

TAL et Littérature comparée. Détection automatique des correspondances textuelles entre les réécritures d'un mythe

Karolina Suchecka, Nathalie Gasiglia

▶ **To cite this version:**

Karolina Suchecka, Nathalie Gasiglia. TAL et Littérature comparée. Détection automatique des correspondances textuelles entre les réécritures d'un mythe. Traitement Automatique des Langues Naturelles, 2022, Avignon, France. pp.88-98. hal-03701475

HAL Id: hal-03701475
https://hal.science/hal-03701475

Submitted on 24 Jun 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

TAL et Littérature comparée. Détection automatique des correspondances textuelles entre les réécritures d'un mythe

Karolina Suchecka[1] Nathalie Gasiglia[2]
(1) ALITHILA (ULR 1061), Université de Lille, France
(2) STL (UMR 8163), Université de Lille, France
karolina.suchecka@univ-lille.fr, nathalie.gasiglia@univ-lille.fr

RÉSUMÉ

L'idée de pouvoir détecter automatiquement des relations intertextuelles est stimulante, pour la recherche littéraire et linguistique, et pour l'édition numérique. Cependant, si les logiciels employés pour notre projet, TextPAIR et Tracer, sont très performants pour les correspondances proches, grâce à des techniques de l'intelligence artificielle, ils ne détectent pas (bien) des réutilisations et évocations plus complexes. Nous proposons d'améliorer les résultats en faisant coopérer l'herméneutique spécifique des études littéraires avec des méthodes talistes, linguistiques et informatiques. Nous rencontrons toutefois quelques difficultés en traitant notre corpus avec des outils du TAL.

ABSTRACT

NLP and Comparative Literature. Automatic detection of textual similarity between the rewritings of a myth.
Being able to automatically detect intertextual relations is stimulating, both for literary and linguistic research and digital publishing. However, while the software used, TextPAIR and Tracer, are very efficient for close matches, thanks to IA techniques, they do not detect (well) more complex textual similarities and allusions. We propose to improve the initial results by making the specific hermeneutics of literary studies cooperate with linguistic and computational methods. However, we encounter some difficulties in interrogating our corpus with the tools proposed by the NLP.

MOTS-CLÉS : TextPAIR, Tracer, XML-TEI, série traductive, intertextualité quantitative, visualisation de données, graphes de connaissance.

KEYWORDS: TextPAIR, Tracer, XML-TEI, chain of translation, quantitative intertextuality, data visualization, knowledge graphs.

Introduction

La détection automatique de reformulations dans des textes est utile pour l'identification du plagiat (Ferrero & Simac-Lejeune, 2015), l'attribution de l'auctorialité d'œuvres anonymes (Rybicki, 2016) ou la génération de résumés (Barzilay & Elhadad, 1997). Des projets littéraires appuyés sur des outils informatiques mettent en relation des textes anciens, en grec (Büchler *et al.*, 2014) et en latin (Coffee *et al.*, 2012a), alignent différentes traductions d'une œuvre (Reboul, 2017), ou encore recherchent des

sources (Allen & Cooney, 2010), voire des inspirations d'un auteur (Ganascia *et al.*, 2014; Ganascia, 2020). Nous, nous souhaitons éditer 94 traductions et réécritures du mythe d'Orphée et Eurydice en français au sein d'une interface enrichie d'annotations et de visualisations des relations intertextuelles (§ 1). Il s'agit de détecter des extraits textuels contenant des réemplois ou reformulations d'extraits d'un ou plusieurs mots à l'aide de logiciels TextPAIR (Horton *et al.*, 2010) et Tracer (Büchler, 2013), puis d'améliorer ces détections et de les rendre mieux exploitables pour les chercheurs littéraires, notamment au moyen de visualisations adaptées (§ 2). Cela implique des défis particuliers tant éditoriaux (non abordés ici) que linguistiques et informatiques (§ 3).

1 Présentation du projet

De nombreux chercheurs littéraires se sont penchés sur les relations intertextuelles. Leurs analyses, souvent divergentes, offrent une terminologie riche (Kristeva, 1969; Barthes, 1973; Laurent, 1976; Genette, 1982), mais dont certains termes désignent des concepts mal articulables (Limat-Letellier, 1998; Gignoux, 2006). En Littérature comparée, l'analyse des réécritures de mythes produit des cadres d'inspirations diverses (Greimas, 1966; Durand, 1979; Brunel, 1992; Heidmann, 2003). La complexité de l'analyse des relations intertextuelles s'illustre exemplairement dans notre corpus.

