

École doctorale Sciences de l'Homme et de la Société (ED 473),
Université de Lille, laboratoire ULR 1061 ALITHILA
École doctorale Littératures françaises et comparée (ED 019),
Sorbonne Université, laboratoire UMR 8599 CELLF

Thèse de doctorat

DISCIPLINE : Littérature comparée

Analyse et édition d'un corpus littéraire dans l'esprit des humanités numériques.

**Vers une édition comparative innovante accompagnée d'une res-
source pour le repérage intertextuel**

Thèse présentée et soutenue publiquement le 27 novembre 2025 par :

Karolina SUCHECKA,

sous la direction de :

Karl ZIEGER et Andrea DEL LUNGO

et devant le jury composé de :

Andrea DEL LUNGO, professeur à Sorbonne Université (codirecteur de thèse)

Ioana GALLERON, professeure à l'Université Sorbonne Nouvelle (rapporteuse)

Françoise LAVOCAT, professeure à l'Université Sorbonne Nouvelle (présidente du jury)

Matthieu MARCHAL, maître des conférences HDR à l'Université de Lille (examinateur)

Glenn ROE, professeur à l'Université d'Oxford (rapporteur)

Karl ZIEGER, professeur émérite à l'Université de Lille (codirecteur de thèse)

Résumé

Le projet propose une approche innovante en recherche littéraire en privilégiant des outils numériques d'analyse dans un contexte plutôt classique (les réécritures du mythe d'Orphée et d'Eurydice). Il s'agit de combiner un héritage culturel spécifique avec les technologies les plus récentes dans le but de permettre une présentation novatrice des œuvres et de leur structure.

Trois axes seront principaux dans ce projet : (i) la recherche et la conception d'outils et de modalités permettant d'améliorer la numérisation d'un corpus intermédial, sa structuration et sa présentation dans le cadre d'une édition comparative, (ii) les possibilités de recherche (semi-)automatique de correspondances entre les différentes œuvres qui réécrivent ou transposent un *topos* commun et la réflexion sur la mise en forme numérique des résultats, (iii) la conception d'un outil qui s'appuiera sur une base de données résultant des études précédentes pour rechercher et typer automatiquement les allusions au mythe présentes dans tout autre corpus textuel.

Dans un premier temps, il s'agit d'approcher le corpus sans a priori théoriques, génétiques ou historiques, afin d'observer les apports de cette nouvelle méthodologie de recherche et de confronter les résultats obtenus à ce qui a déjà pu être observé par des analyses littéraires traditionnelles. La conception d'outils d'affichage et d'analyse appuyée sur un corpus d'auteurs reconnus doit également déboucher sur la création d'un nouveau support bénéficiant à l'enseignement secondaire, à la préparation des concours comme le CAPES ou l'Agrégation et à l'enseignement du FLE.

Mots-clés : édition numérique savante, intermédialité, humanités numériques, intertextualité, réécriture, adaptation

Abstract

Analysis and editing of a literary corpus in the spirit of the Digital Humanities: Towards an Innovative Comparative Edition Accompanied by a Resource for Intertextual Mapping

The project proposes an innovative approach in literary research by privileging digital tools of analysis in a rather classical context (the rewritings of the myth of Orpheus and Eurydice). Its aim is to combine a specific cultural heritage with the latest technologies in order to allow an innovative presentation of the works and their structure.

Three central axes will be considered: (i) research and design of tools and modalities in order to improve the digitization of an intermedial corpora, its structuring and its presentation as part of a comparative edition, (ii) possibilities of (semi-)automatic search for correspondences between the different works that rewrite or transpose a common topos and reflection on the numerical formatting of the results, (iii) conception of a tool based on a database resulting from previous studies that permits to automatically find and categorize the allusions to myth present in any other textual corpora.

First of all, the corpora will be approached without theoretical, genetic or historical presumptions, in order to observe the contributions of this new research methodology and to compare the results obtained with what has already been observed by traditional literary analyzes. The design of editorial and analysis tools based on literary corpora of recognized authors must also lead to the creation of a new support for secondary education, the preparation of competitions such as CAPES or Aggregation and teaching French as a Foreign Language (FLE).

Keywords: scholarly digital edition, intermediality, digital humanities, intertextuality, rewriting, adaptation

Remerciements

Retracer le fil des rencontres et des échanges qui m'ont menée vers ce projet de thèse et m'ont aidée à l'achever est une entreprise vertigineuse. Tant de personnes sont à remercier, pour tant d'attentions et de partages.

Je remercie d'abord mes directeurs de thèse, qui ont été des véritables guides pendant toutes mes études françaises.

Merci, Karl Zieger, de m'avoir fait confiance pour suivre cette idée quelque peu expérimentale, de votre bienveillance et ouverture d'esprit très comparatiste, même face aux chapitres les plus techniques.

Merci, Andrea Del Lungo, de m'avoir encouragé à revenir après mon échange Erasmus et de m'avoir embarqué dans la belle aventure balzacienne, dont l'expérience a façonné plusieurs aspects de cette thèse. Un immense merci également d'avoir accepté de prendre le train en marche. Sans votre aide, l'aboutissement de ce projet aurait été bien plus difficile.

Merci, enfin, Nathalie Gasiglia, pour tout cet accompagnement autant exigeant et formateur que chaleureux et bienveillant. Le plaisir de travailler avec vous, votre esprit et votre intelligence m'ont donné envie de faire cette thèse. Les compétences inestimables que vous m'avez transmises m'ont permis de m'épanouir tant professionnellement que personnellement. Merci pour votre disponibilité sans faille, même lors d'un réveillon de Noël, pour relire, retravailler et perfectionner mes travaux. Merci pour les explications patientes des règles d'usage des articles, alors que je faisais encore et encore les mêmes erreurs. Merci enfin d'être restée présente, autant que possible, ces deux dernières années.

Je remercie ensuite mes collègues et enseignants qui m'ont apporté une aide inestimable dans ce parcours de thèse.

Merci à Ioana Galleron, Françoise Lavocat, Matthieu Marchal et Glenn Roe d'avoir accepté de siéger dans le jury de ma soutenance. Je suis honorée de présenter ma thèse devant celles et ceux dont les travaux ont été essentiels pour ma recherche.

Merci à Fiona McIntosh et Antonio Balvet d'avoir accepté de faire partie de mon comité de suivi individuel. Nos rencontres annuelles étaient une véritable source de motivation, me permettant de relancer la machine avec un nouvel élan d'enthousiasme. Merci pour vos conseils et retours très précieux.

Merci à Valentin De Craene, Victoria Le Fournier et Mónica Michel Rodriguez pour l'accompagnement dans la gestion de mes données de thèse. C'est grâce à vous que j'atteins avec fierté cet objectif (si nécessaire !) de la science ouverte.

Merci à Fabien Torre et Matthieu Marchal pour la transmission de leurs compétences en informatique éditoriale et la transcription de textes médiévaux, et pour leur présence bienveillante tout au long de mon parcours d'étudiante et d'enseignante au sein du Master Éditions numériques et imprimées de textes littéraires.

Merci à Pierre Corbin pour son immense expertise, bien au-delà du domaine lexicographique.

Merci aux comparatistes lilloises, Aude Ameille, Alison Boulanger, Fiona McIntosh, Émilie Picherot et Jessica Wilker, pour leur accompagnement pendant mes années ATER et les bouts de chocolat qui remontaient bien le moral après un cours difficile. Un merci tout particulier à Jessica qui m'a fait aimer Orphée à travers Rilke et Mallarmé et qui a toujours été d'un soutien sans faille.

Merci aux collègues et ami(e)s doctorant(e)s pour le privilège de partager les peines et les joies de la thèse : Anis Baghi, Alice Dancerelle-Bourlon, April Dupont, Philippe-Alexandre Gonçalves, Amina Houara, Rebecca Legrand et Fanny Van Exaerde. Ça y est, c'est enfin mon tour !

Merci enfin aux chercheurs et chercheuses rencontrés ces dernières années, qui m'ont aidé à mûrir ma réflexion : Motasem Alrahabi, Francesa Frontini, Jean-Gabriel Ganascia, Clovis Gladstone, Martin Grandjean, Fatiha Idmhand, Giovanni Pietro Vitali, Elena Pierazzo, Marianne Reboul, Simone Rebori, Michael Sinatra... et bien d'autres. La découverte de vos travaux m'a souvent permis de replonger dans les miens avec une nouvelle énergie.

Et je remercie enfin ma famille et mes amis qui m'ont aidé à garder le cap et à ne pas épuiser toutes mes ressources. Je n'ai pas à les nommer, je n'ai qu'à regarder autour de moi.

Merci à mes familles françaises et polonaises d'avoir toujours été là, merci pour les joyeuses pauses et les questions intéressées.

Et merci surtout à toi, Sam, pour tous les monologues que tu m'as laissé tenir, pour toutes les pauses que tu m'as forcée à prendre. Merci pour ton immense patience et ton soutien qui m'ont permis de continuer sereinement ce travail au fil des ans. Je n'aurais réussi ni à commencer ni à finir cette thèse sans toi.

Arthur, mon petit perturbateur arrivé sur la fin de ce parcours : je ne regrette aucun moment passé avec toi au lieu de rédiger et je te dédie solennellement cette thèse.

Table des matières

Introduction	10
1. Mythe d'Orphée et Eurydice.....	12
2. Corpus littéraires et humanités numériques : mariage de convenance ?.....	13
a) <i>Collecte du corpus : enjeux interdisciplinaires</i>	13
b) <i>Description du corpus : entre principes littéraires et contraintes techniques</i>	14
c) <i>Numérisation et partage du corpus : défis de la science ouverte</i>	16
3. Détection automatique de l'intertextualité : une perspective ambivalente	17
a) <i>Réécriture, reformulation, allusion : préoccupations terminologiques</i>	17
b) <i>Intertextualité quantitative : jeu d'équilibriste entre close et distant reading</i>	19
c) <i>Tracer et TextPAIR : outils perfectibles</i>	21
4. Édition numérique comparative : un nouvel outil d'analyse	22
a) <i>Faire parler les données : défis de la modélisation</i>	23
b) <i>Notes intertextuelles et graphes des relations : outils éditoriaux innovants</i>	24
c) <i>Édition numérique : perspectives illimitées, pérennité incertaine</i>	27
5. Plan de la rédaction	28
I. Présentation du corpus des réécritures du mythe d'Orphée et Eurydice	32
1. Lecture distante et rêve de littérature mondiale	36
a) <i>Google Books et Google Ngram Viewer : hégémonie anglosaxonne</i>	38
b) <i>Frantext : base de données textuelles françaises</i>	43
c) <i>Gallica et Gallicagram : remède à tous les maux ?</i>	48
<i>Conclusion du chapitre</i>	51
2. Établissement du corpus : entre l'exhaustivité quantitative et la représentativité herméneutique	53
a) <i>Heidmann, Brunel et Genette : définition de la notion de réécriture</i>	55
b) <i>Point de départ : exploration de l'archive numérique Gallica</i>	59
c) <i>Point d'arrivée : corpus de 53 réécritures</i>	64
<i>Conclusion du chapitre</i>	69
3. Représentation de la variété des réécritures du mythe	71
a) <i>Répartition chronologique</i>	74
b) <i>Répartition générique</i>	77
c) <i>Répartition fonctionnelle</i>	82
<i>Conclusion du chapitre</i>	90

4.	Numériser un corpus dans l'esprit des humanités numériques.....	92
a)	<i>Éditions critiques numériques</i>	93
b)	<i>Protocole de production des données</i>	104
c)	<i>Description du jeu de données déposé dans Nakala</i>	117
	<i>Conclusion du chapitre</i>	126
	Conclusion de la partie	127
II.	Intertextualité : de la théorie littéraire aux méthodes computationnelles.....	129
1.	Intertextualité dans la théorie littéraire	133
a)	<i>Kristeva et Barthes : approches philosophiques aux sources linguistiques</i>	135
b)	<i>Jenny et Riffaterre : d'une pratique textuelle à un phénomène de réception</i>	139
c)	<i>Compagnon et Genette : concepts "périphériques"</i>	142
d)	<i>Les théories intertextuelles et l'observation du corpus</i>	146
	<i>Conclusion du chapitre</i>	155
2.	Procédés linguistiques propres des relations intertextuelles.....	157
a)	<i>Investigation à partir de 21 traductions d'Ovide</i>	161
b)	<i>Classement des paraphrases : phénomènes linguistiques impliqués</i>	174
c)	<i>Classement des paraphrases : proposition d'une méthode</i>	182
	<i>Conclusion du chapitre</i>	198
3.	Prologomènes des approches computationnelles de l'intertextualité.....	200
a)	<i>Stylométrie</i>	202
b)	<i>Détection et modélisation des topiques (topic modeling)</i>	208
c)	<i>Alignement automatique des corpus parallèles et la comparaison des versions</i>	214
	<i>Conclusion du chapitre</i>	231
	Conclusion de la partie	232
III.	Logiciels dédiés à la détection automatique des réemplois textuels.....	234
1.	Établissement des critères de l'évaluation avec le logiciel Phœbus	238
a)	<i>Fonctionnement du logiciel</i>	238
b)	<i>Analyse des correspondances détectées</i>	240
c)	<i>Établissement des graphes de relations</i>	244
	<i>Conclusion du chapitre</i>	246
2.	Expérimentation avec BLAST, Tesseract, Textrease et Passim	248
a)	<i>BLAST (Basic Local Alignment Search Tool)</i>	248
b)	<i>Textrease R</i>	253
c)	<i>Tesseract</i>	260
d)	<i>Passim</i>	261
	<i>Conclusion du chapitre</i>	269

3.	Présentation des logiciels retenus, TextPAIR et Tracer.....	271
a)	<i>Recherches existantes</i>	271
b)	<i>Critères techniques</i>	277
c)	<i>Format</i>	279
d)	<i>Paramétrage</i>	287
e)	<i>Qualité des résultats</i>	288
	<i>Conclusion du chapitre</i>	298
4.	Présentation des résultats initiaux et problèmes rencontrés.....	302
a)	<i>Évaluation des résultats</i>	303
b)	<i>Problèmes de préparation du corpus et du paramétrage des logiciels</i>	318
c)	<i>Problèmes d'exploration et de visualisation</i>	324
	<i>Conclusion du chapitre</i>	343
	Conclusion de la partie.....	345
IV.	Édition numérique intertextuelle : propositions et perspectives.....	347
1.	Post-traitement des résultats de la détection automatique des correspondances.....	350
a)	<i>Unification des formats</i>	351
b)	<i>Détection des noyaux des syntagmes et les unités lexicales liés, puis calcul du score de ressemblance</i>	355
c)	<i>Exploitation des résultats enrichis et du réseau global des réécritures</i>	362
	<i>Conclusion du chapitre</i>	371
2.	Annotation des entités nommées et leurs périphrases.....	372
a)	<i>Détection automatique des entités nommées</i>	373
b)	<i>Création des lexiques et d'un algorithme à base de règles</i>	377
c)	<i>Exploitation des données obtenues</i>	380
	<i>Conclusion du chapitre</i>	393
3.	Présentation de prototype d'une édition comparative numérique.....	395
a)	<i>Première visualisation des relations intertextuelles</i>	396
b)	<i>Édition hyperannotée</i>	399
c)	<i>Édition hypertextuelle</i>	403
	<i>Conclusion du chapitre</i>	410
4.	Quelques perspectives pour le futur du projet.....	412
a)	<i>Prise en compte des séries traductives</i>	413
b)	<i>Amélioration du traitement des synonymes</i>	420
	<i>Conclusion du chapitre</i>	438
	Conclusion de la partie.....	439
	Conclusion.....	440

Bibliographie.....	449
<i>Corpus principal</i>	<i>450</i>
<i>Corpus secondaire.....</i>	<i>453</i>
<i>Études consacrées au mythe d'Orphée et ses réécritures.....</i>	<i>455</i>
<i>Théorie littéraire et littérature comparée.....</i>	<i>457</i>
<i>Linguistique et traitement automatique du langage.....</i>	<i>461</i>
<i>Études consacrées à la détection automatique de l'intertextualité.....</i>	<i>465</i>
<i>Édition numérique et numérisation.....</i>	<i>470</i>
<i>Humanités numériques et informatique.....</i>	<i>472</i>
<i>Lexicographie ; dictionnaires et encyclopédies.....</i>	<i>476</i>
<i>Autres</i>	<i>477</i>
Sitographie	479
<i>Ressources du projet de thèse</i>	<i>480</i>
<i>Logiciels, scripts et interfaces.....</i>	<i>480</i>
<i>Corpus d'auteurs, projets de recherche et éditions numériques</i>	<i>482</i>
<i>Corpus linguistiques, bibliothèques numériques et leurs interfaces d'exploration</i>	<i>483</i>
<i>Dictionnaires, lexiques et WordNets.....</i>	<i>484</i>
<i>Standards, entrepôts et outils de gestion des données</i>	<i>485</i>
<i>Structures et institutions</i>	<i>485</i>
<i>Notices et autres ressources</i>	<i>486</i>

Introduction

Dans son ouvrage, *Pour un humanisme numérique*, Milad Doueïhi écrit : « L’humanisme numérique est [...] le résultat d’une convergence entre notre héritage culturel complexe et une technique devenue un lieu de sociabilité sans précédent »¹. Un des enjeux des humanités numériques littéraires, cette nouvelle discipline étonnante mariant la littérature avec les supports numériques, est de manipuler des œuvres, avec leurs caractéristiques et différences, à l’aide de langages informatiques qui, eux, ne supportent pas d’ambiguïté. Heureusement, ce n’est pas la littérature qui doit se plier aux règles du numérique. C’est surtout ce dernier qui doit évoluer pour répondre aux besoins des chercheurs, voire les susciter. Et si les humanités numériques, cette transdiscipline se situant au croisement de plusieurs domaines et dont les possibilités semblent inépuisables, gagnent de l’importance, c’est entre autres grâce au développement et à la diffusion d’outils ainsi qu’à une politique dynamique de soutien à la recherche. Nous souhaitons que notre thèse s’inscrive dans cette optique de collaboration interdisciplinaire tout en rendant hommage à la tradition de recherche purement littéraire. Nous espérons démontrer que le numérique est propice aux innovations pour la recherche et l’enseignement et qu’il peut profiter tout autant aux chercheurs qu’aux éditeurs, lecteurs et auteurs.

Ainsi, nous examinons les possibilités offertes par le numérique pour l’édition comparative et la recherche automatique des relations intertextuelles à partir d’un corpus littéraire de traductions, parodies, modernisations et adaptations du mythe d’Orphée et Eurydice. C’est à dessein que nous nous tournons vers ce *topos* qui n’a pas cessé d’être une source d’inspiration depuis l’Antiquité, de nombreuses œuvres y faisant référence ou y étant consacrées. Premièrement, la richesse et la variété du corpus exigent la conception d’outils capables de traiter une certaine diversité de textes : des traductions d’Ovide et de Virgile, des transcriptions des manuscrits du Moyen-Âge aux adaptations de ce mythe en bande dessinée, en passant par des textes plus classiquement pris en compte quand on évoque la production littéraire (romans, poèmes, mais aussi pièces de théâtre, livrets d’opéra et partitions de cantates). Ainsi, les outils développés pourront être réemployés sur des corpus différents et profiter à l’édition comparative plus généralement. Simultanément, les outils que nous mettons en place constitueront un nouveau support à l’enseignement littéraire dans le secondaire, à la préparation des concours (CAPES, Agrégation) et même à l’enseignement du FLE, les auteurs qui font partie de notre corpus étant très présents dans les programmes. Par ailleurs, la richesse des recherches déjà menées sur le sujet nous permet d’une part, d’effectuer des analyses outillées en confrontant nos résultats et ceux des chercheurs qui ont déjà traité le point étudié, et, d’autre part, de bénéficier de toute la richesse des stratégies éditoriales déjà mises en œuvre afin de tenter de gagner en lisibilité pour les annotations et variantes fournies aux lecteurs.

¹ Milad Doueïhi, *Pour un humanisme numérique*, Paris, Seuil, 2011, p. 9.

Ce projet, qui se situe dans le domaine des humanités numériques littéraires, vise donc avant tout à interroger les possibilités offertes par l'informatique pour l'analyse et l'édition comparative, en proposant des outils et méthodes réemployables au-delà de notre recherche personnelle. Le travail sous la direction et en collaboration avec des chercheurs qui n'ont pas forcément l'habitude de côtoyer des outils numériques garantit un équilibre entre la partie technique et littéraire de notre recherche et permet de veiller à ce que les outils proposés soient adaptés à tous les acteurs intéressés, qu'ils soient des spécialistes, des élèves, français ou étrangers, ou des lecteurs passionnés. Ainsi, à partir de l'interrogation du corpus des réécritures du mythe d'Orphée et d'Eurydice, nous réfléchissons à des outils de numérisation, de comparaison et d'analyse des relations intertextuelles capables de traiter des œuvres d'époques et de genres différents, permettant de créer des cartes hypertextuelles et interactives des relations détectées dans un ensemble de textes.

Après avoir présenté le mythe d'Orphée et des recherches déjà effectuées sur ses réécritures, nous exposons les trois étapes de notre projet. Premièrement, nous introduisons les modalités de la collecte, de la description et de la numérisation d'un corpus littéraire dans l'environnement numérique. Deuxièmement, nous évoquons les possibilités de recherche (semi-)automatique de correspondances entre les différentes œuvres qui réécrivent ou transposent ce *topos* et nous réfléchissons à la mise en forme numérique des résultats obtenus. Ensuite, nous introduisons nos recherches sur la conception d'outils et de modalités permettant l'amélioration des résultats obtenus automatiquement, leur structuration et leur présentation dans le cadre d'une édition comparative. La dernière section est consacrée à la présentation du plan de cette rédaction.

1. Mythe d'Orphée et Eurydice

Rappelons-nous : inconsolable de la mort d'Eurydice, sa femme, le chanteur prodige Orphée descend aux Enfers pour prier les dieux qu'elle lui soit rendue. Émus par son chant, Hadès et Perséphone lui permettent de l'emmener à condition qu'Eurydice marche derrière Orphée et que celui-ci ne se retourne pas pour la regarder. Cependant, juste avant d'arriver à la surface, le poète regarde derrière lui et perd sa femme à jamais.

L'histoire d'Orphée et Eurydice, dont les principales sources anciennes sont le livre IV des *Géorgiques* de Virgile (37-30 av. J.-C.) et les livres X-XI des *Métamorphoses* d'Ovide (1^{er} s.), a constitué une source d'inspiration traversant l'histoire littéraire depuis le Moyen-Âge jusqu'à nos jours. Elle a aussi été adaptée, parodiée, modernisée ou réinterprétée pour un public plus jeune.

La figure d'Orphée est sujette à de nombreuses études qui regroupent leur corpus diachroniquement. Ainsi par exemple, la présence de ce *topos* au Moyen-Âge est l'objet de l'analyse de John Block Friedman (1970), *Orpheus in the Middle Ages* ou de celle de Florence Lavaud (1995), *Le personnage*

d'Orphée : du mythe latin au roman médiéval. Alors que Françoise Joukovsky (1970) se concentre sur le XVI^e siècle dans *Orphée et ses disciples dans la poésie française et néolatine du XVI^e siècle* et Léon Cellier (1958) étudie les représentations du XIX^e dans son intervention « Le romantisme et le mythe d'Orphée », Emilie Bleschet (2016) tente de tracer l'évolution des représentations à travers les siècles dans *Les représentations du mythe d'Orphée du XVI^e au XIX^e siècle*. Nous observons aussi une effervescence d'études sur la présence du mythe dans la littérature moderne et contemporaine : en 1961, Eva Kushner publie *Le mythe d'Orphée dans la littérature contemporaine*. En 2002 paraît la thèse de Georgios Archimandritis *Le mythe d'Orphée dans le théâtre et le cinéma du XX^e siècle* et en 2013 celle de Barunka Hauptmann « *Son corps déchiqueté* » : les genres à l'œuvre dans le mythe d'Orphée au XX^e siècle. Enfin, le mythe est régulièrement employé pour démontrer les principales méthodes d'analyse des réécritures mythologiques : la mythanalyse de Gilbert Durand (« Les nostalgies d'Orphée. Petite leçon de mythanalyse », 1997), la mythocritique de Pierre Brunel (« Orphée moderne », 1997) ou encore la comparaison différentielle et discursive d'Ute Heidmann (« Différencier au lieu d'universaliser. Comparer les façons de (r)écrire des mythes », 2018). L'article compare la version du mythe des *Géorgiques* de Virgile avec le roman pour la jeunesse *Orphée l'enchanteur* de Guy Jimenes).

2. Corpus littéraires et humanités numériques : mariage de convenance ?

Si, dans ce travail, nous nous focalisons sur les réécritures du mythe d'Orphée et Eurydice, c'est justement pour bénéficier de la richesse des travaux traditionnels qui portent sur ce mythe afin d'expérimenter certaines méthodes informatiques, de valider ou d'invalidier leur pertinence pour la critique littéraire et, enfin, de délimiter des garde-fous permettant d'effectuer plus sereinement des traitements comparables sur un autre corpus qui ne bénéficierait pas d'une telle richesse critique.

Dans notre thèse, il n'est pas question d'affirmer avec certitude que les technologies informatiques permettent de faire de nouvelles découvertes – la lecture attentive d'un nombre restreint des textes ne saura être dépassée par des outils informatiques, et heureusement. Cependant, la manipulation d'un corpus plus grand, et surtout d'un corpus numérisé, offre tout de même de nouvelles possibilités d'analyse, tout en permettant d'observer les relations potentiellement présentes dans notre corpus sans *a priori* théoriques.

a) Collecte du corpus : enjeux interdisciplinaires

Nous ambitionnons à ce que notre corpus soit simultanément cohérent du point de vue littéraire, qu'il réponde aux exigences de la linguistique computationnelle et qu'il puisse être interrogé par des outils informatiques. Par ailleurs, nous souhaitons suivre la déontologie commune établie

en matière de numérisation, description et partage des corpus, notamment en vue de leur publication au sein d'une édition numérique. Nous nous imposons le respect optimal de ces différentes contraintes disciplinaires afin d'assurer l'utilité de notre corpus pour d'autres projets et recherches.

Ainsi, la collecte du corpus est un processus qui fait partie intégrante de notre thèse. Elle inclut la réflexion sur la manière d'explorer l'environnement numérique (fouille des archives, bibliothèques, projets et dépôts, qu'ils soient privés ou institutionnels), tout en évitant les écueils qu'il présente et sans faire abstraction des travaux empiriques déjà existants.

Pour réunir un corpus aussi complet que possible, nous définissons donc un protocole objectif de collecte des œuvres, qui prend en compte la disponibilité variable des textes issus de différentes époques littéraires et le déséquilibre entre le canon et les œuvres oubliées ou dépréciées. Une exhaustivité du corpus n'étant pas atteignable, nous optons pour la conservation optimale de sa représentativité. Ces règles permettent de minimiser les écueils liés aux particularités de l'environnement numérique sans oublier l'objectif principalement littéraire de la recherche.

De cette manière, nous aboutissons à un corpus principal regroupant 52 réécritures françaises du mythe d'Orphée et Eurydice. Publiées depuis le XVI^e siècle, elles se caractérisent par la diversité des états de la langue (du moyen français au français moderne) et des genres (poésie, prose, théâtre, opéra, bande dessinée).

b) Description du corpus : entre principes littéraires et contraintes techniques

Les textes recueillis sont référencés au sein d'une base de données et classés de manière à représenter la variété des réécritures du mythe présentes au sein du corpus. Mises à part les informations bibliographiques et les précisions sur la provenance du document (archives numériques, version papier, projet de recherche, etc.), nous proposons également plusieurs catégories pour décrire le genre de l'œuvre et son aspect intermédial (présence des partitions pour certaines cantates, des illustrations dans les bandes dessinées, etc.). Nous puisons également dans la théorie genettienne² de l'hypertexte pour préciser le type de réécriture (transposition ou parodie), sa diégèse par rapport au mythe (hétero- ou homodiégétique) et la transformation du récit source (transmodalisation, détournement, modernisation, transformation idéologique ou pragmatique).

Cela permet une description détaillée des œuvres collectées. Alors que certaines, comme *La descente d'Orphée aux Enfers* de Charles de l'Espine (1623) ou la cantate *Orphée* de Berlioz (1827) approchent le mythe de manière très classique, d'autres, comme *Eurydice* de Jean Anouilh (1942) ou *Orphée* de Jean Cocteau (1927), sont difficilement classables dans une seule catégorie. Nous

² Gérard Genette, *Palimpsestes. La littérature au second degré*, Paris, Seuil, coll. « Poétique », 1982.

tâchons de faire en sorte à ce que chaque ensemble soit riche et diversifié, et qu'il ne soit limité ni à une époque ni à un genre précis.

Certaines de ces œuvres, comme *Orphée* de Ballanche (1827-1829), *Orphée. Cantate* de Rameau (1721) ou *Orphée et Eurydice* de Moline et Gluck (1774) sont connues et très largement analysées par les recherches littéraires. D'autres ne bénéficient pas d'une telle renommée : alors qu'il est difficile d'évaluer la popularité de certaines cantates du XVIII^e s., nous pouvons affirmer qu'au moins trois poèmes du XX^e font partie du non-canon : *Orphée* de Ferdinand Divoire (1922), *Orphée : poème* de Louis de Launay (1922) et *Orphée, poème* de Jean Loyson (1948) ont été mis à disposition des lecteurs par la Bibliothèque nationale de France au sein du projet de numérisation des indisponibles³ et sont considérés comme œuvres orphelines.

En outre, les liens entre certains textes sont avérés : c'est le cas par exemple du livre 5 « Orphée » des *Grands initiés* de Schuré (1889) qui a, selon Brunel⁴, inspiré l'opéra inédit *Orphée-roi* de Segalen (1921), ou « Sixième Fragment » de Ballanche qui peut être considéré comme prolégomènes des œuvres postérieures de l'auteur. Ces textes nous permettent de confronter les observations empiriques des chercheurs littéraires avec les résultats de notre recherche automatique.

D'autres composantes du corpus représentent les courants qui se profilent dans la recherche diachronique : les réécritures sentimentales au théâtre et à l'opéra entre XVII^e et XVIII^e s. ou les réécritures symboliques s'approchant davantage de l'aspect religieux de l'orphisme (présent notamment dans les réécritures du XIX^e s.). Pour le XX^e s. nous retenons principalement deux tendances : les modernisations au théâtre de Cocteau et d'Anouilh et les œuvres se focalisant sur le mythe de l'Orphée noir (le roman *Orphée-dafric* de Werewere Liking et *Orphée d'Afrique. Théâtre-rituel* de Manuna Ma Njock, 1981).

Enfin, nous nous intéressons également aux réinterprétations particulières du mythe. Ainsi, nous recueillons ses adaptations contemporaines pour le jeune public, comme le roman *Orphée l'enchanteur* de Guy Jimenes (2004), l'album de jeunesse *Orphée et la morsure du serpent* d'Yvan Pommaux (2009) ou la bande dessinée *Orphée et Eurydice* de Béatrice Bottet et Ariane Pinel (2019). Nous cherchons aussi les différents détournements humoristiques du mythe, comme la mazarinade anonyme *Orphée grotesque avec le bal rustique en vers burlesques* (1649) ou l'opéra burlesque *Orphée aux enfers* de Jacques Offenbach (1858). Les parodies des autres œuvres du corpus sont également présentes, comme la cantate de Grandval (1729) qui reprend celle de Clérembault (1717) ou la pièce *Roger-*

³ Voir <https://reliure.bnf.fr/projet-relire-dispositif-inedit>. Tous les URL ont été consultés en juillet 2025. Les liens obsolètes ont été remplacés par la dernière version archivée par *Internet Archive Wayback Machine* (<https://web.archive.org/>).

⁴ Pierre Brunel, « *Orphée-roi* de Victor Segalen, ou le miracle de la lyre », *Mythocritique : théorie et parcours*, Paris, Presses Universitaires de France, coll. « Écriture », 1992, p. 190-202.

Bontems et Javotte, parodie d'Orphée et Eurydice de Dorvigny et Moline (1775) qui parodie l'opéra de Gluck.

Dans notre recherche, nous effectuons des analyses et des comparaisons sur différents ensembles et entre les éléments qui les constituent en veillant au traitement égalitaire des œuvres de genres, d'époques et destinées à des publics différents, afin de pouvoir appuyer notre étude sur la diversité du corpus. Ce dernier est accompagné par un corpus secondaire composé de 36 traductions du livre IV des *Géorgiques* de Virgile et des livres X et XI des *Métamorphoses* d'Ovide, des traductions académiques, littéraires et juxta-linéaires, comme celles de Jacques Delille ou de Thomas Corneille, et même deux adaptations anonymes du tournant des XIII^e et XIV^e siècles, *Ovide moralisé* et *Ovide moralisé en prose*.

c) Numérisation et partage du corpus : défis de la science ouverte

Pour constituer et étudier un corpus aussi important dans le temps imparti d'une thèse, nous avons recours à des outils de collecte des textes disponibles en versions numériques et de numérisation (semi-)automatique de ceux qui ne le sont pas, dans un format propice à l'extraction et la manipulation de données textuelles. Pour ce faire, nous nous tournons, entre autres, vers le langage XML, pleinement adapté au traitement des textes et des images en l'appuyant sur la syntaxe TEI (*Text Encoding Initiative*), un format ouvert et ayant vocation à faciliter les réemplois des ressources enrichies. La TEI est actuellement exploitée par les chercheurs (en littérature, linguistique, humanités numériques, etc.) et les éditeurs universitaires, ce qui leur donne accès à des outils de traitement communs. Nous souhaitons adhérer à cet esprit de partage en créant des outils de structuration et de mise en forme des contenus dédiés aux éditions comparatives, qu'elles soient destinées à des spécialistes ou à d'autres lecteurs curieux et passionnés.

Pour ce faire, il nous semble important de concevoir des outils (semi-)automatiques permettant la structuration des œuvres, la distinction d'éléments textuels, comme les champs lexicaux ou les entités nommées (antonymes, toponymes, etc.). Les antonomases et périphrases de noms propres présents dans le mythe, comme « Thracien », « harpeur », « doux-sonnant », sont identifiées avec leurs référents (Orphée). Les champs lexicaux récurrents du *topos* orphique, comme celui de la musique, de la poésie, de l'amour ou de la perte, sont aussi dégagés du corpus des traductions en prenant en compte le vocabulaire des différentes époques.

La richesse des annotations résultant de ces traitements nous engage également à sonder les nouvelles possibilités que l'affichage numérique peut offrir à l'édition comparative et à l'enseignement littéraire en ce qui concerne la présentation des œuvres et de leur appareil critique.

À l'aboutissement de ce projet, en plus du corpus XML-TEI annoté, tous les livrables de notre recherche seront accessibles à la communauté des chercheurs grâce au dépôt de données dans des entrepôts et plateformes dédiés. Ainsi, ils pourront ensuite être enrichis et réutilisés par d'autres chercheurs afin d'élargir les aboutissements de ce projet et ouvrir de nouvelles perspectives de recherche.

3. Détection automatique de l'intertextualité : une perspective ambivalente

Alors que la comparaison et le repérage intertextuel manuels d'un corpus aussi imposant et diversifié ne sont pas possibles, le traitement automatique, appuyé sur des techniques d'analyse linguistique et taliste, comme la lexicométrie, la stylométrie et la fouille textuelle, peut donner des résultats fructueux, par le repérage de blocs communs, du vocabulaire récurrent dans chaque sous-corpus ou des rapprochements syntaxiques.

Le corpus des réécritures du mythe d'Orphée et Eurydice est particulièrement adapté à l'expérimentation des techniques informatiques de la détection automatique des similarités, de par la richesse des œuvres et des relations qu'elles peuvent entretenir entre elles. D'abord, plusieurs "chaînes de réécriture" peuvent être observées, comme pour l'opéra *Orphée et Eurydice* de Gluck (1774), qui s'inspire des sources anciennes et, sans doute et entre autres, d'*Orfeo* de Monteverdi (1609), et qui a été à son tour parodié dans les opéras comiques *Roger-Bontems et Javotte* (Dorvigny et Moline, 1775), *Le Petit Orphée* (Deshayes et Rouhier-Deschamps, 1792) ou encore *Orphée aux Enfers* (Crémieux et Offenbach, 1858). Ensuite, le rapport au mythe dépend du type de réécriture : alors que les parodies ridiculisent la musique d'Orphée ou son amour pour Eurydice, les modernisations incluent des amalgames, des personnages additionnels ou des décors modernes qui réorientent l'interprétation du mythe⁵. Enfin, chez certains auteurs, comme Ballanche, Orphée devient une figure récurrente et le mythe se module d'une œuvre à l'autre pour devenir une constante créatrice, parfois implicite, parfois centrale.

a) Réécriture, reformulation, allusion : préoccupations terminologiques

La définition des relations entre les textes de notre corpus n'est pas toujours aisée. Ces difficultés peuvent être macrostructurelles (identification du texte source, détermination de la manière dont le mythe est abordé, etc.) ou microstructurelles (identification des citations, modification ou suppression de certains myèmes, etc.). Les concepts théoriques qui peuvent nous servir des guides,

⁵ Voir notamment Pierre Brunel, « Orphée moderne », *Apollinaire entre deux mondes. Le contrepoint mythique dans Alceste. Mythocritique II*, Paris, Presses Universitaires de France, coll. « Écriture », 1997, p. 63-82.

comme *intertextualité* et *réécriture*, sont complexes et, souvent, sujets à controverses. Bien des chercheurs se sont penchés sur les relations qui peuvent être observées entre les textes littéraires sans toujours arriver à une vue convergente et une panoplie de termes a été proposée. La frontière entre eux est floue et la critique littéraire ne la délimite pas toujours clairement. À cela s'ajoute l'aspect linguistique de notre recherche qui ne peut pas être ignoré. Les sources de la conceptualisation de l'intertextualité sont linguistiques, l'intertextualité reste une branche de la linguistique textuelle, etc.

Ainsi, différentes théories intertextuelles peuvent être purement littéraires (structuralistes, comparatistes, etc.), hybrides ou encore uniquement linguistiques, et s'intéresser de manière générale aux processus de reformulation. Par ailleurs, d'autres chercheurs ont fondé des théories spécifiques, focalisées sur les mythes et issues principalement de la littérature comparée (mythanalyse, mythocritique, comparaison différentielle et discursive, etc.).

Il est impossible de proposer dès l'introduction une terminologie fixe adoptée pour cette thèse. Nous y consacrons un chapitre entier en rappelant l'histoire critique de l'intertextualité et en réfléchissant sur la manière dont les différentes théories s'appliquent à l'observation pratique de notre corpus. Ce faisant, nous constatons que l'hétérogénéité des approches critiques coïncide dans notre corpus avec l'impossibilité fréquente d'assigner sans aucune ambiguïté une relation observée à un concept spécifique. Par exemple, nous observons des relations hypertextuelles, telles que définies par Genette, entre la cantate *Orphée* de Clérembault (1710) et sa parodie par Grandval (1729). Nous remarquons aussi des allusions plus ponctuelles dans la pièce de Cocteau qui reprend le personnage d'Aglaonice, introduit en 1889 par Schuré. Nous percevons aussi des relations plus subjectives, qui peuvent relever de l'intertextualité telle que la définit Roland Barthes, comme dans « Une histoire renouvelée des Grecs » de D.L. Eimann (1854) où la ridiculisation humoristique du poète fait écho à un poème anonyme *L'Orphée grotesque* écrit presque 200 ans plus tôt :

<i>L'Orphée grotesque</i> , 1649	« Une histoire renouvelée des Grecs », 1854
« L'écho se plaint d'être étourdie de sa crierde mélodie [...]. Sa plainte jouant de son reste, il maugrée, fulmine, il peste [...]. Mais il ne s'en prend désormais qu'à la barbe qui n'en peut rien, et s'arrachant sa heure fauve, de male rage devient chauve » ⁶	« Pendant sept mois, il vécut dans les bois, dans les antres, jetant aux échos ses gémissements, ses grincements de dents, oubliant probablement de peigner sa barbe et ses cheveux, attendant que la mort mît fin à sa douleur »

⁶ Nous accompagnons cette rédaction des annexes, qui incluent le corpus numérisé dans différents formats (PDF, texte, XML..., voir [I 4 Corpus-Nakala](#)). La pagination des éditions sources est reproduite. Pour des questions pratiques et pour rendre compte de la spécificité du traitement numérique massif, nous ne donnons pas de références exactes des extraits cités et invitons le lecteur à consulter le corpus numérisé pour retrouver la page et/ou le contexte de la citation.

Toutefois, pour la majorité des textes de notre corpus, le mythe d'Orphée et Eurydice est le seul point commun identifié. Il faut donc revenir vers ses sources antiques pour observer et analyser tant les proximités que les différences dans la réinterprétation du récit original.

Par ailleurs, dans sa mythocritique, Brunel parle de la flexibilité des mythes dépendant des expériences de lecture et souvenirs, parfois altérés, de l'auteur. Alors que certains auteurs ont recours à la citation d'une traduction française, d'autres composent leur récit sans s'appuyer directement sur un texte source précis. Ils en appellent plutôt à leur mémoire, qui parfois s'avère défaillante. Tous peuvent aussi confondre différentes sources, modifier les épisodes originaux ou encore mélanger les sources anciennes avec des réécritures plus connues.

Nous découvrons que l'intégration des outils numériques permet de rendre compte de cette complexité observable au sein du corpus. Notre objectif est alors de démontrer qu'une combinaison des méthodes et analyses empiriques avec les traitements (semi-)automatiques apporte une nouvelle dimension à l'intertextualité, que nous définissons comme rhizomatique, en opposition aux conceptualisations plutôt dualistes du terme (interdépendance entre deux textes ou extraits). Dans ce modèle, différents composants du « tissu »⁷ peuvent être liés selon différents traits, tantôt directement, tantôt indirectement et indépendamment de la chronologie. Cette conceptualisation ouvre la voie à des traitements informatiques permettant de situer et d'apprécier différentes relations dans un réseau non ordonné et élargit les possibilités d'analyse vers des horizons encore inexploités. Ainsi, nous ne visons pas à redéfinir le concept d'intertextualité, mais à démontrer qu'elle gagne à être considérée dans sa pluralité.

b) Intertextualité quantitative : jeu d'équilibriste entre *close* et *distant reading*

Plusieurs inquiétudes ont été formulées concernant l'application des méthodes informatiques aux études littéraires, notamment la lecture distante (*distant reading*) de Franco Moretti et les approches quantitatives. Ces critiques sont tout à fait légitimes, surtout que l'efficacité de ces méthodes par rapport aux approches traditionnelles reste à prouver.

Certes, notre recherche est en partie quantitative, mais elle ne fait pas le deuil de la qualité. Les exploitations numériques que nous présentons dans ce travail ne visent pas à savoir *combien* il y a de relations dans un texte, mais à essayer de comprendre *comment* ces relations se réalisent et quels indices permettent de les identifier.

Nous ne sommes pas la première à s'intéresser à la détection automatique des relations intertextuelles. Tout au long de cette thèse nous citons de nombreux travaux, théoriques et pratiques, effectués sur différents corpus et ayant, chacun, des objectifs de recherche bien particuliers. Ce qui

⁷ Nous faisons allusion à la définition du « Texte » proposée par Barthes dans l'*Encyclopædia Universalis* (1973).

semble commun à tous, c'est la curiosité interdisciplinaire de ces expérimentations : ce n'est pas la méthodologie informatique qui importe, mais de savoir comment les outils et méthodes des autres disciplines peuvent participer à la création de nouveaux matériaux interprétatifs.

Dans l'introduction de l'ouvrage *Quantitative Intertextuality. Analysing the Markers of Information Reuse*, Forstall et Scheirer définissent l'intertextualité quantitative comme l'étude de l'intertextualité avec des méthodes informatiques⁸. Les auteurs se positionnent

en faveur de l'unification d'un vaste ensemble de phénomènes apparemment disparates dans un seul modèle. Ce modèle unificateur est en grande partie construit sur la compréhension structuraliste et post-structuraliste des œuvres littéraires, mais il est soutenu et renforcé par une réflexion plus récente dans les sciences cognitives. Non moins important, le modèle est également ancré dans le résultat pratique de la recherche en humanités numériques.⁹

Pour s'inscrire pleinement dans l'optique de l'intertextualité quantitative, Forstall et Scheirer préconisent trois étapes : (i) l'extraction des caractéristiques du corpus, (ii) la détection automatique des relations et (iii) l'interprétation des résultats¹⁰. Nous suivons ces recommandations.

Pour extraire les caractéristiques de notre corpus, nous usons d'une part des autres techniques issues des humanités numériques, comme la stylométrie et la détection des topiques et nous exploitons, d'autre part, notre corpus secondaire, composé des différentes traductions des sources anciennes du mythe. Ce corpus nous permet par exemple, en amont de la détection automatique des correspondances, de comprendre et de structurer les types des relations que nous sommes susceptibles de trouver dans notre corpus des réécritures.

Parmi les phénomènes pouvant être regroupés sous l'égide de l'intertextualité quantitative, nous trouvons notamment la détection automatique des réutilisations textuelles (*text reuse*¹¹), qui vise à identifier, dans un texte, des fragments identiques à un autre texte, ou suffisamment proches pour pouvoir considérer qu'il s'agit d'un emprunt. Cette détection est effectuée au moyen d'un ensemble d'algorithmes, pour la plupart rassemblés au sein d'un logiciel.

Ces outils sont destinés originellement aux sciences dures, comme la bio-informatique pour l'algorithme d'alignement par fragments grâce à la détection des similarités. Cette méthode de détection des séquences de macromolécules (ADN ou protéines) permet de faire ressortir les régions

⁸ « the study of intertextuality through computational means » (nous traduisons). Christopher Forstall et Walter Scheirer, « What is Quantitative Intertextuality? », *Quantitative Intertextuality: Analyzing the Markers of Information Reuse*, Cham, Springer International Publishing, 2019, en ligne, <https://doi.org/10.1007/978-3-030-23415-7_3>, p. 3.

⁹ « The first part is an argument for uniting a broad set of seemingly-disparate phenomena in a single model. That unifying model is in large part built on structuralists and post-structuralist understanding of literary works, but is supported and strengthened throughout by more recent thought in cognitive science. No less important, the model is also grounded in the practical result of research in the Digital Humanities. We call this approach *quantitative intertextuality* » (nous traduisons). *Ibid.*, p. 4.

¹⁰ Christopher Forstall et Walter Scheirer, « What is Quantitative Intertextuality? », *op. cit.*, p. 6.

¹¹ Nous utilisons plusieurs termes pour décrire ce phénomène (*homologies*, *similarités*, *correspondances*) sans que le sens change de manière significative.

homologues. Elle est exploitée par exemple par les créateurs du logiciel d'alignement de versions MEDITE (laboratoires ITEM et ACASA). Limité aux comparaisons entre deux œuvres en format texte brut, ce logiciel est insuffisant pour nos besoins, mais il peut guider vers les méthodes à mettre en place, tout comme les logiciels de détection de plagiat (Compilatio, Scibbr), de fouilles textuelles (TXM, Hyperbase, Lexico3), de visualisation des relations (Gephi), et certainement d'autres encore.

c) Tracer et TextPAIR : outils perfectibles

Nous identifions également des développements créés avec l'ambition de permettre la détection des similarités plus subtiles, allant des amplifications, réductions ou reformulations jusqu'aux allusions et métaphorisations. Nous trouvons au total sept logiciels spécifiquement dédiés à la détection des relations textuelles, dont six sont encore fonctionnels. À notre connaissance, cette thèse propose la première étude comparative de ces outils.

Même si les critères que nous établissons sont étroitement liés à nos questions de recherche et à la spécificité de notre corpus, nous sommes persuadée que ce travail facilitera les choix d'outillage pour d'autres projets. Nos résultats sont par ailleurs surprenants, car ce ne sont pas les logiciels les plus récents qui se montrent les plus performants et les plus aptes à détecter les relations plus subtiles qu'un simple copier-coller.

Nous n'entrons pas ici dans les détails techniques du fonctionnement des deux gagnants, TextPAIR et Tracer, et de la différence entre eux – un chapitre y est consacré. Il est toutefois important de souligner que les deux outils permettent un large spectre de paramétrages. On peut notamment ignorer les mots les plus fréquents ou les mots faibles, appuyer la recherche sur les mots (*plaindre~plaindre*), les lemmes (*plaint~plaignent*) ou les racines (*plaindre~plainte*), uniformiser les mots similaires (*plainte~plaintie* [ancien français]) ou encore définir le nombre minimal des éléments communs et de la distance entre eux.

Nous soumettons notre corpus des réécritures à ces deux logiciels et évaluons manuellement un échantillon constitué de 1 000 couples d'extraits aléatoirement choisis parmi ceux mis en relation par TextPAIR et Tracer. Nous constatons alors que presque la moitié lie 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, dans le cadre de cette recherche particulière, peu pertinentes. Par ailleurs, pour environ 15 % des couples d'extraits, le contexte fourni par les deux logiciels ne permet pas d'apprécier leur pertinence.

Malgré des résultats probants, que nous présentons, le taux d'erreur des mises en relations d'extraits textuels est trop élevé pour pouvoir se fier à eux, et le nombre d'extraits impliqués est trop important pour examiner manuellement chaque couple. Donc, pour dégager les relations les plus

pertinentes, nous proposons un développement qui vise l'amélioration des résultats initiaux, avant leur visualisation et interprétation.

La linguistique et les outils du traitement automatique des langues s'avèrent particulièrement propices pour l'exploration et l'enrichissement de ces résultats. De l'annotation morpho-syntaxique à l'exploration des périphrases des entités nommées, ces manipulations permettent d'adapter les traitements généralistes proposés par TextPAIR et Tracer à la spécificité de notre corpus et de se focaliser sur les facteurs communs balisés, comme les entités nommées et leurs périphrases (« Hyménée » / « Hymen » / « dieu de noçoiement »), la synonymie (« elle fut replongée à nouveau au même lieu » / « elle roule au lieu où elle était avant »), l'hyperonymie (« il fut présent à la vérité » / « il est là ») ou encore la métaphorisation (« elle ne put s'animer bien qu'il l'agite » / « le dieu qui l'agite ne peut ranimer ses mourantes clartés »), etc.

Plusieurs difficultés peuvent également être surmontées grâce aux ressources déjà créées ou numérisées par des chercheurs, comme les lexicographes ou les généticiens. Ainsi, les différences entre des textes résultant de l'évolution de la langue peuvent être atténuées si nous tâchons de moderniser (semi-)automatiquement les graphies des textes à l'aide de bases lexicales comme le *Dictionnaire du Moyen Français* (de l'ATILF), et le *Lexique d'ancien français* de Douglas C. Walker (1985). Les repérages de blocs communs peuvent également être étendus aux reformulations synonymiques si nous ajoutons à notre traitement des dictionnaires appropriés comme celui des synonymes développé au CRISCO (Université de Caen).

Notre développement, en accès libre et adaptable à d'autres corpus¹², permet dans un premier temps d'unifier les différents formats des résultats fournis par TextPAIR et Tracer (compilation, filtrage, structuration), initialement très peu lisibles. Il propose ensuite une analyse détaillée de chaque couple de correspondances pour détecter les noyaux des syntagmes et les unités lexicales alignées, calculer le score de ressemblance, puis écarter les couples trop faiblement liés. Enfin, il propose également d'ajouter d'appariements plus fins, notamment entre diverses dénominations d'entités nommées.

4. Édition numérique comparative : un nouvel outil d'analyse

L'édition numérique comparative des réécritures du mythe d'Orphée et Eurydice au sein d'une interface enrichie d'annotations et de visualisation des relations intertextuelles, qui est le produit

¹² Pour son application dans un autre projet, voir Andrea Del Lungo et Karolina Suchecka, « Projet eBalzac : construire une bibliothèque hypertextuelle des sources intellectuelles », dans Clarisse Bardiot, Émilien Ruiz et Esther Dehoux (dir.), *La Fabrique numérique des corpus en sciences humaines et sociales*, Presses universitaires du Septentrion, 2022; Andrea Del Lungo et Karolina Suchecka, « Le projet eBalzac : un hypertexte des sources scientifiques », *Balzac et les disciplines du savoir*, Classiques Garnier, à paraître.

final de ce travail doctoral, vise à combiner un héritage culturel spécifique avec les technologies les plus récentes, et de valoriser les réécritures et l'intertextualité. Ce projet éditorial n'a pas seulement pour but d'exploiter les possibilités offertes par les outils informatiques, mais aussi de réfléchir aux modalités d'adaptation des contenus enrichis à la lecture numérique, qui diffère de la lecture papier de manière significative, notamment par sa non-linéarité. Il ne se restreint pas à l'exploitation, à la mise en valeur ou à la numérisation d'un corpus. L'originalité de son approche numérique tient à l'investigation fine opérée sur des corpus massifs qu'il eut été impossible d'exploiter manuellement.

a) Faire parler les données : défis de la modélisation

En travaillant avec les logiciels Tracer et TextPAIR, nous sommes amenée à nous interroger sur la possibilité d'apprécier la pertinence des résultats obtenus face à un fichier tabulé, le résultat brut du traitement avec les deux outils, souvent de plusieurs milliers de lignes, présentant les couples de similarités hors contexte et dépourvus d'informations concernant les éléments qui ont permis leur mise en correspondance. Comment, ensuite, les présenter aux lecteurs de manière lisible et attrayante ?

Pour répondre à ces questions, nous explorons dans un premier temps les possibilités offertes par des outils de visualisation des relations intertextuelles déjà existants, comme TRAViz¹³, Collatex¹⁴ ou Intertextual Hub¹⁵. Nous réalisons alors qu'une édition numérique est un projet interdisciplinaire par excellence. Elle combine les tâches éditoriales classiques (établissement du texte, conception de l'apparat critique) et celles du Web design, de la visualisation de l'information, voire du génie logiciel. Simultanément, les besoins du public cible ne sont pas toujours faciles à cerner, alors que les principes de l'édition numérique sont de plus en plus formalisés au sein de la communauté des humanités numériques.

À partir de ces observations et de la spécificité de notre corpus, nous concevons une interface modulable qui puise dans la théorie des graphes tout en se préoccupant des habitudes de lecture des chercheurs littéraires et des exigences éditoriales traditionnelles. Ce prototype vise à être soumis à l'évaluation des experts à l'issue de ce projet de thèse, puis mis en ligne et étendu à un public plus large.

¹³ Voir <http://www.traviz.vizcovery.org/index.html>.

¹⁴ Voir <https://collatex.net/>.

¹⁵ Voir <https://intertextual-hub.uchicago.edu/>.

b) Notes intertextuelles et graphes des relations : outils éditoriaux innovants

Ainsi, la première vue, au niveau de l'œuvre, permet à l'utilisateur une lecture linéaire de chaque texte du corpus dont la mise en ligne n'est pas restreinte par les droits d'auteur. Les parties qui présentent les similitudes avec d'autres textes sont signalées au moyen des notes "intertextuelles" qui affichent les textes correspondants en vis-à-vis et proposent un détail des unités lexicales liées (Figure 1 et Figure 2¹⁶).

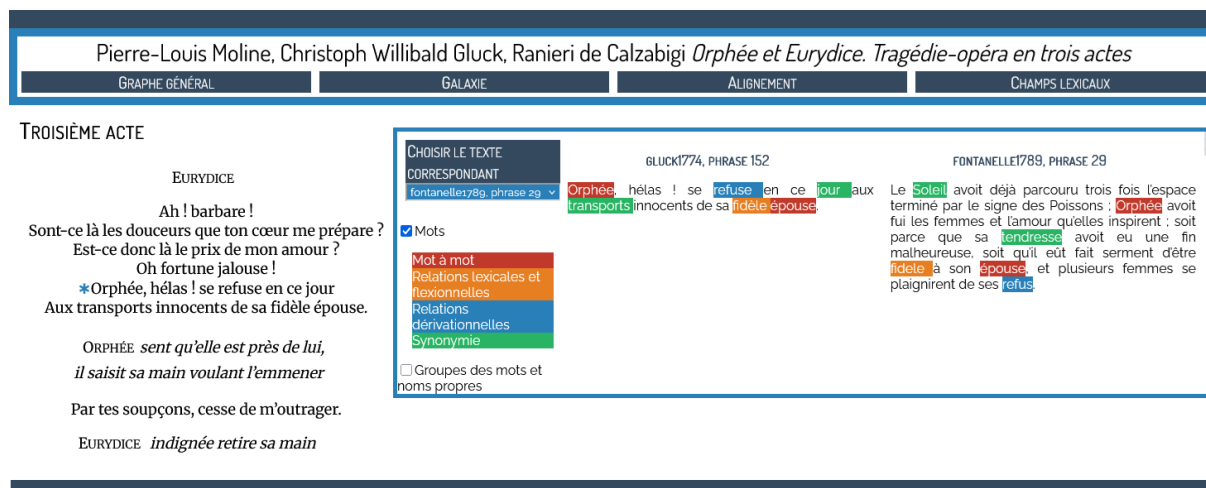


Figure 1. Interface d'édition, note intertextuelle : relation entre l'opéra *Orphée et Eurydice* et une traduction d'Ovide

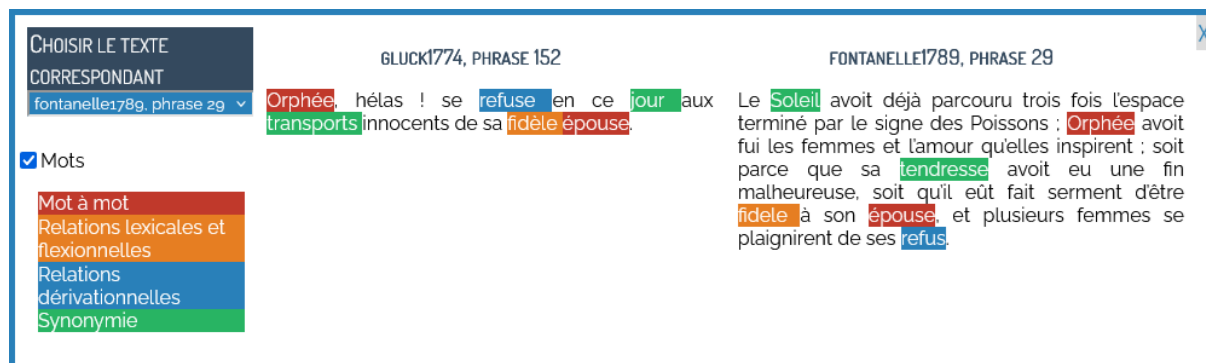


Figure 2. Détail de la note intertextuelle

Les formes identiques sont signalées en rouge (*Orphée, épouse*), les lemmes en orange (*fidèle~fidele*), les unités des différentes classes grammaticales, mais ayant la racine commune sont en bleu (*refuse~refus*) et les synonymes sont en vert (*jour~Soleil, transports~tendresse*). Ce couple, présenté en détail à la Figure 2, relie un extrait de l'opéra *Orphée et Eurydice* de Moline et Gluck (1774) avec une traduction d'Ovide de Fontanelle, presque contemporaine à la pièce (1789).

¹⁶ Chaque œuvre est dotée d'un identifiant unique, composé du nom du premier auteur et de la date de publication. Nous introduirons progressivement ces identifiants dans notre rédaction en utilisant une police dédiée pour les diffrencier du corps du texte (par exemple Fontanelle1789).

Les deux textes ne relatent pas le même épisode du mythe : dans la traduction de Fontanelle, il s'agit de l'épisode suivant la seconde mort d'Eurydice, le deuil d'Orphée et son refus des avances des Bacchantes. Chez Moline et Gluck, la réplique est donnée pendant la sortie des Enfers. Cette relation forte, avec cinq unités communes, s'explique par la réécriture particulière de la pièce où c'est Eurydice qui tente de convaincre Orphée de regarder en arrière. Cela prouve donc l'efficacité de notre développement à détecter des réécritures particulières des mythes classiques.

Nous couplons cette édition plutôt traditionnelle à un autre affichage qui nous paraît tout aussi lisible et intuitif et qui se fait sous forme d'une carte des relations interactive et hypertextuelle. Cette visualisation se fait à l'aide de graphes relationnels (Figure 3), une méthode de modélisation des données qui s'applique le mieux à notre sens pour montrer le réseau rhizomatique des relations intertextuelles. C'est une présentation innovante, mais en même temps ergonomique et simple à manipuler.

Chaque nœud correspond à un extrait présent dans les résultats, identifié à l'aide d'une étiquette qui fournit l'identifiant de l'œuvre et l'emplacement de l'extrait (le numéro de la phrase). Les formes des nœuds donnent des repères supplémentaires quant au type de réécriture (traduction, modernisation, transmodalisation et parodie) et les couleurs des bordures permettent d'identifier le type de l'œuvre (poésie, prose, opéra, théâtre, bande dessinée et autre).

Le graphe à la Figure 3 montre les relations intertextuelles détectées pour trois extraits de la cantate parodique *Orphée* de Grandval¹⁷. La source directe de cette parodie, la cantate de Clérembault, est correctement identifiée par notre développement, car les trois extraits de Grandval sont mis en correspondance avec les vers qui les ont inspirés¹⁸. Parmi les relations détectées, nous comptons principalement d'autres cantates et opéras du XVIII^e s., ce qui montre que Grandval ne parodie pas seulement Clérembault, mais puise dans l'esthétique commune à toute la production générique de son époque¹⁹. Par ailleurs, une autre parodie, cette fois-ci à l'opéra, détourne le mythe de manière similaire : deux extraits du *Petit Orphée* de Deshayes et Rouhier-Deschamps ont été mis en relation

¹⁷ Phrases 18-20 (à gauche du graphe) : « Je commence à prendre courage ; / Ce gouffre obscur m'ouvre un passage / pour pénétrer aux sombres bords : / C'est trop faire icy le Jocrisse ; / J'aime autant rester chez les morts / Que d'être au lit sans Euridice. / Allez Orphée, allez au Diable, / Servez d'exemple à l'univers ; / Pour rejoindre un objet aimable / Il est beau d'aller aux Enfers : » ; phrases 27-28 (en haut du graphe) : « Pour en perdre la mémoire / Dans le fleuve d'oubly va t'en boire. / Du sort réparez le caprice, / Laissez-vous toucher par mes pleurs : / Rendez-moy ma chère Euridice, / Ne séparez pas nos deux cœurs. » et phrases 31-33 : « Laissez-vous toucher par mes pleurs, / Rendez-moy ma chère Euridice, / Ne séparez pas nos deux cœurs. / Vous avez ressenti la flâme / Du Dieu dont j'éprouve les traits, / Quand vous prites pour vostre femme / L'aimable fille de Cérés. / Par cette Épouse si chère / Lachez l'objet de mes vœux. ».

¹⁸ Phrases 11-12 : « Ce gouffre obscur m'offre un passage / Pour pénétrer aux sombres bords / Portons-y mon amour, ma douleur, ma rage », phrases 14-15 : « Allez Orphée, allez, allez, / Que votre amour extrême / Serve d'exemple à l'univers. / Il est beau qu'un mortel / Passe jusqu'aux Enfers / Pour se rejoindre à ce qu'il aime. » et phrases 24-25 : « Laissez-vous toucher par mes pleurs, / D'un sort affreux réparez le caprice, / Rendez-moi ma chère Euridice, / Ne séparez pas nos deux cœurs. / Vous avez ressenti la flamme / Du Dieu dont j'éprouve les traits / L'aimable fille de Cérés / Par ses divins appas / Sut embrasser votre âme. »

¹⁹ Nous y retrouvons d'ailleurs la réplique célèbre de Gluck : « Laissez-vous toucher par mes pleurs, spectres... ».

avec ceux de Grandval²⁰. Enfin, la présence de deux traductions, une de Virgile de Guillois (1863) et une d'Ovide par Massac (1617) montre que les sources anciennes du mythe peuvent bien être identifiées automatiquement.

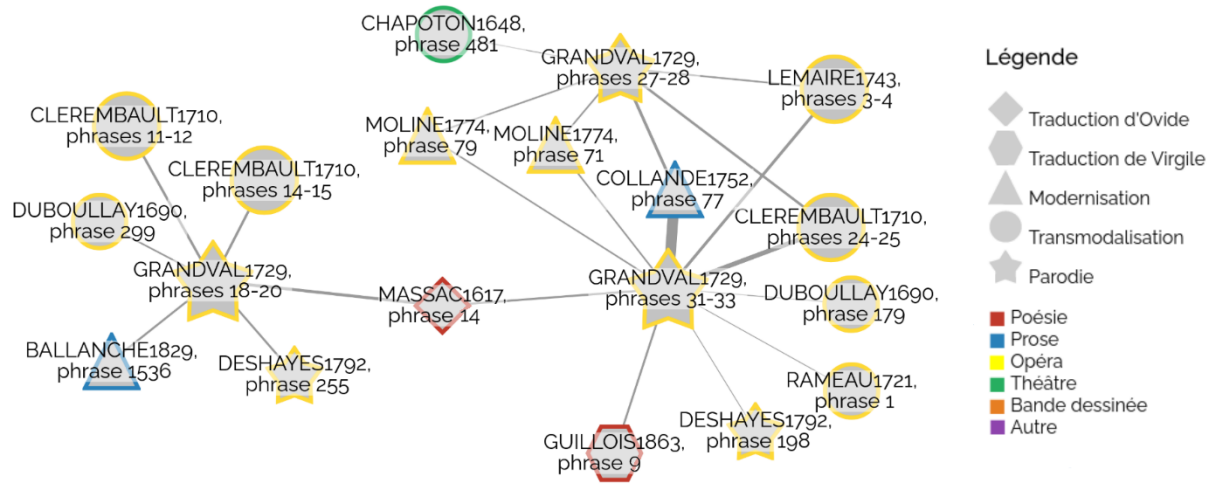


Figure 3. Visualisation en graphe : carte hypertextuelle des relations

CHOISIR LE TEXTE CORRESPONDANT

martin1649, phrase 1

☒ Mots

Mot à mot

Relations lexicales et flexionnelles

Relations dérivationnelles

Synonymie

☒ Groupes des mots et noms propres

Changement ou périphrase d'un nom propre

GRANDVAL1729, PHRASE 40

[40] Je viens expres au **Sabat**, en robbe longue, en rabat, chercher **mon Euridice** qui manque à mon **grabat**.

MARTIN1649, PHRASE 1

[1] Un Vilon yure à la Feste la nuit m'a tant rompu la teste, m'a tant lassé dans mon **grabat** par sa musique de **sabat**, qu'en dépit de sa serenade dont j'ay l'oreille encor malade, je peins d'ancre et non de couleur, ce Menestrier de malheur, qui sonnoit pour feu **sa Donzelle**, sur sa lyre en forme de vielle, donzelle morte à ce qu'on dit, par un lazard qui la mordit, et chantoit non l'epitalame, mais l'epitaphe de **sa Dame** :

GRANDVAL1729, phrase 40

MARTIN1649, phrase 1

Figure 4. Apparition de la fenêtre de comparaison en cliquant sur le lien qui relie deux nœuds.

Les deux visualisations peuvent être combinées, car la note intertextuelle s'affiche également lors d'un clic sur le lien qui relie deux nœuds (Figure 4). Cet affichage combiné facilite l'exploration du réseau des correspondances et l'identification des relations pertinentes. La Figure 4 présente l'unique relation entre la cantate de Grandval et la première partie de la mazarinade anonyme *Orphée grotesque en vers burlesques* publiée par Sébastien Martin. Tout d'abord, nous y trouvons le vocabulaire

²⁰ Il s'agit la phrase 198 « Eurydice, objet de mes vœux est inutile dans ces lieux » et la phrase 255 : « Ah ! Si l'objet de ma constante flamme bravait pour moi la mort et les enfers, heureuse encor de régner sur son âme je le suivrais au bout de l'Univers ».

commun extérieur au mythe d'Orphée et Eurydice, *grabat* et *sabat*, ce qui prouve que notre développement est en mesure de procurer des indices sur le type de réécriture (détournement humoristique dans ce cas précis). Deux périphrases d'Eurydice sont également détectées automatiquement : « sa Donzelle » et « sa Dame », la première étant également un indice du détournement humoristique du mythe²¹.

c) Édition numérique : perspectives illimitées, pérennité incertaine

Encore à l'état de prototype, notre interface s'éloigne des habitudes de l'édition numérique : nous observons plutôt une standardisation généralisée des interfaces, grâce aux outils comme Omeka²² ou TEI Publisher²³. Ces solutions correspondent aux habitudes de lecture numérique tout en restant exploitables par le public habitué à l'imprimé. Elles augmentent aussi la pérennité informatique du projet : les changements technologiques sont gérés collectivement et le financement du développement est plus stable que dans le cadre d'un projet éditorial isolé.

Nous proposons plutôt des visualisations appuyées sur des technologies potentiellement moins stables, notamment en ce qui concerne les bibliothèques en charge de la modélisation des données à l'aide de graphes²⁴. Le développement « fait maison » nous expose également à plus de risque d'erreur d'encodage, d'affichage et d'adaptabilité de notre édition. Enfin, l'interface que nous gérons peut rapidement devenir obsolète du fait de mises à jour des langages informatiques, des navigateurs et des systèmes d'exploitation.

Ces préoccupations sont centrales dans le domaine des humanités numériques. Mais nous considérons qu'il doit être permis de ne pas toujours devoir choisir entre pérennité et innovation. Nous séparons donc clairement, dans ce travail, les préoccupations liées à la pérennité et à la réutilisabilité des données collectées et les expérimentations liées à la conception des nouveaux outils de la détection et de la visualisation des relations intertextuelles.

Nos données sont décrites à l'aide des métadonnées standard (Dublin Core, TEI) structurées à l'aide des formats ouverts et durables (texte, JSON, XML), et partagées dans des entrepôts dédiés. Nous concevons également, en concertation avec les experts de la science ouverte, le plan de gestion des données²⁵ que nous respectons tout au long de notre travail doctoral. Enfin, une vitrine d'exposition est conçue pour notre corpus à l'aide de la plateforme de publication Omeka.

²¹ Une autre périphrase d'Eurydice est présente : « donzelle morte à ce qu'on dit, par un hasard qui la mordit », qui n'a pas pu être détecté automatiquement. En revanche, « ce Menestrier de malheur » est correctement relié à Orphée, mais n'est pas signalé dans cet extrait, car le texte de Grandval ne mentionne pas le poète (si ce n'est que par le « je » lyrique).

²² Voir <https://omeka.org/>.

²³ Voir <https://teipublisher.com/index.html>.

²⁴ Il s'agit de Cytoscape.js (voir <https://js.cytoscape.org/>), qui est dépendant de l'évolution du langage JavaScript.

²⁵ Voir <https://doranum.fr/plan-gestion-donnees-dmp/>.

Alors que nos données sont facilement réutilisables telles quelles, pour nos développements, nous nous préoccupons surtout de la documentation facilitant leur reproductibilité, par exemple en cas de changements technologiques qui les rendraient obsolètes. Les outils de gestion de version de code, comme GitLab, facilitent la reprise d'un projet informatique pour la mise à jour et l'enrichissement du développement. La documentation détaillée accompagnée des captures d'écran dans un format pérenne (Markdown, PDF) permet, quant à elle, la conservation des traces de notre conception éditoriale originale. En revanche, de par leur aspect expérimental, nous acceptons, sans l'ignorer, la fragilité informatique de notre interface éditoriale et du développement qui enrichit les résultats des logiciels Tracer et TextPAIR.

5. Plan de la rédaction

Dans cette rédaction, nous exposons notre projet à l'aide de quatre parties, chacune divisée en trois à quatre chapitres.

La première partie se consacre à la présentation détaillée de notre corpus des réécritures.

- Nous proposons dans un premier temps une réflexion sur la lecture distante et la littérature mondiale, telles qu'elles ont été conceptualisées par Franco Moretti, en nous plaçant dans le contexte des archives et bibliothèques numériques.
- Nous délimitons ensuite les modalités de la collecte de notre corpus en partant d'une tentative de définition de la notion de réécriture et d'une exploration des archives numériques. Nous démontrons sa représentativité des points de vue chronologique, générique et fonctionnelle.
- Enfin, le dernier chapitre est consacré aux questions de la numérisation, de la description et de la diffusion du corpus dans l'esprit des humanités numériques. Nous y rappelons quelques bases théoriques concernant l'édition critique numérique, la philologie numérique et la construction du champ des humanités numériques. Nous détaillons aussi notre protocole de production des données, du format brut à la structuration au moyen du XML-TEI, et leur dépôt dans l'entrepôt numérique Nakala.

La deuxième partie interroge la notion d'intertextualité dans le contexte interdisciplinaire.

- Le premier chapitre rappelle les travaux littéraires concentrés autour de cette notion, des approches philosophiques de Kristeva et Barthes, par la pratique textuelle de Jenny et Riffaterre jusqu'aux concepts dits « périphériques » de Compagnon et Genette. Nous proposons également une brève démonstration de la manière dont ces théories se retrouvent dans l'observation pratique de notre corpus.

- Le deuxième chapitre propose un point de vue des sciences du langage afin de déterminer quels procédés linguistiques sont caractéristiques des relations intertextuelles. Pour ce faire, nous proposons une étude traductologique à partir de 21 traductions d'Ovide avant de rappeler les principales théories linguistiques liées à la notion de réutilisation textuelle et les classements existants du phénomène de la paraphrase. Cela nous amène à proposer une méthode de classement de ce que nous dégageons de nos observations empiriques. Nous réfléchissons alors tant à la complexité des procédés présents au sein du corpus qu'à la faisabilité de la reproduction automatique des analyses et classements empiriques des périphrases.
- Le troisième chapitre enfin s'intéresse aux trois méthodes et approches computationnelles pouvant être considérées comme prolégomènes de la détection automatique des réutilisations textuelles : la stylométrie, la détection et modélisation des topiques et l'alignement automatique des corpus parallèles. Pour chaque approche, nous proposons une expérimentation pratique afin de révéler tant son intérêt pour notre recherche que son insuffisance pour la détection automatique des relations intertextuelles.

