Quel alignement textuel pour l'étude des réécritures théâtrales ? Le cas du Pédant joué de Cyrano de Bergerac (1878-1935)

Côme Saignol

26 juin 2021

Résumé

À partir d'un état de l'art présentant l'histoire et la diversité des logiciels d'alignement textuel, cet article entend décrire différentes conceptions de ce socle technique pour l'étude des réécritures théâtrales. Le cas particulier des adaptations du Pédant joué de Cyrano de Bergerac (1878-1935) est mobilisé afin d'explorer les gestes d'écritures associés aux différentes granularités de l'œuvre littéraire.

Mots-clés : Cyrano de Bergerac Le Pédant joué Alignement textuel FindAlign AutoMEDITE XML-TEI

Table des matières

Liste des figures

Liste des tableaux

Introduction

Si l'usage des outils numériques a mis du temps à s'implanter dans les recherches littéraires, force est de constater que de nombreuses initiatives voient le jour pour explorer de manière automatique les réécritures d'œuvres. La comparaison de plusieurs versions d'un texte est en effet une pratique scientifique et intellectuelle commune, en particulier dans les domaines de la philologie et de la critique génétique. Son automatisation devient un enjeu pour les équipes de recherche qui développent des outils capables d'analyser, dans un laps de temps restreint, ce qui aurait nécessité jadis une vie de travail laborieux. On parle d'*alignement* pour désigner ce socle technique qui fait désormais le miel des humanités numériques.

Définies à l'origine comme une « transdiscipline[1] » liant les traditions critiques en Arts, Langues, Lettres, Sciences humaines et sociales (ALLSHS) à la Science informatique (SI)[2] ; puis, comme une communauté de pratiques prônant « autant le faire que le dire[3] », les humanités numériques se montrent aujourd'hui prolixes dans la création de logiciels d'alignement. Plusieurs universités françaises y ont ainsi consacré des développements sur le long cours en prenant appui sur la création des instituts et Labex associant chercheurs en humanités classiques et ingénieurs informaticiens. Au-delà de l'enthousiasme et des échanges féconds nés de cette rencontre[4], les logiciels qui en sont issus semblent très variables dans leurs fonctions, à tel point que la notion d'*alignement* peut apparaître diluée ou tout du moins sujette à discussion. Après tout, comment les recherches littéraires considèrent-elles cette technologie et quel usage concret en est-il fait ? Bien que les logiciels d'alignement permettent une exploration immédiate des corpus au travers d'une annotation liminaire, l'exploitation de ces résultats dans une démarche scientifique semble plus délicate et engendre des questionnements de nature technique autant qu'épistémologique.

Le cas des réécritures théâtrales se montre catalyseur de ces difficultés. En effet, toute étude des adaptations d'une pièce de théâtre se trouve immédiatement confrontée aux problèmes

[1] Selon l'article 3 du manifeste des *Digital Humanities*, « [l]es *digital humanities* désignent une transdiscipline, porteuse des méthodes, des dispositifs et des perspectives heuristiques liés au numérique dans le domaine des Sciences humaines et sociales ». Voir M. Dacos, « Manifeste Des Digital Humanities », THAT Camp Paris, 2010.

[2] Le 9 juillet 2019, le Journal officiel publie une définition proposée par le Comité national de Terminologie de l'expression française allant dans ce sens. Selon ce dernier les humanités numériques sont un « [d]omaine de recherche et d'enseignement au croisement de l'informatique et des lettres, des arts, des sciences humaines et des sciences sociales, visant à produire et à partager des savoirs, des méthodes et de nouveaux objets de connaissance à partir d'un corpus de données numériques ». Un débat public sur la pertinence de cette définition a vu le jour dans les mois qui ont suivi, voir X.-L. Salvador, « La France Va-t-Elle En Finir Avec Les Humanités Classiques ? », *Le Figaro*, août 2019.

[3] O. Le Deuff et F. Clavert, « Petite histoire des humanités digitales », dans *Le temps des humanités digitales. La mutation des sciences humaines et sociales*, Limoges, Fyp éditions, 2014, p. 15-31.

[4] Une rencontre toutefois difficile tant le cloisonnement disciplinaire hérité du XXe siècle semble présent. Dans *La République des savants*, Jean-François Picard mettait en lumière l'état de la recherche au début du XXe siècle : « Une recherche fondamentale victime du conservatisme universitaire, mais capable à l'occasion de brillants feux individuels, une recherche appliquée pénalisée par les frilosités d'une industrie faiblement innovatrice et qui cherche déjà ses modèles outre-Atlantique, telle est brossée à grands traits la situation du pays au moment de la réalisation d'une organisation moderne de la science ». J.-F. Picard, *La République Des Savants. La Recherche Française et Le C.N.R.S.*, Flammarion, Paris, 1990, p. 29.

classiques que rencontre la critique génétique pour l'étude du théâtre, auxquels s'ajoutent en plus les limites des logiciels d'alignement. D'une part, l'hétérérogénéité des matériaux tend à complexifier la démarche critique : que faire des documents témoignant du travail de texte et de leurs multiples statuts (manuscrits d'acteurs, cahiers de metteur en scène, notes éparses)[5] ? D'autre part, les logiciels d'alignement semblent peu adaptés aux spécificités du genre, aux différentes granularités qui le composent et s'enchevêtrent de manière à former le tissu textuel ; ainsi, l'analyse des unités dramatiques semble être une piste inexplorée. Enfin, l'emploi d'un traitement computationnel pourrait accentuer une démarche textocentriste alors même que l'étude du théâtre se fonde sur une certaine « alliance entre le texte et la scène[6] » selon la formule manifeste de Bernard Dort.

Ces outils suscitent donc de réelles problématiques pour un corpus théâtral. Et sans prétendre les résoudre intégralement, les mises en scène successives du *Pédant joué* de Cyrano de Bergerac constituent un éclairant point d'observation, un laboratoire où on réfléchit plus largement sur ce que pourrait être un alignement textuel pour les réécritures théâtrales. Car si les outils existants ne parviennent à répondre aux enjeux que posent ces corpus, c'est peut-être qu'une facette de l'alignement a été peu explorée et qu'elle appelle des tentatives de réponses, tout du moins des essais techniques. C'est ce point précis qui motive l'écriture de cet article : quel alignement est-il possible de concevoir pour l'étude des réécritures théâtrales ?

Après une discussion sur l'histoire de l'alignement considérée comme un exemple d'interdisciplinarité, les lignes qui suivent proposeront une cartographie des outils existants, en s'attardant particulièrement sur la base de données FindAlign qui les référence, avant de finir par une série d'expérimentations ayant pour but de modéliser les réécritures à l'œuvre dans notre corpus.

Soulignons également que la visée est ici moins de formaliser des résultats de la recherche que de proposer un travail de contextualisation et de construction d'hypothèses, préalables à toute étude plus méthodique. Dans cette perspective, une publication au format numérique prend tout son sens, puisqu'elle autorise l'intégration des interfaces présentées au cœur même de l'article scientifique. La démarche s'inscrit ainsi dans la lignée de la « science ouverte[7] » en rendant accessible l'ensemble des données de la recherche, qu'elles soient textuelles, bibliographiques ou encore logicielles.

[5] Pour une présentation détaillée de ces enjeux, voir l'avant-propos dans A. Grésillon, M.-M. Mervant-Roux et D. Budor, *Genèses théâtrales*, Paris, France, CNRS éditions, 2010.