1.1 Présentation du corpus

Le récit de l'amour du poète Orphée et de la nymphe Eurydice, dont les principales sources antiques sont le livre IV des *Géorgiques* de Virgile (37-30 av. J.-C.) et les livres X-XI des *Métamorphoses* d'Ovide (I^er^ s.), a inspiré nombre d'auteurs depuis le Moyen-Âge. En poésie ou en prose, à l'opéra ou au théâtre, ce mythe est adapté, parodié, modernisé ou réinterprété. Notre premier corpus réunit 53 réécritures en français du mythe, publiées entre le XV^e^ et le XXI^e^ s. Leur proximité par rapport au mythe est variable, et souvent difficile à évaluer. Une étape importante est de revenir vers ses sources antiques afin d'analyser les points communs et les différences de réinterprétation du récit.

Dans sa mythocritique, Brunel (1992) parle de la flexibilité des mythes qui dépend des expériences de lecture et souvenirs des auteurs. Certains de ceux du corpus citent une traduction française, d'autres composent leur récit sans s'appuyer sur un texte source précis, mais en se fiant à leur mémoire, parfois défaillante. Tous peuvent aussi mêler différentes sources antiques, éventuellement en modifiant leurs épisodes, ou ces sources et d'autres réécritures, parfois mieux connues.

Afin d'analyser les réinterprétations du mythe et d'améliorer la prise en compte des relations allusives entre réécritures, nous étudions spécifiquement les séries traductives (ST) des sources antiques en traitant un second corpus de 41 textes (distinct de celui des 53 réécritures) : 17 traductions de Virgile (ST-V) et 24 d'Ovide (ST-O) publiées depuis le XV^e^ s. (donc avec une part d'ancien français), en vers ou en prose, et, selon les publics visés, fidèles, ou adaptées et annotées.

1.2 Détection des correspondances avec les logiciels TextPAIR et Tracer

Détecter des relations intertextuelles au moyen d'outils informatiques est stimulant, mais leurs performances varient (Manjavacas *et al.*, 2019). TextPAIR et Tracer exploitent peu de connaissances linguistiques, mais des techniques comme le plongement lexical (Mikolov *et al.*, 2013) ou l'alignement séquentiel (Wise, 1993; Bergroth *et al.*, 2000) les rendent performants pour des relations proches (reformulation simple, paraphrase lexicale). TextPAIR, par exemple, génère pour chaque texte des *n-grammes* (Jurafsky & Martin, 2009; Forstall *et al.*, 2015) de mots pleins, de stemmes (cf. § 3) ou de lemmes (à lui fournir) qu'il indexe et confronte à ceux d'autres textes.

L'observation de 1 000 couples d'extraits aléatoirement choisis (Suchecka & Gasiglia, 2022) parmi ceux mis en relation par TextPAIR (500/8 851) et Tracer (500/10 414) dans les corpus des 53 réécritures et des 41 traductions montre que presque la moitié lient des mots vides ou des éléments formels des œuvres (didascalies, etc.), des expressions très fréquentes et souvent figées, ou des noms propres considérés isolément. Ces relations sont peu pertinentes. Par ailleurs, pour environ 15 % des couples d'extraits, le contexte fourni par les 2 logiciels ne permet pas d'apprécier leur pertinence. Au final, seulement 36 % des couples de l'échantillon présentent des relations effectives, explicites ou pas, et elles lient majoritairement des traductions et très rarement des réécritures.

Malgré des résultats probants (Suchecka & Gasiglia, 2021), le taux d'erreurs des mises en relations d'extraits textuels est trop élevé pour pouvoir se fier à elles et le nombre d'extraits impliqués est trop important pour examiner chaque couple. Donc, pour dégager les relations les plus pertinentes, nous exploitons ce qui les motive : identité de mots, de racines ou de lemmes, ou synonymie.

2 Amélioration des résultats initiaux

Pour le corpus des ST, qui réunit 41 traductions (donc *a priori* le plus riche en relations intertextuelles), la précision est élevée : parmi 860 couples détectés par Tracer, seulement 12 sont invalides. Les 848 pertinents sont xmlisés et enrichis (FIG. 1) avec, pour chaque mot plein (<w>), le lemme (@lemma) et la catégorie grammaticale (@pos) – produits par TreeTagger (Schmidt, 1994) puis corrigés manuellement –, et (cf. § 3) jusqu'à 3 synonymes (@sameAs) en partie issus du *Dictionnaire électronique des synonymes* (*DES*) (Chardon & François, 2020). Les syntagmes sont également balisés (<phr>) avec mention de leur référent (@select). Les mots pleins de chaque couple d'extraits correspondants sont comparés automatiquement afin d'enregistrer le détail des proximités lexicales (<xr> avec @corresp indiquant l'@xml:id du <w> lié, le @type de lien et une évaluation de son degré de certitude, @cert). Ces annotations permettent, par l'examen du nombre et la nature des relations lexicales[1] au sein de chaque couple, d'écarter ceux liés uniquement par des mots ou expressions vides et de visualiser les couples non écartés afin de faciliter l'évaluation de l'effectivité de leur pertinence.