Ces deux parties plutôt théoriques, mais qui présentent tout de même quelques démonstrations pratiques, nous fournissent les bases pour les expérimentations au cœur de notre recherche. Les deux parties suivantes sont consacrées alors à l'exposition des résultats, propositions et perspectives liés à nos objectifs principaux, à savoir permettre la détection des similitudes entre les œuvres, leur alignement et le repérage intertextuel à l'aide d'outils numériques existants et conçus par nos soins, et concevoir une édition numérique qui facilite l'analyse comparative d'un corpus des œuvres littéraires.

Pour ce faire, la troisième partie se concentre spécifiquement sur les outils dédiés à la recherche automatique des réemplois textuels, leur évaluation et la présentation critique des résultats initiaux.

- Nous présentons d'abord notre expérimentation initiale avec le logiciel Phoebus, le seul outil doté d'une interface graphique et capable de fournir des informations lexicales sur la nature des relations détectées, mais qui n'est plus fonctionnel. En partant de ce que Phoebus nous a permis d'observer sur un petit corpus des traductions d'Ovide, nous concevons une grille d'évaluation, composée de cinq axes principaux concernant les recherches existantes, les critères techniques, le format, les possibilités de paramétrage et la qualité des résultats.
- Nous confrontons ensuite ces critères à six logiciels, BLAST¹, Textreuse R, Tesserae, Passim, TextPAIR et Tracer, en utilisant le même corpus que celui soumis initialement

à Phoebus. Nous présentons d’abord les outils écartés puis, de manière comparative et plus détaillée, les deux gagnants : TextPAIR et Tracer.

- Nous effectuons ensuite la recherche avec TextPAIR et Tracer sur la totalité du corpus des réécritures, en le comparant à lui-même et au corpus des traductions (nous pouvons donc détecter des relations entre réécritures et entre réécritures et traductions, mais pas au sein des traductions). Nous proposons une méthode d’évaluation manuelle d’un échantillon des résultats, puis exposons les problèmes que celle-ci a permis d’observer.

Cette partie nous paraît constituer un réel apport à la communauté intéressée par l’intertextualité quantitative. Toutefois, pour faciliter la reproduction de notre expérimentation, une certaine technicité, plus importante sans doute que dans les autres parties, est nécessaire. Nous tâchons d’accompagner au maximum le lecteur non-spécialiste des méthodes et traitements informatiques et de coupler nos explications techniques à des démonstrations pratiques des résultats obtenus. Alors que dans les parties précédentes, l’axe disciplinaire de chaque chapitre est clairement identifiable, à partir de la troisième partie, il s’agit d’une réelle recherche interdisciplinaire, où la littérature, la linguistique et l’informatique sont sollicitées conjointement.

La quatrième partie enfin présente notre apport à la détection des relations intertextuelles au moyen d’outils informatiques.

- Le premier chapitre explique notre manière de post-traiter les résultats obtenus avec TextPAIR et Tracer : l’unification des formats, la détection des unités communes et le calcul du score de ressemblance nous permettent d’écarter les couples peu pertinents et de mettre en avant ceux qui sont les plus susceptibles de contenir des relations effectives. Nous démontrons l’efficacité de cette méthode à l’aide d’un exemple d’exploration des résultats enrichis et de l’analyse du réseau global des réécritures que nous avons pu effectuer grâce à notre traitement.
- Le deuxième chapitre se consacre spécifiquement au développement que nous avons conçu pour permettre l’annotation des entités nommées et leurs périphrases. La présentation des méthodes existantes de la détection automatique des entités nommées est suivie par l’explication des algorithmes et des lexiques que nous mettons en place. Nous prouvons ensuite l’efficacité de cette méthode en proposant une analyse focalisée sur la figure d’Eurydice.
- Le troisième chapitre présente le prototype de notre édition comparative, dont nous avons montré quelques exemples dans cette introduction. Le but du chapitre est non seulement d’expliquer les différentes interfaces et leurs fonctionnalités, mais aussi de montrer notre raisonnement et de défendre nos choix éditoriaux.

- Le dernier chapitre propose enfin quelques perspectives pour le futur du projet. Nous y présentons deux expérimentations que nous avons menées dans la perspective de l'amélioration du post-traitement des résultats, mais que nous n'avons pas pu ni su intégrer dans le développement : l'exploitation du corpus des traductions pour améliorer la qualité des résultats obtenus (tant initialement qu'après l'étape de post-traitement) et l'optimisation du traitement des synonymes grâce à l'exploitation des dictionnaires de la synonymie distinctive.

I. Présentation du corpus des réécritures du mythe
d'Orphée et Eurydice

La figure d'Orphée est l'objet de nombreuses études diachroniques. Ainsi, la présence de ce *topos* au Moyen Âge est analysée par John Block Friedman¹. Alors que Françoise Joukovsky² se concentre sur le XVI^e s. et que Léon Cellier³ étudie les représentations du XIX^e s., Jacqueline Bellas⁴ retrace l'évolution des représentations à travers les XIX^e et XX^e siècles. Enfin, la présence du mythe dans la littérature moderne et contemporaine est un sujet prisé. Nous recensons notamment les études d'Eva Kushner⁵ et de Pierre Brunel⁶, ainsi que les thèses de doctorat de Georgios Archimandritis⁷ et de Barunka Hauptmann⁸.

Les études synchroniques et comparatives sont également nombreuses : Arlette Bouloumié⁹ et Julie Dekens¹⁰ s'intéressent à l'évolution de la figure d'Eurydice dans la littérature contemporaine, Véronique Porra¹¹ observe la présence des Orphées noirs dans la littérature postcoloniale, Michel Décaudin¹² compare le mythe d'Orphée chez Apollinaire et chez Cocteau. D'autres chercheurs retracent la résurgence de la figure orphique chez un auteur en particulier, chez Pierre-Simon de Ballanche dans la thèse de doctorat de Romijn Helene Boom¹³, chez Victor Segalen dans l'article d'Eva Kushner¹⁴ ou encore chez Marcel Proust selon l'analyse d'Émilie Turmel¹⁵.

Concluons ce bref état des lieux en constatant la présence de travaux sur le mythe d'Orphée et Eurydice chez les chercheurs qui ont forgé les principales méthodes actuellement employées pour

¹ John Block Friedman, *Orphée au Moyen Âge*, trad. Jean-Michel Roessli, Fribourg, Cerf, coll. « Vestigia », 1999.

² Françoise Joukovsky, *Orphée et ses disciples dans la poésie française et néolatine du XVI^e siècle*, Droz, 1970.

³ Léon Cellier, « Le Romantisme et le mythe d'Orphée », *Cahiers de l'AIEF*, vol. 10, n° 1, 1958, p. 138-157, en ligne, <https://www.persee.fr/doc/caief_0571-5865_1958_num_10_1_2128>.

⁴ Jacqueline Bellas, « “Orphée” au XIX^e et au XX^e siècles : interférences littéraires et musicales », *Cahiers de l'AIEF*, vol. 22, n° 1, 1970, p. 229-246.

⁵ Eva Kushner, *Le mythe d'Orphée dans la littérature française contemporaine*, Paris, A.G. Nizet, 1961.

⁶ Pierre Brunel, « Orphée moderne », *Apollinaire entre deux mondes. Le contrepoint mythique dans Alcools. Mythocritique II*, Paris, Presses Universitaires de France, coll. « Écriture », 1997, p. 63-82.

⁷ Georgios Archimandritis, « Le mythe d'Orphée dans le théâtre et le cinéma du XX^e siècle », Presses universitaires du Septentrion, 2002.

⁸ Barunka Hauptmann, « “Son corps déchiqueté” : les genres à l'œuvre dans le mythe d'Orphée au XIX^e siècle », Thèse de doctorat, Université de Paris IV, 2013.

⁹ Arlette Bouloumié, « La Résurgence du mythe d'Eurydice et ses métamorphoses dans l'œuvre d'Anouilh, de Pascal Quignard, d'Henri Bosco, de Marguerite Yourcenar, de Michèle Sarde, et Jean Loup Trassard », *Loxias*, Université de Nice, 2004, en ligne, <<http://revel.unice.fr/loxias/index.html?id=1244>>.

¹⁰ Julie Dekens, « Rester aux Enfers : le bonheur paradoxal d'Eurydice », *TRANS- Revue de littérature générale et comparée*, n° 17, Presses Sorbonne Nouvelle, février 2014, en ligne, <<http://journals.openedition.org.ressources-electroniques.univ-lille.fr/trans/910>>.

¹¹ Véronique Porra, « Sur quelques Orphées noirs. Reproduction, adaptation et hybridation du mythe d'Orphée en contexte post(-)colonial », *Revue de littérature comparée*, vol. 344, n° 4, Klincksieck, 2012, p. 441-455, en ligne, <<https://www.cairn.info/revue-de-litterature-comparee-2012-4-page-441.htm>>.

¹² Michel Décaudin, « Deux aspects du mythe orphique au XX^e siècle : Apollinaire, Cocteau », *Cahiers de l'AIEF*, vol. 22, n° 1, 1970, p. 215-227.

¹³ Romijn Helene Boom, « Pierre-Simon Ballanche et l'orphisme : une tentative d'herméneutique », Thèse de doctorat, University of Cape Town, 1988.

¹⁴ Eva Kushner, « Orphée et l'orphisme chez Victor Segalen », *Cahiers de l'AIEF*, vol. 22, n° 1, 1970, p. 197-214, en ligne, <https://www.persee.fr/doc/caief_0571-5865_1970_num_22_1_960>.

¹⁵ Émilie Turmel, « Orphée et l'orphisme dans *À la recherche du temps perdu* de Marcel Proust », *Études françaises*, vol. 51, n° 1, Les Presses de l'Université de Montréal, 2015, p. 141-161, en ligne, <<https://www.erudit.org/fr/revues/etudfr/2015-v51-n1-etudfr01694/1028526ar/>>.

analyser les réécritures des mythes antiques. D'abord, Gilbert Durand illustre sa “mythanalyse” en proposant, entre autres, une segmentation du mythe en 17 épisodes et en observant la récurrence du *leitmotiv* de la nostalgie dans tous les récits qui reprennent l'histoire d'Orphée¹⁶. Ensuite, Brunel met en pratique sa mythocritique en proposant une analyse particulière de la pièce de théâtre *Orphée-Roi* de Segalen (1921)¹⁷. Enfin, Ute Heidmann fait la démonstration de sa comparaison différentielle et discursive en confrontant *Les Géorgiques* de Virgile avec *Orphée l'enchanteur*, un roman de jeunesse de Guy Jimenes (2004). Face à une telle richesse de recherches déjà menées sur le sujet, il est légitime de s'interroger quant à l'intérêt de proposer un nouveau travail sur les réécritures de ce mythe.

Cette question mérite une considération approfondie. Des éléments de réponses peuvent déjà être fournis par rapport à nos choix d'établissement du corpus. Dans l'intégralité des travaux cités, le nombre d'œuvres analysées reste très restreint et aucune étude n'a été proposée jusqu'alors avec des méthodes computationnelles. Or, nous rassemblons plus de 50 œuvres en langue française qui réécrivent le mythe d'Orphée et Eurydice. Publiées depuis le XVI^e s., elles se caractérisent par leur diversité, tant formelle, que générique et fonctionnelle. Notre objectif n'est pas de faire *table rase* du passé, mais de faire communiquer cette nouvelle recherche avec tout ce qui a été observé auparavant. Pour ce faire, il est indispensable d'observer la manière dont des domaines à première vue très différents, comme la littérature et l'informatique, ou au contraire, moins proches que l'on pourrait le croire, comme la littérature et la linguistique, peuvent devenir complémentaires au sein d'une démarche qui s'efforce d'exploiter au mieux les spécificités de chaque discipline, au lieu de lisser leurs différences. Par ailleurs, le nombre de travaux existants sur le sujet nous permettra, d'une part, de confronter nos résultats à ceux des recherches antérieures et, d'autre part, de bénéficier de toute la richesse des théories, méthodes et stratégies de la recherche littéraire “traditionnelle”, mise en pratique par les plus grands experts du domaine.

Dans cette partie, les réécritures recueillies sont présentées en mettant en exergue les enjeux de la collecte et de l'élaboration d'un corpus littéraire appuyés sur des méthodes linguistiques et opérés dans un environnement numérique, régi par des phénomènes particuliers. Dans un premier temps, nous rappelons la théorie de la lecture distante de Franco Moretti qui pose les bases de l'analyse littéraire appuyée sur des outils informatiques. Afin d'illustrer quelques difficultés épistémologiques de l'application de cette approche dans l'environnement numérique, nous exposons les résultats de trois cas d'exploitation concrète d'outils informatiques destinés, entre autres, aux chercheurs littéraires. Ensuite, à partir de différentes conceptualisations par la critique littéraire du phénomène de

¹⁶ Gilbert Durand, « Les nostalgies d'Orphée. Petite leçon de mythanalyse », *Religiologiques*, n° 15, 1997, en ligne, <<https://www.religiologiques.uqam.ca/no15/durand.html>>.

¹⁷ Pierre Brunel, « *Orphée-roi* de Victor Segalen, ou le miracle de la lyre », *Mythocritique : théorie et parcours*, Paris, Presses Universitaires de France, coll. « Écriture », 1992, p. 190-202.

réécriture, nous cherchons à établir des critères de recherche objectivables nous permettant de trouver toutes les réécritures françaises du mythe d'Orphée et Eurydice disponibles au sein d'archives numériques, et notamment celles de la Bibliothèque Nationale de France, Gallica. Nous décrivons notre démarche de manière détaillée pour montrer l'importance de garder un équilibre entre la recherche quantitative, permettant d'aboutir à une exhaustivité relative, et la recherche qualitative, garante de la représentativité herméneutique. Nous proposons également une réflexion plus approfondie sur la notion de représentativité au sein d'un corpus comme le nôtre et exposons les difficultés de catégorisation et de description formelle des œuvres qui le composent. En effet, le choix d'étiquettes conformes tout autant aux usages de la linguistique de corpus qu'aux postulats des théories littéraires, qui montrent que la plupart des phénomènes littéraires échappent à une tentative de classement quelconque, n'est pas aisé. Enfin, nous expliquons la manière dont les œuvres qui composent notre corpus sont préparées, non seulement pour l'exploitation avec des outils computationnels au sein de notre projet, mais aussi pour leur réutilisation par nos pairs.

Comme Alexandre Gefen¹⁸, nous considérons que, que ce soit au sein de la critique littéraire en particulier ou des sciences humaines et sociales en général, « l'extraction de connaissances, le repérage et l'agrégation des faits culturels, la visualisation et la navigation à l'intérieur de corpus textuels plus ou moins massifs sont également des instruments nouveaux de connaissance et de perception » (§ 1). En effet, notre démarche de collecte de corpus peut être appliquée à d'autres projets qui visent le regroupement de textes littéraires de nature hétérogène. De même, notre corpus peut être enrichi par les travaux de nos pairs, voire réutilisé pour des objectifs qui diffèrent des nôtres. Toutefois, pour que tout cela soit possible, il est indispensable de documenter notre démarche de manière détaillée et sans omettre les différents écueils auxquels nous étions confrontée. C'est ce que nous nous attachons à faire tout au long de cette partie.

¹⁸ Alexandre Gefen, « Les enjeux épistémologiques des humanités numériques », *Socio. La nouvelle revue des sciences sociales*, n° 4, Éditions de la maison des sciences de l'homme, avril 2015, p. 61-74, en ligne, <<https://journals.openedition.org/socio/1296>>.

1. Lecture distante et rêve de littérature mondiale

Dans « Conjectures on World Literature »¹⁹, Franco Moretti propose le *distant reading* (« lecture distante »²⁰) comme la continuité – ou la réalisation – de la *Weltliteratur*. Cette belle idée empruntée à Goethe vise à étudier la littérature dans sa globalité, en dépassant les diverses frontières, qu'elles soient géographiques, culturelles ou linguistiques. Un regard englobant porte une promesse de recherche qui n'est pas restreinte par les canons littéraires et qui se libère des influences des cultures qui s'imposent comme dominantes depuis des siècles. Toutefois, selon Moretti, cela nécessite de prendre une certaine distance vis-à-vis de son objet de recherche. La démarche traditionnelle de l'analyse littéraire, définie, par opposition, comme *close reading* (« lecture rapprochée », « microlecture ») n'en serait jamais capable, car le choix extrêmement restreint des œuvres qu'elle impose implique forcément un jugement qualitatif des textes sélectionnés. Si l'on veut véritablement étudier la littérature mondiale, la microlecture ne s'y prête guère :

[I]l s'agit d'un exercice de nature théologique – un traitement très solennel de très peu de textes pris très au sérieux – alors que ce dont nous avons besoin, c'est un petit pacte avec le diable : nous savons comment lire des textes, apprenons maintenant comment nous pourrions *ne pas* les lire²¹.

Ce « petit pacte avec le diable » consiste à considérer chaque texte comme partie indissociable d'un ensemble plus grand – dans l'idéal des milliers de textes publiés à la même époque partout dans le monde. La proposition est à l'opposé de l'approche généralement inductive et exploratoire de la lecture rapprochée, qui consiste à observer en détail quelques textes pour en tirer des conclusions globalisantes. La lecture distante, au contraire, est avant tout quantitative et porte un regard objectif sur un large ensemble de textes, visant à confirmer ou à réfuter les hypothèses formulées au préalable.

L'objectivité du regard implique également le traitement égalitaire de l'ensemble des textes qui composent de larges corpus littéraires, sans se limiter au canon. En effet, selon Moretti chaque

¹⁹ Franco Moretti, « Conjectures on World Literature », *New Left Review*, n° 1, février 2000, p. 54-68. Dans le corps du texte, nous citons la traduction française de l'article, Franco Moretti, « Hypothèses sur la littérature mondiale », *Études de lettres*, n° 258 (2), trad. Raphaël Micheli, 2001, p. 9-24, en ligne, <<https://enseignement.typepad.fr/files/moretti-hypotheses.pdf>>. Le texte original est fourni en note de bas de page.

²⁰ Plusieurs traductions du terme ont été proposées dans les travaux critiques, comme « lecture distante », « lecture à distance », « distance aux textes » ou encore « macro-critique ». C'est toutefois le terme anglais, *distant reading*, qui semble le plus répandu dans l'usage francophone, ou du moins le plus immédiatement associé aux propos de Moretti (voir par exemple *Lire de près, de loin. Close vs distant reading*, Paris, Classiques Garnier, 2014.). Pourtant, dans la notice introduisant une nouvelle publication de l'article « *Conjectures on World Literature* », le chercheur semble regretter le choix du terme, qu'il qualifie de « formule funeste », son choix initial étant « lecture sérielle ». Voir Franco Moretti, *Distant reading*, London ; New York, Verso, 2013, p. 44.

²¹ « At bottom, it's [close reading] a theological exercise – very solemn treatment of very few texts taken very seriously – whereas what we really need is a little pact with the devil: we know how to read texts, now let's learn how not to read them », Franco Moretti, « Conjectures on World Literature », *loc. cit.*, p. 57. Pour la traduction française, voir Franco Moretti, « Hypothèses sur la littérature mondiale », *loc. cit.*, p. 3.

texte doit être mis en parallèle avec les littératures étrangères qui lui sont contemporaines et avec la totalité de la production éditoriale de son temps, incluant les œuvres oubliées ou dépréciées par la critique. Dans « The Slaughterhouse of Literature »²², le chercheur observe qu'uniquement 0,5 % de la production éditoriale est réellement étudiée (d'où la métaphore de *slaughterhouse*, « l'abattoir » de la littérature). L'approche lansonienne²³ de l'histoire littéraire rend ce constat d'autant plus vrai dans le contexte français. En principe, l'établissement du canon littéraire sert à attirer l'attention sur les œuvres de la plus haute qualité. Cependant, la restriction de la recherche à ces textes seuls mène, entre autres, à ignorer leur contexte de production et à la limiter à la prise en compte de la totalité de l'univers intellectuel ayant influencé la création d'un chef-d'œuvre canonique.

Moretti vise également à encourager la collaboration internationale des chercheurs littéraires. L'extension du corpus aux œuvres de différents horizons culturels invite à comparer ses propres analyses et résultats avec ceux réalisés par d'autres. L'internationalité de l'approche va de pair avec son interdisciplinarité. En effet, la proposition d'adapter des modèles et usages des autres sciences – des sciences humaines, comme l'histoire ou l'archéologie, et des sciences dures, comme les statistiques et les mathématiques – à l'analyse comparatiste est une véritable révolution théorique du début du XXI^e s. Et Moretti va encore plus loin, car, selon lui, la place du texte n'est plus centrale dans les études littéraires :

[L]a distance, je le répète, est une *condition de savoir*. Elle permet de se concentrer sur des unités qui sont bien plus petites ou bien plus grandes que des textes : des procédés, des thèmes, des tropes – ou des genres et des systèmes. Et si le texte devait se dissoudre dans cette tension nouvelle entre l'échelle infime des composantes stylistiques et l'échelle globale des agencements discursifs et des transferts culturels, il s'agirait alors de l'un de ces cas où l'on peut soutenir avec raison, pour reprendre le fameux « *less is more* » de Mies van der Rohe, que ce qui semble une perte est en réalité un gain.²⁴

Les propos de Moretti peuvent faire peur : si nous ne travaillons plus sur le texte, quel est notre objet ? Quel est notre rôle en tant que chercheurs et chercheuses littéraires ?²⁵ Il ne s'agit pas seulement de proposer une nouvelle méthode, mais aussi de questionner l'ancienne et de pointer du

²² Franco Moretti, « The Slaughterhouse of Literature », *MLQ: Modern Language Quarterly*, vol. 61, n° 1, Duke University Press, 2000, p. 207-227, en ligne, <<https://muse.jhu.edu/article/22852>>.

²³ Il s'agit d'une approche normative de l'histoire littéraire, telle qu'elle est présentée dans l'ouvrage *Histoire de la Littérature française* de Gustave Lanson (1895).

²⁴ « [D]istance, let me repeat it, is a condition of knowledge: it allows you to focus on units that are much smaller or much larger than text: devices, themes, or tropes—or genres and systems. And id, between the very small and the very large, the text itself disappears, well, it is one of those cases when one can justifiably say, *Less is more* », Franco Moretti, « Conjectures on World Literature », *loc. cit.*, p. 57. Pour la traduction française, voir Franco Moretti, « Hypothèses sur la littérature mondiale », *loc. cit.*, p. 3.

²⁵ Concernant le débat autour de la littérature mondiale en général et de la lecture distante en particulier, voir par exemple : Françoise Lavocat, « Le comparatisme comme herméneutique de la défamiliarisation », *Vox-Poetica*, avril 2012, en ligne, <<http://www.vox-poetica.org/t/articles/lavocat2012.html>> ; Emily S. Apter, *Against world literature: on the politics of untranslatability*, London ; New York, Verso, 2013 ; Haun Saussy, « La lecture, pratique dissidente », *Le Comparatisme comme approche critique*, vol. 3. Objets, méthodes et pratiques comparatistes, Paris, Classiques Garnier, coll. « Rencontres », 2017, p. 59-73, en ligne, <<https://classiques-garnier.com/le-comparatisme-comme-approche-critique->

doigt son aspect paradoxal : « plus on est attentif au fonctionnement d'un texte, à ses « détails » [...] plus, en somme, on a le *nez collé sur l'objet*, moins cet objet devient *accessible* »²⁶.

Quel est l'objectif du *distant reading* ? Il s'agit avant tout, du moins initialement, d'identifier des patrons et modèles permettant de proposer une vue différente de l'histoire de la littérature, plus proche de l'idée universalisante de *Weltliteratur*. Les approches quantitatives sont combinées avec différentes vues analytiques, les objets d'étude étant comparés selon des critères diachroniques, synchroniques ou cycliques²⁷. Pour en rendre compte, Moretti préconise de recourir aux différents types de visualisations, comme des graphes ou des arborescences, voire des cartes géographiques qui permettent de superposer un espace fictionnel à l'espace géographique réel.

Même si la théorie de la lecture distante de Moretti n'est pas nécessairement liée à un environnement de travail numérique, elle en est difficilement dissociable. La disponibilité des corpus numérisés, notamment dans l'espace anglo-saxon, invite à les explorer. De nombreux outils d'analyse sont également mis à disposition, dont une grande partie ne nécessite pas de compétences informatiques. Toutefois, un cas concret d'interrogation de quelques-uns de ces outils montrera certains problèmes auxquels se heurte une tentative d'application des principes de la lecture distante à une question littéraire particulière dans un environnement numérique. En effet, avant de procéder au recensement aussi complet que possible des réécritures du mythe d'Orphée et Eurydice, il est utile d'interroger les possibilités offertes par l'environnement numérique. Pour ce faire, observons concrètement la présence des mentions du nom d'Orphée au sein de trois collections des textes qui contiennent des ouvrages numérisés, Google Books²⁸, Frantext²⁹ et Gallica³⁰.

a) Google Books et Google Ngram Viewer : hégémonie anglosaxonne

Le géant informatique Google est un des premiers à initier, dès 2004, la numérisation et l'indexation des livres à grande échelle. Le corpus ainsi recueilli, appelé Google Books, regroupe tant des ouvrages du domaine public, accessibles alors à la lecture et au téléchargement, que ceux encore sous droits. Dans ce dernier cas, le texte complet n'est pas disponible, mais l'utilisateur dispose d'une description bibliographique complète, il peut accéder à un nombre limité de pages, faire une

comparative-literature-as-a-critical-approach-tome-3-objets-methodes-et-pratiques-comparatistes-objects-methods-practices-la-lecture-pratique-dissidente-en.html>.

²⁶ *Ouverture à trois voix*, Paris, Classiques Garnier, 2014, en ligne, <<https://classiques-garnier.com/lire-de-pres-de-loin-close-vs-distant-reading-ouverture-a-trois-voix.html>>, p. 8.

²⁷ Dans ce contexte, un "cycle" est un cadre temporel défini ainsi par Moretti : « les cycles constituent des *structures temporaires à l'intérieur du flux de l'histoire*. Il s'agit, après tout, de la logique sous-jacente à la tripartition de Braudel : le temps bref est du pur flux sans structure, la longue durée est la pure structure sans flux, et les cycles sont la zone frontalière – instable – entre les deux », Franco Moretti, « Graphes », *Graphes, cartes et arbres. Modèles abstraits pour une autre histoire de la littérature*, trad. Étienne Dobenesque, Paris, Les Prairies ordinaires, 2008, p. 47.

²⁸ <https://books.google.fr/>.

²⁹ <https://www.frantext.fr/>.

³⁰ <https://gallica.bnf.fr/>.

recherche de mots dans la totalité du texte et accéder au contexte relativement restreint de l'occurrence. L'idée est très belle : Google Books promet de devenir une bibliothèque d'Alexandrie qui ne craint pas le feu, de rendre mondialement accessibles les livres épuisés ou rares et d'élargir l'accès aux publications plus récentes. Il est aussi destiné à devenir un bon lieu de coprésence de différents horizons culturels. L'idée de Moretti semble se réaliser dans ce projet. En outre, Google ne se limite pas à numériser le patrimoine culturel. L'archive est accompagnée d'outils d'interrogation et de visualisation, comme Google Books Ngram Viewer³¹. Ce logiciel génère des diagrammes permettant d'analyser l'évolution diachronique des occurrences d'une unité lexicale ou d'une combinaison d'unités (comme une phrase) dans un corpus de l'aire linguistique choisie et sur une période définie.

Google Ngram Viewer permet par exemple de rechercher les mentions du nom d'Orphée dans les ouvrages anglais, français, allemands, espagnols et italiens parus entre 1500 et 2020 et numérisés par Google Books. Une telle requête, portant sur des millions d'œuvres, permet d'interroger la pertinence de cet outil pour analyser la fréquence des mentions du protagoniste principal du mythe qui nous intéresse dans la production éditoriale offerte dans ces langues. Bien évidemment, le nom du poète est traduit dans la langue du corpus : *Orfeo* pour l'italien et l'espagnol, *Orpheus* pour allemand et l'anglais (seuls des textes de l'aire britannique sont pris en compte). La visualisation en Figure 5 présente les résultats obtenus. Une couleur spécifique est assignée à chaque trait désignant la fréquence du mot dans chaque corpus (comme le rouge pour le français). L'axe horizontal correspond aux années (avec des graduations chaque demi-siècle) et l'axe vertical présente la fréquence du mot recherché par rapport à la totalité des mots du corpus à chaque instant.

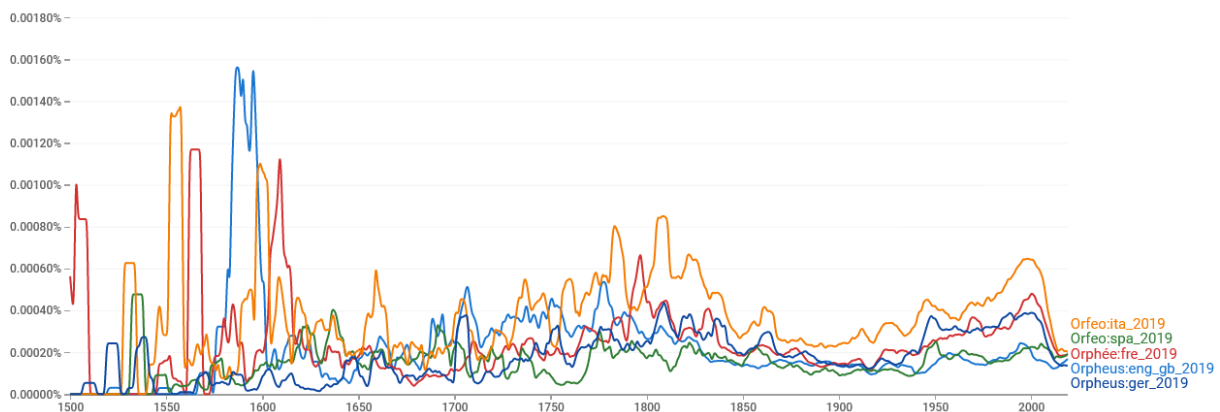


Figure 5. Occurrences du mot Orphée dans 5 aires linguistiques. Google Books Ngram Viewer³²

³¹ Voir <https://books.google.com/ngrams>. Pour l'exemple de l'exploration de cet outil dans la recherche littéraire, voir Jean-Baptiste Michel et al., « Quantitative Analysis of Culture Using Millions of Digitized Books », *Science*, vol. 331, n° 6014, American Association for the Advancement of Science, janvier 2011, p. 176-182, en ligne, <<https://www.science.org/doi/10.1126/science.1199644>>.

³² Ce diagramme peut être reproduit avec la requête « Orpheus:eng_gb_2019, Orphée:fre_2019, Orfeo:ita_2019, Orfeo:spa_2019, Orpheus:ger_2019. »

La lecture du diagramme dans sa totalité permet d'observer des tendances comparables, quoique d'intensité variable, à partir de la deuxième moitié du XVII^e s. Le mot *Orphée* semble généralement plus régulièrement évoqué dans l'aire italienne et reste plutôt faible dans le corpus espagnol. Une hausse relative des fréquences peut être observée dans les aires italienne, française et allemande dans le courant de la première moitié du XX^e s., mais elle chute autour de l'an 2000. Les corpus britannique et espagnol sont moins fluctuants et restent bas depuis plus de 150 ans, le premier présentant toutefois une légère hausse ces dernières années, dont l'importance est difficile à évaluer pour le moment.

Des tendances plus particulières sont également à noter. Observons concrètement, en Figure 6, les ondulations presque synchronisées des corpus français, italien et anglais entre 1550 et 1650. Aucune occurrence d'*Orphée* n'est enregistrée en français en 1559 et en 1569. En revanche, la fréquence du nom est significativement en hausse entre ces deux années. Entre 1563 et 1565, elle atteint le score le plus élevé pour la totalité de la production française prise en compte, puis rechute drastiquement jusqu'à revenir à aucune mention en 1569. Cette courbe semble suivre de près la vague observée dans l'aire italienne 10 ans plus tôt et dont le pic est atteint en 1553, puis en 1557.

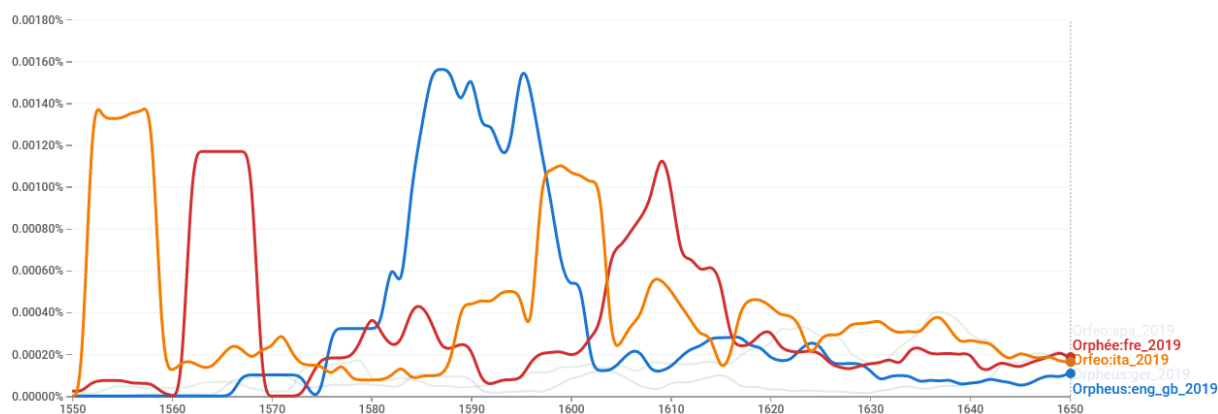


Figure 6. Focus sur la période 1550-1650 et les corpus français, italien et anglais. Google Books Ngram Viewer

La fluctuation comparable apparaît presque 30 ans plus tard dans l'aire anglophone. Une hausse des occurrences est à noter à partir de 1580, elle atteint son apogée en 1595. La même année, les fréquences grimpent dans le corpus italien. En trois ans, elles ont plus que doublé (0,000498% en 1595 pour 0,00107 % en 1598). Pour finir, la hausse des occurrences pour les corpus français suit la même tendance à partir de 1600 pour atteindre son apogée en 1609.

Nous hasarderons-nous à soumettre des interprétations de ces observations ? Quelques thèses paraissent plausibles. La réappropriation des récits mythologiques dans l'aire européenne au XX^e s. – que ce soit par le théâtre français de Cocteau et d'Anouilh, ou par la poésie allemande de Rilke et italienne de Campana – est-elle moins importante dans les aires anglophone et hispanique, du moins pour Orphée ? Les récits des dieux et héros ont-ils davantage d'importance dans la culture italienne, qui hérite plus directement de la mythologie romaine ? La fréquence particulièrement élevée au début du XV^e s. indique-t-elle le rôle précurseur de la littérature française dans la (re)découverte de la figure orphique ? Enfin, les vagues synchroniques observées entre 1550 et 1650, reflètent-elles la circulation internationale de quelques textes “déclencheurs”, par exemple grâce à la parution de leurs traductions dans d'autres aires linguistiques ou par le biais des représentations des pièces de théâtre ou d'opéra à l'étranger ?

Ces questions mériteraient une investigation approfondie. Toutefois, plusieurs problèmes épistémologiques, liés à la particularité du corpus, sont à observer ici. Tout d'abord, l'hétérogénéité des contenus rend impossible l'analyse des résultats dans le contexte littéraire uniquement, car chaque sous-corpus linguistique regroupe indifféremment des ouvrages de natures très diverses : articles scientifiques, traités, catalogues... Si les œuvres littéraires sont majoritaires dans les périodes antérieures au XVIII^e s., elles sont de moins en moins représentées dans les siècles postérieurs. En outre, la place des traductions parmi les résultats est importante. Parmi les dix premiers ouvrages trouvés dans chaque corpus correspondant à la période entre 1950 et 2019, les traductions de *Die Sonette an Orpheus* de Rainer-Maria Rilke (1922) sont présentes dans le corpus britannique (*Sonnets di Orpheus*, traduits par Carlyle F. MacIntyre, 1960), italien (*Sonetti a Orfeo*, traduits par Giacomo Cacciapaglia, 1995) et espagnol (*Sonetos a Orfeo*, traduits par Otto Dörr, 2002). Pour le corpus italien, quatre ouvrages sur dix sont des traductions. Enfin, parmi les dix premiers résultats allemands, une seule œuvre est littéraire (le thriller *Die Orpheus-Prophezeiung* d'Oliver Buslau), les autres étant des ouvrages scientifiques. Par conséquent, il est impossible de questionner des corpus aussi hétérogènes uniquement du point de vue littéraire.

Ensuite, la possibilité de comparer des aires linguistiques différentes est enthousiasmante, mais les disproportions dans la taille des corpus demandent une vigilance accrue. Certes, la question du canon ne se pose plus. En revanche, celle des cultures dominantes est davantage présente, même si, dans le contexte numérique, elle est plus linguistique que territoriale.

D'ailleurs, le débat concernant l'impérialisme linguistique de Google n'est pas nouveau. En 2005, Jean-Noël Jeanneney, alors président de la Bibliothèque nationale de France, publie une tribune dans *Le Monde* intitulée « Quand Google défie l'Europe : plaidoyer pour un sursaut ». Le livre éponyme qui paraît la même année chez l'éditeur Mille et une nuits connaît trois éditions

successives en France (dont la dernière en 2010) et sa traduction anglaise est publiée en 2007 dans les presses de l'université de Chicago. L'intérêt pour le débat est indéniable et l'engagement de Jeanneney influence indubitablement la direction du développement des archives numériques en France et en Europe³³. Parmi de nombreux dangers possibles perçus à l'époque par l'historien, la question de l'emprise de l'anglais (américain) est centrale :

Danger que, du côté des ouvrages patrimoniaux tombés dans le domaine public, la liste des priorités favorise la pesée de la culture anglo-saxonne. Danger que, du côté de ceux qui sont encore sous droits [...], le poids des éditeurs américains soit écrasant. Danger que, pour les revues et les livres diffusant la recherche en mouvement, la domination des travaux d'outre-Atlantique devienne plus lourde encore que ce n'est le cas aujourd'hui.

L'enjeu est langagier, évidemment, et l'on voit bien comment risque de s'accélérer l'emprise de l'anglais (sous sa forme américaine) aux dépens de tous les autres parlers de l'Europe.³⁴

La taille exacte de différents corpus de Google Books est difficile à trouver. Toutefois, il convient de noter que seul le corpus anglais peut être partitionné pour l'exploration avec Ngram Viewer. Parmi 12 sous-corpus disponibles, 5 sont anglais : l'anglais américain, l'anglais britannique, la compilation des deux, et deux sous-ensembles particuliers, *English Fiction*, un corpus des ouvrages déclarés par les éditeurs comme livres de fiction, et *English One Million*, ou *Google Million*, qui retient 6 000 livres maximum par l'an entre 1500 et 2008³⁵. S'il est possible de restreindre la recherche au corpus britannique, les résultats ne concernent pas pour l'autant uniquement l'Europe. De même, le corpus hispanophone inclut des publications d'Amérique latine, le corpus français, celles du Québec. Par conséquent, il est erroné de les comparer et de les analyser en tant que productions européennes, d'autant plus que les aires linguistiques prises en compte sont uniquement celles de l'Europe occidentale.

³³ Voir notamment <https://gallica.bnf.fr/edit/und/a-propos> : « En janvier 2005 dans *Le Monde*, Jean-Noël Jeanneney lance le projet d'une bibliothèque numérique européenne pour répondre à Google Books, qui inaugure une accélération dans le développement de Gallica et un changement d'échelle et de rythme de numérisation ». Le projet Europeana est lancé dans la continuité de ce mouvement, voir <https://www.europeana.eu/fr/about-us>.

³⁴ Jean Noël Jeanneney, *Quand Google défie l'Europe : plaidoyer pour un sursaut*, 3^e éd., Paris, Fayard/Mille et une nuits, coll. « Essai », 2010, p. 13.

³⁵ « All are in English with dates ranging from 1500 to 2008. No more than about 6 000 books were chosen from any one year, which means that all of the scanned books from early years are present, and books from later years are randomly sampled. The random samplings reflect the subject distributions for the year (so there are more computer books in 2000 than 1980). », [Tous sont en anglais, avec des dates de parution allant de 1500 à 2008. 6 000 livres maximum ont été choisis pour chaque année, ce qui implique que, tous les livres anciens [numérisés par Google Books] sont présents, et que les livres récents sont échantillonnés de manière aléatoire. Cet échantillonnage montre la répartition des sujets récurrents pour chaque année (il y a donc plus de livres sur les ordinateurs en 2000 qu'en 1980)] (nous traduisons). Voir <https://books.google.com/ngrams/info#>, § Corpora.

b) Frantext : base de données textuelles françaises

Pour comprendre cette disproportion quantitative, il faut revenir au début des années soixante, marquées notamment par l'apparition de la littérature et la linguistique computationnelles (*Literary and Linguistic Computing*)³⁶. Ce courant de recherche se caractérise par la création des grands corpus textuels, la génération d'index et le développement d'outils de traitement automatique des langues (TAL) et de linguistique du corpus (pour produire des concordances, extraire des cooccurrences lexicales régulières et effectuer des traductions automatiques ou assistées par ordinateur). Tel qu'il est compris et défini par la linguistique, le corpus regroupe un ensemble d'énoncés langagiers (données primaires), éventuellement enrichis d'annotations et de métadonnées, voire modélisés selon une politique de structuration prédéfinie³⁷. Son rôle est différent de celui d'une archive numérique. Quelle que soit la nature des énoncés regroupés (textes littéraires, transcription des entretiens oraux...), le corpus linguistique est avant tout dédié à l'étude (diachronique ou synchronique) d'une langue donnée, voire des multiples langues afin de comparer leur fonctionnement. Toutefois, dès les débuts de la linguistique du corpus, les ressources ne se développent pas de manière égalitaire et la majorité des premiers grands corpus textuels est en anglais³⁸. Actuellement, le clivage est moins

³⁶ Ou « informatique littéraire et linguistique ». Ce courant méthodologique est généralement considéré comme la première période du développement du champ des humanités numériques. Pour son histoire complète, voir notamment : Lou Burnard, « Du *literary and linguistic computing* aux *digital humanities* : retour sur 40 ans de relations entre sciences humaines et informatique », dans Pierre Mounier (dir.), *Read/Write Book 2 : Une introduction aux humanités numériques*, Marseille, OpenEdition Press, coll. « Read/Write Book », 2012, p. 45-58 ; Michaël E. Sinatra et Marcello Vitali-Rosati, « Chapitre 3. Histoire des humanités numériques », dans Michael E. Sinatra (dir.), *Pratiques de l'édition numérique*, Montréal, Presses de l'Université de Montréal, coll. « Parcours numérique », 2014, p. 49-60 ; Pierre Mounier, *Les humanités numériques Une histoire critique*, Paris, Éditions de la Maison des sciences de l'homme, 2018.

³⁷ « A corpus is any collection of speech recordings which is accessible in computer readable form and which comes with annotation and documentation sufficient to allow re-use of the data-in-house, or by people in other organisations », [Un corpus est une collection d'énoncés langagières disponible dans un format qui peut être interprété par un ordinateur. Ces énoncés sont annotés et suffisamment documentés pour permettre la réutilisation des données en interne ou au sein des projets externes.] (Nous traduisons). Dafydd Gibbon, Roger Moore et Richard Winski (dir.), *Spoken language system and corpus design*, Berlin, New York, De Gruyter, coll. « Handbook of standards and resources for spoken language systems », 1998, p. 93.

³⁸ Notamment :

- *Brown University Standard Corpus of Present-Day American English* (1963, voir W. Nelson Francis, « A Standard Corpus of Edited Present-Day American English », *College English*, vol. 26, n° 4, National Council of Teachers of English, 1965, p. 267-273, en ligne, <<https://www.jstor.org/stable/373638>>.) ;
- *Susanne* (1994, voir Geoffrey Sampson, *English for the computer: the SUSANNE corpus and analytic scheme*, Oxford : New York, Clarendon Press ; Oxford University Press, 1995.) ;
- *London-Lund* (1982, voir *The London-Lund corpus of spoken English: Description and research*, vol. 82, Lund University Press, coll. « Lund Studies in English », 1990.) ;
- *Helsinki* (1991, voir Merja Kytö (dir.), *Manual to the diachronic part of the Helsinki corpus of English texts: coding conventions and lists of source texts*, 3. ed, Helsinki, Dep. of Engl. Univ. of Helsinki, 1996.) ;
- *Archer* (1993, voir Douglas Biber, Edward Finegan et Dwight Atkinson, « ARCHER and its challenges: Compiling and Exploring a Representative Corpus of Historical English Registers », dans Udo Fries, Peter Schneider et Gunnel Tottie (dir.), *Creating and using English language corpora. Papers from the 14th International Conference on English Language Research on Computerized Corpora, Zurich 1993*, Amsterdam, Rodopi, 1994, p. 1-13, en ligne, <[doi:10.1515/9783110867275.213](https://doi.org/10.1515/9783110867275.213)>.) ;
- *British National Corpus* (1991, BNC Consortium, *British National Corpus, XML edition*, Oxford Text Archive, 2007.)

important par exemple pour le français, l'allemand ou l'espagnol, dont les utilisateurs sont nombreux et où les structures nationales soutiennent la collecte et l'archivage du patrimoine linguistique. Cependant, ce n'est pas le cas des langues dites "mineures" (ou "peu dotées"), régionales ou dialectales, employées par des minorités et, à ce titre, marginalisées en faveur des langues standards.

Dans l'aire francophone, Frantext est un projet pionnier, quoique restreint au français normé. Cette base de textes, majoritairement littéraires³⁹ et parus depuis le XVI^e s., peut d'abord être considérée comme un corpus destiné à des études linguistiques. Elle est initialement développée dans les années 1960 à l'Institut National de la Langue Française pour documenter les rédacteurs du *Trésor de la Langue Française*. Elle est ensuite enrichie pour regrouper plus de 160 millions de mots. Simultanément, Frantext joue le rôle d'archive numérique destinée à l'exploration par des chercheurs littéraires. L'interrogation est facilitée grâce à une interface Web. Le corpus peut être échantillonné selon les genres, les auteurs ou les périodes, et les résultats des recherches peuvent être sauvegardés en quelques clics⁴⁰. En outre, l'outil est prévu pour accompagner de potentiels utilisateurs non initiés à l'exploration de corpus. La mise en garde concernant des disproportions en son sein est affichée avant d'accéder aux outils d'exploration.

Le diagramme en bâtons (en Figure 7) permet de rendre compte de la plus grande représentativité des textes publiés au XX^e s. pour prévenir d'éventuels biais d'analyse. Outre la transparence de la composition des corpus explorés, Frantext donne une plus grande exploitabilité des résultats que Google Ngram Viewer, en particulier en permettant de les télécharger. Par exemple, l'interrogation de l'intégralité du corpus *Frantext* permet de trouver 1 234 mentions du nom d'*Orphée*⁴¹, pour la plupart dans sa forme canonique (1 038), mais d'autres graphies ont également été relevés (*Orphee*, 83 occ., *Orpheus*, 64 occ.). La comparaison des années de parution des œuvres et des occurrences trouvées révèle, sans grande surprise, que la graphie normalisée n'apparaît qu'en 1552, et coexiste avec d'autres formes jusqu'à 1650. Néanmoins, des occurrences de la forme *Orpheus* sont tout de même recensées en 1832 (2 occ. dans les dialogues de *Notre-Dame de Paris* d'Hugo), 1896 (3 occ. dans les poèmes de Verlaine) et en 1936 (1 occ. qui fait partie du titre *Sonnette An Orpheus* non traduit au sein du texte en français).

³⁹ Voir le diagramme « Genres textuels » dans l'onglet « Corpus » (<https://www-frantext-fr.ressources-electroniques.univ-lille.fr/repository/frantext/corpora/stats>).

⁴⁰ Toutefois, il convient de noter que seulement une version de démonstration, restreinte à 40 textes ; est librement accessible. Grâce à l'abonnement institutionnel qui nous octroie l'accès au corpus intégral, nous décrivons la recherche menée sur 5 555 textes au total.

⁴¹ Cette recherche peut être reproduite en choisissant le mode de recherche « Assisté » et en renseignant « ORPHÉE|ORPHEUS » dans le champ « Lemme ». Les cases « Sensible à la casse » et « Sensible aux diacritiques » ont été décochées. Le tableur Excel avec la totalité des résultats est disponible en annexes de cette partie, [I-2 Resultats-fouille.xlsx](#).

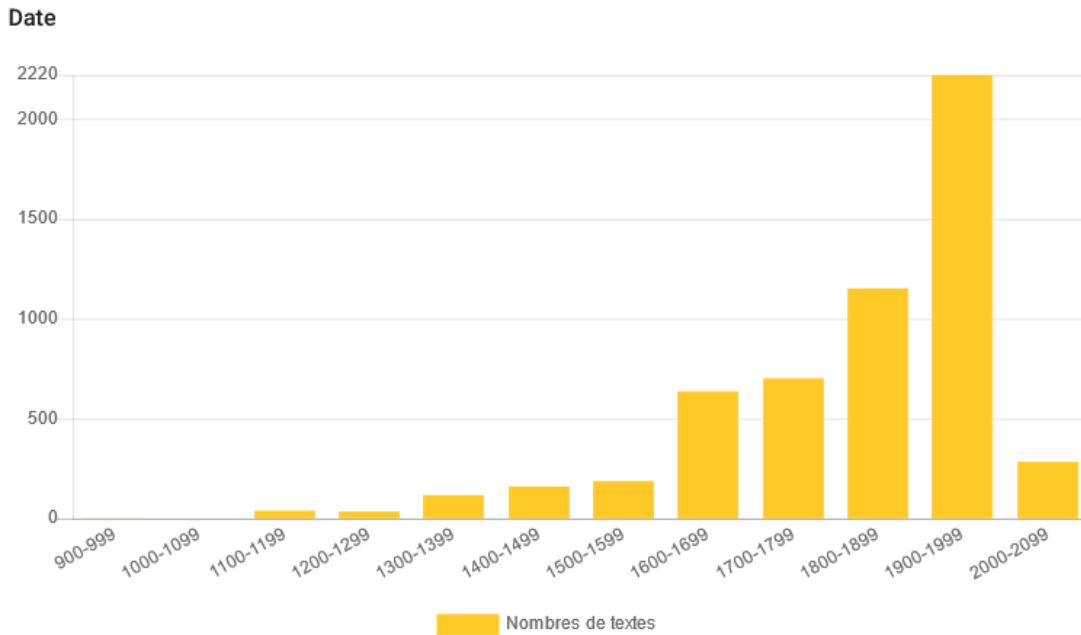


Figure 7. Répartition chronologique du corpus intégral de Frantext⁴²

La requête avec Google Ngram Viewer ne prend pas en compte les différentes graphies. Pourtant, la courbe pour le français est particulièrement haute en 1504. En effet, l'exploration des résultats dévoile qu'il s'agit d'un même ensemble d'occurrences, au sein du commentaire de Nicolas Richelet au sonnet 52 des *Amours* de Ronsard, comptée trois fois, car tel est le nombre d'exemplaires du premier tome des *Œuvres de P. de Ronsard, gentil-homme Vandomois* numérisés par Google. Cela suffit pour que le ratio de la fréquence atteigne le troisième meilleur résultat pour le français, tous siècles confondus. Or, il s'agit d'une erreur d'indexation, car même si le premier volume de cette édition paraît en 1567, celle numérisée par Google date de 1604⁴³ et la graphie moderne n'apparaît qu'au sein d'un commentaire ajouté par Richelet à cette nouvelle édition.

D'ailleurs, les erreurs de ce type au sein du corpus Google Books sont bien connues et assez nombreuses. C'est ce que constate Geoffrey Nunberg dans son article « Google's Book Search: A Disaster for Scholars »⁴⁴. Des erreurs que l'auteur observe en 2009 restent toujours non corrigées,

⁴² Voir <https://www-frantext-fr.ressources-electroniques.univ-lille.fr/repository/frantext>. Ce diagramme est accompagné d'un bref texte expliquant la raison d'une telle disparité : « [h]istoriquement dévolue à la recherche d'exemples pour le *Trésor de la Langue Française*, et d'abord centrée sur la langue des XIX^e et surtout XX^e siècle[s] la base Frantext a poursuivi son évolution en élargissant son empan diachronique : ajout de textes médiévaux, de textes en moyen français, français préclassique et classique. » Nous constatons par ailleurs que, si la volonté de dépasser le canon littéraire présente chez Moretti est également au cœur du projet, son objectif principal reste linguistique : « [s]on ambition [du projet] n'est pas d'opérer une sélection en suivant le canon littéraire, mais de représenter l'usage du français écrit dans sa diversité ».

⁴³ En effet, l'automate a reproduit l'erreur de datation originale : sur la page de titre, nous lisons bien MDIII, mais la bonne date a été inscrite au crayon gris en chiffres arabes, tant sur la page du titre que sur la page de garde, voir <https://page.bn/y5gvie>.

⁴⁴ « A search on "Internet" in books published before 1950 produces 527 results; "Medicare" for the same period gets almost 1,600. Or you can simply enter the names of famous writers or public figures and restrict your search to works

car, pour Google, un livre recèle avant tout des données informationnelles qui peuvent être indexées, hiérarchisées, extraites puis exploitées, non pas pour en faire une bibliothèque, mais pour satisfaire la demande d'information et pour en tirer profit⁴⁵. Google Ngram Viewer n'est donc pas destiné à fouiller Google Books pour approfondir nos connaissances sur la *Weltliteratur*, mais, comme le constate Alexandre Gefen, pour conduire des « expériences de “culturonomique” (quantification des tendances culturelles) les plus simples et les plus spectaculaires »⁴⁶.

À cela s'ajoute la non-transparence du contenu des corpus fouillés, qui est un des principaux défauts de Google Books et de Google Ngram Viewer. Ce type de système, souvent qualifié de « boîte noire », ne donne pas suffisamment d'informations concernant son fonctionnement interne, ou, du moins, ne les fournit pas de manière suffisamment accessible pour un utilisateur capable de l'exploiter. Le logiciel étant extrêmement facile à utiliser, ses résultats peuvent paraître enthousiasmants pour un explorateur non averti. Or, une quantité d'informations manque. Par exemple, comment se répartissent les textes numérisés dans les différentes aires linguistiques, tant par période qu'en fonction des genres ? Quelles sont les proportions entre les ouvrages déjà numérisés et ceux qui sont en attente de traitement, ou du moins recensés dans les catalogues ? A-t-on perdu des occurrences à cause d'une mauvaise reconnaissance du texte par les automates d'océrisation⁴⁷ ? Il nous est impossible, ou au moins très difficile, de répondre à ces questions en exploitant le corpus Google Books. Comme le souligne Michel Wieviorka, l'emploi des outils de ce type doit être accompagné d'une réflexion portant sur les « questions d'échantillonnage, de représentativité, d'homogénéité du corpus, [et sur les] conditions dans lesquelles les données sont produites, agrégées ou combinées »⁴⁸. Or, ces informations ne sont pas fournies de manière systématisée et explicite.

Alors que nous avons vu que Google Ngram Viewer ne permet pas de contrôler précisément quel corpus est exploré, observons que les informations concernant la répartition périodique des

published before the year of their birth. “Charles Dickens” turns up 182 results for publications before 1812, the vast majority of them referring to the writer. The same type of search turns up 81 hits for Rudyard Kipling, 115 for Greta Garbo, 325 for Woody Allen, and 29 for Barack Obama », [La recherche du mot « Internet » dans les livres publiés avant 1950 donne 527 résultats ; « Medicare » en obtient 1 600 pour la même période. On peut aussi chercher des noms d'auteurs ou de personnalités publiques et restreindre la requête aux œuvres publiés avant l'année de leur naissance. On trouve 182 occurrences de « Charles Dickens » dans les publications datant d'avant 1812, et la grande majorité d'entre elles renvoie effectivement à l'écrivain. Le même type de requête fait apparaître 81 occ. de « Rudyard Kipling », 115 occ. de « Greta Garbo », 325 de « Woody Allen » et 28 de « Barack Obama »] (nous traduisons). Voir Geoffrey Nunberg, « Google's Book Search: A Disaster for Scholars », *The Chronicle of Higher Education*, section The Review, 31 août 2009, en ligne, <<https://www.chronicle.com/article/googles-book-search-a-disaster-for-scholars/>>.

⁴⁵ Voir Roger Chartier, « Qu'est-ce qu'un livre ? Grandeurs et misères de la numérisation », dans Gérard Fussman (dir.), *La mondialisation de la recherche : Compétition, coopérations, restructurations*, Paris, Collège de France, coll. « Conférences », 2013, en ligne, <<http://books.openedition.org/cdf/1579>>.

⁴⁶ Alexandre Gefen, « Les enjeux épistémologiques des humanités numériques », *loc. cit.*, paragr. 3.

⁴⁷ Ce terme est dérivé de l'abréviation du nom anglais *Optical Character Recognition* (OCR), reconnaissance optique des caractères. L'équivalent français, ROC, reste très peu utilisé.

⁴⁸ Michel Wieviorka, *L'impératif numérique ou La nouvelle ère des sciences humaines et sociales*, Paris, CNRS, 2013, p. 26.

œuvres du corpus Frantext nous permettent d’avoir une appréciation plus juste de la distribution des occurrences du lemme *Orphée*. Voyons concrètement le diagramme proposé en Figure 8.

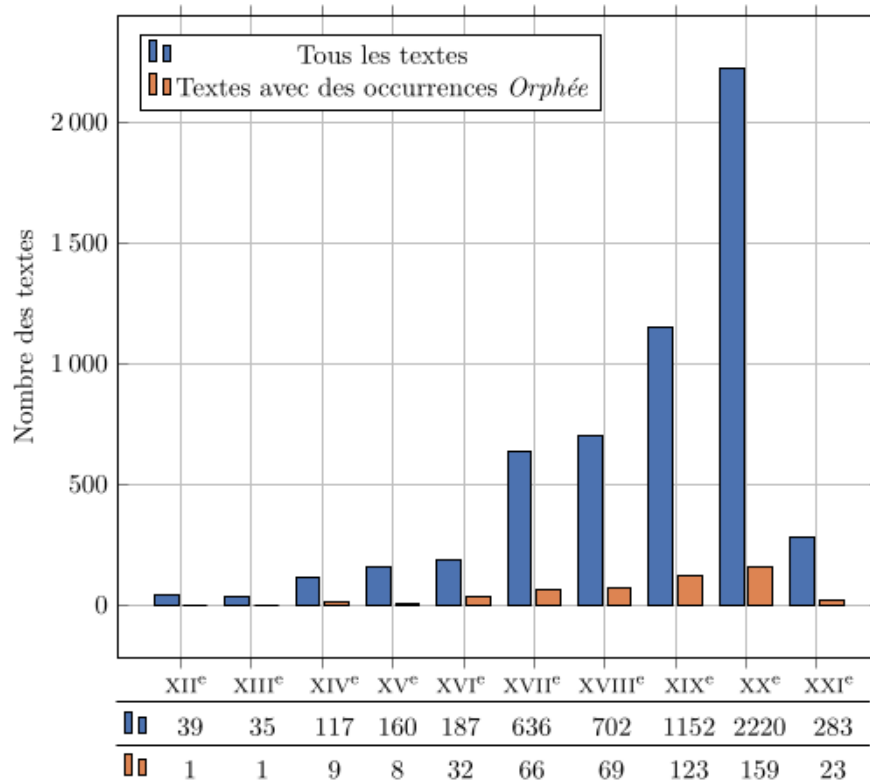


Figure 8. Répartition chronologique du corpus intégral Frantext (en bleu) comparée à la répartition chronologique des textes contenant au moins une occurrence du lemme *Orphée* (en orange)

Comme nous disposons uniquement des chiffres concernant le nombre d’œuvres qui composent le corpus intégral Frantext, au lieu de décompter les occurrences, il peut être intéressant de prendre en compte le nombre des œuvres qui mentionnent *Orphée* au moins une fois⁴⁹. Toutefois, les données ainsi représentées sont peu parlantes. Les occurrences sont généralement minoritaires, et le sous-corpus le plus fourni (celui du XX^e s.) en contient le plus, contrairement à ceux les moins fournis (XII^e et XIII^e s.), ce qui ne semble pas surprenant. De ce point de vue, la prise en compte de la fréquence relative de Google Ngram Viewer est plus pertinente. Cet outil calcule la fréquence relative des occurrences du mot en divisant le nombre des occurrences du mot recherché par le nombre total des mots dans les textes numérisés de chaque année. Nous adaptons la méthode de Google pour visualiser les résultats de notre requête avec Frantext. Pour ce faire, nous calculons la

⁴⁹ Nous verrons *infra* les écueils de la prise en compte non différenciée des occurrences épisodiques et celles plus récurrentes, voire de la recherche par occurrences en général, notamment dans le contexte d’une recherche portée sur les réécritures d’un mythe.

proportion des textes avec au moins une occurrence d'Orphée par rapport au nombre total des textes numérisés pour chaque siècle, comme montré en Figure 9.

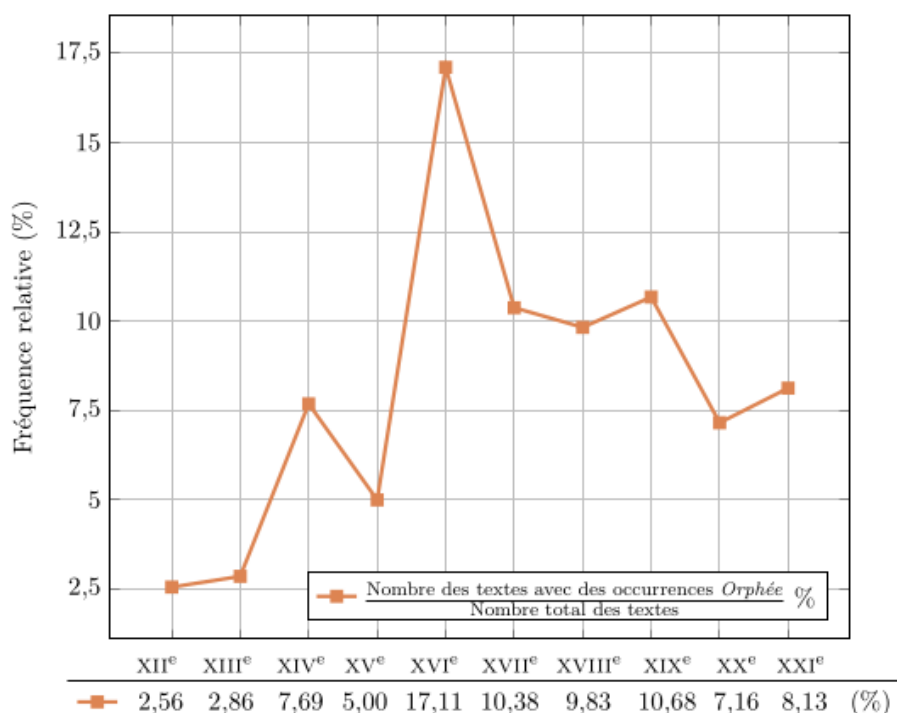


Figure 9. Répartition chronologique du pourcentage des textes avec au moins une occurrence d'Orphée par rapport au nombre total des textes pour chaque période

Ce diagramme montre que les proportions des textes avec au moins une mention du lemme *Orphée* sont particulièrement élevées au XVI^e s., et restent relativement importantes aux XVII^e et XVIII^e s. Des fréquences comparables sont observables au XIV^e et au XX^e s., et une légère hausse (1 %) survient au XXI^e s. Même s'il était intéressant de comparer ces résultats avec l'exploration de Google Ngram Viewer, l'insuffisance des informations concernant le corpus français de Google Books et l'impossibilité de télécharger les résultats ne permettent pas de le faire correctement.

c) Gallica et Gallicagram : remède à tous les maux ?

L'exploration de corpus sélectionnés de manière plus transparente⁵⁰ et l'emploi d'outils dédiés conçus avec rigueur permettent de pallier une bonne partie des problèmes épistémologiques que pose Google Books. Toutefois, d'autres problèmes, sans doute moins généraux et dépendant de l'objectif spécifique de la recherche, persistent. Par exemple, il est regrettable que la restriction de la recherche dans des parties de textes (titres, dédicaces...) en particulier ne soit pas permise. En

⁵⁰ Même si l'exploration de Frantext est moins problématique que celle de Google Books, elle n'est pas pour autant dépourvue des difficultés épistémologiques, notamment concernant le choix des éditions numérisées.

effet, une occurrence du nom d'Orphée dans le titre de l'ouvrage n'a pas la même importance que sa mention en note de bas de page, au sein d'une citation, voire dans une référence bibliographique. La connaissance du corpus interrogé reste également limitée, que ce soit dans le cas déjà discuté de Frantext, ou dans celui de Gallicagram⁵¹, l'outil spécifiquement dédié à l'exploration de Gallica, la bibliothèque numérique de la BnF, qui donne accès gratuitement et librement à plus de 8,5 millions de documents, dont des numéros de presse (5,5 millions), des livres (1,2 million) et des manuscrits (200 000)⁵². D'une extrême transparence et libre d'accès, Gallicagram est une réponse à Google Ngram Viewer, dont il s'inspire sans ignorer ses écueils.

En effet, l'utilisation du corpus de Gallica garantit une plus grande qualité de description des œuvres, cataloguées et indexées par les experts de la BnF, et la transparence concernant les occurrences trouvées. Un clic sur un point du diagramme renvoie aux résultats de la recherche dans Gallica pour la période choisie et permet de localiser immédiatement le nombre d'occurrences, leur contexte et les œuvres dont elles font partie. La maîtrise du corpus est également plus grande, d'autant plus que Gallicagram permet de restreindre la recherche au sous-corpus des livres ou celui de la presse uniquement. La requête qui recherche les graphies *Orphée*, *Orpheus* ou *Orfée* dans le corpus des livres de Gallica génère le diagramme de la Figure 10.

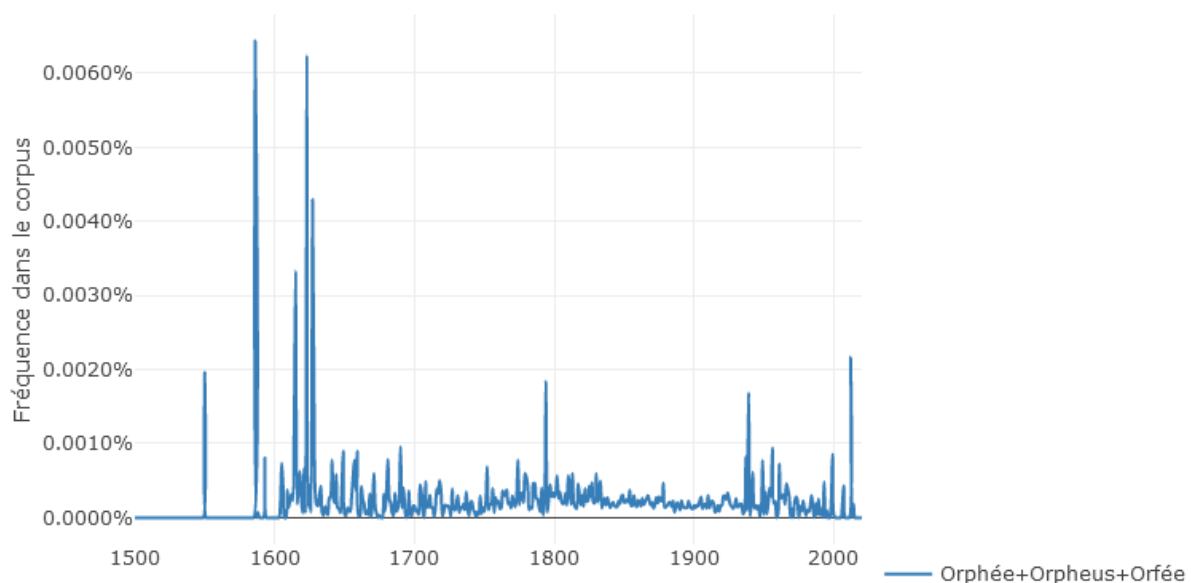


Figure 10. Occurrences des trois graphies d'Orphée dans le corpus des livres disponibles sur Gallica. Gallicagram⁵³

⁵¹ Voir <https://shiny.ens-paris-saclay.fr/app/gallicagram>, et Benjamin Azoulay et Benoît de Courson, *Gallicagram : un outil de lexicométrie pour la recherche*, SocArXiv, 2021.

⁵² Voir <https://web.archive.org/web/20221228212743/https://gallica.bnf.fr/GallicaEnChiffres>.

⁵³ Pour reproduire cette visualisation, rechercher « Orphée+Orpheus+Orfée » dans la langue française, source : Gallica, corpus : Livres / Gallica, Mode de recherche : par n-gramme, début : 1500, fin : 2020.

Il est intéressant de comparer ce diagramme avec celui obtenu avec Frantext (Figure 9). Une certaine corrélation peut être observée concernant une baisse du nombre d'occurrences entre le XIX^e et le XX^e s., puis une remontée entre le XX^e et le XXI^e s. (qui paraît toutefois plus importante en Figure 10). En revanche, alors que le diagramme en Figure 9 montre que les proportions des œuvres qui mentionnent Orphée sont les plus importantes au XVI^e s. La visualisation de Galligram révèle aussi un pic de fréquences à cette période, mais il est isolé et survient plutôt à la fin du siècle (1586⁵⁴). Par ailleurs, cette hausse marque le début de l'augmentation de la fréquence des occurrences, avec trois pics successifs, en 1615, 1623 et 1627. Toutefois, force est de constater que parmi quinze livres au sein desquels sont trouvées les occurrences de ces trois années, un seul est littéraire, *Le mariage d'Orphée, sa descente aux enfers et sa mort par les bacchantes*, une tragédie de Charles Lespine (1623). Parmi les autres ouvrages, nous trouvons des essais médicaux (*Le royal syrop de pommes, antidote des passions melancholiques* du docteur Gabriel Droyn, 1615) et religieux (*Discours de l'estat et des grandeurs de Jésus* par Pierre de Bérulle, 1623), ou des mythographies (traduction de la *Mythologie, ou Explication des fables* de Natale Conti, revue par Jean Baudoin et traduite par Jean de Montlyard, 1627). D'autres œuvres littéraires sont présentes, mais ne font que mentionner Orphée. Par exemple, dans « L'Ode au Roy » de Vincent Borée (*Les princes victorieux, tragédies françoises*, 1627), l'unique occurrence a la valeur d'une antonomase, car le nom d'Orphée, mis au pluriel, désigne les poètes inspirés : « En cest espoir les doctes Fées / Vont animant dix mille Orphées » (p. 686).

Outre le problème de la hiérarchisation des occurrences déjà relevé, les résultats aussi peu fournis pour les XVI^e et XVII^e s. incitent à la prudence dans l'analyse de cette période. Fort heureusement, la composition de la totalité du corpus peut être analysée de manière approfondie pour mieux rendre compte d'éventuels biais d'interprétation des résultats. L'onglet « Distributions » de Galligram permet d'afficher des graphiques qui montrent la répartition du corpus exploité selon plusieurs critères, comme la distribution chronologique, la qualité de l'OCR, etc.⁵⁵ Cette transparence dévoile que le corpus est, forcément, restreint aux textes océrisés, soit à environ un tiers de la totalité des ouvrages mis à disposition par la bibliothèque numérique. La reconnaissance du texte est d'ailleurs plus régulièrement appliquée aux œuvres plus récentes, publiées à partir du XIX^e s. Enfin, le graphique « État de la numérisation »⁵⁶ montre que moins de 10 % de toute la production éditoriale a été numérisée (au moins en format image) jusqu'alors.

⁵⁴ Il est dû à un seul texte, *Le Théâtre des divers cerveaux du monde* de Tommaso Garzoni, traduit d'italien par Gabriel Chapuis.

⁵⁵ Voir <https://shiny.ens-paris-saclay.fr/app/galligram>, onglet « Distributions », puis « Corpus de livres ».

⁵⁶ *Idem*.

Conclusion du chapitre

En analysant les approches numériques de la littérature du XIX^e s., Andrea Del Lungo et Giovanni Pietro Vitali⁵⁷ constatent que le milieu universitaire français est davantage tourné vers l'édition numérique, et, par conséquent, particulièrement concentré sur le texte, contrairement à d'autres aires culturelles, plus enclines à « traiter les problématiques littéraires avec des méthodologies typiques du *distant reading* »⁵⁸. Certes, l'importance de la philologie dans la déontologie critique française y est sans doute pour quelque chose. Mais les auteurs relèvent aussi une autre raison :

Afin de développer des analyses typiques du *distant reading* à travers des instruments numériques, il est nécessaire de posséder les textes au format numérique ; or, surtout pendant la phase de construction de beaucoup d'instruments d'analyse numérique dont on profite aujourd'hui, les premiers textes accessibles librement et en grande quantité étaient en anglais. Les ressources textuelles, principalement en anglais, ont facilité des approches textuelles de type interprétatif, alors que dans le monde francophone, où les textes étaient difficiles à repérer, l'attention a été portée sur la collection de corpus en français, avec un intérêt particulier pour les pratiques acquisitionnelles du texte.⁵⁹

C'est également ce que nous observons, en particulier lors de notre exploitation du corpus Google Books (mais aussi, quoique dans une moindre mesure, celui de Frantext et de Gallica). Toutefois, comme les auteurs, nous choisissons de rester optimiste, car « les possibilités à venir sont nombreuses, surtout si les différents projets menés jusqu'ici pouvaient croiser leurs expériences ou mettre en commun leurs méthodes et leurs résultats, suivant le nouvel esprit de partage et de collaboration typique des humanités numériques »⁶⁰. Les exemples de cette collaboration ne manquent pas. Parmi eux se trouve le projet international *Distant Reading COST Action European Literary Text Collection (ELTeC)*⁶¹, une des initiatives clés des dernières années qui vise à produire un corpus littéraire international et plurilingue particulièrement proche des idées de la *Weltliteratur* de Franco Moretti. La première phase du projet consistait à construire six sous-corpus, chacun constitué de 100 romans écrits entre 1840 et 1920 dans différents pays européens. Actuellement, dix sous-corpus sont achevés, et sept autres sont en cours de création⁶². La prochaine étape prévoit l'extension des collections à la période de 1780 à 1850. Une politique aussi rigoureuse permettra, nous semble-t-il, de surmonter les faiblesses pointées dans cette critique formulée par Gefen :

⁵⁷ Andrea Del Lungo et Giovanni Pietro Vitali, « La littérature du XIX^e siècle au prisme des humanités numériques », *Romantisme*, vol. 191, n° 1, 2021, p. 64-75, en ligne, <<http://www.cairn.info/revue-romantisme-2021-1-page-64.htm>>.

