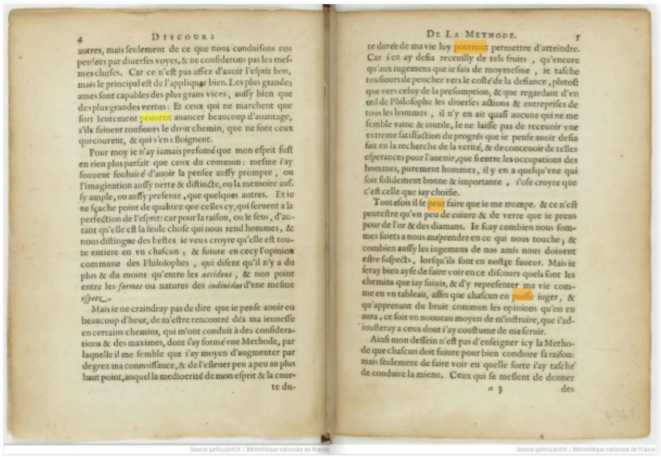

FIGURE 3 – Recherche de formes et mise en contexte des occurrences sur les pages de l'ouvrage

6 Perspectives

Dans le domaine des humanités numériques et de l'édition d'œuvres anciennes, l'influence de la Text Encoding Initiative (TEI) a porté ses fruits et n'est plus à démontrer. Le contexte technologique évolue cependant très rapidement et est propice à l'émergence de nouveaux modes de consultation et de diffusion. En travaillant à la conception d'une d'une nouvelle interface de consultation des œuvres de Descartes, des questions liées à la traçabilité des opérations, à l'interopérabilité des ressources de TAL, et à l'interprétabilité ont émergé. Nous avons montré comment un document au format TEI, constituant en tant que telle une référence, conservait une place tout à fait centrale dans un tel dispositif, les questions de traçabilité, d'interoperabilité et d'interprétabilité trouvant réponse en enrichissant cette colonne vertébrale à la fois souple et robuste, tout en permettant de récréer l'expérience d'une certaine matérialité retrouvée.

Références

ANDRO M. & SALEH I. (2015). La correction participative de l'ocr par crowdsourcing au profit des bibliothèques numériques. *Bulletin des bibliothèques de France*, p. Contribution–du.

BERMES E. (2020). *Le numérique en bibliothèque : naissance d'un patrimoine : l'exemple de la Bibliothèque nationale de France (1997-2019)*. Theses, Paris, Ecole nationale des chartes. HAL : tel-02475991.

BERTRAND S. & GIRARD A. (2016). Gallica (1997– 2016). de la bibliothèque de « l'honnête homme » à celle du gallicanaute. *Bulletin des bibliothèques de France*, **juillet**(9), 48–59.

BLUMENTHAL P., DIWERSY S., FALAISE A., LAY M.-H., SOURVAY G. & VIGIER D. (2017). Presto, un corpus diachronique pour le français des xvie-xxe siècles. *TALN 2017 : Actes de l'atelier «ACor4French–Les corpus annotés du français»(ACor4French2017)*, p. 18–26.

BROWN P. F., DELLA PIETRA S. A., DELLA PIETRA V. J. & MERCER R. L. (1993). The mathematics of statistical machine translation : Parameter estimation. *Computational Linguistics*, **19**(2), 263–311.

BUARD P.-Y. (2015). *Modélisation des sources anciennes et édition numérique* . Theses, Université de Caen. HAL : tel-01279385.

CAMPS J.-B., GABAY S., FIÈVRE P., CLÉRICE T. & CAFIERO F. (2021). Corpus and Models for Lemmatisation and POS-tagging of Classical French Theatre. *Journal of Data Mining and Digital Humanities*. DOI : 10.46298/jdmdh.6485, HAL : halshs-02591388.

CARRAUD V. (2009). Coordinateur. corpus descartes. projet d'édition en ligne des oeuvres et de la correspondance de descartes (anr prodescartes 2009-2014), programme blanc csd 9, nt09_439558.