[1] Notre méthode de comparaison et d'évaluation automatique des proximités entre les couples est exposée dans nos publications mentionnées au § 1.2.

```
<phr type="GN" select="mari">
        <w xml:id="2400028_14" n="14" lemma="son" pos="DET:POS">son</w>
        <w xml:id="2400028_15" n="15" lemma="mari" pos="NOM" sameAs="époux conjoint homme">mari
                <xr corresp="6900027_17" type="forme" cert="5"/>
                <xr corresp="7400048_48" type="lemme" cert="5"/>
                <xr corresp="1300018_19" type="lemme_synonyme" cert="2"/>
                <xr corresp="1600020_13" type="synonyme_synonyme" cert="1"/>
        </w>
</phr>
```

FIGURE 1 : Exemple de balisage XML d'un groupe nominal dont le <w> *mari* est lié à 4 items

2.1 Graphes pour délimiter des segments du mythe

Dans nos visualisations sous forme de graphes, chaque nœud correspond à une phrase ou plusieurs présentant des réemplois. Quand plusieurs phrases successives sont en relation avec d'autres, elles sont regroupées et représentées par un nœud de taille proportionnelle à la longueur de l'extrait. La totalité des relations dans la ST-V détectées par Tracer constitue un seul graphe complexe, au sein duquel certains textes sont regroupés en leur quasi-totalité dans un même nœud, ce qui prouve une forte proximité entre eux et une précision satisfaisante des résultats obtenus pour la ST-V.

Les relations identifiées (par Tracer toujours) dans la ST-O, elles, se répartissent en 7 grands graphes de plus de 10 nœuds chacun et en une grande quantité de petits graphes de 2 à 4 nœuds. Cette répartition montre qu'il faut améliorer la détection y compris pour les traductions seules. Mais, elle nous permet déjà de dégager les proximités thématiques des nœuds de chaque graphe. Celui en FIG. 2 renvoie dans sa totalité à l'épisode de la seconde mort d'Eurydice. Des relations liant 1 à 3 phrases y sont établies parmi 14 traductions publiées à partir du XVI^{e} s.

Comme 6 traductions, dont les 4 en ancien français, n'apparaissent pas dans ce graphe, afin de voir s'il y a lieu de prendre en compte d'autres extraits relatifs à cet épisode, nous couplons cette visualisation cumulative à une autre, qui limite la taille de chaque nœud à une phrase et qui liste au-dessus du nœud les lemmes pour lesquels de relations lexicales ont été détectées (FIG. 3).

Une partie des phrases du graphe cumulatif de 18 nœuds, représentant des extraits, se retrouve dans un réseau de 11 nœuds, dont les mots-clés, c'est-à-dire les lemmes les plus fréquemment mis en relation au sein des extraits, sont, notamment, *aimer, Eurydice, mourir* et *plaindre* (en haut de la FIG. 3). Nous les retrouvons aussi dans un graphe de 9 nœuds (en dessous). Celui-ci regroupe les 4 manuscrits anciens, qui n'ont pas de relations directes avec les traductions modernes. Nous constatons donc que certaines phrases présentent des relations lexicales non détectées par Tracer, et que les nœuds qui les représentent ne sont par conséquent pas liés. La présence de mots-clés identiques, ou proches, nous permet bien, en revanche, de regrouper les 2 ensembles de textes et de considérer qu'ils évoquent tous la seconde mort d'Eurydice.

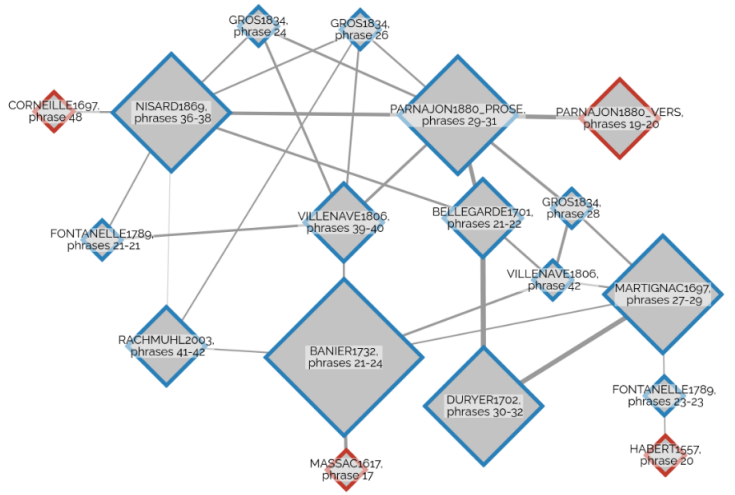


FIGURE 2 : Graphe de 18 nœuds évoquant la seconde mort d'Eurydice chez Ovide. Nœud : extrait textuel (phrase(s)) ; bordures rouges : textes en vers ; bordures blues : textes en prose ; diamants : traductions d'Ovide ; et degré de similarité de 2 extraits liés représenté par l'épaisseur des liens.