[6] B. Dort,« «Le Texte et La Scène : Pour Une Nouvelle Alliance » », *Encyclopædia Universalis*, 1984.

[7] Au cours de l'écriture de cet article, Madame la ministre de l'Enseignement supérieure, de la Recherche et de l'Innovation a annoncé le lancement du « 2e Plan national pour la science ouverte » qui se déploiera jusqu'en 2024. Outre l'objectif des 100 % de publications scientifiques en accès ouvert d'ici 2030, ce plan est marqué par un triplement du budget qui y est dédié. Le débat sur la science ouverte est, plus que jamais, sujet d'actualité scientifique. Voir F. Vidal, « 2e Plan National Pour La Science Ouverte », https://www.enseignementsup-recherche.gouv.fr/cid159134/la-ministre-de-l-enseignement-superieur-de-la-recherche-et-de-l-innovation-frederique-vidal-annonce-un-2e-plan-national-pour-la-science-ouverte.html, 2021.

L'alignement : histoire d'un socle technique interdisciplinaire

Si l'on souhaite envisager les méthodes d'alignement pour l'étude des réécritures théâtrales, il importe de s'abstraire un instant de l'alignement textuel qui concentre naturellement l'attention des recherches littéraires, pour envisager l'histoire des échanges interdisciplinaires autour de la notion d'alignement.

L'histoire de l'alignement débute en 1965 lorsque Vladimir Levenshstein, informaticien russe, formalise la célèbre « distance d'édition[8] » : c'est-à-dire un algorithme calculant le nombre d'étapes successives minimales pour passer d'une séquence de caractères à une autre, en les modifiant un à un. Trois opérations sont formalisées dans ce traitement, l'algorithme est capable : 1) d'insérer un caractère ; 2) de le supprimer, ou bien ; 3) de le remplacer. Afin d'illustrer ce calcul, prenons deux termes simples : « Théâtre » et « Théâtral ». La distance d'édition est ici égale à 2 puisque le passage d'un terme à l'autre se fait *a minima* avec deux opérations : on effectue tout d'abord une substitution (du caractère -e- vers le caractère -a-) suivie d'une insertion (celle du caractère -l-). En d'autres termes, l'intérêt de cet algorithme est de mesurer le coût minimal à effectuer pour passer d'un mot à l'autre. Il marque, plus particulièrement, l'intérêt de l'époque pour la comparaison automatique de séquences textuelles : la distance d'édition est en effet fille de la célèbre *théorie de l'Information* de Shannon et voit donc le jour après l'âge d'or de la cybernétique.[9]

Dans la continuité de ces travaux, notons également l'algorithme de Needleman-Wunsch[10] capable dès 1970 d'aligner des séquences de macromolécules biologiques, à partir de protéines et d'acides aminés qui les constituent. Bien qu'*a priori* très éloigné des corpus textuels, cet algorithme constitue un acte de naissance pour la bio-informatique et met au jour de nouvelles perspectives de recherche en distinguant les alignements « globaux » (sur l'ensemble de la longueur des séquences) ou « locaux » (limités à certaines régions).

Les premiers « aligneurs » voient ainsi le jour dans les années 1970 en se donnant pour objectif de comparer deux versions du code source d'un même logiciel. Le programme *Diff*, conçu par Hunt et Mcllroy en collaboration avec les laboratoires Bell, mettaient en lumière les ambitions et difficultés rencontrées par les chercheurs :

> *The program diff reports differences between two files, expressed as a minimal list of line changes to bring either file into agreement with the other. Diff has been engineered to*

[8] Nommé également « distance de Levenshstein » du nom de son auteur. Voir V. I. Levenshtein, « Binary Codes Capable of Correcting Deletions, Insertions and Reversals. », *Soviet Physics Doklady*, vol. 10, nº 8, 1966, p. 707-710.

[9] La cybernétique se définit essentiellement comme une pratique interdisciplinaire de modélisation des phénomènes de rétroaction dans les systèmes biologiques, sociaux ou cognitifs. Pour une présentation exhaustive de son histoire et de sa réception, voir R. Le Roux, *Une histoire de la cybernétique en France (1948-1975)*, Paris, Classiques Garnier, 2018.

[10] S. B. Needleman et C. D. Wunsch, « A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins », *Journal of Molecular Biology*, vol. 48, nº 3, 1970, p. 443-453.

> *make efficient use of time and space on typical inputs that arise in vetting version-to-version changes in computer-maintained or computer-generated documents. Time and space usage are observed to vary about as the sum of the file lengths on real data, although they are known to vary as the product of the file lengths in the worst case.*[11]

Si la perspective était bien de calculer le coût « minimal » du passage d'une séquence à l'autre, la longueur des fichiers analysés se montrait contraignante en termes de temps de traitement et d'espace disponible sur les machines. L'optimisation ainsi que le « passage à l'échelle[12] » des algorithmes devenaient dès lors impératifs pour l'étude des corpus textuels qui présentaient des caractéristiques bien distinctes des corpus informatiques. Tandis que pour ces derniers le vieil adage « *one line, one instruction*[13] » se vérifiait et justifiait une analyse à l'échelle de la ligne, les corpus d'écrivains invitent au contraire à changer de granularité en passant à celle du mot, voire à celle du caractère pour les études les plus fines. Dans l'étude d'une réécriture théâtrale par exemple, la moindre flexion morphologique ou l'ajout de signe de ponctuation constituent des indices importants du travail de l'écrivain que le logiciel ne peut écarter.

Ces obstacles des origines s'accentuent dans les années 1980 avec les vastes corpus multilingues constitués dans le cadre des projets transnationaux. Au-delà du bouleversement épistémologique que ces objets entraînent, les algorithmes d'alignement fournissaient difficilement des résultats probants et le passage à l'échelle apparaissait hors de portée. Il a fallu attendre le début des années 1990 pour que l'alignement connaisse un véritable tournant, lorsque de nouveaux algorithmes, plus performants, sont créés en bio-informatique. L'étude de macromolécules biologiques, riches de centaines de milliers de protéines, généraient en effet des corpus de grandes tailles, sans commune mesure avec les corpus d'auteurs qui sont de taille infiniment plus réduite. Conçu dans ce cadre disciplinaire, le logiciel BLAST (Basic Local Alignment Search Tool)[14] a rendu par exemple possible le passage à l'échelle et constitue, encore aujourd'hui, une nette avancée pour le domaine[15].

Les aligneurs se sont ainsi peu à peu imposés comme des outils précieux pour l'écriture sur ordinateur en intégrant les systèmes d'exploitation : le programme avant-gardiste Diff

[11] Nous traduisons : « Le programme Diff signale les différences entre deux fichiers, exprimées sous la forme d'une liste minimale de changements de lignes pour les mettre en correspondance l'un avec l'autre. Diff a été conçu pour permettre une utilisation efficace du temps et de l'espace sur les entrées typiques qui se présentent lors de la vérification de version à version dans les documents maintenus ou générés par ordinateur. On observe que l'utilisation du temps et de l'espace varie à peu près comme la somme de la taille des fichiers sur des données réelles, bien qu'on sache qu'elles varient comme le produit de la taille des fichiers dans le pire des cas. » J. W. Hunt et M. D. Mcllroy, « An Algorithm for Differential File Comparison », *Computing Science Technical Report*, nº 41, 1975, p. 1.