⁵⁸ *Ibid.*, p. 70.

⁵⁹ *Ibid.*, p. 71.

⁶⁰ *Ibid.*, p. 70.

⁶¹ Voir <https://www.distant-reading.net/eltec/>.

⁶² Voir <https://distantreading.github.io/ELTeC/>.

[D]ans l'anarchie des corpus numériques des bibliothèques, le rêve d'exhaustivité d'une bibliothèque de Babel est [...] éloigné de sa réalisation [...]. Ainsi, souvent, le travail [...] doit se faire par bricolage, à partir de corpus partiellement représentatifs et partiellement organisés, imposant des biais considérables à toute analyse quantitative et requérant des chercheurs une conscience méthodologique aguerrie.⁶³

En s'inspirant de méthodes et de développements existants, de plus en plus de projets tentent tantôt de résoudre des difficultés épistémologiques repérées par la communauté, tantôt de se les approprier. Pour ce faire, il est crucial d'adapter les approches, outils et méthodes généralistes aux besoins particuliers d'une recherche particulière. Dans un de ses derniers articles, Moretti constate qu'« [i]l s'agit de comprendre ce qu'un paradigme de recherche *fait*, plutôt que ce qu'il déclare vouloir faire »⁶⁴. Comme Gallicagram s'efforce de montrer ses limites pour faciliter leur dépassement, les difficultés liées aux grands corpus textuels et à l'adaptation des outils et méthodes aux besoins spécifiques des chercheurs révèlent la nécessité d'une herméneutique spécifique. Il est indispensable de trouver un équilibre entre la recherche empirique et l'exploitation d'outils informatiques, entre les méthodes traditionnelles de son domaine d'expertise et les usages et méthodes computationnelles quantitatives. Moretti conclut son article en disant qu'« il y a une logique à notre pratique, et nous devrions chercher à la comprendre – à comprendre ce que nous faisons véritablement lorsque nous étudions la littérature – plutôt que d'évoquer une synthèse que personne n'a jamais vue »⁶⁵. C'est ce que nous nous attachons à faire en constituant notre corpus des réécritures selon une herméneutique spécifique qui fait son deuil du rêve d'exhaustivité d'une bibliothèque de Babel.

⁶³ Alexandre Gefen, « Les enjeux épistémologiques des humanités numériques », *loc. cit.*, paragr. 3.

⁶⁴ Franco Moretti, « Les études littéraires entre herméneutique et quantification », *Revue italienne d'études françaises. Littérature, langue, culture*, n° 10, trad. Anne-Frédérique Schläpfer, *Revue italienne d'études françaises*, novembre 2020, paragr. 1, en ligne, <<http://journals.openedition.org/rief/6508>>.

⁶⁵ *Ibid.*, paragr. 24.

2. Établissement du corpus : entre l'exhaustivité quantitative et la représentativité herméneutique

Sans la prise de conscience de son héritage heuristique, la lecture distante risque de se limiter à « une déssentialisation du texte, désormais dénudé dans sa texture et placé au centre d'un empire hypertextuel et de systèmes variationnels qui le décentrent »¹ et de ne manipuler que des « données qui restent des artéfacts muets en l'absence d'une herméneutique spécifique »². La question de l'équilibre entre l'étude quantitative et qualitative est présente dans la plupart des réflexions sur la lecture distante et les humanités numériques littéraires. Nous nous attachons particulièrement à la conclusion de l'article de Del Lungo et Vitali :

La réintroduction d'une forme de subjectivité paraît en effet nécessaire afin de combattre le risque d'une dérive quantitative à laquelle s'exposent les études littéraires, les humanités numériques fournissant des outils qui pourraient être conçus comme des instruments de mesure au sein d'une approche prétendument objective du texte. L'enjeu actuel, que doivent affronter les recherches littéraires dans le domaine, consiste à réaliser une alliance du quantitatif et de l'herméneutique dont la clé réside dans la place laissée à la subjectivité critique, mais aussi dans une nouvelle conception de l'édition : sans suivre davantage la chimère de l'exhaustivité qui fait miroiter le numérique, celle-ci gagnerait peut-être à se défaire de la logique fortement auctoriale qui a jusqu'ici présidé aux projets en humanités numériques littéraires, pour s'ouvrir plus amplement à la logique de l'hypertexte.³

Alors que la préconisation de « l'alliance du quantitatif et de l'herméneutique » dans la phase analytique du projet est discutable⁴, celle-ci est particulièrement probante lors de l'étape de l'établissement d'un corpus qui vise à regrouper de manière cohérente un ensemble hétérogène d'œuvres de genres et d'époques différents.

En effet, nous tâchons de combiner deux méthodes d'établissement du corpus qui, à première vue, paraissent antithétiques. La première est quantitative et hérite tout autant de la démarche du *distant reading* que de méthodes de la linguistique du corpus et du traitement automatique des langues (TAL). La méthode qualitative, quant à elle, reste fidèle aux usages et traditions de la recherche empirique en littérature, et plus particulièrement en littérature comparée.

¹ Alexandre Gefen, « Les enjeux épistémologiques des humanités numériques », *loc. cit.*, paragr. 1.

² *Ibid.*

³ Andrea Del Lungo et Giovanni Pietro Vitali, « La littérature du XIX^e siècle au prisme des humanités numériques », *loc. cit.*, p. 74-75.

⁴ Nous renvoyons en particulier à Franco Moretti, « Les études littéraires entre herméneutique et quantification », *loc. cit.* : « La quantification peut bien donner de nouveaux objets à l'activité herméneutique ; les interprétations se prêter à la falsification quantitative : cela est clair. Mais ici, la question porte sur la relation entre *les catégories* de l'analyse : est-ce que les catégories quantitative et herméneutique ont la capacité d'interagir afin d'*unifier conceptuellement* les deux approches ? A-t-on la possibilité de formuler une synthèse théorique ? » (paragr. 17).

En suivant les postulats de Moretti et de ses successeurs, ainsi que les usages linguistiques et talistes, trois enjeux principaux de la méthode quantitative se dégagent pour l'établissement du corpus littéraire en vue de l'interrogation computationnelle :

- un protocole de collecte et de choix des œuvres clairement défini et objectivable ;
- une exhaustivité ou une représentativité⁵ ;
- un équilibre entre le canon et les œuvres oubliées ou dépréciées (le *non-canon*).

Ces règles peuvent permettre de minimiser les écueils liés aux particularités de l'environnement numérique sans oublier l'objectif principalement littéraire de la recherche. Toutefois, nous sommes rapidement amenée à faire le deuil de l'exhaustivité du corpus, impossible à prétendre faute de ressources et de temps, dans le cadre d'un projet doctoral. Nous décidons également de laisser de côté, pour le moment, l'ambition d'étudier la littérature mondiale, et de nous restreindre aux œuvres en langue française uniquement. Les corpus multilingues collectés dans le cadre des recherches linguistiques, regroupent des documents traduits dans plusieurs langues (il s'agit alors des corpus *parallèles*). Dans certains cas, les passages, paragraphes, phrases ou mots correspondants sont reliés (les corpus sont alors *alignés*). Regrouper les réécritures d'un mythe dans des langues différentes au sein d'un même corpus rendrait difficile l'exploitation de la plupart des outils d'interrogation computationnelle, conçus majoritairement pour travailler avec une langue à la fois. Nous décidons également d'écarter les traductions françaises des ouvrages étrangers, car elles posent des questions épistémologiques particulières et risquent d'introduire des biais d'analyse. Par exemple, les différences entre le style de l'auteur et du traducteur sont difficilement traçables et la dissemblance de leurs horizons culturels, voire même des époques de leur existence, peut rendre leurs sources respectives assez disparates.

Afin de choisir quels textes intégrer au corpus et de limiter autant que possible différents biais d'exploitation des ressources numériques, comme les difficultés observées dans la partie précédente, nous nous tournons vers des méthodes de recherche littéraire traditionnelle. Cela implique, tout d'abord, la compréhension et la définition de notre objet d'étude : qu'est-ce qu'une réécriture ? Et, plus particulièrement, une réécriture d'un mythe ? Comment la reconnaître et la définir ? Finalement, une définition universelle existe-t-elle ?

⁵ Nous nous focalisons davantage sur cette notion, et de manière plus générale sur les enjeux de la linguistique du corpus, au I.3 « Représentation de la variété des réécritures du mythe », p. 65.

a) Heidmann, Brunel et Genette : définition de la notion de réécriture

Jean Ricardou appelle *réécriture* un « ensemble des manœuvres qui vouent un écrit à se voir supplanté par un autre »⁶. Pour Henri Béhar, la *réécriture* est « toute opération consistant à transformer un texte de départ A pour aboutir à un nouveau texte B »⁷. Antoine Compagnon, lui, considère que « [l]e travail de l'écriture est une réécriture dès lors qu'il s'agit de convertir des éléments séparés et discontinus en un tout continu et cohérent, de les rassembler, de les comprendre (de les prendre ensemble), c'est-à-dire de les lire »⁸. Par ailleurs, Ute Heidmann⁹ propose une méthode appelée « comparaison différentielle et discursive » qui vise à analyser et à comparer de multiples façons de récrire les mythes pour comprendre le processus de création de sens dans chaque œuvre. Pour Heidmann, une réécriture est toujours étroitement liée avec son contexte d'énonciation, qui lui octroie à chaque fois un nouvel effet de sens, différent de celui qu'elle aurait eu dans un autre contexte. Pour figurer cette dynamique, et la convergence entre l'acte d'écriture et celui de réécriture, la chercheuse propose de recourir à un préfixe contracté et parenthésé pour souligner son optionnalité, (*r*)écriture.

La divergence des approches¹⁰, qui se manifeste dès le choix du terme (*réécriture*, *réécriture*, voire (*r*)écriture) rend particulièrement difficile la tâche d'application des théories scientifiques au cas concret de l'établissement du corpus. Toutefois, nous pouvons retenir la récurrence du caractère globalisant du terme : contrairement à des concepts comme la reprise, l'allusion ou la citation, la réécriture implique une opération effectuée sur l'ensemble d'un texte source, et non pas sur un extrait restreint. Cela rapproche la réécriture des travaux sur la transtextualité de Gérard Genette¹¹, et plus particulièrement ceux sur les relations hypertextuelles, définies comme « toute relation unissant un texte B à un texte antérieur A sur lequel il se greffe d'une manière qui n'est pas celle du commentaire »¹².

Pour illustrer le concept de l'hypertextualité, Genette utilise la métaphore d'un palimpseste, où l'hypotexte serait la première écriture effacée par le nouveau texte, à savoir l'hypertexte, écrit par-dessus l'ancien. Le chercheur propose une typologie détaillée des pratiques hypertextuelles, selon le régime (ludique, satirique ou sérieux) et la relation (de transformation, qui vise un texte, ou de

⁶ Jean Ricardou, « Pour une théorie de la réécriture », *Poétique*, n° 77, 1989, p. 3.

⁷ Henri Béhar, « La réécriture comme poétique – ou le même et l'autre », *Romantic Review*, n° 72/1, janvier 1981, p. 51.

⁸ Antoine Compagnon, *La Seconde main, ou le travail de la citation*, Paris, Seuil, 1979, p. 32.

⁹ Ute Heidmann, « Différencier au lieu d'universaliser. Comparer les façons de (*r*)écrire des mythes », *Interférences littéraires/Littéraire interférenties*, n° 17, 2015, p. 15-34, en ligne, <<http://interferenceslitteraires.be/index.php/illi/article/view/117>>.

¹⁰ Pour une étude de la notion de la réécriture et ses théorisations par la critique, voir Anne Garcia, « La réécriture dans l'œuvre de José Emilio Pacheco. Une poétique du déjà-lu », Thèse de doctorat, Université Paris-Est, 2016.

¹¹ Gérard Genette, *Palimpsestes. La littérature au second degré*, Paris, Seuil, coll. « Poétique », 1982.

¹² *Ibid.*, p. 13.

l'imitation, qui reproduit un style, une manière) qui lient l'hypo- et l'hypertexte. La parodie, par exemple, est classée comme une relation de transformation dans un régime ludique (Tableau 1).

Tableau 1. Tableau des relations hypertextuelles de Genette

Régime Relation	Ludique	Satirique	Sérieux
Transformation	Parodie (<i>Chapelain décoiffé</i>)	Travestissement (<i>Virgile travesti</i>)	Transposition (<i>Doctor Faustus</i>)
Imitation	Pastiche (<i>L'Affaire Lemoine</i>)	Charge (<i>À la manière de...</i>)	Forgerie (<i>Suite d'Homère</i>)

L'hypertextualité, telle qu'elle est conçue par Genette, « englobe entièrement certains genres cano- niques (quoique mineurs) comme le pastiche, la parodie, le travestissement, et en traverse d'autres – probablement tous les autres »¹³. Cette exhaustivité est liée à une délimitation très précise de chaque pratique hypertextuelle et illustrée par des exemples concrets¹⁴.

À cela s'ajoute la prise en compte de l'aspect spéculatif, et par extension trop subjectif, de l'ana- lyse et de la recherche des relations qui lient différentes œuvres et qui risquent de se limiter à « tra- quer dans n'importe quelle œuvre les échos partiels, localisés et fugitifs de n'importe quelle autre, antérieure ou postérieure. Une telle attitude aurait pour effet de verser la totalité de la littérature universelle dans le champ de l'hypertextualité, ce qui en rendrait l'étude peu maîtrisable »¹⁵. La ri- gueur de classement apparaît alors comme un garde-fou qui limite le rôle interprétatif du lecteur. Simultanément, Genette insiste sur le fait que l'auteur peut déclarer l'influence de l'hypotexte « au moyen d'un indice paratextuel qui a valeur contractuelle : *Virgile travesti* est un contrat explicite de travestissement burlesque, *Ulysse* est un contrat implicite et allusif qui doit au moins alerter le lecteur sur l'existence probable d'une relation entre ce roman et l'*Odyssée* »¹⁶.

Nous trouvons dans l'approche genettienne des éléments d'informations permettant de recon- naître une réécriture avec des moyens relativement objectivables (indice auctorial dans le paratexte, dès le titre). Cependant, son application aux mythes peut paraître problématique. Par exemple, dans son article « De l'intertextualité à la réécriture » Anne-Claire Gignoux critique l'élargissement du concept vers les domaines mythiques (et historiques), qu'elle qualifie d'une « simple référence cul- turelle ou artistique » (paragr. 5). Elle s'oppose ainsi à ce que propose Marc Eigeldinger en 1987 :

¹³ *Ibid.*, p. 17.

¹⁴ Dans son article, Nathalie Limat-Letellier propose un tableau général et exhaustif des pratiques hypertextuelles qui offre une synthèse extrêmement claire du travail de classement effectué par Genette. Voir Nathalie Limat-Letellier, « Historique du concept d'intertextualité », dans Marie Miguët-Ollagnier (dir.), *L'intertextualité*, Besançon, Presses uni- versitaires de Franche-Comté, coll. « Annales littéraires », 1998, p. 17-64, en ligne, <<http://books.openedition.org/pufc/4507>>.

¹⁵ Gérard Genette, *Palimpsestes. La littérature au second degré*, op. cit., p. 18-19.

¹⁶ *Ibid.*, p. 17.

Mon projet est de ne pas limiter la notion d'intertextualité à la seule littérature, mais de l'étendre aux divers domaines de la culture. Elle peut être liée à l'émergence d'un autre langage à l'intérieur du langage littéraire ; par exemple celui des beaux-arts et de la musique, celui de la Bible ou de la mythologie, ainsi que celui de la philosophie.¹⁷

Pouvons-nous réellement exclure la mythologie des réflexions autour du concept de réécriture ? Dans l'ouvrage collectif *La littérature dépliée. Reprise, répétition, réécriture*¹⁸, six articles sont regroupés dans un chapitre dédié spécifiquement aux mythes. Plusieurs autres articles analysent les réécritures mythologiques au théâtre (« Phèdre et Nora : convergence de deux figures dans une réécriture contemporaine (Insolation de W. Palka) » de Béatrice Jongy) ou au cinéma (« Orphée chez Ophuls » de Laurence Moinereau). L'ouvrage collectif dirigé par Peter Schnyder, *Métamorphoses du mythe. Réécritures anciennes et modernes des mythes antiques*¹⁹, qui paraît la même année, regroupe plus de 60 articles dédiés spécifiquement à ce sujet. Ce foisonnement de travaux prouve que les concepts de réécriture et de mythe sont fréquemment liés au sein des recherches littéraires, et en particulier celles inscrites dans le champ de la littérature comparée.

Par ailleurs, une théorisation spécifique du réemploi des mythes dans les textes littéraires a été progressivement forgé au sein de ce domaine de recherche. À commencer par la théorie de la recherche morphologique d'André Jolles²⁰, par la mythologie comparée de Georges Dumézil²¹, puis la mythanalyse de Denis de Rougemont²², la notion de la mythocritique se crée peu à peu dans la deuxième moitié du XX^e s., grâce à Durand²³ d'abord, qui la fait évoluer vers ce qu'il appellera à la fin du siècle la « mythodologie »²⁴, et à Brunel²⁵ ensuite, qui propose une autre mythocritique, plus éloignée de la psychocritique et appartenant pleinement à la déontologie comparatiste.

En outre, les approches généralement applicables à la littérature, comme celle de Genette, gagnent à être combinées avec les méthodes spécifiques aux mythes, comme celle de Brunel. Cette dernière apporte des indices supplémentaires quant aux facteurs paratextuels facilitant la reconnaissance d'une réécriture d'un mythe :

¹⁷ Marc Eigeldinger, « Introduction », *Mythologie et intertextualité*, Genève, Slatkine, 1987, p. 15.

¹⁸ Jean-Paul Engélibert et Yen-Mai Tran-Gervat (dir.), *La littérature dépliée : reprise, répétition, réécriture*, Rennes, Presses Universitaires de Rennes, coll. « Interférences », 2008.

¹⁹ Peter Schnyder (dir.), *Métamorphoses du mythe : réécritures anciennes et modernes des mythes antiques*, Paris, Orizons, coll. « Universités – domaine littéraire », 2008.

²⁰ André Jolles, *Formes simples*, trad. Antoine Marie Buguet, Paris, Seuil, 1972.

²¹ Georges Dumézil, *Mythe et épopée*, Paris, Gallimard, coll. « Bibliothèque des sciences humaines », 1968.

²² Denis de Rougemont, *Lumières du mythe*, Paris, Presses Universitaires de France, 1983.

²³ Gilbert Durand, *Figures mythiques et visages de l'œuvre : de la mythocritique à la mythanalyse*, Paris, Berg international, coll. « L'Île verte », 1979.

²⁴ Gilbert Durand, *Introduction à la mythodologie : mythes et sociétés*, Paris, Albin Michel, coll. « La pensée et le sacré », 1996.

²⁵ Pierre Brunel, *Mythocritique : théorie et parcours*, Paris, Presses universitaires de France, coll. « Écriture », 1992.

Un texte peut reprendre un mythe, il entretient une relation avec lui. Mais la mythocritique s'intéressera surtout à l'analogie qui peut exister entre la structure du mythe et la structure du texte. [...] Un texte à lui seul peut être une répétition. Son titre redit souvent le nom d'un personnage mythique : *Médée*, ou *Ulysse*, ou « Artémis ». ²⁶

Bien sûr, le titre ne peut pas être le seul indice de réécriture mythologique. Par exemple, « El Desdichado » de Gérard de Nerval propose une excellente réécriture du mythe orphique au moyen des répétitions plus subtiles, ce que Brunel définit comme *émergence* des mythes : « [u]ne analyse de ce genre [(mythocritique)] paraît plus légitime si elle part de l'examen d'occurrences mythiques dans le texte. Sans doute ne peut-on s'en tenir à une description de la surface du texte. Mais sans elle le danger est grand de fabuler » ²⁷. L'émergence est un de trois piliers de la mythocritique, à côté de la *flexibilité* et de l'*irradiation* des mythes.

Plutôt que de chercher des indices paratextuels, Brunel conseille de commencer l'analyse par le recensement des occurrences mythiques dans le texte. Ces occurrences peuvent être de natures très différentes (évocation des noms propres, allusion à un des épisodes du mythe...) : « [i]l suffit d'un nom, qui est le premier à émerger, d'une caractéristique [...], d'un acte fondamental », dit le comparatiste en observant l'émergence d'un épisode odysseén dans le poème « Voyage » de Baudelaire (p. 73). Toutefois, le recensement des occurrences dans un corpus numérique pose de nombreux problèmes, tant techniques (qualité et taux d'océrisation) qu'épistémologiques (inclusion des occurrences dans les références bibliographiques, notes, etc.) ²⁸. Par exemple, quelle requête serait optimale pour différencier les réécritures des allusions ponctuelles ? Nous pourrions fixer un ratio minimum d'occurrences qu'une œuvre doit contenir pour être considérée comme étant une réécriture, mais la pertinence d'une telle requête, trop arbitraire, est incertaine. Par ailleurs, quels mots-clés faut-il décompter ? Le nom *Orphée* est-il un indice suffisant ? Si l'apparition du nom *Eurydice* est rendue obligatoire, ne risquons-nous pas d'omettre les œuvres qui ne l'évoquent que par périphrase ? Quel que soit notre choix d'interrogation, il reste difficile d'évaluer le nombre d'œuvres manquantes et d'assurer un taux satisfaisant de résultats pertinents.

Finalement, nous considérons qu'une requête suffisamment objective, et qui fait écho aux théories de Genette et de Brunel, est celle qui permet de restreindre notre recherche à toutes les œuvres mentionnant Orphée ou Eurydice dans leur titre. Ce faisant, nous n'oublions pas la mise en garde de Brunel quant à l'insuffisance de cette approche : en effet, une exploitation "objective" des ressources numériques devra être complétée par des approches qualitatives.

²⁶ *Ibid.*, p. 66-67.

²⁷ *Ibid.*, p. 72-73.

²⁸ Voir I.1 « Lecture distante et rêve de littérature mondiale », p. 30.

b) Point de départ : exploration de l'archive numérique Gallica

Pour commencer, nous interrogeons les documents disponibles sur Gallica. Nous avons déjà évoqué les avantages liés à l'exploitation d'une ressource adossée à une institution nationale telle que la BnF : catalogage minutieux, transparence du contenu de l'archive, outils d'interrogation innovants et conçus en concertation avec les chercheurs, etc. Nous pouvons y ajouter des avantages plus techniques, comme la bonne qualité des images numérisées et du texte obtenu par océrisation, ou encore la richesse des fonctions avancées du moteur de recherche. Par ailleurs, même si la stratégie de la numérisation des ressources de la BnF a probablement priorisé le canon littéraire²⁹, Gallica est particulièrement fournie en ressources moins connues.

La fouille du contenu de Gallica à la recherche des œuvres dont le titre comporte l'un des noms *Orphée* ou *Eurydice*, peut être faite grâce au moteur de recherche avancée³⁰. Nous demandons à ce que le titre contienne une des graphies les plus fréquentes d'un des deux noms (*Orphée*, *Orfée*, *Orpheus* et *Eurydice*, *Euridice*, *Erudice*). La recherche doit porter sur des œuvres en langue française uniquement (champ « Par langue ») et n'inclure que des livres, des partitions ou des manuscrits (champ « Par type de document »).

Une telle requête aboutit à 91 résultats, dont 69 sont des œuvres mises à disposition sur Gallica et 21 sont trouvées dans d'autres bibliothèques numériques collaborant avec l'archive de la BnF : des bibliothèques municipales ou spécialisées de Bordeaux, Dijon, Agen, Strasbourg et Paris. Les ouvrages en plusieurs tomes sont référencés volume par volume, ceux ayant connu plusieurs éditions ou ceux dont le manuscrit a été numérisé par la BnF sont recensés plusieurs fois. Finalement, 56 documents différents ont été trouvés au total. Parmi eux, sept sont classés dans les manuscrits, mais seul le tapuscrit du scénario du film « Légende d'Orphée », par Pathé Frères (1909), pourrait être intégré au corpus. Trois recueils, contenant des coupures de presse, programmes, etc. des pièces *Eurydice* de Jean Anouilh et *Orphée et son amour* de Népomucène Jonquille, et du film *Orphée* de Jean Cocteau, sont intéressants pour envisager l'inclusion des œuvres qu'ils traitent dans le corpus (à condition de disposer des ressources textuelles, comme les scénarios publiés). En revanche, les trois documents restants sont à écarter. Il s'agit d'un journal de bord de la corvette *L'Eurydice* de Francis Garnier, d'un reçu pour le prêt des parties d'orchestre fait à Antoine Banès par

²⁹ « Un des critères de sélection des collections à numériser est la prise en compte de l'intérêt que les collections peuvent susciter auprès du plus grand nombre possible de segments de publics : actifs, enseignants et médiateurs, scolaires et étudiants, jeunes et enfants... », voir BnF, *Charte documentaire de la numérisation des collections de la Bibliothèque nationale de France*, 2017, p. 7.

³⁰ Voir <https://gallica.bnf.fr/services/engine/search/advancedSearch/>. Les résultats de notre requête en format tabulé sont disponibles en annexes de cette partie, fichier [I 2 Resultats-fouille.xlsx](#).

Léon Carvalho, et d'un brouillon de l'introduction d'une édition d'*Orphée et Eurydice* de Gluck (qui fait partie des résultats de la requête) rédigée par Camille Saint-Saëns.

Quelques résultats sont également à écarter parmi les monographies imprimées. Certains sont des documents descriptifs, comme une lettre anonyme, intitulée *Lettre à M. le chevalier de M***, sur l'opéra d'Orphée* et publiée en 1774 chez M^{lle} Lamarche, ou le catalogue commercial *Cache-radiateur Orphée : l'une des créations les plus intéressantes et les plus pratiques dans l'art décoratif moderne*. D'autres semblent apparaître parmi les résultats à cause d'une approximation graphique mise en place dans l'implémentation du moteur de recherche, comme le roman *L'épopée serbe* de Léo d'Orfer, dont la présence est sans doute causée par la proximité entre le nom d'auteur, inclus dans le titre, et le nom d'Orphée. Enfin, conformément à nos choix initiaux, nous écartons les traductions françaises des œuvres étrangères, comme la tragédie lyrique italienne *Orphée* d'Ange Politien ou le recueil des poèmes grecs publié en 1838 par Ernest Falconnet. 47 entrées sont donc retenues à ce stade.

Parmi les documents restants, lesquels sont des réécritures du mythe d'Orphée ? Combien incluent l'épisode de la descente aux Enfers ? Pour la plupart, il n'est pas possible d'en juger sans porter un regard attentif sur chaque texte. Mais pour certains, le titre donne suffisamment d'indices. C'est notamment le cas de *L'Esprit d'Orphée ou de l'influence respective de la musique, de la morale et de la législation. Seconde étude ou dissertation* de Jean de Dieu d'Olivier (1802), du *Miracle d'Orphée : lettres de détenus* (Stock, 1928), ou encore du *Trésor d'Orphée, livre de tablature de luth* d'Antoine Francisque (1600) dont le sous-titre indique qu'il s'agit d'un recueil de « plusieurs fantaisies [et] préludes passé[s] », au sein duquel aucun titre n'évoque le mythe orphique.

Ces quelques exemples montrent la complexité du traitement de la figure orphique et de son mythe. Les auteurs de l'ouvrage *Les Visages d'Orphée*³¹ distinguent trois fonctions de ce personnage mythique :

- celle de l'amoureux, l'exemple par excellence de la fidélité conjugale ;
- celle du poète, chanteur prodigieux capable d'enchanter les animaux, les arbres, et même les rochers, et qui apaise les âmes tourmentées dans les Enfers ;
- et celle du législateur, inventeur de la lyre à neuf cordes, fondateur des cultes, initié aux divers mystères.

Liées, concomitantes, ou successives, ces trois fonctions sont susceptibles de se retrouver dans les textes de notre corpus. En revanche, seule la réalisation de la fonction de l'amoureux garantit, *a priori*, la présence d'au moins une mention de l'épisode de la perte d'Eurydice.

³¹ Annick Béague et al., *Les visages d'Orphée*, Villeneuve d'Ascq, Presses universitaires du Septentrion, coll. « Savoirs Mieux », 2019.

Les deux autres fonctions, quant à elles, sont liées avec le débat, aussi antique qu'animé, autour de l'existence d'Orphée. On se demande par exemple si son nom dérive d'une tradition philosophique et théologique, connue sous l'appellation de l'orphisme ? Ou est-ce plutôt l'individu qui donne son nom au culte qu'il fonde ? Des questions se posent également quant à son œuvre. A-t-il écrit les *Argonautiques* et les *Hymnes orphiques* ? Dans *De natura deorum* de Cicéron, nous lisons : « Aristote nous apprend que le poète Orphée n'a jamais existé et le poème orphique que nous connaissons est, selon les phrygiens, l'œuvre d'un certain Cercops. Pourtant Orphée, c'est-à-dire selon vous son image, se présente souvent à mon esprit » (I, 107³²). Pour Cicéron, et pour Aristote, ce n'est donc qu'un mythe (*nunquam fuisse*, « n'a jamais existé »). En revanche, les dictionnaires du XVII^e au XIX^e s. s'accordent sur le fait qu'il s'agisse d'un personnage réel, dont l'existence se situe environ 100 ans avant la guerre de Troie (il serait donc, selon ces sources, contemporain à Hercule). Pour l'abbé Banier, le fait que le nom d'Orphée soit attesté par tous les philosophes, historiens et mythologues prouve son existence, quoiqu'il soit fort probable que la tradition ne condense plusieurs personnages en un³³. Enfin, au XIX^e s., dans *Les dieux et les héros, conte mythologique*, George Cox affirme qu'un homme nommé Orphée a existé, mais que sa vie a été enrichie par des récits symboliques.

Parmi les œuvres retenues à ce stade de notre exploration de Gallica, au moins quatre textes se focalisent sur l'aspect initiatique d'Orphée-personne réelle : *Essais de palingénésie sociale. Orphée* de Pierre-Simon de Ballanche (1827-1829), *Les grands initiés : esquisse de l'histoire secrète des religions* d'Édouard Schuré (1889), *Orphée, civilisateur de l'humanité* de Jean Coman (1939) et *Orphée et l'orphisme* de Reynal Sorel (1995). Au sujet du texte de Coman et de Sorel, nous apprenons dès l'introduction qu'il s'agit d'une étude analytique plutôt que d'une œuvre littéraire. Le statut des œuvres de Ballanche et de Schuré est plus ambigu. Dans les deux textes, la vie d'Orphée est narrée et la trame principale de son mythe y est présente, quoique réécrite (et réinterprétée) sous un angle spécifique et particulièrement métaphysique. Par conséquent, il est pertinent de les inclure dans notre corpus.

Progressivement, le nombre d'entrées est de plus en plus réduit. Or, l'exploitation plus détaillée des œuvres restantes nous montre que plusieurs d'entre elles ne sont pas en accès libre. Il s'agit notamment de neuf monographies publiées entre 1948 et 1999. Ces ouvrages, encore concernés par les droits d'auteur, mais non disponibles à l'achat en format papier, ont été numérisés dans le

³² « *Orpheum poetam docet Aristoteles numquam fuisse Cercopis ; at Orpheus, id est imago eius, ut nos uultis, in animum meum saepe incurrit.* ». Pour la traduction française, voir l'édition numérique de Clara Auvray-Assayas (ERAC EA 4705, université de Caen Normandie), <https://www.unicaen.fr/puc/sources/ciceron/accueil>.

³³ Voir *La mythologie et les fables expliquées par l'histoire*, tome VII, 1740. Nous appuyons cette brève synthèse principalement sur le chapitre « L'Orphée des savants » des *Visages d'Orphée* (Annick Béague et al., *Les visages d'Orphée*, op. cit., paragr. 67-71.).

cadre du projet de numérisation des indisponibles³⁴. Gallica permet de consulter 15 % des pages gratuitement à distance, et d'accéder à la numérisation complète à la BnF. Les livres sont aussi disponibles à l'achat en format numérique (en format image, PDF ou ePub), à un coût modéré (entre cinq et dix euros). Sur neuf ouvrages, six sont des romans et tous font partie du non-canon (comme en témoigne notamment leur inaccessibilité). Si nous décidons de nous engager dans la longue tâche de la numérisation de ces œuvres, qui ne pourront être mises à disposition qu'une fois entrées dans le domaine public, il faudra inclure au corpus d'autres romans de l'époque, et notamment ceux considérés comme canoniques : *La Nouvelle Eurydice* de Marguerite Yourcenar (1931), *Histoire d'Eurydice pendant la remontée* de Michèle Sarde (1979), etc. Le temps imparti pour notre recherche doctorale ne permet ni de constituer un sous-corpus des romans modernes et contemporains suffisamment riche, ni de s'adonner à l'analyse particulière qu'il mérite. Le projet est à considérer pour de futures études sur ce sujet.

Pour des raisons comparables, nous faisons le même choix concernant la poésie moderne et contemporaine, à une seule exception : celle d'inclure dans le corpus l'œuvre *Orphée, poème* de Jean Loyson (1929). Le premier paragraphe de la « Préface » indique : « Cela doit paraître incomparablement ridicule, en l'an de grâce 1929, de voir paraître un poème lyrique de cinquante pages, agrémenté de chœurs et de tout l'appareil désuet de la tragédie antique »³⁵. S'il est étonnant de retrouver l'esthétique classique dans la poésie de l'entre-deux-guerres, il serait intéressant d'examiner la proximité de ce texte avec, par exemple, les poèmes de Tristan l'Hermite. En outre, parmi les résultats de notre requête, nous trouvons un recueil de poèmes éponyme, écrit par Louis de Lau-nay et publié en 1922. Des recherches plus étendues nous révèlent un autre recueil *Orphée*, paru à Paris la même année et écrit par Fernand Divoire. Nous incluons les trois textes à notre corpus afin d'examiner d'éventuelles relations entre eux et d'observer leur place particulière au sein des réécritures du mythe. Par ailleurs, l'interrogation préliminaire de la performance des outils informatiques à l'aide de quelques poèmes relativement peu symboliques pourrait permettre l'ajout progressif des œuvres qui abordent le mythe de manière beaucoup plus allusive, métaphorique, voire ponctuelle, comme *Le Bestiaire* d'Apollinaire, « Orphée » de Victor Hugo et le poème éponyme de Paul Valéry, ou encore « El Desdichado » de Nerval.

Pour finir, nous trouvons également quelques documents difficilement exploitables, comme le *Répétiteur ballet d'Orphée*, présent dans le catalogue des Bibliothèques de Bordeaux, mais indisponible à la consultation numérique, ou *Orphée ballet meslé de recits pour servir d'intermedes a la tragedie d'Alexandre qui sera representée au College de Louis le Grand le deuxième jour d'aoust à une heure après midy*, dont l'un des

³⁴ Voir <https://web.archive.org/web/20230327143905/https://gallica.bnf.fr/edit/und/numerisation-des-livres-in-disponibles>.

³⁵ Jean Hyacinthe-Loyson, *Orphée, poème*, Paris, Nouvelle Librairie de France, 1948, p. 7.

auteurs présumés serait Joseph de Jouvancy (1690) et qui ne contient que des titres des scènes accompagnés de l'énumération des danseurs présents. Par ailleurs, le document *Par Privilège du Roy, l'esprit d'Orphée, machine nouvelle et surprenante, ou le Concert nouveau composé de voix et instruments de musique... par le sieur Cadet de Beaupré* présente une nouvelle machine. C'est une tête parlante, qui s'avère être, malgré le titre de l'ouvrage, celle du philosophe Démocrite. Elle est mise en scène dans un dialogue humoristique avec une personne réelle. Même si le nom de cette invention rappelle sans doute l'épisode de la mort d'Orphée, où sa tête, séparée du corps, continue à chanter son amour pour Eurydice³⁶, le dialogue ne fait aucune mention du mythe d'Orphée. Il n'est donc pas pertinent de l'intégrer dans notre corpus. Enfin, nous décidons également de ne pas inclure les œuvres cinématographiques, et écartons en fin de compte l'ébauche du scénario du film *Légende d'Orphée*.

Finalement, 22 œuvres seulement sont conservées à l'issue de cette requête initiale, soit environ un quart des résultats (Tableau 2³⁷).

Tableau 2. Les œuvres intégrées dans le corpus à l'issue de la première fouille de Gallica

Auteur	Titre	Date
Espine, Charles de l'	<i>La Descente d'Orphée aux Enfers</i>	1614
Puget de La Serre, Jean	<i>Les amours des dieux, de Cupidon et Psyché, du Soleil et Clytie, de Jupiter et Danaé, de Jupiter et Io, de Jupiter et Calisto, de Neptune et Amphitrite, avec celles d'Orphée & sa descente aux enfers</i>	1624
L'Hermite, Tristan	<i>L'Orphée [et Meslanges poétiques]</i>	1641
Buffequin, Denis	<i>Dessein du poème et des superbes machines du mariage d'Orphée et d'Eurydice</i>	1648
Chapoton, François	<i>La grande Journée des machines, ou Le mariage d'Orphée et d'Euridice</i>	1648
	<i>L'Orphée grotesque, avec le bal rustique. En vers burlesques</i>	1649
	<i>Suite de l'Orphée avec les Bacchantes, ou les rudes joyeuses, en vers burlesques</i>	1649
L'Hermite, Tristan	<i>Recueil de diverses poésies héroïques et burlesques. Contenant la Belle recluse, la Vieille layde, l'Amour honneste, le Doute amoureux, la Nuit amoureuse, l'Inquiétude amoureuse, les Soupirs de Silvie, Caprice burlesques, Orphée aux enfers, l'Aurore du bois de Vincennes et autres pièces curieuses</i>	1652
Du Boullay, Michel et Lully, Louis de	<i>Orphée, tragédie en musique</i>	1690
	<i>Le Ballet des machines représentant le V^e acte de la Mort d'Orphée et d'Euridice</i>	1700
Collande, Alexandre-Jacques-Pierre Le Gendre	<i>Opuscule d'un célèbre auteur égyptien. Contenant l'histoire d'Orphée, par laquelle on pourroit soupçonner qu'il est peu de femmes fideles</i>	1752
Hus, Jean-Baptiste	<i>La Mort d'Orphée ou les Fêtes de Bacchus, ballet héroïque</i>	1759
Moline, Pierre-Louis et Gluck, Christophe Willibald von	<i>Orphée et Euridice, drame héroïque en 3 actes</i>	1774

³⁶ D'ailleurs, Cocteau introduit la tête parlante d'Orphée dans les dernières scènes de sa pièce de théâtre *Orphée*. Toutefois, ce n'est pas une tête mécanique, mais celle de l'acteur, remplacée, lors des déplacements, par une tête en carton. Voir les scènes IX-XII et les notes de la mise en scène. Nous renvoyons à l'édition du scénario publiée chez Stock en 2005.

³⁷ Pour le tableau complet, incluant les liens d'accès et les informations détaillées, voir les annexes de cette partie, fichier [I 2 Corpus.xlsx](#).

Auteur	Titre	Date
Dorvigny et Moline, Pierre-Louis	<i>Roger-Bontems et Javotte, parodie d'Orphée et Eurydice ; pièce en un acte, mêlée d'ariettes</i>	1775
Ballanche, Pierre-Simon	<i>Essais de palingénésie sociale. Orphée</i>	1827-1829
Defontenay, Charlemagne Ischir	<i>Études dramatiques. I. Barkokébas. II. Le Vieux de la montagne. III. Orphée. IV. Prométhée</i>	1854
Crémieux, Hector et Offenbach, Jacques	<i>Orphée aux enfers : opéra-féerie en 4 actes et douze tableaux</i>	1847
Grandmougin, Charles	<i>Orphée : drame en 4 actes 16 tableaux</i>	1882
Schuré, Édouard	<i>Les grands initiés : esquisse de l'histoire secrète des religions : Rama, Krishna, Hermès, Moïse, Orphée, Pythagore, Platon, Jésus</i>	[1889]1921
Segalen, Victor	<i>Orphée-roi</i>	1921
Launay, Louis de	<i>Orphée : poème</i>	1922
Hyacinthe-Loyson, Jean	<i>Orphée, poème</i>	[1929]1948

Pour certaines de ces œuvres, nous ne nous appuyons pas sur la version mise à disposition sur Gallica. C'est notamment le cas des opéras de Gluck et d'Offenbach, pour lesquels nous privilégions la numérisation du libretto plutôt que celle de l'édition des partitions. La disparité des résultats et les proportions relativement basses, entre des documents peu pertinents et ceux qui correspondent à nos besoins, montrent la nécessité d'élargir la recherche à d'autres ressources numériques. Il nous paraît également important d'associer l'épluchage des archives à une enquête plus spécifique, portant sur des travaux menés sur les réécritures du mythe d'Orphée, afin d'améliorer la représentativité d'œuvres pouvant être considérées comme canoniques et pour bien identifier celles oubliées ou marginalisées.

c) Point d'arrivée : corpus de 53 réécritures

Tout d'abord, il serait erroné de considérer que l'exploitation initiale de Gallica a permis de collecter la totalité des réécritures en français du mythe d'Orphée et Eurydice. Par exemple, les cantates très connues, *Orphée* de Jean-Philippe Rameau (1721) et *La mort d'Orphée* d'Hector Berlioz (1827), manquent parmi les résultats. En effet, la restriction de la recherche aux documents uniquement en français a omis les ouvrages dont la langue n'est pas mentionnée dans la notice. C'est souvent le cas des partitions, principalement constituées d'annotations musicales, régulièrement manuscrites, et qui ne sont accompagnées de texte que de manière irrégulière. Ainsi, nous complétons notre première requête par une nouvelle, faite avec le même motif mais restreinte aux partitions, mais sans sélection de langue. Pour optimiser cette exploitation, nous excluons les partitions relatives aux opéras de Gluck et d'Offenbach, déjà recensées par la première requête.

Sur 22 ouvrages différents ainsi obtenus, trois sont déjà recensés, quatre ne comportent pas de paroles, quatre autres sont en langue étrangère et enfin, deux sont des traductions. Nous écartons aussi la chanson *Les canotiers de l'Orphée* écrite par Edouard Prevel, dont seul le nom du bateau fait référence au mythe. Les huit autres partitions sont des cantates et des cantatilles relatives au mythe.

Elles sont incluses dans notre corpus à l'issue de cette requête (Tableau 3). Cette recherche, plus ciblée et plus restreinte grâce à l'interrogation initiale de Gallica, s'avère beaucoup plus efficace (50 % d'œuvres incluses au corpus, 36,4 % en considérant uniquement de nouveaux ajouts).

Tableau 3. Les œuvres intégrées dans le corpus à l'issue de la fouille spécifique de Gallica (partitions)

Auteur	Titre	Date
Clérembault, Nicolas	<i>Orphée, cantate française, à voix seule avec symphonie</i>	1710
Lacroix	<i>Orphée</i>	1714-1730
Piroye, Charles	<i>Le retour d'Eurydice aux enfers, suite de la cantate d'Orphée</i>	1717
Rameau, Jean-Philippe	<i>Orphée</i>	1721
Grandval, Nicolas de	<i>Orphée. Cantate avec symphonie</i>	1729
Lemaire, Louis	<i>Orphée, cantatille nouvelle pour un dessus avec accompagnement de violons et flûtes</i>	1743
Guillion, Albert	<i>Orphée. Cantate composée à l'occasion de la fête de Henry Monton Berton</i>	1770-1799
Berlioz, Hector	<i>La mort D'Orphée. Monologue et Bacchanale à grands chœurs et à grand orchestre</i>	1828

Dans un deuxième temps, nous nous tournons également vers le corpus Google Books, malgré les difficultés conceptuelles et épistémologiques liés à l'utilisation de cette ressource³⁸. L'exploitation préalable de Gallica fournit suffisamment d'informations pour repérer et éviter d'éventuelles erreurs d'indexation, et pour ne pas craindre d'omettre une partie trop importante d'œuvres. Le moteur de recherche de Google Books, plus avancé que ce que permet Google Ngram Viewer seul³⁹, rend possible la restriction de la recherche aux livres en français publiés avant 1920 et dont le titre comporte soit le nom d'Orphée, soit celui d'Eurydice. Toutefois, les résultats s'avèrent assez minces. Sur 57 entrées, nous trouvons beaucoup d'éditions différentes du même texte (neuf ouvrages relatifs à l'opéra de Gluck, cinq à celui d'Offenbach, etc.), 30 œuvres distinctes étant recensées au total. Parmi elles, 13 ont déjà été trouvées par les requêtes précédentes, six sont des ouvrages extralittéraires, deux ne sont pas disponibles à la consultation, et quatre autres sont des traductions. Nous trouvons également un texte étranger, quoique francophone, *Orphée aux champs-élysées : vaudeville en cinq actes* de l'auteur belge Auguste de Peellaert. Une entrée relève d'une erreur d'indexation : *Orphée, tragédie, en trois actes* est attribuée à Jean de la Fontaine, alors que c'est la pièce de Du Boullay et Lully, déjà recensée avec la première requête de Gallica. Enfin, nous écartons deux autres textes. Le premier, *L'Orphée sacré du paradis, qui, par les mélodieux accords de plusieurs préceptes moraux, sentences exquises et conceptions théologiques, enchante doucement les brutales affections du vice et en désabuse les esprits mondains* par Claude Girard (1627), est focalisé uniquement sur la fonction initiatique d'Orphée. Le texte anonyme *Dialogue entre Lulli, Rameau et Orphée, dans les Champs* (1774), traite de sa musique en le faisant dialoguer avec deux compositeurs, Lully et Rameau. Un texte seulement

³⁸ Voir I.1a) « Google Books et Google Ngram Viewer : hégémonie anglosaxonne », p. 32.

³⁹ Voir https://books.google.fr/advanced_book_search?hl=fr.

est retenu suite à cette interrogation, *Le petit Orphée : opéra-comique en quatre actes et en vaudevilles* de Prosper-Didier Deshayes et de Jean Rouhier-Deschamps (1792, Tableau 4).

Tableau 4. L'œuvre intégrée dans le corpus à l'issue de la fouille du corpus Google Books

Auteur	Titre	Date
Deshayes, Prosper Didier et Rouhier-Deschamps, Jean	<i>Le petit Orphée : opéra-comique en quatre actes et en vaudevilles</i>	1792

Nous obtenons des résultats comparables en explorant Internet Archive⁴⁰, l'archive numérique privée états-unienne, dont le but principal est l'archivage du Web, mais qui recense également une quantité importante de numérisations (28 millions des livres au total, dont 332 582 en français). Nous y trouvons le poème *Orphée* de Fernand Divoire (1922, Tableau 5).

Tableau 5. L'œuvre intégrée dans le corpus à l'issue de la fouille du corpus Internet Archive

Auteur	Titre	Date
Divoire, Fernand	<i>Orphée</i>	1922

Contrairement au Gallica et de Google Books, Internet Archive permet de rechercher des mots-clés dans les notices des ouvrages. Cela implique que notre requête n'est pas restreinte à un champ bibliographique précis (auteur, titre, etc.). Cette requête étendue dévoile que la notice du deuxième volume des œuvres complètes de Jean-François de Regnard, présentes dans les résultats de notre requête, précise le contenu des six tomes de cette collection. Grâce à cela, nous réalisons qu'une œuvre intitulée *Orphée aux enfers* se trouve dans le quatrième volume. Ce dernier, absent de la collection d'Internet Archive, a été numérisé par Gallica, mais n'est pas apparu dans les résultats initiaux, car son titre est *Œuvres complètes de J.-F. Regnard*. Une fouille plus approfondie de l'archive de la BnF, combinée à la recherche des ouvrages recensés dans des travaux analytiques menés sur le mythe, nous ont permis d'ajouter à notre corpus six autres documents (Tableau 6).

Tableau 6. Les œuvres intégrées dans le corpus à l'issue de la fouille spécifique de Gallica

Auteur	Titre	Date
Tournes, Jean-François	« Eurydice morse du serpent », « Orphée aus Enfers », « Orphee Harpeur excellent », « Orphee mis en pieces » et « La langue & La Lyre d'Orphee mort, font pleints ». In <i>La métamorphose d'Ovide figurée</i>	1557
Ronsard, Pierre	« L'Orphée ». In <i>Les œuvres de Pierre Ronsard</i>	[1563] 1609
Regnard, Jean-François	<i>Orphée aux enfers, opéra</i> . In <i>Œuvres complètes de J.-F. Regnard</i>	1699
Courbois, Philippe et Fuzelier, Louis	« Orphée. Canate à voix seule et un violon ». In <i>Cantates françaises à I et II voix sans symphonie et avec symphonie</i>	1710
Esquiros, Alphonse et Esquiros, Adèle	<i>Histoire des amants célèbres dans l'antiquité</i> , chap. V	1849
Eimann, D. L.	« Une histoire renouvelée des Grecs ». In <i>Histoires cavalières</i>	1854

⁴⁰ Voir <https://archive.org/about/>.

Nous cherchons également à élargir notre corpus aux autres ouvrages des auteurs déjà recensés. Comme démontré par des recherches analytiques, Orphée est une figure récurrente chez certains auteurs majeurs de notre corpus, comme Pierre-Simon de Ballanche⁴¹ ou Victor Segalen⁴², le mythe se modulant et devenant une constante créatrice, parfois implicite, parfois centrale. Ballanche, par exemple, se réfère au mythe orphique pour prôner la force civilisatrice de la musique (*Du sentiment considéré dans ses rapports avec la littérature et les arts*, 1801), pour déplorer l'espoir perdu de la fidélité conjugale (*Sixième Fragment*, 1809) et pour conter les exploits d'un Orphée Argonaute (*Antigone*, 1819). Ces motifs, réunis, étoffés et enrichis, illustrent la doctrine sociale palingénésique de l'auteur dans l'épopée *Orphée* (1827), déjà recensée grâce à l'interrogation initiale de Gallica. Le chantre y est prophète de liberté et d'égalité, successeur de Prométhée et fondateur des cultes. Nous décidons d'ajouter à notre corpus uniquement *Sixième Fragment*, un bref essai qui réfère directement à l'épisode avec Eurydice, sans le parodier, ni le moderniser. Nous écartons toutefois *Du sentiment considéré dans...*, tout comme la nouvelle *Dans un monde sonore* de Segalen (1907), qui traitent le mythe de manière allusive ou très symbolique, et comme *L'Antigone*, car nous ne nous intéressons pas aux aventures d'Orphée et les Argonautes.

Une fouille ciblée de ressources numériques révèle qu'au moins trois autres auteurs déjà recensés ont repris le mythe d'Orphée et Eurydice dans plusieurs de leurs œuvres (Tableau 7). La pièce *Arlequin aux Enfers* de Jean-François de Regnard est incluse dans le cinquième tome des œuvres complètes de l'auteur. *Orphée ou Arlequin aux Enfers, divertissement* de Louis Fuzelier est compris dans *Scaramouche pedant*. Toutes les deux ont été mises à disposition sur Gallica et leurs éditions numériques ont été établies au sein du projet *Théâtre classique* dirigé par Paul Fièvre⁴³. Le texte de la cantate *Orphée* de Berlioz, présentée à la séance publique de l'Académie Royale en 1827, a été retranscrit dans le compte rendu de cette séance, publié chez Didot la même année et numérisé par Google Books. Nous verrons la pertinence de l'inclusion de cette version dans le corpus, en la comparant à la cantate *La mort d'Orphée* du même auteur, exécutée à l'École Royale de musique en 1828, mais aussi avec les textes des autres cantates présentes dans notre corpus. En revanche, faute d'accès au libretto, nous n'incluons pas la revisite berliozienne de l'opéra de Gluck de 1859, qui combine les

⁴¹ Concernant la présence du mythe orphique chez Ballanche, voir par exemple : Brian Juden, « Particularités du mythe d'Orphée chez Ballanche », *Cahiers de l'AIEF*, vol. 22, n° 1, 1970, p. 137-152 ; William R. Paulson, « Fragment et autobiographie dans l'œuvre de Ballanche : étude et textes inédits », *Nineteenth-Century French Studies*, vol. 15, n° 1/2, 1986, p. 14-32.

⁴² Eva Kushner, « Orphée et l'orphisme chez Victor Segalen », *loc. cit.*

⁴³ Voir <http://www.theatre-classique.fr/index.html> ; Christof Schöch, « Théâtre Classique – review », *RIDE*, n° 8, « Text Collections », 2018, en ligne, <<https://ride.i-d-e.de/issues/issue-8/theatre-classique/>>.

versions italienne (1732) et française (1774), et procède plus à une réinterprétation, qu'à une réécriture de l'œuvre⁴⁴.

Tableau 7. Des œuvres d'auteurs déjà présents dans le corpus qui traitent aussi du mythe d'Orphée (Google Books et le projet Théâtre classique)

Auteur	Titre	Date
Regnard, Jean-François	<i>La Descente d'Arlequin aux Enfers</i>	1689
Fuzelier, Louis	<i>Orphée ou Arlequin aux Enfers, divertissement</i>	1711
Ballanche, Pierre-Simon	<i>Sixième Fragment. In Œuvres de M. Ballanche : Antigone. L'homme sans nom. Élégie. Fragments.</i>	1809
Berlioz, Hector	<i>Orphée. Cantate. In Séance publique de l'Académie Royale des Beaux-Arts</i>	1827

Pour finir, nous complétons le corpus de quelques œuvres modernes et contemporaines, encore sous droits, mais dont l'inclusion est nécessaire, afin d'aboutir à une relative représentativité des réécritures recueillies (Tableau 8). Il s'agit notamment des modernisations particulières du mythe au théâtre de Jean Cocteau et de Jean Anouilh⁴⁵, ainsi que celles se focalisant sur le mythe de l'Orphée noir⁴⁶. Enfin, dans le prolongement des travaux d'Ute Heidmann⁴⁷, nous nous intéressons également à l'adaptation du mythe pour le jeune public.

Tableau 8. Les œuvres récentes et sous droits intégrées au corpus suite aux recherches empiriques

Auteur	Titre	Date
Cocteau, Jean	<i>Orphée. Tragédie en un acte et un intervalle</i>	1927
Anouilh, Jean	<i>Eurydice</i>	1941
Liking, Werewere	<i>Orphée-dafric, roman</i>	1981
Ma Njock, Manuna	<i>Orphée d'Afrique. Théâtre-rituel</i>	1981
Goudot, Marie	<i>Les mystères d'Orphée</i>	2003
Jimenes, Guy	<i>Orphée l'enchanteur</i>	2004
Pommaux, Yvan	<i>Orphée et la morsure du serpent</i>	2009
Montardre, Hélène	<i>Orphée aux Enfers</i>	2013
Laffon, Martine	<i>Orphée l'ensorceleur</i>	2016
Bottet, Béatrice et Pinel, Ariane	<i>Orphée et Eurydice. In Métamorphoses d'Ovide en BD</i>	2019
Ferry, Luc, Bruneau, Clotilde et Oddi, Diego	<i>Orphée et Eurydice. Précédé de Déméter et Perséphone</i>	2019

Nous recensons deux bandes dessinées et un album de jeunesse qui présentent le mythe au moyen de la combinaison du texte et des images, et visent, *a priori*, un public du même âge (entre 8 et 11 ans). Le nombre restreint des productions de ce type nous a permis d'aboutir à une relative exhaustivité, du moins en ce qui concerne le XXI^e s. Mais ce n'est pas le cas pour les romans, dont

⁴⁴ Voir par exemple Joël-Marie Fauquet, « Berlioz and Gluck », dans Peter Bloom (dir.), *The Cambridge Companion to Berlioz*, Cambridge, Cambridge University Press, 2000, p. 199-210.

⁴⁵ Ces œuvres canoniques mériteraient d'être complétées, par exemple, par *Orphée et son amour* de Népomucène Jonquille (1941), mais nous ne trouvons aucune ressource permettant d'établir le texte de cette pièce.

⁴⁶ Voir Jean-Paul Sartre, « Orphée noir », dans Léopold Sédar Senghor (dir.), *Anthologie de la nouvelle poésie nègre et malgache de langue française*, Paris, Presses Universitaires de France, 1948.

⁴⁷ Ute Heidmann, « Différencier au lieu d'universaliser. Comparer les façons de (r)écrire des mythes », *loc. cit.*

la production a été particulièrement fertile sur le marché français de la littérature de jeunesse, notamment depuis l'inclusion des *Métamorphoses* d'Ovide au programme de la sixième. Faute de pouvoir traiter la totalité des ouvrages existants, nous en choisissons quatre, dont l'analyse conjointe donne une vision claire des différentes tendances à observer dans les réécritures du mythe destinées à la jeunesse et parues au XXI^e s. Pour combler partiellement la subjectivité de nos choix, nous prenons comme point de départ le roman analysé par Heidmann, *Orphée l'enchanteur* de Jimenes (2004). Du point de vue synchronique, il est pertinent de comparer cette œuvre publiée chez Nathan avec celle parue un an auparavant chez un autre éditeur de jeunesse notoire, l'École des loisirs. Il s'agit des *Mystères d'Orphée* de Marie Goudot. Nous y ajoutons également une autre réécriture du mythe sous forme du roman, présente dans le catalogue de Nathan, *Orphée aux Enfers* d'Hélène Montandre, publié presque 10 ans plus tard. Pour finir, nous complétons le recensement par le roman le plus récent au moment de l'établissement du corpus, *Orphée l'ensorceleur* de Martine Laffon, paru chez Flammarion Jeunesse en 2016⁴⁸.

Conclusion du chapitre

Pour allier le quantitatif et l'herméneutique, Del Lungo et Vitali préconisent de réintroduire « une forme de subjectivité » et de « s'ouvrir plus amplement à la logique de l'hypertexte »⁴⁹. Nous suivons ces conseils en constituant notre corpus. Pour ce faire, nous partons initialement de l'interrogation quantitative de plusieurs archives numériques, que nous complétons par une fouille approfondie et plus subjective visant à combler les manques observés. Toutefois, en constituant notre sous-corpus des romans de jeunesse, nous choisissons de commencer plutôt par l'observation qualitative, que nous étendons avec des méthodes plus objectivables. À cinq reprises, nous laissons de côté notre restriction principale concernant la présence obligatoire des noms d'Orphée ou d'Eurydice dans le titre. Nous verrons que ces œuvres nous permettent des observations très particulières qui justifient cette entorse à la règle préétablie. Ces choix, comme nous espérons le démontrer par la suite, nous garantissent l'ouverture du corpus, préconisée par Del Lungo et Vitali, à la logique de l'hypertexte, qu'il soit entendu dans le sens genettien ou selon son acception informatique.

⁴⁸ Parmi les publications plus récentes, il convient de citer, par exemple, *Orphée dans la cité/Icare sur les toits* de Benjamin Desmares (Rouergue, 2020) qui a obtenu le prix Millepages Jeunesse en 2021, ou encore *Orphée à la recherche d'Eurydice* de Clémentine Beauvais (Nathan, 2021). La fréquence aussi élevée des nouvelles publications sur ce sujet prouve l'intérêt de l'inclusion de ce sous-corpus dans notre recherche.

⁴⁹ Andrea Del Lungo et Giovanni Pietro Vitali, « La littérature du XIX^e siècle au prisme des humanités numériques », *loc. cit.*, p. 74-75. Voir note 4.

Pour respecter les usages talistes et linguistiques, nous nous imposons aussi trois règles principales : de définir un protocole clair et objectivable de collecte du corpus, de veiller à ce que l'ensemble des réécritures retenues soit représentatif, à défaut d'être exhaustif, et, enfin, de garder un équilibre entre les œuvres canoniques et celles dépréciées ou oubliées. Comment déterminer si ces règles sont respectées ? Nous espérons avoir exposé avec clarté les principes de l'établissement de notre corpus. La spécificité des archives numériques explorées nous garantit également la présence du non-canon parmi les réécritures collectées. En effet, contrairement à nos inquiétudes initiales, celui-ci semble plus représenté que le canon, même si nous sommes très loin de 0,5 % annoncé par Moretti⁵⁰, car 13 œuvres seulement peuvent être considérées comme canoniques. Toutefois, les outils permettant de mesurer avec exactitude et de manière objective l'intérêt suscité par nos œuvres manquent. Nous pouvons observer leur nombre d'éditions ou la présence des travaux qui se consacrent à elles, mais il est regrettable que les archives numériques ne décomptent pas le nombre de consultations.

En revanche, la question de la représentativité mérite d'être approfondie par une observation plus détaillée de notre corpus. Il convient, d'une part, de nous intéresser aux informations concernant le genre ou l'époque de publication de nos œuvres, et, d'autre part, de nous pencher sur les relations, très hétérogènes, qui peuvent exister entre elles. Certaines de ces relations sont complexes, et invitent à chercher à relier plusieurs œuvres de notre corpus. Par exemple, l'opéra-comique *Le petit Orphée* de Deshayes et Rouhier-Deschamps (1792) parodie celui de Gluck, Calzabigi, et Moline (1774), lui-même influencé par les récits anciens, ceux de Virgile et d'Ovide, et par d'autres réécritures, comme sans doute l'opéra italien *Orfeo* de Monteverdi (1609). Il est alors difficile de conformer la pluralité des relations observées à l'approche binaire de l'hypotexte et l'hyper-texte proposée par Genette. Par ailleurs, est-il toujours possible de définir le texte source sans aucune ambiguïté ? Cette question mérite de s'y pencher, surtout dans le contexte des réécritures des mythes.

⁵⁰ Franco Moretti, « The Slaughterhouse of Literature », *loc. cit.*

3. Représentation de la variété des réécritures du mythe

En 1982, W. Nelson Francis utilise le terme de *représentativité* dans sa définition de *corpus*¹. Celle-ci devient alors un critère de constitution évoquée dans la majorité des travaux issus de la linguistique du corpus². En 1993, Douglas Biber précise que « la représentativité désigne la mesure dans laquelle un échantillon comprend toute la gamme de la variabilité d'une population »³. Trois ans plus tard, John Sinclair définit un corpus de référence comme « un ensemble suffisamment grand pour représenter toutes les variétés pertinentes de la langue et du vocabulaire qui les caractérisent »⁴. Ces ensembles sont généralement subdivisés (échantillonnées) selon des critères inhérents à chaque projet.

Même si les enjeux de la linguistique du corpus diffèrent de nôtres, certaines préoccupations, notamment celles concernant la difficulté à juger de la pertinence des méthodes de collecte et de la représentativité de l'ensemble recueilli, sont communes. Dans « Collection strategies and design decisions », Susan Hunston constate que l'établissement du corpus est « un domaine enclin aux paradoxes, où même les décisions les plus simples peuvent avoir de vastes ramifications »⁵. *A priori*, chaque corpus est constitué et décrit selon des critères spécifiques, comme le domaine (économie, histoire, littérature, etc.) ou le médium (livres, journaux, enregistrements sonores, etc.), ce qui permet d'effectuer des comparaisons entre différentes catégories et de s'assurer que les textes sélectionnés sont suffisamment variés. Quels que soient les choix pris lors de l'établissement de corps, selon Hunston, uniquement le contexte de création et les objectifs envisagés permettent d'évaluer leur pertinence⁶ :

¹ « I have previously defined the linguistic corpus as "a collection of texts assumed to be representative of a given language, dialect, or other subset of a language, to be used for linguistic analysis" » [J'ai précédemment défini le corpus linguistique comme une collection de textes censée être représentative d'une langue donnée, d'un dialecte ou d'un autre sous-ensemble du langage, et destinée à être utilisée pour l'analyse linguistique] (nous traduisons). Voir W. Nelson Francis, « Language corpora B.C. », dans Jan Svartvik (dir.), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991*, Berlin, New York, De Gruyter Mouton, 2011, p. 17, en ligne, <<https://www.degruyter.com/document/doi/10.1515/9783110867275.17/html>>.

² Voir le travail de synthèse de Najib Arbach et Saandia Ali, « Aspects théoriques et méthodologiques de la représentativité des corpus », *Corela. Cognition, représentation, langage*, n° HS-13, Université de Poitiers, décembre 2013, en ligne, <<https://journals.openedition.org/corela/3029>>.

³ « Representativeness refers to the extent to which a sample includes the full range of variability in any population » (nous traduisons). Voir Douglas Biber, « Representativeness in Corpus Design », *Literary and Linguistic Computing*, vol. 8, n° 4, janvier 1993, p. 243, en ligne, <<https://doi.org/10.1093/lc/8.4.243>>.

⁴ « [A reference corpus] aims to be large enough to represent all the relevant varieties of the language, and the characteristic vocabulary » (nous traduisons). John Sinclair, « Reference corpora », *EAGLES. Preliminary recommendations on Corpus Typology*, 1996, en ligne, <<http://www.ilc.cnr.it/EAGLES96/corpus/typ/node18.html>>.

⁵ « [T]his is an area prone to paradox, where even the apparently simplest decisions can have extensive ramifications » (nous traduisons). Voir Susan Hunston, « Collection strategies and design decisions », dans Anke Lüdeling et Merja Kytö (dir.), *Corpus Linguistics. An International Handbook*, Berlin, De Gruyter Mouton, 2008, p. 154, en ligne, <<https://www.degruyter.com/document/doi/10.1515/9783110211429.1.154/html>>.

⁶ « [A] corpus cannot be judged except in the context of its purpose », *Ibid.*, p. 155.

Si un corpus est compilé dans le but d'effectuer des recherches uniquement sur son propre contenu, il est facile de déterminer ce contenu. [...] Cependant, dans de nombreux cas, il est clairement impossible de regrouper au sein d'un corpus tous les textes qui font l'objet de la recherche. Même un sujet d'étude assez restreint [...] implique trop de textes pour constituer un corpus de tous les textes pertinents. À la place, un sous-ensemble des textes candidats possibles est choisi. Les enjeux de la création de corpus (*corpus design*) résident dans cette sélection.⁷

Finalement, les stratégies mises en œuvre lors de la création d'un corpus dépendent tout autant du type de corpus (spécialisé, général, etc.) que des raisons pour lesquelles il est créé. Par ailleurs, Hunston souligne que l'acte de création d'un corpus est, en grande partie, un compromis entre ce que l'on veut idéalement obtenir et ce qui est réellement possible face aux contraintes pratiques, techniques, et autres. De point de vue méthodologique, la représentativité fait partie des trois difficultés le plus fréquemment prises en compte lors de la création du corpus (à côté de l'équilibre et de la taille du corpus à concevoir).

Dans « New resources, or just better old ones? The Holy Grails of representativeness », Geoffrey Leech constate que malgré l'importance accordée à la représentativité, peu de précisions ont été apportées quant à la manière de déterminer si celle-ci est respectée. L'existence d'un corpus entièrement représentatif d'une langue est également mise en doute. Najib Arbach et Saandia Ali⁸ observent trois difficultés majeures. Tout d'abord, les stratégies de la collecte du corpus établies au sein du projet et leur documentation influent de manière importante sur l'appréciation de la représentativité. Ensuite, un corpus seul ne peut prétendre représenter le langage. Enfin, la transcription des données orales est une tâche longue et fastidieuse, nécessitant d'importantes ressources humaines, économiques et temporelles. Il en résulte que ces données sont généralement sous-représentées. Selon Leech, des difficultés particulières s'ajoutent dans le contexte numérique, au sein duquel certains types de discours sont plus présents que d'autres. Par exemple, parmi les données accessibles en ligne très peu sont manuscrites ou orales. Pourtant, avec l'avènement des réseaux sociaux, le discours familier, qui se rapproche des échanges informels et quotidiens, est davantage présent. Toutefois, le corpus ainsi constitué est-il réellement équilibré, notamment par rapport aux écrits académiques ou aux textes officiels, très représentés sur le Web ? Il est difficile d'en juger avec les outils disponibles actuellement. Comme le constate Leech :

⁷ « If a corpus is compiled in order to carry out research only on its own content, identifying what the content should be is straightforward. [...] In many cases, however, it is clearly impossible to pull all the texts that are the object of research into a corpus. Even a fairly restricted topic of study [...] involves too many texts for a corpus of all the relevant texts to be compiled. Instead of a sub-set of the possible candidate texts is selected, and in that selection lies corpus design. » *Ibid.*

⁸ Najib Arbach et Saandia Ali, « Aspects théoriques et méthodologiques de la représentativité des corpus », *loc. cit.*

Un moteur de recherche comme Google utilise des algorithmes qui sont totalement mystérieux pour l'utilisateur. Google n'offre rien de tel qu'une recherche complète sur le Web, et les rapports [...] montrent à quel point les chiffres que l'on obtient de Google sont instables, du moins à l'heure actuelle. Ce que nous obtenons est un énorme échantillon du Web, mais sa représentativité reste un mystère. Il semble y avoir consensus sur le fait que la fréquence des informations obtenues de Google est actuellement gravement trompeuse.⁹

Outre le cas spécifique et déjà discuté de Google, des biais similaires sont généralement relevés dans les réflexions autour de la création des corpus. Selon Hunston, « il n'est pas possible d'identifier une liste complète des "catégories" qui rendraient compte de manière exhaustive de tous les textes produits dans une langue donnée »¹⁰. Nous avons rencontré une difficulté semblable en explorant les archives numériques à la recherche des réécritures du mythe d'Orphée et Eurydice. Le fait d'avoir collecté (ou au moins recensé) la totalité des textes français numérisés au sein des archives ne garantit ni la représentativité ni l'exhaustivité de notre corpus.

Lors du premier consortium du projet *ELTeC*¹¹, une des préoccupations principales du groupe de travail concernait les questions de la conception du corpus littéraire international et plurilingue. Les difficultés de la représentativité soulevées rejoignent majoritairement les problèmes déjà évoqués par des travaux issus de la linguistique du corpus. De même qu'il n'est pas possible de représenter toute une langue au sein d'un seul corpus, de même, il n'existe pas de connaissances sur toute la littérature mondiale, sur chaque œuvre publiée, lue ou discutée sur un territoire donné et pendant une période étudiée. Finalement, il s'agit d'un choix : « souhaitons-nous utiliser chaque critère pour représenter la variété des valeurs possibles ou plutôt concevoir des échantillons permettant de représenter la distribution de ces valeurs entre chaque population ? »¹². Alors que, entre la variété (représenter tous les critères possibles d'une population) et la distribution (représenter la répartition statistique de la population à une moindre échelle), le projet *ELTeC* choisit plutôt cette dernière, nous verrons dans cette partie que, pour notre projet, nous optons plutôt pour la variété.

Dans son article, Gefen parle de l'« herméneutique intégrative », rendue possible grâce au numérique et qui « ne considère donc le texte littéraire comme un objet informationnel ordinaire et

⁹ « A search engine like Google employs algorithms which are totally mysterious to the user. Google provides nothing like a complete search of the Web, and reports [...] show how unstable and inconsistent are the counts that one gets from Google, at least at the present time. What we get is an enormous sample of the Web, but how representative it is remains a mystery. The consensus seems to be that frequency information obtained from Google is at present seriously misleading » (nous traduisons), voir Geoffrey Leech, « New resources, or just better old ones? The Holy Grail of representativeness », *Corpus Linguistics and the Web*, vol. 59, Brill, coll. « Language and Computers », 2007, p. 145, en ligne, <<https://brill.com/view/book/edcoll/9789401203791/B9789401203791-s009.xml>>.

¹⁰ « It is not possible to identify a complete list of categories that would exhaustively account for all the texts produces in a given language », Susan Hunston, « Collection strategies and design decisions », *op. cit.*, p. 161.

¹¹ *COST Action Distant Reading, Working group first expert meeting*, Prague, 2018. Toutes les ressources et informations produites dans le cadre du projet sont disponibles sur <https://distantreading.github.io/>. Nous nous appuyons sur le document *Sampling Criteria for ELTeC – Proposal*, COST Action Distant Reading, WG 1 expert meeting, Prague, 13 février 2018, en ligne, <https://distantreading.github.io/sampling_slides.pdf>.

¹² « Would we like to use each criterion, with the intention to represent the variety of possible values, or should the sample represent the distribution of those values across the population? » (nous traduisons), *Ibid.*, p. 15.

empirique analysable par une machine qu'au prix d'une sélection *a priori* délicate [...] et d'une décomposition structurelle et sémantique très complexe »¹³. Cela implique que le travail n'est pas terminé une fois les œuvres du corpus sélectionnées. Avant de procéder à une recherche particulière, soutenue par des outils informatiques, il faut se consacrer au travail de la décomposition structurelle et sémantique des textes recueillis. Il importe, d'une part, d'observer la diversité des œuvres ainsi regroupées et d'en établir le classement empirique et, d'autre part, d'appliquer des méthodes particulières de la collecte et de conception du corpus optimisées pour l'interrogation computationnelle.