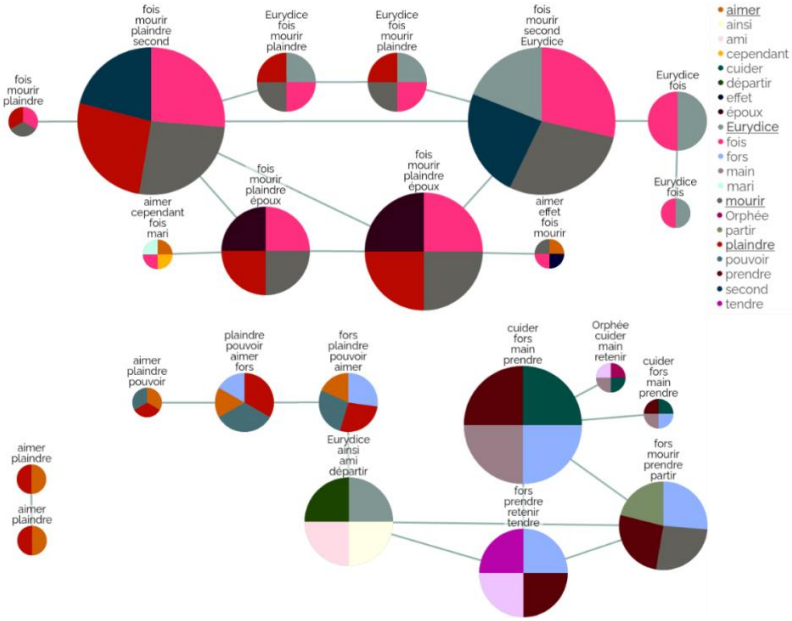


FIGURE 3 : Rapprochement de 3 graphes. Nœud : phrase ; nombre de lemmes communs représenté par l'épaisseur des liens ; codes-couleur pour les 21 lemmes communs les plus fréquents.

2.2 Exploitation des mots-clés extraits des relations entre traductions

Un tel regroupement permet une analyse fine des 445 relations lexicales trouvées pour 57 phrases de 18 textes, correspondantes (au sein des ST) à la seconde mort d'Eurydice. Parmi elles, 130 lemmes sont recensés. 30 d'entre eux ont au moins 10 occurrences. *Second*, *mourir* et *plaindre* sont les plus fréquents (32 à 34 occ.), suivis des noms et des périphrases des deux amants (*Orphée*, *Eurydice*, *ami* et *époux*, 13 à 14 occ.) et de beaucoup de verbes (*tendre*, *entendre*, *retourner*, *aimer*, *embrasser* et *regarder*, entre 12 à 21 occ.). Ces lemmes seront exploités lors de l'analyse des réécritures, surtout en cas de différences de tailles de récit de cet épisode. C'est le cas par exemple des opéras qui se focalisent sur l'amour d'Orphée et Eurydice. Chez Gluck (1774), la sortie des Enfers et la seconde mort d'Eurydice sont narrées par 38 répliques découpées en 110 phrases alors que dans les ST-V et ST-O, le traitement de l'épisode ne dépasse pas respectivement 15 et 8 phrases. Cette différence d'échelle handicape TextPAIR et Tracer. En revanche, nous constatons que sur 358 mots pleins présents dans les 110 phrases de Gluck, 175 se trouvent parmi les relations lexicales détectées dans les traductions pour cet épisode. L'analyse de leur répartition permet de distinguer les extraits de Gluck qui se rapprochent le plus des traductions, mais il est aussi intéressant d'observer les lemmes qui ne font pas partie des mots-clés, car ils révèlent des modifications introduites dans la réécriture. Chez Gluck, Eurydice ne comprend pas pourquoi Orphée refuse de la regarder. Elle interprète cela comme un manque d'amour, et refuse de le suivre sans obtenir d'explication. En écho à cette mécompréhension, nous recensons, d'un côté, une quantité importante de lemmes à consonance négative, exprimant l'incompréhension et le refus d'Eurydice (*indifférence*, *barbare*, *ingrat*) et, de l'autre, beaucoup de lemmes illustrant le désarroi d'Orphée (*contrainte*, *effroi*, *implorer*).