[12] En informatique, le *passage à l'échelle* désigne communément la faculté d'un algorithme à pouvoir fonctionner suivant un changement de taille ou de volume conséquent dans les données entrées.

[13] Nous traduisons : « Une instruction, une ligne ».

[14] S. F. Altschul, « Basic Local Alignment Search Tool (BLAST) », National Center for Biotechnology Information, 2021.

[15] L'article scientifique qui en est issu a été très largement cité en 1990, voir S. F. Altschul *et al.*, « Basic Local Alignment Search Tool », *Journal of Molecular Biology*, vol. 215, nº 3, 1990, p. 403-410.

constitue aujourd'hui un utilitaire important de l'écosystème GNU[16] avec la commande Unix « diff » qui fait par ailleurs l'objet d'une documentation dédiée[17]. L'ouverture au grand public est telle que la plupart des suites bureautiques (comme Word ou LibreOffice) proposent nativement des outils permettant la comparaison de deux versions d'un document ; néanmoins leur exploitation dans un cadre littéraire s'avère peu efficiente dès que les réécritures deviennent trop différentes ou de taille trop importantes[18].

Les années 2000 sont par la suite marquées par l'adaptation des nouveaux algorithmes en bio-informatique pour les recherches littéraires. Bien que le fondement algorithmique des premiers puisse être repris, les corpus textuels mettent au jour des besoins inédits en alignement qui appellent de nouveaux développements. Conçu conjointement par le laboratoire d'informatique Lip6 de l'Université Pierre et Marie-Curie[19] et l'Institut des Textes et Manuscrits Modernes (ITEM) de l'École Normale Supérieure[20], le logiciel MEDITE[21] était ainsi prototypique de cette volonté d'adapter des algorithmes issus de la bio-informatique. Afin d'analyser les manuscrits et brouillons d'écrivains et de mettre au jour les processus de création à l'œuvre, l'équipe a déplacé les trois opérations classiques de la distance d'édition (insertion / suppression / remplacement) vers les quatre gestes scripturaires fondateurs de la critique génétique. Il a donc fallu intégrer le déplacement comme une nouvelle opération à annoter dans les logiciels, au prix d'une recherche coûteuse qui se poursuit encore de nos jours.

Le début des années 2010 voit enfin l'amorce d'une réflexion sur les questions de visualisation et d'interopérabilité pour les aligneurs qui tendent à devenir des plateformes d'édition à part entière. L'enjeu n'est plus seulement d'analyser des réécritures et de présenter des annotations dans une interface, il s'agit aussi d'éditer des corpus ayant circulé sous plusieurs versions : on serait tenté ici de parler de textes en *situation de variance*. La plateforme éponyme Variance[22], en collaboration avec la maison d'édition Slaktine[23], fournit par exemple à ses partenaires les moyens techniques d'interpréter des transformations textuelles par le biais d'un alignement éditorialisé. Plus généralement, la perspective tend à s'ouvrir sur les nouveaux *big corpus*[24]

[16] GNU désigne une famille de logiciels d'exploitation disponibles en libre accès. Ils se fondent sur le noyau Linux créé en 1991 par Linus Torvalds.

[17] D. MacKenzie, P. Eggert et R. Stallman, « Comparing and Merging Files », GNU Operating System, 2018.

[18] Dans le cadre du développement de l'aligneur MEDITE, Julien Bourdaillet et Jean-Gabriel Ganascia ont notamment comparé les résultats obtenus par leur logiciel et WORD. Voir J. Bourdaillet et J.-G. Ganascia, « MEDITE:A Unilingual Textual Aligner, Finland », dans *Advances in Natural Language Processing: 5th International Conference on NLP, FinTAL 2006*, Turku, Springer, 2006, p. 458-469.

[19] Depuis le regroupement des universités Paris-Sorbonne (Paris-IV) et Pierre-et-Marie-Curie (Paris-VI), le laboratoire Lip6 a intégré la Faculté des Sciences et de l'Ingénierie de l'université Sorbonne Université. Voir son site officiel : https://www.lip6.fr/.

[20] Le site officiel de l'ITEM présentant l'actualité du laboratoire et en particulier le calendrier des séminaires est accessible à l'adresse suivante : http://www.item.ens.fr/.

[21] Le nom MEDITE trouve ainsi son inspiration dans l'EDITE, la fameuse machine imaginée par Claude Shannon.

[22] R. Mahrer et J. Zufferey, « Variance », s. d.

[23] Slatkine est une maison d'édition établie à Genève et spécialisée dans les travaux en recherches littéraires, et plus particulièrement en Littérature française et en philologie. Leur catalogue est consultable à l'adresse suivante : https://www.slatkine.com/fr/homecategory/editions-slatkine.

[24] Par opposition aux *big datas* dont l'appellation a pour origine les recherches en sciences des calculs des données

afin d'envisager les phénomènes de circulation, de réemplois voire de plagiats à grande échelle à partir des standards en édition numérique.

L'histoire de l'alignement souligne ainsi une tradition riche d'échanges interdisciplinaires associant les sciences naturelles, la science de l'informatique et les recherches littéraires. Elle met également au jour la valeur heuristique de l'alignement, c'est-à-dire la possibilité pour le concept de susciter de la recherche fondamentale, tout en constituant une passerelle entre plusieurs champs disciplinaires. Ainsi la constitution de nouveaux corpus comme les réécritures théâtrales stimule de nouvelles interrogations scientifiques, et donc de nouveaux besoins logiciels. De la même manière que la philologie tirait son inspiration au XIXe siècle des arbres de la phylogénétique pour constituer les *stemma codicum*,[25] les recherches littéraires se sont emparées des dernières avancées en bio-informatique pour la comparaison automatique de séquence textuelle. La communauté de pratiques qu'on nommera *humanités numériques littéraires* est aussi née en partie de ces échanges.

Des typologies et définitions instables pour les recherches littéraires

Considérant cette histoire, les champs d'application de l'alignement sont vastes pour la recherche en humanités classiques : on l'a vu pour l'étude des réécritures dont la collation des variantes se trouve facilitée ; mais elle est aussi précieuse pour les études de réceptions, notamment médiatiques, qui trouvent leur intérêt dans l'analyse de réemplois de grande ampleur. La communauté scientifique est unanime en ce qui concerne l'intérêt de ces logiciels : ils ouvrent de nouvelles perspectives en même temps qu'ils constituent un appui précieux à la recherche. Certains aligneurs comme Juxta[26] proposent par exemple un accompagnement à la rédaction d'apparats critiques en exportant les résultats de l'annotation au format HTML, immédiatement lisible dans un navigateur Web[27]. Bien que cette fonctionnalité soit pertinente dans une démarche éditoriale, elle reste néanmoins à la marge des pratiques courantes des chercheurs. Pour beaucoup, l'alignement textuel se résume au temps gagné lors de l'annotation.

Cette conception de l'alignement se fait néanmoins au détriment d'une réflexion sur les fonctions et présupposés théoriques de chaque logiciel. Leur usage apparaît en réalité peu anodin étant donné que ces outils transforment en profondeur les méthodes et processus d'établissement des textes, en y insufflant des gestes issus d'autres champs disciplinaires.

et en intelligence artificielle, les *big corpus* renvoient davantage aux données textuelles constituées dans le cadre des humanités numériques. En ce sens, la notion invite davantage à considérer l'établissement des corpus en lien avec les pratiques d'éditions et d'organisation conceptuelle propres aux recherches littéraires.