Dans l'article « Literary Criticism and Literary Computing: The Difficulties of a Synthesis »¹⁴, Rosanne Potter affirme que « ce dont nous avons besoin, c'est d'une utilisation principale de la technologie et de la critique pour former un nouveau type d'étude littéraire absolument à l'aise avec les méthodes scientifiques tout en imprégnant complètement les valeurs des sciences humaines ». Plus de 30 ans après, ce besoin reste toujours d'actualité et son application dans les projets de recherche ne fait aucun doute. Simultanément, le corpus est, certes, destiné à l'exploitation par des outils informatiques, mais il est avant tout appelé à enrichir la recherche littéraire. Pour exposer la coexistence des valeurs littéraires et des usages numériques dans notre projet dès l'étape de la collecte du corpus, nous présentons ici une réflexion épistémologique concernant la description et la classification des textes de notre corpus.

a) Répartition chronologique

Les 53 œuvres de notre corpus peuvent être classées selon plusieurs critères¹⁵. Alors que, à première vue, certains de ces critères paraissent assez habituelles et ne semblent poser aucune difficulté particulière, nous verrons que l'assignation univoque d'une telle caractéristique à une telle œuvre n'est pas toujours aisée. Le premier critère qui vient assez naturellement à l'esprit concerne la répartition chronologique du corpus, comme celle que nous avons observée avec Google Books, Frantext et Gallica. La date de parution est, généralement, un critère objectif. Néanmoins, pour les manuscrits des deux cantates, celle de Lacroix et celle de Guillion, nous ne disposons pas de dates d'édition fixes, mais d'une estimation sur la fourchette allant d'une quinzaine à une trentaine d'années. Par ailleurs, aucune date n'est renseignée pour le texte anonyme *Le Ballet des machines représentant le V^e*

¹³ Alexandre Gefen, « Les enjeux épistémologiques des humanités numériques », *loc. cit.*, parag. 6.

¹⁴ « *What we need is a principal use of technology and criticism to form a new kind of literary study absolutely comfortable with scientific methods yet completely suffused with the values of the humanities* » (nous traduisons), voir Rosanne G. Potter, « Literary Criticism and Literary Computing: The Difficulties of a Synthesis », *Computers and the Humanities*, vol. 22, n° 2, Springer, 1988, p. 91-92, en ligne, <<https://www.jstor.org/stable/30200105>>.

¹⁵ Pour le tableau détaillé des œuvres du corpus, voir le fichier [I 3 Corpus-détail.xlsx](#).

acte de la mort d'Orphée et d'Eurydice. Il peut s'agir de celui d'Auguste Hus¹⁶, *La Mort d'Orphée, ou les Fêtes de Bacchus* représenté en 1759, mais ce dernier est qualifié d'un ballet « héroïque » et non pas de ballet « des machines » et le contenu des deux documents présente relativement peu de similitudes. Il est en revanche très probable que ce ballet des machines est postérieur à la pièce *La grande Journée des machines, ou Le mariage d'Orphée et d'Euridice* de Chapoton (1648). En effet, dans l'ouvrage *Théâtres lyriques de Paris : l'académie impériale de musique* de François Henri Joseph Blaze (1855), nous lisons : « *Le Mariage d'Orphée et d'Eurydice, ou la grande journée des machines*, tel est le titre d'une tragédie de Chappoton, représentée, en 1640, par la troupe royale. Ce mélodrame, où le chant et la danse étaient unis au dialogue écrit en alexandrins, est le premier que l'on offrit aux Parisiens avec un grand appareil de mise en scène »¹⁷. Par ailleurs, Blaze mentionne plus loin que la pièce est reprise en 1648 sous le titre *La Grande Journée des Machines, ou la Descente d'Orphée aux enfers et sa mort par les Bacchantes*. Le ballet *La mort d'Orphée* étant également écrit en alexandrin, il est peut-être à mettre en lien avec cette reprise. L'interrogation computationnelle du corpus peut apporter des informations supplémentaires, et contribuer à mieux comprendre l'histoire de cette œuvre. Cependant, au moment de la collecte du corpus, la fourchette concernant la date de création est particulièrement vaste, allant de 1650 à 1750. Si nous optons pour la date médiane de [1700], elle est à considérer comme une borne très approximative et incertaine. Or, cette nuance ne peut pas être rendue visible sur le diagramme de la Figure 11 qui présente la répartition chronologique des œuvres et les proportions entre le canon (en orange) et le non-canon (en bleu). Les disproportions observables sur ce diagramme ne sont pas étonnantes.

Tout d'abord, la présence majoritaire des œuvres non canoniques est généralisée, ce qui est conforme aux attentes et correspond à la réalité éditoriale décrite par Moretti. Toutefois, pour les périodes peu fournies, les proportions tendent à être plus équilibrées, car elles résultent d'ajouts d'œuvres qui ont succédé à la fouille des archives numériques. Cela constitue une preuve du fait que notre stratégie d'établissement du corpus vise à représenter la variété et non pas la distribution.

Concernant la répartition chronologique, les écarts entre 1600-1649 et 1900-1949 sont assez équilibrés. Un pic est observé au début du XVIII^e s., après une baisse relative dans la deuxième

¹⁶ Sur la page de garde, nous lisons uniquement « de la composition de M^r Hus ». Dans le catalogue de la BnF, l'auteur est identifié comme Auguste Hus (père). Toutefois, dans le *Dictionnaire des danseurs : chorégraphes et maîtres de danse à Bruxelles de 1600 à 1830* de Jean-Philippe Van Aelbrouck, nous trouvons, à la page 254, « *Mort (la) d'Orphée, ou les Fêtes de Bacchus*, ballet héroïque, par Jean-Baptiste Hus. Représenté à Bruxelles le 9 mai 1762 (Première à la Comédie-Française de Paris le 6 juin 1759. Sources : Gazette des Pays-Bas ; Livret parisien (Paris, Bibliothèque Nationale) ». De nombreuses œuvres, principalement des traités historiques et philosophiques, sont signées par Auguste Hus (fils). L'œuvre *La fête de l'Hymen est celle de l'Amour* (1775) est attribué à Hus père dans le catalogue de la BnF, mais sur la page de titre, nous trouvons la précision « M. Hus, fils » (voir <https://catalogue.bnf.fr/rechercher.do?index=AUT3&numNotice=10473022&typeNotice=p>). Le nombre restreint de mentions tant d'Auguste Hus père, que de Jean-Baptiste Hus, voire du ballet *Le Mort d'Orphée* en général, rend difficile la détermination univoque de l'identité de l'auteur.

¹⁷ François Henri Joseph Blaze, *Théâtres lyriques de Paris : l'académie impériale de musique, histoire littéraire, musicale, chorégraphique, pittoresque, morale, critique, facétieuse, politique et galante de ce théâtre*, Paris, Castil-Blaze, 1855, p. 14.

moitié du XVII^e s. Le sous-corpus du XXI^e s. est particulièrement semblable à celui de 1600-1649 (7 œuvres, dont une canonique). Toutefois, il a été collecté avec des méthodes empiriques et le recensement des œuvres contemporaines est rendu plus aisé par leur meilleur référencement non seulement par la BnF ou Google Books, mais aussi les catalogues des bibliothèques et des librairies numériques.

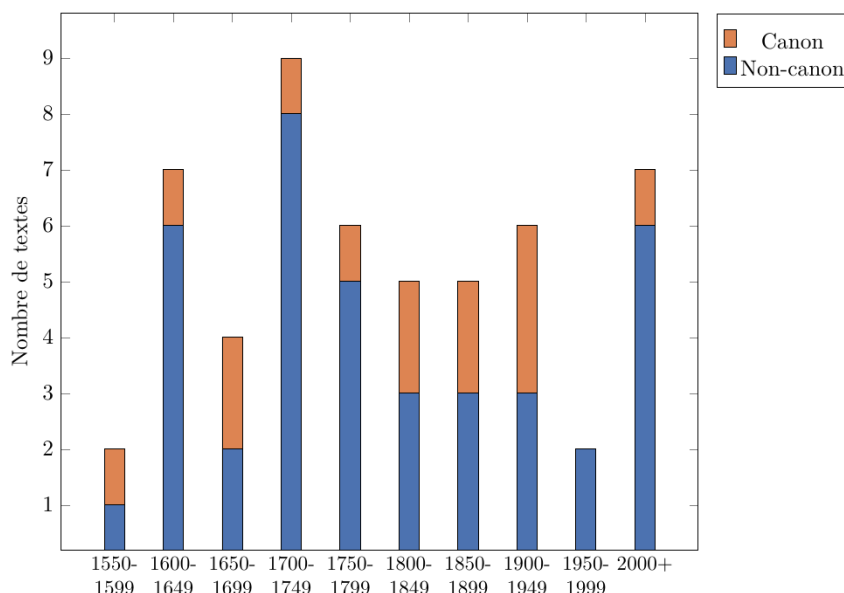


Figure 11. Répartition chronologique et proportions entre le canon et le non-canon

Finalement, l'observation des tendances semble pertinente uniquement entre 1600 et 1949. En effet, les tranches qui précèdent et qui suivent sont particulièrement peu fournies. Deux œuvres seulement ont été trouvées en 1550-1599 et en 1950-1999. Ce dernier résultat reflète notre choix d'écarter la majorité des œuvres sous droits et n'est pas analytiquement pertinent. En revanche, le fait qu'aussi peu d'œuvres aient été trouvées à la fin du XVI^e s. est un constat intéressant, d'autant que les deux seules œuvres recensées ont été ajoutées au corpus suite à une fouille approfondie de l'archive Gallica. La première, *La métamorphose d'Ovide figurée* contient cinq huitains relatifs au mythe d'Orphée et Eurydice¹⁸. La deuxième est l'élégie de Pierre de Ronsard, « Orphée », publiée pour la première fois en 1563 dans *Les quatre saisons de l'an*¹⁹. Ces minces résultats impliquent-ils que le

¹⁸ « Eurydice morse du serpent », « Orphée aus Enfers », « Orphée Harpeur excellent », « Orphée mis en pieces » et « La langue et La Lyre d'Orphee mort, font pleints ».

¹⁹ Émilie Seris, « Ronsard, imitateur de Politien ? », dans Perrine Galand (dir.), *Camēnae*, n° 22, « La réception d'Ange Politien en France au XVI^e siècle », décembre 2018, p. 10, en ligne, <<https://www.saprat.fr/toutes-les-revues-en-ligne-camenae/camenae-n-22-decembre-2018-la-reception-drange-politien-en-france-au-xvie-siecle-699.htm>>. Uniquement l'édition posthume est disponible sur Gallica. Initialement dédiée au Seigneur de Bray, la dédicace passe ultérieurement à Jacques-Auguste de Thou, historien et bibliophile. Quatre vers sont également ajoutés à la fin de l'élégie. Les autres variantes sont peu nombreuses et principalement d'ordre lexical.

mythe n'a pas intéressé les poètes de la Pléiade ? De nombreuses recherches sur ce sujet²⁰ prouvent le contraire, du moins pour Ronsard, Du Bellay et Baïf²¹, mais n'invite pas pour autant à enrichir le corpus.

En effet, s'il est vrai que ces poètes réfèrent régulièrement à la figure orphique dans des poèmes comme « À deux damoysselles »²², « Du jour des Bacchanales »²³ ou encore « L'Hymne de France »²⁴, ils ne procèdent que très rarement à la réécriture du mythe en tant que tel et ils renvoient principalement à la fonction de poète d'Orphée. L'élégie « Orphée » de Ronsard ou le sonnet « Que n'ai-je encor la harpe thracienne » de Du Bellay sont les rares poèmes qui le font. Toutefois, le sonnet de Du Bellay ne remplit pas nos critères de sélection : non seulement il ne comporte pas de titre, mais en plus, il ne mentionne directement ni Orphée ni Eurydice.

Peu de poèmes de cette époque portent des intitulés, ce qui constitue un des aspects problématiques de notre méthode de recensement. Nous aurions pu effectuer une fouille des archives numériques restreinte aux textes publiés avant 1600 et procéder à une recherche par occurrences. Cependant, très peu d'ouvrages concernés sont disponibles en format texte, et, par conséquent, interrogeables par les moteurs de recherche. Enfin, l'inclusion de chacun dans le corpus provoquerait une disparité, puisque nous avons décidé, lors de l'interrogation initiale de Gallica, d'écarter les poèmes qui abordent le mythe de manière allusive, métaphorique et ponctuelle, comme ceux d'Apollinaire, d'Hugo et de Valéry. Nous faisons donc le même choix pour les poèmes de la Pléiade, et nous ne retenons qu'« Orphée » de Ronsard.

b) Répartition générique

Le corpus peut également être partitionné en fonction du genre littéraire de chaque œuvre. Comme le constate Laurent Jenny, « nous avons besoin de savoir à quelle catégorie un texte appartient pour le comprendre tout à fait »²⁵. L'inscription générique est également un des quatre plans de la comparaison différentielle et discursive préconisée par Ute Heidmann²⁶. Selon la chercheuse,

²⁰ Voir par exemple Eva Kushner, « Le Personnage d'Orphée chez Ronsard », *Lumières de la Pléiade*, Paris, Vrin, coll. « Neuvième stage international d'études humanistes, Tours 1965 », 1966, p. 271-302; Françoise Joukovsky, *Orphée et ses disciples dans la poésie française et néolatine du XVI^e siècle*, op. cit.; Émilie Seris, « Ronsard, imitateur de Politien ? », loc. cit.

²¹ Françoise Joukovsky, *Orphée et ses disciples dans la poésie française et néolatine du XVI^e siècle*, op. cit., p. 7.

²² « Qui de la harpe Thracienne / A ouy la voix ancienne, / Des forestz l'ebahissement, / La votre luy fera pareille, / Qui fait des plus rudes oreilles, / Voyre des cœurs, ravissement » (Joachim du Bellay, « À deux damoysselles », *Œuvres complètes*, vol. 2, Champion, 2003, p. 63).

²³ « Celuy qui sceut les boys / Et les rochers attraire / Qui fist les trois aboys / Tous ebahiz se taire / Sceut au prix de sa teste / Combien est perilleux / Blamer la sainte feste / De ton nom merveilleux » (*Ibid*, p. 67-68).

²⁴ « Tu peuz tirer les forez de leur place, / Fleschir l'enfer, mouvoir les monts de Thrace, / Voire appaiser le feu, qu'il ne saccaige / Les verds cheveux d'un violé boucaige » (Pierre de Ronsard, « L'Hymne de France », *Œuvres complètes*, vol. 5, Paris, P. Jannet, 1857-1867, p. 283).

²⁵ Laurent Jenny, « Les genres littéraires », *Méthodes et problèmes*, Genève, Département de français moderne, 2003, en ligne, <<https://www.unige.ch/lettres/framo/enseignements/methodes/genres/index.html>>.

²⁶ Ute Heidmann, « Différencier au lieu d'universaliser. Comparer les façons de (r)écrire des mythes », loc. cit.

l'auteur (l'énonciateur) doit forcément s'inscrire dans des pratiques connues et reconnues de son époque, afin de pouvoir se faire entendre et comprendre par le public visé. Toutefois, ce qui est considéré comme genre littéraire reste relativement flou. Par exemple, les cinq textes de *La métamorphose d'Ovide figurée* sont des poèmes, et plus précisément, des huitains, présentés comme la traduction lyrique des extraits d'Ovide. Cependant, le titre insiste sur l'importance de l'aspect figuratif de l'œuvre. En effet, les miniatures ne sont pas uniquement là pour illustrer chaque poème. Au contraire, les deux éléments sont tout à fait complémentaires, comme cela se constate, en Figure 12, avec le poème « Orphée aus Enfers » et la miniature qui l'introduit.

Orphee avoit jà fort ploré la perte
De sa nouvelle espouse bien aymee,
Quand prend sa harpe, et d'une main
experte
Bien jointe avec sa voix vive animee,
Il vient jouer en la sale enfumee,
Du Roy d'Enfer, et de la Royne aussi :
Desquels obtient sa requeste estimee,
D'avoir sa femme : et cela, sous un si.



Figure 12. Le poème et la miniature d'« Orphée aus Enfers », dans Aneau, Barthelemy, et al., *La métamorphose d'Ovide figurée*. Lyon : Jan de Tournes, 1557, f° 61v.

Alors que le huitain aborde la catabase d'Orphée et son plaidoyer devant Pluton et Proserpine de manière très générale, la miniature, elle, apporte davantage des détails, notamment en second plan : nous y voyons les âmes damnées, comme Ixion, Sisyphe ou les Danaïdes. Leurs représentations renvoient à l'épisode de l'apaisement des peines des mortels (punis dans le Tartare) grâce à la musique d'Orphée, qui précède le plaidoyer du poète devant les dieux infernaux.

Ainsi, en fonction de la convention discursive choisie, la répartition et la perception des genres sont différentes. Dans *Qu'est-ce qu'un genre littéraire ?*, Jean-Marie Schaeffer²⁷ distingue différents types de conventions discursives (constituantes, régulatrices et traditionnelles) qui peuvent toutes se retrouver au sein d'une œuvre donnée. Pour Schaeffer, la transgression de certaines conventions met en échec la réalisation du genre visé par l'auteur. Toutefois, une approche aussi normative s'applique mal notamment à la littérature moderne et contemporaine.

²⁷ Jean-Marie Schaeffer, *Qu'est-ce qu'un genre littéraire ?*, Paris, Seuil, coll. « Poétique », 1989.

Dans *Le Texte littéraire. Pour une approche interdisciplinaire*, Ute Heidmann et Jean-Michel Adam²⁸ considèrent que chaque texte participe à plusieurs genres, ce qui empêche toute tentative de classement : « un texte relevant généralement de plusieurs genres, il ne s'agit plus de le classer dans une catégorie – son appartenance –, mais d'observer les potentialités génériques qui le traversent »²⁹. Heidmann et Adam distinguent la généricité auctoriale, éditoriale et lectoriale³⁰. Alors que cette dernière dépend des grilles interprétatives instables, appliquées par le lecteur dans des situations discursives différentes, la généricité auctoriale est plus stable : « [t]out au long de l'histoire de sa production et de sa médiation matérielle [...], un texte subit des modifications autoriales et est parfois accompagné des commentaires autoriaux qui affectent ce que nous pouvons appeler [...] son régime de généricité auctoriale »³¹. Ainsi, certains auteurs procèdent plutôt à une reconfiguration générique³², ce qui rend difficile l'attribution de l'œuvre à un genre canonique particulier de manière univoque. C'est notamment le cas d'*Eurydice* de Jean Anouilh. La présence de répliques et de didascalies, ainsi que la répartition en actes sont des conventions permettant d'identifier une œuvre dramatique. Le manque des directives musicales montre qu'il ne s'agit pas d'un opéra, mais plutôt d'une pièce de théâtre. En revanche, aucune convention discursive ni aucun indice auctorial ne permettent de préciser s'il s'agit d'une tragédie, d'une comédie ou autre.

D'autres auteurs utilisent régulièrement les paratextes pour préciser le genre visé. Parfois, l'indice est inclus dans le titre, comme dans *Le petit Orphée : opéra-comique* de Deshayes et Rouhier-Deschamps. Toutefois, ces informations peuvent être très générales (*Orphée-dafnic, roman* de Liking, *Orphée : poème* de Launay) ou, au contraire, très spécifiques (*Orphée aux enfers : opéra bouffon en deux actes et quatre tableaux* de Crémieux et Offenbach). Pour d'autres, il s'agit d'un jeu (voire d'un piège) conventionnel, comme pour le titre, étonnement classique, de la pièce moderne de Cocteau, *Orphée. Tragédie en un acte et un intervalle*. Pour d'autres œuvres, plusieurs indices (éditoriaux, autoriaux, formels) peuvent être trouvés, ce qui rend difficile la détermination univoque du genre visé par l'auteur. C'est par exemple le cas des œuvres de Ballanche dont les titres comportent la définition du genre (*Sixième Fragment*, *Essais de la palingénésie sociale*), mais les conventions régulatrices ne sont pas respectées : le *Sixième Fragment* est un texte assez long. Dans *Orphée*, la préface de plus de 130 pages prend effectivement les caractéristiques d'un essai. Le narrateur est homodiégétique, le ton est argumentatif, etc. En revanche, les conventions discursives d'*Orphée* sont davantage romanesques. Par ailleurs, les travaux analytiques portés sur l'œuvre tendent régulièrement à la qualifier d'épopée.

²⁸ Jean-Michel Adam et Ute Heidmann, « Des genres de discours à la généricité : six propositions », *Le texte littéraire : pour une approche interdisciplinaire*, Louvain-la-Neuve, Bruylant-Academia, coll. « Au cœur des textes », 2009, p. 11-23.

²⁹ *Ibid.*, p. 12.

³⁰ Jean-Michel Adam et Ute Heidmann, « Des genres de discours à la généricité : six propositions », *op. cit.*

³¹ *Ibid.*, p. 12.

³² Ute Heidmann, « Différencier au lieu d'universaliser. Comparer les façons de (r)écrire des mythes », *loc. cit.*, p. 19.

Finalement, les genres que nous avons déterminés de cette manière sont extrêmement hétérogènes (ballet des machines, ballet héroïque, comédie-ballet ; drame antique, drame héroïque, drame lyrique, étude dramatique...). Alors qu'il est important d'apprécier une telle richesse, nous avons également besoin des critères plus restreints et plus objectivables pour bien observer la composition de notre corpus.

Les réflexions menées sur la théorie et la classification des genres littéraires sont très riches³³. Le caractère normatif des premières classifications avec la différenciation platonicienne entre *diégèsis* et *mimèsis* se complexifie davantage chez Aristote, qui propose une grille typologique organisée selon les moyens, les objets et les modes de représentation. La poésie lyrique, ni narrative ni imitative, ne fait partie des genres littéraires qu'à partir de son inclusion, quelque peu artificielle, dans le système aristotélicien par Charles Batteux en 1746³⁴. Comme le constate Antoine Compagnon, « le système classique des genres faisait des modes de l'énonciation des archétypes génériques et des universaux poétiques. Il se situait hors de l'histoire dans une typologie abstraite et essentialiste. Avec le romantisme, on passe à des conceptions évolutionnistes et historiques des genres »³⁵. En effet, la libération de la typologie du système imitatif par les romantiques, les frères Schlegel d'abord, puis Höderlin et Schelling, aboutit à un système triadique des genres lyriques, épiques et dramatiques. La doxa contemporaine se construit à partir de cette typologique romantique, modale et thématique, et non pas de celle normative du système latin. Même si la triade roman-poésie-théâtre, en vigueur depuis le milieu du XIX^e s. environ, manque de théorisation et mélange des critères somme toute assez hétérogènes, « c'est une classification courante rendant compte de l'état présent de la distribution littéraire »³⁶.

Ainsi, nous proposons, outre les indices auctoriaux et éditoriaux, une classification plus générale, limitée à trois catégories : narrative, poétique et dramatique. Le diagramme de la Figure 13 montre que ces trois catégories sont représentées au sein du corpus de manière équilibrée.

Cela prouve que, pour ce critère, notre corpus est non seulement varié, mais aussi équilibré, ce qui est satisfaisant. En revanche, du point de vue de la théorie littéraire, la méthode reste problématique. Tout d'abord, nous constatons que l'approche triadique ne rend plus tout à fait compte de la distribution littéraire actuelle. Trois œuvres du corpus n'ont pas pu être inscrites dans aucune catégorie : il s'agit des bandes dessinées *Orphée et Eurydice* de Béatrice Bottet et d'Ariane Pinel (2019) et *Orphée et Eurydice. Précédé de Déméter et Perséphone* de Luc Ferry, Clotilde Bruneau et Diego Oddi

³³ Voir notamment Gérard Genette, *Introduction à l'architexte*, Paris, Seuil, coll. « Poétique », 1979 ; Antoine Compagnon, « La notion de genre », *Fabula*, 2001, en ligne, <<https://www.fabula.org/compagnon/genre.php>>.

³⁴ Charles Batteux, *Les Beaux arts réduits à un même principe*, Paris, Durand, 1746.

³⁵ Antoine Compagnon, « Septième leçon. Esthétique des genres : triade romantique », *La notion de genre*, Fabula, 2001, en ligne, <<https://www.fabula.org/compagnon/genre7.php>>.

³⁶ Antoine Compagnon, « Huitième leçon. Le système d'aujourd'hui », *La notion de genre*, Fabula, 2001, en ligne, <<https://www.fabula.org/compagnon/genre8.php>>.

(2019), et de l'album de jeunesse *Orphée et la morsure du serpent* d'Yvan Pommaux (2019). Par ailleurs, cette classification passe sous silence le caractère intermédial de certaines œuvres, comme les opéras et les cantates. Pour pallier ces manques, nous ajoutons une autre catégorie qui précise, pour chaque œuvre, si le texte est destiné à coexister avec d'autres médias (uniquement image ou son dans notre corpus). De ce point de vue, dans presque la moitié des œuvres du corpus (26/53), le texte est accompagné tantôt d'images (dont la fonction dépasse celle d'une simple illustration, 4/53), tantôt des partitions ou d'autres indications musicales (en plus des indications scéniques, dont la présence est déjà impliquée dans la catégorie « (genre) dramatique », 22 /53).

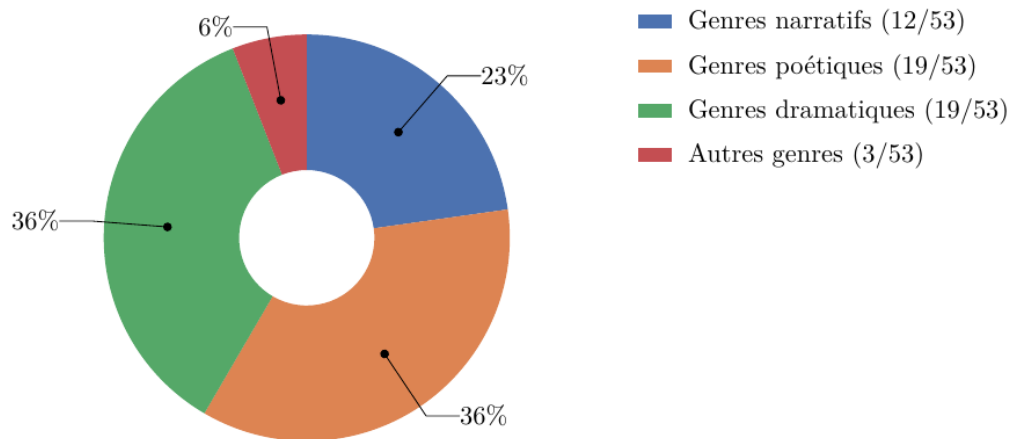


Figure 13. Représentativité des genres canoniques

Compagnon conclut son séminaire en disant :

Depuis Aristote, la notion de genre est de fait hétérogène : elle condense un faisceau de critères formels (comme le mode) et sémantiques (comme l'objet), suivant des combinaisons infinies. En pratique, l'analyse multicritère a toujours été dominante. On peut être tenté, comme Croce, suivant une démarche nominaliste et terroriste, de théoriser les œuvres modernes en dénonçant le concept de genre. Mais ce sont seulement les définitions *normatives* et *prescriptives* de genres qui sont dépassées, beaucoup moins les descriptions *formelles* et *sémantiques*. Et, au début du XXI^e siècle, les œuvres sont toujours identifiées *négativement*, par ce qu'elles ne sont pas, plutôt que par ce qu'elles sont : ainsi le système de genres reste pertinent, fut-ce modérément, même dans la violation généralisée aujourd'hui.

La difficulté à faire coopérer le formalisme inhérent à la collecte des corpus avec les postulats de la théorie littéraire est un problème qui ne doit pas être négligé. Nous avons besoin de critères objectifs et homogènes pour décrire et catégoriser notre corpus. Toutefois, celui-ci est très hétérogène et les réflexions issues de la critique et de la théorie littéraire ne permettent que très rarement d'aboutir à un classement stable et accepté par l'ensemble de la communauté.

Ainsi, nous proposons trois catégorisations liées à la caractérisation du point de vue du genre. La première, « genre », est très générale, appuyée sur un système triadique globalement reconnu et

qui bénéficie d'une longue histoire de théorisation et de critique. La deuxième, « sous-genre », s'appuie sur les indices éditoriaux et auctoriaux pour proposer un étiquetage relativement objectivable, mais peu cohérent et très hétérogène. Enfin, la troisième catégorie (« intermédialité ») ajoute des précisions concernant la coexistence, au sein de certaines œuvres, du texte et d'autres médias (musique et images dans le corpus, mais aussi, potentiellement, vidéo ou éléments interactifs comme hyperliens). Consciente des imperfections de cette méthode, nous l'utilisons alors en tant que *pharmakon*, en considérant que la multiplication des étiquettes incomplètes permet d'aboutir à une description satisfaisante.

c) Répartition fonctionnelle

Force est de constater que le dernier type de classement, visant à différencier les types des réécritures, est encore plus problématique. Ce phénomène fait l'objet d'une théorisation très complexe et hétérogène³⁷. Curieusement, malgré cette richesse, les tentatives de classement des relations pouvant être considérées comme celles de réécriture ont bien souvent opté pour les systèmes dichotomiques ou triadiques. Dans « La réécriture du texte littéraire. Mythe et Réécriture », Maurice Domino propose de différencier la « copie » (*rewriting*), qui vise la fidélité optimale au texte source, de la « réécriture » qui, elle, a pour objectif l'amélioration, « sa règle n'est pas conforme au texte premier ou au modèle prescrit par des modèles fixés, mais satisfaction d'une exigence virtuelle, réalisation d'un projet en train de s'élaborer »³⁸. Cette approche, très générale, est également fortement subjective, voire flottante. Il est légitime de se demander, par exemple, quel est le statut des traductions, dont la fidélité au texte source est variable ? Qu'en est-il des parodies, qui visent plutôt la dégradation que l'amélioration ?

La dichotomie de Anne-Claire Gignoux³⁹, quant à elle, s'appuie sur des facteurs plus objectifs, focalisés sur l'auctorialité. La chercheuse propose d'utiliser la variation orthographique pour distinguer deux phénomènes critiques différents : la *réécriture* génétique et la *réécriture* intertextuelle :

Il serait préférable d'utiliser le terme de « réécriture » avec un seul « é » pour désigner des transformations qui, même quand elles s'apparentent à celles de la réécriture génétique, en sont pourtant par essence différentes. [...] Notre réécriture, comme volonté manifeste d'un auteur de récrire le livre d'autrui ou de récrire un de ses propres livres déjà publiés, ou un de ses propres textes à l'intérieur d'un livre, n'est pas de la réécriture génétique.⁴⁰

³⁷ Anne Garcia, « La réécriture dans l'œuvre de José Emilio Pacheco. Une poétique du déjà-lu », *op. cit.*

³⁸ Maurice Domino, « La réécriture du texte littéraire Mythe et Réécriture », *Semen. Revue de sémio-linguistique des textes et discours*, n° 3, Presses universitaires de Franche-Comté, 1987, paragr. 5, en ligne, <<http://journals.openedition.org/semen/5383>>.

³⁹ Anne-Claire Gignoux, « De l'intertextualité à la réécriture », *Cahiers de Narratologie. Analyse et théorie narratives*, n° 13, REVEL, 2006, en ligne, <<http://journals.openedition.org.ressources-electroniques.univ-lille.fr/narratologie/329>>.

⁴⁰ *Ibid.*, paragr. 15.

Ainsi posée, la *réécriture* relève de la critique génétique, et s'intéresse à un ensemble de manuscrits, documents préparatoires et variantes successives d'un même texte chez un auteur en particulier. L'investigation menée sur la réécriture d'un texte permet alors de comprendre la genèse d'une œuvre particulière.

En revanche, la *réécriture*, comprise comme la « volonté manifeste d'un auteur de récrire le livre d'autrui ou de récrire un de ses propres livres déjà publiés, ou un de ses propres textes à l'intérieur d'un livre »⁴¹, peut être divisée en trois sous-catégories⁴². Dans la réécriture intratextuelle, le renvoi se fait à l'intérieur du même livre (passages récrits ou *leitmotif*). La réécriture macrotextuelle concerne toujours le même auteur, mais les relations sont construites entre différentes œuvres. Enfin, la réécriture intertextuelle amène l'auteur à puiser dans les œuvres d'autrui. Pour ce dernier type de réécriture, Gignoux propose trois sous-types spécifiques : la traduction, la citation indexée (référéncée) et la citation intégrée (cachée).

Cette typologie nécessite de déterminer l'auctorialité du texte source, ce qui implique que ce dernier doit être identifié sans équivoque. Or, contrairement à la classification générique, les indices auctoriaux sont beaucoup plus rares au sein de notre corpus. Un titre seulement contient la précision du type de réécriture, *Roger-Bontems et Javotte, parodie d'Orphée et Eurydice* de Dorvigny et Moline. Toutefois, cet indice n'est pas suffisant pour déterminer s'il s'agit de la parodie d'une œuvre particulière ou du mythe en général. Par ailleurs, il n'y a pas qu'une seule version dans les sources antiques. Au contraire, plusieurs versions sont dépendantes et apportent chacune des variantes particulières. Comme le constate Domino, « la dyade Mythe et Réécriture est pour le moins incongrue, voire monstrueuse »⁴³. En effet, pour arriver à réconcilier ces deux concepts, si incompatibles en apparence, il est nécessaire de concéder que la variance fait partie intégrale de chaque écriture (l'héritage littéraire médiévale en est un exemple par excellence) et d'accepter que l'identification univoque d'un texte source n'est pas toujours (voire n'est jamais) possible. Selon Domino, « [s]i on admet en effet que réécriture peut se dire alors même qu'un premier texte écrit n'est pas assignable, la dyade mythe et réécriture, moins monstrueuse qu'il ne paraissait d'abord se fait proposition épistémologique »⁴⁴. Reste qu'une classification relativement stable des relations aussi hétérogènes – souvent non linéaires, plus présumées qu'avérées –, doit prendre en compte des relations plus distendues et permettre une classification initiale sans exiger l'identification précise du texte source.

La deuxième difficulté concerne l'étendue du phénomène de réécriture. Alors que l'approche de Domino est davantage macrostructurelle et générale, et qu'elle concerne la totalité d'un ouvrage,

⁴¹ *Ibid.*

⁴² Anne Claire Gignoux, *Initiation à l'intertextualité*, Paris, Ellipses, coll. « Initiation à », 2005, p. 127-147.

⁴³ Maurice Domino, « La réécriture du texte littéraire Mythe et Réécriture », *loc. cit.*, paragr. 11.

⁴⁴ *Ibid.*, paragr. 17.

Gignoux propose une typologie davantage microstructurale ; sa réécriture peut porter sur des extraits et des fragments. Pour éviter la confusion notionnelle, nous choisissons de différencier le phénomène de réécriture (macrostructurel), sur lequel nous nous appuyons lors de la collecte et de la description de notre corpus, de celui de l'intertextualité (microstructurel), pour lequel nous proposons une investigation approfondie avec les outils et méthodes computationnelles⁴⁵. Par conséquent, notre perception de la réécriture s'approche davantage de ce que Genette qualifie d'hypertextualité. C'est déjà l'approche genettienne, qui, mise en commun avec la mythocritique de Brunel, nous a facilité l'identification des indices de réécriture pour délimiter nos critères de la collecte du corpus. Sa typologie exhaustive et clairement délimitée ne pourrait-elle pas être exploitée également pour caractériser les réécritures ?

Genette propose une typologie détaillée des pratiques hypertextuelles selon le régime (ludique, satirique ou sérieux) et la relation (de transformation ou d'imitation, voir Tableau 1). Cependant, l'assignation d'une des six relations ainsi identifiées à notre corpus montre que seulement deux catégories s'appliquent réellement : 10 transformations dans le régime ludique (parodies) et 43 transformations dans un régime sérieux (transpositions) sont identifiables. Une telle description n'est pas suffisante pour valoriser la richesse des réécritures observées. Fort heureusement, Genette propose une typologie spécifique de chaque catégorie principale. Par exemple, aux chapitres VIII-XI, le chercheur délimite six sous-types de transformations ludiques :

1. Parodie à transformation minimale (par exemple *Chapelain décoiffé* de Boileau qui est une paraphrase parodique du *Cid* de Racine) ;
2. Déformation de proverbes et renversement des aphorismes (par exemple *Cent cinquante-deux proverbes mis au goût du jour* de Paul Eluard et Benjamin Peret) ;
3. Calembour ou allusion parodique, notamment dans les titres (par exemple, *Lazarillo de Tormes* initie l'apparition d'une série des titres à la valeur d'indice générique : *Guzmán d'Alfarache*, *Moll Flanders*, *Gil Blas*, etc. ; l'émission radio *Le Masque et la Plume* devient *Le Casque et la Plume* dans l'ouvrage d'Olivier Kempf) ;
4. Lipogramme transformationnel (par exemple *Les Chats* de Baudelaire, réécrits sans *e* par Perec) ;
5. Transformation homophonique (*traducson*, par exemple du « Tombeau d'Edgar Poe » par Léon Robel⁴⁶) ;
6. Translation lexicale, qui consiste en un remplacement systématique de chaque substantif du texte par celui qui se trouve en vedette, dans un dictionnaire, d'un article qui précède ou suit ce substantif selon la distance convenue (traditionnellement, il s'agit du septième

⁴⁵ Nous abordons cette question en détail à partir de la partie suivante, « Intertextualité : de la théorie littéraire aux méthodes computationnelles », p. 122.

⁴⁶ Ainsi, les premiers vers du « Tombeau », « Tel qu'en Lui-même enfin l'éternité le change, / Le Poète suscite avec un glaive nu / Son siècle épouvanté de n'avoir pas connu » deviennent « Quelque ennui mène en vain le Termite et le Singe / L'appau est un suicide avec l'anglais venu / Sans socle, époux vanté donne à voir Pâques aux nues ».

article suivant, comme dans « La ciboule et la fourchette » de Nadirpher) ; et la permutation lexicale, issue principalement de la littérature combinatoire, comme *Cent mille milliards de poèmes* de Queneau.

Pour préciser de quel type de transformation ludique il s'agit, il est nécessaire, cette fois-ci, de bien identifier le texte source. Cela a été possible pour les dix œuvres de notre corpus classées comme parodies, mais seulement trois d'entre elles trouvent une place dans l'un des sous-types ainsi délimités. En effet, la cantate *Orphée* de Grandval (1729) est une paraphrase parodique de la cantate de Clérembault (1710), ce qui permet de la considérer sans aucune équivoque comme une parodie à transformation minimale. Les deux parties de la mazarinade anonyme, *L'Orphée grotesque, avec le bal rustique. En vers burlesques* et *Suite de l'Orphée avec les Bacchantes, ou les rudes joyeuses, en vers burlesques* reprennent, quant à elles, assez fidèlement la version ovidienne du mythe. Toutefois, même si certains extraits peuvent être considérés comme des paraphrases parodiques, les ajouts et modifications sont majoritaires. Pour quatre autres œuvres, le sens du mot « parodie » s'approche plutôt de la manière dont le mot est perçu dans la langue générale (« texte, ouvrage qui, à des fins satiriques ou comiques, imite en le tournant en ridicule, une partie ou la totalité d'une œuvre sérieuse connue »⁴⁷). Enfin, trois œuvres ont un statut particulier. « Une histoire renouvelée des Grecs » de D. L. Eimann place le mythe au sein d'un récit-cadre moderne qui oriente le mythe vers un sens complètement différent. *Orphée aux Enfers* de Crémieux et d'Offenbach et *Orphée* de Cocteau font intervenir de nouveaux personnages, pour ce dernier dans un décor tout à fait moderne. Ces caractéristiques correspondent aux cinq procédés les plus courants de la modernisation du mythe d'Orphée déterminés par Brunel⁴⁸. Pour mieux rendre compte de cette richesse, nous ajoutons à la description de notre corpus le critère « transformation ». Pour les transformations ludiques, trois sous-catégories sont retenues : « transformation minimale », « détournement humoristique », plus neutre que le terme « parodie », et « modernisation ». Cette dernière est également utilisée pour deux transformations sérieuses, les pièces *Eurydice* d'Anouilh et *Orphée-roi* de Segalen.

Par ailleurs, les transformations sérieuses, appelées « transpositions », regroupent un vaste champ des productions textuelles souvent marginalisées. Du chapitre XL au chapitre LXXX, Genette détaille plusieurs sous-types de transposition, dont la plupart peuvent être à leur tour partagés en sous-catégories⁴⁹ :

⁴⁷ Voir « Parodie », sens B.1., *Trésor de la Langue Française informatisé*. URL : <https://www.cnrtl.fr/definition/parodie>.

⁴⁸ Pierre Brunel, « Orphée moderne », *Apollinaire entre deux mondes. Le contrepoint mythique dans Alcools. Mythocritique II*, Paris, Presses Universitaires de France, coll. « Écriture », 1997, p. 18-19.

⁴⁹ Nous suivons le tableau général des pratiques hypertextuelles proposé par Nathalie Limat-Letellier, « Historique du concept d'intertextualité », *op. cit.*, paragr. 39. Toutefois, nous regroupons les transmodalisations inter- et intramodales (4.).

1. Traduction
2. Transstylistique (mise en vers, mise en prose, correction et autocorrection substitutives)
3. Transformation quantitative (réduction, augmentation, ajout, contamination, expansion)
4. Transmodalisation (dramatisation, narrativisation, transformation intramodale)
5. Transposition sémantique (diégétique, pragmatique, idéologique)

En outre, plusieurs cas de figure particuliers sont également observés, sans que Genette parvienne à les assigner à une des classes (par exemple, les hypertextes à hypotexte inconnu). Cette richesse de classement semble problématique : Limat-Letellier, par exemple, déplore que Genette « n'analyse pas dans quel contexte interviennent et quels effets de sens produisent ces pratiques fort dissemblables, tantôt utilitaires et tantôt créatrices, tantôt fidèles et tantôt infidèles à l'hypotexte »⁵⁰. En effet, pour appliquer cette typologie à notre corpus, quelques adaptations s'avèrent nécessaires.

Tout d'abord, nous reprenons la transposition diégétique en tant qu'une catégorie à part, « diégèse »⁵¹, applicable à toutes les œuvres du corpus. Deux étiquettes sont admises : les œuvres qui conservent le cadre mythique (Grèce antique, Enfers...) sont qualifiées d'« homodiégétiques » (39 textes), les œuvres qui transposent le mythe dans un autre décor sont « hétérodiégétiques » (14 textes). Parmi ces dernières se trouvent non seulement les modernisations, mais également les récits comme *Histoire des amants célèbres dans l'antiquité* d'Alphonse et Adèle Esquiros, qui remplace la catabase par un parcours initiatique dans un temple de l'ancienne Égypte, la parodie *Roger-Bontems et Javotte*, au sein de laquelle les personnages évoluent dans une forge, ou encore l'album de jeunesse *Orphée et la morsure du serpent* d'Yvan Pommaux qui inscrit le mythe dans un récit-cadre moderne.

Nous ajoutons également trois autres étiquettes à la catégorie « transformation ». Tout d'abord, les œuvres destinées à la jeunesse, qui emploient des procédés particuliers visant la préparation du texte pour le jeune public (omission de certains épisodes, comme les amours homosexuels d'Orphée, adoucissement des scènes brutales, comme le déchirement par les Bacchantes, etc.), sont étiquetées comme « transformations pragmatiques ». Ce faisant, nous détournons quelque peu l'emploi genettien, qui entend par transformations pragmatiques des pratiques correctives, souvent exercées par les traducteurs classiques. Toutefois, le chercheur mentionne les transpositions pour le jeune lecteur parmi les pratiques de transformation quantitative, comme l'*expurgation*⁵². Or, l'adaptation du mythe au jeune public implique régulièrement des modifications sémantiques majeures (par exemple pour aborder les thèmes de l'amitié, du deuil et de la perte). Par conséquent, si nous

⁵⁰ *Ibid.*, paragr. 38.

⁵¹ La diégèse est comprise, en suivant Genette, comme « l'univers où advient l'histoire », voir Gérard Genette, *Palimpsestes. La littérature au second degré*, op. cit., p. 419.

⁵² *Ibid.*, p. 330-331.

choisissons le qualifiant « pragmatique », c'est également pour son sens « pratique, voire utilitaire »⁵³.

Ensuite, les réécritures qui mettent en premier plan la fonction initiatique d'Orphée, en mettant son amour pour Eurydice en second plan, sont étiquetées comme des transformations idéologiques (9 textes). Genette évoque ce type de manière assez allusive, en mentionnant le *Vendredi* de Michel Tournier : « Ainsi, le *Vendredi* de Michel Tournier ressortit à la fois (entre autres) à la transformation thématique (retournement idéologique), à la *transvocalisation* (passage de la première à la troisième personne) et à la translation spatiale (passage de l'Atlantique au Pacifique) ; je l'évoquerai seulement, ou essentiellement, à propos de la première, qui est sans doute la plus importante »⁵⁴. Nous retrouvons une analyse plus détaillée de cette œuvre au chapitre LXXV, consacré à la *transvalorisation*⁵⁵. Cependant, la fonction initiatique d'Orphée n'est pas un ajout, mais une modification par rapport aux sources anciennes qui ne l'ignorent pas totalement, surtout dans la version virgilienne. Son importance est, pour la plupart, amplifiée, sans pour autant effacer complètement les fonctions du poète et de l'amoureux. Si nous privilégions l'étiquette « transformation idéologique », c'est, d'une part, pour ne pas impliquer obligatoirement un changement de valeur, et, d'autre part, pour insister sur un système d'idées particulier, propre à ces réécritures. Par exemple, quand *Orphée* de Ballanche et *Orphée-Dafric* de Liking se focalisent sur Orphée fondateur des cultes, les systèmes symboliques et idéologiques évoqués sont complètement différents, car Ballanche utilise le mythe pour illustrer sa doctrine relevant d'une philosophie métaphysique, la palingénésie sociale, alors que Liking expose les traditions africaines appuyées sur le rituel et l'initiation.

Enfin, les œuvres qui, au moins à la première vue, n'apportent pas de modifications sémantiques à aucune des versions antiques du mythe sont considérées comme des « transmodalisations » (25 œuvres). Nous verrons, lors de l'analyse détaillée du corpus, que les cas de transformations uniquement modales, sans modification d'aucun épisode du mythe, sont extrêmement rares. Dans certaines œuvres, comme le poème *Orphée* de Ronsard et la plupart des cantates, le narrateur devient autodiégétique. À partir du XVIII^e s., certains opéras et pièces de théâtre font intervenir l'Amour pour réunir les deux amants, malgré le non-respect de la condition imposée par les dieux infernaux. Dans d'autres œuvres encore, c'est Eurydice qui incite Orphée à se retourner. Certainement, l'étiquette « transmodalisation » n'est qu'une description superficielle. Toutefois, elle nous permet de limiter l'aspect interprétatif de la catégorisation initiale de notre corpus, susceptible d'introduire des biais et des présupposés qui peuvent influencer notre traitement avec les outils informatiques et

⁵³ Voir *s.v.* **pragmatique**, sens B. – 3.b), *Trésor de la langue française informatisé*. URL : <https://www.cnrtl.fr/definition/pragmatique>.

⁵⁴ Gérard Genette, *Palimpsestes. La littérature au second degré*, op. cit., p. 292.

⁵⁵ *Ibid.*, p. 514-524.

l'analyse qui en découle. En érigeant des frontières trop strictes, nous risquons de nous limiter à l'observation de sous-ensembles restreints plutôt que de porter un regard globalisant, plus apte à dévoiler des liens surprenants et inattendus.

Ainsi, trois catégories ont été introduites pour préciser la nature de chaque réécriture en suivant une adaptation du classement de Genette à la spécificité du corpus. La première catégorie, « type », est la plus générale et différencie les transpositions (transformations sérieuses) des parodies (transformations dans le régime ludique). La deuxième catégorie, « diégèse », exprime d'éventuels changements du cadre mythique, tel qu'il est établi dans les sources antiques. Enfin, six étiquettes particulières sont proposées pour la troisième catégorie, « transformation ». Le Tableau 9 montre la répartition du nombre d'œuvres du corpus au sein de ces trois catégories.

Tableau 9. La répartition des trois caractéristiques de réécriture dans le corpus

Type et diégèse Transformation	Parodie		Transposition		Total
	Hétéro-diégétique	Homo-diégétique	Hétéro-diégétique	Homo-diégétique	
Détournement humoristique	3	1	0	0	4
Modernisation	1	2	1	1	5
Parodie minimale	0	3	0	0	3
Transformation idéologique	0	0	7	2	9
Transformation pragmatique	0	0	1	6	7
Transmodalisation	0	0	1	24	25
Total	4	6	10	33	53

Tout d'abord, ce tableau dévoile qu'il y a peu de porosité des types de transformations entre le registre ludique et le registre sérieux (seulement les « modernisations » concernent les deux). Ensuite, d'une part, il n'y a pas d'équilibre quantitatif entre les parodies et les transpositions, et, d'autre part, les transformations du type « transmodalisation » sont beaucoup plus nombreuses que les autres. Enfin, d'une part, alors que les proportions entre les parodies hétérodiégétiques et les parodies homodiégétiques sont assez équilibrées, ce n'est pas le cas des transpositions (celles homodiégétiques sont trois fois plus nombreuses que celles hétérodiégétiques). Toutefois, ce déséquilibre, problématique de point de vue de la linguistique de corpus, est analytiquement important de point de vue littéraire, car il montre potentiellement les tendances saillantes dans le traitement du mythe à travers les époques. Par ailleurs, la combinaison de plusieurs catégories pour classer le contenu de notre corpus nous permet de formuler quelques hypothèses initiales, et de bien identifier les biais introduits par nos choix initiaux.

Nous observons, par exemple, que 16 transmodalisations sur 25 sont intramodales (œuvres poétiques). Toutefois, nous avons choisi d'écarter les poèmes symboliques de notre corpus. Par conséquent, il serait erroné de considérer que les poèmes restent généralement plus fidèles à la trame principale du mythe. En revanche, il est intéressant de constater, par exemple, que la majorité des transmodalisations ont été publiées avant le XVIII^e s. Ce type de transformation regroupe d'ailleurs presque 80 % d'œuvres musicales (17/22). Des tendances particulières sont également observables au sein des différents sous-ensembles. Par exemple, la comparaison de la date de publication et du type de transformation pour les 19 œuvres dramatiques montre la répartition en trois cycles des deux ou trois transmodalisations suivies d'un ou deux détournements humoristiques de 1614 jusqu'à 1792. Ensuite, à partir de 1854 jusqu'à 1981, uniquement des transformations idéologiques (trois textes, publiés en 1854, 1882 et 1981) et des modernisations (deux parodies, publiées en 1858 et trois transpositions, publiées en 1921, 1927 et 1941) sont relevées. Ces résultats concordent avec les travaux analytiques sur les réécritures du mythe d'Orphée et Eurydice au théâtre et à l'opéra⁵⁶. Enfin, il est aussi possible d'observer quelles œuvres échappent aux tendances ou semblent être pionnières. Par exemple, parmi les transformations idéologiques, nous retrouvons principalement des œuvres publiées à partir du XIX^e s., sauf *Opuscule d'un célèbre auteur égyptien* de Collande, publiée en 1752.

⁵⁶ Voir par exemple Jacqueline Bellas, « « Orphée » au XIX^e et au XX^e siècles », *loc. cit.*

Conclusion du chapitre

Certes, notre corpus n'est pas exhaustif. Il ne constitue pas non plus un échantillon équilibré et représentatif de la distribution des réécritures du mythe d'Orphée et Eurydice dans toute l'histoire littéraire française. En revanche, nous veillons à ce qu'il soit aussi varié que possible et proposons de multiples critères permettant d'apprécier cette variété. Si certains types, comme poèmes symboliques ou romans modernes, ne sont pas représentés, ces omissions sont réfléchies et argumentées. Comme le constate Hunston, « chaque corpus est un compromis entre ce qui est souhaitable [...] et ce qui est possible »⁵⁷. Les contraintes pratiques, techniques et méthodologiques sont nombreuses. Parmi elles se trouve la difficulté d'assigner des critères stables et objectifs de classement et de description des œuvres de notre corpus sans entraver les postulats des critiques littéraires. Par ailleurs, le peu de consensus dans la théorisation de ces phénomènes s'adapte mal à une expérimentation concrète et objectivable, appuyée sur un corpus spécifique.

Pour résoudre ces difficultés épistémologiques, nous diversifions et multiplions les critères permettant de décrire chaque œuvre. Nous concevons également une documentation détaillée qui, outre expliquer notre démarche, sert de garde-fou lors de l'étape analytique. En effet, c'est l'analyse du corpus en fonction des objectifs visés qui permettra de vérifier la validité des étiquettes assignées à chaque œuvre.

Dans leur article, Arbach et Ali constatent que « la représentativité ne saurait dépendre des critères linguistiques des données, mais est définie par les métadonnées du corpus, soit la documentation du corpus ; autrement dit, la représentativité d'un corpus ne saurait être appréciée sans que les critères de sélection des auteurs des textes, ainsi que les textes eux-mêmes soient clairement définis et disponibles »⁵⁸. Cependant, comme le souligne Hunston, même si la description détaillée et explicite du contenu de chaque corpus et des méthodes de collecte est indispensable, elle reste obligatoirement biaisée :

[L]a description [du corpus] ne peut être que partiellement explicite. En fait, les arguments selon lesquels un corpus particulier est représentatif, ou équilibré, sont inévitablement circulaires, en ce sens que les catégories que nous sommes invitées à observer sont des artefacts de la procédure de conception. Les catégories qui ont constitué la base de la conception du corpus sont, en effet, représentatives ou équilibrées, mais d'autres catégories peuvent être moins représentatives ou équilibrées, ou moins observables.⁵⁹

⁵⁷ « All corpora are a compromise between what is desirable, that is, what the corpus designer has planned, and what is possible » (nous traduisons). Voir Susan Hunston, « Collection strategies and design decisions », *op. cit.*, p. 156.

⁵⁸ Najib Arbach et Saandia Ali, « Aspects théoriques et méthodologiques de la représentativité des corpus », *loc. cit.*, paragr. 55.

⁵⁹ « The explicitness of description, however, can only be partial. In fact, arguments that a particular corpus is a representative, or balanced, are inevitably circular, in that the categories we are invited to observe are artefacts of the design

L'idée de représentativité *participative* apparaît pour faire face à ces problèmes. Arbach et Ali considèrent qu'« il ne s'agit plus de prétendre représenter le langage en un seul corpus, mais de participer à la représentation du langage par le corpus constitué, les différents corpus constitués indépendamment pourraient alors être regroupés ou consultés simultanément »⁶⁰.

Que ce soit « l'herméneutique intégrative » préconisée par Gefen, la « représentativité participative » des linguistes, ou encore « les principes *FAIR* » que prônent les humanités numériques, pour rendre le corpus (ré)exploitable, il est nécessaire de s'inscrire dans quelques usages partagés concernant la collecte, la numérisation et le dépôt des ressources produites au sein du projet. L'interopérabilité et l'adaptabilité à d'autres projets, comme l'accès ouvert aux données, motivent les choix d'annotation et de structuration. C'est ce que nous nous attachons à exposer dans le dernier chapitre de cette partie.

procedure. The categories that have formed the basis of the corpus design are, indeed, representative or balanced, but other categories may be less representative or balanced or less observable. », Susan Hunston, « Collection strategies and design decisions », *op. cit.*, p. 164.

⁶⁰ Najib Arbach et Saandia Ali, « Aspects théoriques et méthodologiques de la représentativité des corpus », *loc. cit.*, p. 58.

4. Numériser un corpus dans l'esprit des humanités numériques

Dans les chapitres précédents, nous nous sommes concentrée sur l'établissement d'un corpus aussi représentatif que possible des réécritures du mythe d'Orphée et Eurydice, sans nous préoccuper de son format. Or, nous avons besoin d'un corpus numérisé que nous pouvons interroger avec des outils numériques et éditer au sein d'une interface web. Pour ce faire, plusieurs étapes sont à accomplir.

Tout d'abord, nous devons établir le texte de chaque œuvre de notre corpus. La majorité de ces œuvres est disponible en format image, car il s'agit des numérisations de l'original à l'aide d'un scanner. Certains sont disponibles en format texte (généralement en plus du fac-similé), mais la fiabilité des transcriptions reste variable. Alors que nous pouvons généralement nous fier à celles établies au sein d'autres projets de recherche, celles issus des campagnes de transcription collaborative, disponibles par exemple sur Wikisource¹, demandent une relecture attentive. Parfois, il s'agit plutôt d'un texte issu d'une reconnaissance optique des caractères à partir d'un scan. La qualité de celle-ci est très variable et dépend principalement des logiciels et modèles utilisés, de la qualité de l'image et de la date de l'édition (les résultats seront bien meilleurs pour les éditions modernes que pour celles du XVII^e s.).

Un protocole d'édition rigoureux est nécessaire pour l'établissement du texte philologiquement exacte et respectueux de l'original. Mais alors que le texte seul peut être suffisant pour l'interrogation de notre corpus avec les outils numériques, pour pouvoir le visualiser ou en extraire des informations (concernant, par exemple, les références bibliographiques de l'œuvre, son classement au sein du corpus ou sa structure interne), nous devons également procéder à la structuration des textes en format XML (*eXtensible Markup Language*, ou langage de balisage extensible). Comme le constatent Benoît Epron et Marcello Vitali-Rosati dans leur ouvrage *L'édition à l'ère numérique*,

[l]e balisage XML, par l'introduction d'éléments sémantiques standardisables et manipulables de façon algorithmique, ouvre la voie à une forme nouvelle de construction éditoriale fondée sur la modularité. En combinant les formats numériques et le balisage sémantique porté par le XML, il devient en effet possible et nécessaire de redéfinir une unité de base pour la production, le traitement et la distribution éditoriaux. Dans l'édition sur support papier, cette unité de base est évidemment le livre : c'est à partir et autour de celui-ci que l'ensemble de l'écosystème éditorial se met en place. Avec l'édition numérique, il devient possible de modifier cette unité de base pour la ramener à une partie, à un chapitre ou à un élément précis, comme une figure ou une illustration.²

¹ Voir <https://fr.wikisource.org/wiki/Wikisource:Accueil>.

² Benoît Epron et Marcello Vitali-Rosati, *L'édition à l'ère numérique*, Paris, la Découverte, coll. « Repères », 2018, p. 62.

Ainsi, les auteurs distinguent deux dimensions de l'activité éditoriale dans l'environnement numérique : « la structuration fine et stable des ouvrages d'une collection » et « l'intégration de logiques de chaîne de publication multisupports »³.

De nombreux protocoles et guides existent pour faciliter la publication des corpus numériques. Certains d'entre eux sont créés au sein de groupes de travail et spécifiquement dédiés à la création des éditions numériques de qualité. Des chaînes éditoriales sont proposées pour faciliter le partage des pratiques et des outils prêt-à-porter existent pour automatiser certaines étapes (numérisation, encodage, transformation, etc.).

Dans ce chapitre, nous commençons par nous intéresser aux théories, modèles et méthodes de l'édition numérique dans le milieu universitaire. Nous présentons ensuite notre manière de nous inscrire dans cette réflexion commune en décrivant notre protocole de l'édition des textes du corpus (de l'établissement du texte à la transformation automatique en format XML-TEI, puis HTML) et en présentant le dépôt du corpus ainsi numérisé au sein de l'entrepôt Nakala⁴ dans le but de le préserver et le partager avec la communauté de recherche.

a) Éditions critiques numériques

Comme le souligne Joana Casenave dans *L'édition critique numérique : une nouvelle approche du patrimoine littéraire et documentaire*, « [l']édition critique numérique bénéficie d'un temps de recul suffisant pour dresser les premiers bilans des travaux édités sur le web »⁵. Les premières éditions critiques numériques datent d'il y a plus de 30 ans : Peter Robinson commence à travailler sur l'édition des *Canterbury Tales* de Chaucer en 1993. La première version (en CD-ROM) est disponible en 1996⁶. Au même moment, Jerome McGann débute le projet de l'édition des œuvres de Dante Rossetti au sein du projet *The Rossetti Archive*⁷. Les deux éditeurs sont régulièrement revenus sur ces projets éditoriaux précurseurs, tantôt pour bénéficier des nouvelles possibilités de publication (notamment l'arrivée du web), tantôt pour porter un regard critique sur leurs expériences⁸.

Dans *La « philologie numérique » : tentative de définition d'un nouvel objet éditorial*, Guillot et al. énumèrent certains avantages de l'édition numérique par rapport à celle papier. Les capacités de stockage

³ *Ibid.*

⁴ Voir <https://nakala.fr/about>.

⁵ Joana Casenave, *L'édition critique numérique : une nouvelle approche du patrimoine littéraire et documentaire*, Paris, Honoré Champion, coll. « Lettres numériques », 2023, p. 11.

⁶ Voir <http://canterburytalesproject.com/index.html>.

⁷ Voir <https://www.rossettiarchive.org/>.

⁸ Voir notamment Peter Robinson, « Editing Without Walls », *Literature Compass*, vol. 7, n° 2, 2010, p. 57-61, en ligne, <<https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1741-4113.2009.00676.x>> ; Jerome McGann, « Electronic Archives and Critical Editing », *Literature Compass*, vol. 7, n° 2, 2010, p. 37-42, en ligne, <<https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1741-4113.2009.00674.x>> ; Peter Robinson, « Toward a Theory of Digital Editions », *Variants*, vol. 10, 2012, p. 105-131.

sont accrues, les données sont plus facilement accessibles, le lecteur est plus impliqué par l'interactivité et le dynamisme des pages web. Les points jusqu'alors implicites, voire relativement flous dans les éditions traditionnelles sont clarifiés, car l'édition numérique repose sur la définition très précise des choix et procédures éditoriales. Par ailleurs, grâce aux principes d'ouverture, d'échange et de pérennité des données « l'éditeur numérique est, assez naturellement, amené à rechercher et à appliquer des standards et normes d'édition qui soient aussi larges et partagés que possible »⁹. Enfin, le numérique rend également possible la réalisation des éditions multi-facettes (simultanément normalisées, diplomatiques et fac-similaires), voire à de variantes multiples.

Néanmoins, dans *Pratiques de l'édition numérique*, Michael Sinatra et Marcello Vitali-Rosati soulignent que l'édition numérique dans l'environnement académique constitue un véritable défi :

[L]es choix techniques qui sont à la base de la structuration des contenus ne sont pas neutres et [...] témoignent et promeuvent des idées et des valeurs particulières. Pratiquer l'édition numérique signifie prendre en compte ce lien étroit entre la technique et la culture. Le risque du numérique est ce qu'on pourrait appeler un « déterminisme technologique » : nos pratiques et notre façon de penser pourraient finir par être déterminées par les outils.¹⁰

Ainsi, selon Sinatra et Vitali-Rosati, aux tâches habituelles de l'édition (choix du contenu, sa mise en forme, diffusion, etc.) s'ajoutent celles de l'éditorialisation, définie comme « l'ensemble des pratiques d'organisation et de structuration de contenus sur le web »¹¹. Il s'agit, entre autres, des questions de « [l]a structuration des contenus, leur organisation, la mise en place de dispositifs permettant leur validation et assurant leur visibilité ainsi que leur accessibilité »¹².

Afin d'adresser pleinement ces questions, il nous incombe de nous intéresser tout d'abord au champ de réflexion scientifique qui influence grandement les questions d'édition et d'éditorialisation des publications numériques, à savoir les humanités numériques (HN). En effet, la création d'éditions critiques numériques fait partie intégrante de la réflexion interdisciplinaire liée à la modification profonde de différentes disciplines des sciences humaines et sociales avec l'arrivée de l'informatique et de la publication web. Avant de proposer une tentative de définition de ce champ, il convient de présenter rapidement son évolution historique et les étapes-clés qui ont participé à son développement.

⁹ Céline Guillot et al., *La « philologie numérique » : tentative de définition d'un nouvel objet éditorial*, ATILF/SLR, 2017, p. 3.

¹⁰ Marcello Vitali-Rosati et Michael E. Sinatra, *Pratiques de l'édition numérique*, Montréal, Presses de l'Université de Montréal, coll. « Parcours numérique », 2014, p. 29.

¹¹ *Ibid.*, p. 32.

¹² *Ibid.*

*Du Literary and Linguistic Computing au Digital X*¹³

L'histoire des HN a été retracée notamment par Lou Burnard¹⁴, un de ses pionniers. Spécialisé dans l'encodage des textes numérisés et la modélisation des données, Burnard adopte, dans son étude diachronique, une approche centrée sur ses domaines d'expertise. L'histoire des HN peut être vue différemment si elle se focalise davantage sur d'autres champs des SHS, mais les trois périodes de développement distingués par Burnard semblent très fréquemment repris par les recherches portant sur l'édition numérique¹⁵.

La première période, née dans les années 1960 est appelée « informatique littéraire et linguistique » (*Literary and Linguistic Computing*). En 1951, Roberto de Busa, jésuite et linguiste italien, commence une collaboration avec IBM pour créer automatiquement un index d'œuvres de saint Thomas d'Aquin. Le projet aboutit à la création de *L'Index Thomisticus*, l'un des premiers projets de linguistique informatique. Disponible en ligne depuis 2005¹⁶, il a été initialement imprimé (en 56 volumes), puis transféré sur un CD-ROM (en 1989). Le fait que des informaticiens décident à se mettre au service des humanistes mène au développement de puissants outils en traitement automatique des langues (TAL) et de linguistique du corpus (pour produire des concordances, extraire des cooccurrences lexicales régulières et effectuer des traductions automatiques ou assistées par ordinateur). Dominée par des modèles quantitatifs et des mesures statistiques, cette période se caractérise également par la constitution des corpus de référence pour les langues dotées, comme l'anglais américain (Brown Corpus¹⁷) et britannique (British National Corpus¹⁸) ou le français (Frantext¹⁹).

La deuxième période, qui débute dans les années 1980, oriente les HN dans une direction différente : les « humanités computationnelles » (*Humanities Computing*) s'intéressent en particulier à la modélisation des données et à la conception des programmes informatiques. La TEI (*Text Encoding Initiative*) est élaborée en 1987²⁰ (et régulièrement amplifiée depuis) pour proposer un standard de structuration et d'encodage des textes numérisés. Le métalangage commun ainsi créé, le XML-TEI²¹, vise à faciliter l'interopérabilité des données, peu importe le champ dont elles sont issues. Il

¹³ Dans les deux sections suivantes nous reprenons partiellement l'introduction de notre article « Focus sur les traitements linguistiques d'un projet d'humanités numériques. Édition et cartographie hypertextuelle de réécritures », *Essais*, vol. 21, « Linguistique et interdisciplinarité », 2024, en ligne, <<https://journals.openedition.org/essais/13158>>.

¹⁴ Lou Burnard, « Du literary and linguistic computing aux digital humanities », *op. cit.*

¹⁵ Voir par exemple Joana Casenave, *L'édition critique numérique*, *op. cit.*; Marcello Vitali-Rosati et Michael E. Sinatra, *Pratiques de l'édition numérique*, *op. cit.*

¹⁶ Voir <http://www.corpusthomisticum.org/wintrofr.html>.

¹⁷ Voir <https://www.sketchengine.eu/brown-corpus/>.

¹⁸ Voir <http://www.natcorp.ox.ac.uk/>.

¹⁹ Voir <https://www.frantext.fr/>.

²⁰ Le langage de balisage alors employé était le SGML (Standard Generalized Markup Language ; norme ISO 8879:1986), le prédécesseur du XML. Les recommandations de la TEI ont ensuite rapidement été converties en XML.

²¹ Voir <https://tei-c.org/>.

permet de modéliser tant des œuvres littéraires, comme les pièces de théâtre²², que des dictionnaires²³, des corpus oraux²⁴ et d'autres données linguistiques²⁵. Le fait que ce standard soit toujours largement employé et continue d'évoluer prouve le besoin d'un dispositif permettant de faire un pont entre les disciplines, mais dévoile aussi que les préoccupations majeures des HN sont à ce stade orientées vers la numérisation, l'archivage et la diffusion.

La troisième période, les « humanités numériques » (*Digital Humanities*) proprement dites, débute avec le Web (1994). L'apparition des bibliothèques numériques, de sources indexées par des non-spécialistes et la naissance d'objets d'étude nativement numériques font basculer ce champ de l'informatique (outil qui sert à la recherche) vers le numérique (environnement au sein duquel celle-ci s'effectue). Progressivement, le mouvement qui part de la linguistique et des lettres s'étend vers d'autres disciplines et se rapproche de celles qui mobilisent des moyens informatiques depuis longtemps, comme la géographie. L'arrivée d'Internet ajoute une dimension communicationnelle, notamment avec le développement des médias sociaux et le basculement en ligne des corpus numériques. Il est alors indispensable d'adresser la question de la diffusion auprès d'un public plus large, de repenser le travail éditorial et la conception des interfaces ou encore de poser un cadre juridique et économique de cette nouvelle réalité éditoriale. C'est également à ce moment-là que commence à s'articuler la collaboration entre des chercheurs et des institutions patrimoniales, bibliothèques et musées au sein des projets numériques. Enfin, l'apparition des données nativement numériques (*born digital*) incite à réfléchir aux protocoles de leur exploitation et préservation. Cette troisième période marque donc le début de la théorisation et de la recherche épistémologique portant sur le numérique, l'outil de travail devenu l'objet central de la recherche.

La dynamique des HN proposée par Burnard peut être complétée, selon Pierazzo²⁶, par une quatrième période (l'actuelle), *Digital X*, le X étant une variable que chacun peut remplacer par le nom de son propre champ d'étude, ce qui exprime la multiplicité croissante des méthodes et approches au sein des HN au point que, paradoxalement, dans ce domaine, on pourrait presque considérer que seul chaque chercheur peut lui-même dire s'il contribue aux HN. Le nom de cette dernière période évoque également un certain cloisonnement disciplinaire des HN, contraire à la dimension collaborative du champ.

²² Voir <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DR.html>.

²³ Voir <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html>.

²⁴ Voir <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/TS.html>.

²⁵ Voir <https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/CC.html>.

²⁶ Elena Pierazzo, « Les Humanités Numériques : une perspective historique et sociale », *Formation 2020 des GIS aux Humanités Numériques*, 9 juin 2020, <<http://www.gis-reseau-asie.org/fr/formation-2020-des-gis-aux-humanites-numeriques-les-videos-des-interventions-sont-en-ligne>>.

En effet, Pierazzo observe deux modèles d'intégration des chercheurs HN au sein de leur domaine de rattachement principal²⁷. Dans le premier, les humanités voient le numérique comme un service, les chercheurs HN étant alors injustement considérés comme des prestataires, qui développent sans vraiment faire de recherche²⁸. Dans le deuxième modèle les HN sont acceptées comme étant une sous-discipline, mais les rapports avec la discipline de rattachement diffèrent. Dans certains domaines, comme l'archéologie, les méthodes numériques, devenues indispensables, sont employées naturellement par tous les chercheurs. Dans d'autres, comme la paléographie, les chercheurs traditionnels et numériques collaborent à pied d'égalité. Enfin, le troisième modèle est qualifié d'"antagoniste". En son sein, les recherches traditionnelles et informatisées ne se mélangent pas ou peu. Elles constituent pratiquement deux disciplines indépendantes et dont les objectifs diffèrent. C'est par exemple le cas de la linguistique générale et la linguistique computationnelle (le TAL)²⁹.

Dans notre projet, le dialogue interdisciplinaire est indispensable. Malgré un rattachement principal en littérature comparée, le projet que nous portons inclut également des questions linguistiques ou plus proches du domaine des sciences d'information et de communication. Il nous importe d'appartenir à une communauté qui accepte, voire promeut, cette ouverture disciplinaire et les nouvelles expérimentations appuyées sur les outils numériques. Cependant, si nous nous intéressons à la définition de ce champ, nous constatons sa complexité et les tensions qui persistent tant au sein de la communauté des chercheurs HN qu'entre elle et celle des chercheurs traditionnels.

Humanités numériques : définition difficile

L'application *What is Digital Humanities?*³⁰ recense plus de 800 définitions de la notion d'HN et en propose une aléatoirement à chaque chargement de la page. Pour Federico Meschini de l'université de Montfort, par exemple : « la meilleure manière [...] pour décrire les HN est [...] d'utiliser la citation de la *Vente à la criée du lot 49* de Thomas Pynchon : "les miracles sont bien les intrusions

²⁷ Elena Pierazzo, *The disciplinary impact of the digital: DH and « the Others »*, DHSI, University of Victoria, Canada, 2017.

²⁸ Concernant l'accueil mouvementé des chercheurs HN au sein de leurs domaines, voir par exemple Matthew Kirschenbaum, « What Is "Digital Humanities", and Why Are They Saying Such Terrible Things about It? », *differences*, vol. 25, n° 1, janvier 2014, p. 46-63, en ligne, <<https://mkirschenbaum.files.wordpress.com/2014/04/dh-terriblethingskirschenbaum.pdf>>.

²⁹ De nombreux articles étudient ce schisme entre la linguistique et le TAL. Par exemple, Karen Spärck Jones considère que « la linguistique computationnelle n'a pas besoin de la linguistique générale, non computationnelle, que ce soit pour lui accorder de la crédibilité intellectuelle ou pour en assurer le progrès. La linguistique computationnelle n'est pas seulement de la linguistique accompagnée de quelques extensions *geeks*, utiles dans la pratique, mais théoriquement non pertinentes et déconcertantes » (« [C]omputational linguistics does not *need* mainstream, non-computational linguistics, whether to supply intellectual credibility or to ensure progress. Computational linguistics is not just linguistics with some practically useful but theoretically irrelevant and obfuscating nerdy add-ons », nous traduisons). Voir Karen Spärck Jones, « Computational Linguistics: What About the Linguistics? », *Computational Linguistics*, vol. 33, n° 3, 2007, p. 437-441, en ligne, <<https://www.aclweb.org/anthology/J07-3008>>.

³⁰ Voir <https://whatisdigitalhumanities.com/>.

d'un autre monde dans le nôtre" »³¹. Si les définitions que nous pouvons découvrir ainsi sont très différentes, elles évoquent régulièrement l'aspect interdisciplinaire de ce domaine de recherche qui invite les chercheurs des sciences humaines et sociales à échanger avec les sciences informatiques et entre leurs domaines, même si ce dialogue interdisciplinaire est souvent complexe.