3 Corpus littéraire et outils du TAL. Quelques difficultés

Des traitements linguistiques et talistes, destinés à suppléer aux limites des logiciels TextPAIR et Tracer ou réalisés pour eux-mêmes, améliorent les résultats. Par exemple, restreindre la recherche des synonymes à ceux déjà attestés et mis en relation dans les ST-V et ST-O, améliore la pertinence de la détection des relations synonymiques entre réécritures. Les synonymes, issus d'un dictionnaire de synonymie cumulative comme le *DES* pour les mots pleins de tout le corpus ou d'une ontologie comme WOLF (Sagot & Fišer, 2012) localement, notamment pour mettre en relations des extraits de textes poétisés ou très métaphoriques, comme (L'Hermite, 1662), sont progressivement mieux sélectionnés et leur liste est complétée au fil des analyses. Par exemple, pour l'expression « sur son visage il détournât ses yeux », qui implique qu'Orphée se retourne pour voir Eurydice, nous avons détecté d'abord le mot *œil* parmi les mots-clés de l'épisode, puis la synonymie entre *détourner* et *se retourner*. Mais, dans ce contexte, il est aussi pertinent d'aligner l'expression « détourner les yeux » avec « regarder en arrière », « se retourner [et] poser les yeux » et « tourner le regard vers ». Or, pour ce faire, au moins pour des verbes très polysémiques, comme *(se) retourner*, il faut prendre en compte les cooccurrents et les constructions syntaxiques qui contribuent à l'interprétation en contexte de l'occurrence verbale. Si nous apprécions la richesse et la facilité d'exploitation des ressources disponibles, leur caractère cumulatif (une liste de synonymes, pas une par sens) pose problème. Par

ailleurs, le recours à des représentations distributionnelles, comme des plongements lexicaux, qui est probant dans certains projets de détection automatique des relations intertextuelles (Burns *et al.*, 2021), ne l'est pas pour notre corpus. Les résultats d'une expérimentation avec *word2vec* sont décevants : les mots proches proposés pour *Orphée* ou *Eurydice* sont des noms d'autres entités (*Aristée* est en première position pour les deux), le verbe le plus proche de *mourir* est *consoler* et celui d'*aimer*, *obéir.* Même si la pertinence de cette méthode pour des techniques d'extraction d'information est avérée, nous cherchons donc plutôt à exploiter une ressource qui propose des sous-listes des synonymes, accompagnées de sélecteurs sémantiques, cooccurrentiels ou syntaxiques susceptibles de faciliter l'identification (semi-)automatique des sens des mots-occurrences, comme des dictionnaires de synonymie distinctive.

Par ailleurs, pour enrichir la pré-annotation du corpus, la prise en compte des relations morphologiques (flexionnelles et dérivationnelles) au sein de notre corpus s'avère être un véritable défi. La racinisation proposée par Snowball n'est pas faite selon des règles morphologiques, mais par coupe de fins des mots. Les "pseudo-tiges" proposées sont rarement les radicaux des analyses morphologiques, ce qui nous prive de détections pertinentes – les stemmes de 3 formes plurielles du verbe *savoir* (*savons*, *savez* et *savent*) sont différents (*savon / sav / savent)* –, et augmente même le nombre de fausses détections (par exemple en liant *fer / fermer*). Mais, en cherchant une alternative, nous constatons que l'exploitation de 2 outils conçus pour le français, DériF (Namer, 2009) et Morphonette (Hathout, 2010), est peu probante. Même en cumulant les résultats des 2 outils, peu de lemmes peuvent être liés à leur base morphologique (2 007 sur 11 653), et de nombreuses erreurs persistent. Morphonette, par exemple, indique que les couples *noceur* / *nocif* ou *voyant* / *voyage* appartiennent à la même famille morphologique et l'analyse automatique mise en œuvre par DériF l'amène à apparier *mauve* et *mauvais* ou *vin* et *divin*. Finalement, dans 13 cas seulement l'appariement de dérivés de classes grammaticales différentes serait facilité par ces outils.

Comme nous voulons montrer des visualisations fines des relations intertextuelles effectives, notre annotation doit être aussi exacte que possible. Nous privilégions donc des outils qui sont peut-être moins performants, mais dont nous pouvons comprendre clairement le traitement afin d'intervenir aisément sur leurs sorties si des corrections manuelles sont nécessaires. Cette procédure est fréquente au sein des projets HN, travaillant régulièrement avec des corpus spécifiques, comme en témoigne par exemple la mise à disposition de la plateforme de correction des annotations Pyrrha (Clérice *et al.*, 2021). Ainsi, par exemple, nous privilégions TextPAIR et Tracer, des outils désormais assez anciens, mais pour lesquels la documentation est riche et détaillée, les développeurs actifs et à l'écoute des utilisateurs, et le code source librement disponible. Par ailleurs, ils sont spécifiquement destinés aux recherches littéraires, et ont servi dans de projets très diversifiés, sur des corpus de langues, d'époques et d'auteurs différents (Franzini, 2016 ; Kokkinakis & Malm, 2016 ; Gladstone, 2018, O'Neill *et al.*, 2020). Tous ces facteurs nous amènent à privilégier ces 2 outils pour explorer notre corpus, sans que cela nous empêche d'apprécier les méthodes et techniques d'autres outils, comme Tesserae (Coffee *et al.*, 2012b), Phœbus (Boukhaled, *et al.*, 2015) ou Textreuse (Mullen, 2015).