CARRAUD V. (2011). Nihil esse certi. punto e a capo ? *Alvearium*, **juillet**(4), 49–55.

CHARTIER R. (1997). Du livre au lire. *Sociologie de la communication*, **1**, 271–290.

GALLERON I., DEMONET M.-L., MEYNARD C., FATIHA I., PIERAZZO E., WILLIAMS G., ROGER J. & BUARD P.-Y. (2018). *Les publications numériques de corpus d'auteurs - Guide de travail, grille d'analyse et recommandations (V1-Novembre 2018)*. Research report, Huma-Num. HAL : halshs-01932519.

GARCIA E. (2012).

HAENTJENS DEKKER R., VAN HULLE D., MIDDELL G., NEYT V. & VAN ZUNDERT J. (2015). Computer-supported collation of modern manuscripts : Collatex and the beckett digital manuscript project. *Digital Scholarship in the Humanities*, **30**(3), 452–470.

HATHOUT N., SAJOUS F. & CALDERONE B. (2014). GLÀFF, a Large Versatile French Lexicon. In *Conference on Language Resources and Evaluation (LREC)*, p. 1007–1012, Reykjavik, Iceland. HAL : hal-00998467.

HEIDEN S. & PRÉVOST S. (2002). Étiquetage d'un corpus hétérogène de français médiéval : enjeux et modalités. In *Romance Corpus Linguistics - Corpora and Spoken Language, Tübingen, Gunter Narr Verlag Tübingen*, p. p. 127–136. C.D. Pusch et W. Raible. HAL : halshs-00087995.

IDE N. & VÉRONIS J. (1995). *Text encoding initiative : Background and contexts*, volume 29. Springer Science & Business Media.

IFLA S. G. (1998). *Functional Requirements for Bibliographic Records : Final Report*, volume 19. International Federation of Library Associations and Institutions (IFLA). ISBN 978359811382-6 https://repository.ifla.org/handle/123456789/811.

KIESSLING B. (2019). Kraken-an universal text recognizer for the humanities. alliance of digital humanities organizations (adho), utrecht, the netherlands.

KRAIF O. (1999). Identification des cognats et alignement bi-textuel : une étude empirique. *Conférence TALN*, p. 12–17.

LEFEUVRE F. (2014). Étude grammaticale du français classique. Diachronie - 17es - grammaire, HAL : halshs-01138852.

MELOT M., Éd. (2006). *Livre*. L'Œil neuf Éditions.

OCH F. J. & NEY H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, **29**(1), 19–51. DOI : 10.1162/089120103321337421.

ROGER J. (2015). *Descartes et ses livres. L'édition comme geste philosophique*. Theses, Université de Caen Basse-Normandie. HAL : tel-03337305.

SIMARD M., FOSTER G. F. & ISABELLE P. (1992). Using cognates to align sentences in bilingual corpora. In *Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*.

TANGUY J.-B. (2020). Exploiter des modèles de langue pour évaluer des sorties de logiciels d'OCR pour des documents français du XVIIe siècle. In C. BENZITOUN, C. BRAUD, L. HUBER, D. LANGLOIS, S. OUNI, S. POGODALLA & S. SCHNEIDER, Éds., *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 3 : Rencontre des Étudiants Chercheurs en Informatique pour le TAL*, p. 205–217, Nancy, France : ATALA. HAL : hal-02786201.

TRANOUEZ P., NICOLAS S., DOVGALECS V., BURNETT A., HEUTTE L., LIANG Y., GUEST R. & FAIRHURST M. (2012). Docexplore : Overcoming cultural and physical barriers to access ancient documents. In *Proceedings of the 2012 ACM Symposium on Document Engineering*, DocEng '12, p. 205–208, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/2361354.2361399.

TSCHICHOLD J. & PARIS M. (1994). *Livre et typographie : essais choisis*. Éditions Allia.