La définition d'un nouveau champ disciplinaire, qui est encore en train de se structurer et ne cesse pas d'évoluer, est une tâche ardue, mais nécessaire³². La première tentative date de 2004 avec la publication du *Compagnon to Digital Humanities*³³. Cet ouvrage a été réédité en 2016³⁴ dans une version très largement revue et commentée qui consacre toute une partie de la préface aux débats autour de la définition du champ. Les auteurs constatent en particulier qu'« il reste discutable si les HN devraient être considérées comme une "discipline à part entière", plutôt qu'un ensemble de méthodes connexes »³⁵.

Dans *Defining digital humanities*, Melissa Terras passe en revue des définitions qui ont été proposées entre 2009 et 2012. Elle constate ainsi la grande variabilité des points d'intérêt : alors que certaines des définitions recensées portent sur des méthodes, théories, outils ou résultats de recherche, d'autres se focalisent plutôt sur les questions communautaires. En effet, si certains considèrent que les HN consistent en l'application des outils et méthodes numériques aux méthodes de recherche en SHS³⁶, d'autres mettent en avant les enjeux collaboratifs et interdisciplinaires³⁷. Les réflexions épistémologiques se multiplient, car il paraît nécessaire « de produire l'appareil théorique pour interpréter les structures conceptuelles fondamentales qui caractérisent notre culture actuelle »³⁸.

Une certaine éthique communautaire et inclusive se dégage effectivement de la plupart des tentatives de définition. Lincoln Muller propose de parler d'un « spectre » des HN pour insister sur la mouvance de la frontière entre chercheurs traditionnels et "humanistes numériques". La métaphore du « grand chapiteau des humanités numériques » (*The Big Tent*)³⁹ est également évoquée pour souligner le fait que les HN regroupent des chercheurs de domaines différents, avec des objectifs et

³¹ « The best way I've found to describe what Digital Humanities is [...] using a quote from Thomas Pynchon's *Crying of Lot 49*: 'You know what a miracle is... another world's intrusion into this one' ». Nous traduisons.

³² Pour le relevé de différents éléments de définition des HN, nous nous appuyons sur l'excellent travail de Joana Casenave, *L'édition critique numérique*, *op. cit.*, p. 134-137.

³³ Susan Schreibman, Ray Siemens et John Unsworth (dir.), *A Companion to Digital Humanities*, Oxford, Blackwell, 2004.

³⁴ Susan Schreibman, Ray Siemens et John Unsworth (dir.), *A New Companion to Digital Humanities*, 1^{re} éd., Wiley, 2016.

³⁵ « It remains debatable whether digital humanities should be regarded as a "discipline in its own right", rather than a set of related methods » (nous traduisons). *Ibid.*, p. XVII.

³⁶ Voir par exemple Dominique Vinck, *Humanités numériques : la culture face aux nouvelles technologies*, 2e éd. revue et augmentée, Paris, le Cavalier bleu éditions, coll. « Idées reçues », 2020.

³⁷ Anne Burdick et al., *Digital Humanities*, Cambridge, Mass, MIT Press, 2012.

³⁸ Marcello Vitali-Rosati et Michael E. Sinatra, « Pratiques de l'édition numérique », *op. cit.*, p. 26.

³⁹ Voir par exemple Melissa Terras, « Un regard jeté sous le chapiteau : les humanités numériques et la crise de l'inclusion », in Pierre Mounier (éd.), *Read/Write Book 2 : Une introduction aux humanités numériques*, Marseille, OpenEdition Press, coll. « Read/Write Book », 2012, p. 89-97.

méthodes qui leur sont propres. La diversité et l'ouverture du champ sont parfois élevés au rang d'un combat éthique. Dans *Debates in the Digital Humanities*, par exemple, Lisa Spiro déclare que les chercheurs en HN doivent se battre pour les valeurs suivantes : l'ouverture, la collaboration, la collégialité, l'appartenance, la diversité et l'expérimentation⁴⁰.

Mise à part de nombreuses études et publications, déterminantes pour le développement du champ, d'autres structures et initiatives prouvent cette éthique du "grand chapiteau". Il convient notamment de mentionner l'inauguration en 2009 de la Journée des HN (*Day of DH*) par le réseau international des centres de recherche dédiés à cette discipline, *CenterNet*⁴¹. Chaque année, les humanistes numériques choisissent un jour pour partager leurs travaux, soit *via* des réseaux sociaux, soit grâce aux nombreux événements organisés ce jour-là au sein des centres partenaires. La journée de 2023, par exemple, a été coordonnée par le Centre de recherche interuniversitaire sur les humanités numériques⁴² de Québec sous le thème de de l'"humanisation des données" (*humanizing data*).

Plusieurs associations sont également consacrées aux HN, comme *Alliance for Digital Humanities Organization*⁴³ créé en 2005 ou la première association francophone, *Humanistica*⁴⁴, en activité depuis 2014. En France, l'Infrastructure de Recherche « étoile » (IR*) Huma-Num est depuis 2013 mise à disposition des chercheurs et chercheuses SHS. Elle a pour but « de développer, de réaliser et de préserver sur le long terme les programmes de recherche – leurs données et outils – dans un contexte de science ouverte et de partage des données »⁴⁵. Ce travail est majoritairement fait par le biais de consortiums, qui regroupent des communautés scientifiques, et grâce à la présence des agents d'IR* Huma-Num au sein de différentes Maisons des sciences de l'Homme.

Parmi les consortiums d'IR* Huma-Num, nous partageons nos intérêts de recherche en particulier avec Ariane (Analyses, Recherches, Intelligence Artificielle et Nouvelles Éditions numériques)⁴⁶, spécifiquement dédié à l'édition numérique, et avec CORLI (Corpus, Langues et Interaction)⁴⁷ qui travaille sur les corpus de langage. Nous bénéficions également de l'accompagnement de l'ingénieur d'études CNRS chargé du traitement des données scientifiques au sein de la MESHS de Lille Nord de France⁴⁸ qui assure le relais entre les chercheurs et les services Huma-Num. Ces

⁴⁰ *Openness, Collaboration, Collegiality, Connectedness, Diversity et Experimentation*. Voir Lisa Spiro, « "This Is Why We Fight": Defining the Values of the Digital Humanities », dans Matthew K. Gold (dir.), *Debates in the Digital Humanities*, University of Minnesota Press, 2012, p. 16-35, en ligne, <<https://www.jstor.org/stable/10.5749/j.ctttv8hq.6>>.

⁴¹ Voir <https://dhcenter.net.org/>.

⁴² Voir <https://www.crihn.org/?redirwww>.

⁴³ Voir <https://adho.org/?lang=fr>.

⁴⁴ Voir <https://www.humanistica.ca/>.

⁴⁵ Voir <https://www.huma-num.fr/quest-ce-que-l-ir-huma-num/>.

⁴⁶ Voir <https://cst-ariane.huma-num.fr>.

⁴⁷ Voir <https://corli.huma-num.fr/fr/accueil/>.

⁴⁸ Durant la période de notre recherche nous avons pu bénéficier de l'aide et de soutien de Victoria Le Fournier et Valentin De Craene qui ont successivement occupé ce poste entre 2019 et 2025. Nous les remercions très chaleureusement pour toute leur aide.

services offrent des solutions de stockage, du traitement, de la diffusion, de l'archivage, du signalement et de l'exposition des données. Ils sont accessibles à condition de disposer d'un HumanID qu'on peut obtenir en prouvant l'inscription de notre projet dans le champ des HN⁴⁹.

Pour que les ressources que nous produisons soient réutilisables, nous nous attachons à suivre les principes établis par la communauté des chercheurs et chercheuses des HN pour garantir l'interopérabilité, la pérennité et l'exploitabilité de notre recherche et des données produites en son sein. En cela, nous nous attachons par exemple à respecter les principes FAIR qui décrivent la manière dont les données doivent être organisées pour qu'elles soient Faciles à trouver, Accessibles, Interopérables et Réutilisables⁵⁰. Pour ce faire, nous disposons de nombreuses publications collectives qui font office de guide pour la publication des corpus dans l'esprit des HN.

Concevoir des publications numériques de qualité

Dans *Les publications numériques de corpus d'auteurs*, Ioana Galleron *et al.*⁵¹ distinguent trois types de publications numériques :

- L'archive éditorialisée met à disposition des fac-similés, parfois accompagnés d'une collection de textes bruts (tantôt générés par l'océrisation, tantôt obtenus par saisie manuelle). Son objectif est de donner accès au texte, « mettre en avant la cohérence d'une série de documents, souvent provenant de sources diverses, et dont la publication facilite la perception comme un ensemble »⁵².
- L'édition de lecture propose un texte relu, corrigé et nettoyé. Ce texte est mis en perspective et les choix éditoriaux (transcription, modalités de lecture, interprétation, etc.) sont rigoureusement documentés. D'enrichissements numériques sont également mis à disposition. Ils peuvent inclure l'analyse et la fouille à l'aide d'outils numériques ou l'extraction des informations pouvant répondre à des questions de recherche particulières. Les fonctionnalités de cette édition peuvent évoluer et se développer.
- L'édition enrichie comprend le texte édité accompagné des informations documentaires et contextuelles. Il est possible de moduler l'affichage en fonction de différents critères éditoriaux. Son enjeu principal consiste en travail d'encodage complexe pour faciliter l'exploitation approfondie des données. Par ailleurs, « ce type d'édition suit nécessairement des pratiques harmonisées et se réfère à des standards soutenus par de larges communautés internationales, comme celle du consortium TEI »⁵³. Il est attendu que cette

⁴⁹ Voir <https://humanid.huma-num.fr/>.

⁵⁰ Voir par exemple <https://www.ccsd.cnrs.fr/principes-fair/>.

⁵¹ Ioana Galleron *et al.*, *Les publications numériques de corpus d'auteurs. Guide de travail, grille d'analyse et recommandations*, Huma-Num CAHIER, 2018. Ici, nous nous appuyons principalement sur les guides de travail conçus dans le cadre du consortium Huma-Num CAHIER (Corpus d'Auteurs pour les Humanités : Informatisation, Édition, Recherche), le pré-décès d'Ariane en activité entre 2011 et 2022 (voir <https://cahier.hypotheses.org/> et, pour la totalité des guides, <https://cahier.hypotheses.org/guides>). Au moment où nous rédigeons ce chapitre, un nouveau guide relatif aux éditions numériques est en cours de préparation au sein du consortium Ariane.

⁵² *Ibid.*, p. 5.

⁵³ *Ibid.*, p. 4.

édition, après une évaluation scientifique par des pairs, soit déposée dans une archive ouverte et dotée d'un identifiant permettant de la citer de manière pérenne.

Dans notre projet, nous visons ce troisième type de publication numérique, ce qui nous incite, avant la conception de l'interface de visualisation, à structurer et à organiser les textes de notre corpus selon les principes partagés par la communauté des HN. Cela implique également de veiller à ce que notre édition inclue les cinq dimensions fondamentales des publications numériques décrites par Galleron *et al* : (i) un texte, (ii) les métadonnées et l'annotation, (iii) la description du projet scientifique, (iv) l'interface de consultation, (v) la gestion des données.

Tout d'abord, le texte doit être complet (les omissions sont signalées et justifiées) et établi de manière méthodique et uniforme, suivant des règles claires. Si possible, il convient à ce que la version publiée soit confrontée aux autres témoins du même texte (qu'ils le précèdent ou suivent).

Ensuite, chaque texte doit être accompagné de métadonnées, définies comme « un ensemble structuré d'informations permettant de décrire la ressource, de la classer, de l'organiser et de caractériser des données ou du contenu »⁵⁴. Chaque type de métadonnée doit respecter les normes et standards internationaux (Dublin Core, ISO, DOI, etc.). Cinq types de métadonnées sont attendues pour l'édition numérique :

- Les métadonnées descriptives qui identifient la source sans ambiguïtés ;
- Les métadonnées administratives qui précisent les modalités d'accès et d'exploitation des fichiers et décrivent leur processus de création ;
- Les métadonnées structurelles qui détaillent l'organisation du fichier (pages, chapitres, table de matières, etc.)
- Les métadonnées techniques qui précisent quels logiciels ont été utilisés pour la production et la manipulation des données, quelles sont les versions des fichiers, etc.
- Les métadonnées d'enrichissement (ou de balisage) qui comportent des annotations pour analyser et interpréter la ressource.

En outre, la description du projet scientifique est nécessaire, car « [u]ne édition qui n'expose pas sa question de recherche et ne déclare pas ses critères de numérisation et de gestion des sources, n'est pas une entreprise scientifique »⁵⁵. La documentation minimale inclut la description des enjeux scientifiques du projet (quelle est sa motivation, ses apports, son contexte scientifique ?). Les publications antérieures sur le sujet sont référencées, l'équipe et leur responsabilités respectives sont présentées. Il est également nécessaire de détailler la composition du corpus et de faciliter la localisation des sources numérisées, de décrire les critères de transcription et la gestion des erreurs présentes dans la source, le traitement de la ponctuation et des graphies anciennes, etc. Enfin, il est attendu que le choix de l'encodage soit détaillé et argumenté.

⁵⁴ *Ibid.*, p. 6.

⁵⁵ *Ibid.*, p. 10.

L'interface de consultation, quant à elle, doit permettre au plus grand nombre d'utilisateurs d'accéder librement au texte et à toutes les fonctionnalités de l'édition. Cela concerne également l'extraction du texte dans un format non-propriétaire et l'accès libre au code source et aux schémas d'encodage. Pour ce faire, il est préconisé d'utiliser des plateformes de partage du type « git »⁵⁶. Différentes versions du texte (originale et modernisée, diplomatique, normalisée, etc.) et formats numériques (texte brut, document XML, livre numérique en format ePub, etc.) devraient également être proposés aux utilisateurs. Par ailleurs, en concevant l'interface, la priorité doit être donnée à sa réutilisabilité et non pas à l'esthétique et aux aspects visuels. Le texte doit avant tout être lisible.

Enfin, en ce qui concerne la gestion des données, un plan de conservation et de gestion est nécessaire pour éviter le risque, extrêmement présent dans l'environnement numérique, de perte de travail. Ce document, spécifiant la manière dont les données sont collectées, gérées, partagées et préservées pendant le projet, est par ailleurs exigé par certains organismes et institutions de financement, comme l'ANR⁵⁷. Pour respecter les principes FAIR, il est également préconisé de déposer les données dans un dépôt assurant leur pérennité. L'entrepôt des données Nakala⁵⁸, géré par l'infrastructure IR* Huma-Num, est recommandé pour les données numériques issues de la recherche en SHS. Pour les disciplines qui ne bénéficient pas d'un autre entrepôt disciplinaire reconnu, il est préconisé d'utiliser la plate-forme Recherche Data Gouv⁵⁹.

La dernière partie du guide de travail de Galleron *et al.* présente une grille d'évaluation des publications numériques⁶⁰, que nous reproduisons ici (Tableau 10) pour montrer ce qui est attendu pour la conception d'une édition enrichie que nous ambitionnons. Nous ajoutons le signe étoile « * » derrière les exigences optionnelles. Toutes les autres sont considérées comme essentielles.

Dans la suite de cette section, nous nous attachons à montrer comment notre édition répond à ces attendus. Nous expliquons comment nous numérisons et enrichissons notre corpus, nous présentons notre projet scientifique et explicitons la manière dont nos données sont gérées durablement. En revanche, nous n'abordons pas, pour le moment, la question de l'interface de consultation.

⁵⁶ Git est un outil libre de gestion de versions (<https://git-scm.com/>) qui, couplé avec des plateformes dédiées d'hébergement de projets, comme GitHub (<https://github.com/>) ou GitLab (<https://about.gitlab.com/>), facilite le développement collaboratif des sites web, logiciels, éditions numériques, etc.

⁵⁷ Un outil d'accompagnement pour la conception des projets de gestion de données, DMP OPIDoR, est disponible sur <https://dmp.opidor.fr/>.

⁵⁸ Voir <https://nakala.fr/>.

⁵⁹ Voir <https://entrepot.recherche.data.gouv.fr/dataverse/root/?q=>.

⁶⁰ Ioana Galleron et al., *Les publications numériques de corpus d'auteurs. Guide de travail, grille d'analyse et recommandations*, op. cit., p. 14-15.

Tableau 10. Grille d'évaluation des publications numériques de corpus d'auteurs selon Galleron et al.

Traitement des sources – Fac-similé (jpeg, pdf, tiff, etc. ; sauf restrictions liées aux droits)* – Transcription (ou OCR, ou similaire), au format brut* – Transcription selon des critères scientifiques établis par le projet (diplomatique, semi-diplomatique, normalisée)
Métadonnées – Format standard (normalisé) – Métadonnées descriptives (bibliographiques, y compris les responsabilités dans la création de l'édition) – Métadonnées administratives (techniques, de droits, etc.) – Métadonnées structurelles – Enrichissements (annotation et/ou balisage)
Description du projet scientifique – Enjeux scientifiques (motivation, apports, etc.) – Présentation de l'équipe et des responsabilités de chacun – Éventuellement, critères ayant accompagné le choix des sources (témoins) – Présentation du corpus* – Présentation des critères de transcription – Précisions sur le traitement des erreurs présentes dans la source (scribes, typographes, lectures antérieures, etc.) – Choix du traitement de la ponctuation et des graphies – Choix d'encodage – Informations sur la genèse du texte*
Interface de consultation – Version de lecture (interface minimale) – Possibilité de lire le texte en différentes versions – Possibilité de télécharger les sources ○ En format texte* ○ En format image (si libres de droits)* ○ En format XML – Possibilité de télécharger le texte en différents formats*
Plan de gestion des données – Usage des standards – Archivage pérenne – Présence d'un identifiant permanent – Programme de maintenance des données (périodicité, responsabilité)

b) Protocole de production des données⁶¹

Dans cette section, nous nous intéressons spécifiquement à la chaîne éditoriale conçue pour l'établissement du corpus des réécritures (corpus principal) en format XML-TEI, puis sa transformation vers d'autres formats (notamment HTML et ePub). Le corpus des traductions (corpus secondaire), que nous introduisons dans la partie suivante⁶², suit les mêmes principes éditoriaux. D'autres données sont également produites dans le cadre de la détection des correspondances avec les logiciels dédiés et lors de l'enrichissement et la visualisation des résultats. Nous les présentons au fur et à mesure de notre rédaction. Notre démarche est également décrite au sein de notre projet de gestion des données⁶³.

Établissement du texte

La méthode d'acquisition des données varie en fonction de l'accessibilité des œuvres. Pour la plupart, nous disposons de la numérisation du fac-similé mis à disposition par Gallica. Certaines sont accompagnés du texte ocrisé, plus rarement d'une édition en format ePub. Pour d'autres, nous disposons du texte brut en accès libre (*Orphée-roi* de Segalen et *Orphée aux Enfers* d'Offenbach, par exemple, sont disponibles sur Wikisource⁶⁴), voire même d'un fichier édité et structuré en format XML-TEI (*Orphée, ou Arlequin aux Enfers* de Fuzelier et *Arlequin aux Enfers* de Regnard ont été édités au sein du projet Théâtre classique⁶⁵). Les œuvres les plus récentes, on revanche, ont été soit numérisées par nos soins à partir d'un livre papier⁶⁶ (*Eurydice* de Jean Anouilh, *Orphée et la morsure du serpent* d'Yvan Pommaux), soit extraits à partir du format ePub⁶⁷.

Quelle que soit la manière d'obtention du fac-similé, dans la plupart des cas nous devons procéder à la reconnaissance automatique du texte (l'ocrisation). De nombreux projets scientifiques

⁶¹ Dans cette section et la section suivante, nous reprenons partiellement notre article « Phoebus e-Balzac : édition numérique exhaustive d'un monument littéraire », dans Clarisse Bardiot, Émilien Ruiz et Esther Dehoux (dir.), *La Fabrique numérique des corpus en sciences humaines et sociales*, Presses universitaires du Septentrion, 2022. Nous enrichissons notre propos et les adaptons aux spécificités de notre corpus.

⁶² Voir partie 2 « Intertextualité : de la théorie littéraire aux méthodes computationnelles », chap. 2 « Procédés linguistiques propres des relations intertextuelles », p. 149.

⁶³ Ce document est disponible dans les annexes ([I 4 Projet-gestion-donnees.pdf](#)). Il a été revu et validé par trois experts dans le domaine, Joachim Schöpfel (GERiiCO), Bernard Jacquemin (GERiiCO) et Valentin De Craene (CNRS) de l'université de Lille. Nous les remercions chaleureusement pour leur expertise.

⁶⁴ Voir <https://fr.wikisource.org/wiki/Orph%C3%A9e-Roi> et https://fr.wikisource.org/wiki/Orph%C3%A9e_aux_Enfers.

⁶⁵ Voir http://www.theatre-classique.fr/pages/programmes/edition.php?t=../documents/FUZELIER_OR-PHEE.xml et http://www.theatre-classique.fr/pages/programmes/edition.php?t=../documents/REGNARD_AR-LEQUINAUXENFERS.xml.

⁶⁶ La qualité de l'image obtenue par un scanner à l'usage personnelle n'est pas suffisante pour les exigences éditoriales, mais les œuvres sous droits sont destinées uniquement à l'interrogation avec les outils de la détection des réutilisations textuelles et ils ne seront pas publiés.

⁶⁷ Le format ePub est en réalité un dossier compressé contenant plusieurs fichiers en format HTML. Plusieurs outils existent pour convertir un livre électronique dans un format structuré, comme le XML-TEI, et notamment Debook conçu par Frédéric Glorieux (voir <http://obvil.lip6.fr/Debook/>).

et commerciaux s'intéressent à ce procédé relativement nouveau (même si les premières tentatives datent d'il y a presque 100 ans⁶⁸). Des logiciels et modèles issus de ces projets diffèrent en fonction de l'objectif visé. Nous présentons rapidement trois outils en accès libre et gratuits.

L'outil Tesseract OCR⁶⁹, par exemple, est un logiciel libre de reconnaissance optique. Largement utilisé, il propose nombreux modèles pré-entraînés⁷⁰ pour pouvoir travailler avec de nombreuses langues et polices. Il peut être facilement installé sur MacOS et Linux et doit être exécuté en ligne de commande. Tesseract a également été implémenté avec une interface web dans la chaîne de traitement conçue au sein du laboratoire ObTIC, Pandore Toolbox⁷¹. Il reste un outil d'océrisation de base qui permet d'obtenir le texte brut rapidement. Cependant, ce dernier nécessite de nombreuses corrections, comme la suppression des fins de lignes et des titres courants. Par ailleurs, il est plutôt efficace pour les textes modernes, mais la qualité de la transcription baisse rapidement pour les éditions plus anciennes.

Transkribus⁷², au contraire, est très efficace pour la transcription des œuvres manuscrites, qu'elles soient anciennes ou modernes, car l'implémentation des méthodes de l'intelligence artificielle facilite l'entraînement des modèles personnels. Il dispose d'une riche sélection de modèles déjà entraînés et très efficaces. L'application en ligne est facile à prendre en main. Elle propose une interface de correction de segmentation de l'image et de transcription. Il est également possible d'ajouter des annotations, par exemple des entités nommées. Enfin, la transcription peut être téléchargée dans différents formats (Word, texte brut, XML...). Cent pages par an peuvent être numérisées gratuitement.

Le projet ANR eScriptorium⁷³, quant à lui, est reconnu notamment pour la transcription des manuscrits anciens⁷⁴. Il s'agit d'une application qui fournit une interface graphique pour le logiciel d'océrisation Kraken⁷⁵, développé à l'École Pratique des Hautes Études et qui bénéficie d'un nombre impressionnant de modèles déjà entraînés spécifiquement sur les textes en français et créés dans le cadre de projets d'édition numérique. La qualité du texte océrisé est excellente, même pour les éditions plus anciennes, et respecte les règles ortho-typographiques françaises, notamment les

⁶⁸ S. Mori, C.Y. Suen et K. Yamamoto, « Historical review of OCR research and development », *Proceedings of the IEEE*, vol. 80, n° 7, Proceedings of the IEEE, juillet 1992, p. 1029-1058, en ligne, <<https://ieeexplore.ieee.org/document/156468>>.

⁶⁹ Voir <https://tesseract-ocr.github.io/tessdoc/>.

⁷⁰ Voir <https://github.com/tesseract-ocr/tessdata>.

⁷¹ Voir <https://pandore-toolbox.isir.upmc.fr/numeriser>.

⁷² Voir <https://www.transkribus.org/fr>.

⁷³ Voir <https://escriptorium.inria.fr/>.

⁷⁴ Nous pouvons mentionner par exemple l'obtention de la prestigieuse bourse Schmidt DataX Fund de l'université de Princeton pour l'application de logiciel sur la collection Cairo Geniza des manuscrits médiévaux écrits en hébreu. Voir Peter Stokes, « eScriptorium recognised in prestigious DataX grant from Princeton University », dans *eScripta*, [text/html], eScripta, 28 mars 2022, en ligne, <<https://escripta.hypotheses.org/486>>.

⁷⁵ Voir <https://kraken.re/main/index.html>.

apostrophes courbes et les espaces insécables précédant les ponctuations doubles. Ce logiciel est gratuit et libre, mais, pour eScriptorium, la création du compte doit être demandée et son installation et sa prise en main restent assez complexes.

Certaines œuvres de notre corpus sont disponibles sur Gallica en format texte. Le logiciel de l'océrisation utilisé n'est pas mentionné, car celle-ci est effectuée par un prestataire externe⁷⁶. La qualité de l'océrisation est très variable : pour la nouvelle « Une histoire renouvelée des Grecs » de D. L. Eimann (1854)⁷⁷ le taux de reconnaissance est estimé à 99,98 %, alors que le poème « Orphée aux Enfers parlant à Pluton » de Tristan l'Hermite (1652)⁷⁸ n'est qu'à 67,87 %. En effet, les deux œuvres ne présentent pas les mêmes défis éditoriaux (Figure 14 et Figure 15).

— Allons chez Julia, se disait Alfred A. par
une froide soirée de décembre.

Mais il se trouvait devant sa cheminée, en
face d'un feu pétillant et gai, à demi-noyé dans
un excellent confortable, et il ne se sentit pas
le courage d'aller affronter la bise, qui, au

Figure 14. Première page de la nouvelle « Une
histoire renouvelée des Grecs » de D. L.
Eimann (1854)

 *E* feu qui m'embrase le sein
Ne vient point dissiper tes ombres ;
Et s'il paroist en ces lieux sombres,
C'est pour un plus noble destin ;
Amour de qui l'humeur altière
Te força de voir la lumière,
Me l'a fait quitter aujourd'huy ;
Et ce Dieu qui vainquit Alcide,
M'ayant abatu comme luy,
De mon vainqueur, devient mon guide.

Figure 15. Première page du poème « Orphée
aux Enfers parlant à Pluton » de Tristan
l'Hermite (1652)

L'édition de 1854 présente quelques défis assez mineurs : il s'agit d'un tapuscrit, certaines lettres sont moins lisibles que les autres (« f » notamment) et la numérisation comporte quelques tâches (un trait au-dessous de « pétillant », un point au-dessous de « sa »). L'œuvre de 1652, en revanche, s'éloigne de manière importante des usages éditoriaux modernes : le texte est en italiques, l'initiale est ornée et les s sont longs (f). Nous observons également deux ligatures : « ft » pour « est » et « destin » ; « ß » pour « dissiper ». Les lettres ne sont pas accentuées (« lumière », « altière »). Enfin,

⁷⁶ « Pour chaque document numérisé par la BnF, le taux de qualité calculé automatiquement par le logiciel est vérifié manuellement par le prestataire sur un échantillon de mots [...]. Pour tous ces documents, quel que soit le taux de qualité après OCR, le prestataire doit garantir ce taux en employant tous les moyens de corrections nécessaires, y compris manuels. », voir § « Taux qualité » dans l'article « Techniques et formats de conversion en mode texte », dans BnF - Site institutionnel, en ligne, <<https://www.bnf.fr/fr/techniques-et-formats-de-conversion-en-mode-texte>>.

⁷⁷ Voir <https://gallica.bnf.fr/ark:/12148/bpt6k5699364p/f26.item>.

⁷⁸ Voir <https://gallica.bnf.fr/ark:/12148/bpt6k5540284p/f42.item>.

nous retrouvons encore quelques traces des lettres ramistes *u* et *v* (« vn », « deuient ») et des graphies anciennes (« abatu », « aujourd'huy », « luy »).

Tout cela influence forcément la qualité des océrisations, mais les différents logiciels n'ont pas les mêmes performances. Pour mieux en rendre compte, nous comparons les transcriptions de Gallica pour la première page de ces deux textes avec les résultats obtenus avec Tesseract, Transkribus et Kraken (Tableau 11). Nous cherchons un outil qui demande le moins de corrections, qu'il s'agit de la transcription du texte que de la ponctuation. En revanche, à ce stade de la transcription, nous ne nous souhaitons pas que le texte soit modernisé (ajout des accents, etc.).

Tableau 11. Comparaison des outils d'océrisation pour un texte de 1854 et un texte de 1652

	Gallica	Tesseract	Transkribus	Kraken
1854	— Allons chez Julia, se disait Alfred A. par une froide soirée de décembre. Mais il se trouvait devant sa cheminée, en face d'un feu pétillant et gai, à demi-noyé dans un excellent confortable, et il ne se sentit pas le courage d'aller affronter la bise, qui, au	— Allons chez Julia, se disait Alfred A. par une froide soirée de décembre. Mais il se trouvait devant sa cheminée , en face d'un feu pétillant et gai, à demi-noyé dans un excellent confortable, et il ne se sentit pas le courage d'aller affronter la bise, qui, au	Allons chez Julia, se disait alfred A. par une froide soirée de décembre. Mais il se trouvait devant sa cheminée, en face d'un feu pétillant et gai, à demi- nové dans un excellent confortable, et il ne se sentit pas le courage d'aller affronter la bise, qui, au	— Allons chez Julia, se disait Alfred A. par une froide soirée de décembre. Mais il se trouvait devant sa cheminée, en face d'un feu pétillant et gai, à demi-noyé dans un excellent confortable, et il ne se sentit pas le courage d'aller affronter la bise, qui, au
1652	SffijÉS?^- seu °lu'17^ embrafe le sein wÉ i?mÊi ~^e 'vient point diffiper tes ombres 5 JE; ÌIJBI? Et s'il paro ifi en ces lieux sombres 3 C efi pour vn plus noble destin j Amour de qui l'humeur alti ère F^e fors a de voir la lumière, Me l'a fait quitter aujourd'huy \$ Et ce Dieu qui vainquit Alcide s M'ayant abat u comme luy, De mon vainqueur, deuient mon guide.	VE feu qu im 'embraf è le fein de CA NN Ne vient point diffiper tes ombres ; KW Ê 7) Et s'il paro ift en ces lieux ombres , 7 Cf pour Vn plus noble de ffins Amour de qui l'humeur alti ère Te for ca de voir la lumière , Lie l'a fait quitter aujourd'huy; Et ce Dieu qui vainquit Alcide, M'ayant abatu comme luy, De mon vainqueur, deuient mon guide.	Le feu qui m'embrase le sein Ne vient point dissiper tes ombres; Et s'il paro ist en ces lieux sombres, C'est pour vn plus noble destin; Amour de qui l'humeur alti ère Te for ça de voir la lumière Me l'a fait quitter aujourd huy; Et ce Dieu qui vainquit Alcide M'ayant abatu comme luy De mon vainqueur, deuient mon guide.	E feu qui m'embrase le sein Ne vient point di ßiper tes ombres; Et s'ïl paro ist en ces lieux sombres, C'est pour vn plus noble destin; Amour de qui l'humeur alti ere Te for ça de voir la lumière, Me l'a fait quitter aujourd'huy; Et ce Dieu qui vainquit Alcide, M'ayant abatu comme luy, De mon vainqueur, deuient mon guide.

Nous constatons que pour le texte d'Eimann les quatre transcriptions sont plutôt satisfaisantes et les corrections nécessaires restent minimales (remplacement de l'apostrophe droit par le courbe). Les bruits présents sur l'image ne semblent pas poser problème, sauf pour Tesseract qui ajoute « | » à la fin du premier paragraphe. Les seules erreurs de reconnaissance du texte sont présentes dans les résultats de Transkribus : il manque le tiret cadratin qui introduit la réplique, le nom d'Alfred est en minuscule, et il y a une confusion entre « u » et « y » dans « demi-noyé ».

Curieusement, Transkribus propose une transcription très satisfaisante du texte plus ancien : mis à part les apostrophes droites, nous observons uniquement l'ajout de l'accent grave à « altièr » et le manque d'apostrophe dans « aujourd'huy ». Alors que les résultats de Kraken sont quasiment parfaits (le manque de l'initiale ornementé « L » est la seule erreur observée, et encore il est plutôt favorable que celle-ci ne soit pas considérée comme texte, mais comme image), ceux de Gallica et de Tesseract ne sont pas tout à fait satisfaisants. La tentative de transcrire en texte l'initiale ornementée ajoute énormément du bruit aux quatre premiers vers. Les *s* longs ne sont pas reconnus, mais confondus avec des « f », les ligatures « st » sont transcrites par « fi », les ponctuations ne sont pas reconnues, etc. En somme, pour Gallica et Tesseract, il est envisageable que la transcription manuelle soit plus rapide que la correction des résultats obtenus.

Kraken propose donc indubitablement les meilleurs résultats, quelle que soit l'époque de la publication du texte. Nous le privilégions pour la totalité des œuvres du corpus, même quand l'océrification est déjà disponible sur Gallica. Nous utilisons le modèle CATMuS Print (Large) mis à disposition par Simon Gabay et Thibault Clérice⁷⁹. Le texte ainsi obtenu est ensuite revu et corrigé en appliquant les règles éditoriales suivantes :

- Les *i* et les *u* ayant valeur de consonnes sont à remplacer respectivement par *j* et par *v* (et inversement, le cas échéant) ;
- Les cédilles et apostrophes sont rétablis ;
- Les mots qui sont (ou semblent) accolés sont séparés ;
- L'emploi des majuscules suit l'usage actuel (noms propres, début des phrases, etc.) ;
- Pour les éditions prémodernes et les manuscrits, l'accentuation suit les règles de l'École nationale de Chartes : seulement l'accent aigu est introduit, et ce seulement sur les mots se terminant en *e* ou *es*, pour distinguer des *e* muets (*porté, né, mais pere, portée*) ;
- Les abréviations sont résolues et le développement respecte la graphie originale (*cõe* est résolu en *come* et non pas *comme*, etc.) ;
- Le *et* est développé en *et* ;
- Les espaces insécables sont introduits avant les ponctuations doubles, entre guillemets français et après le tiret demi-cadratin ;

⁷⁹ Simon Gabay et Thibault Clérice, « CATMuS-Print [Large] », *Reconnaissance des écritures dans les imprimés. CATMuS print: un modèle générique, multilingue et diachronique*, Zenodo, janvier 2024, en ligne, <<https://zenodo.org/records/10592716>>.

- Tous les tirets, sauf les tirets simples des mots composés sont remplacés par les tirets demi-cadratin ;
- Toutes les apostrophes droites sont remplacées par des apostrophes courbes ;
- L’accentuation sur les majuscules est ajoutée, même en cas d’absence dans le fac-similé ;
- Les graphies ne sont pas modernisées, mais la ponctuation est rétablie suivant l’usage courant (points d’interrogation, guillemets, etc.) ;
- Les chiffres romains qui composent les noms des rois sont précédés d’un espace insécable (Louis XIII, François 1^{er}) ;
- Pour transcrire les siècles, les chiffres romains sont en petites majuscules suivis d’un « e » en exposant (XIX^e s.) ;
- Les abréviations de Madame et de Mademoiselle sont uniformisées en M^{me} et M^{lle}.
- Les coquilles sont corrigées et signalées avec les notes de l’éditeur.

Stylage

Après l’établissement d’une version satisfaisante du texte, nous suivons la chaîne de traitement Odette/Teinte conçue par Frédéric Glorieux⁸⁰. Appuyée sur une structure restreinte issue du standard XML-TEI, Teinte, une feuille de style dédiée a été proposée⁸¹ pour permettre la structuration à l’aide des logiciels du traitement de texte (Word ou OpenOffice). Même si les outils WYSIWIG⁸² sont peu performants pour les structures complexes, ils sont suffisants pour la structuration « de surface » des œuvres (chapitres, paragraphes, citations, numéros de pages, etc.) et plus confortables à utiliser que les logiciels de structuration XML où l’on travaille directement avec les balises. Cette méthode a aussi l’avantage de faciliter le partage des textes en cours de transcription dans un format plus accessible aux chercheurs non-initiés au langage XML (pour relecture, consultation concernant les coquilles, etc.).

Optimisé pour les textes romanesques, le stylage avec le schéma Teinte compose avec les fonctionnalités initiales des outils (mise en italique, style du paragraphe par défaut, etc.) et les noms de styles particuliers (mis entre chevrons) permettant l’annotation sémantique d’un bon nombre d’éléments textuels (citations, illustrations, correspondance, poèmes, etc.). Le développement est accompagné d’un guide d’emploi détaillé⁸³. Le nombre d’éléments limite l’hétérogénéité des annotations choisies par les éditeurs impliqués dans le processus, dont l’interprétation de la structure textuelle peut varier.

Voici par exemple ce que nous obtenons pour une œuvre en vers, la cantate *Orphée* de Nicolas Clérembault (Figure 16). Les vers sont balisés avec le style <l>, les précisions musicales (« Récitatif »,

⁸⁰ Frédéric Glorieux, « Le traitement de textes (odt) pour produire des documents structurés (XML/TEI) — Odette », dans *J’attends des résultats*, 2015, en ligne, <<https://resultats.hypotheses.org/267>>.

⁸¹ Celle utilisée au sein de notre projet se trouve dans les annexes : [I 4 Styles odette.docx](#). Le tableau des styles est accompagné de quelques exemples.

⁸² *What You See Is What You Get*, acronyme qui renvoie aux éditeurs visuels de texte.

⁸³ Voir <https://obvil.github.io/Teinte/teinte.html>.

« Air tendre et piqué ») sont stylées avec <label>. Enfin, le début de chaque page (accompagné de numéro) est signalé avec <pb>.

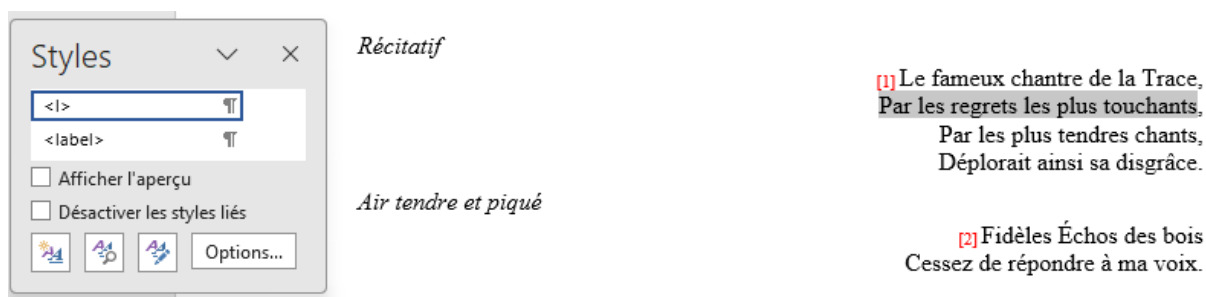


Figure 16. Extrait du stylage d'Orphée de Nicolas Clérembault

Le stylage se complexifie davantage pour les pièces de théâtre et les opéras. Dans l'extrait de la tragédie en musique *Orphée* de Michel Du Boullay et Louis de Lully présenté à la Figure 17, en plus des vers et les numéros de pages, nous trouvons également les didascalies (en italiques) stylées avec <stage>. Le locuteur qui prononce la réplique est signalé avec <speaker>. La feuille de style permet également de combiner les deux styles pour signaler les didascalies qui suivent immédiatement le nom du locuteur (ORPHÉE, *en chantant*). La didascalie est alors stylée avec le style de caractères <stage.c>.

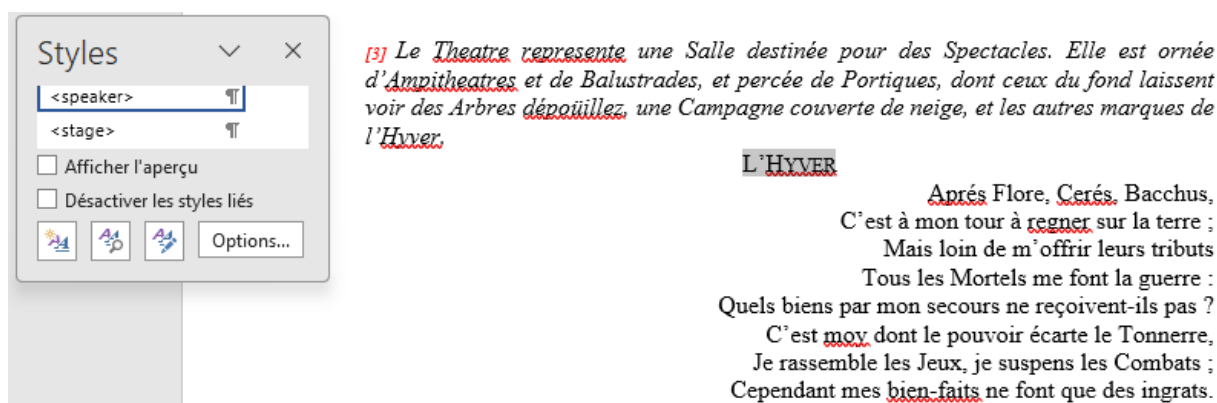


Figure 17. Extrait du stylage de la tragédie Orphée de Du Bellay et Lully.

Enfin, alors que les textes en prose nécessitent généralement un stylage minimal, il convient de noter quelques cas particuliers. Dans « Une histoire renouvelée des Grecs », par exemple, nous trouvons plusieurs citations des œuvres existantes, notamment deux vers de poème XVI de *Mardoche* d'Alfred de Musset (1829), « Une lune de miel n'a pas trente quartiers / Comme un baron saxon ». Nous les signalons, en suivant le standard XML-TEI Teinte, par <quote>. Mais un autre élément est

également différencié du corpus du texte par une mise en page spécifique⁸⁴ sans pour autant pouvoir être considéré comme citation. Il s'agit d'un billet envoyé à Alfred, le narrateur de l'histoire, par son amie Julia (Figure 18).

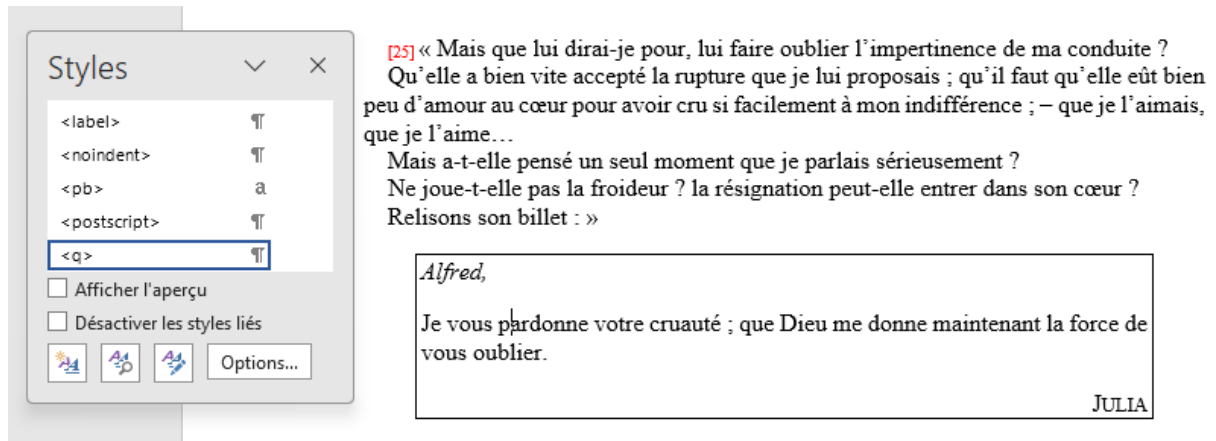


Figure 18. Extrait de stylage d'« Une histoire renouvelée des Grecs » de D. L. Eimann.

Du point de vue philologique, une citation est définie comme une partie de texte externe à l'œuvre, alors qu'ici il s'agit d'une lettre écrite par un des personnages. Nous choisissons donc de le signaler avec l'élément `<q>`, qui, dans le standard TEI,

contient un fragment qui est marqué (visiblement) comme étant d'une manière ou d'une autre différent du texte environnant, pour diverses raisons telles que, par exemple, un discours direct ou une pensée, des termes techniques ou du jargon, une mise à distance par rapport à l'auteur, des citations empruntées et des passages qui sont mentionnés, mais non employés.⁸⁵

Pour répondre aux besoins spécifiques de plusieurs projets ayant rencontré ce type d'entités, le schéma Teinte a modifié cette définition en précisant que « le contenu de `<q>` ne peut pas être attribué à une origine extérieure au texte (origine fictionnelle ou non identifiable). Exemples : récit enchâssé, lettre insérée [...] »⁸⁶. La structure interne de l'élément textuel codé par `<q>` est également balisée à l'aide de styles dédiés. En Figure 18, par exemple, la formule de politesse est signalée avec `<salute>` et la signature de la femme est marquée avec `<signed>`.

Transformation automatique

Le texte stylé dans un traitement de texte et enregistré en format ODT est ensuite transformé en format XML-TEI avec le logiciel Odette⁸⁷. À nouveau, l'emploi d'un outil intuitif et disponible en ligne augmente l'autonomie des éditeurs sans demander des compétences particulières.

⁸⁴ Voir <https://gallica.bnf.fr/ark:/12148/bpt6k5699364p/f28.item>.

⁸⁵ Voir <https://tei-c.org/release/doc/tei-p5-doc/fr/html/ref-q.html>.

⁸⁶ Voir https://obvil.github.io/Teinte/teinte.html#el_q_2.

⁸⁷ Voir <http://obvil.lip6.fr/Odette/>.

Grâce à Odette, trois étapes simples (chargement du fichier stylé, choix du format d'export, puis téléchargement du résultat en format choisi) suffisent pour transformer rapidement un document ODT vers le format XML-TEI Teinte. Pour l'extrait textuel présenté à la Figure 18, par exemple, nous obtenons une structuration suivante :

```
<p><pb n="25" xml:id="p25"/><p>« Mais que lui dirai-je pour, lui faire oublier l'impertinence de
ma conduite ?</p>
<p>Qu'elle a bien vite accepté la rupture que je lui proposais ; qu'il faut qu'elle eût bien
peu d'amour au cœur pour avoir cru si facilement à mon indifférence ; – que je l'aimais,
que je l'aime...</p>
<p>Mais a-t-elle pensé un seul moment que je parlais sérieusement ?</p>
<p>Ne joue-t-elle pas la froideur ? la résignation peut-elle entrer dans son cœur ?</p>
<p>Relisons son billet : »</p>
<q type="letter">
<opener>Alfred,</opener>
<p>Je vous pardonne votre cruauté ; que Dieu me donne maintenant la force de vous
oublier.</p>
<signed>Julia</signed>
</q>
```

Figure 19. Structuration XML-TEI du début de la page 25 d'« Une histoire renouvelée des Grecs » de D. L. Eimann

Pour les utilisateurs expérimentés, le code source du logiciel est disponible en ligne⁸⁸. Le logiciel est conçu principalement à l'aide des feuilles de style XSL et scripts PHP. Il est possible, en l'utilisant localement, d'appliquer massivement la transformation à l'aide d'une ligne de commande exécutée à partir du dossier avec les ressources Odette⁸⁹.

La qualité de la transformation est très satisfaisante (à condition, bien sûr, que le stylage soit fait correctement et en respect du schéma Teinte). Sporadiquement, des balises auto-fermantes <anchor/> peuvent apparaître dans le XML. Il est également conseillé de revoir les éléments de l'emphase⁹⁰, car la mise en italiques de longs fragments textuels produit parfois l'accumulation inutile des balises.

Ajout des métadonnées

Ensuite, nous modifions et complétons l'élément <teiHeader> contenant les métadonnées de chaque œuvre. Le <teiHeader> comporte trois parties : la description du fichier (<fileDesc>), la description de l'encodage (<encodingDesc>) et la précision d'autres aspects non bibliographiques (<profileDesc>).

⁸⁸ Voir <https://github.com/oeuvres/odette>.

⁸⁹ « php Odt2tei.php "*" .odt XML/ ». La variable « *.odt » indique que tous les fichiers en format ODT présents dans le dossier doivent être traités. En revanche, il n'est pas permis de les regrouper dans un sous-dossier : la variable « sous-dossier/*.odt » ne sera pas reconnue. Les résultats, eux, peuvent être sauvegardés dans un sous-dossier spécifique (« XML/ » dans notre exemple).

⁹⁰ Cela peut être fait avec les deux requêtes XPath : « //hi[@rend="i"] » et « //emph ».

Dans la première partie, nous précisons le titre et l’auteur de l’œuvre (<titleStmt>). Nous présentons ensuite l’édition et les acteurs engagés dans sa création (<editionStmt>), puis décrivons la façon dont le texte est distribué (<publicationStmt>). Nous indiquons enfin la référence bibliographique complète du document (<sourceDesc>). La deuxième partie dédiée à l’encodage comporte uniquement la précision suivante :

Nous utilisons la feuille de style Teinte pour les premières transformations ODT/XML. Nous nettoyons et enrichissons ensuite la structure avec des algorithmes Python conçus au sein du projet. La structure des métadonnées a été établie en référence à la manière dont le logiciel TextPAIR (The ARTFL Project) les extrait ensuite pour son traitement.

En effet, dans la troisième partie du <teiHeader> (<profilDesc>), nous ajoutons plusieurs informations destinées spécifiquement à être traitées par TextPAIR, un des logiciels que nous utilisons pour la détection automatique des réutilisations textuelles⁹¹. Cela nous permet par la suite d’identifier et de catégoriser automatiquement les résultats de la détection. Ainsi, comme présenté à la Figure 20, les différentes catégorisations des œuvres du corpus (genre, sous-genre, intermédialité, type, diégèse, transformation, public, canon, accès et langue) ainsi que l’identifiant unique (composé d’un nom d’auteur et de la date de publication) sont précisés en plus des métadonnées plus usuelles, présentes dans la première partie (<fileDesc>).

```
<textClass>
  <keywords>
    <list>
      <item ana="identifiant">eimann1854</item>
      <item ana="genre">narratif</item>
      <item ana="subgenre">nouvelle / histoire cavalière</item>
      <item ana="intermedialite"></item>
      <item ana="type">parodie</item>
      <item ana="diégèse">hétérodiégétique</item>
      <item ana="transformation">modernisation</item>
      <item ana="public">adultes</item>
      <item ana="canon">non-canon</item>
      <item ana="accès">libre</item>
      <item ana="langue">fr</item>
    </list>
  </keywords>
</textClass>
```

Figure 20. Annotation des différentes catégorisations dans <teiHeader> : l’exemple d’« Une histoire renouvelée des Grecs » de D. L. Eimann.

⁹¹ Voir III. « Logiciels dédiés à la détection automatique des réemplois textuels », notamment à partir du chap. 3, « Présentation des logiciels retenus, TextPAIR et Tracer », p. 281 où, après avoir présenté ce logiciel, nous reviendrons plus en détail aux ajustements effectués automatiquement et aux éléments dédiés au traitement avec TextPAIR.

La transformation vers des format éditoriaux

Il est également possible, dans la continuité de la chaîne du traitement Odette/Teinte, de générer des fichiers HTML à partir du document XML ainsi obtenu. La procédure devient, à ce point, plus complexe, notamment si l'on souhaite utiliser le développement sur un serveur Web⁹². Toutefois, pour une utilisation basique et locale, les compétences de base en XSLT sont suffisantes⁹³.

La visualisation ainsi générée est basique, mais offre tout de même quelques fonctionnalités facilitant la lecture dans l'environnement numérique. Par exemple, le clic sur le numéro de page affiché sur la marge gauche permet d'afficher le fac-similé correspondant⁹⁴, la génération de la table des matières accélère la navigation dans le texte, et la mise en forme des éléments balisés facilite la révision et le contrôle qualité du document. Ce fichier HTML de base peut également repris pour une visualisation plus complexe. L'avantage principal de la transformation XML vers HTML avec Teinte est qu'il évite la création des feuilles de style propres à chaque projet éditorial. C'est une tâche complexe et souvent chronophage, qui nécessite des compétences avancées de la programmation en langage XSL et de nombreux tests pour s'assurer que tous les éléments de la structure des œuvres ont été pris en compte correctement.

Cas particuliers

Parmi 53 œuvres du corpus, 26 ne se composent pas uniquement du texte, mais aussi de la musique (22) ou des images (5). Les ouvrages musicaux sont des cantates, opéras et ballets, il est donc relativement aisé de séparer la dimension textuelle et musicale de l'œuvre grâce aux livrets d'opéra, des partitions (quelles soient manuscrites ou imprimées), etc. Seules les annotations descriptives (« air léger », « ritournelle », etc.) gardent la trace de l'intermédialité de l'œuvre.

En ce qui concerne les images, dans deux cas sur cinq, il s'agit des illustrations. Dans *La métamorphose d'Ovide figurée*, par exemple, à chaque page une illustration est accompagnée d'un bref poème. Cela peut aisément être transcrit en XML-TEI (Figure 21).

⁹² Dans ce cas, nous renvoyons les lecteurs vers la documentation détaillée disponible sur <https://github.com/oeuvres/teinte>. Pour l'utilisation avec Omeka, voir <https://github.com/oeuvres/Bookmeka>.

⁹³ À partir du dossier des sources, le document XML traité doit être transformé à l'aide de la feuille de style « `tei2html.xsl` ». Pour ce faire, il est possible d'utiliser une ligne de commande à exécuter à partir du dossier Teinte : « `xsltproc tei2html.xsl texte.xml > ../HTML/texte.html` ». Pour que le fichier HTML ainsi généré puisse être mis en forme (à l'aide d'une feuille de style « `tei2html.css` »), il doit être enregistré dans un dossier frère de celui contenant les ressources Teinte.

⁹⁴ À condition que le lien vers l'image soit renseigné en tant que valeur de l'attribut `@facs` de l'élément `<pb>`.

```

<div>
  <pb n="61r" xml:id="f61r"/>
  <head>Eurydice morte du Serpent</head>
  <figure>
    <graphic url="tourmes1557_graphic1.jpg"/>
  </figure>
  <div type="poem">
    <|>Orphee ayant pour espouse Eurydice</|>
    <|>Pas n'eut faveur du Dieu de mariage :</|>
    <|>Car un Serpent, à son grand prejudice,</|>
    <|>Lui ha tollu son amoureux partage :</|>
    <|>Sans qu'en amours il ust grant avantage</|>
    <|>En ce point donq, du faus Serpent pinsee</|>
    <|>Par le talon, en un champ sur l'herbage,</|>
    <|>Tot trespasa cette povre espousee.</|>
  </div>

```

Figure 21. Transcription XML-TEI du poème « Eurydice morte du Serpent » de La métamorphose d'Ovide figurée (folio 61r).

Néanmoins, pour les trois œuvres restantes, la coexistence de deux médias (l'image et le texte) est plus substantielle, car il s'agit de deux bandes dessinées et d'un album de jeunesse. Le texte et l'image ne sont pas séparés de manière distincte, les répliques des personnages étant transcrites dans les bulles et les onomatopées incluses dans le dessin (Figure 22⁹⁵).



Figure 22. L'image de la deuxième vignette d'Orphée et Eurydice de Bottet et Pinet.

La TEI ne prévoit pas de schéma pour la transcription de ce type d'ouvrages, mais une proposition d'encodage a été faite par John A. Walsh⁹⁶. *Comic Book Markup Language* (CBML) permet d'une part d'utiliser les éléments existants du standard TEI pour décrire des structures caractéristiques de la bande dessinée, comme <sound> pour les onomatopées, et, de l'autre part, d'ajouter de nouveaux

⁹⁵ Béatrice Bottet et Aurore Pinet, « Les Métamorphoses d'Ovide », *La Mythologie en BD*, vol. 7, Paris, Casterman, 2016, p. 42.

⁹⁶ John A. Walsh, « Comic Book Markup Language: An Introduction and Rationale », *Digital Humanities Quarterly*, vol. 6, n° 1, 2012, p. 43.

éléments spécifiquement dédiés à ce genre. Les noms de ces éléments sont alors précédés par le préfixe « cbml: » (<cbml:balloon> pour le contenu d'une bulle, etc., Figure 24).

```
<cbml:panel characters="#hymenee" n="2" xml:id="bottet2019_1-2">
  <cbml:caption>
    Pour la noce, les Dieux avaient envoyé Hyménée, le dieu du mariage en
    personne.
  </cbml:caption>
  <cbml:balloon type="thought" who="#hymenee">
    Qu'a donc ma torche nuptiale ? Elle brûle mal, ce n'est pas un bon présage
    pour les mariés...
  </cbml:balloon>
  <sound>
    Kof Kof
  </sound>
</cbml:panel>
```

Figure 23. Transcription de la deuxième vignette d'Orphée et Eurydice de Bottet et Pinel à l'aide du schéma CBML

Toutefois, même si la question de l'annotation de ce genre d'ouvrages est importante, elle n'est pas au centre de notre recherche. CBML n'est pas un schéma intégré pleinement dans le standard TEI et les usages des utilisateurs. La transcription aussi détaillée reste très chronophage sans pour autant présenter de valeur ajoutée à notre principal objectif de recherche, à savoir la détection des réutilisations textuelles. Nous décidons alors de séparer artificiellement le texte et l'image et d'encoder les bandes dessinées de manière similaire à celle adoptée pour les œuvres dramatiques.

Ainsi, chaque planche est insérée dans une <div>, qui contient à son tour autant de <div> que de vignettes. C'est à l'intérieur de chaque vignette que nous employons l'encodage théâtral : <sp> pour la bulle, <speaker> pour le nom du personnage qui prononce la réplique, <stage> pour le cartouche (l'encadré narratif). Nous ne transcrivons pas les onomatopées et ne gardons pas les images dans la transcription XML. Ainsi, la vignette présentée à la Figure 22 est encodée de manière plus simple et surtout plus facilement comparable avec les autres œuvres du corpus (Figure 24).

```
<div corresp="#bottet2019_1-2">
  <stage>
    Pour la noce, les Dieux avaient envoyé Hyménée, le dieu du mariage en
    personne.
  </stage>
  <sp who="hymenee">
    <speaker>Hyménée</speaker>
    <p>
      Qu'a donc ma torche nuptiale ? Elle brûle mal, ce n'est pas un bon
      présage pour les mariés...
    </p>
  </sp>
</div>
```

Figure 24. Transcription simplifiée de la deuxième vignette d'Orphée et Eurydice de Bottet et Pinel.

Certes, nous perdons des informations, notamment concernant le type de la bulle (réplique, pensée, etc.) et les onomatopées, mais ces éléments ne nous semblent pas pouvoir constituer d'indices substantielles pour la recherche des réutilisations textuelles. Par ailleurs, comme les œuvres concernées sont restreintes par les droits d'auteur, les questions éditoriales ne se posent pas ici, car les deux bandes dessinées et l'album de jeunesse ne seront pas intégrés dans notre publication numérique.

c) Description du jeu de données déposé dans Nakala

Le corpus principal, composé de 53 réécritures, et le corpus secondaire, regroupant 21 traductions d'Ovide et 16 traductions de Virgile, ont été déposés dans l'entrepôt de données Nakala. Ce faisant, nous souhaitons nous inscrire dans la politique nationale de science ouverte⁹⁷ tout en respectant les principes FAIR, cruciales au sein des projets issus des humanités numériques. Pour compléter le dépôt, nous consacrons cette section à la description du jeu de données et de la modélisation spécifique des métadonnées préparée dans ce cadre. Nous espérons que cela facilitera la réutilisation de notre corpus par d'autres chercheurs et chercheuses.

Choix de l'entrepôt

Comme mentionné supra⁹⁸, Nakala est un service de l'IR* Huma-Num permettant aux chercheurs, enseignants-chercheurs ou équipes de recherche de partager, publier et valoriser tous types de données numériques documentées (fichiers texte, sons, images, vidéos, objets 3D, etc.) dans un entrepôt sécurisé afin de les publier en accord avec les principes des données FAIR (*FAIR data*). Choisir Nakala comme entrepôt de données permet de s'inscrire dans la logique du Web de données ouvertes (*Linked Open Data*), qui rend possible la connexion à d'autres entrepôts existants. C'est en quelque sorte le prolongement de notre travail sur l'intertextualité.

Une fois les données publiées sur Nakala, il est possible de les récupérer, moissonner⁹⁹ et exposer ailleurs grâce au protocole documentaire OAI-PMH¹⁰⁰, au modèle RDF¹⁰¹ ou à une API REST¹⁰². Par ailleurs, Nakala fait partie d'un dispositif cohérent de services mis en place par Huma-

⁹⁷ Voir <https://www.ouvrirlascience.fr/accueil/>.

⁹⁸ Voir § « Concevoir des publications numériques de qualité », p. 94.

⁹⁹ Cela veut dire que les données peuvent être récoltées et incluses dans une base de données regroupant les documents aux références bibliographiques similaires. Seulement les données encodées avec les mêmes procédés techniques peuvent être moissonnées.

¹⁰⁰ *Open Archives Initiative – Protocol for Metadata Harvesting*, protocole pour la collecte des métadonnées de l'initiative pour les archives ouvertes, est un dispositif permettant l'échange des métadonnées entre plusieurs institutions qui donnent accès aux documents numériques (voir <https://www.openarchives.org/OAI/openarchivesprotocol.html>). Il est utilisé, par exemple, par la BnF (voir <https://www.bnf.fr/fr/les-entrepots-oai-de-la-bnf>).

¹⁰¹ *Resource Description Framework* est un modèle descriptif des données Web et des métadonnées attachées.

¹⁰² *Application Programming Interface Representational State Transfer*, une interface de programmation d'information qui respecte un ensemble de règles et de consignes pour la création des interfaces Web.

Num en suivant la chaîne de traitement des données. L'accès, le signalement, la conservation et l'archivage des données de la recherche en SHS sont facilités par les outils développés par Huma-Num.

Nakala est destiné au stockage de données stabilisées (décrites et complètes) et est utilisé par les porteurs de projets une fois le processus de collecte des données terminé. Afin de réaliser une collecte optimale, les projets sont encouragés à :

- mettre en sécurité les données sur un outil de stockage externe (disques durs...) ;
- ordonner et organiser les données ;
- consigner toutes les informations disponibles sur les données ;
- mettre en place un plan de nommage harmonisé des fichiers¹⁰³ ;

Nakala accepte tous les types de formats de fichiers pour le dépôt et permet d'en visualiser certains¹⁰⁴. Les données sont respectées dans leur intégrité et ne subissent pas de modification une fois déposées. De plus, Huma-Num effectue une copie sécurisée des données et des métadonnées au sein de son infrastructure. Celles-ci sont identifiées grâce à l'attribution d'identifiants pérennes. Il est ainsi possible de citer chaque donnée de manière précise et exacte grâce au DOI¹⁰⁵. Par exemple, les données relatives à la nouvelle d'Eimann sont connues sous l'identifiant 10.34847/nkl.3932s9o3¹⁰⁶. Il est également possible de citer la référence complète de la donnée¹⁰⁷. Même si cette manière de référencer est encore à retravailler, notamment en ce qui concerne la reprographie des textes imprimés, l'identifiant alphanumérique assure que la citabilité de chaque donnée et de chaque collection soit pérenne et aussi exacte que possible. Afin de suivre les recommandations FAIR, nous nous sommes également assurée que les ressources déposées ne contiennent pas d'informations personnelles, dont la diffusion est encadrée par le Règlement général sur la protection des données (RGPD). Par conséquent, il n'est pas nécessaire de mettre en place des procédures de pseudonymisation¹⁰⁸ ou de *mash-up* (application composite) des données. Elles doivent tout de même être mises sous une licence. Tant que pour le dépôt de notre corpus sur Nakala que pour notre projet en général, nous appliquant la licence *Creative Commons Attribution Non Commercial No Derivatives 4.0 International* (CC-BY-NC-ND-4.0)¹⁰⁹.

Outre la facilitation du respect des principes FAIR, d'autres fonctionnalités de Nakala visent à simplifier la gestion collective du dépôt. Tous les utilisateurs pourvus des droits d'édition peuvent

¹⁰³ Cela implique que les fichiers sont toujours nommés de la même manière, par exemple : « Type_date_numéro ».

¹⁰⁴ Elle intègre, par exemple, la visionneuse d'images OpenSeadragon (voir <https://openseadragon.github.io/>).

¹⁰⁵ *Digital Object Identifier*, identifiant d'objet numérique. Voir <https://www.doi.org/>.

¹⁰⁶ Voir <https://nakala.fr/10.34847/nkl.3932s9o3>.

¹⁰⁷ D. L., E. (2024). Une histoire renouvelée des Grecs [dans :] *Histoires cavalières*, avec une préface d'Alexandre Dumas. NAKALA - <https://nakala.fr> (Huma-Num - CNRS). <https://doi.org/10.34847/NKL.3932S9O3>.

¹⁰⁸ Moins forte que l'anonymisation, qui empêche totalement et de manière irréversible la possibilité d'identifier une personne, la pseudonymisation permet d'empêcher l'identification d'une personne en fonction des données présentes.

¹⁰⁹ Voir <https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>.

déposer dans les collections et modifier les métadonnées avant la publication. Cette gestion fine des droits d'accès aux dépôts, collections et fichiers est un atout de taille pour répartir le travail entre deux personnes n'appartenant pas à la même institution et ne pouvant pas se voir régulièrement. Les droits sur les données sont partagés entre le déposant et les administrateurs ajoutés. La vue et l'édition de la donnée dépendent ainsi des rôles attribués par la suite, qui peuvent être modifiés par le déposant. Il est aussi possible d'ajouter des listes d'utilisateurs autorisés à avoir certains rôles. Cette fonctionnalité est certes plus utile pour les projets regroupant plusieurs collaborateurs, mais même dans le cadre de notre projet, elle nous permet, par exemple, de bénéficier d'une révision des données par une personne spécialisée dans la gestion des données. Nous accordons les droits d'administration à notre direction de thèse, ainsi qu'à l'ingénieur chargé du traitement des données scientifiques au sein de la MESHS de Lille Nord de France, Valentin de Craene, qui nous a accompagnée lors de la préparation et du dépôt de nos données.

Métadonnées

La modélisation des métadonnées a été effectuée à partir des informations présentes dans le `<teiHeader>` des fichiers XML et les métadonnées obligatoires pour chaque dépôt dans Nakala. Chaque donnée déposée peut contenir plusieurs fichiers et est décrite par un certain nombre de métadonnées. La clarté dans l'exposition de ces dernières est fondamentale pour une réutilisation et une compréhension par autrui. Aussi, nous visons la précision optimale de la description de chaque donnée.

Cinq métadonnées sont rendues obligatoires pour le dépôt : le type de la donnée, le titre, l'auteur, la date de création et la licence (voir Tableau 12). Il est admis qu'un auteur soit anonyme ou qu'une date soit inconnue. Le choix du type de dépôt est restreint par les valeurs présentes au sein d'une liste déroulante. Le type « Édition de sources » convient à la majorité des données du projet.

Pour nos choix de structure des métadonnées, nous avons suivi le modèle du projet Nénufar¹¹⁰ qui a déposé toutes ses données sur Nakala avec une description assez complète¹¹¹. Nous avons également consulté les recommandations du consortium CAHIER pour les éditions numériques¹¹². Cela nous a permis, par exemple, de savoir s'il était permis de garder la structuration des informations contenues dans la `<teiHeader>` ou s'il était préconisé de respecter les usages du standard Dublin Core¹¹³. La multiplication des mots-clés nous a semblé importante pour un meilleur référencement dans le moteur de recherche de Nakala et pour un moissonnage extérieur.

¹¹⁰ Voir <https://nenufar.huma-num.fr/presentation/>.

¹¹¹ Voir <https://nakala.fr/collection/10.34847/nkl.be428511>.

¹¹² Fatiha Idmhand et Ioana Galleron, *Guide pour la FAIRisation des données des corpus d'auteurs*, Rapport de recherche, version 2, Huma-Num, 6 septembre 2020.

¹¹³ Voir <https://www.dublincore.org/>.

Ci-après (Tableau 12) nous présentons le tableau des métadonnées retenues, leur type, l'explication de ce qu'elles codent et l'exemple de la valeur de chaque métadonnée pour une des œuvres du corpus (« Une histoire renouvelée des Grecs » de D. L. Eimann).

Tableau 12. Liste des métadonnées remplies pour chaque donnée¹¹⁴.

Nom de la métadonnée	Type	Explication	Exemple (D. L. Eimann)
Type de dépôt* nakala:type	Liste déroulante	Type de la donnée déposée	Édition de sources
Titre* nakala:title	Champ de texte	Titre de la version déposée	Une histoire renouvelée des Grecs. In Histoires cavalières, avec une préface d'Alexandre Dumas
Auteurs* nakala:creator	Nom, Prénom	Auteur de la source	Eimann, D. L.
Date de création* nakala:created	Format YYYY-MM-JJ, YYYY-MM ou YYYY	Date de création de la source. Celle-ci peut être inconnue	1854
Licence* nakala:licence	Liste déroulante	Licence attribuée sur le dépôt, en lien avec le type de la donnée	Creative Commons Attribution Non Commercial No Derivatives 4.0 International
Description dcterms:description	Champ de texte	Description de l'édition électronique reprise des indications du <teiHeader> du fichier	Cette édition électronique fait partie du projet de thèse intitulé « Analyse et édition d'un corpus littéraire dans l'esprit des humanités numériques. Vers une édition comparative intermédiaire innovante accompagnée d'une ressource pour le repérage intertextuel » réalisé par Karolina Suchecka au sein du laboratoire ALTHILA à l'Université de Lille.
Mots-clés dcterms:subject	Liste déroulante pour choisir la langue et champ de texte pour chaque mot-clé		édition en ligne édition électronique intertextualité réutilisation textuelle réécriture Orphée Eurydice TEI traitement automatique des langues TextPAIR Tracer
Langues dcterms:language	Liste déroulante	Langue pour les fichiers des données	Français
dcterms:identifier	Champ de texte	Identifiant de l'œuvre	eimann1854
dcterms:publisher	Champ de texte	Éditeur de la source	Charpentier
dcterms:bibliographicCitation	Champ de texte	Référence bibliographique de la source	D.-L. Eimann / Une Histoire renouvelée des Grecs / Histoires

¹¹⁴ L'astérisque (*) indique le caractère obligatoire de la métadonnée lors du dépôt.

			cavalières / Paris / Charpentier / 1854 / p. 23-35.
dcterms:rights	Champ de texte	Droits d'auteur de la source	Accès libre
dcterms:source	Lien URL	Lien vers l'emplacement de la source (fac-similé)	https://gallica.bnf.fr/ark:/12148/bpt6k5699364p/
dcterms:contributor	Champ de texte	Éditeur(s) de la version numérique	Éditeur : Karolina Suchecka Directeurs du projet : Nathalie Gasiglia, Karl Zieger, Andrea Del Lungo Dépôt Nakala : Karolina Suchecka et Valentin De Craene
dcterms:issued	Format YYYY-MM-JJ, YYYY-MM ou YYYY	Date de publication du fichier électronique	2024
dcterms:isPartOf	Lien URL	Lien vers les collections qui contiennent cette donnée	Lien des collections « Corpus » « Corpus principal : Réécritures » « Genre : Narratif » « Type de réécriture : Parodie »

En plus de cinq premières métadonnées rendues obligatoires par Nakala, nous en ajoutons 11 issues du Dublin Core (préfixe dcterms). Elles permettent de décrire la source de l'œuvre (l'année de publication, la maison d'édition, le lien vers le fac-similé...) et notre projet éditorial (contributeurs, description du projet, licence...). La dernière métadonnée permet de placer l'œuvre dans des collections thématiques (voir § « L'organisation en collections », p. 122).

Nous choisissons de mettre dans une donnée tous les types de fichiers générés à partir du XML initial (HTML, TXT et DOCX), plutôt de créer une collection par type de fichier. Ce choix est motivé par la volonté de faciliter la recherche et l'extraction ciblée des informations par un utilisateur externe. Par ailleurs, le type de fichiers présents dans les données varie en fonction du type de l'ouvrage : pour les quatre œuvres qui n'ont pas pu être structurées selon le schéma Teinte¹¹⁵, nous ne fournissons pas le fichier DOCX. Les fichiers XML, HTML et texte contiennent uniquement les répliques pour les trois bandes dessinées, mais les illustrations sont bien présentes dans le HTML de *La métamorphose d'Ovide figurée* de Jan de Tournes. Au total, 90 fichiers XML sont publiés, 90 fichiers texte, 90 pages HTML, 87 documents DOCX, 68 fac-similés en format PDF, 46 illustrations en format JPG et 16 autres en PNG. Le volume total des données publiées est de 390 Mo. Elles ont été déposées en novembre 2024 et publiées en décembre 2024. Toutes les données sont

¹¹⁵ Voir § « Cas particuliers », p. 108.

disponibles dans l'entrepôt Nakala du projet sous l'identifiant 10.34847/nkl.ba7766ru¹¹⁶, sauf les 12 œuvres encore sous droit qui ont été déposées sous embargo. L'accès aux données relatives sera débloqué après l'expiration de l'embargo (au plus tôt en 2033 pour *Orphée et Eurydice* de Cocteau).

Organisation en collections

En choisissant Nakala comme entrepôt de données, il faut composer avec la spécificité de celui-ci : les collections n'ont pas de niveaux hiérarchiques. En revanche, il est possible de mettre les données dans autant de collections que nécessaire (une collection publique n'acceptant que des données publiées et non seulement déposées¹¹⁷). Sans hiérarchisation, il est plus difficile de rendre compte de la répartition conceptuelle de notre corpus.