[25] Un *stemma codicum* est un arbre généalogique présentant les filiations entre les différentes sources manuscrites d'une même œuvre. Cette représentation graphique trouve son inspiration dans les arbres de la phylogénétiques tels que conceptualisés par Darwin. Voir C. Darwin, *On the Origin of Species*, London, John Murray, 1859, p. 116-117. Et en particulier l'arbre suivant : https://en.wikipedia.org/wiki/On_the_Origin_of_Species#/media/File:Darwin_divergence.jpg.

[26] D. Wheeles et K. Jensen, « Juxta », 2021.

[27] D. Wheeles et K. Jensen, « Juxta Commons », dans *Proceedings of the Digital Humanities*, s. l., Alliance of Digital Humanities Organizations (ADHO), 2013.

La technologie incarne de manière extrêmement vivante les débats que pose la nouvelle « philologie numérique[28] » et notamment la place qu'elle occupe dans les réflexions sur les pratiques de la recherche.

On remarque ainsi que les publications scientifiques en langue française se montrent peu loquaces sur les multiples origines de l'alignement : le propos se restreint de manière significative à l'alignement textuel, objet il est vrai de nos disciplines, mais qui est en réalité nourri par une certaine conception de l'interdisciplinarité. Dès lors il serait tentant de proposer une première définition de l'alignement textuel qui désignerait : *la mise en correspondance de séquences textuelles par le biais d'un traitement computationnel*.

La proposition écarte pourtant toute particularité disciplinaire et les usages concrets faits par les chercheurs. Afin de surmonter cette difficulté de définition, Jean-Gabriel Ganascia et Jean-Louis Lebrave distinguaient deux types de logiciels d'alignement dans un article dressant le bilan de plusieurs décennies de collaboration scientifique autour des manuscrits littéraires[29]. Les premiers seraient davantage centrés sur « l'analyse et le traitement des données » ; les seconds sur la « structuration et la visualisation des résultats de l'alignement ». Bien que cette typologie soit commode, elle nous permet par exemple de considérer The Versioning Machine[30] (un logiciel centré sur l'édition de plusieurs *témoins* textuels[31]), et donne davantage matière à penser la variabilité de cette grande famille, elle ne rentre pas dans les détails du fonctionnement standard d'un aligneur.

Pourtant, il est possible de discerner des traitements communs d'un logiciel à l'autre comme le soulignent les auteurs de CollateX.[32] Pour ces derniers, un alignement textuel procéderait successivement par :

1. **Tokénisation**. C'est-à-dire le découpage du texte en plusieurs unités textuelles minimales, appelées « tokens ».
2. **Normalisation**. La correction des textes en supprimant les informations peu pertinentes et génératrices de bruits (la ponctuation par exemple).
3. **Alignement**. La mise en correspondance de séquences textuelles identiques, certains logiciels mettant en place plusieurs étapes consécutives.

[28] Jean-Baptiste Camps la définit de la manière suivante : « La philologie numérique peut alors être définie comme une transformation dans les méthodes d'établissement du texte et d'analyse, par l'intégration d'outils computationnels, dès que ceux-ci peuvent apporter des gains dans le processus d'établissement du texte (gains de temps, de finesse, de granularité dans la transcription, la collation…), sa représentation (enrichissement par la représentation de phénomènes graphiques, linguistiques, sémantiques…, et enregistrement des opérations éditoriales dans l'édition même) ou son analyse scientifique. » Voir : J.-B. Camps, « Où va La Philologie Numérique ? », *Fabula-LhT [Le Moyen Âge pour laboratoire]*, nº 20, janvier 2018.

[29] J.-G. Ganascia et J.-L. Lebrave, « Trente Ans de Traitements Informatiques Des Manuscrits de Genèse », dans *Critique Génétique. Concepts, Méthodes, Outils*, Paris, IMEC éditeur, 2009.

[30] S. Schreibman, « The Versioning Machine », 2021.

[31] Dans le sens philologique du terme, en anglais *witness*. Voir S. Schreibman, A. Kumar et J. McDonald, « The Versioning Machine », *Literary and Linguistic Computing*, vol. 18, nº 1, 2003, p. 101-107.

[32] T. I. D. Group, « CollateX », 2021.

4. **Analyse**. L'analyse de l'ensemble du corpus soumis à l'étude.
5. **Visualisation**. La visualisation des résultats de l'analyse à l'aide d'une interface dédiée.

Cette définition efficace hérite directement du domaine du Traitement Automatique des Langues Naturelles (TALN) dont elle emprunte les méthodes et la terminologie. Pour l'étude de réécritures théâtrales, un aligneur pose par exemple la question de la « normalisation » du texte si l'œuvre originale et la réécriture comportent des états de langues divers. Néanmoins la définition proposée fait l'impasse sur les socles techniques utilisés par ces logiciels et appelle une définition générale, capable de rassembler les implicites des recherches littéraires. L'alignement textuel pourrait alors être défini comme : *la mise en correspondance de segments textuels entre une ou plusieurs langues naturelles par le biais d'un traitement computationnel, afin d'annoter leurs variations ou similitudes*.

L'ajout de deux distinctions par rapport à notre première définition permet de cartographier les contours disciplinaires au sein des recherches littéraires, et donc de délimiter les périmètres d'usage de chaque logiciel. La mention d'« une ou plusieurs langues naturelles » distingue à terme les *corpus monolingues* appropriés pour l'étude d'une œuvre littéraire singulière — comme c'est le cas généralement en Littérature française — des *corpus multilingues*, davantage circonscrits dans le périmètre de la Littérature comparée. Ensuite, la séparation entre le relevé de « variations ou [de] similitudes » distingue d'une part les logiciels capables d'aligner les textes deux à deux — et donc d'annoter finement leurs variations[33] — avec, d'autre part, les aligneurs relevant les segments communs dans des corpus de grandes tailles. Elle permet à elle seule d'envisager le travail de collation de variantes (dont la méthode se rapproche davantage des études philologiques ou génétiques), sans la confondre avec l'analyse de réemploi à grande échelle (qui touche généralement les études de réception ou d'intertextualité).

Table 1: Usages de l'alignement textuel dans les recherches littéraires

	VARIATION	**SIMILITUDE**
MONOLINGUE	Philologie et/ou critique génétique	Intertextualité
MULTILINGUE	Littérature comparée	Études de réception

Le tableau ci-dessus résume de manière schématique ces répartitions : il constitue une hypothèse de départ devant être complétée par une enquête sur les usages réels des logiciels par les projets de recherche. Il invite également à penser les logiciels d'alignement textuel comme un socle technique habité par des implicites théoriques et des conceptions du texte qui ne sont pas nécessairement le propre des recherches littéraires.

[33] Notons également que le logiciel Ediff autorise la comparaison entre trois fichiers, et non deux comme l'immense majorité des logiciels. Voir M. Kifer, « Ediff User's Manual », 1998.

Une difficulté qui subsiste est d'arriver à identifier les logiciels adaptés aux problématiques de la recherche. Pour l'étude de réécritures théâtrales, le chercheur serait à la recherche d'un aligneur « monolingue » capable de relever les « variations » entre les textes. Mais quels logiciels répondent à cette définition ? Comment les découvrir alors même que ces connaissances échappent généralement au cœur de métier du chercheur ?