Par ailleurs, nous aimerions disposer d'un outillage cohérent nous permettant de cumuler les enrichissements sans avoir à lisser les différences de structuration des entrées et sorties de chaque

logiciel. TextPAIR peut travailler directement à partir de fichiers XML-TEI, mais il n'exploite pas le balisage initial ni ne le restitue. Tracer, lui, demande de partitionner au préalable le corpus en segments (phrases, paragraphes, etc.) identifiés par un numéro unique et intégrés dans un tableau respectant une structuration spécifique. Après avoir récupéré les résultats des 2 logiciels, nous procédons à leur xmlisation, à leur fusion et à l'enrichissement linguistique afin d'obtenir le balisage présenté en FIG. 1.

Par ailleurs, il existe de nombreux annotateurs morphosyntaxiques qui traitent conjointement ou pas différents états de la langue, mais le choix de l'un d'eux doit anticiper la suite des manipulations pour qu'ils s'inscrivent efficacement dans la chaîne de traitement. Certains types de programmes sont en effet plus ou moins aptes à travailler ensemble (application Web ou logiciel autonome et algorithme Python), de même certains logiciels ne sont compatibles qu'avec certains systèmes d'exploitation (IramuteQ est difficile à faire fonctionner sur Linux, alors que TextPAIR n'est pas optimisé pour Windows). Au niveau du découpage des syntagmes, plusieurs parseurs syntaxiques ont été testés, mais leurs résultats étaient inégaux non seulement en fonction de l'outil, mais aussi de la complexité des phrases et du type d'œuvre dont elles étaient extraites (la syntaxe de la poésie ou du théâtre étant particulièrement problématique). C'est finalement un traitement moins sophistiqué, à base de règles et fait maison, qui est retenu. Toutefois, si nous pouvons remanier les résultats de certains outils, voire concevoir un algorithme simple pour pallier certaines difficultés, nous ne cherchons pas à faire du développement, nous tenons à travailler dans une optique littéraire, linguistique et éditoriale.

Conclusion

Si la flexibilité des mythes, un concept théorique issu de la recherche littéraire, peut être observée et analysée au sein de notre corpus à l'aide d'outils informatiques, nous avons montré l'utilité de les faire coopérer et de travailler avec eux en enrichissant leurs résultats initiaux, en affinant les détections, en annotant les unités lexicales communes et en générant des visualisations sous forme de graphes qui rendent possible l'observation et l'analyse des correspondances détectées, même quand elles sont subtiles et implicites. Les manipulations entreprises sont motivées par des objectifs littéraires, mais sollicitent des compétences linguistiques-informatiques à acquérir (ou à trouver chez des partenaires).

Les difficultés exposées montrent la complexité des questions auxquelles peuvent faire face les chercheurs des HN qui exploitent des outils informatiques. Nous sommes convaincues que, comme nous, ils ont besoin de chaînes de traitement cohérentes et dont les produits soient compréhensibles et manipulables. Ce qui est foncièrement nécessaire, c'est un dialogue interdisciplinaire, qui aurait d'ailleurs au moins deux contreparties pour les développeurs, puisque les utilisateurs (littéraires, linguistes, lexicographes, etc.) pourraient d'une part leur faire des retours sur des emplois effectifs de leurs outils et, d'autre part, partager avec eux des corpus enrichis d'annotations linguistiques et littéraires plus riches et pertinentes, éventuellement exploitables comme corpus d'entraînement.

Références

ALLEN, T. & COONEY, C. (2010). Plundering philosophers: identifying sources of the *Encyclopédie*. *Journal of the Association for History and Computing*, vol. 13, nº 1. URL : http://hdl.handle.net/2027/spo.3310410.0013.107.

BARTHES, R. (1973). Texte (Théorie du). In *Encyclopædia Universalis*, vol. 15, p. 1013-1017.

BARZILAY, R. & ELHADAD, M. (1997). Using lexical chains for text summarization. In *Intelligent scalable text summarization*. URL : https://www.aclweb.org/anthology/W97-0703.

BERGROTH, L., HAKONEN, H. & RAITA, T. (2000). A survey of longest common subsequence algorithms. In *Proceedings of the 7th International symposium on string processing information retrieval*, Corogne, p. 39‑48. DOI : 10.1109/SPIRE.2000.878178.

BOUKHALED, M.-A., SELLAMI, Z. & GANASCIA, J.-G. (2015). Phoebus : un logiciel d'extraction de réutilisations dans des textes littéraires. Communication présentée à la *22e Conférence sur le traitement automatique des langues naturelles*, Caen, 22-25 juin. HAL : hal-01198411.