En effet, celui-ci se divise en corpus principal (réécritures) et secondaire (traductions). Ensuite, nous proposons plusieurs catégories permettant de classer chaque ouvrage : le genre de l'œuvre (poétique, dramatique, narratif ou autre), le type de réécriture (parodie ou transposition), la transformation du mythe (transmodélisation, modernisation, détournement, idéologique, minimale, pragmatique), la diégèse du récit (homodiégétique ou hétérodiégétique) et le public visé (adultes, jeunesse). Nous différencions également les œuvres canoniques et non-canoniques et précisons, pour certaines, d'autres médias coexistant avec le texte (musique, image). Au sein du corpus secondaire (des traductions), la catégorisation est plus restreinte, car nous différencions seulement les traductions d'Ovide de celles de Virgile et précisons si elles sont écrites en vers ou en prose.

Une partie de cette hiérarchie est pensée comme une arborescence, alors que Nakala demande de la représenter avec un graphe. Pour reconstituer les liens entre collections, nous explorons les champs de Dublin Core comme `dcterms:hasPart`, `dcterms:isPartOf` ou encore `dcterms:relation`. L'utilisateur peut alors retrouver plus aisément l'appartenance d'une collection à une autre. Par exemple, les collections « Traduction d'Ovide » et « Traduction de Virgile » sont en relation entre elles et font partie de la collection « Corpus secondaire : Traductions ».

Il est également nécessaire de mettre chaque donnée dans différentes collections afin de restituer clairement sa place au sein du corpus. Par exemple, la donnée Eimann1854 appartient aux collections « Orphidys », « Corpus », « Corpus principal : Réécriture », « Genre : Narratif » « Type de récri-

¹¹⁶ Voir <https://nakala.fr/collection/10.34847/nkl.ba7766ru>. Les données et que le fichier des métadonnées ayant servi au versement via le service Mynkl (<https://mynkl.huma-num.fr/>) peuvent aussi être consultées en annexes (I 4 Corpus-Nakala).

¹¹⁷ Une donnée déposée est une donnée accessible uniquement au déposant et aux personnes ayant un accès à cette donnée. Une donnée publiée est une donnée visible par n'importe quel utilisateur de Nakala. Si la donnée est visible, cela signifie que ses métadonnées sont visibles, mais le contenu peut être masqué et mis sous embargo. Les utilisateurs sont toutefois au courant de son existence.

ture : Parodie », « Transformation du mythe : Modernisation », « Diégèse du récit : Hétérodiégétique », « Public visé : Adultes » et « Autorité : Non-canon ». Ainsi, les 90 données sont disséminées dans 25 collections (Tableau 13).

Tableau 13. Liste des collections, leur population (nombre de données qu'elles regroupent) et leurs descriptions

Nom de la collection	Nb	Description
Corpus	90	La totalité des œuvres collectées au sein du corpus
Corpus principal : Réécritures	53	Le corpus principal regroupant 53 réécritures du mythe d'Orphée et Eurydice
Corpus secondaire : Traductions	37	Le corpus secondaire regroupant 37 traductions du mythe d'Orphée et Eurydice
Traductions d'Ovide	21	Traduction française des extraits livres X et XI des <i>Métamorphoses</i> d'Ovide relatifs au mythe d'Orphée et Eurydice
Traductions de Virgile	16	Traduction française des extraits du livre IV des <i>Géorgiques</i> de Virgile relatifs au mythe d'Orphée et Eurydice
Type de réécriture : Parodie	10	Parodie, ou transformation hypertextuelle dans le régime ludique, du mythe d'Orphée et Eurydice
Type de réécriture : Transposition	43	Transposition, ou transformation hypertextuelle dans le régime sérieux du mythe d'Orphée et Eurydice
Transformation du mythe : Transmodélisation	25	Transmodélisation (sous-catégorie de transposition) : œuvres qui (à la première vue) n'apporte pas de modifications sémantiques à aucune des versions antiques du mythe
Transformation du mythe : Modernisation	5	Modernisation (sous-catégorie de transposition et de parodie) : œuvre qui place le mythe dans un récit-cadre moderne, réoriente son sens, introduit de personnages modernes, actualise le décor, etc.
Transformation du mythe : Détournement	4	Détournement humoristique (sous-catégorie de parodie) : « texte, ouvrage qui, à des fins satiriques ou comiques, imite en le tournant en ridicule, une partie ou la totalité d'une œuvre sérieuse connue », <i>s.v.</i> parodie , TLFi)
Transformation du mythe : Minimale	9	Transformation minimale (sous-catégorie de parodie) : paraphrase parodique d'une œuvre bien identifiée (Ovide, Virgile ou une autre réécriture du corpus), parodie à transformation minimale (voir Genette, 1982)
Transformation du mythe : Idéologique	3	Transformation idéologique (sous-catégorie de transposition) : œuvre qui met en premier plan la fonction initiatique d'Orphée, en mettant son amour pour Eurydice en second plan
Transformation du mythe : Pragmatique	7	Transformation pragmatique (sous-catégorie de transposition) : œuvre destinée à la jeunesse, qui emploie des procédés particuliers visant l'adaptation du mythe pour le jeune public (omission, adoucissement, modification)
Diégèse : Homodiégétique	39	Transposition homodiégétique : œuvre conserve le cadre mythique (Grèce antique, Enfers, etc.)
Diégèse : Hétérodiégétique	14	Transposition hétérodiégétique : œuvre transpose le mythe dans un autre décor
Genre : Poétique	29	Œuvre qui répond aux caractéristiques du genre poétique (ex. versification), ou traduction en vers
Genre : Dramatique	19	Œuvre qui répond aux caractéristiques du genre dramatique, comme la présence des didascalies
Genre : Narratif	35	Œuvre qui répond aux caractéristiques du genre narratif (ex. la présence du narrateur), ou traduction en prose
Genre : Autre	3	Œuvre qui ne répond pas de manière univoque aux caractéristiques des trois autres genres (bande dessinée)

Nom de la collection	Nb	Description
Intermédialité : Image	15	La présence des images au sein de l'œuvre, de manière accessoire (illustrations) ou substantielle (bande dessinée)
Intermédialité : Musique	22	La présence de la musique au sein de l'œuvre (cantate, opéra, ballet, etc.)
Public : Adultes	46	L'œuvre est destinée à un public adulte
Public : Jeunesse	7	L'œuvre est destinée à la jeunesse
Autorité : Canon	13	L'œuvre considérée comme canonique a fait objet des adaptations, réédition, analyses scientifiques, etc.
Autorité : Non-canon	40	L'œuvre considérée comme non-canonique est difficile d'accès et généralement ignorée par la critique

La multiplication des collections permet, d'une part, de catégoriser et décrire chaque donnée de manière très précise et, d'autre part, de faciliter le regroupement des œuvres du corpus selon différents critères. Par exemple, la collection « Genre : narratif » permet de rassembler tant les réécritures du genre narratif que les traductions en prose d'Ovide et de Virgile, alors que le filtrage des œuvres du genre dramatique appartenant à la collection « Intermédialité : musique » recense toutes les opéras, cantates et ballets du corpus. La collection « Corpus » contient la totalité des œuvres. Nakala permet également d'ajouter une description à chaque collection et de définir les mots-clés la décrivant, ce qui est particulièrement utile notamment pour certaines catégories qui ne sont pas explicites et restent très spécifiques à notre recherche (« Transformation : idéologique », « Diégèse : hétérodiégétique »).

Navigation dans l'entrepôt avec une vitrine Web

Une fois l'ensemble des collections publiées, il est possible de mettre en place un site Nakala. Ce site n'a pas pour vocation de remplacer l'édition numérique du projet, mais de faciliter la lecture et la navigation au sein des données partagées. Dans la même veine que les expositions du *Research Data Journal for the Humanities and Social Sciences* (Brill)¹¹⁸, le but de cette présentation des données est de rendre la recherche des ressources dans l'entrepôt moins fastidieuse pour les personnes non-initiées aux services et outils d'Huma-Num. Ainsi, les données peuvent être consultées sur le site <https://orphidys.nakala.fr>.

Différents types de contenu peuvent être créés à partir de la page d'accueil (lien vers une ressource externe, liste de données, métadonnées, visualisation du contenu...). Les données contenues dans la collection et les métadonnées sont mises en forme, ce qui facilite leur lecture. En cliquant sur « Chronologie des publications », il est également possible de filtrer les données par année de création. Ainsi, un utilisateur extérieur peut facilement récupérer uniquement un échantillon chronologique du corpus qui l'intéresse. De même, il est possible de voir les différentes collections

¹¹⁸ Voir par exemple la page dédiée aux données du projet *The Huguenots Online* : <https://primarysources.brillonline.com/browse/huguenots>.

auxquelles les données sont liées. Une donnée pouvant être contenue dans plusieurs collections, il est plus aisé pour un utilisateur extérieur de naviguer, par exemple, directement dans la collection « Genre : narratif » à partir de la collection principale du projet depuis laquelle le site Nakala est généré.

Des pages annexes ont été ajoutées pour structurer davantage le site. Il s'agit notamment des liens vers le repository GitLab du projet et vers les outils exploités dans l'établissement du corpus (Teinte, Odette, Deteint, etc.). Un moteur de recherche est également disponible. Ainsi, il s'agit non seulement de rendre les données disponibles, mais aussi visibles, et surtout lisibles par des personnes extérieures au projet, voire peu familières des outils et méthodes des humanités numériques. La consultation de l'entrepôt n'est peut-être pas encore un réflexe, mais la navigation sur un site Web est une pratique généralement démocratisée.

Conclusion du chapitre

En présentant de manière détaillée la façon dont nous numérisons, transformons et partageons notre corpus, nous ambitionnons que notre travail soit réutilisable et répliquable, que ce soit au niveau du jeu de données, que des pratiques et méthodes décrites.

C'est pour cela que l'établissement rigoureux du format XML-TEI a été priorisé, puisque celui-ci est considéré comme un format pivot, à partir duquel nous procédons aux transformations automatiques vers d'autres formats éditoriaux. La transformation du XML-TEI dans les formats HTML, DOCX et TXT permet la diffusion des textes au public plus large et l'utilisation de l'édition au-delà du milieu universitaire (dans le cadre scolaire, voire pour une lecture de plaisir). L'emploi du standard TEI garantit l'interchangeabilité et l'ouverture des données produites dans le cadre du projet et l'emploi de la syntaxe Teinte, plus restreinte, mais adaptée au projet et développée avec lui, permettent de proposer des textes dotés d'une structuration sémantique.

L'automatisation de la chaîne du traitement garantit par ailleurs la conformité des versions des fichiers et facilite la correction plus rigoureuse d'éventuelles coquilles. La description de chaque donnée par des métadonnées aussi détaillées que possible améliore son indexation en la rendant plus facile à trouver et, par conséquent, à réutiliser. Enfin, la multiplication des collections Nakala permettant de classer chaque donnée de manière détaillée et nous la présentons dans des formats différents.

Comme le constatent Vitali-Rosati et Sinatra,

[I]l ne s'agit pas seulement de choisir, de légitimer, de mettre en forme et de diffuser un contenu, mais il s'agit aussi de réfléchir à l'ensemble des techniques que l'on va utiliser ou créer pour le faire, ainsi qu'aux contextes de circulation produits par l'espace numérique. Si les humanités numériques s'occupent de produire des outils et de réfléchir à leur impact sur la production et la circulation du savoir, alors l'éditorialisation devient l'objet central de leur travail.¹¹⁹

Le concept de la science ouverte invite à donner accès non seulement aux résultats, mais aussi aux chaînes de traitement qui ont permis leur production. Nous sommes persuadée que le dépôt des données collectées et produites au sein de chaque projet, accompagné d'une documentation facilitant leur réutilisation, permet de remettre l'éditorialisation au centre de la réflexion sur les objets et outils numériques.

¹¹⁹ Marcello Vitali-Rosati et Michael E. Sinatra, « Pratiques de l'édition numérique », *op. cit.*, p. 60.

Conclusion de la partie

Dans cette première partie, qui présente notre manière de collecter, classer et éditer le corpus de réécritures françaises du mythe d'Orphée et Eurydice, nous avons combiné des méthodes et théories littéraires et linguistiques avec des usages et outils numériques. Cette interdisciplinarité, qui accompagne l'entièreté de nos travaux, n'est pas toujours aisée. Tout au cours de cette partie, nous nous sommes employée à chercher un consensus entre les méthodes traditionnelles et computationnelles, entre la recherche empirique et l'exploitation des outils informatiques, ou encore entre la théorie littéraire et la pratique de ce qui est observable concrètement au sein d'un corpus.

L'alliance du quantitatif et de l'herméneutique est nécessaire, mais elle implique d'adapter et de combiner certaines méthodes à la spécificité d'une recherche particulière. Par exemple, pour respecter au maximum les usages linguistiques et talistes, nous avons défini un protocole clair et objectivable de la collecte du corpus, nous avons veillé à ce que l'ensemble des réécritures soit représentatif et nous avons gardé un équilibre entre les œuvres canoniques et celles dépréciées ou oubliées. Toutefois, nous avons parfois laissé de côté ces restrictions et principes préétablis en intégrant au corpus les œuvres dont le titre ne porte pas le nom d'Orphée ou d'Eurydice ou en complétant les méthodes objectivables de la collecte du corpus par une fouille manuelle et plus subjective. Ces décisions, motivées et explicitées, ont réintroduit une forme de subjectivité de notre corpus, nécessaire dans les travaux littéraires. Inversement, nous avons réalisé que le classement et description du corpus collecté n'est pas facilité par la critique littéraire. Face à une expérimentation concrète et objectivable, appuyée sur un corpus spécifique, le manque de consensus dans la théorisation des phénomènes de réécriture ou encore du classement des genres littéraires nous a incitée à multiplier différents critères, quitte à combiner de théories parfois antagonistes.

Malgré ces problèmes épistémologiques, nous avons réussi à collecter un corpus de 53 réécritures françaises du mythe d'Orphée et Eurydice publiées entre XVI^e et XXI^e s. Documenté et classé à l'aide de huit étiquettes différentes, il constitue une ressource précieuse tant pour notre thèse que, nous l'espérons, pour d'autres recherches, qu'elles soient littéraires, linguistiques ou informatiques.

Ainsi, nous avons veillé à exploiter les outils, méthodes et usages des humanités numériques pour faciliter l'accessibilité et la réutilisabilité de ce corpus. Après avoir établi le texte de chaque œuvre, nous l'avons structuré en format XML-TEI, ce qui a garanti l'interchangeabilité du corpus avec d'autres projets utilisant le même standard et nous a permis d'exploiter les outils conçus pour générer automatiquement une édition numérique à partir de ce format pivot. Nous avons déposé la totalité des données ainsi conçues et générées au sein de l'archive de dépôt Nakala où chaque œuvre éditée est stockée de manière pérenne, dotée d'un identifiant unique et mise à disposition de

la communauté des chercheurs. En cela, nous avons respecté les attendus de nos pairs concernant la conception, la préservation et le partage des corpus littéraires destinés à être édités numériquement.

La meilleure manière d'exposer la pertinence de nos choix d'établissement du corpus est de le confronter à nos questions de recherche. Ces différentes réécritures du mythe d'Orphée et Eurydice sont-elles suffisamment hétérogènes et riches pour l'expérimentation que nous menons avec des outils de la détection automatique des relations intertextuelles ? Avons-nous omis des œuvres majeures, reprises dans les textes du corpus ? L'établissement du texte, et notamment le nettoyage des océrisations, a-t-il été fait avec suffisamment de soin pour ne pas introduire de bruits dans nos traitements informatiques ? La pertinence du classement manuel du corpus pourrait-elle être confirmée par des traitements objectivables et (semi-)automatiques ? Enfin, la manière dont les données sont décrites et présentées au sein de l'archive Nakala est-elle optimale pour faciliter leur réutilisation ?

Alors que la réponse à la dernière question dépend des facteurs externes, comme l'évaluation et l'appréciation de notre projet par la communauté scientifique, dans les parties suivantes nous n'aurons de cesse de démontrer la pertinence de la collecte, du classement et de l'établissement de notre corpus. À présent, il est impératif de nous pencher de manière approfondie sur une des notions clés de notre recherche, l'intertextualité. Sa définition, qu'elle soit littéraire ou linguistique, a suscité beaucoup de querelles. Nous espérons que notre corpus nous permettra encore une fois de pallier le manque de consensus théorique en ayant recours à l'observation pratique des textes. Plusieurs méthodes et outils numériques peuvent faciliter cette analyse sans pour autant être dédiés spécifiquement à la détection automatique des relations intertextuelles. Nous les expérimentons dans la deuxième partie de cette thèse, après avoir rappelé la théorisation littéraire de l'intertextualité et avoir observé les procédés linguistiques caractéristiques des relations intertextuelles.

II. Intertextualité : de la théorie littéraire aux méthodes computationnelles

Dans « L'intertextualité : enquête sur l'émergence et la diffusion d'un champ notionnel »¹, Marc Angenot considère que l'intertextualité « conduit à ouvrir des “réseaux de connexions” qui mettent en évidence l'hétérogénéité des matériaux prélevés à d'autres textes ». Dans l'appendice au *Dictionnaire encyclopédique des sciences du langage*, François Wahl² parle d'un « réseau de connexions multiples à hiérarchie variable », grâce auquel le texte échappe aux règles prédéterminantes de la langue. Cette notion fait écho à ce que nous observons en analysant, à l'aide d'outils numériques, les relations entre les textes de notre corpus, du contenu comparable – puisqu'ils sont tous, d'une manière ou d'une autre, relatifs au mythe d'Orphée et Eurydice –, mais pour lesquels le sujet mythologique semble souvent être le seul point commun.

La définition des relations entre les textes de notre corpus est problématique pour plusieurs raisons. Dans la partie précédente, nous avons évoqué les difficultés macrostructurelles et nous avons montré qu'il n'est pas aisé d'établir un classement homogène et exhaustif des relations que les réécritures réunies dans notre corpus entretiennent entre elles et avec les sources antiques du mythe. Dans cette partie, nous nous intéressons à l'observation des relations plus détaillées, que nous pouvons considérer comme microstructurelles. Le concept qui semble s'y prêter le mieux, l'intertextualité, est très vaste et souvent sujet à controverses. Certains chercheurs, par exemple, considèrent que les œuvres relatives au même mythe ne peuvent pas être qualifiées d'intertextuelles. D'autres auront plutôt tendance à en exclure les traductions. Par ailleurs, la frontière entre l'intertextualité, la réécriture, et la critique traditionnelle de sources littéraires ne paraît pas clairement délimitée. Enfin, des questions et problèmes particuliers surgissent quand les concepts théoriques sont confrontés à ce qui est observé concrètement au sein d'un corpus particulier.

Rappelons également que les sources théoriques de l'intertextualité ne sont pas littéraires, mais linguistiques : Julia Kristeva³ appuie sa théorie sur le dialogisme de Mikhaïl Bakhtine ; Michel Arrivé puise dans les théories linguistiques de Greimas, Hjelmslev et Chomsky et définit l'intertexte comme « les isotopies connotées »⁴ ; dans *La Seconde main, ou le travail de la citation*⁵, Antoine Compagnon propose une « sémiologie de la citation » où il présente une typologie formelle évoquant les théories sémiotiques de Peirce et renvoyant aux travaux de Benveniste. Henri Béhar⁶, lui, conçoit un tableau d'opérations intertextuelles qui reprend les six fonctions du langage de Jakobson.

¹ Marc Angenot, « L'intertextualité : enquête sur l'émergence et la diffusion d'un champ notionnel », *Revue des sciences humaines*, n° 189, 1983, p. 121-135.

² François Wahl, « Texte [375]. Le texte comme productivité », dans Oswald Ducrot et Cvetan Todorov (dir.), *Dictionnaire encyclopédique des sciences du langage*, Paris, Seuil, coll. « Points Série essais », 1993, p. 443-448.

³ Julia Kristeva, *Séméiotikè. Recherches pour une sémanalyse*, Paris, Seuil, coll. « Tel quel », 1969.

⁴ Michel Arrivé, *Les Langages de Jarry : essai de sémiotique littéraire*, Paris, Klincksieck, 1972.

⁵ Antoine Compagnon, *La seconde main, op. cit.*

⁶ Henri Béhar, « Le collage, ou le pagure de la modernité », *Cahiers du XX^e siècle*, n° 5, 1975, p. 43-68.

L'intertextualité reste également présente dans quelques branches des sciences du langage, et notamment la linguistique textuelle : Jean-Michel Adam, par exemple, inclut le “dialogisme intertextuel” comme une des opérations du liage qui assure la continuité textuelle⁷. Nous constatons ainsi que l'étude de notre corpus ne peut se faire qu'en combinant les méthodes littéraires et linguistiques. Aux théories intertextuelles, purement littéraires ou hybridant les approches des deux disciplines, s'ajoutent celles uniquement linguistiques, qui s'intéressent de manière générale aux processus de reformulation, tout autant au niveau global d'un texte qu'à l'échelle des phrases, des syntagmes, et même des unités simples.

Si l'intertextualité gagne à être considérée conjointement du point de vue littéraire et linguistique, elle peut être observée sous un axe très différent quand elle est étudiée à l'aide d'outils numériques. Nombreuses disciplines s'intéressent à la détection automatique des réutilisations et reformulations dans des textes, que ce soit pour l'identification du plagiat⁸, l'attribution de l'auctorialité des œuvres anonymes⁹ ou encore la génération des résumés¹⁰. Dans le cadre des études littéraires appuyées sur des outils informatiques, des projets se focalisent sur la détection des relations dans des textes anciens, en grec¹¹ et en latin¹², l'alignement de différentes traductions d'une même œuvre¹³, ou encore la recherche des sources, voire des inspirations intellectuelles d'un auteur. Différents algorithmes ont été expérimentés pour améliorer la pertinence des relations détectées automatiquement, et plusieurs logiciels ont vu le jour. Ces approches se révèlent utiles pour étudier l'intertextualité en tant qu'un « réseau des connexions » et pour révéler des phénomènes globaux qu'il eût été difficile d'analyser avec des outils empiriques seuls. Toutefois, comme nous le démontrons, cela ne peut se faire qu'en combinant des traitements (semi-)automatiques avec des analyses empiriques effectuées à l'aide des méthodes linguistiques et littéraires.

⁷ Jean-Michel Adam, *La linguistique textuelle*, 3^e éd., Paris, A. Colin, coll. « Cursus », 2011 [2005].

⁸ Jérémy Ferrero et Alain Simac-Lejeune, « Détection automatique de reformulations – Correspondance de concepts appliquée à la détection du plagiat », *Revue des Nouvelles Technologies de l'Information*, vol. Extraction et Gestion des Connaissances, RNTI-E-28, 2015, p. 287-298.

⁹ Jan Rybicki, « Vive la différence : Tracing the (authorial) gender signal by multivariate analysis of word frequencies », *Digital Scholarship in the Humanities*, vol. 31, n° 4, 2016, p. 746-761.

¹⁰ Regina Barzilay et Michael Elhadad, « Using Lexical Chains for Text Summarization », *Intelligent Scalable Text Summarization*, 1997, en ligne, <<https://www.aclweb.org/anthology/W97-0703>>.

¹¹ Voir par exemple Marco Büchler et al., « Detection of citations and text reuse on ancient greek texts and its applications in the classical studies: eAQUA project », *Proceedings of the Digital Humanities*, 2010 ; Greta Franzini, Emily Franzini, Marco Büchler, et al., « Towards a Historical Text Re-use Detection », dans C. Biemann et A. Mehler (dir.), *Text Mining. From Ontology Learning to Automated Text Precession Applications*, Suisse, Springer International Publishing, 2014, p. 221-238.

¹² Neil Coffee et al., « The Tesseract Project: intertextual analysis of Latin poetry », *Literary and Linguistic Computing*, vol. 28, n° 2, 2012, p. 221-228, en ligne, <https://www.academia.edu/3830928/The_Tesseract_Project_Intertextual_Analysis_of_Latin_Poetry>.

¹³ Les travaux portant sur l'alignement des versions ou traductions d'une œuvre sont nombreux. Parmi ceux appuyés sur des techniques de détection automatique de relations, cf. par exemple Marianne Reboul, « Comparaison semi-automatique des traductions en langue française de l'Odyssée d'Homère (1547-1955) », Thèse de doctorat, 2017.

Dans cette partie, nous étudions le concept de l'intertextualité en insistant sur sa pluridisciplinarité. Tout d'abord, nous observons la complexité conceptuelle du terme, sa définition difficile, son lien avec d'autres concepts, ou encore les imprécisions concernant les phénomènes qu'il inclut. Ensuite, nous couplons les réflexions épistémologiques et l'observation pratique de notre corpus de point de vue littéraire en répertoriant quelques types de relations intertextuelles présentes en son sein et nous demandons comment ce qui est observé est analysable selon divers postulats des approches théoriques.

Nous nous intéressons également aux caractéristiques linguistiques des textes, en partant des différences entre les états de la langue, qui ne se limitent pas à celles entre graphies, anciennes ou modernes, mais concernent aussi le vocabulaire employé, la morphologie, la syntaxe et la sémantique. Nous examinons ensuite la présence des phénomènes caractéristiques de la reformulation, comme la synonymie ou la paraphrase, et nous nous demandons si des règles de la langue générale, formalisées par exemple dans des dictionnaires, peuvent être appliquées à la spécificité de notre corpus.

Enfin, nous observons le renouvellement du concept de l'intertextualité dans la recherche littéraire appuyée sur des outils computationnels en introduisant quelques méthodes les plus utilisées dans le traitement computationnel de textes littéraires, comme la stylométrie, la modélisation thématique ou l'alignement des séquences. Comme nous le constatons en les appliquant sur notre corpus, certains outils et certaines méthodes semblent plus aptes que d'autres à répondre à nos besoins.

1. Intertextualité dans la théorie littéraire

« [T]out texte est un *intertexte* ; d'autres textes sont présents en lui, à des niveaux variables, sous des formes plus ou moins reconnaissables ; les textes de la culture antérieure et ceux de la culture environnante ; tout texte est un tissu nouveau de citations révolues » : cette célèbre citation de Roland Barthes, tirée de son article « Texte (théorie du) » de l'*Encyclopædia Universalis* (1973)¹ présente l'intertexte comme un phénomène obligatoire de chaque écriture. Mais elle est aussi, d'une certaine manière, une reformulation de son analyse morphologique : rappelons que le radical de la *textualité* est dérivé du latin classique *texere*, « tisser », avant de devenir, en latin médiéval, *textualis*, « relatif au texte ». Selon Nathalie Limat-Letellier², « la textualité évoque la qualité du texte comme "tissage", "trame" ; d'où un redoublement sémantique de l'idée de réseau, intersection » (§ 1). À cela s'ajoute le préfixe *inter-*, qui implique la réciprocité des échanges.

Le recours à l'analyse morphologique nous permet-il de comprendre pleinement le phénomène qui a tant occupé la critique littéraire tout au long du XX^e s. ? Sans qu'il soit erroné ou dépourvu d'intérêt, il reste, indubitablement, une reconstitution incomplète et artificielle. Comme le souligne Limat-Letellier, « pour avoir été reconnue comme une dimension essentielle de la communication verbale, l'intertextualité renvoie surtout à des enjeux cognitifs, à l'élaboration de méthodes d'analyse littéraire actuellement très usitées »³.

15 ans après l'article de Barthes, Pierre-Marc de Biais propose, pour la première fois dans l'*Encyclopædia Universalis* (1989), une définition consacrée spécifiquement à l'intertextualité⁴. Avant de procéder à une genèse de ce concept, de Biais constate qu'« en un quart de siècle, ce concept a suscité beaucoup de controverses et ne s'est finalement imposé qu'après plusieurs refontes définitionnelles »⁵. La controverse principale semble être liée à la simplification, voire la « vulgarisation pédagogique » du terme :

[L]'intertextualité devient un concept plus maniable, et plus rassurant aussi, puisque son champ d'application ne paraît plus si éloigné du domaine traditionnel de la critique des « sources », et qu'on peut de proche en proche y adjoindre les secteurs tout aussi classiques de l'étude du pastiche, de la parodie, et – pourquoi pas ? – plusieurs des grandes problématiques de la littérature comparée. Mais un tel élargissement, tout en contribuant beaucoup à généraliser l'usage de ce concept, ne sera pas étranger à un certain flou théorique où l'intertextualité finira par perdre, pour un temps, l'essentiel de sa spécificité notionnelle.⁶

¹ Roland Barthes, « Texte (Théorie du) », dans *Encyclopædia Universalis*, 1973.

² Nathalie Limat-Letellier, « Historique du concept d'intertextualité », *op. cit.*

³ *Ibid.*

⁴ Pierre-Marc de Biais, « Intertextualité (Théorie de l') », dans *Encyclopædia Universalis*, 1989.

⁵ *Ibid.*, p. 514.

⁶ *Ibid.*

La simplification malvenue, le flottement définitionnel, voire même le désenchantement conceptuel semblent traverser la plupart des travaux de synthèse. Dans son intervention d'ouverture du colloque *Le Plaisir de l'intertexte* en 1986, Michel Arrivé est attendu pour proposer un résumé des théories intertextuelles qui occupaient la critique littéraire depuis les vingt années précédentes. Mais il n'en est rien :

On l'a compris : sauf à entrer dans d'infinis détails, ou à se contenter d'approximations, il était impossible, ici, de procéder à une synthèse. [...] La synthèse d'ensemble exclue, restait une autre possibilité : une mise au point terminologique. J'ai essayé. Et j'ai renoncé. Parce que je rencontrais, au niveau de la terminologie, les problèmes insupportables que j'avais préalablement rencontrés.⁷

La synthèse historique en trois temps que nous proposons ne permet ni de lever le flou terminologique ni de proposer une classification stable de différents phénomènes. En revanche, pour chaque approche, qu'elle soit philosophique ou pratique, orientée vers l'auteur ou vers le lecteur, nous nous efforçons de comprendre comment elle peut être appliquée à notre corpus des réécritures. Certaines conceptualisations n'admettent pas les références mythologiques parmi les relations intertextuelles, d'autres proposent des typologies qui brouillent la frontière entre l'intertextualité et la réécriture. Pour d'autres encore, le recensement des relations intertextuelles au sein d'un texte est malvenu et qualifié de « vieille critique des sources »⁸. Il n'en reste pas moins que, pour chaque théorie, des exemples concrets peuvent être trouvés au sein de notre corpus, prouvant l'hétérogénéité de ce qui est réellement observable et rendant difficile la restriction de notre compréhension de l'intertextualité à une théorie particulière.

Les auteurs de l'introduction de l'ouvrage *Le mythe en littérature. Essais en hommage à Pierre Brunel*, insistent, en évoquant la méthodologie que Brunel applique au mythe d'Électre, sur le plaisir, pour un lecteur, de disposer « d'un dossier, naturellement sélectif et ordonné, qui, en aucune façon, ne s'efforçait de remanier tout à un dénominateur commun, mais à partir duquel il était loisible d'envisager des prolongements, des rapprochements, des interrogations »⁹. Nous nous laissons guider par cet idéal dans la synthèse qui suit.

⁷ Michel Arrivé, « Intertextualité et intratextualité chez Ferdinand de Saussure », *Le Plaisir de l'intertexte*, Berne, Peter Lang, 1986, p. 11-36.

⁸ Voir par exemple Daniel Ferrer, « Quelques remarques sur le couple intertextualité-génèse », *La création en acte. Devenir de la critique génétique*, vol. 289, Brill, 2007, p. 203, en ligne, <<https://brill.com/view/book/9789401203531/B9789401203531-s015.xml>>.

⁹ Yves Chevrel et Camille Dumoulié (dir.), « Introduction », *Le mythe en littérature. Essais en hommage à Pierre Brunel*, éd. 1, Paris, Presses universitaires de France, coll. « Ecriture », 2000, p. 3.

a) Kristeva et Barthes : approches philosophiques aux sources linguistiques

Le dialogisme de Mikhaïl Bakhtine¹⁰ est généralement considéré comme fondement de la théorie de l'intertextualité, exposée pour la première fois par Julia Kristeva dans deux articles : « Le mot, le dialogue et le roman » (1966) et « Le texte clos » (1966-67), tous les deux repris ensuite dans *Séméiotikè* (1969)¹¹. La théorie bakhtinienne se caractérise par une certaine bivalence. Le dialogisme est, d'une part, inhérent à la lecture et à la spécificité de la langue : tout mot ayant déjà servi, chaque énoncé linguistique renvoie à d'autres textes, et est donc obligatoirement intertextuel. Il s'agit alors du dialogisme interne (*dialogue*). L'autre facette, le dialogisme externe, est plus volontaire, elle peut procéder d'un choix formel, par exemple pour faire appartenir le texte à une esthétique précise et globalement reconnue. Bakhtine prend l'exemple du roman dialogique et de la satire ménippée pour montrer que la polyphonie peut être introduite dans le texte grâce à la répétition des textes antérieurs. Dans ce cas-là, il s'agit d'une répétition ambivalente, pour disqualifier l'énoncé parodié, voire pour l'imiter de manière détournée (stylisation, hybridation, etc.).

Pour illustrer le rôle de l'intertextualité, Kristeva rapproche ces deux types du dialogisme à la dichotomie, fondamentale en linguistique, entre l'axe paradigmatique et l'axe syntagmatique :

[L]'axe horizontal (sujet-destinataire) et l'axe vertical (texte-contexte) coïncident pour dévoiler un fait majeur : le mot (le texte) est un croisement de mots (de textes) où on lit au moins un autre mot (texte). Chez Bakhtine d'ailleurs, ces deux axes, qu'il appelle respectivement *dialogue* et *ambivalence*, ne sont pas clairement distingués. Mais ce manque de rigueur est plutôt une découverte que Bakhtine est le premier à introduire dans la théorie littéraire : tout texte se construit comme une mosaïque de citations, tout texte est absorption et transformation d'un autre texte. À la place de la notion d'intersubjectivité s'installe celle d'*intertextualité*.¹²

Le remplacement de l'intersubjectivité par l'intertextualité fait écho aux théories critiques en plein essor à l'époque. Alors que la première notion implique une certaine complaisance entre l'écrivain et ses lecteurs, la Nouvelle Critique tend vers l'abolition de la figure de l'auteur : « il s'agit avant tout de désacraliser l'autorité de l'auteur, de le destituer de son illusion d'originalité, et de récuser par là même les prérogatives de l'œuvre finie, achevée, autonome ; le déni de l'individualité, l'impersonnalité de l'acte d'écriture, tels sont les postulats de l'intertextualité dans sa première acception »¹³.

¹⁰ Voir notamment Mikhaïl Bakhtine, *Problèmes de la poétique de Dostoïevski*, traduit par Guy Verret, Lausanne, L'Âge d'Homme, 1970 [1929] ; Mikhaïl Bakhtine, *La Poétique de Dostoïevski*, traduit par Isabelle Kolitcheff, Paris, Seuil, 1970 [1929] ; Mikhaïl Bakhtine, *L'œuvre de François Rabelais et la culture populaire au Moyen Âge et sous la Renaissance*, traduit par Andrée Robel, Paris, Gallimard, coll. « Tel », n° 70, 1970 [1965].

¹¹ Julia Kristeva, *Séméiotikè. Recherches pour une sémanalyse*, *op. cit.*

¹² *Ibid.*, p. 145-146.

¹³ Nathalie Limat-Letellier, « Historique du concept d'intertextualité », *op. cit.*, paragr. 6.

Kristeva comprend l'intertextualité en tant que phénomène très vaste : la littérature est, pour elle, essentiellement intertextuelle. Simultanément, l'application de cette notion à certains types d'expressions (les mythes), voire aux expressions hybrides (l'adaptation d'un mythe au théâtre) paraît admise par l'auteure¹⁴. L'intertextualité pourrait alors être assimilée à l'évolution diachronique des arts, à condition que ces pratiques soient, à la fin, incorporées dans un texte. Cette extension vers d'autres supports est toutefois considérée comme dangereuse. Selon Limat-Letellier, « d'une manière toujours ambivalente, l'extension démesurée du concept en affaiblit la validité : ce trop puissant levier de la sémiotique lui ôte la rigueur que lui garantissait son sens étymologique »¹⁵.

Certes, en souhaitant s'éloigner de l'intersubjectivité, voire de l'interdiscursivité, Kristeva, de par le choix du terme, a impliqué la restriction du phénomène aux relations entre des textes uniquement. Mais, comme le stipule Barthes, « [p]assent dans le texte, redistribués en lui, des morceaux de codes, de formules, des modèles rythmiques, des fragments de langages sociaux, etc., car il y a toujours du langage avant le texte et autour de lui »¹⁶. La conception du texte de Kristeva, nous semble-t-il, est à mettre en relation avec ses influences linguistiques, le dialogisme, la linguistique structurale, et aussi la grammaire générative de Noam Chomsky. Cette dernière inspiration est particulièrement visible dans l'article « Pour une sémiologie des paragrammes » (1967) où la chercheuse propose une réflexion sur la signifiante considérée comme un système paragrammatique en se référant aux modèles formels des mathématiques¹⁷. Kristeva propose aussi, lors d'une communication au Colloque de Cluny (1968)¹⁸, une dichotomie entre un phéno-texte, le texte accompli, et le géno-texte, le texte en cours de conception¹⁹. L'intertextualité appartiendrait à ce dernier niveau de structuration, plus profond, qui peut être assimilé au concept de la compétence virtuelle proposée par Chomsky²⁰. L'inclusion de l'intertexte au sein du géno-texte impliquerait davantage l'aspect non-obligatoire de l'écrit (le texte peut être pensé à partir d'un objet pictural, un récit oral, etc., ou représenté, en cours de conception, en tant que suite d'images, schémas, voire enregistrements audiovisuels).

¹⁴ « [Le texte peut être assimilé à] une *productivité*, ce qui veut dire : 1. son rapport à la langue est redistributif (destructivo-constructif), par conséquent il est abordable à travers des catégories logiques plutôt que purement linguistiques ; 2. il est une permutation de textes, une intertextualité : dans l'espace d'un texte plusieurs énoncés, pris à d'autres textes, se croisent et se neutralisent. », Julia Kristeva, *Sémiotikè. Recherches pour une sémanalyse*, *op. cit.*, p. 113.

¹⁵ Nathalie Limat-Letellier, « Historique du concept d'intertextualité », *op. cit.*, paragr. 9.

¹⁶ Roland Barthes, « Texte (Théorie du) », *op. cit.*

¹⁷ Julia Kristeva, « Pour une sémiologie des paragrammes », *Tel Quel*, n° 29, 1967, p. 53-75.

¹⁸ Julia Kristeva, « Problèmes de la structuration du texte », *La Nouvelle Critique*, n° spécial, « Linguistique et littérature », 1968, p. 55-64.

¹⁹ « Aux deux types de structures, celles de compétence et de performance, correspondraient le géno-texte, c'est-à-dire le niveau où le texte est pensé, transformé, produit, généré, et le phéno-texte, c'est-à-dire le niveau du texte accompli, du phénomène textuel, de ce résidu dans lequel bascule le processus de production et qui est toujours moins que le processus de transformation antérieur au produit », *Ibid.*, p. 60.

²⁰ Noam Chomsky, *Structures syntaxiques*, trad. Michel Braudeau, Paris, Seuil, 1969 [1957].

Finalement, quand Kristeva finit par s'éloigner de l'intertextualité, c'est à de la simplification du concept dans les usages postérieurs. Dans *La Révolution du langage poétique* (1974), la notion de l'intertextualité n'apparaît pas dans la table de matières, et est remplacée par celle de *transposition* :

Nous connaissons les deux « procédés » fondamentaux que Freud désigne dans le travail de l'inconscient : le *déplacement* et la *condensation*. Kruszewski et Jakobson les ont, d'une autre façon, introduits dans les débuts de la linguistique structurale par les concepts de *métonymie* et de *métaphore* qu'on a pu interpréter à la lumière de la psychanalyse.

Il nous faut ajouter un troisième « procédé » : le *passage d'un système de signe à un autre*. Que le déplacement et la condensation s'y conjuguent pour l'effectuer ; n'épuise pas l'ensemble de l'opération. Il s'y ajoute une transformation de la *position* théique : la destruction de l'ancienne et la formation d'une autre. Le nouveau système signifiant peut être produit dans le même matériau signifiant : par exemple, dans le langage, le passage peut s'effectuer de la narration au texte ; mais il peut être emprunté à un matériau signifiant différent : par exemple, le passage peut s'effectuer de la scène carnavalesque au texte écrit [...]. Le terme d'*inter-textualité* désigne cette transposition d'un (ou de plusieurs) système(s) de signes en un autre ; mais, puisque ce terme a été souvent entendu dans le sens banal de « critique des sources » d'un texte, nous lui préférons celui de *transposition*, qui a l'avantage de préciser que le passage d'un système signifiant à un autre exige une nouvelle articulation du théique – de la positionnalité énonciative et dénotative.²¹

Toujours dans le sillage de la linguistique moderne, Kristeva propose donc de compléter la poétique de Jakobson par le procédé qu'elle a conceptualisé jusqu'alors sous la notion de l'intertextualité. La *transposition* ne définit pas un nouveau phénomène, mais, tout en prenant la distance de l'intertextualité, devenue trop éloignée de sa théorie initiale, elle paraît convenir mieux à un phénomène qui ne concerne pas uniquement les relations entre textes.

Même si, comme Kristeva l'admet dans sa préface de *La Poétique de Dostoïevsky*, pour appuyer la théorie de l'intertextualité, le sens des propos de Bakhtine a été légèrement détourné²², l'importance du concept au sein de la recherche littéraire dépasse le cercle initial de la Nouvelle Critique et donne naissance à de nombreuses conceptualisations postérieures, dont certaines, comme celle que Michel Arrivé a exposée en 1972²³, continuent à être fortement inspirées par les théories linguistiques.

Au contraire, dans le chapitre « Intertextualité » du *Plaisir du texte*²⁴, Barthes laisse de côté les théories scientifiques de ses prédécesseurs :

²¹ Julia Kristeva, *La révolution du langage poétique : l'avant-garde à la fin du XIX^e siècle, Lautréamont et Mallarmé*, Paris, Seuil, coll. « Points », 1974, p. 60.

²² Mikhaïl Bakhtine, *La Poétique de Dostoïevski*, op. cit., p. 10. À ce sujet, voir également Serge Rolet, « L'affleurement du contre-texte chez Volochinov et Bakhtine », dans Alain Trouvé (dir.), *La lecture littéraire*, n° 12, Reims, « Le contre-texte », 2014, p. 17-30, n. 3 et 4.

²³ Michel Arrivé, *Les Langages de Jarry*, op. cit.

²⁴ Roland Barthes, *Le Plaisir du texte*, Paris, Seuil, 1975. Rappelons que *La révolution du langage poétique* de Kristeva, qui prend ses distances avec le terme de l'intertextualité, paraît un an plus tôt.

Tout ce qui est à peine toléré ou carrément refusé par la linguistique (comme science canonique, positive), la signifiante, la jouissance, c'est précisément là ce qui retire le texte des imaginaires du langage.

Sur le plaisir du texte, nulle "thèse" n'est possible ; à peine une inspection (une introspection), qui tourne court.²⁵

Désormais, la jouissance devient le mot-clé de la théorie intertextuelle de Barthes. Pour lui, il existe deux formes de répétition. La première ne produit pas de jouissance, elle est honteuse, elle provoque des stéréotypes et elle s'engage idéologiquement. L'autre engendre la jouissance de par son excessivité, elle est formelle et tellement littérale qu'elle en devienne excentrique. Comparable, par exemple, à la danse tantrique, elle est marginalisée dans la société contemporaine. Chez Barthes, l'intertextualité devient alors entièrement subjective, dépendant des expériences de lecture, non contrainte par la chronologie, impossible à démontrer avec des preuves logiques :

Lisant un texte rapporté par Stendhal (mais qui n'est pas de lui) j'y retrouve Proust par un détail minuscule [...]. Je comprends que l'œuvre de Proust est, du moins pour moi, l'œuvre de référence, la *mathésis* générale, le *mandala* de toute la cosmogonie littéraire [...]. Et c'est bien cela l'inter-texte : l'impossibilité de vivre hors du texte infini – que ce soit Proust, ou le journal quotidien, ou l'écran télévisuel : le livre fait le sens, le sens fait la vie.²⁶

Ce sentiment d'un déjà-vu / déjà-lu (car Barthes mentionne aussi des médias audiovisuels), assez vague et très personnel, contribue à la désacralisation de la figure d'auteur, visée par la Nouvelle Critique. Mais, simultanément, le concept de l'intertextualité, et notamment son rôle, semble devenir plus discret : il n'est plus un aspect obligatoire de chaque écriture, qui prime sur le texte en soi. Loin de mettre l'intertextualité au centre de sa théorie du texte, comme il le fait dans son article dans *Encyclopadia Universalis*²⁷, ici Barthes manie le concept de manière plutôt allusive. En outre, selon Limat-Letellier, le chercheur « tient un discours volontiers déceptif et intransitif sur le rôle de l'intertextualité, alors que J. Kristeva insistait en 1974 sur la désacralisation efficace, le rejet subversif des modèles »²⁸.

Une nouvelle fois, la réception de cette théorie intertextuelle n'est pas unanime. Hans-George Ruprecht, en 1983, reproche à Barthes de rendre la notion imprécise et « scientifiquement

²⁵ *Ibid.*, p. 55-56.

²⁶ *Ibid.*, p. 58-59.

²⁷ Rappelons, par exemple, ce qu'il écrit par rapport à l'étymologie du mot texte : « [A]lors que précédemment la critique (seule forme connue en France d'une théorie de la littérature) mettait unanimement l'accent sur le « tissu » fini (le texte étant un « voile » derrière lequel il fallait aller chercher la vérité, le message réel, bref le *sens*), la théorie actuelle du texte se détourne du texte-voile et cherche à percevoir le tissu dans sa texture, dans l'entrelacs des codes, des formules, des signifiants, au sein duquel le sujet se place et se défait, telle une araignée qui se dissoudrait elle-même dans sa toile » (Roland Barthes, « Texte (Théorie du) », *op. cit.*, p. 1017.). En le mettant aux antipodes des théories littéraires traditionnelles (comme celle de Lanson), le chercheur souligne alors la modernité de ce nouveau terme, et de l'outil critique qu'il propose.

²⁸ Nathalie Limat-Letellier, « Historique du concept d'intertextualité », *op. cit.*, paragr. 15.

irrecevable »²⁹. Gignoux considère même que le chercheur a participé au déclin de la popularité du concept :

Finally, il n'est resté de l'intertextualité qu'une définition très large, très théorique, dont on ne voit pas en quoi elle pourrait aider une poétique ou une stylistique des textes. [...] [L]e sentiment de déjà-lu, l'idée de la bibliothèque universelle, s'ils sont incontestables, n'en sont pas moins décevants en stylistique. Le concept permet de mieux cerner la lecture et ses enjeux, mais il entre difficilement en jeu dans une description matérielle. À cause de ce flou, l'intertextualité, après avoir été à la mode pendant les années 70 et 80, a lassé peu à peu les critiques et les universitaires.³⁰

Au vu de la réception mouvementée et critiquée de chaque nouvelle théorie intertextuelle, il n'est pas étonnant de constater que, un an après la publication du *Plaisir du texte*, quand Laurent Jenny propose sa conception de l'intertextualité, il considère avoir hérité d'un terme déjà banalisé³¹.

b) Jenny et Riffaterre : d'une pratique textuelle à un phénomène de réception

En s'éloignant du champ de la philosophie pour se focaliser sur la pratique textuelle, Jenny propose, pour la première fois, un système de gradation de ce phénomène. Le chercheur différencie l'intertextualité proprement dite, définie comme des « rapports de texte à texte en tant qu'ensembles structurés »³², et l'intertextualité faible, à assimiler plutôt aux allusions ou réminiscences ponctuelles, identifiables « chaque fois qu'il y a emprunt d'une unité textuelle abstraite de son contexte et insérée telle quelle dans un nouveau syntagme textuel, à titre d'élément paradigmatique »³³. En 1971, Jean Ricardou³⁴ propose déjà une dichotomie préliminaire entre un régime externe (rapport d'un texte à l'autre) et un régime interne (l'apport d'un texte à lui-même), ce qui provoque l'apparition d'autres distinctions (intertextualité / intratextualité, intratextualité / extratextualité, etc.³⁵). Jenny se différencie par la fluidité de sa typologie, qui met les intertextualités faible et forte aux antipodes, et accepte l'existence des phénomènes plus nuancés entre ces deux notions extrêmes. Le chercheur exclut toutefois les analogies thématiques des phénomènes intertextuels, et considère que tout texte réfère aux autres soit de manière implicite (par exemple avec l'emploi des mêmes codes), soit explicitement (citation, imitation, parodie...). Son approche semble aussi plus

²⁹ Hans-George Ruprecht, « Intertextualité », *Texte*, n° 2, 1983, p. 16.

³⁰ Anne-Claire Gignoux, « De l'intertextualité à la réécriture », *loc. cit.*, paragr. 11-12.

³¹ « Nous héritons donc d'un terme « banalisé » et qu'il nous appartient de rendre aussi plein de sens que possible » Laurent Jenny, « La Stratégie de la forme », *Poétique*, n° 27, 1976, p. 262.

³² *Ibid.*, p. 263.

³³ *Ibid.*, p. 262.

³⁴ Jean Ricardou, *Pour une théorie du nouveau roman*, Paris, Seuil, coll. « Tel quel », 1971.

³⁵ Voir par exemple : Tzvetan Todorov, *Symbolisme et interprétation*, Paris, Seuil, coll. « Poétique », 1978, p. 61-62.; Jean Verrier, « Segalen lecteur de Segalen », *Poétique*, n° 27, 1976, p. 338-339. Brian T. Fitch propose même une notion de « l'intra-intertextualité », cf. Brian Fitch, « L'Intra-intertextualité interlinguistique de Beckett, la problématique de la traduction en soi », *Texte*, n° 2, 1983, p. 85-100. Nous suivons Nathalie Limat-Letellier, « Historique du concept d'intertextualité », *op. cit.*, n. 18.

opérationnelle qu'épistémologique, et l'intertextualité « répond toujours à une vocation critique, ludique et exploratoire »³⁶. En somme, l'intertextualité devient chez Jenny plus une pratique textuelle qu'un concept philosophique.

Cette approche restrictive se situe à l'opposé de la théorie plutôt généralisante de Michael Riffaterre, qui ancre son concept dans les études de la réception. Dans « La Trace de l'intertexte », il déclare que ses objectifs principaux sont « de chercher une définition plus précise de l'intertexte, et de décrire la manière dont le lecteur le perçoit, du moins dans le cas où le mode de perception gouverne la fonction cognitive du texte ou la subordonne à la fonction esthétique »³⁷. Telle qu'elle est conçue par les théories précédentes, l'intertextualité paraît trop variable, excessivement dépendante du bagage culturel et intellectuel du lecteur. Sans pour autant nier l'existence d'un intertexte en tant que « la perception, par le lecteur, des rapports entre une œuvre et d'autres qui l'ont précédée ou suivie », et de l'intertextualité en tant que « le mécanisme propre de la lecture littéraire »³⁸, Riffaterre propose de différencier l'intertextualité *aléatoire*, un arrière-plan culturel enrichissant le sens sans l'affecter, et l'intertextualité *obligatoire*, impliquant que le décodage du message intertextuel est indispensable pour que le texte soit lu correctement.

C'est ce dernier type d'intertextualité qui occupe particulièrement le chercheur :

Il ne s'agit plus d'une complémentarité à éclipses, il ne s'agit plus de répétitions dont la perception enrichit notre sentiment du texte, mais dont l'occultation accidentelle n'affecte pas le sens, ou en tout cas ne suspend pas la compréhension. [...] Ce type d'intertextualité [obligatoire] fonctionne même si le lecteur ne parvient pas à retrouver l'intertexte : dans ce cas, sa lecture cerne un inconnu, elle en subit l'influence sans pouvoir l'éluder, puisqu'il est aussi présent comme question qu'il le serait comme réponse. Vide à combler, attente du sens, l'intertexte n'est alors qu'un postulat, mais le postulat suffit, à partir duquel il faut construire, déduire la signifiante.³⁹

Contrairement à la théorie de Barthes, le savoir littéraire n'est pas la *jouissance* qui permet de savourer le texte, mais une nécessité indispensable pour le déchiffrement des messages textuels. Selon Limat-Letellier, « [i]l lui importe [à Riffaterre] de déchiffrer les énigmes de ce principe générateur de l'écriture tandis que Barthes se plaisait à décrire les effets de brouillage ainsi produits, la confusion de sources »⁴⁰. Francis Goyet⁴¹, quant à lui, souligne que l'aspect obligatoire de l'intertextualité implique qu'aucun texte ne peut lui échapper, ce qui présente le risque de dogmatisme. Umberto Eco,

³⁶ Laurent Jenny, « La Stratégie de la forme », *loc. cit.*, p. 262.

³⁷ Michael Riffaterre, « La trace de l'intertexte », *La Pensée*, n° 215, octobre 1980, p. 4.

³⁸ *Ibid.*

³⁹ *Ibid.*, p. 5-6.

⁴⁰ Nathalie Limat-Letellier, « Historique du concept d'intertextualité », *op. cit.*, paragr. 19.

⁴¹ Francis Goyet, « "Imitatio" ou intertextualité ? », *Poétique*, n° 71, septembre 1987, p. 313-320.

enfin, voit dans cette approche un risque de surinterprétation, pouvant amener le lecteur à considérer qu'il n'y ait qu'une seule manière à lire le texte et que le message transmis par l'auteur soit unique⁴².

La distinction entre l'intertextualité aléatoire et obligatoire, et surtout le fondement de celle-ci sur la compétence culturelle du lecteur, semble problématique. Certains chercheurs paraissent également réticents quant à l'approche plutôt microstructurale de Riffaterre, qui se focalise sur des allusions et citations ponctuelles, d'un vers, voire même un mot dont la non-compatibilité au reste du texte est l'indice d'un recours à l'intertexte. Genette, par exemple, proclame dans l'introduction des *Palimpsestes* que

L'état implicite (et parfois tout hypothétique) de l'intertexte est depuis quelques années le champ d'études privilégié de Michael Riffaterre, qui définit, en principe, l'intertextualité d'une manière beaucoup plus vaste que je ne le fais ici [...], allant jusqu'à identifier dans sa visée l'intertextualité (comme je fais la transtextualité) à la littérarité elle-même [...] Mais cette extension de principe s'accompagne d'une restriction de fait, car les rapports étudiés par Riffaterre sont toujours de l'ordre de microstructures sémantico-stylistiques, à l'échelle de la phrase, du fragment ou du texte bref, généralement poétique. La « trace » intertextuelle selon Riffaterre est donc davantage (comme allusion) de l'ordre de la figure ponctuelle (du détail) que de l'œuvre considérée dans sa structure d'ensemble.⁴³

Certes, Genette s'oppose à l'approche de Riffaterre. De Biaisi, dont la critique est moins modérée, considère qu'« en pratique le concept d'intertextualité est utilisé par Riffaterre dans les limites d'une instrumentalité stylistique et sémiotique qui reconduit les hypothèses formulées par Kristeva en les lestant d'une riche et pleine expérience des textes »⁴⁴. Le foisonnement des critiques s'explique partiellement par la position dominante que la théorie de Riffaterre occupe, depuis la fin des années 80, au sein de la recherche littéraire. Le rapprochement du concept initialement issu du structuralisme à la théorie de réception le modifie fortement : au lieu de servir à la mort de l'auteur, l'intertextualité obligatoire renforce désormais la position de celui-ci et écarte définitivement l'accidentalité de la main qui écrit.

Cependant, comme nous l'avons vu, la conceptualisation initiale d'intertextualité proposée par Kristeva est modifiée (voire, pour certains, lestée ou reconduite) bien avant la théorie de Riffaterre. Éloignée de ses sources linguistiques chez Barthes, dépourvue de son engagement « telquelien »,

⁴² « [L]a seule alternative à une théorie d'interprétation radicale orientée vers le lecteur est celle que prônent ceux pour qui la seule interprétation valide est celle qui vise à saisir l'intention primitive de l'auteur », Umberto Eco, *Interprétation et surinterprétation*, trad. Jean-Pierre Cometti, Paris, Presses universitaires de France, coll. « Formes sémiotiques », 1996, p. 23.

⁴³ Gérard Genette, *Palimpsestes. La littérature au second degré*, op. cit., p. 8.

⁴⁴ Pierre-Marc de Biaisi, « Intertextualité (Théorie de l') », op. cit.

plus pratique que philosophique chez Jenny, ponctuelle et obligatoire chez Riffaterre, l'intertextualité a tant évolué que, pour certains chercheurs comme Antoine Compagnon et Gérard Genette, il paraît nécessaire de s'éloigner du terme afin de reconceptualiser efficacement le concept.

c) Compagnon et Genette : concepts “périphériques”

Dans *La Seconde main ou le Travail de la citation*⁴⁵, Antoine Compagnon renvoie aux principaux penseurs de la modernité, comme Derrida, Foucault, Jakobson, Lacan, ou encore Lévi-Strauss. Par ailleurs, le chercheur a initialement traité l'objet de l'ouvrage dans sa thèse dirigée par Kristeva⁴⁶. Mais le terme *intertextualité* (ou *intertexte*) n'apparaît qu'à 18 reprises et les travaux de Riffaterre ne sont jamais mentionnés. Plus encore, l'objet principal de l'étude, la citation, a été négligé par la théorisation initiale de l'intertextualité, « trop associée aux privilèges institutionnels de l'auteur »⁴⁷, et donc incompatible avec les postulats telquelien. Or, nous voyons dans la définition de la citation proposée par Compagnon le retour de l'influence de théories linguistiques. Considérée comme « la forme simple d'une relation interdiscursive de répétition »⁴⁸, la citation est envisagée sous un angle particulièrement technique : la deuxième « séquence » de l'ouvrage, intitulé « Structures élémentaires », propose une sémiologie de la citation, en tant que fait de langage, et renvoie régulièrement aux travaux de Benveniste et de Peirce. Compagnon propose en effet une typologie formelle qui évoque les théories sémiotiques de Peirce⁴⁹ et qui est présentée sous forme d'une combinatoire de formules algébriques (Tableau 14).

Tableau 14. Typologie de la citation selon Compagnon

Citation symbolique, symbole	T_1 (texte cité) - T_2 (texte citant)	L'auteur n'est pas cité (ou n'existe pas), ex. les proverbes
Citation indicielle, indice	A_1 (auteur cité) - T_2 (texte citant)	On fait référence à l'auteur, mais pas au texte
Citation iconique, icône	$S_1(A_1, T_1)$ - A_2 (auteur citant)	Le système cité est complet (S_1) et l'auteur citant est présent, ex. pastiche, réminiscence
?	S_1 - S_2 (système complet : A_2, T_2)	Relation déconcertante, où les deux composantes sont difficiles à reconnaître

⁴⁵ Antoine Compagnon, *La seconde main*, *op. cit.*

⁴⁶ « Une version précédente de ces pages fut une thèse de doctorat de troisième cycle, réalisée à l'occasion d'un séjour à la fondation Thiers, et soutenue en décembre 1977 à l'université de Paris VII. Julia Kristeva suivit et dirigea [...] ce travail », *Ibid.*, n. 1.

⁴⁷ Nathalie Limat-Letellier, « Historique du concept d'intertextualité », *op. cit.*, paragr. 22.

⁴⁸ Antoine Compagnon, *La seconde main*, *op. cit.*, p. 55.

⁴⁹ Voir par exemple David Savan, « La sémiotique de Charles S. Peirce », *Langages*, vol. 14, n° 58, 1980, p. 9-23, en ligne, <https://www.persee.fr/doc/lgge_0458-726x_1980_num_14_58_1844>.

Une telle approche structurante rappelle les sources linguistiques du concept de l'intertextualité. De Biaisi considère même la théorie de Compagnon comme étude de cas d'une des pratiques intertextuelles :

[L]'ouvrage d'Antoine Compagnon [...] donnait pour la première fois une vaste étude systématique de la pratique intertextuelle de la citation. [...] En systématisant cette description du processus citationnel, Antoine Compagnon propose de penser ce processus comme modèle de l'écriture littéraire qui serait structurellement aux prises avec la même exigence transformationnelle et combinatoire [...]. Saisie dans sa nature hybride (à la fois lecture et écriture), la citation est ainsi posée comme ce cas de figure intertextuel par lequel se révélerait un processus beaucoup plus profond dont il ne serait lui-même qu'un effet remarquable : le travail de l'écriture, l'énergie qui circule dans cette structure mobile.⁵⁰

Nous retrouvons d'ailleurs chez Compagnon une dimension affective déjà observée chez Barthes (mais placée, cette fois-ci, du côté de l'auteur et pas du lecteur) : « il y a toujours un livre avec lequel j'ai l'envie que mon écriture entretienne une relation privilégiée, "relation" valant ici pour son double sens, celui du récit (de la récitation), et celui de la liaison (de l'affinité élective) »⁵¹. Plutôt que de se focaliser sur la définition de la citation en tant qu'objet, un produit fini, le chercheur s'intéresse à l'acte de citer et, par l'extension, à l'écriture et à la réécriture : « Récrire, réaliser un texte à partir de ses amorces, c'est les arranger ou les associer, faire les raccords ou les transitions qui s'imposent entre les éléments mis en présence : toute l'écriture est collage et glose, citation et commentaire »⁵².

Limat-Letellier formule plusieurs critiques quant à la manière dont Compagnon conceptualise la citation. Tout d'abord, l'effacement de la frontière entre la citation et le plagiat, bien délimitée dans l'usage courant, inquiète la chercheuse. Ensuite, le discours rapporté, y compris intradiégétique, et la citation représentée seule (gravure, graffiti, publicité...), sont difficiles à associer à l'intertextualité :

[U]ne définition aussi large de la citation finit par déborder tout le champ notionnel de l'intertextualité [...]. C'est le dialogisme bakhtinien lui-même qui expliquerait cette extrapolation : si toute activité verbale évoque la trace omniprésente des discours antérieurs, le "déjà dit" semble orienter tout acte d'énonciation vers la réitération. Au contraire, si l'intertextualité est définie comme une activité d'écriture et de la lecture qui implique la transformation d'allusions et d'emprunts à des textes antérieurs, elle apparaît comme plus restrictive.⁵³

Le manque de délimitation claire et univoque du "champ notionnel" de l'intertextualité paraît être un reproche récurrent de la plupart des critiques que nous avons évoquées jusqu'alors. Simultanément, la conceptualisation proposée par Gérard Genette est systématiquement vue comme une

⁵⁰ Pierre-Marc de Biaisi, « Intertextualité (Théorie de l') », *op. cit.*

⁵¹ Antoine Compagnon, *La seconde main*, *op. cit.*, p. 34.

⁵² Antoine Compagnon, *La seconde main*, *op. cit.*, p. 32.

⁵³ Nathalie Limat-Letellier, « Historique du concept d'intertextualité », *op. cit.*, paragr. 30.

tentative d'éclaircissement tant espérée et qui « lève beaucoup des obscurités dans lesquelles le métadiscours critique se débattait jusque-là »⁵⁴.

Rappelons que dans *Palimpsestes*⁵⁵, Genette propose de délimiter précisément le domaine de l'intertextualité et de l'inscrire dans un concept théorique plus large, la *transtextualité*, qu'il définit en tant que « tout ce qui met [un texte] en relation, manifeste ou secrète, avec d'autres textes »⁵⁶. Cette définition dépasse la conception initiale de l'intertextualité de Kristeva pour se rapprocher de la définition de la *littérarité*⁵⁷ ou de la *poéticité*⁵⁸. L'intertextualité genettienne est présentée comme une des cinq catégories de la transtextualité et est définie comme « une relation de coprésence entre deux ou plusieurs textes, c'est-à-dire, eidétiquement et le plus souvent, par la présence effective d'un texte dans un autre »⁵⁹. Trois modalités seulement sont incluses : la citation (avec ou sans guillemets), le plagiat et l'allusion, ce qui rend ce concept particulièrement restreint.

La conceptualisation de l'intertextualité a débuté par une définition relativement large proposée par Kristeva. Depuis, certains théoriciens ont proposé des outils plus pratiques pour étudier la présence de l'intertextualité dans les œuvres, alors que d'autres ont insisté sur l'aspect subjectif et abstrait du phénomène. Des difficultés particulières ont été formulées par les critiques, liées par exemple à la compétence du récepteur, les domaines et relations à inclure ou exclure, ou encore le rôle de l'auteur. Comme le souligne Gignoux, « les divergences d'opinion [...] et un certain flou conceptuel que l'on remarque dans les études intertextuelles n'ont pas aidé le concept à devenir un réel outil pratique et incontesté »⁶⁰. Selon la chercheuse, les problèmes liés à la définition de l'intertextualité sont multiples.

Tout d'abord, de nombreux critiques soulignent le danger de limiter l'analyse intertextuelle à la critique des sources traditionnelle, autrement dit de « faire du neuf avec du vieux, et par exemple à appeler *analyse intertextuelle*, une bien traditionnelle critique philologique des sources et des influences littéraires »⁶¹. Étant donnée la fréquence du signalement de cette difficulté, ces deux modes d'analyse sont régulièrement confondus, ce qui prouve l'aspect problématique de leur distinction. Comme l'explique Limat-Letellier,

⁵⁴ Pierre-Marc de Biiasi, « Intertextualité (Théorie de l') », *op. cit.*

⁵⁵ Gérard Genette, *Palimpsestes. La littérature au second degré*, *op. cit.* Nous avons déjà eu recours aux théories genettiennes au I.2a) « Heidmann, Brunel et Genette : définition de la notion de réécriture ».

⁵⁶ *Ibid.*, p. 7.

⁵⁷ Voir par exemple Mircea Marghescu, *Le concept de littérarité : critique de la métalittérature*, Paris, Kimé, coll. « Détours littéraires », 2009 [1974].

⁵⁸ Voir Roman Jakobson, *Questions de poétique*, Paris : Seuil, 1973.

⁵⁹ Gérard Genette, *Palimpsestes. La littérature au second degré*, *op. cit.*, p. 8.

⁶⁰ Anne-Claire Gignoux, « De l'intertextualité à la réécriture », *loc. cit.*, paragr. 10.

⁶¹ Marc Angenot, « L'intertextualité : enquête sur l'émergence et la diffusion d'un champ notionnel », *loc. cit.*

[L]es recherches sur l'intertextualité ne visent pas des influences mais à construire une analyse des modes d'insertion et surtout de transformation, d'altération des emprunts. Tandis que l'idée de la filiation "naturelle" privilégiait le modèle au détriment des successeurs, l'intertextualité n'est plus entravée par cet héritage de la philologie : elle s'intéresse à des pratiques d'écriture à la fois prospectives et rétrospectives, à la plurivocité et à la réversibilité des effets de sens.⁶²

Toutefois, cette différenciation épistémologique reste relativement vague quant à la méthodologie pratique de l'analyse intertextuelle qui permettrait d'échapper à la pratique de la "vieille" critique des sources. La définition de l'intertextualité en tant que « tissu » ou « croisement » de textes laisse de la place à des interprétations très diversifiées. Cette subjectivité du terme, et de l'interprétation des phénomènes qu'il désigne, semble nécessiter un cadre conceptuel très large, trop large peut-être pour pouvoir l'outiller. Selon Gignoux,

Il est, en fait, impossible de décider du moment où l'intertexte est délibérément convoqué par l'auteur, inconsciemment provoqué par lui, ou redevable au lecteur seulement. Le passage de l'une à l'autre de ces catégories est évidemment graduel. On ne peut nier que l'intertextualité soit un phénomène de réception. C'est pourquoi il faut accepter ce sens très large de l'intertextualité, bien que le concept, probablement, risque alors de se désintégrer totalement.⁶³

En effet, le reproche de l'acception trop large du terme de l'intertextualité revient fréquemment dans les travaux critiques. D'autres chercheurs, comme les traductologues, déplorent l'exclusion de certaines formes de la réflexion autour de l'intertextualité. Finalement, nous observons que l'histoire de ce terme est parsemée de disputes successives entre théoriciens, de voix critiques et de conceptualisations par opposition. L'agglomération des néologismes telles que l'*autotextualité*, la *mimotextualité* ou encore l'*interintermimotextualité* rend le concept plus difficile à s'approprier et à exploiter dans une analyse d'un ensemble concret de textes. Enfin, comme le remarque Limat-Letellier,

L'intertextualité a été partagée et reconstruite selon diverses orientations. Nous en voulons pour preuve les dictionnaires et manuels méthodologiques destinés [...] aux étudiants de lettres : leurs notices sur l'intertextualité résument les définitions successives (ou certaines d'entre elles) ; elles intègrent parfois en dernier lieu les propositions de *Palimpsestes*, mais une synthèse fait souvent défaut. En outre, il importe aussi de tenir compte d'un regard critique, de l'usure des théories dans la réflexion contemporaine.⁶⁴

Si l'existence d'un réseau des relations qui relie différents textes littéraires est indubitable, un certain flou conceptuel rend difficile l'application d'analyses intertextuelles à une recherche littéraire, notamment quand il s'agit du domaine de la littérature comparée⁶⁵. Nous observons d'ailleurs

⁶² Nathalie Limat-Letellier, « Historique du concept d'intertextualité », *op. cit.*, paragr. 70.

⁶³ Anne-Claire Gignoux, « De l'intertextualité à la réécriture », *loc. cit.*, paragr. 4.

⁶⁴ Nathalie Limat-Letellier, « Historique du concept d'intertextualité », *op. cit.*, p. 64.

⁶⁵ Certaines réflexions consacrées à l'activité comparante tâchent de différencier l'*intertextualité* de la *comparaison*, de la *réception* ou encore de la *parallèle*. Par exemple, Jürgen Siess remarque que « les concepts opératoires retenus par Didier Soullier et Wladimir Troubetzkoy [...] prennent bien en considération "réception" et "intertextualité", mais ne relèvent ni "comparaison" ni "parallèle" » (« "Parallèle", un concept opératoire en Littérature comparée ? », *Revue de Littérature comparée*, no 298, LXXV, avril-juin 2001, p. 226-227.) ; selon Bernard Franco, « [l]a parallèle se situerait donc entre

la présence décroissante de ce terme dans les travaux universitaires publiés après 1985 et son remplacement progressif par d'autres termes : *hypertexte*, *palimpseste*, *reprise*, *transformation textuelle*, *réécriture*... Comme le souligne Lombez, « ces termes, parfois en usage dans la réflexion théorique contemporaine menée autour des phénomènes de textualité semblent avoir au moins un élément en commun. Tous renvoient en effet à une intervention [...] qui, à partir d'un texte original, amène la constitution d'un nouveau produit textuel, un « dérivé » du premier. »⁶⁶. Pour autant, même si les publications sur le sujet sont moins nombreuses et le débat moins mouvementé que dans la seconde moitié du XX^e siècle, l'intertextualité n'est pas oubliée⁶⁷. Au contraire, elle est régulièrement adaptée à des créations artistiques et culturelles très variées, comme les web-séries⁶⁸ ou la chanson publicitaire⁶⁹. Le numéro 51 de la revue *Genesis* (2020), intitulé « Intertextualité – Exogenèse », est une bonne démonstration de l'évolution du concept, initialement spécifiquement littéraire, vers des phénomènes autres que textuels (architecture, sciences, histoire de l'art).

d) Théories intertextuelles et observation du corpus

L'hétérogénéité des approches critiques coïncide dans notre corpus avec l'impossibilité fréquente d'assigner sans aucune ambiguïté une relation observée à un concept spécifique. Prenons comme exemple la nouvelle parodique de D.-L. Eimann, « Une histoire renouvelée des Grecs », parue en 1854 dans *Histoires cavalières*. Ce recueil de dix nouvelles préfacé par Alexandre Dumas père est l'unique publication de l'auteur. Par ailleurs, l'identité de ce dernier, dont le nom est très proche de l'allemand *ein Mann* (un homme), reste mystérieuse. Plusieurs nouvelles d'Eimann ont paru dans la revue *Mousquetaire*, le quotidien dirigé par Dumas de novembre 1853 à février 1857. Avant d'y publier, Eimann s'illustre déjà en tant que lecteur fidèle et critique de la revue. Plusieurs de ses lettres y ont été publiées (ou au moins citées) et commentées. Initialement désigné comme « le premier venu », Eimann devient le moteur de plusieurs « causeries » permettant à Dumas d'asseoir son projet éditorial. Mombert et Durand⁷⁰ sont catégoriques quant à l'identité d'Eimann :

l'identité, qui ôte toute possibilité de comparaison, et l'incomparable, qui rendrait la comparaison arbitraire, réduisant même les singularités de chaque auteur et oubliant ce qui se joue dans la création, c'est-à-dire l'individualité, voire l'unicité » (*La Littérature comparée. Histoire, domaines, méthodes*, Paris, Armand Colin, 2016, p. 143-144.).

⁶⁶ Christine Lombez, « Réécriture et traduction », dans Jean-Paul Engélibert et Yen-Mai Tran-Gervat (dir.), *La littérature dépliée : Reprise, répétition, réécriture*, Rennes, Presses universitaires de Rennes, coll. « Interférences », 2008, p. 71-72.

⁶⁷ La recherche du terme « intertextualité » dans les titres des ouvrages et articles publiés sur le site OpenEdition entre 2015 et 2022 trouve 30 articles des revues, 17 chapitres et 2 livres.

⁶⁸ Voir par exemple l'article « Réappropriation et intertextualité dans la web-série *Girltrash* » d'Émilie Marolleau (*TV/Series*, 2014).