FindAlign ou une cartographie des logiciels d'alignement

C'est dans cette perspective que FindAlign[34] constitue une ressource précieuse. Pensé à l'origine comme un état de l'art des logiciels d'alignement au cours d'un travail de doctorat, le projet prend désormais la forme d'un moteur de recherche explorant leurs fonctionnalités à l'aide d'une base de données SQLite[35] ouverte. Lors de sa constitution, il a été décidé de répertorier les logiciels propriétaires autant que ceux en accès libres et ouverts, afin d'offrir une vision d'ensemble du sujet. Même si l'essor de la science ouverte invite les chercheurs à privilégier ces derniers, une équipe de recherche peut être amenée à utiliser des logiciels propriétaires pour des raisons externes (compétence et formation de l'équipe, équipements du laboratoire, etc.).

La base de données se donne ainsi pour objectif de répertorier les logiciels et de faciliter leur découverte suivant les besoins de la recherche. Pour y parvenir, FindAlign est structurée suivant 17 critères contrôlés à différentes étapes de leur élaboration[36] :

Table 2: Critères et valeurs de la base de données FindAlign

#	Critère	Champ	Valeurs possibles
1.	Nom de l'aligneur	*name*	Texte
2.	Type de logiciel	*type*	Propriétaire ; Open source
3.	Type de données	*data*	Textuel ; Organique ; Code source
4.	Langue des données	*langue*	Monolingue ; Multilingue
5.	Grain analysé	*grain*	Phrase ; Ligne ; Mot ; Caractère
6.	Traitement des données	*process*	Variation ; Similitude
7.	Interface d'utilisation	*interface*	GUI[37] ; Console

[34] C. Saignol, « FindAlign », 2021. Les données brutes du projet sont en accès ouvert sur GitHub : https://github.com/comesaignol/autoMedite/findAlign.

[35] SQLite est un système de gestion de base de données (SGBD) relationnelle accessible par le langage SQL. Sa particularité est de regrouper l'ensemble des données stockées dans la base au sein d'un unique fichier.

[36] La base de données SQLite est en effet alimentée par classeur OpenDocument exporté au format CSV. C'est ensuite un script Python qui se charge de transférer les données brutes dans la base, il vérifie que chacune des valeurs transmises est conforme à la documentation.

#	Critère	Champ	Valeurs possibles
8.	Visualisation des résultats	*visualisation*	Multifenêtrage ; Coloration syntaxique ; Graph ; Apparat critique ; Aucune
9.	Format d'entrée	*input*	TXT ; XML ; Binaire ; Image ; Répertoire
10.	Export des annotations	*output*	HTML ; XML-TEI ; Aucun
11.	Institution	*institution*	Texte
12.	Première version	*release*	Texte
13.	Dernière version	*version*	Texte
14.	Site web officiel	*site*	Texte
15.	Téléchargement	*download*	Texte
16.	Tutoriel et démo	*demo*	Texte
17.	Bibliographie	*bibliography*	Texte

Dès lors que les paramètres souhaités ont été sélectionnés et validés dans le moteur de recherche, les résultats de la requête se présentent sous la forme d'une fiche synthétique mettant en lumière les spécificités de chaque logiciel. À l'heure actuelle, différentes informations utiles comme l'institution d'appartenance, la date de la dernière version, les liens de téléchargement ou encore les références bibliographiques font l'objet d'une veille scientifique continue qui encourage les collaborations inter-établissements.

Figure 1: Exemple de requête dans FindAlign

Mais au-delà de l'intérêt socioéconomique que cette base représente, que nous apprend FindAlign sur les logiciels d'alignement ?

[37] En informatique, l'acronyme GUI renvoie à une interface graphique, c'est-à-dire un dispositif visuel permettant à l'être humain d'interagir avec une machine. Il est issu du terme anglais *Graphical User Interface* (= GUI).

En premier lieu, il existe une frontière nette entre les logiciels issus de la recherche privée qui disposent d'une forte visibilité, et ceux développés au sein de la recherche publique dont la maintenance et la communauté d'utilisateurs apparaissent très fragiles. L'exemple des documentations est à cet égard emblématique : quand elles existent, elles sont parcellaires ou non traduites pour les logiciels issus de la recherche académique. Celle du logiciel TUSTEP (*Tuebingen System of Text Processing tools*)[38], qui a pourtant fait l'objet de nombreuses publications, n'est par exemple disponible qu'en langue allemande, alors que les logiciels issus de la recherche privée disposent d'une page Wikipédia comparant leur fonctionnalités.[39] Dans cette liste riche en ressources, on découvre en particulier ExamDiff Pro[40], un logiciel propriétaire qui propose un ensemble de fonctionnalités avancées couplé à un temps de traitement rapide. Inconnu du monde académique, ce type de logiciel pourrait remplir de nombreux services pour une équipe de recherche en humanités classiques.

Outre ces obstacles inséparables du mode de financement de la recherche publique en France[41], la base de données FindAlign nous apprend également que les fonctions propres à un usage scientifique, comme la gestion des standards de l'édition, sont sous-représentées par rapport à celles envisagées pour le développement logiciel. Nous ne pouvons guère nous en étonner : le socle technique utilisé pour l'alignement de code source a une origine très lointaine comparé à celui utilisé dans les recherches littéraires. Pour ces dernières, l'heure semble aujourd'hui au développement d'outils *ad hoc* capables de répondre aux questionnements toujours plus précis des chercheurs. Odysseus[42] est par exemple un logiciel développé dans le cadre d'un doctorat qui se donnait pour objectif d'aligner les différentes traductions de *L'Odyssée* en utilisant les noms de divinités comme pivot. La perspective s'inscrivait ainsi au carrefour des humanités numériques et de la traductologie.

En dernière analyse, FindAlign nous montre finalement que le calcul de statistiques dynamiques, indispensables pour la formalisation de résultats scientifiques, est aujourd'hui inexistant dans les logiciels d'alignement. C'est pourtant là un enjeu central pour l'étude de textes fortement structurés comme le sont les réécritures théâtrales. La possibilité de quantifier l'ajout ou la suppression de texte par personnages s'avérerait éclairante pour la compréhension des adaptations du *Pédant joué* de Cyrano de Bergerac.

D'une granularité à l'autre : Le Pédant joué par AutoMedite

Publié en 1654 en seconde partie d'un recueil de lettres, *Le Pédant joué* est généralement considéré comme une des premières comédies en prose du XVIIe siècle. Cette pièce

[38] University of Tübingen, « Tuebingen System of Text Processing Tools (TUSTEP) », 2021.

[39] Anonyme, « Comparison of File Comparison Tools », dans *Wikipédia*, https://en.wikipedia.org/wiki/Comparison_of_file_comparison_tools, 2021.

[40] PrestoSoft, « ExamDiff Pro », 2021.

[41] Le mode de financement par projet apparaît en particulier un obstacle majeur à la maintenance des logiciels informatiques.

[42] M. Reboul, « Odysseus », 2021.

régulière[43] met en scène un pédant de campagne, le proviseur Granger, qui se trouve confronté à trois personnages se disputant la main de sa fille : Châteaufort (un matamore), Gareau (un paysan) et La Tremblaye (un gentilhomme). Au-delà d'une réception critique qui fait peu état de l'histoire de sa mise en scène,[44] on dispose de trois témoins mettant en lumière des adaptations originales :

1. Un manuscrit de metteur en scène pour l'adaptation du Théâtre de la Gaîté-Lyrique en 1878[45].
2. Une édition imprimée pour l'adaptation de l'Université d'Harvard en 1899[46].
3. Et, enfin, un tapuscrit d'acteur pour l'adaptation de la Radio Tour Eiffel en 1935[47].