BRUNEL, P. (1992). *Mythocritique : théorie et parcours*, Presses universitaires de France.

BÜCHLER, M. (2013). *Informationstechnische Aspekte des historical Text Re-use*. Thèse de doctorat, Université de Leipzig.

BÜCHLER, M., BURNS, P., MÜLLER, M., FRANZINI, E. & FRANZINI, G. (2014). Towards a historical text re-use detection. In C. BIEMANN & A. MEHLER, Éds., *Text mining. From ontology learning to automated text precession applications*, Springer International Publishing, Suisse, p. 221‑238. DOI : 10.1007/978-3-319-12655-5_11.

BURNS, P., BROFOS, J., LI, K., CHAUDHURI, P., DEXTER, J. (2021). In *Proceedings of the 2021 Conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, En ligne, Association of Computational Linguistics, p. 4900-4907. DOI : 0.18653/v1/2021.naacl-main.389.

CHARDON, L. & FRANÇOIS, J. (2020). Les vedettes du *Dictionnaire Électronique des Synonymes* et les relations d'adjacence entre leurs synonymes. In *Lexique*, n°27, p. 21-45. HAL : halshs-03192815.

CLÉRICE, T., PILLA, J., FRFERRY., CAMPS J.-B., NGAWANGTRINLEY, ARCHITEXTE, JETELY, A., PINCHE, A., S. SIDDHANT & JHRDT. (2021). *hipster-philology/pyrrha* (version 3.0.0). Zenodo. DOI : 10.5281/zenodo.5144781.

COFFEE, N., KOENIG, J.-P., POORNIMA, S., FORSTALL, C., OSSEWAARDE, R. & JACOBSON, S. (2012a). Intertextuality in the digital age. In *Transactions of the American Philological Association*, vol. 142, n°2, p. 383-422. DOI : 10.1353/apa.2012.0010.

COFFEE, N., KOENIG, J.-P., POORNIMA, S., FORSTALL, C., OSSEWAARDE, R. & JACOBSON, S. (2012b). The Tesserae Project: intertextual analysis of Latin poetry. In *Literary and Linguistic Computing*, vol. 28, n°2, p. 221‑228. DOI : 10.1093/llc/fqs033.

DURAND, G. (1979). *Figures mythiques et visages de l'œuvre : de la mythocritique à la mythanalyse*, Berg international, Paris.

FERRERO, J. & SIMAC-LEJEUNE, A. (2015). Détection automatique de reformulations – Correspondance de concepts appliquée à la détection du plagiat. In *Actes de la 15ème conférence internationale sur l'extraction et la gestion des connaissances*, Luxembourg. HAL : hal-01108061.

FORSTALL, C., COFFEE, N., BUCK, T., ROACHE, K. & JACOBSON, S. (2015). Modelling the scholars: Detecting intertextuality through enhanced word-level n-gram matching. In *Digital Scholarship in the Humanities*, vol. 30, n°4, p. 503‑515. DOI : 10.1093/llc/fqu014.
GANASCIA, J.-G., GLAUDES, P. & DEL LUNGO, A. (2014). Automatic detection of reuses and citations in literary texts. In *Digital scholarship in the Humanities*, vol. 29, n°3, p. 412‑421. HAL : hal-00977310, DOI : 10.1093/llc/fqu020.
GANASCIA, J.-G. (2020). Détection automatique de phénomènes intertextuels. *Genesis. Manuscrits*, nº 51, p. 63-77. DOI : https://doi.org/10.4000/genesis.5671.
GENETTE, G. (1982). *Palimpsestes. La littérature au second degré*, Seuil, Paris.
GIGNOUX, A.-C. (2006). De l'intertextualité à la récriture. In *Cahiers de Narratologie. Analyse et théorie narratives*, n°13. DOI : 10.4000/narratologie.329.
GLADSTONE, C. (2018). Evaluating the practices and legacy of the enlightenment on 19th-century print culture. In *ARTFL* (blog). URL : https://artfl.blogspot.com/2018/11/evaluating-practices-and-legacy-of.html.
GLUCK, C. W., CALZABIGI, R. DE & MOLINE, P.-L. (1774). *Orphée et Eurydice*, Académie Royale, Paris.
GREIMAS, A. J. (1966). Éléments pour une théorie de l'interprétation du récit mythique. In *Communications*, vol. 8, n°1, p. 28‑59. DOI : 10.3406/comm.1966.1114.
HATHOUT, N. (2010). *Morphonette: a morphological network of French*. HAL : hal-00485503.
HEIDMANN, U. (2003). (Ré)écritures anciennes et modernes des mythes : la comparaison pour méthode. L'exemple d'Orphée. In *Études de Lettres : revue de la Faculté des lettres de l'Université de Lausanne*, n°3, p. 47‑64. DOI : 10.5169/seals-870182.
HORTON, R., OLSEN, M. & ROE, G. (2010). Something borrowed: Sequence alignment and the identification of similar passages in large text collections. In *Digital Studies / Le Champ numérique*, vol. 2, n°1. DOI : 10.16995/DSCN.258.
JURAFSKY, D. & MARTIN, J. H. (2009). N-gram language models. In *Speech and language processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice-Hall, Inc., New York, p. 189‑232.
KOKKINAKIS, D. & MALM, M. (2016). Detecting reuse of biblical quotes in Swedish 19th-century fiction using sequence alignment. In F. MAMBRINI, M. PASSAROTTI & C. SPORLEDER, Éds, *Proceedings of the workshop on corpus-based research in the Humanities (CRH)*, Varsovie, Pologne : Polish Academy of Sciences, p. 79-86.
KRISTEVA, J. (1969). *Sèméiotikè. Recherches pour une sémanalyse*, Seuil, Paris.
LAURENT, J. (1976). La Stratégie de la forme. In *Poétique*, n°27, p. 257‑281.
L'HERMITE, T. (1662). La Lyre d'Orphée. In *Les Amours de feu Mr Tristan et autres pièces très-curieuses*, Chez Gabriel Quinet, Paris, p. 195‑216.
LIMAT-LETELLIER, N. (1998). Historique du concept d'intertextualité. In M. MIGUET-OLLAGNIER, Éd., *L'intertextualité*, Presses universitaires de Franche-Comté, Besançon, p. 17‑64. DOI : 10.4000/books.pufc.4507.
FRANZINI, G. (2016). English translations of *Pan Tadeusz*: a comparison with TRACER. In *eTRAP* (blog). URL : https://www.etrap.eu/english-translations-of-pan-tadeusz-a-comparison-with-tracer/.
MANJAVACAS E., LONG, B. & KESTEMONT, M. (2019). On the feasibility of automated detection of allusive text reuse. In *Proceedings of the 3rd joint SIGHUM workshop on Computational Linguistics*