⁶⁹ Voir par exemple le chapitre « Intertextualités à l'œuvre dans la chanson publicitaire : le cas de la parodie » de Christophe Magis (*Chanson et intertextualité*, 2012).

⁷⁰ Sarah Mombert et Pascal Durand (dir.), « Ce que causer veut dire », *Entre presse et littérature : Le Mousquetaire, journal de M. Alexandre Dumas (1853-1857)*, Liège, Presses universitaires de Liège, 2019, p. 45-63, en ligne, <<http://books.openedition.org/pulg/5385>>.

À l'image et à l'intention de ce public protéiforme, Dumas crée un alter ego auquel il donne d'entrée de jeu la parole, endossable par quiconque, sorte de case vide : un « premier venu » dont l'entrée en scène se confond avec le numéro spécimen du journal dans le « programme artistique » duquel il apporte la contradiction au romancier saisi par le démon du journalisme.

[...] Et c'est en suscitant encore l'intervention intempestive de ce « premier venu » que Dumas, régisseur de tout ce petit théâtre, va s'employer à définir ce qu'il entend par causerie. Le numéro 10, daté du 29 novembre 1853, publie ainsi [une lettre] que ce providentiel quidam adresse à « M. Alexandre Dumas » tout en réajustant son propre masque (« D.-L. Eimann », c'est-à-dire n'importe qui).⁷¹

Si, dans la préface des *Histoires cavalières*, Dumas maintient le mythe de l'existence de D.-L. Eimann⁷², une série d'indices, plus ou moins directs, permet de lever le masque. En effet, alors que, pour Kristeva et pour Barthes, la mort de l'auteur est un des enjeux majeurs de l'intertextualité, dans le cas des *Histoires cavalières*, c'est elle, entre autres, qui permet de l'identifier. Mise à part l'étude du style d'écriture d'Eimann et sa comparaison avec celui de Dumas, une série de références (littéraires, théâtrales ou sociétales) est très proche du cercle des relations de l'auteur des *Trois Mousquetaires*. Certaines de ces références sont facilement identifiables et apparaissent dès le paratexte. C'est notamment le cas de la postface de l'ouvrage qui est une lettre adressée à Roger de Beauvoir, l'ami et le témoin du mariage de Dumas. Cependant, d'autres références sont plus subtiles, comme nous le verrons avec l'exemple de la nouvelle « Une histoire renouvelée des Grecs ».

Ce récit reprend sous des traits humoristiques l'épisode de l'amour d'Orphée et Eurydice. Alfred A., un dandy désabusé et volage, évoque le mythe pour accuser les femmes de jouer avec le cœur des hommes. Le récit mythologique est tantôt raconté par le narrateur, tantôt évoqué par le biais de trois citations, une de six vers et deux de cinq vers. La dernière est accompagnée d'une note : « Pardon pour ces citations. Nous aurions préféré vous donner du Virgile pur ; mais nous ne le comprenons pas trop bien, et puis nous écrivons pour tout le monde »⁷³. Les conventions de la mise en page indiquent clairement qu'il s'agit de citations, et la note donne l'illusion que l'auteur est identifié. Or, même si le lecteur est orienté vers les *Géorgiques* de Virgile, l'identité du traducteur n'est pas révélée⁷⁴. Ce cas de figure s'inscrit tout à fait dans la conception très vaste de Genette (« sous sa forme la plus explicite, c'est la pratique traditionnelle de la *citation* (avec guillemets, avec ou sans référence précise) »⁷⁵), mais se retrouve moins bien dans le classement assez binaire de

⁷¹ *Ibid.*, paragr. 7-8.

⁷² « Vous avez lu ses articles, et vous restez convaincu que, pour être un des derniers venus dans l'arène littéraire, il n'en a pas moins le droit d'y revendiquer une large place. De plus, il est devenu un de mes amis ; et, pour ne pas dater de vingt ans, son amitié ne m'en est pas moins chère. Permettez-moi donc de faire quelque chose pour lui, quoique mon gaillard soit de force à marcher tout seul ; permettez-moi de vous le présenter et de vous dire : « Mon premier-venu, – D.-L. Eimann, – a de l'esprit comme un démon. », voir Alexandre Dumas, « Introduction », *Histoires cavalières*, Paris, Charpentier, 1854, p. 6-8, en ligne, <<https://gallica.bnf.fr/ark:/12148/bpt6k5699364p>>.

⁷³ D.-L. Eimann, « Une Histoire renouvelée des Grecs », dans Alexandre Dumas (dir.), *loc. cit.*, p. 34, en ligne, <<https://gallica.bnf.fr/ark:/12148/bpt6k5699364p>>.

⁷⁴ Il s'agit de la traduction en vers de Jacques Delille (1770).

⁷⁵ Gérard Genette, *Palimpsestes. La littérature au second degré*, *op. cit.*, p. 8.

Compagnon, appuyé sur la présence et l'absence tant de l'auteur que du texte cité. Nous pouvons également considérer qu'il s'agit ici de l'intertextualité forte de Jenny, voire de l'intertextualité obligatoire de Riffaterre. En effet, pour que la parallèle entre la situation du personnage principal de la nouvelle et l'histoire mythologique soit comprise par le lecteur, il est indispensable que ce dernier arrive à bien identifier les sources auxquelles le narrateur se réfère.

Outre le mythe, le texte est truffé de références, plus ou moins explicites, à d'autres œuvres. À la page 29, par exemple, deux vers sont sortis du corps du texte : « Une lune de miel n'a pas trente quartiers, / Comme un baron saxon... »⁷⁶. L'auteur n'est pas cité, mais les conventions de la mise en page constituent l'indice de la provenance extérieure de l'extrait. Dans le classement de Compagnon, le manque du référencement de l'auteur est caractéristique des citations symboliques. Toutefois, pour le chercheur, dans ce cas de figure, « la citation est un opérateur logique, exempt de toute connotation »⁷⁷, comme un proverbe ou un théorème. Or, ce n'est pas le cas ici. L'auteur n'est certes pas évoqué clairement, mais il semble être sous-entendu, ou du moins supposément connu par les lecteurs de l'époque, puisqu'il s'agit d'un des poèmes d'Alfred de Musset⁷⁸. De ce point de vue, notre exemple se rapproche plutôt de la citation iconique :

L'icône est une citation qui qualifie le citateur lui-même, quand il assume une énonciation propre malgré la reprise ; le pastiche, la citation implicite, la réminiscence sont d'ordre iconique : des paroles secondes qui revendiquent leur identité, leur spécialité dans la similarité, par l'appropriation, la mainmise, la violence.⁷⁹

Le chercheur souligne, pour ce type de citation, que le système cité est complet, c'est-à-dire que tant le texte que l'auteur sont présents. Cette présence doit-elle obligatoirement être explicite ? Compagnon ne semble pas le préciser, et la dimension algébrique de son classement, dépourvu d'exemples, rend difficile son application à des cas de figure concrets.

Une autre référence présente dans la nouvelle est encore plus implicite. Eimann détourne l'épisode d'Aristée en accusant Eurydice de séduire volontairement le berger pour échapper à l'ennui de sa vie conjugale. Ainsi, en croisant Aristée dans la rue, elle pose sur lui un regard entendu :

Pardonnez-lui ce regard. Elle était femme, elle s'ennuyait, et, comme toutes les femmes qui s'ennuient, elle voulait un combat – pour avoir un vainqueur.

Cette pensée est très-jolie ; mais elle n'est pas de nous.⁸⁰

Les conventions typographiques ne sont pas respectées : le texte cité n'est pas mis entre guillemets ni sorti du corps du texte. L'auteur n'est pas référencé. En revanche, le lecteur dispose d'un indice lui permettant de réaliser qu'au moins une partie de la phrase précédente ne vient pas de l'auteur.

⁷⁶ D.-L. Eimann, *op. cit.*, p. 29.

⁷⁷ Antoine Compagnon, *La seconde main, op. cit.*, p. 78.

⁷⁸ Alfred de Musset, « Mardoche. XVI », *Premières poésies 1829-1835*, Larousse, 1907, p. 83.

⁷⁹ Antoine Compagnon, *La seconde main, op. cit.*, p. 79.

⁸⁰ D.-L. Eimann, *op. cit.*, p. 30.

Il s'agit de nouveau d'un extrait de Musset, cette fois-ci tiré de la comédie *À quoi rêvent les jeunes filles* ?⁸¹ En considérant la typologie de Genette, il est difficile de déterminer sans aucune équivoque s'il s'agit d'un *plagiat* (« une forme moins explicite et moins canonique ») ou d'une *allusion* (« un énoncé dont la pleine intelligence suppose la perception d'un rapport entre lui et un autre, auquel renvoie nécessairement telle ou telle de ses inflexions, autrement non recevable »⁸²). La reformulation est très légère et peut être schématisée de manière suivante :

(il leur faut)[Musset] | [elle voulait]Eimann) un combat [–]Eimann pour avoir un vainqueur.

Les modifications d'Eimann sont-elles suffisantes pour pouvoir parler d'une allusion ? Dans le cas contraire, pouvons-nous réellement considérer que qu'il s'agit d'un plagiat, alors que l'indice de citation est bien présent (même si l'auteur n'est pas évoqué clairement) ?

Par ailleurs, du point de vue de la théorie de Riffaterre, comment déterminer s'il s'agit d'une intertextualité obligatoire ou aléatoire ? Les deux citations de Musset paraissent autosuffisantes, leur sens est déjà compréhensible sans accéder au contexte original. Le renvoi complice au répertoire littéraire contemporain du lecteur peut être interprété comme un enrichissement du récit par l'introduction d'un arrière-plan culturel. Nous le voyons à plusieurs reprises dans le texte par la mention de noms propres comme Coquardeau ou Joseph Prudhomme⁸³. Cependant, la répétition des renvois à Musset à si peu de distance et la présence d'indices de citation interpellent. *A minima*, elles permettent au lecteur de comprendre la proximité entre le protagoniste, Alfred A., et le poète homonyme (éponyme ?), connu pour son dandysme. La nouvelle pourrait-elle être une boutade que Dumas lance à Musset, son proche ami ?

Il ne s'agit pas tant de trouver la réponse à cette question que d'observer la fluidité de la dichotomie de Riffaterre. Du moins, l'interprétation de l'intertexte présent dans la nouvelle est affectée par cette thèse de l'écriture adressée avant tout à un cercle amical. Par exemple, le nom de Félicie, la femme chez laquelle Alfred A. décide finalement de se rendre, ne renvoie-t-il pas à Félicie de Fauveau, la sculptrice féministe admirée tant par Dumas et Musset⁸⁴ que par l'ensemble des romantiques de l'époque ? Par ailleurs, pour comprendre la fonction du nom propre Arthur dans

⁸¹ « Les femmes cependant demandent autre chose. / Bien plus, sans les aimer, du moment que l'on ose, / On leur plaît. La faiblesse est si chère à leur cœur, / Qu'il leur faut un combat pour avoir un vainqueur. », Alfred de Musset, « À quoi rêvent les jeunes filles ? », *Premières poésies 1829-1835*, Larousse, 1907, p. 197.

⁸² Gérard Genette, *Palimpsestes. La littérature au second degré*, op. cit., p. 8.

⁸³ « Hélas ! les Aristées sont de tous les temps ; On rencontre encore *quelques* Eurydices ; Mais il n'est plus d'Orphées, – ou, s'il en existe encore, ils répondent au nom de Coquardeau ou à celui non moins harmonieux de Joseph Phrudhomme ! », D.-L. Eimann, op. cit., p. 35.

⁸⁴ Voir cette notice de la sculpture « L'âme se détachant des sentiments terrestres » (§ Précision auteur) : « Son œuvre a eu une grande influence sur des auteurs romantiques comme Alfred de Musset et Alexandre Dumas [...]. Alexandre Dumas visite le Salon de 1827 et s'émeut devant un de ses bas-reliefs, *Le Meurtre de Monaldeschi*. La pièce du dramaturge de vingt-cinq ans, inspirée de cette tragique histoire, est définitivement reçue au Théâtre-Français à peine six mois plus tard. » (<https://www.pop.culture.gouv.fr/notice/joconde/000SC019415>).

la phrase « Arthur est bien joli, mais Orphée vaut mieux, n'est-ce pas ? »⁸⁵, ne faut-il pas puiser à nouveau dans le réseau des connaissances communes de Dumas et de Musset ? Le cas échéant, le roman *Arthur* d'Ulric Guttinger semble être un bon candidat. En 1829, Musset écrit un poème intitulé « À Ulric G. », qui est l'expression de leur amitié de longue date. Le roman *Arthur* paraît en 1832 et rencontre un franc succès. C'est un récit inspiré de la peine amoureuse de l'auteur, peine qui l'a poussé à se retirer de la vie sociale dans un chalet en forêt normande. En somme, si la nouvelle est à lire comme un clin d'œil aux amis de l'auteur, l'intertexte obligatoire n'est plus lié au mythe d'Orphée et Eurydice, mais aux références partagées du cercle intellectuel dumasien.

Enfin, nous observons également, dans cette brève nouvelle, des relations davantage allusives, plus proches de l'intertextualité telle que la définit Barthes et comparables à un sentiment subjectif d'un déjà-lu. Voyons, par exemple, cette ridiculisation du deuil d'Orphée chez Eimann :

Pendant sept mois, dit Virgile, – et nous ne le lui faisons pas dire, soyez-en convaincu,
– pendant sept mois, il vécut dans les bois, dans les antres, jetant aux échos ses gémissements, ses grincements de dents, oubliant probablement de peigner sa barbe et ses cheveux, attendant que la mort mît fin à sa douleur.⁸⁶

Dans cet extrait, l'épisode virgilien est dévalorisé, alors que le narrateur assure le lecteur de sa fidélité à la source antique. Plusieurs attributs d'Orphée sont mis à mal. Ses qualités de poète, chanteur prodige, capable d'enchanter les arbres et les animaux, sont ridiculisées. Sa voix divine devient des « gémissements » et des « grincements de dents ». Le long deuil après la mort d'Eurydice perd de sa noblesse : Orphée est dépeint comme un sauvage habitant des bois, négligé et décoiffé.

Curieusement, cet épisode est détourné de manière similaire dans une mazarinade anonyme écrite presque 200 ans plus tôt, *L'Orphée grotesque, avec le bal rustique. En vers burlesques* :

L'escho se plaint d'estre estourdie
De sa criarde melodie [...] ;
Sa plainte joüant de son reste,
Il maugrée, il fulmine, il peste [...].
Mais il ne s'en prend desormais
Qu'à sa barbe qui n'en peut [rien],
Et s'arrachant sa heure fauve,
De male rage devient chauve.⁸⁷

Le détournement porte sur les mêmes attributs : la voix d'Orphée devient « une criarde mélodie », le poète « maugrée », « fulmine » et « peste ». Comme chez Eimann, l'écho est l'unique témoin des plaintes du poète. Enfin, la ridiculisation passe une nouvelle fois par les aspects capillaires : décoiffé chez Eimann, Orphée tord sa barbe et s'arrache tous ses cheveux dans la mazarinade.

⁸⁵ D.-L. Eimann, *op. cit.*, p. 28.

⁸⁶ D.-L. Eimann, *op. cit.*, p. 33.

⁸⁷ Anonyme, *L'Orphée grotesque, avec le bal rustique. En vers burlesques. Suivi de Suite de l'Orphée avec les Bacchantes ou Les Rudes Joyeuses*, Paris, Sebastien Martin, 1649, p. 6-7.

L'éloignement, tant historique que linguistique, des deux extraits rend incertaine la thèse de l'influence directe de l'auteur de la nouvelle par ce poème burlesque. Les autres détournements humoristiques présents dans le corpus ne recourent pas aux mêmes méthodes pour ridiculiser le deuil du poète. Certes, la dépoétisation du chant d'Orphée est présente, mais s'articule généralement au remplacement de son pouvoir enchanteur par celui d'endormir le public. Ainsi, par exemple, dans *Roger-Bontems et Javotte*, le poète s'exclame : « [S]i je ne puis les enchanter avec ma vielle, fais du moins que je les endorme », alors que dans l'opéra *Le Petit Orphée*, le chœur des Démons chante « Il nous endormira / Tout comme à l'Opéra ». C'est également le reproche que l'Eurydice d'Offenbach formule dans le premier acte de *l'Orphée aux Enfers* : « [J]'ai cru épouser un artiste et [...] je me suis unie à l'homme le plus ennuyeux de la création », s'écrit-elle, avant d'ajouter : « Il m'ennuie, comme vos vers, votre violon ! ». Dans l'opéra-bouffe, la musique d'Orphée devient un véritable moyen de punition sadique que le poète exerce sur sa femme avec un malin plaisir.

En revanche, alors que la ridiculisation de la pilosité du chantré apparaît exclusivement dans « Une nouvelle renouvelée des Grecs » et dans *Orphée grotesque*, la cantate de Grandval, une parodie minimale de l'œuvre de Clérembault, contient la ridiculisation du deuil d'Orphée semblable à celle présente chez Eimann. La cantate commence par ces mots : « Le fameux chantré de la Trace par les regrets les plus touchans déplorait ainsi sa disgrâce, courant comme un fou par les champs », puis, la plainte d'Orphée est interrompue par l'interrogation suivante :

Mais à qui tien-je un tel langage ?
Il n'est icy ny chien ny chat.
[...] J'avois des choses magnifiques, plaignons-nous du moins aux Écos.
Imitons les tendres heros.

Le poète apparaît à nouveau comme un sauvage affolé, errant dans les bois, avec l'écho pour unique compagnon. Toutefois, l'épisode traité par Eimann est celui du deuil d'Orphée après la seconde mort d'Eurydice. Or, Grandval renvoie au désespoir qui résulte de la première mort de la femme, poussant le poète à descendre aux Enfers. Un autre détournement humoristique exploite, à son tour, l'absurdité de la plainte d'Orphée qui l'empêche de passer à l'action. Chez Moline et d'Orvigny, Orphée dit : « Pour gagner du temps, mettons-nous en règle, et commençons par apostropher l'aurore, le jour, la nuit, les échos, les parques, les démons et la nature entière. Allons. », avant de constater : « Mais, je m'amuse à chanter, lorsque je dois agir... ».

Il est possible de modéliser ces proximités de traitement au sein du sous-corpus des parodies avec un graphe comme celui que nous présentons en Figure 25.

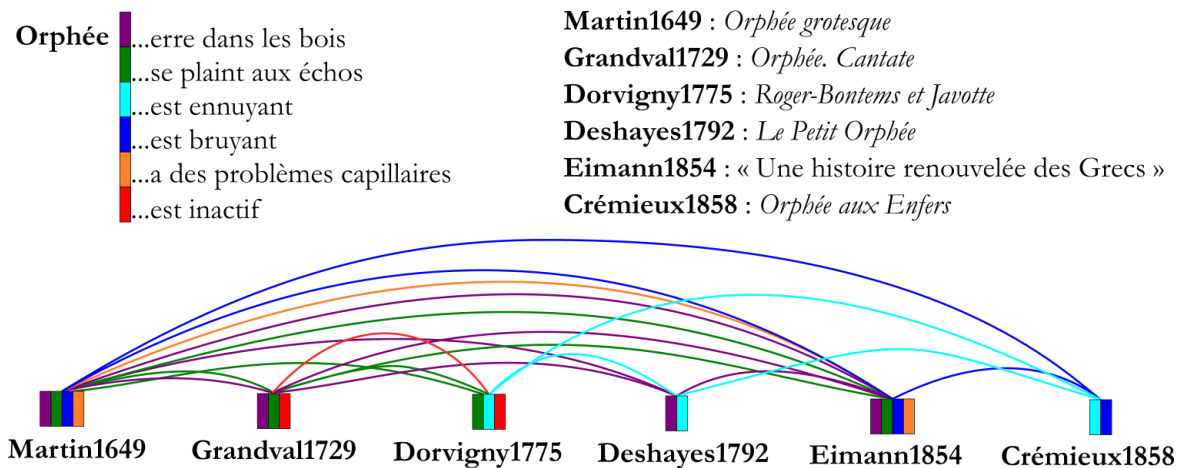


Figure 25. Modélisation de la ridiculisation du deuil d'Orphée dans le sous-corpus des détournements humoristiques

Dans ce schéma, nous différencions six traits humoristiques liés à la ridiculisation du deuil d'Orphée. Chacun de ces traits est représenté à l'aide d'un code couleur. L'inactivité du poète, par exemple, est marqué en rouge. Nous renvoyons à chaque texte, représenté par un rectangle multicolore, à l'aide d'un identifiant composé du nom du premier auteur et de la date de publication (la nouvelle « Une histoire renouvelée des Grecs », par exemple, est identifiée par Eimann1854). Les couleurs composant chaque rectangle indiquent les traits de ridiculisation présents dans chaque texte. Enfin, les arcs qui relient les rectangles dévoilent la complexité des relations.

Cette représentation visuelle montre la grande proximité entre Martin1649 et Eimann1854. Les opéras de Deshayes1792 et de Crémieux1858 sont les plus spécifiques, car ils ne possèdent qu'un trait commun chacun avec une autre réécriture. En revanche, ce sont eux qui peuvent être reliés au plus grand nombre de textes pris en compte (quatre œuvres chacun alors que tous les autres textes ne peuvent être liés qu'avec trois autres œuvres). Dans le cas de Martin1649 et d'Eimann1854, la thèse qu'il s'agit de relations intertextuelles paraît plausible, notamment en envisageant le concept dans le sens barthesien. Mais en est-il de même pour les œuvres reliées par un seul trait ? Où faut-il tracer la frontière entre une influence directe, un sentiment de déjà-lu, et une proximité topique, liée au traitement du même mythe ?

Il est important d'apprécier l'hétérogénéité des phénomènes observés sans s'efforcer de la cloisonner dans une définition particulière. En effet, les approches dualistes de l'intertextualité, qui conçoivent le concept en tant que lien d'interdépendance entre deux textes, s'appliquent mal à

notre exemple de comparaison du sous-ensemble des détournements humoristiques. Nous proposons de conceptualiser les relations intertextuelles sous forme d'un rhizome⁸⁸, où tous les composants peuvent être liés selon des traits multiples, tantôt directement, tantôt indirectement, indépendamment de la chronologie. Cette conceptualisation rhizomatique est davantage saillante quand le schéma présenté en Figure 25 est remodelé au sein d'une spatialisation moins linéaire, comme celle en Figure 26.

Alors que les informations présentées sont identiques, leur appréciation est légèrement modifiée. Par exemple, il est plus aisé de constater que la concentration des liens, très forte à gauche du schéma, entre les œuvres étroitement liées de Martin1649 et d'Eimann1854, est beaucoup moins importante à droite du graphe où sont représentées les relations plus allusives (Dorvigny1775, Deshayes1792 et Crémieux1858). Par ailleurs, cette spatialisation montre mieux le manque de lien entre Grandval1729 et Crémieux1858 et permet de bien observer différents chemins pour les lier indirectement. Les deux œuvres ont des relations communes avec les quatre autres textes, mais leur force varie, notamment pour Grandval1729 qui a deux traits communs avec Martin1649 et avec Dorvigny1775.

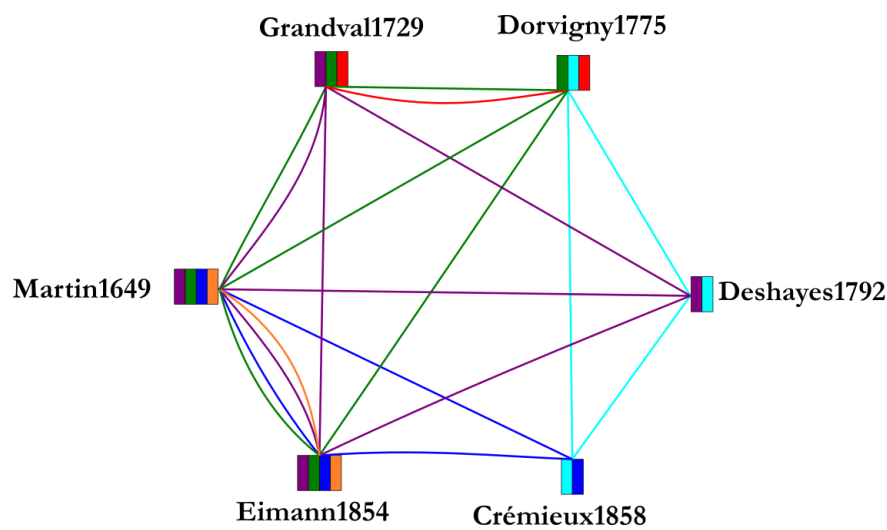


Figure 26. Spatialisation non-linéaire de la modélisation présentée en Figure 25

Les deux chemins, une qui passe par Martin1649 et l'autre qui traverse Dorvigny1775, ont la force égale (2+1 traits communs), contrairement au chemin qui passe par Eimann1854, pour lequel le lien est plus faible (1+1). Cela implique que, pour comparer plus généralement le texte de Crémieux1858 et celui de Grandval1729, il serait pertinent de prendre en compte également deux autres œuvres, celle de Martin1649 et celle de Dorvigny1775, qui peuvent, par moments, servir de passerelles. Cela peut

⁸⁸ Pour la théorie qui reprend ce terme botanique en philosophie, voir Gilles Deleuze et Félix Guattari, *Rhizome : introduction*, Paris, Éditions de Minuit, 1976. Nous privilégions ce concept, que nous opposons à celui de l'arborescence, pour insister sur la dimension non-ordonnée et simultanée des relations qui lient les composants du modèle, mais nous n'incluons pas toutes les propriétés du rhizome statuées par Deleuze et Guattari.

paraître curieux, car le texte d'Eimann¹⁸⁵⁴ est plus proche, chronologiquement, de celui de Crémieux¹⁸⁵⁸. Cependant, l'éloignement de leur inscription générique⁸⁹, entre une nouvelle et un opéra-bouffe, et la différence de leurs modalités de l'énonciation⁹⁰ (différence de la diégèse, présence du récit-cadre chez Eimann) rend leur bagage référentiel très différent, contrairement à l'opéra-comique de Dorvigny¹⁷⁷⁵, et même le poème de Martin¹⁶⁴⁹, plus proche à son tour de la cantate lyrique de Grandval¹⁷²⁹.

⁸⁹ Voir Ute Heidmann, « Différencier au lieu d'universaliser. Comparer les façons de (r)écrire des mythes », *loc. cit.*

⁹⁰ *Ibid.*

Conclusion du chapitre

Un traitement semblable, qui permet d'apprécier le degré de proximité et de situer les différentes relations observées entre les textes dans un réseau non ordonné, offre un nouveau regard sur le concept de l'intertextualité et ouvre l'analyse vers des horizons encore inexplorés. En proposant d'inclure les outils computationnels à notre traitement, nous espérons qu'ils nous aideront à approfondir, voire à nuancer, nos analyses empiriques. Les axes de recherches sont multiples et les possibilités offertes par le numérique nombreuses. Toutefois, pour en profiter pleinement, il nous semble important d'approcher les relations potentiellement présentes dans le corpus sans *aprioris* théoriques et de ne reconsidérer le concept de l'intertextualité dans son contexte critique qu'une fois devant les résultats de notre traitement. Nous espérons que cette approche nous permettra non pas de redéfinir le concept, mais de montrer qu'il gagne à être considéré dans sa pluralité.

Avant de nous intéresser davantage aux traitements computationnels dédiés à la détection de l'intertextualité, il nous paraît impératif de revenir vers les sources linguistiques de l'intertextualité, tant du point de vue théorique que, plus concrètement, par le biais de l'observation pratique de notre corpus. Nous considérons en effet que les phénomènes lexicaux, syntaxiques et sémantiques jouent un rôle important dans la détection des relations potentiellement intertextuelles. Dans les exploitations informatiques à venir, notre but n'est pas de savoir *combien* de relations textuelles chaque texte contient, mais *comment* ces relations se réalisent et quels indices permettent de les identifier.

Par exemple, les périphrases, métaphorisations ou changements de registre sont de précieux indices, dans le sous-corpus comique, de la manière dont la langue véhicule l'humour et le rire. Ils permettent également de mieux déterminer ce qui caractérise les dépréciations de la trame principale du mythe. Nous considérons aussi que la détermination des mots-clés caractéristiques des œuvres parodiques peut faciliter l'identification des passages humoristiques dans des textes qui ne sont pas, *a priori*, repérés comme en contenant. Dans ce contexte, nous nous attendons à ce que les traitements (semi-)automatiques nous accompagnent au mieux dans l'identification des particularités linguistiques de ces textes, notamment en nous permettant de nous focaliser sur les facteurs communs et les unités linguistiquement pertinentes, introduisant souvent une dépréciation ou une ridiculisation au sein des parodies, comme certaines périphrases des entités (« Pluton » / « Prince enfumé »), certaines substitutions synonymiques (« se précipiter » / « dégringoler ») ou hyperonymiques (« femme » / « femelle »), certaines métaphorisations (« descendre aux Enfers » / « aller au Diable »), etc. Pour mieux observer ces phénomènes, nous présentons, dans la partie suivante, la manière dont le concept de l'intertextualité est réinvesti dans les travaux des linguistes, notamment ceux issus de la linguistique textuelle et de la linguistique du corpus, avant de nous intéresser aux

caractéristiques des microstructures linguistiques, potentiellement porteuses de l'intertextualité, comme les paraphrases ou les reformulations. Nous proposons ensuite quelques exemples de travaux particuliers, issus du traitement automatique des langues (TAL) qui nous permettront de mettre en lumière le lien étroit, au sein de notre recherche, entre la littérature, la linguistique et l'informatique.

2. Procédés linguistiques propres des relations intertextuelles

Est-il possible de concevoir pleinement le phénomène de l'intertextualité sans s'aventurer vers le domaine de la linguistique ? Dans le chapitre précédent, nous avons rappelé que la théorisation préliminaire de Kristeva a été influencée par le dialogisme de Bakhtine. Dans ses travaux postérieurs, notamment dans *La Révolution du langage poétique*, la chercheuse semble renvoyer plus volontiers à la poétique de Jakobson et à la grammaire générative de Chomsky. Cette dernière inspiration est particulièrement visible dans l'article « Pour une sémiologie des paragrammes » (1967) où Kristeva propose une réflexion sur la signifiante considérée comme un système paragrammatique en se référant aux modèles formels des mathématiques¹. Les travaux de Michel Arrivé, quant à eux, se focalisent tantôt sur les théories chomskiennes et hjemsleviennes², tantôt sur les fondements du concept de l'intertextualité dans la recherche sur les anagrammes de Ferdinand de Saussure³. Henri Béhar, enfin, dans « Le collage, ou la pagure de la modernité »⁴, propose un classement des opérations intertextuelles qui renvoie au modèle de communication de Jakobson⁵. Même si les influences linguistiques paraissent moins importantes dans les travaux postérieurs au courant structuraliste, comme ceux de Jenny ou de Riffaterre, ceux-ci s'intéressent forcément aux phénomènes linguistiques décelables dans des microstructures lexicales, grammaticales, sémantiques et stylistiques.

En revanche, l'intérêt pour l'intertextualité paraît moins prononcé dans les travaux issus du domaine de la linguistique. Ils ne sont pas pour autant inexistantes : nous nous attachons en particulier aux recherches issues du champ de la linguistique textuelle⁶, et notamment à celles de Jean-Michel Adam⁷. Dans son ouvrage *Souvent textes varient : génétique, intertextualité, édition et traduction*, le chercheur constate que « les signes et les énoncés ne sont [...] pas un stock neutre de mots et de formules ; ils sont lourds d'emplois antérieurs et associés à des énoncés plus ou moins longs [...] qui circulent dans la mémoire d'une communauté discursive »⁸. Ce constat, qui dépasse le cadre littéraire, renvoie le concept à ses origines dialogistes. En effet, dans son livre, Adam définit l'in-

¹ Julia Kristeva, « Pour une sémiologie des paragrammes », *loc. cit.*

² Michel Arrivé, *Les Langages de Jarry*, *op. cit.* Le chercheur y vérifie notamment, à partir de l'exemple de *La Revue Blanche* de Jarry, la théorie des langages de dénotation et de connotation qui se réalisent tous deux dans le texte littéraire.

³ Michel Arrivé, « Intertextualité et intratextualité chez Ferdinand de Saussure ».

⁴ Henri Béhar, « Le collage, ou la pagure de la modernité », *loc. cit.*

⁵ Roman Jakobson, *Essais de linguistique générale*, trad. Nicolas Ruwet, Paris, Éditions de Minuit, 1963.

⁶ Eugenio Coseriu, *Textlinguistik : Eine Einführung*, Francke, Tübingen-Basel, 1994 [1980].

⁷ Jean-Michel Adam, *La linguistique textuelle*, *op. cit.*

⁸ Jean-Michel Adam, *Souvent textes varient : génétique, intertextualité, édition et traduction*, Paris, Classiques Garnier, coll. « Investigations stylistiques », 2018, p. 213.

tertextualité comme « un phénomène linguistique dialogique qui fait qu'un énoncé réécrit des énoncés empruntés à un autre texte, littéraire ou non »⁹. Le terme de « dialogue intertextuel », présent également dans les travaux sur la comparaison référentielle et discursive d'Ute Heidmann¹⁰, insiste particulièrement sur l'aspect dynamique du phénomène. Tant pour Adam que pour Heidmann, le rapport intertextuel gagne à être considéré « comme une *réponse* à une proposition de sens d'un texte ou d'un discours antérieur »¹¹, plutôt que comme un emprunt, une influence ou une imitation. Par conséquent, alors que certaines théories insistent sur l'aspect hiérarchique des relations intertextuelles, l'approche dialogique implique plutôt une relation de réinterprétation de sens. Cette vision non hiérarchique permet de se libérer de l'approche chronologique et d'un système d'interdépendance entre une œuvre moderne et sa source plus ancienne. C'est d'autant plus vrai dans le cas des mythes, car, comme le constate Heidmann,

Le fait de considérer le phénomène des mythes à partir de leur énonciation [...] permet de quitter le rapport hiérarchique dans lequel on place communément leurs représentations anciennes et modernes. L'énonciation constitue en effet un plan d'analyse commun à *tous* les ouvrages (r)écrivant (ou reconfigurant) des mythes, anciens *et* modernes. Au lieu de considérer les représentations anciennes principalement dans leur rôle de sources et au lieu de limiter le potentiel sémantique des modernes à la reprise et à l'imitation des anciennes, on peut dès lors les comparer par rapport à leur potentialité commune de créer des effets de sens dans le rapport complexe avec leurs contextes d'énonciation respectifs.¹²

La complexité des phénomènes engagés dans les relations intertextuelles dépasse en effet la dichotomie hiérarchique entre la réécriture et le texte source, ce qui rejoint la conception rhizomique que nous avons proposée au chapitre précédent¹³.

En outre, la conceptualisation interdisciplinaire d'Adam et d'Heidmann montre qu'il est possible de combiner avec pertinence les recherches littéraires et linguistiques. Toutefois, pour que cette collaboration soit fructueuse, elle doit s'éloigner « de la confortable juxtaposition pluridisciplinaire que certains pratiquent en évitant soigneusement tout risque de réfutation des positions et d'ébranlement des certitudes qu'un véritable dialogue entre disciplines entraîne inévitablement »¹⁴. Par ailleurs, dans son article « Méthodes sans questions, questions sans méthodes », Laurent Jenny déclare :

⁹ *Ibid.*, p. 211.

¹⁰ Ute Heidmann, « Différencier au lieu d'universaliser. Comparer les façons de (r)écrire des mythes », *loc. cit.* Heidmann et Adam collaborent régulièrement pour prôner l'intérêt d'une approche interdisciplinaire d'un texte littéraire. Voir notamment Jean-Michel Adam et Ute Heidmann, *Le texte littéraire : pour une approche interdisciplinaire*, Louvain-la-Neuve, Bruylant-Academia, coll. « Au cœur des textes », 2009.

¹¹ Ute Heidmann, « Différencier au lieu d'universaliser. Comparer les façons de (r)écrire des mythes », *loc. cit.*, p. 20.

¹² *Ibid.*, p. 16-17.

¹³ Voir II.1d) « Théories intertextuelles et observation du corpus », p. 139.

¹⁴ Jean-Michel Adam, *Souvent textes varient*, *op. cit.*, p. 14.

[L]e problème tient à la dissociation, dans les études littéraires, des méthodes et des questions. D'un côté, les études littéraires élaborent les outils descriptifs sans questions à poser aux contenus. De l'autre, les discours experts des sciences humaines posent des questions aux « contenus » littéraires sans réflexion de méthode sur leur mise en discours. Si l'on veut sauver les études littéraires, il faut réconcilier en elles méthodes et questions.¹⁵

Outre cette position éclairée concernant le dialogue interdisciplinaire, nous constatons que tant Adam¹⁶ que Jenny¹⁷ rejettent catégoriquement la pratique de la lecture distante et l'approche numérique des textes littéraires. Nous pensons prouver que, à condition de ne pas faire abstraction de l'herméneutique et de ne pas « placer les moyens avant les fins »¹⁸, la lecture distante, considérée en tant qu'étape de recherche et mobilisée pour mieux explorer et exploiter le corpus, ne va à l'encontre ni du dialogue intertextuel ni du dialogue interdisciplinaire.

Pour ce faire, nous proposons de nous éloigner, temporairement, des dimensions philosophique et psychologique, forcément arbitraires, du concept d'intertextualité et de nous intéresser à ce qui est directement observable, du point de vue de la linguistique, au sein d'un ensemble concret de textes. Dans la section « Linguistique du corpus et herméneutique » de son article « La linguistique du corpus : une alternative », Wolfgang Teubert constate :

En faisant référence à ce qui a été dit précédemment, la fonction de l'intertextualité ajoute une nouvelle couche d'interprétation à quelque chose qui a été dit auparavant. Ma communauté interprétative peut maintenant négocier dans quelle mesure ce que j'ai dit est acceptable au titre d'une nouvelle interprétation. Ce qui signifie que le concept d'intertextualité est arbitraire.¹⁹

¹⁵ Laurent Jenny, « Méthodes sans questions, questions sans méthodes », *Versants. Revue suisse des littératures romanes*, vol. 1, n° 63, « À quoi bon l'enseignement de la littérature ? », octobre 2017, p. 81, en ligne, <<https://bop.unibe.ch/versants/article/view/3574>>.

¹⁶ « [J]e me démarquerai de tendances critiques comme la théorie des “textes possibles”, la “méthode poststructurelle [sic]” est surtout le *distant reading*, sociologisant et positiviste, prôné par Franco Moretti et ses émules. [...] [C]e livre a pour but de penser les variations non pas virtuelles, mais réalisées, attestées et, de ce fait, directement étudiables de certains textes. », Jean-Michel Adam, *Souvent textes varient*, *op. cit.*, p. 15.

¹⁷ « Trop souvent [...] les littéraires ont renoncé à faire apercevoir la valeur cognitive de la littérature pour se replier sur des territoires restreints : travaux minutieux de génétique nourrissant les éditions savantes, biographisme, historicisme, affinement des outils méthodologiques, particulièrement narratologiques, et aujourd'hui *digital humanities* », Laurent Jenny, « Méthodes sans questions, questions sans méthodes », *loc. cit.*, p. 79.

¹⁸ *Ibid.* Jenny formule cette critique concernant les travaux issus des humanités numériques à partir d'une expérience personnelle : « [L]e responsable d'un gros projet de numérisation de corpus littéraire [...] me confiait récemment qu'au terme de trois ans de travail acharné, et au prix de considérables investissements financiers, on commençait à voir “à quoi pourrait servir” cet immense effort. C'est donc que la démarche a consisté à chercher non pas quel outil pourrait répondre à une question préalable mais plutôt quelle question pourrait bien être conforme à l'outil mis en place. Comment les questions ainsi “trouvées” pourraient-elles ne pas répondre à l'outil (à sa logique et à son réductionnisme) bien davantage qu'à une inquiétude ou un désir authentiquement partagé par une communauté humaine ? ». Même si nous comprenons qu'il est difficile, pour la recherche littéraire traditionnelle, de concevoir la dimension analytique et scientifique de l'établissement d'un corpus numérique qui prévaut régulièrement sur les préoccupations techniques, nous sommes étonnée de constater avec quelle facilité les lacunes méthodologiques d'un projet sont attribuées à toute une discipline.

¹⁹ Wolfgang Teubert, « La linguistique de corpus : une alternative », *Semen. Revue de sémio-linguistique des textes et discours*, n° 27, trad. Aurélie Lebaud, Presses universitaires de Franche-Comté, juin 2009, paragr. 56, en ligne, <<https://journals.openedition.org/semen/8923#tocto1n10>>.

En effet, divers outils peuvent nous aider à repérer d'éventuelles similarités entre des unités textuelles (textes entiers, paragraphes, phrases, unités lexicales...), mais aucun mécanisme formel n'est en mesure de nous garantir que deux textes sont en relation intertextuelle. Peu importe, au fond, si les relations-candidates ont été repérées automatiquement ou empiriquement, puisque « [l']intertextualité est établie non par un ordinateur, mais par son interprétation »²⁰. En revanche, alors que le chercheur peut "pressentir" des liens entre des textes grâce à ses connaissances, sa culture et son intelligence, la sélection des relations-candidates à l'aide d'un outil numérique nécessite une approche plus formelle. Selon Teubert,

[t]out texte nouveau répète, dans une large mesure, des choses qui ont déjà été dites. Les segments de texte, collocations, items lexicaux simples et complexes qui ont été utilisés auparavant sont recombinaisonnés, permutés et insérés dans de nouveaux contextes, contenant quelques nouveaux contenus paraphrastiques joints avec les anciens. Un nouveau texte reprend, accepte, commente, modifie, réinterprète ou rejette des textes préalables. L'essentiel de ce que nous disons est une reformulation de choses dites avant, mais chaque production, normalement, apporte une petite innovation.²¹

Les « segments de texte, collocations, items lexicaux simples et complexes » mentionnés par Teubert constituent des indices formels précieux pour la recherche des relations-candidates, qui peuvent ensuite être soumis à l'analyse empirique. Mais notre corpus est très hétérogène. La proximité des réécritures par rapport au mythe est variable, et souvent difficile à évaluer. Certains auteurs citent une traduction française, alors que d'autres composent leur récit sans s'appuyer sur un texte source précis. Souvent, il se fient à leur mémoire, qui est parfois défaillante. Tous peuvent aussi mêler différentes sources antiques, éventuellement en modifiant leurs épisodes, ou ces sources et d'autres réécritures, parfois mieux connues.

Pour analyser les réinterprétations du mythe avec les outils numériques et pouvoir prendre en compte des relations allusives entre réécritures, intéressons-nous à un plus petit corpus, plus homogène car composé de différentes traductions françaises du livre X des *Métamorphoses* d'Ovide. Elles font partie de notre corpus secondaire, bien distinct de celui des réécritures (que nous appelons le corpus principal) et laissé de côté jusqu'alors, mais qui est pourtant crucial pour notre recherche. Son importance sera bien mieux valorisée après avoir observé les limites du traitement des seules réécritures du mythe d'Orphée et Eurydice. Cependant, nous mobilisons l'ensemble des traductions d'Ovide ici, d'une part, parce que les séries traductives sont particulièrement adaptées à l'observation du phénomène de reformulation et de paraphrase et, d'autre part, pour ne pas dévoiler, pour le moment, les relations-candidates présentes dans notre corpus principal.

²⁰ *Ibid.*, paragr. 58.

²¹ *Ibid.*, paragr. 53.

Ainsi, avant de nous pencher davantage sur les phénomènes linguistiques, nous analysons quelques relations particulières entre ces traductions en nous demandant si nos constats coïncident avec la théorie des périodes cycliques dans l'histoire de la traduction exposée par Johann Wolfgang von Goethe²². En comparant 21 traductions de la première phrase du livre X des *Métamorphoses*, nous proposons tout d'abord une analyse "distante", appuyée sur une schématisation visuelle. À partir de ces observations générales, nous analysons plus en détail les relations effectives recensées dans notre échantillon et les classons selon la méthode proposée par Marta Vila, Antonia Martí et Horacio Rodríguez²³.

Outre l'investigation littéraire et linguistique, cette étude nous aide à construire un cadre formel nous permettant de mieux comprendre la nature des phénomènes linguistiques engagés dans les relations intertextuelles et de préciser nos besoins concernant la méthode pour les détecter (semi-)automatiquement.

a) Investigation à partir de 21 traductions d'Ovide

Si notre corpus principal, que nous avons présenté dans la partie précédente²⁴, se compose des 53 réécritures françaises du mythe d'Orphée et Eurydice, le corpus secondaire est composé de deux séries traductives. La première regroupe 16 traductions de Virgile, la deuxième rassemble 21 traductions d'Ovide²⁵. Parmi les traductions de Virgile, la grande majorité date du XIX^e s., alors que les dates de publication de celles d'Ovide sont beaucoup plus variées et se répartissent de manière plus égalitaire entre le XV^e s. et le XIX^e s. Si nous ne recensons pas davantage de traductions contemporaines, c'est parce que notre corpus des réécritures regroupe très peu d'œuvres publiées entre le XX^e s. et le XXI^e s. Celles-ci se limitent aux romans et albums de jeunesse et aux bandes dessinées, ce qui nous incite à ajouter tout de même au corpus secondaire la traduction des *Métamorphoses* de Françoise Rachmuhl²⁶, destinée au jeune public. Au niveau de la forme, les traductions de Virgile sont plus diversifiées que celles d'Ovide : sept parmi elles sont en vers (dont quatre rimées) et neuf sont en prose, alors que celles d'Ovide sont majoritairement en prose (six traductions seulement sur 21 sont en vers, dont quatre contiennent des rimes). Enfin, ce sous-corpus inclut le cycle d'*Ovide moralisé*, soit une série de quatre œuvres, traduisant (et adaptant) Ovide en ancien puis

²² Johann Wolfgang von Goethe, « Traductions », *Divan occidental-oriental*, trad. Henri Lichtenberger, Paris, Aubier, 1940 [1819], p. 430-433.

²³ Marta Vila, M. Antonia Martí et Horacio Rodríguez, « Paraphrase Concept and Typology. A Linguistically Based and Computationally Oriented Approach », *Procesamiento de Lenguaje Natural*, n° 46, 2011, p. 83-90.

²⁴ Voir I. « Présentation du corpus des réécritures du mythe d'Orphée et Eurydice ».

²⁵ Pour la liste complète des composants de ce corpus, voir le fichier [II_2_Corpus-secondaire.xlsx](#).

²⁶ Françoise Rachmuhl, « Vie et mort d'Orphée », *16 métamorphoses d'Ovide*, Paris, Castor Poche Flammarion, 2003, p. 139-147. Si nous ne disposons pas de textes analogues pour les *Géorgiques*, c'est parce que cette œuvre n'est pas au programme du second cycle.

en moyen français²⁷. Pour les *Géorgiques*, un texte seulement est en moyen français, celui de Guillaume Michel (dit de Tours)²⁸.

Une telle hétérogénéité garantit la richesse des phénomènes linguistiques, traductologiques et littéraires qui peuvent être observés au sein de chaque série traductive, voire en comparant les traductions de Virgile à celles d'Ovide. Voyons, par exemple, les relations entre les 21 traductions de la première phrase du livre X des *Métamorphoses* (a-v). Nous distinguons six segments dans le texte latin (a.).

Comme il s'agit d'une segmentation expérimentale et intuitive faite en prenant en compte les 21 traductions, le découpage en segments est davantage sémantique que syntaxique (nous ne pouvons donc pas parler des syntagmes ici). Par exemple, le segment 2, « per immensum æthera digreditur » est interrompu par l'incise « croceo velatus amictu », comme l'adjectif « ciconumque » du segment 4 est séparé du syntagme prépositionnel « ad oras » qu'il qualifie par le sujet « Hymenæus ». Un tel regroupement reste problématique et tend à masquer l'importance de certaines unités plus petites. Par exemple *nequicquam* (littéralement « en vain, inutilement »²⁹) est intégré dans le segment 6, mais, sémantiquement, il est à relier plutôt avec la phrase suivante du livre X, « Adfuit ille quidem, sed nec solemnia verba, / Nec lætos vultus, nec felix allulit omen. », puisque ce n'est pas l'appel d'Orphée qui est vain (Hyménée est bien présent, « adfuit ille quidem »), mais la venue du dieu de mariage, qui n'apporte ni la parole consacrée (« solemnia verba »), ni regard favorable (« lætos vultus »), ni heureux présage (« felix omen »)³⁰. De même *inde*, « de là », renvoie au mariage d'Iphis et Ianthé qui clôt le livre IX³¹.

Consciente de ces approximations, nous restons donc vigilante dans notre analyse aux variations et subtilités que notre schématisation risque de cacher tout en profitant des capacités de cet approche à révéler quelques régularités.

A. Ovide, *Métamorphoses* : texte latin *vs* 21 traductions françaises ([segments correspondants]₁₋₆) :

a. Ovide : [Inde]₁ [per immensum]₂ [croceo velatus amictu]₃ / [æthera digreditur]₂ [ciconumque]₄ [Hymenæus]₅ [ad oras / tendit ;]₄ [et Orpheà nequicquam voce vocatur.]₆

²⁷ Anonyme, *Ovide moralisé*, Rouen, 1315, 246v-248r p.; Maître de Marguerite d'York, *Ovide moralisé en prose II*, Bruges, 1470, 132v-135r, 147r-148r p.; Thomas Walleys, *La Bible des poètes, Métamorphose [d'Ovide moralisée]*, trad. Colard Mansion, Paris, Antoine Vêrard, 1493, 107v-109v, 119r-199v; Ovide, *Le grand Olympe des histoires poétiques du prince de poésie Ovide Naso en sa Métamorphose, œuvre authentique et de hault artifice, pleine de honneste récréation, traduyct de latin en françois et imprimé nouvellement*, Lyon, R. Morin, 1532, 81r-83r, 2r-3r.

²⁸ Virgile, « Les Georgiques », *Les œuvres de Virgile. Translatees de latin en [vers] francoyses*, trad. Michel Guillaume (dit de Tours), Paris, Jean Petit, Pierre Vidoue, 1540 [1519], p. 19-29, en ligne, <<https://books.google.fr/books?id=EeksmgevUeoC&hl=fr&pg=PA3#v=onepage&q&f=false>>.

²⁹ Voir s. v. *nēquicquam*, *nēquidquam*, *nēquīquam*, Félix Gaffiot, *Dictionnaire illustré latin-français*, Paris : Hachette, 1934, p. 1025.

³⁰ Voir aussi notre analyse à la page 177.

³¹ Nous nous y penchons davantage à partir de la page 183.

- b. Legouais1301 : [Hymen diex de nocoïement]₅ [d'illuec]₁ [s'en part grant aleure par l'air]₂ [en janne couverture]₃ [en Ciconie]₄ [ou semons iere a nocés d'estrange maniere ; Orpheus semons li avoit qui nouvele espouse devoit espouser gente et joenne et bele Euridice la damoisele.]₆
- c. York1470 : [Le dieu Hymen]₅ [fut a ses nopces et quant d'illec]₁ [se parti grant erre s'en ala par l'air]₂ [couvert de janne couverture]₃ [en Cyconie]₄ [ou semons estoit aux nopces de estrange maniere de par Orpheus qui nouvelle femme devoit espouser : c'estoit Erudice qui moult estoit belle, gente et jone.]₆
- d. Vêrard1493 : [Et quant]₁ [Hymen]₅ [se partit, par l'air s']₂[en]₁ [vola grant erre]₂ [en Cyconie,]₄ [où semons estoit de par Orpheus qui devoit espouser Erudice.]₆
- e. Morin1532 : [Et quant]₁ [Hymen]₅ [se partit, par l'air s']₂[en]₁ [volla grant erre]₂ [en Cyconie,]₄ [où semons estoit de par Orpheus qui devoit espouser Eurydice.]₆
- f. Habert1557 : [Le Dieu Hymen]₅ [de Crete]₁ [s'en alla, et parmy l'air legerement vola,]₂ [vestu de Pourpre]₃ [et sans long intervalle droict au pays Thracien il devalle,]₄ [où par Orphée aux nopces appelé,]₆
- g. Renouard1606 : [Quand]₁ [le Dieu Nopcier]₅ [se retira]₂ [du festin qui se fit aux espousailles d'Iphis et d'Ianthe,]₁ [il se jeta dans l'air]₂ [couvert de sa robe jaune,]₃ [et pris le chemin de Thrace]₄ [où l'attiroit la voix charmeresse d'Orphée, qui l'appeloit.]₆
- h. Massac1617 : [De là]₁ [par l'infini]₂ [le nocier Hymenee,]₅ [couvert de son manteau de couleur safrance,]₃ [guide son vol en Thrace au terroir Ciconin,]₄ [par Orphée imploré, qui l'esperoit benin,]₆
- i. Corneille1687 : [Hymen]₅ [quitte]₂ [la Crete,]₁ [et volant vers la Thrace,]₄ [appelé par Orphée, il va prendre sa place, où ce Chantre fameux par des mots solempnels doit livrer sa franchise à des nœuds éternels.]₆
- j. Martignac1697 : [Le Dieu des nopces]₅ [vêtu de jaune]₃ [s'envola]₂ [d'Égypte]₁ [en Thrace]₄ [où la voix d'Orphée l'appelloit.]₆
- k. Bellegarde1701 : [Le Dieu du mariage]₅ [vêtu d'une robe couleur d'or]₃ [s'élevant en l'air,]₂ [prit le chemin de la Thrace,]₄ [attiré par les charmes de la voix d'Orphée.]₆
- l. Duryer1702 : [De là]₁ [Hymen le Dieu des nocés,]₅ [vêtu d'une robe de jaune-doré,]₃ [s'éleva en l'air]₂ [pour aller en Thrace,]₄ [où l'appelloit la voix d'Orphée pour assister à son mariage.]₆
- m. Banier1732 : [Le Dieu Hyménée,]₅ [couvert d'une Robbe couleur de feu,]₃ [aïant pris son essor au milieu des airs,]₂ [vola du côté de la Thrace ;]₄ [où il étoit attiré par le charme de la voix d'Orphée, qui le prioit de rendre son mariage heureux.]₆
- n. Fontanelle1789 : [L'Hymen,]₅ [vêtu d'une robe de pourpre,]₃ [quittant]₂ [la Crete]₁, [traversant les airs,]₂ [se rendit dans la Thrace,]₄ [attiré par la voix d'Orphée ;]₆
- o. Villenave1806 : [L'Hymen,]₅ [vêtu d'une robe de pourpre,]₃ [s'élève]₂ [des champs de Crète]₁ [dans les airs,]₂ [et vole vers la Thrace,]₄ [où la voix d'Orphée l'appelle en vain à ses autels.]₆
- p. Gros1834 : [Des champs de la Crète]₁ [l'Hymen,]₅ [vêtu d'une robe de pourpre,]₃ [s'élance dans les vastes plaines de l'air]₂ [et dirige son vol vers la Thrace,]₄ [où, d'une voix impuisante, Orphée invoque son appui.]₆
- q. Nisard1869 : [De là,]₁ [par les champs de l'espace,]₂ [Hyménée,]₅ [couvert de tissus éclatants,]₃ [s'élance vers les rives de l'Hèbre.]₄ [Il vient : Orphée l'appelle, mais il l'appelle en vain.]₆
- r. Parnajon1880-1 : [De là]₁ [Hyménée,]₅ [vêtu d'une robe couleur de safran,]₃ [s'éloigne à travers les plaines immenses de l'air,]₂ [et se dirige vers les rivages des Ciconiens]₄ [où l'appelle en vain la voix d'Orphée.]₆

- s. Parnajon1880-2 : [De là]₁ [Hyménée]₅ [voilé d'un manteau couleur de safran]₃ [s'éloigne à travers l'air immense]₂ [et il se dirige vers les rivages des Ciconiens]₄ [et il est appelé en vain par la voix d'Orphée].₆
- t. Beague1998 : [De là]₁ [à travers l'immensité de l'éther]₂ [couvert de son vêtement safran]₃ [Hyménée]₅ [s'éloigne]₂ [se dirige vers le territoire des Ciconiens]₄ [et la voix d'Orphée l'y appelle en vain].₆
- u. Rachmuhl2003 : [Hyménée, le dieu du mariage]₅ [se dirigeait]₄ [à travers le ciel immense]₂ [vers la Thrace].₄ [Son manteau, couleur de safran comme un voile de mariée, flottait au vent].₃ [Il se rendait aux noces du poète Orphée].₆
- v. Boxus2018 : [S'éloignant]₂ [de là]₁ [enveloppé d'un voile couleur de safran]₃ [Hyménée]₅ [traverse l'immense éther]₂ [et gagne les rives des Cicones]₄ [tandis que l'appelle, bien inutilement, la voix d'Orphée].₆

Nous observons par exemple que la syntaxe particulière de la phrase latine incite presque la moitié des traducteurs à modifier sa construction en plaçant le sujet « Hyménée » en première position (a.-b., f., i.-k., n.-o., u.). En revanche, le texte de Béague1998 (B.b.), destiné au public universitaire, suit très fidèlement la construction du texte latin (B.a.) en ne proposant que des adaptations minimales, indispensables pour respecter les règles de la grammaire française. Il s'agit notamment du segment 5, « Hyménée », qui est remonté de deux positions, devant la seconde partie du segment 2 (« digreditur » / « s'éloigne »).

B. Ovide, *Métamorphoses* : texte latin *vs* Béague *et al.* (trad.), 1998 *vs* Rachmuhl (trad.), 2003 ([segments correspondants]₁₋₆) :

- a. Ovide : « [Inde]₁ [per immensum]₂ [croceo velatus amictu]₃ / [æthera digreditur]₂ [ciconumque]₄ [Hymenæus]₅ [ad oras / tendit]₄ [et Orpheà nequicquam voce vocatur].₆ »
- b. Beague1998 : [De là]₁ [à travers l'immensité de l'éther]₂ [couvert de son vêtement safran]₃ [Hyménée]₅ [s'éloigne]₂ [se dirige vers le territoire des Ciconiens]₄ [et la voix d'Orphée l'y appelle en vain].₆
- c. Rachmuhl2003 : [Hyménée, le dieu du mariage]₅ [se dirigeait]₄ [à travers le ciel immense]₂ [vers la Thrace].₄ [Son manteau, couleur de safran comme un voile de mariée, flottait au vent].₃ [Il se rendait aux noces du poète Orphée].₆

La stratégie traductionnelle du texte de Rachmuhl2003 (B.c.), destiné au jeune public, est très différente. Non seulement la construction de la phrase d'Ovide est complètement adaptée à la langue française, mais aussi, par souci de simplification, elle est découpée en trois phrases. De nombreux ajouts peuvent être constatés, visant principalement à expliciter les noms propres (« Hyménée, **le dieu du mariage** », « du **poète** Orphée ») et les allusions symboliques (« couleur de safran **comme un voile de mariée** »). Toutefois, il est surprenant de constater que le verbe *velo, as, are* (« se couvrir », « s'envelopper ») est remplacé par « flotter au vent ». Nous observons également que le dernier segment s'éloigne de manière importante du texte latin : le sujet de la subordonnée, la voix d'Orphée (« Orpheà voce ») disparaît. Le pronom « il », référant à Hyménée, devient le sujet

de la phrase et « le poète Orphée » prend la fonction du complément du nom *noces*, noyau du complément d'objet indirect du verbe *se rendre*. Il n'est donc plus question ni de l'invocation orale d'Orphée ni de l'inutilité de cet appel (*nequicquam*, « en vain »).

Outre les différences liées au public cible, c'est également la forme de chaque traduction qui influe sur la fidélité au texte source. En effet, les traductions en vers (C.) sont plus remaniées, notamment pour respecter la rime et le rythme des vers.

C. Ovide, *Métamorphoses* : texte latin *vs* Habert (trad.), 1557 *vs* Massac (trad.), 1617 *vs* Corneille (trad.), 1687 ([segments correspondants]₁₋₆) :

- a. Ovide : « [Inde]₁ [per immensum]₂ [croceo velatus amictu]₃ / [æthera digreditur]₂ [ciconumque]₄ [Hymenæus]₅ [ad oras / tendit ;]₄ [et Orpheà nequicquam voce vocatur.]₆ »
- b. Habert1557 : « [Le Dieu Hymen]₅ [de Crete]₁ [s'en alla, / et parmy l'air legerement vola,]₂ / [vestu de Pourpre]₃ [et sans long intervalle / droict au pays Thracien il devalle,]₄ / [où par Orphee aux nopces appelé]₆ »
- c. Massac1617 : « [De là]₁ [par l'infini]₂ [le nocier Hymenee,]₅ / [couvert de son manteau de couleur safrane,]₃ / [guide son vol en Thrace au terroir Ciconin,]₄ / [par Orphee imploré, qui l'esperoit benin,]₆ »
- d. Corneille1687 : « [Hymen]₅ [quitte]₂ [la Crete]₁, & volant vers la Thrace,]₄ / [appellé par Orphée, il va prendre sa place, / où ce Chantre fameux par des mots solennels / doit livrer sa franchise à des nœuds éternels.]₆ »

Alors que l'extrait de Rachmuhl2003 (B.c.) est le seul qui propose une traduction en plusieurs phrases, pour quatre textes, dont deux en vers (Habert1557, C.b. et Massac1617, C.c.), au moins deux phrases latines sont rassemblées au sein d'une phrase française. Curieusement, il ne s'agit pas des traductions les plus concises. Au contraire, nous recensons de nombreux ajouts au texte source, comme la précision du lieu de départ d'Hyménée (Crète) et de son arrivée (« au pays Thracien », C.b. ; « en Thrace au terroir Ciconin », C.c.), l'ajout d'adverbes et de locutions adverbiales (« legerement », « droict », « sans long intervalle »), voire l'introduction de nouvelles informations (« par Orphee imploré, **qui l'esperoit benin** », C.c.). Les ajouts et les reformulations polylexicales sont particulièrement présents pour le segment 6, « et Orpheà nequicquam voce vocatur ». Dans l'extrait de Corneille1687 (C.d.), quatre segments sont traduits de manière plutôt minimaliste, notamment avec l'omission de « per immensum æthera » et la simplification de « ciconumque ad oras » par « vers la Thrace ». Par ailleurs, le segment 3, « croceo velatus amictu », est supprimé alors que le sixième est particulièrement long. Le nom d'Orphée est remplacé par la périphrase « ce Chantre fameux » et le mariage est métaphorisé par « livrer sa franchise à des nœuds éternels ». Simultanément, toutes les informations de la phrase source ne sont pas transmises, car l'adverbe *nequicquam* n'est pas traduit.

Outre des comparaisons ponctuelles, des traductions destinées au publics différents (B.), comme celles de Rachmuhl2003 et Béague1998, ou celles composés en vers (C.), comme Habert1557, Mas-

sac1617 et Corneille1687, l'observation plus globale de toutes les traductions à la fois permet potentiellement de répondre à des questions particulières et de confronter quelques constats théoriques avec ce qui est directement observable au sein d'un corpus spécifique. Nous pouvons par exemple nous demander si les traductions en vers présentent globalement des structures syntaxiques différentes de celles en prose, ou encore si des corrélations particulières sont à observer entre des traductions qui omettent certains segments et celles qui en développent d'autres de manière importante.

Par exemple, le segment 6, « et Orpheà nequicquam voce vocatur » incite à être observé en détail, d'une part parce que la présence de l'adverbe *nequicquam* est surprenante et, d'autre part, parce que les traductions de cet extrait sont extrêmement hétérogènes, ce qui prouve l'ambiguïté de la formulation originale. Il est intéressant d'observer une certaine tendance chronologique dans le traitement de ce segment. En effet, les quatre traductions les plus anciennes, de Legouais1301 (« où semons iere a nocces d'estrange maniere ; Orpheus semons li avoit qui nouvele espouse devoit espouser gente et joenne et bele Euridice la damoisele », a.) à Morin1532 (« où semons estoit de part Orpheus qui devoit espouser Eurydice », e.), sont particulièrement explicatives et introduisent dès à présent le personnage d'Eurydice et le motif de l'appel d'Orphée. En revanche, il n'y est plus question de la voix du poète (sauf à interpréter ainsi la formulation « d'estrange maniere ») ni de l'inutilité de son appel. Parmi les neuf traductions suivantes (de Habert1557, f., à Fontanelle1789, n.) sept traduisent *voce* (« la voix »), parfois en ajoutant des informations supplémentaires (Renouard1606, g. : « la voix charmeresse d'Orphée », Corneille1687, i. : « par des mots solempnels », Bellegarde1701, k., et Banier1730, : « attiré par le(s) charme(s) de la voix d'Orphée »). Certaines maintiennent l'ajout du motif de l'appel d'Orphée (Habert1557, f. : « par Orphée aux nopces appellé », Duryer1702, l. : « pour assister à son mariage », Banier1732, m. : « qui le prioit de rendre son mariage heureux »), une propose un ajout particulier : Massac1617 (h.) traduit le segment 6 par « par Orphée imploré qui l'espéroit bénin ». Aucun texte de ce sous-ensemble ne traduit *nequicquam*. En revanche, l'adverbe « en vain » est bien présent dans toutes les traductions suivantes, à partir de celle de Villenave1806 (o.) et jusqu'à celle de Boxus2018 (v.), exception faite de la traduction de Gros1834 (p.), qui opte pour « d'une voix impuissante », et de celle de Rachmuhl2003 (u.), destinée à la jeunesse, qui rassemble plutôt aux traductions antérieures à 1806 : « Il se rendait aux nocces du poète Orphée ».

La présence de ces trois séries chronologiques incite à observer notre corpus en parallèle avec certaines théories traductionnelles. Par exemple, dans le chapitre « Traductions » du *Divan occidental-oriental*, Johann Wolfgang von Goethe³² différencie trois tendances traductionnelles qui émergent

³² Johann Wolfgang von Goethe, « Traductions », *op. cit.* Nous remercions Alison Boulanger dont le séminaire « Pratique de la traduction littéraire » (M1 Lettres modernes, Université de Lille, 2015) a été d'une aide précieuse pour appuyer et documenter nos observations.

en périodes cycliques dans l'histoire littéraire afin d'opérer « la transition de l'étranger à l'indigène, du connu à l'inconnu »³³. La période initiale permet un premier contact avec l'étranger et consiste, pour la plupart, en une traduction simple, en prose. On ne cherche pas spécialement à rendre la poésie, mais à produire un texte aussi accessible que possible. La deuxième période, appelée *parodistique* s'engage dans l'interprétation et cherche à tout prix l'équivalent dans la culture du public cible. Cela implique la perte de tout ce qui fait la différence par rapport au contexte littéraire et culturel du texte source. Ce type de traduction ne s'intéresse pas tellement à l'étranger, mais cherche à se conformer au goût des lecteurs. Enfin, la dernière période « voudrait rendre la traduction identique à l'original, en sorte qu'elle pût non seulement le rendre de façon approximative, mais en tenir lieu »³⁴, c'est la seule qui prête attention à l'étrangeté de la forme originale en essayant à tout prix de l'imiter, et qui n'a pas pour tâche de faciliter le travail du lecteur. En choisissant de s'y inscrire, le traducteur se condamne à être incompris jusqu'à ce que d'autres le rejoignent pour participer à l'évolution des habitudes du public.

Nous pouvons tout d'abord nous demander quels critères permettraient de déterminer la présence de ces périodes dans notre corpus, pour voir ensuite comment (et si) la théorie de Goethe s'y réalise. En effet, selon le poète, « ces trois périodes se reproduisent, parfois en sens inverse et [...] les trois méthodes de traduction peuvent être pratiquées en même temps. »³⁵. Le cas échéant, rien n'exclut de les exploiter conjointement :

On [se] servirait [de la version en prose] pour la lecture cursive destinée à nous donner une idée sommaire du contenu, nous prendrions plaisir au côté historique, fabuleux, moral et nous nous familiariserions toujours davantage avec la façon de sentir et de penser, jusqu'au moment où nous serions en état de fraterniser complètement avec ces œuvres.³⁶

Si cette complémentarité est effectivement observable au sein des séries traductives recueillies, leur apport à notre analyse des réécritures du mythe sera conséquent, notamment si les traitements informatiques mis en place les exploitent comme un ensemble cohérent.

La schématisation des traductions de la première phrase du livre X des *Métamorphoses* d'Ovide (Tableau 15) permet d'apporter des éléments de réponse à notre investigation. Pour chaque œuvre, les six segments correspondants à ceux de la phrase latine ont été mis en couleur respectivement en bleu foncé, rouge, orange, violet, vert et bleu azur, comme montré ci-dessous avec l'exemple de la phrase originale d'Ovide.



³³ *Ibid.*, p. 433.

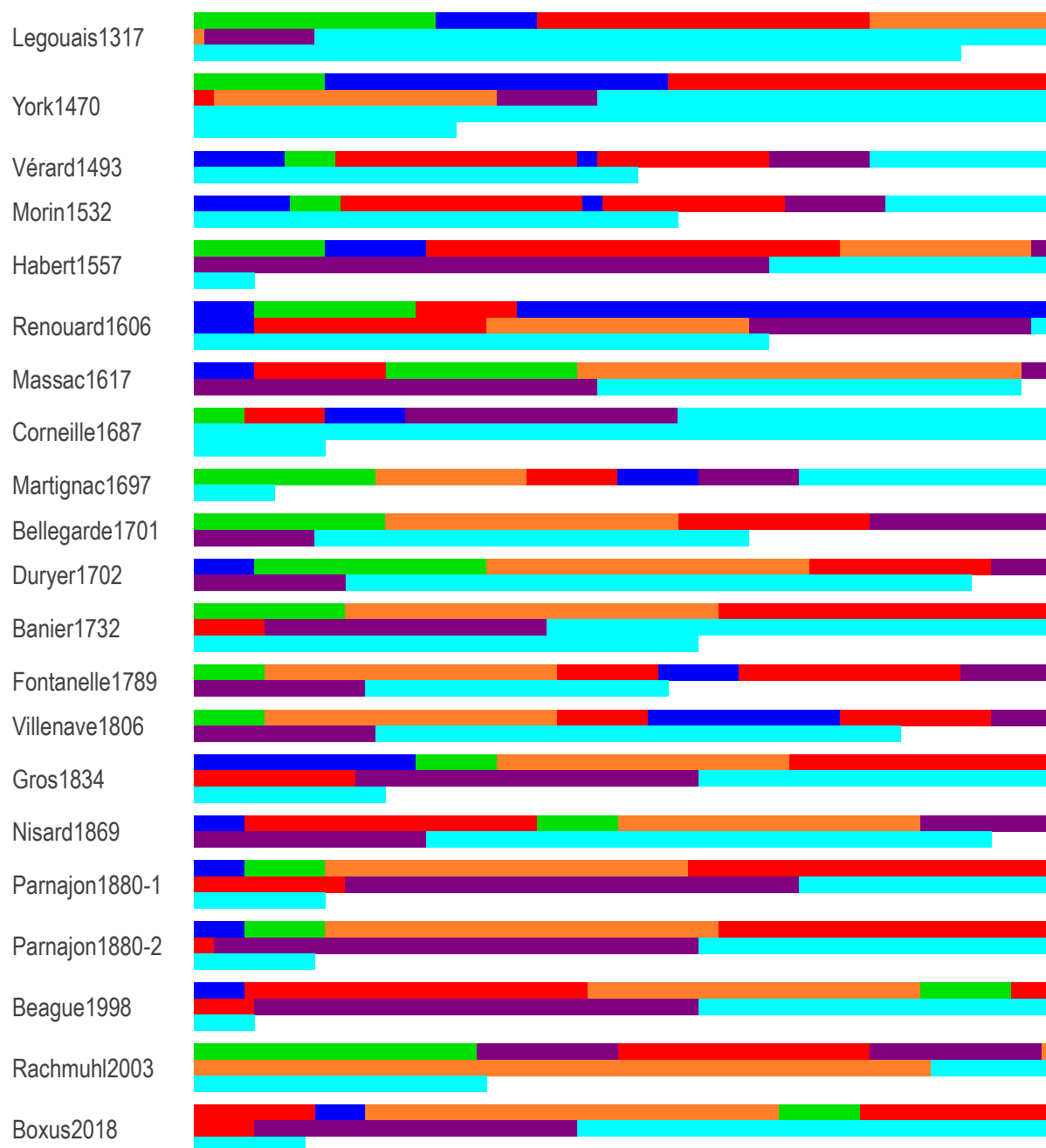
³⁴ *Ibid.*, p. 431.

³⁵ *Ibid.*, p. 432.

³⁶ *Ibid.*

Une telle schématisation permet, potentiellement, de dévoiler quelques tendances ou, au contraire, des irrégularités, notamment concernant l'ordre dans lequel les segments sont traduits, la suppression de certains extraits, voire la variation de la longueur de chaque traduction et de chaque segment (pour permettre la comparaison relativement objectivable de ces facteurs, chaque graphème ou espace a été remplacé par le même espace colorisé).

Tableau 15. Schématisation des longueurs et positions correspondant aux 6 segments de la première phrase du livre X des Métamorphoses d'Ovide dans les 21 traductions³⁷



³⁷ Pour une visualisation qui compile la schématisation à l'aide des codes-couleurs avec le texte des traductions, voir les annexes de cette partie, le fichier [II_2_phrase-1-Ovide.pdf](#).

Nous voyons que, pour plusieurs traductions, la couleur bleu azur, symbolisant le segment 6 (« et Orpheà nequicquam voce vocatur »), est prédominante. Alors que la traduction la plus courte de ce segment, celle de Fontanelle1789, ne compte que 30 caractères (« attiré par la voix d'Orphée »), celle de York1470 (« où semons estoit de par Orpheus qui nouvelle femme devoit espouser gente et joenne et bele Euridice la damoisele ») en contient 155 et pour celle de Corneille1687, la longueur du dernier segment constitue 75 % des caractères de la traduction (136/183, « appelé par Orphée, il va prendre sa place, où ce Chantre fameux par des mots solennels doit livrer sa franchise à des nœuds éternels »). Si nous considérons que les traductions qui développent de manière importante certains segments sont soit des “belles infidèles” (les adaptations “parodistiques”, qui cherchent à tout prix l'équivalent, ce qui correspond à la deuxième période de Goethe) ou des adaptations explicatives destinées à un public non-expert (première période de Goethe, initiation au texte étranger, inconnu par le public), cette visualisation nous donne au moins cinq candidats : les deux traductions du cycle *Ovide moralisé*, Legouais1301 et York1470, la traduction en vers de Corneille1687, celle en prose de Banier1732 et l'adaptation pour la jeunesse de Rachmuhl2003. Comme le but principal de cette dernière est de rendre le texte d'Ovide aussi accessible que possible, elle doit forcément s'adapter à son public, ce qui introduit des modifications importantes par rapport à l'original.