Une première lecture des documents témoigne d'un grand nombre de suppressions pour chacune de ces adaptations, ce qui nous a interrogé : réécrire une pièce du XVIIe siècle, est-ce nécessairement en supprimer la moitié ? Comment étudier ces suppressions massives par le biais d'un traitement automatique ? Pour répondre à ces questions, deux choix méthodologiques se sont rapidement imposés : l'analyse des réécritures pouvait être envisagée à partir d'un alignement textuel ; ce qui appelait nécessairement une éditorialisation des textes par la mise à plat de leurs aspérités. Ce dernier choix, difficile à accepter dans une optique philologique, s'expliquait aussi par le besoin de dépasser les contraintes liées aux divers états de la langue qui ne constituent pas le cœur de cette étude.

Après avoir transcrit de manière diplomatique les documents, nous avons rédigé un protocole de modernisation inspiré des normes en vigueur dans l'édition du théâtre de la première modernité. Il prenait le parti d'établir l'orthographe contemporaine tout en conservant la ponctuation et les majuscules. À ce premier protocole a été adjoint un second, dit de neutralisation, dont la vocation était de limiter le nombre de variantes non significatives entre les textes : la numérotation des actes et des scènes, ou encore l'écriture des didascalies, respectent ainsi les mêmes conventions d'une adaptation à une autre. Nous avons également uniformisé la forme des didascalies ainsi que les noms des personnages quelles qu'en soient les variations (Corbinelli avec un unique -l- au lieu de deux par exemple). L'édition a enfin suivi un standard de l'édition électronique le XML-TEI[48] suivant les recommandations (*guidelines*) de la TEI P5 et plus particulièrement celles de la

[43] Dans le sens où elle suit les règles du théâtre classique telles qu'elles furent théorisées par des auteurs du XVIIe siècle comme La Ménardière ou encore l'Abbé d'Aubignac.

[44] H. S. de Cyrano de Bergerac, *Œuvres Complètes. Théâtre*, A. Blanc (éd.), Paris, Honoré Champion, 2001, vol. III.

[45] H. S. de Cyrano de Bergerac, « Le Pédant Joué », 1878.

[46] S. de C. de Bergerac, *Le Pédant Joué*, H. B. Stanton et F. Bôcher (éd.), Boston, J. de Peiffer, 1899.

[47] S. de C. de Bergerac, « Le Pédant Joué », Radio Tour Eiffel, 1935.

[48] La TEI (*Text Encoding Initiative*) est une recommandation pour l'encodage de documents textuels créés en 1987. La dernière version (P5) voit le jour en 2007 et est utilisée dans de nombreux projets de recherche en humanités numériques.

TEI Lite[49] capables de satisfaire « *90 % of the needs of 90 % of the TEI user community*[50] ».

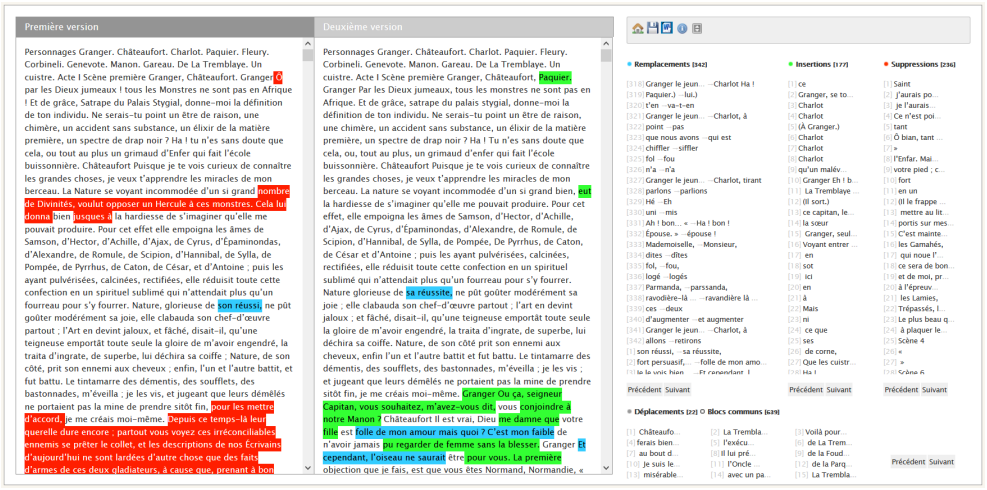

Figure 2: Alignement textuel du *Pédant joué* avec MEDITE (1654-1878), 6 janvier 2020.

Concernant le choix du logiciel, notre intérêt s'est porté sur MEDITE, un « logiciel monolingue avec recherche de déplacement[51] ». La version disponible en ligne compare automatiquement deux versions d'une même œuvre à partir de paramètres réglant la sensibilité de l'algorithme selon les besoins de la recherche (à la casse ; aux signes diacritiques ; aux séparateurs) et selon deux grains distincts (au caractère ou au mot). Une fois ces paramètres sélectionnés, MEDITE opère alors en deux temps : il détecte tout d'abord les similitudes nommées *homothéties*, c'est-à-dire les séquences textuelles communes entre les deux textes, avant de noter les remplacements, insertions, suppressions et déplacements. L'interface présente enfin les résultats de l'alignement sous trois colonnes avec à gauche la version source avec les suppressions colorisées en rouge ; au centre, la version cible avec les ajouts colorisés en vert ; et à droite, l'énumération des variantes détectées sous forme de listes. Pour les remplacements et déplacements, les données ont été colorisées dans les deux versions, en bleu pour les premiers et en gris pour les seconds.

Pour un premier essai de comparaison, nous avons coché l'ensemble des paramètres, mais il s'est avéré que les résultats obtenus comportaient un bruit trop conséquent. Nous avons alors décoché les options « Sensible à la case » et « Sensible aux séparateurs diacritiques (éêçè…) » et coché « Sensible aux signes diacritiques (éêçè…) » en choisissant le grain « mot ».

[49] La TEI Lite est une restriction de la TEI P5 afin de la rendre plus accessible à la communauté des chercheurs. Elle est en ce sens allégée (= *lite*) des recommandations plus spécifiques qui ne concernent qu'un faible nombre de projets scientifiques. Voir L. Burnard et C. M. Sperberg-McQueen, « La TEI Simplifiée : Une Introduction Au Codage Des Textes Électroniques En Vue de Leur Échange », *Cahiers GUTemberg [TEI : Text Encoding Initiative]*, nº 24, juin 1996, p. 23-151

[50] Nous traduisons : « 90 % des besoins de 90 % de la communauté d'utilisateurs de la TEI ». Voir *Id.*

[51] Pour une présentation exhaustive des fondements algorithmiques de ce logiciel, voir J. Bourdaillet, « Alignement Monolingue Avec Recherche de Déplacements Pour La Critique Génétique », *Traitement Automatique des Langues*, vol. 50, 2009, p. 61-85.

Le nouvel alignement présenté ci-dessus offre des résultats intéressants avec par exemple la détection d'une variation significative : la disparition du ô vocalique dans la première réplique de la pièce. Elle fournit en plus des résultats bruts sur l'ensemble de l'adaptation : MEDITE relève 342 remplacements, 177 insertions, 256 suppressions et 22 déplacements.