for Cultural Heritage, Social Sciences, Humanities and Literature, Association for Computational Linguistics, Minneapolis, USA, p. 104-114. DOI : 10.18653/v1/W19-2514.
MIKOLOV, T., CHEN, K., CORRADO, G. & DEAN, J. (2013). Efficient estimation of word representations in vector space. In *ICLR 2015,* Vancouver, Canada. DOI : 10.48550/arXiv.1301.3781.
MULLEN, L. (2015). *textreuse: detect text reuse and document similarity* (version 0.1.4.9000). URL : https://hdl.handle.net/1920/10077.
NAMER, F. (2009). *Morphologie, Lexique et Traitement Automatique des Langues : l'analyseur DériF*, Hermès-Lavoisier, Paris. HAL : hal-00413337.
O'NEILL, H., WELSH, A., SMITH, D., ROE, G. & TERRAS, M. (2021). Text mining Mill: Computationally detecting influence in the writings of John Stuart Mill from library records. In *Digital scholarship in the Humanities*, vol. 36, n°4, p. 1013‑1029.
REBOUL, M. (2017). *Comparaison semi-automatique des traductions en langue française de l'Odyssée d'Homère (1547-1955)*. Thèse de doctorat, Université Paris IV.
RYBICKI, J. (2016). *Vive la différence* : Tracing the (authorial) gender signal by multivariate analysis of word frequencies. In *Digital scholarship in the Humanities*, vol. 31, n°4, p. 746‑761.
SAGOT, B. & FIŠER, D. (2012). *WordNet Libre du Français (WOLF)* (version 1.0b4), Institut national de recherche en sciences et technologies du numérique (Inria). URL : http://pauillac.inria.fr/~sagot/index.html#wolf.
SCHMIDT, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International conference on new methods in language processing*, Manchester. URL : https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf.
SUCHECKA, K. & GASIGLIA, N. (2021). Réécritures d'un mythe et outils de détection des réutilisations. De l'Orphée de Virgile à celui de Ballanche. In *Humanités numériques*, n°4. DOI : 10.4000/revuehn.2467.
SUCHECKA, K. & GASIGLIA, N. (2022). On digital comparative editions and textual similarity detection tools: towards a hypertextual cartography of a rewritten myth. In P. PLECHÁČ, R. KOLÁR, A.-S. BORIES & J. ŘÍHA, Éds., *Tackling the toolkit: Plotting poetry through Computational Literary Studies*, Institute of Czech Literature of the Czech Academy of Sciences, p. 167‑182. DOI : 10.51305/ICL.CZ.9788076580336.11.
WISE, M. J. (1993). Neweyes: a system for comparing biological sequences using the running Karp-Rabin greedy string-tiling algorithm. *ISMB-95 Proceedings*, AAAI, p. 393-401. URL : https://www.aaai.org/Papers/ISMB/1995/ISMB95-047.pdf.