Si ces résultats permettent d'obtenir des statistiques de première main, deux limites apparaissent immédiatement. Les statistiques relevées sont tout d'abord inexploitables en l'état, étant donné que MEDITE comptabilise comme geste scripturaire tout bloc textuel, sans le pondérer par sa taille. À un signe de ponctuation supprimé est accordé, par exemple, le même « poids » que la suppression d'une scène entière, ce qui va à l'encontre de la valeur qu'on lui accorde généralement : le geste de supprimer une scène apparaît en effet très significatif en ce qu'il constitue une réécriture large et rare dans la réduction du texte. Autrement dit, il apparaît impossible de déduire des résultats bruts de MEDITE le geste scripturaire principal d'une adaptation. La deuxième limite du logiciel concerne l'export des résultats de la comparaison sous un format structuré, le XML-TEI par exemple. On aurait attendu la possibilité de travailler ces annotations avec les méthodes issues du Traitement Automatique des Langues Naturelles (TALN) comme avec un étiquetage morpho-syntaxique[52] qui serait lui-même automatique ; or MEDITE ne possède pas cette fonction qui aurait aidé à l'organisation de sessions de travail.

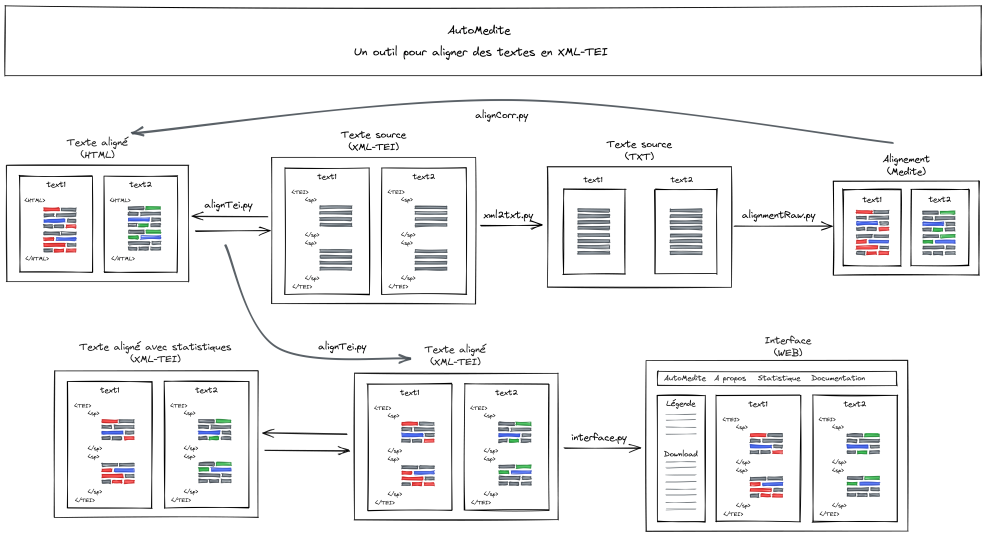

Figure 3: Fondements algorithmiques d'AutoMedite

Il a donc fallu étendre l'usage normal du logiciel par-delà ces contraintes techniques. La principale procédure a été d'enregistrer les résultats visualisés dans la page Web et de les transférer dans la version source éditée au format XML-TEI. Pour ce faire, nous avons

[52] L'étiquetage morpho-syntaxique désigne le processus associant les informations grammaticales (genre et nombre) à chaque terme d'un texte. De nombreux outils existent comme TreeTagger par exemple, voir https://www.cis.uni-muenchen.de/schmid/tools/TreeTagger/.

procédé à une tokénisation, c'est-à-dire une division du texte en tokens ou unités textuelles (mots) ; puis, nous avons remplacé les tokens de l'édition XML-TEI par les tokens HTML préalablement annotés. Ce transfert d'annotation nous a permis d'obtenir un texte édité comportant, pour chaque token, une annotation précisant s'il s'agit d'un mot remplacé, inséré, supprimé ou déplacé.

L'étape de la visualisation a ensuite été l'occasion d'imaginer une interface de lecture des éditions XML-TEI annotées, associée à la consultation de statistiques. Les scripts Python développés à cette fin ont été rassemblés, formalisés, documentés et enfin mis à disposition de la communauté scientifique sous le nom d'AutoMedite au cours de l'année 2021[53]. Son principal intérêt est, comme nous l'avons vu, d'étendre les fonctionnalités du logiciel MEDITE d'origine : il s'agit donc d'une surcouche applicative (figure 3).

Le transfert des annotations de l'interface HTML dans la structure XML-TEI autorise l'établissement de statistiques et donc l'étude de modèle de réécritures propre à chaque adaptation du *Pédant joué*. Si ces dernières sont en apparence similaires en procédant à une importante réduction du texte, elles se distinguent toutefois par la densité moyenne des réécritures. La taille moyenne des suppressions est par exemple deux fois plus importante pour la version de 1878 que de celle de 1899. En d'autres termes, les suppressions de la version de 1878 sont en moyenne « deux fois plus grandes » que celles de la version de 1899, ce qui témoigne d'un geste de réécriture plus large de la part des compositeurs (figure 2).

Accompagnée des éléments de la terminologie génétienne[54], l'analyse statistique nous invite à poser des hypothèses pour qualifier ces modèles de réécritures :

- L'adaptation du théâtre de la Gaîté-Lyrique (1878) proposerait des suppressions par amputation à l'échelle de la scène.
- L'adaptation de l'université d'Harvard (1899) des suppressions par élagage à l'échelle de la tirade.
- Et enfin, pour l'adaptation de la radio Tour Eiffel (1935), des corrections stylistiques à l'échelle infra-phrastique[55].

Conclusion

À l'heure où se multiplient les formations qui proposent une composante d'humanités numériques, cette ébauche d'une histoire de l'interdisciplinarité par le prisme de l'alignement invite les chercheurs à davantage s'intéresser à ce socle technique dans sa diversité[56]. Si nous nous focalisons sur l'alignement textuel, la distinction entre les outils

[53] C. Saignol, « AutoMedite », 2021.
[54] G. Genette, *Palimpsestes*, Seuil, Paris, 2014, p. 266.
[55] L'étude statistique détaillée de ce corpus fera l'objet d'une publication dédiée courant l'année 2022.
[56] C'est dans cette direction que nous souhaitons poursuivre nos travaux sur la notion d'alignement : l'alignement d'image, absent de cet article, constitue par exemple un enjeu à part entière pour l'étude des ouvrages à destination de

qui autorisent la comparaison des variantes deux à deux et ceux qui analysent les similitudes dans un corpus de grande taille est un critère central pour les recherches littéraires, et permet déjà de qualifier des méthodes de recherche. La liste d'outils répertoriés dans FindAlign se veut une contribution concrète éclairant les discussions et enjeux de développement associés à la notion d'alignement qui appelle une poursuite collaborative[57].

Concernant le cas particulier des réécritures théâtrales, l'analyse de la granularité de l'œuvre est un enjeu présentant un double intérêt pour la recherche : heuristique, pour le transfert des annotations dans un format structuré ; mais aussi théorique pour penser le propre de l'écriture dramatique. Au sujet des réécritures du *Pédant joué* de Cyrano de Bergerac, il est ainsi mis au jour un principe général de réduction qui n'est pas sans faire écho à un passage de la pièce. Dans la scène finale, Charlot trompe son père Granger en jouant une comédie avec l'aide de son valet. Granger, abusé par « le fourbe » Corbineli s'étonne alors de la brièveté de la pièce représentée, qui n'était qu'un stratagème pour le duper :

> GRANGER
> Comment, marier, c'est une comédie ?
> CORBINELI
> Hé bien, ne savez-vous pas que la conclusion d'un poème comique est toujours un mariage ?
> GRANGER
> Oui, mais comment serait-ce ici la fin, il n'y a pas encore un acte de fait.
> CORBINELI
> Nous avons uni tous les cinq en un, de peur de confusion : cela s'appelle pièce à la polonaise.[58]

Si la référence fait aujourd'hui défaut au lecteur moderne — le mariage « à la polonaise » est une allusion historique seulement explicite à l'aide d'une note de bas de page — elle peut nous aider à qualifier ce geste de réécriture singulier qui réduit la longueur de la pièce. Par ses coupures très frustes du texte, l'adaptateur de la version de 1878 aurait ainsi fait de la comédie une pièce « à la polonaise » à la manière de l'ingénieux Corbineli. La nouvelle adaptation prend ainsi le risque de perdre en clarté par une composition malmenée, ce qu'elle espérait gagner par une concentration de l'intrigue.

Bibliographie

ALTSCHUL Stephen F., « Basic Local Alignment Search Tool (BLAST) », National Center for Biotechnology Information, 2021.

la jeunesse. Voir notamment une analyse du logiciel *Beyond Compare* : A. MALONEY, « Comparing Nearly Identical Images Using "Beyond Compare" », *Journal of Forensic Identification*, vol. 63, n° 2, 2013, p. 153-164.

[57] La perspective de compléter la base par des tutoriels et notices d'installation en langue française est par exemple un enjeu important pour la suite de la recherche.

[58] H. S. de CYRANO DE BERGERAC, *Œuvres Complètes. Théâtre*, *op. cit.*, p. 181.

Altschul Stephen F., Warren Gish, Webb Miller, Eugene W. Myers et David J. Lipman, « Basic Local Alignment Search Tool », *Journal of Molecular Biology*, vol. 215, nº 3, 1990, p. 403-410.

Anonyme, « Comparison of File Comparison Tools », dans *Wikipédia*, https://en.wikipedia.org/wiki/Comparison_of_file_comparison_tools, 2021.

Bergerac Savinien de Cyrano de, « Le Pédant Joué », Radio Tour Eiffel, 1935.

Bergerac Savinien de Cyrano de, *Le Pédant Joué*, H. B. Stanton et Ferdinand Bôcher (éd.), Boston, J. de Peiffer, 1899.

Bourdaillet Julien, « Alignement Monolingue Avec Recherche de Déplacements Pour La Critique Génétique », *Traitement Automatique des Langues*, vol. 50, 2009, p. 61-85.

Bourdaillet Julien et Jean-Gabriel Ganascia, « MEDITE:A Unilingual Textual Aligner, Finland », dans *Advances in Natural Language Processing: 5th International Conference on NLP, FinTAL 2006*, Turku, Springer, coll. « Lecture Notes in Computer Science », 2006, p. 458-469.

Burnard Lou et C. Michael Sperberg-McQueen, « La TEI Simplifiée : Une Introduction Au Codage Des Textes Électroniques En Vue de Leur Échange », *Cahiers GUTemberg [TEI : Text Encoding Initiative]*, nº 24, juin 1996, p. 23-151.

Camps Jean-Baptiste, « Où va La Philologie Numérique ? », *Fabula-LhT [Le Moyen Âge pour laboratoire]*, nº 20, janvier 2018.

Cyrano de Bergerac Hercule Savinien de, *Œuvres Complètes. Théâtre*, André Blanc (éd.), Paris, Honoré Champion, coll. « Sources Classiques », 2001, vol. III.

Cyrano de Bergerac Hercule Savinien de, « Le Pédant Joué », 1878.

Dacos Marin, « Manifeste Des Digital Humanities », THAT Camp Paris, 2010.

Darwin Charles, *On the Origin of Species*, London, John Murray, 1859.

Dort Bernard,« «Le Texte et La Scène : Pour Une Nouvelle Alliance » », *Encyclopædia Universalis*, 1984.

Ganascia Jean-Gabriel et Jean-Louis Lebrave, « Trente Ans de Traitements Informatiques Des Manuscrits de Genèse », dans *Critique Génétique. Concepts, Méthodes, Outils*, Paris, IMEC éditeur, coll. « Inventaires », 2009.

Genette Gérard, *Palimpsestes*, Seuil, Paris, coll. « Poétique », 2014.

Grésillon Almuth, Marie-Madeleine Mervant-Roux et Dominique Budor, *Genèses théâtrales*, Paris, France, CNRS éditions, 2010.

Group The Interedition Development, « CollateX », 2021.

Hunt J. W. et M. D. Mcllroy, « An Algorithm for Differential File Comparison », *Computing Science Technical Report*, n° 41, 1975.

Kifer Michael, « Ediff User's Manual », 1998.

Le Deuff Olivier et Frédéric Clavert, « Petite histoire des humanités digitales », dans *Le temps des humanités digitales. La mutation des sciences humaines et sociales*, Limoges, Fyp éditions, 2014, p. 15-31.

Le Roux Ronan, *Une histoire de la cybernétique en France (1948-1975)*, Paris, Classiques Garnier, coll. « Histoire des techniques », 2018.

Levenshtein Vladimir Iosifovich, « Binary Codes Capable of Correcting Deletions, Insertions and Reversals. », *Soviet Physics Doklady*, vol. 10, n° 8, 1966, p. 707-710.

MacKenzie David, Paul Eggert et Richard Stallman, « Comparing and Merging Files », GNU Operating System, 2018.

Mahrer Rudolf et Joël Zufferey, « Variance », sans date.

Maloney Andy, « Comparing Nearly Identical Images Using "Beyond Compare" », *Journal of Forensic Identification*, vol. 63, n° 2, 2013, p. 153-164.

Needleman Saul B. et Christian D. Wunsch, « A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins », *Journal of Molecular Biology*, vol. 48, n° 3, 1970, p. 443-453.

Picard Jean-François, *La République Des Savants. La Recherche Française et Le C.N.R.S.*, Flammarion, Paris, 1990.

PrestoSoft, « ExamDiff Pro », 2021.

Reboul Marianne, « Odysseus », 2021.

Saignol Côme, « AutoMedite », 2021.

Saignol Côme, « FindAlign », 2021.

Salvador Xavier-Laurent, « La France Va-t-Elle En Finir Avec Les Humanités Classiques ? », *Le Figaro*, août 2019.

Schreibman Susan, « The Versioning Machine », 2021.

Schreibman Susan, Amit Kumar et Jarom McDonald, « The Versioning Machine », *Literary and Linguistic Computing*, vol. 18, n° 1, 2003, p. 101-107.

University of Tübingen, « Tuebingen System of Text Processing Tools (TUSTEP) », 2021.

Vidal Frédérique, « 2e Plan National Pour La Science Ouverte », https://www.enseignementsup-recherche.gouv.fr/cid159134/la-ministre-de-l-enseignement-superieur-de-la-recherche-et-de-l-innovation-frederique-vidal-annonce-

un-2e-plan-national-pour-la-science-ouverte.html,
2021.

WHEELES Dana et Kristin JENSEN, « Juxta », 2021.

WHEELES Dana et Kristin JENSEN, « Juxta Commons », dans *Proceedings of the Digital Humanities*, sans lieu, Alliance of Digital Humanities Organizations (ADHO), 2013.