Si l'expansion des segments peut être signe d'une traduction adaptative, qu'en est-il de leur suppression ? Trois traductions omettent le segment 1 (« Inde »), Bellegarde1701, Banier1732 et Rachmuhl2003, trois également ne traduisent pas le segment 3 (« croceo velatus amictu »), Vêrard1493, Morin1532, Corneille1687. La moitié de ces textes (Corneille1687, Banier1732 et Rachmuhl2003) a été identifiée comme étant des traductions adaptatives au paragraphe précédent. La structure des trois textes restants (Vêrard1493, Morin1532 et Bellegarde1701) est particulière.

Les traductions de Vêrard1493 (D.d.) et de Morin1532 (D.e.) qui font partie du cycle d'*Ovide moralisé* sont identiques, mises à part deux différences de graphies (« vola » / « volla » et « Erudice » / « Eurydice »).

D. Ovide, *Métamorphoses* : texte latin *vs* le cycle d'*Ovide moralisé* ([segments correspondants]₁₋₆) :

- a. Ovide : « [Inde]₁ [per immensum]₂ [croceo velatus amictu]₃ / [æthera digreditur]₂ [ciconumque]₄ [Hymenæus]₅ [ad oras / tendit ;]₄ [et Orpheà nequicquam voce vocatur.]₆ »
- b. Legouais1301 : [Hymen diex de nocoïement]₅ [d'illuec]₁ [s'en part grant aleure par l'air]₂ [en janne couverture]₃ [en Ciconie]₄ [ou semons iere a nocés d'estrange marié ; Orpheus semons li avoit qui nouvele espouse devoit espouser gente et joenne et bele Euridice la damoisele.]₆
- c. York1470 : [Le dieu Hymen]₅ [fut a ses nopces et quant d'illec]₁ [se parti grant erre s'en ala par l'air]₂ [couvert de janne couverture]₃ [en Cyconie]₄ [ou semons estoit aux nopces de estrange maniere de par Orpheus qui nouvelle femme devoit espouser : c'estoit Erudice qui moult estoit belle, gente et jone.]₆

- d. Vérard1493 : [Et quant]₁ [Hymen]₅ [se partit, par l'air s']₂[en]₁ [vola grant erre]₂ [en Cyconie,]₄
[où semons estoit de par Orpheus qui devoit espouser Erudice.
- e. Morin1532 : [Et quant]₁ [Hymen]₅ [se partit, par l'air s']₂[en]₁ [volla grant erre]₂ [en Cyconie,]₄
[où semons estoit de par Orpheus qui devoit espouser Eurydice.]₆

Pour ces deux traductions, le segment 2 (« per immensum aethera digreditur ») est interrompu par le rappel du segment 1 (« Inde »), dont la première partie, « Et quant » est renforcée ensuite par le pronom « en » : « **Et quant** Hymen se partit, par l'air s'**en** vola grant erre ». Mais mise à part la particularité de cette construction et même si le texte source n'est pas traduit très fidèlement (le segment 6 notamment, « où semons estoit de par Orpheus qui devoit espouser Eurydice »), la traduction reste plutôt informative, notamment avec les segments 5 (« Hymenaeus » / « Hymen ») et 4 (« ciconumque ad oras tendit » / « en Ciconie ») qui sont très courts.

Suivant les propos de Goethe, les traductions adaptatives de la deuxième période (potentiellement : celles de Legouais1301, D.b., et de York1470, D.c.) succèdent à celles qui veulent rester les plus proches de l'original. Mais les contextes historique et éditorial interviennent ici pour orienter notre interprétation plutôt vers une stratégie traductionnelle conforme à la première période, et dont le but est la simplification. En effet, la version de Vérard1493 est un post-incunable, un livre imprimé qui conserve l'aspect des plaquettes gothiques, diffusé d'abord par Colard Mansion, un célèbre libraire de Bruges, puis, à Paris, par Antoine Vérard. La structuration du texte suit généralement le modèle bourgeois, et les exemplaires de luxe sont richement illustrés, mais l'ouvrage est quand même plus accessible que le manuscrit de Legouais1301, dédié à Jeanne de Bourgogne, et celui de York1470, commandité, selon toute probabilité, par Louis de Bruges. La thèse de la simplification qui facilite le premier contact de l'auteur avec le public semble donc plausible. C'est d'autant plus le cas pour la version de Morin1532 qui est, selon Maurizio Busca, « la première édition en petit format des *Métamorphoses* vernaculaires et le dernier avatar de l'*Ovide moralisé* jouissant d'une large diffusion »³⁸. Dans ce cas précis, notre visualisation n'est pas suffisante pour aboutir à une telle conclusion, mais elle permet d'identifier certaines tendances ou anomalies qui invitent à une observation détaillée, orientée par une question spécifique. Il est donc important de pouvoir accéder, à partir de cette visualisation schématique, au texte (et au contexte) de chaque extrait.

³⁸ Maurizio Busca, « La mise en recueil des Métamorphoses d'Ovide aux XVI^e et XVII^e siècles en France », *Pratiques et formes littéraires. Cahiers du Gadges*, n° 17, Institut d'Histoire des Représentations et des Idées dans les Modernités, décembre 2020, paragr. 8, en ligne, <<http://publications-prairial.fr/pratiques-et-formes-litteraires/index.php?id=199>>. Voir aussi : Stefania Cerrito, *Une relecture renaissante de l'Ovide moralisé en prose Le Grand Olympe des Histoires poétiques du prince de la poésie Ovide (Lyon, 1530-1532)*, Classiques Garnier, coll. « Ovidiana », 2021, en ligne, <DOI : 10.48611/isbn.978-2-406-11517-5.p.0173>.

La traduction de Bellegarde1701 (E.c.), quant à elle, est très rigoureuse. Mise à part la suppression du premier segment, la longueur des cinq segments restants est assez proportionnelle : elle se situe entre 13,14 % pour le segment 2 (18/137 caractères) et 32,12 % pour le segment 6 (44/137)³⁹.

E. Ovide, *Métamorphoses* : texte latin vs Martignac (trad.), 1697 vs Bellegarde (trad.), 1701 ([segments correspondants]₁₋₆) :

- a. Ovide : « [Inde]₁ [per immensum]₂ [croceo velatus amictu]₃ / [æthera digreditur]₂ [ciconumque]₄ [Hymenæus]₅ [ad oras / tendit ;]₄ [et Orpheà nequicquam voce vocatur.]₆ »
- b. Martignac1697 : [Le Dieu des nopces]₅ [vêtu de jaune]₃ [s'envola]₂ [d'Égypte]₁ [en Thrace]₄ [où la voix d'Orphée l'appelloit.]₆
- c. Bellegarde1701 : [Le Dieu du mariage]₅ [vêtu d'une robe couleur d'or]₃ [s'élevant en l'air,]₂ [prit le chemin de la Thrace,]₄ [attiré par les charmes de la voix d'Orphée]₆

L'ordre et la longueur des segments semblent se stabiliser progressivement depuis la traduction précédente, celle de Martignac1697 (E.b.), la plus courte, alors que tous les segments ont été traduits.

Il est intéressant d'observer la différence quant au lieu du départ d'Hyménée, que Martignac est le seul à situer en Égypte. Ce pays est effectivement bien présent dans l'histoire d'Iphis et Ianthé qui conclut le livre IX des *Métamorphoses*. La mère d'Iphis, Téléthuse, rêve de « la fille d'Inachus, entourée de son cortège sacré [...] ; elle avait autour d'elle l'aboyant Anubis, la sainte Bubastis, Apis [...] et cet Osiris qu'on n'a jamais fini de chercher »⁴⁰. C'est finalement grâce à la prière de Téléthuse à Isis⁴¹ qu'Iphis se transforme en homme et peut épouser Ianthé. Toutefois, le cadre du récit est bien placé en Crète : « Le bruit de [la métamorphose de Byblis en fontaine] eût peut-être rempli les cent villes de la Crète si récemment la Crète elle-même n'avait pas vu plus près se produire une autre merveille par la métamorphose d'Iphis. Sur le territoire de Phestus, voisin du royaume de Gnose, était né jadis Ligdus » (p. 312). Il est difficile de déterminer si la traduction de Martignac relève d'une confusion ou d'un choix analytique.

Jean-Pierre Néraudau voit dans la présence des dieux égyptiens une preuve « de la vogue de la déesse [Isis] à l'époque d'Auguste », qui était assimilée tantôt à Io, tantôt à Cérès. Il est donc probable que des temples d'Isis soient construits en Crète. Le chercheur ne met par ailleurs aucunement en doute le fait que le mariage ait eu lieu en Crète⁴². Cependant, « De là » qui ouvre le livre X ne se réfère pas à un pays, mais au mariage d'Iphis et Ianthé, puisque le livre IX se termine par « Le

³⁹ Les proportions pour la version originale sont : 3,14 % pour le segment 1 (4/126), 28,57 % pour le segment 2 (36/126), 17,46 % pour le segment 3 (22/126), 14,29 % pour le segment 4 (18/126), 7,94 % pour le segment 5 (10/126) et 28,57 % pour le segment 6 (36/127).

⁴⁰ Ovide, *Les métamorphoses*, trad. Jean-Pierre Néraudau, Paris, Gallimard, coll. « Collection Folio Classique », 2006, p. 313. Pour les analyses critiques, nous utilisons cette traduction moderne, destinée au public des chercheurs, et qui n'appartient pas à notre corpus.

⁴¹ « Isis, dit-elle, toi qui chéris Parétonium, les champs Météoriques, Pharos et le Nil divisé en sept branches, viens à notre aide, je t'en supplie, porte remède à nos angoisses », *Ibid.*, p. 316.

⁴² Voir l'introduction du livre X, *Ibid.*, p. 318.

jour suivant avait de ses rayons illuminé toute l'étendue de l'univers, lorsque Vénus, Junon et Hyménée, leur acolyte, viennent se réunir auprès des torches nuptiales et le jeune Iphis possède sa chère Ianthé. » (p. 317). Deux traductions rétablissent cette chaîne référentielle, celle de York¹⁴⁷⁰ (« Le dieu Hymen fut à ses nopces ») et celle de Renouard¹⁶⁰⁶ (« le Dieu Nopcier se retira du festin qui se fit aux espousailles d'Iphis et d'Ianthé »).

En outre, des travaux récents, comme l'article de Rebecca Begum-Lees, « Que(e)r(y)ing Iphis' Transformation » semblent pouvoir expliquer le choix de Martignac :

[Il] est possible qu'Ovide pointe vers un endroit où Iphis et Ianthé peuvent se marier telles qu'elles sont : l'Égypte [...]. Diane Pintabone (2002, p. 277) mentionne rapidement qu'en remplaçant la déesse Latone dans la version de Nicandre [racontée par Antonius Liberalis (*Les Métamorphoses*, livre XVII)] par la déesse égyptienne Isis, Ovide aurait pu amener un public romain à s'attendre à ce que le conte se termine par le mariage de deux femmes, puisque, comme le démontre Bernadette Brooten (1996, p. 332-336), les mariages entre femmes, aussi officieux soient-ils, pouvaient, dans la pensée romaine, être associés à l'Égypte. [...] [E]n multipliant les références à ce pays, non seulement en remplaçant Latone par Isis, mais aussi en faisant référence aux dieux, instruments et paysages égyptiens, Ovide rappelle potentiellement au public un endroit où le désir d'Iphis peut être réalisé sans la solution radicale du changement de sexe.⁴³

Même s'il nous paraît hasardeux d'assigner des interprétations aussi modernes à une traduction qui date de 1697, il convient toutefois d'observer qu'elle constitue, au sein de notre échantillon, une transition entre la traduction en vers, très adaptative, de Corneille¹⁶⁸⁷ et une série de traductions plus modernes, se rapprochant progressivement de la structure et du sens du texte source. Or, rappelons que, selon Goethe, les traductions de la troisième période, visant le rapprochement optimal du texte source, sont initialement isolées, voire rejetées par leurs contemporains.

La proximité relative de la structure des traductions publiées entre 1697 et 1834 évolue à partir de la traduction de Nisard¹⁸⁶⁹. Outre l'adaptation de Rachmuhl²⁰⁰³, les cinq textes les plus modernes, recensées à partir de la seconde moitié du XIX^e s., suivent assez fidèlement la construction du texte latin : Nisard¹⁸⁶⁹ (F.b.), Parnajon¹⁸⁸⁰⁻¹ (en prose, F.c.) et Parnajon¹⁸⁸⁰⁻² (juxtalinéaire, F.d.), Béague¹⁹⁹⁸ (F.e.) et Boxus²⁰¹⁸ (F.f.), même si, comme la plupart des textes antérieurs, ils placent toujours le sujet « Hymenaeus » en première position, voire juste après le segment 1, « Inde ».

⁴³ « It is possible that Ovid gestures to a place in which Iphis and Ianthe can marry as they are: Egypt (for a fuller appreciation of the roles of Isis and Egypt in this passage, see Panoussi 2019). Diane Pintabone (2002:277) fleetingly suggests that by replacing the goddess Leto in Nicander's version with the Egyptian goddess Isis, Ovid could have led a Roman audience to expect the tale to 'end with two female spouses', since as Bernadette Brooten (1996:332-6) shows, female same-sex marriages – however unofficial – may have been associated in Roman thought with Egypt. [...] But by foregrounding Egypt, not just by replacing Leto with Isis but also by referring to Egyptian gods, the Egyptian rattle and Egyptian landmarks, Ovid potentially reminds the audience of a place where Iphis' desire can be realised without the radical solution of sex-change » [nous traduisons], Rebecca Begum-Lees, « Que(e)r(y)ing Iphis' Transformation in Ovid's *Metamorphoses* », dans Allison Surtees et Jennifer Dyer (dir.), *Exploring Gender Diversity in the Ancient World*, Edinburgh University Press, 2020, p. 110, en ligne, <<https://www.cambridge.org/core/books/exploring-gender-diversity-in-the-ancient-world/queerying-iphis-transformation-in-ovid-s-metamorphoses/6DEDAFEB8F7236AF2691C2FD25C3CF38>>.

- F. Ovide, *Métamorphoses* : texte latin *vs* Nisard (trad.), 1869 *vs* Parnajon (trad.), 1880 en prose *vs* Parnajon (trad.), 1880 (juxtalinéaire) *vs* Béague *et al.* (trad.), 1998 *vs* Boxus (trad.), 2018 ([segments correspondants]₁₋₆) :
- a. Ovide : « [Inde]₁ [per immensum]₂ [croceo velatus amictu]₃ / [æthera digreditur]₂ [ciconumque]₄ [Hymenæus]₅ [ad oras / tendit ;]₄ [et Orpheà nequicquam voce vocatur.]₆ »
 - b. Nisard1869 : [De là,]₁ [par les champs de l'espace,]₂ [Hyménée,]₅ [couvert de tissus éclatants,]₃ [s'élance vers les rives de l'Hèbre.]₄ [Il vient : Orphée l'appelle, mais il l'appelle en vain.]₆
 - c. Parnajon1880-1 : [De là]₁ [Hyménée,]₅ [vêtu d'une robe couleur de safran,]₃ [s'éloigne à travers les plaines immenses de l'air,]₂ [et se dirige vers les rivages des Ciconiens]₄ [où l'appelle en vain la voix d'Orphée.]₆
 - d. Parnajon1880-2 : [De là]₁ [Hyménée,]₅ [voilé d'un manteau couleur de safran,]₃ [s'éloigne à travers l'air immense,]₂ [et il se dirige vers les rivages des Ciconiens,]₄ [et il est appelé en vain par la voix d'Orphée.]₆
 - e. Béague1998 : [De là,]₁ [à travers l'immensité de l'éther,]₂ [couvert de son vêtement safran,]₃ [Hyménée]₅ [s'éloigne,]₂ [se dirige vers le territoire des Ciconiens,]₄ [et la voix d'Orphée l'y appelle en vain.]₆
 - f. Boxus2018 : [S'éloignant]₂ [de là,]₁ [enveloppé d'un voile couleur de safran,]₃ [Hyménée]₅ [traverse l'immense éther]₂ [et gagne les rives des Cicones,]₄ [tandis que l'appelle, bien inutilement, la voix d'Orphée.]₆

Seulement les constructions de la traduction de Béague1998 et de celle de Boxus2018 sont différentes. Pour cette dernière, la phrase est introduite par le participe présent « S'éloignant de là ». Cette solution permet de conserver par la suite la structure particulière de la phrase latine, en plaçant le segment 3 en incise.

Même si les visualisations schématiques sont, selon nous, indispensables pour éprouver un corpus de taille relativement conséquente, le retour (voire l'aller-retour) au texte est nécessaire. En outre, alors que notre exploitation des premières phrases de la traduction d'Ovide à l'aide d'une visualisation schématique a été stimulante, imaginons la complexité et la difficulté d'un alignement semblable appliqué à la comparaison d'extraits plus étendus. Par ailleurs, malgré quelques adaptations, suppressions et expansions, les segments présents dans les traductions sont uniquement ceux qui apparaissent dans le texte original d'Ovide. En revanche, dans les réécritures, le texte est non seulement adapté, mais il peut aussi être fortement remanié, modernisé, parodié, etc. Tout ceci rend la schématisation des correspondances difficile, sauf à s'imposer des cadres formels, et à identifier des phénomènes susceptibles d'apparaître dans toutes (ou au moins la plupart) des relations intertextuelles.

Jusqu'alors nous nous sommes intéressée principalement aux enjeux littéraires et avons proposé un travail relativement fin des propriétés et relations traductionnels. Cela nous a permis d'acquérir une connaissance approfondie des 21 phrases au centre de notre expérimentation. À présent, pour chercher à fournir ce cadre formel d'identification des phénomènes linguistiques qui participent à

l'établissement des relations intertextuelles, nous regardons les différentes traductions comme s'il s'agissait des paraphrases (et négligeant donc l'aspect traductionnel). Pour ce faire, nous commençons par observer les différents types de paraphrases délimités par les linguistes au sein des classements plus ou moins exhaustifs qui peuvent, selon nous, fournir le cadre dont nous avons besoin.

b) Classement des paraphrases : phénomènes linguistiques impliqués

En 1982, Catherine Fuchs regrette que les travaux des linguistes sur la paraphrase s'intéressent très peu au co-texte et déplore que les analyses se limitent souvent aux relations inter-phrastiques et à la comparaison de phrases isolées de leur contexte⁴⁴. La prise en compte du texte dans son unité, proclamée par la linguistique textuelle, découle directement de ce constat en souhaitant se distinguer, d'après Eugenio Coseriu, de la « grammaire transphrastique »⁴⁵. En effet, les recherches sur la paraphrase antérieures à celles de Fuchs approchent ce phénomène principalement du point de vue de la syntaxe. C'est notamment le cas des courants transformationnel⁴⁶ et générativiste⁴⁷ qui ne prennent pas en compte les phénomènes lexicaux et se limitent à considérer la paraphrase dans les limites de la grammaire. À partir de 1968, avec l'apparition de la sémantique générative⁴⁸, le cadre syntaxique est dépassé au profit de l'approche sémantique. Ces objectifs sont plus orientés vers le recensement et le classement de différentes formes de paraphrases et vers une analyse approfondie des relations sémantiques qui sont communes à ces phénomènes. Dans sa théorie Sens-Texte, Igor Melčuk⁴⁹ considère le langage comme un système de paraphrasage dont il définit 67 fonctions lexicales. Il ne s'agit pas, pour le chercheur, d'observer ce phénomène, mais de considérer les paraphrases comme un cadre théorique pour construire des modèles universels des

⁴⁴ Catherine Fuchs, « La paraphrase entre la langue et le discours », *Langue française*, vol. 53, n° 1, 1982, p. 22-33, en ligne, <https://www.persee.fr/doc/lfr_0023-8368_1982_num_53_1_5113>.

⁴⁵ Eugenio Coseriu, *Textlinguistik : Eine Einführung*, op. cit.

⁴⁶ Voir notamment Zellig S. Harris, « Co-Occurrence and Transformation in Linguistic Structure », *Language*, vol. 33, n° 3, Linguistic Society of America, 1957, p. 283-340, en ligne, <<https://www.jstor.org/stable/411155>> ; Zellig S. Harris, « Introduction to Transformations », dans Zellig S. Harris (dir.), *Papers in Structural and Transformational Linguistics*, Dordrecht, Springer Netherlands, coll. « Formal Linguistics Series », 1970 [1956], p. 383-389, en ligne, <https://doi.org/10.1007/978-94-017-6059-1_22>.

⁴⁷ Voir notamment Noam Chomsky, *Structures syntaxiques*, op. cit.

⁴⁸ James McCawley, « D. 1968. The Role of Semantics in a Grammar », dans Emmon Bach et R. Harms (dir.), *Universals in Linguistic Theory*, Holt, Rinehart, and Winston, 1968, p. 125-170.

⁴⁹ Igor Melčuk, *Un modèle linguistique fonctionnel : le modèle Sens-Texte*, p. 12 ; Igor Melčuk, « Paraphrase et lexique : la théorie Sens-Texte et le Dictionnaire explicatif et combinatoire », *Lexique*, n° 6, 1992, p. 9-58, en ligne, <http://olst.ling.umontreal.ca/pdf/Mel%27%C4%8Duk_1992_Paraphrase_DEC-3.pdf>.

langues. Des classements plus récents, bien moins exhaustifs que celui de Melčuk, s'efforcent davantage de prendre en compte le contexte, ou du moins de proposer une description linguistique appuyée sur des exemples concrets plutôt que sur une classification universalisante⁵⁰.

Pour établir un classement des phénomènes impliqués dans la paraphrase, il est nécessaire de disposer de données qui en sont évocatrices, ou au moins indicatrices. Plusieurs critères de jugement existent, allant de l'intuition du linguiste à l'établissement de règles de mise en relation. Nous privilégions, dans ce chapitre, l'étude d'un corpus monolingues (quoiqu'incluant des variations diachroniques) et parallèle, car il s'agit de différentes traductions d'un même ouvrage. Une recherche comparable a été menée par Regina Barzilay et Kathleen McKeown à l'aide de différentes traductions en anglais de *Madame Bovary* de Gustave Flaubert⁵¹ ou encore par Ali Ibrahim, Boris Katz et Jimmy Lin qui ont proposé d'étudier des couples de phrases équivalentes à l'aide d'arbres syntaxiques⁵². En effet, l'analyse d'un corpus parallèle de traductions garantit la prise en compte tant du co-texte que de l'architexte, c'est-à-dire le texte source (écrit dans notre cas en latin), et facilite l'alignement objectivable des segments correspondants.

Par exemple, selon la classification des paraphrases de Vila *et al.*⁵³, les noms propres et leurs périphrases (« Hyménée » / « le Dieu du mariage »), ou les reformulations qui touchent à la chaîne référentielle (« de là » / « de la Crète ») peuvent être définis comme, « des types particuliers de paraphrases qui dépassent une simple similarité sémantique pour tomber dans le champ de la pragmatique »⁵⁴. Pour ces relations, deux types de connaissances peuvent être mobilisés : La première est la connaissance encyclopédique, nécessaire pour pouvoir préciser la fonction d'Hyménée ou localiser le peuple des Ciconies en Thrace, à l'embouchure de l'Hèbre. La deuxième est appelée connaissance situationnelle, c'est elle qui permet de rétablir la chaîne référentielle avec la dernière histoire du livre IX des *Métamorphoses*, celle d'Iphis et Ianthé.

Alors qu'il est parfois discutable de considérer ce type d'équivalences comme paraphrases (on comparera « la Crète » et « les champs de Crète », mais non « la Crète » et « les espousailles d'Iphis

⁵⁰ Voir par exemple Marta Vila, M. Antònia Martí et Horacio Rodríguez, « Paraphrase Concept and Typology. A Linguistically Based and Computationally Oriented Approach », *loc. cit.* ; Rahul Bhagat et Eduard Hovy, « What Is a Paraphrase? », *Computational Linguistics*, vol. 39, n° 3, 2013, p. 463-472, en ligne, <https://doi.org/10.1162/COLL_a_00166>.

⁵¹ Regina Barzilay et Kathleen R. McKeown, « Extracting Paraphrases from a Parallel Corpus », *Proceedings of the 39th Annual Meeting of the ACL*, Toulouse, Association for Computational Linguistics, 2001, p. 50-57, en ligne, <<https://www.aclweb.org/anthology/P01-1008>>.

⁵² Ali Ibrahim, Boris Katz et Jimmy Lin, « Extracting Structural Paraphrases from Aligned Monolingual Corpora », *Proceedings of the Second International Workshop on Paraphrasing*, Sapporo, Japon, Association for Computational Linguistics, 2003, p. 57-64, en ligne, <<https://www.aclweb.org/anthology/W03-1608/>>.

⁵³ Marta Vila, M. Antònia Martí et Horacio Rodríguez, « Is This a Paraphrase? What Kind? Paraphrase Boundaries and Typology », *Open Journal of Modern Linguistics*, vol. 4, n° 1, Scientific Research Publishing, 2014, p. 205-218, en ligne, <<http://www.scirp.org/journal/PaperInformation.aspx?PaperID=43894&#abstract>>.

⁵⁴ « Examples like the ones in (6) to (10) are treated by several authors, both in linguistics and computational linguistics, as special types of paraphrases that go beyond pure semantic similarity to fall within the field of pragmatics » [nous traduisons], *Ibid.*, p. 211.

et d'Ianthe »), d'autres relations observables au sein de notre échantillon sont plus aisément reconnues comme telles. Pour décrire les paraphrases, Vila *et al.* différencient trois niveaux de granularité : les classes, les sous-classes et les types. Les quatre classes principales spécifient le mécanisme linguistique déclencheur ; elles sont divisées en sous-classes qui suivent l'organisation de la linguistique formelle :

- (i) modifications morpho-lexicales :
 - a. modifications morphologiques (morph. en Tableau 16), par exemple le changement de flexion (*époux* / *épouse*) ou la dérivation (*époux* / *épouser*),
 - b. modifications lexicales (lex.), comme la synonymie (*époux* / *mari*) ;
- (ii) modifications structurales :
 - a. modifications syntaxiques (syntax.), par exemple les modifications de coordination et de subordination (« Hymen quitte la Crète [...] et volant vers la Thrace, [...] » / « Hymen, [...] quittant la Crète, [...] se rendit dans la Thrace »),
 - b. modifications discursives (prag.), comme la modification de ponctuation (« De là, [...] Hyménée » / « De là Hyménée, ») ;
- (iii) modifications sémantiques (séman.) qui conservent le sens de l'énoncé en le formulant de manière différente (« où, d'une voix impuissante, Orphée invoque son appui » / « tandis que l'appelle, bien inutilement, la voix d'Orphée ») ;
- (iv) modifications pragmatiques (prag.) qui renvoient aux connaissances encyclopédiques ou référentielles du lecteur (« il se dirige vers les rivages des Ciconiens » / « [il] dirige son vol vers la Thrace »).

Deux autres classes, difficiles à assigner à un mécanisme linguistique particulier, s'ajoutent à la typologie :

- (v) modifications diverses (misc) :
 - a. modification du format (« 7,5 » / « sept et demie »),
 - b. changement d'ordre des mots (« l'appelle en vain la voix d'Orphée » / « la voix d'Orphée l'appelle en vain »),
 - c. la suppression ;
- (vi) modifications externes (ext.) :
 - a. formulations identiques (« De là » dans sept traductions),
 - b. implicature⁵⁵ (« appeler en vain » / « implorer »),

⁵⁵ Le terme anglais, *entailment*, devrait être traduit par *implication*, mais le manque du contexte dans l'exemple fourni par Vila *et al.* (« It **was acquiring** the “intellectual property and technology assets” of GeCAD » / « It **intends to acquire** the intellectual property and technology assets of Romanian antivirus firm GeCAD Software Srl ») ne nous permet pas de déterminer précisément s'il s'agit d'une *implication* ou d'une *présupposition*, voire d'une *implicature* (voir H. Paul Grice,

- c. non-paraphrases (« où l'appelloit la voix d'Orphée pour assister à son mariage » / « où il était attiré par le charme de la voix d'Orphée qui le prioit de rendre son mariage heureux »).

Nous privilégions ce classement pour plusieurs raisons. Tout abord, il est relativement récent tout en faisant l'objet d'une réflexion à long terme : le classement sur lequel nous nous appuyons est présenté en 2014 dans l'article « Is this a paraphrase? What kind? Paraphrase boundaries and typology »⁵⁶, qui apporte des corrections à celui initialement conçu, publié dans l'article de 2011, « Paraphrase concept and typology. A linguistically based and computationally oriented approach »⁵⁷ et prolonge le travail de thèse de Marta Vila, *Paraphrase scope and typology. A data-driven approach from computational linguistics*, soutenue en 2013⁵⁸. Par ailleurs, l'établissement de la typologie s'appuie sur un état d'art critique des travaux linguistiques et est appliquée sur un corpus annoté. Enfin, le classement étant conçu et expérimenté dans un contexte de traitement automatique (détection automatique du plagiat), la graduation des phénomènes observés (en classes, sous-classes, puis types) facilite la schématisation et s'adapte assez aisément aux analyses massives, prenant en compte simultanément plusieurs extraits⁵⁹. Observons à présent les résultats de son application, toujours en nous appuyant sur les 21 traductions des *Métamorphoses* d'Ovide.

Les auteurs du classement concèdent qu'une paraphrase peut être concernée simultanément par plusieurs types. En outre, en fonction de la taille des segments pris en compte, l'appréciation de leur similarité sera différente. Les subordonnées « où l'appelloit la voix d'Orphée pour assister à son mariage » de Duryer1702 et « où il était attiré par le charme de la voix d'Orphée qui le prioit de rendre son mariage heureux » de Banier1732 n'ont pas le même sens. Mais, au niveau des unités lexicales, nous trouvons trois noms en commun : « voix », « Orphée » et « mariage ». De plus, « appeler » et « attirer » partagent un sens commun (faire venir quelqu'un quelque part⁶⁰), et pourraient

« Logic and conversation », dans John P. Kimball, Jerry L. Morgan et Peter Cole (dir.), *Syntax and semantics*, New York San Francisco London, Academic press, Harcourt Brace Jovanovich, 1975, p. 41-58). En linguistique pragmatique, une *implicature* est ce qui est signifié, mais pas dit explicitement, une *implication* est une relation conditionnelle (A implique B) et une *présupposition* est une information qui n'est pas dite, mais considérée comme connue par l'interlocuteur. Pour couvrir les phénomènes spécifiques présentes dans notre corpus, nous incluons dans cette sous-classe toutes les relations allusives, comme les métaphorisations.

⁵⁶ Marta Vila, M. Antònia Martí et Horacio Rodríguez, « Is This a Paraphrase? », *loc. cit.*

⁵⁷ Marta Vila, M. Antònia Martí et Horacio Rodríguez, « Paraphrase Concept and Typology. A Linguistically Based and Computationally Oriented Approach », *loc. cit.*

⁵⁸ Marta Vila Rigat, « Paraphrase Scope and Typology. A Data-Driven Approach from Computational Linguistics / Abast i tipologia de la paràfrasi. Una aproximació empírica des de la lingüística computacional », Thèse de doctorat, Universitat de Barcelona, 2013.

⁵⁹ Nous remercions Natalia Grabar dont le séminaire « Reformulations et paraphrases » (M1/M2 Sciences du langage, Université de Lille, 2017) nous a été d'une aide inestimable pour déterminer le classement susceptible de s'adapter le mieux à nos besoins spécifiques.

⁶⁰ Voir Ponglet « Synonymes » du *Dico en ligne* (<https://dictionnaire.lerobert.com/>), repris du *Dictionnaire des synonymes et nuances* (Le Robert, 2019), *s.v.* **appeler I**, 3^e ● « **faire venir**, convier, convoquer, demander, inviter », et *s.v.* **attirer I**, 2^e ● « **faire venir**, amener, [en masse] faire affluer, faire courir ».

donc être considérés, dans notre exemple, comme des substitutions conversives, un type appartenant à la classe des modifications morpho-lexicales (sous-classe : paraphrase lexicale).

Néanmoins, pour ces deux textes du XVIII^e s. les équivalences dans la langue moderne, attestées par un dictionnaire de la langue française contemporain, ne sont pas forcément pertinentes. En effet, pour juger de la relation effective entre *appeler* et *attirer* à l'époque où les traductions ont été publiées, il nous faut nous tourner vers des ressources lexicographiques et des dictionnaires de l'époque. Le premier dictionnaire français des synonymes, *La justesse de la langue françoise ou Les différentes définitions des mots qui passent pour des synonymes* de l'abbé Gabriel Girard, date de 1718. Il est donc postérieur au texte de Duryer1702, mais il reflète sans doute l'état de la langue contemporain, ou au moins proche, de celui du traducteur. Toutefois, l'approche lexicographique de l'époque « résulte d'un constat : la synonymie exacte n'existe pas, il faut alors pouvoir distinguer les nuances de sens existantes entre des termes synonymes »⁶¹. Ainsi, Girard propose des entrées rédigées qui comparent le sens et l'usage des mots sémantiquement proches. Par exemple, pour l'entrée « NOMMER. APPELER », nous lisons :

On Nomme pour distinguer dans le discours. On Appele pour faire venir dans le besoin.

Le Seigneur Appela tous les animaux, & les Nomma devant Adam.

Il ne faut pas toujourns Nommer les choses par leurs noms, ni Appeler toutes sortes de gens à son secours.⁶²

Outre la mention du sens « faire venir dans le besoin » d'*appeler*, cette définition n'est pas utile pour notre échantillon, aucun texte ne proposant de traduire « vocatur » par « nommer ». En revanche, la troisième édition du dictionnaire, parue en 1786, inclut l'entrée « APPELER. ÉVOQUER. INVOQUER. », qui permet de comparer la traduction de Duryer1702, « où l'*appeloit* la voix d'Orphée » avec celle de Gros1834, « où d'une voix impuissante Orphée *invoque* son appui ». En effet, selon Girard :

Nous *appelons* les hommes & les animaux qui vivent avec nous & autour de nous sur la terre. [...] Nous *invoquons* la Divinité, les Saints, les puissances célestes, & tout ce que nous regardons comme au-dessus de nous, soit par l'habitation dans les cieux, soit par la dignité & le pouvoir sur la terre.

On *appelle* simplement par le nom, ou en faisant signe de venir. [...] On *invoque* par les vœux et par la prière.⁶³

⁶¹ Alice Ferrara-Léturgie, « La transformation des dictionnaires de synonymes au XX^e siècle : de la synonymie distinctive à la synonymie cumulative », dans *La révolution du dictionnaire*, Hermann, 2014, p. 157-174.

⁶² Gabriel Girard, *La justesse de la langue françoise ou Les différentes significations des mots qui passent pour synonymes*, Paris, L. d'Houry, 1718, p. 142.

⁶³ Gabriel Girard, *Dictionnaire universel des synonymes de la langue française : contenant les synonymes de Girard et ceux de Beauzée, Ronbaud, d'Alembert, Diderot et autres écrivains célèbres*, Paris, France, E. Penaud, 1846, p. 57-58 (t. 1). Cette édition paraît 100 ans après la mort de Girard. Elle est enrichie de quelques entrées et les distinctions initiales de Girard sont parfois relativisées par ses postérieurs (voir par exemple s. v. « NOMMER. APPELER. », p. 128-129 du t. 2). Le choix de cette version permet, potentiellement, de comparer les relations lexicales au sein des cinq traductions publiées entre 1700 et 1850.

Curieusement, malgré cette différenciation précise du destinataire humain pour *appeler* et divin pour *invoker*, Gros¹⁸³⁴ est bien le seul texte ayant recours à ce dernier verbe. En outre, l'invocation ne concerne pas tant Hyménée directement, mais « son appui ».

En revanche, aucune entrée n'est proposée pour les autres équivalents d'*appeler* dans notre échantillon : *attirer*, *implorer* et *semondre*. Dans ces cas de figure, la comparaison des définitions de ces mots dans des ressources contemporaines à chaque traduction qui les atteste serait une solution fiable, mais très chronophage. Pour le couple « appeler » (Duryer¹⁷⁰²) et « attirer » (Banier¹⁷³²), les trois premières éditions du *Dictionnaire de l'Académie Française* seraient à prendre en compte (1694, 1718 et 1740). Nous constatons que la définition *s.v.* **attirer** n'a presque pas évolué au cours de ces trois éditions, mise à part la modernisation de graphie et l'ajout d'une phrase d'exemple « Un abîme en attire un autre » :

Tirer à soi. Le soleil attire les vapeurs. L'aimant attire le fer. L'ambre attire la paille. Cet onguent a la vertu d'attirer les matières, a la vertu d'attirer. On dit qu'il y a des animaux & des plantes qui attirent le mauvais air.

On dit figur. Attirer quelqu'un à son parti. Attirer l'ennemi dans une embuscade. Attirer les yeux, les regards de tout le monde sur soi. Sa beauté, sa bonne mine, lui attire tous les cœurs. S'attirer l'affection, le mépris, l'estime, l'approbation de tout le monde. Cette action lui a attiré de grandes bénédictions, de grandes louanges. S'attirer de méchantes affaires. S'attirer une querelle sur les bras. S'attirer la haine du public. Ce crime a attiré de grands malheurs sur son auteur. Un malheur en attire un autre. Un abîme en attire un autre.⁶⁴

Nous retrouvons donc l'emploi figuré qui correspond au contexte de notre occurrence, en particulier avec la phrase d'exemple « Attirer quelqu'un à son parti », mais l'établissement du lien avec le sens correspondant pour *s. v.* **appeler**, « Se servir de la voix ou d'un autre signe pour faire venir quelqu'un » (1694, *s. v.* **appeller**, 7^e § ; 1718, 5^e *s.v.* **APPELLER**), devient véritablement aisé seulement à partir de l'édition du 1835 (6^e) au sein de laquelle la définition initiale, « tirer à soi », est complétée par « faire venir à soi ».

Un seul dictionnaire du XVIII^e siècle permet de faire un lien direct entre ces deux verbes. Dans le *Dictionnaire critique de la langue française* (1787)⁶⁵, nous lisons, *s.v.* **appeler**, 3^e : « se dit du cri des animaux, qui *atire* ceux de leur espèce. Voy. *Appeau*. Et de toutes les chûses, dont le son sert de signe pour faire qu'on se trouve à un endroit. « Les cloches *apèlent* à l'Église. ». Ce sens peut s'appliquer à notre contexte à condition que le mot *chûses* est employé par l'opposition aux animaux, et inclut donc les hommes ou si on considère que, comme l'homme active les cloches, la voix d'Orphée est l'agent qui sert à faire savoir qu'Hyménée est attendu quelque part.

⁶⁴ Voir *s.v.* **attirer**, 3^e édition, tome 1, disponible en ligne sur <https://www.dictionnaire-academie.fr/article/A3A1308>.

⁶⁵ Jean-François Féraud, *Dictionnaire critique de la langue française*, Marseille, Chez Jean Mossy Père et Fils, 1787.

Finalement, on pourrait considérer qu'*appeler* et *attirer* sont équivalents au niveau lexical au sein des traductions publiées à partir du début du XIX^e siècle⁶⁶ et présentent plutôt une relation sémantique au sein des trois textes antérieurs dans lesquels les occurrences d'*attirer* ont été relevées. Outre Banier1732 (« où il **étoit attiré** par le charme de la voix d'Orphée »), Renouard1606 utilise les deux (« où l'**attiroit** la voix charmeresse d'Orphée qui l'**appeloit** ») alors que Bellegarde1701, presque contemporain à Duryer1702, propose « **attiré** par les charmes de la voix d'Orphée ».

Dans le cadre des séries traductives, il faudrait compléter cette analyse par une étude comparative des équivalents proposés dans des dictionnaires de latin contemporains à chaque traducteur. Mais, comme notre démonstration s'intéresse principalement aux paraphrases et aux phénomènes potentiellement décelables au sein des réécritures, nous ne l'effectuons pas ici. Notons toutefois que, pour une telle étude comparative, que ce soient des dictionnaires de langue française ou ceux de latin, quelques ressources numériques s'avèrent précieuses. Il s'agit, d'une part, de Gallica qui souvent ne donne accès qu'aux PDF sans recherche lexicale et, d'autre part, notamment, de la nouvelle édition numérique du *Dictionnaire de l'Académie Française* qui permet de comparer l'évolution des articles au fil des neuf éditions (depuis 1694 jusqu'à nos jours)⁶⁷ et du portail *Dictionnaires d'autrefois*, porté par le projet ARTFL de l'Université de Chicago⁶⁸ qui inclut, outre les éditions 1, 4-6 et 8 du *Dictionnaire de l'Académie Française*, le *Thresor de la langue françoise* de Jean Nicot (1606), le *Dictionnaire critique de la langue française* de Jean-François Féraud (1787-1788) et Le *Dictionnaire de la langue française* d'Émile Littré (1872-1877). Une telle ressource comparative consacrée aux dictionnaires des synonymes permettrait sans aucun doute de porter un nouveau regard sur les paraphrases lexicales.

En définitive, si les ressources lexicographiques modernes permettent de considérer le couple « appeler » / « attirer » comme des paraphrases conversives, cette relation pour les textes du XVIII^e s. est moins directe, mais suffisante pour établir une relation sémantique entre un emploi figuratif d'*attirer* et un des sens d'*appeler*. Par ailleurs, nous pouvons affirmer que Duryer1702 (« où l'appelloit la voix d'Orphée pour assister à son mariage ») et Banier1732 (« où il étoit attiré par le charme de la voix d'Orphée, qui le prioit de rendre son mariage heureux ») s'éloignent du sens littéral du segment latin (« et Orpheà nequicquam voce vocatur ») en remplaçant tous les deux l'adverbe « en vain » par une précision concernant le mariage du poète. Par ailleurs, dans le deuxième extrait, les intentions d'Orphée sont présentées de manière beaucoup plus spécifique. En effet, il y a une grande différence d'intentionnalité entre l'action d'appeler quelqu'un pour assister

⁶⁶ La modification de la définition *s.v.* **attirer** dans la 6^e édition du *Dictionnaire de l'Académie Française* (1835) montre uniquement que l'emploi « faire venir » s'est stabilisé ou a été observé par les lexicographes à ce moment-là, sans pour autant indiquer à partir de quelle période il a commencé à être utilisé dans la langue courante.

⁶⁷ Voir <https://www.dictionnaire-academie.fr/lancement>.

⁶⁸ Voir <https://portail.atilf.fr/dictionnaires/>.

à son mariage et celle de l'attirer par le charme de sa voix et de le prier de rendre son mariage heureux.

Par ailleurs, les constructions syntaxiques des deux subordonnées sont très différentes. Le sujet de la première est le syntagme nominal « la voix d'Orphée ». L'infinitif « assister », introduit par la préposition « pour », est un complément de finalité de la phrase, et Hyménée est remplacé dans la subordonnée par le pronom du complément d'objet direct *le*. Dans la deuxième subordonnée, Hyménée devient le sujet (pronom personnel *il*), le verbe *attirer* est employé à la voix passive et le syntagme « voix d'Orphée » devient le complément d'agent. La proposition subordonnée relative « qui le prioit de rendre son mariage heureux » est le complément, cette fois-ci, du nom *voix*.

Dans leur classement, Vila *et al.* proposent cinq types des paraphrases syntaxiques :

1. Altération de la diathèse (« Orphée **l'appelle** » / « **Il est appelé par** Orphée ») ;
2. Commutation de la négation (« il **n'y** montra **pas** un visage riant » / « il **n'y** parut **point** avec un visage riant ») ;
3. Ellipse (« vêtu **d'une robe de** pourpre » / « vêtu **de** pourpre ») ;
4. Modification de coordination (« attiré par la voix d'Orphée ; il vint présider à son union avec Eurydice » / « attiré par les charmes de la voix d'Orphée : il se trouva **en effet** à son mariage ») ;
5. Changement de subordination et d'imbrication (« Hyménée [...] **s'éloigne à travers les plaines immenses de l'air** » / « **à travers l'immensité de l'éther**, [...] Hyménée **s'éloigne** »).

Toutefois, parmi la complexité des modifications syntaxiques observées pour le couple des “non-paraphrases” de Duryer1702 et Banier1732, seule la relation entre les subordonnées « où l'appelloit la voix d'Orphée » / « où il était attiré par [le charme de] la voix d'Orphée » semble pouvoir être assignée à un des types des paraphrases syntaxiques (altération de la diathèse, en plus d'une potentielle substitution conversive déjà abordée).

Même en travaillant sur des équivalences relativement proches, nous observons quelques problèmes liés à la complexité et l'hétérogénéité des phénomènes linguistiques à décrire. Ces difficultés seront encore plus grandes pour des relations plus allusives et incertaines. L'alignement des traductions parallèles pose relativement peu de difficultés concernant le jugement de leur pertinence, car le texte source sert de socle commun et de point de repère relativement stable. Nous constatons toutefois que, pour ce type de relations, plusieurs niveaux d'analyse se superposent. Ils ne touchent pas tous à la langue : les références encyclopédiques, culturelles et littéraires participent à l'évaluation et à la compréhension des relations qui peuvent lier deux extraits textuels. Par ailleurs, du point

de vue méthodologique, nous voyons déjà quelques problèmes concernant l'exploitation des classements préexistants, comme l'appariement lexical au sein d'un corpus diachronique ou la complexité des reformulations syntaxiques. Enfin, des allers-retours incessants semblent nécessaires entre le détail des relations entre les unités lexicales ou entre les syntagmes et la prise en compte d'un contexte textuel plus étendu (une phrase, un paragraphe...), voire la mise en perspective extratextuelle (paratextes, contexte éditorial, sociétal ou culturel...).

c) Classement des paraphrases : proposition d'une méthode

Pour mieux comprendre les enjeux linguistiques des relations intertextuelles et les problèmes qu'une telle recherche implique, observons un segment davantage restreint de la première phrase d'Ovide qui nous a intéressée jusqu'alors, en comparant simultanément les 21 traductions. Tout d'abord, une telle analyse nous permettra de confronter la pertinence d'un classement établi *a priori* à un cas concret d'exploitation d'un corpus littéraire et de recenser le plus de difficultés possible. Elle nous amènera également à nous interroger sur les méthodes de comparaison simultanée des réutilisations textuelles entre plusieurs extraits similaires et sur la faisabilité d'analyses linguistiques très détaillées portant sur la totalité de notre corpus. Enfin, nous essayerons aussi d'identifier, lors de cette investigation, les facteurs linguistiques pertinents pour déterminer si des extraits entretiennent des relations intertextuelles.

Pour ce faire, nous proposons d'expérimenter une méthode particulière pour décrire 21 traductions de différents segments de la première phrase d'Ovide. Parmi 24 types de paraphrases différenciées au sein du classement Vila *et al.*, nous en retenons, pour l'instant, 18. Nous écartons les paraphrases discursives⁶⁹, observables à partir de segments plus étendus que le syntagme, tout comme l'implicature (paraphrase externe) et le changement du format alphanumérique (modifications diverses). La typologie étant conçue pour décrire des couples de relations, nous l'adaptions à la comparaison de segments multiples en proposant une grille de notation qui combine la grille sémique, la méthode permettant de faire apparaître l'ensemble de traits distinctifs qui caractérisent un mot, et un arbre de décision (Tableau 16).

⁶⁹ Elles sont quatre : (i) changement de la ponctuation, (ii) style direct/indirect, (iii) changement de la modalité de la phrase, (iv) changement de la structure syntaxique ou discursive. Voir Marta Vila, M. Antònia Martí et Horacio Rodríguez, « Is This a Paraphrase? », *loc. cit.*, p. 212.

Tableau 16. Typage des paraphrases : traits et parcours

Type de paraphrase	Si Oui, passer à	Si Non, passer à
1. Changement de graphie (lex.)	2	2
2. Changement de flexion (morph.)	9	3
3. Changement de dérivation (morph.)	9	Si 1=Oui→9 ; Si 1=Non→4
4. Suppression (misc)	Fin	5
5. Substitution à polarité identique (lex.)	9	6
6. Substitution à polarité opposée (lex.)	9	7
7. Substitution synthétique ou analytique (lex.)	9	8
8. Substitution converse (lex.)	9	9
9. Ellipse (syntax.)	10	10
10. Commutation de la négation (syntax.)	11	11
11. Modification de coordination (syntax.)	12	12
12. Changement de subordination et d'imbrication (syntax.)	13	13
13. Changement d'ordre des mots (misc)	Si {9 10 11 12 13}= Oui → 15	14
14. Formulation identique (ext.)	Fin	15
15. Paraphrase sémantique (séman.)	Fin	16
16. Paraphrase pragmatique (prag.)	Fin	17
17. Non-paraphrase (ext.)	Fin	Fin

Tout d'abord, même si plusieurs types de paraphrases peuvent s'appliquer à un même segment, certains d'entre eux ne sont pas compatibles. Par exemple, la traduction la plus ancienne, *Ovide moralisé en vers* (1315-1325) traduit *Inde* par « d'illuec ». Une autre traduction en moyen français, *Ovide moralisé en prose II* (1470), propose une graphie différente : « d'illec ». Pour ce cas de figure (Tableau 17), après avoir noté qu'il s'agit d'un changement de graphie sans modification de flexion ni de dérivation, il est inutile de passer par les étapes 4-8 (à condition de statuer que deux lemmes identiques ne sont pas synonymes). Pour ce couple, les deux locutions adverbiales ont la même fonction (COI) et la même position dans la phrase (« d'illuec s'en part grant aleure » / « et quant d'illec se parti grant erre »). Mais, pour la première, la locution adverbiale est renforcée par la présence du pronom *en*. Cette construction particulière est difficile à classer parmi les cinq types des paraphrases syntaxiques, conçus selon la grammaire moderne, et de surcroît celle de la langue anglaise, mais nous pouvons considérer que la disparition du pronom dans la version de York relève d'une ellipse (étape 10). En revanche, si nous comparons cette première traduction avec une plus moderne, « de là », nous suivons le parcours jusqu'au point 5 (substitution à polarité identique, en percevant qu'il y a aussi une modification diachronique, absente du classement de Vila *et al.*), avant de passer aux modifications syntaxiques (à partir du type 9). Six traductions placent « De là » en tête de phrase où il joue la fonction conjonctive (« De là, Hyménée », etc.) et dans une, Boxus2018, la locution devient le COI du participe présent : « s'éloignant de là ». Pour ce dernier cas de figure, si nous

faisons abstraction de l'état de la langue, la modification par rapport à *Ovide moralisé en vers* est principalement lexicale.

Tableau 17. Grille de comparaison : « d'illuec » vs 8 autres traductions

« d'illuec [...] en » vs « d'illec » (1), « De là » (6), « de là » (1)			
Type	d'illec	De là	de là
1. Changement de graphie	Oui	Non	Non
2. Changement de flexion	Non	Non	Non
3. Dérivation	Non	Non	Non
4. Suppression	/	Non	Non
5. Substitution à polarité identique	/	Oui	Oui
6. Substitution à polarité opposée	/	/	/
7. Substitution synthétique ou analytique	/	/	/
8. Substitution converse	/	/	/
9. Ellipse	Oui	Oui	Oui
10. Commutation de la négation	Non	Non	Non
11. Modification de coordination	Non	Non	Non
12. Changement de subordination et d'imbrication	Non	Oui	Non
13. Changement d'ordre des mots	/	/	/
14. Formulation identique	/	/	/
15. Paraphrase sémantique	Non	Non	Non
16. Paraphrase pragmatique	Non	Non	Non
17. Non-paraphrase	Non	Non	Non

Cet exemple révèle le premier problème de l'adaptation de la classification de Vila *et al* à notre corpus. Pour un ensemble de textes aussi hétérogène du point de vue de l'état de la langue, il est impossible de ne pas noter les variations diachroniques. De ce point de vue, plusieurs facteurs sont à prendre en compte. Alors que, pour certains cas de figure, la modification se limite à la modernisation de graphie (« estrange » / « étrange », « appelloit » / « appelait », etc.), d'autres nous amènent à comparer un mot vieilli, qui n'est plus en usage dans la langue courante, et son équivalent moderne (« noçoïement » / « nocés »). En outre, l'évolution de la structure syntaxique, et surtout son instabilité dans les textes anciens, rend parfois difficile l'observation des relations paraphrastiques (« Orpheus semons li avoit qui novele espouse devoit espouser », « où semons estoit de par Orpheus qui devoit espouser Euridice »). Enfin, certains mots et locutions sont toujours présents en français moderne, mais leurs sens, voire leurs constructions syntaxiques dans le cas des verbes, ont évolué. Par exemple, dans le segment « et sans long intervalle droict au pays Thracien il devalle » (Habert1557), l'emploi intransitif du verbe *dévaler*, associé au complément de lieu « au pays Thracien », est surprenant. Certes, le verbe implique la rapidité de déplacement, mais celui d'Hyménée se fait principalement à l'horizontale (d'un pays à l'autre), même si le dieu se déplace dans les airs. Par ailleurs, la Thrace étant situé au nord de la Crète, il ne peut pas s'agir d'une descente figurée, qui s'effectue du nord vers le sud. En revanche, dans *Le Dictionnaire du Moyen Français (1330-1500)*, un des sens n'implique pas la direction du mouvement vers le bas et sied donc bien mieux au

contexte de la traduction d'Habert : « **I.** – Empl. intrans. ou pronom. **C.** – P. ext. “Effectuer un déplacement” »⁷⁰ et qui permet donc de valider *dévaler* comme verbe de déplacement en 1557.

D'autres difficultés peuvent être constatées en analysant les traductions d'un segment plus développé que « Inde ». Prenons comme exemple les 21 traductions du segment 3, « *croceo velatus amictu* », et comparons la traduction chronologiquement médiane, celle de Bellegarde1701 (« vêtu d'une robe couleur d'or ») avec les 10 textes qui la précèdent et les 10 textes qui la suivent. Chez Bellegarde1701, comme dans quatre autres textes, le segment remplit la fonction du complément du syntagme nominal « Le Dieu du mariage », le participe passé du verbe « vêtir » ayant ici la fonction adjectivale. 13 traductions placent la totalité du segment en apposition du nom d'Hyménée ou de sa périphrase, alors que trois textes ne le traduisent pas du tout. Enfin, Rachmuhl2003 développe le syntagme adjectival en une phrase indépendante : « Son manteau, couleur de safran comme un voile de mariée, flottait au vent ». L'observation de la totalité du segment dans son contexte et sa comparaison avec les autres traductions permettent de noter principalement deux modifications discursives par rapport à la version de Bellegarde1701 : (i) changement de la ponctuation et (ii) changement de la structure syntaxique ou discursive. Pour procéder à une description détaillée, suivant la grille de typage que nous proposons, il est pertinent de nous intéresser dans un premier temps séparément à trois composants du syntagme, traduisant chacun un mot latin, et leurs équivalents dans les autres traductions : le participe passé suivi de la préposition *de* (« vêtu de », qui traduit *amictu*), le nom précédé de son déterminant (« une robe » correspondant à *velatus*) et le groupe adjectival épithète du nom (« couleur d'or », *croceo* en latin).

Sept équivalents sont recensés pour le participe passé (Tableau 18), dont le plus fréquent est « couvert de » (7 occ.). Par ailleurs, « vêtu de » est attesté dans quatre autres traductions. Nous observons également des variantes de graphies tant dans les textes plus anciens, que dans ceux contemporains à celui de Bellegarde1701 : « vêtu de » (2 occ.) et « vestu de » (1 occ.). Enfin, deux traductions modernes se démarquent avec « voilé de » (Parnajon1880-1) et « enveloppé de » (Boxus2018).

⁷⁰ s.v. **dévaler**, *Dictionnaire du Moyen Français (1330-1500)*, en ligne, <<http://www.atilf.fr/dmf/definition/dévaler>>. Présenté en ligne avec des surlignements qui soutiennent la lecture structurelle des articles.

Tableau 18. Grille de comparaison : « vêtu de » (Bellegarde1701) vs 20 autres traductions.

« vêtu (de) » vs « couvert de » (7), « vêtu de » (4), Ø (4), « vêtu de » (2), « vestu de » (1), « voilé de » (1), « enveloppé de » (1)							
Type	couvert (de)	vêtu (de)	Ø	vêtu (de)	vestu (de)	voilé (de)	enveloppé (de)
1. Changement de graphie	Non	Non	Non	Oui	Oui	Non	Non
2. Changement de flexion	Non	Non	Non	Non	Non	Non	Non
3. Dérivation	Non	Non	Non	Non	Non	Non	Non
4. Suppression	Non	Non	Oui	Non	Non	Non	Non
5. Substitution à polarité identique	Oui	Non	/	Non	Non	Non	Non
6. Substitution à polarité opposée	/	Non	/	Non	Non	Non	Non
7. Substitution synthétique ou analytique	/	Non	/	Non	Non	Non	Non
8. Substitution converse	/	Non	/	Non	Non	Non	Oui
9. Ellipse	Non	Non	/	Non	Non	Non	Non
10. Commutation de la négation	Non	Non	/	Non	Non	Non	Non
11. Modification de coordination	Non	Non	/	Non	Non	Non	Non
12. Changement de subordination et d'imbrication	Non	Non	/	Non	Non	Non	Non
13. Changement d'ordre des mots	Non	Non	/	Non	Non	Non	Non
14. Formulation identique	Non	Oui	/	Oui	Oui	Non	Non
15. Paraphrase sémantique	Non	/	/	/	/	Oui	Non
16. Paraphrase pragmatique	Non	/	/	/	/	/	Non
17. Non-paraphrase	Non	/	/	/	/	/	Non

Les deux variantes graphiques ont la structure syntaxique identique à celle de Bellegarde1701 (participe passé suivi de la préposition *de*), ce qui nous amène à les considérer, ainsi que l'équivalent « vêtu de », comme des formulations identiques. Parmi les suppressions, nous recensons les trois omissions de la totalité du segment, mais aussi la suppression du participe dans la traduction de Rachmuhl2003 (« Son manteau, couleur de safran comme un voile de mariée, flottait au vent »).

En ce qui concerne les paraphrases lexicales, nous rencontrons ici le problème déjà relevé de l'évolution diachronique de la langue. Mais, dans le dictionnaire des synonymes de Girard, nous ne trouvons que deux entrées avec les équivalents recensés, « VÊTU. REVÊTU. AFFUBLÉ. » et « À COUVERT. À L'ABRI », et seule la première entrée est potentiellement pertinente pour nos analyses. Si nous suivons la distinction de Girard, *vêtu* « se dit des habits ordinaires, faits pour le besoin et la commodité », alors que *revêtu* « s'applique aux habillements établis pour distinguer dans l'ordre civil des emplois, les honneurs et les dignités »⁷¹. Or, aucune traduction ne propose *revêtir*.

⁷¹ Gabriel Girard, *Dictionnaire universel des synonymes de la langue française*, op. cit., p. 421 (t. 2).

Consciente des biais qu'introduit un tel traitement, nous décidons tout de même de nous tourner vers les dictionnaires synonymiques du français moderne. Faute de pouvoir comparer les équivalences en prenant en compte la perception du sens de chaque mot par les traducteurs, nous examinons la manière dont les relations lexicales sont potentiellement appréciées par le lecteur contemporain. Ce faisant, nous défavorisons les textes en moyen français, certains mots hors d'usage n'apparaissant pas dans les dictionnaires modernes (*semondre*, *épousailles*, etc.), mais le premier dictionnaire des synonymes étant publié au début du XVIII^e s., les textes antérieurs ne peuvent être comparés qu'à l'aide de dictionnaires de langue générale. En outre, pour ceux d'avant le XVII^e s., nous ne disposons que des outils qui reconstruisent les sens des mots *a posteriori*, comme le *Dictionnaire de l'ancienne langue française et de tous ses dialectes du IX^e au XV^e s.* de Frédéric Godefroy (1881-1902)⁷².

Ainsi, la consultation de l'onglet « Synonymes » du *Dico en ligne* du Robert nous permet de trouver *couvrir* parmi les synonymes de *vêtir*⁷³, c'est donc une substitution à polarité identique. En revanche, pour *envelopper*, il s'agit plutôt d'une substitution converse : sa relation avec *vêtir* est moins forte, mais attestée dans notre dictionnaire, puisque les deux verbes ont un synonyme commun (*couvrir*)⁷⁴. Alors qu'*envelopper* et *voiler* sont synonymes⁷⁵, aucune relation formelle n'est observée pour *vêtir* et *envelopper*, mais ils ont des traits en commun (entourer quelque chose ou quelqu'un d'une chose souple⁷⁶), ce qui nous incite à les considérer comme paraphrases sémantiques.

À ce point, la comparaison des traductions ne pose pas de problèmes particuliers grâce à leur proximité relative. Les différences diachroniques restent de simples modifications de graphie et la structure syntaxique des participes passé est identique dans toutes les attestations (sauf, bien sûr, pour les suppressions). En revanche, les relations au sein du syntagme nominal qui suit sont beaucoup plus complexes.

Tout d'abord, les équivalents sont bien plus nombreux (Tableau 19) : 11 cas de figure ont été recensés, dont six noms différents (« robe », « couverture », « manteau », « voile », « tissus » et « vêtement »). À cela s'ajoutent des modifications de graphie (« robbe », « Robbe »), des changements

⁷² Il convient de mentionner ici l'œuvre de Robert Estienne, *Dictionnaire françois-latin*, qui paraît en 1539 puis, en seconde édition augmentée, en 1549. Les premiers dictionnaires bilingues sont publiés à partir du XIV^e s., mais Estienne est un des premiers à proposer une nomenclature à partir des mots français et non pas, comme ses prédécesseurs, des mots latins. Voir Gérard Jacquin, « L'Œuvre de Robert Estienne et les dictionnaires de la Renaissance », *Revue 303 arts recherches créations. Patrimoine linguistique*, n° 120, Nantes, Association 303, 2012, p. 28-33. Cette ressource est exploitable pour la comparaison des équivalents traductionnels, mais difficilement applicable à l'analyse des paraphrases.

⁷³ *s.v.* **vêtir**, 1^{er} ● « **habiller**, **couvrir**, fringuer (*familier*), fagoter (*familier, péjoratif*), accoutrer (*péjoratif*), affubler (*péjoratif*) », voir l'onglet « Synonymes » du *Dico en ligne* (<https://dictionnaire.lerobert.com/>).

⁷⁴ *s.v.* **envelopper**, 3^{ème} ● « **couvrir**, entourer, recouvrir », *ibid.*

⁷⁵ *s.v.* **envelopper**, 5^{ème} ● « **cacher**, déguiser, dissimuler, emballer, enrober, farder, **voiler** » et *s.v.* **voiler**, 2^{ème} ● « **masquer**, **envelopper**, estomper », *ibid.*

⁷⁶ *s.v.* **vêtir**, « [Littéraire] Couvrir (qqn) de vêtements ; mettre un vêtement à (qqn) » et *s.v.* **envelopper**, « 1. Entourer d'une chose souple qui couvre de tous côtés », voir l'onglet « Définition » du *Dico en ligne* (*op. cit.*).

du déterminant (article indéfini, déterminant possessif ou article zéro), voire des modifications syntaxiques. Par exemple, « Son manteau » est le sujet d'une proposition chez Rachmuhl2003 alors que chez Massac1617 il fait partie, comme dans toutes les autres traductions, du groupe prépositionnel complément de l'adjectif.

La classification de Vila *et al.* n'inclut pas la modification du déterminant. Si nous souhaitons l'ajouter, parmi quel sous-type de paraphrases faudrait-il le classer ? Le changement de déterminant modifie l'appréciation du référent (« un manteau » / « son manteau »), ce qui concerne la sémantique et la pragmatique, mais son absence constitue une modification syntaxique (ellipse), comme c'est le cas, dans notre exemple, de deux équivalents, *couverture* et *tissus* (pour lequel l'emploi du pluriel nous conduit également à considérer qu'il s'agit d'un changement de flexion). Par ailleurs, pour la traduction particulière de Rachmuhl2003, nous observons aussi la modification de la coordination ainsi que le changement de subordination et d'imbrication (en plus des paraphrases discursives déjà constatées). Enfin, du point de vue lexical, aucune entrée du dictionnaire de Girard n'inclut *robe*⁷⁷, comme aucun synonyme n'est attesté dans le *Dico en ligne* parmi les équivalents traductionnels recensés⁷⁸. En revanche, compris dans le contexte religieux, ce nom a dans ce dernier dictionnaire deux synonymes communs avec *vêtement* (*aube* et *soutane*⁷⁹). Pour les autres noms, nous considérons que ces relations peuvent être perçues par les lecteurs contemporains comme paraphrases sémantiques.

⁷⁷ *Vêtement* est désambiguïsé au sein de l'entrée « VÊTEMENT. HABILLEMENT. HABIT » (p. 421, t. 2), le sens de *tissu* est précisé dans l'entrée « TISSU. TISSURE. TEXTURE. CONTEXTURE » (p. 387, t. 2).

⁷⁸ Mais *enveloppe* et *manteau* sont synonymes de « voile », voir *s.v.* **voile II** 1^{er} ● « **enveloppe, manteau** », *ibid.*

⁷⁹ *s.v.* **robe** 3^{ème} ● « [Religion] **soutane, aube, froc** » et *s.v.* **vêtement** 4^{ème} ● « [Religion] **aube**, chasuble, **soutane**, surplis ». Nous relevons également un synonyme commun avec *manteau*, *cape*, mais alors que, *s.v.* **manteau**, le sens évoqué est effectivement celui du champ vestimentaire, pour *robe*, il s'agit d'une équivalence spécifique entre une cape ou une robe d'un cigare.

Tableau 19. Grille de comparaison : « une robe » (Bellegarde) vs 20 autres traductions.

une robe vs « une robe » (5), ∅ (5), « couverture » (2), « Son manteau » (1), « son manteau » (1), « un manteau » (1), « un voile » (1), « une Robbe » (1), « sa robe » (1), « tissus » (1), « son vêtement » (1)											
Type	une robe	∅	couverture	Son manteau	son manteau	un manteau	un voile	une Robbe	sa robe	tissus	son vêtement
1. Changement de graphie	Oui	Non	Non	Non	Non	Non	Non	Oui	Oui	Non	Non
2. Changement de flexion	Non	Non	Non	Non	Non	Non	Non	Non	Non	Oui	Non
3. Dérivation	Non	Non	Non	Non	Non	Non	Non	Non	Non	Non	Non
4. Suppression	Non	Oui	Non	Non	Non	Non	Non	Non	/	Non	Non
5. Substitution à polarité identique	Non	/	Non	Non	Non	Non	Non	Non	/	Non	Non
6. Substitution à polarité opposée	Non	/	Non	Non	Non	Non	Non	Non	/	Non	Non
7. Substitution synthétique ou analytique	Non	/	Non	Non	Non	Non	Non	Non	/	Non	Non
8. Substitution converse	Non	/	Non	Non	Non	Non	Non	Non	/	Non	Oui
9. Ellipse	Non	/	Oui	Non	Non	Non	Non	Non	Non	Oui	Non
10. Commutation de la négation	Non	/	Non	Non	Non	Non	Non	Non	Non	Non	Non
11. Modification de coordination	Non	/	Non	Oui	Non	Non	Non	Non	Non	Non	Non
12. Changement de subordination et d'imbrication	Non	/	Non	Oui	Non	Non	Non	Non	Non	Non	Non
13. Changement d'ordre des mots	Non	/	Non	Non	Non	Non	Non	Non	Non	Non	Non
14. Formulation identique	Oui	/	/	Non	/	Non	Non	Oui	Non	Non	Non
15. Paraphrase sémantique	/	/	Oui	Oui	Oui	Oui	Oui	/	Non	Oui	Non
16. Paraphrase pragmatique	/	/	Non	Non	Non	Non	/	/	Non	/	Non
17. Non-paraphrase	/	/	Non	Non	Non	Non	/	/	Non	/	Non

Tableau 20. Grille de comparaison : « couleur d'or » (Bellegarde1701) vs 20 autres traductions.

couleur d'or vs Ø (3), « jaune » (2), « janne » (2), « Pourpre » (1), « de pourpre » (1), « de couleur safrancee » (1), « couleur de safran » (1), « safran » (1), « couleur de safran comme un voile de mariée » (1), « de jaune-doré » (1), « couleur de feu » (1), « éclatants » (1)												
Type	Ø	jaune	janne	Pourpre	de pourpre	de couleur safrancee	couleur de safran	safran	couleur de safran comme	de jaune-doré	couleur de feu	éclatants
1. Ch. de graphie	Non	Non	Non	Non	Non	Non	Non	Non	Non	Non	Non	Non
2. Ch. de flexion	Non	Non	Non	Non	Non	Non	Non	Non	Non	Non	Non	Oui
3. Dérivation	Non	Non	Non	Non	Non	Non	Non	Non	Non	Oui	Non	Non
4. Suppression	Oui	Non	Non	Non	Non	Non	Non	Non	Non	Non	Non	
5. Sub. à polarité identique	/	Non	Non	Non	Non	Oui	Oui	Non	Non	Non	Non	Non
6. Sub. à polarité opposée	/	Non	Non	Non	Non	/	/	Non	Non	Non	Non	Non
7. Sub. synthétique / analytique	/	Oui	Oui	Non	Non	/	/	Oui	Oui	Oui	Non	Non
8. Sub. converse	/	/	/	Non	Non	/	/	/	/	/	Non	Oui
9. Ellipse	/	Oui	Oui	Oui	Oui	Non	Non	Oui	Non	Oui	Non	Oui
10. Commut. de la négation	/	Non	Non	Non	Non	Non	Non	Non	Non	Non	Non	Non
11. Mod.de coordination	/	Non	Non	Non	Non	Non	Non	Non	Non	Non	Non	Non
12. Ch. de subordination/imbrication	/	Non	Non	Non	Non	Non	Non	Non	Oui	Non	Non	Oui
13. Ch. d'ordre des mots	/	Non	Non	Non	Non	Non	Non	Non	Non	Non	Non	Non
14. Form. identique	/	/	/	/	/	/	/	/	/	/	Non	/
15. Paraphr. sémantique	/	Non	Non	Non	Non	Non	Non	Non	Non	Non	Oui	Non
16. Paraphr. pragmatique	/	Non	Non	Oui	Oui	Non	Non	Non	Oui	Non	Non	Non
17. Non-paraphrase	/	Non	Non	Non	Non	Non	Non	Non	/	Non	Non	Non

La structure syntaxique des traductions du dernier composant du syntagme, « croceo » (Tableau 20), est relativement schématique. Dans sept textes, il s'agit de l'ajout d'un adjectif seul, avant le nom du vêtement (« janne couverture » chez Legouais1301 et York1470), après (« robe jaune » chez Renouard1606, « tissus éclatants » chez Nisard1869, « vêtement safran » chez Beague1998), voire même avec l'omission du nom (« vestu de Pourpre » chez Habert1557, « vêtu de jaune » chez Martignac1697). Dans quatre textes, la couleur est nominalisée et précédée par une préposition *de* pour former un complément du nom (Duryer1702 : « robe de jaune-doré », Fontanelle1789, Villenave1806 et Gros1834 : « robe de pourpre »). Six autres ont plutôt recours à un groupe prépositionnel « de couleur ADJ » (Massac1671 : « de couleur safrancee ») ou un groupe nominal « couleur de NOM » (Bellegarde1707 : « couleur d'or », Banier1732 : « couleur de feu », Parnajon1880-1,2 et Boxus2018 : « couleur de safran »). Enfin, la traduction de Rachmuhl2003 reste particulière car le syntagme nominal « couleur de safran » est apposé au nom du vêtement (« Son manteau ») et complété par la comparaison introduite par l'adverbe *comme* (« comme un voile de mariée »). En comparant la traduction de Bellegarde1701, « couleur d'or », aux 20 autres segments, nous notons donc, pour les paraphrases syntaxiques, sept cas d'ellipse (omission du nom *couleur* et /ou de la préposition *de*) et un changement d'imbri- cation pour la paraphrase descriptive de Rachmuhl2003 (un complément de nom / incise).

Pour ce segment, les modifications foncières se trouvent plutôt au niveau lexical, même si nous notons deux paraphrases morphologiques : un changement de flexion pour le pluriel « éclatants » (« tissus éclatants », Nisard1869) et un changement de dérivation entre « d'or » et « de jaune-doré » (Duryer1702). Les couleurs et leurs nuances semblent peu affectées par l'évolution diachronique de la langue. Par exemple, la phrase « qui est de couleur d'or, de citron, de safran » est présente dans la définition du *s. v.* **jaune** pour toutes les éditions du *Dictionnaire de l'Académie française* (1694-ac- tuelle). Alors qu'aucune couleur ne figure parmi les entrées du dictionnaire de Giraud, *safran*, *doré*, et *or* sont également proposés comme synonymes de *jaune* dans l'onglet « Synonymes » du *Dico en ligne*. Nous pourrions donc considérer que, pour la traduction « d'or » de Bellegarde1701, la substi- tution à polarité identique a lieu pour les équivalents « jaune » (et la graphie ancienne « janne »), « (de) safran », « safranée » et « jaune-doré » (en plus du changement de dérivation pour la deuxième partie du nom composé). Néanmoins, chez Bellegarde1701, la précision « couleur de » introduit une inclusion lexicale¹ et est donc plus spécifique que l'adjectif de couleur seul, ce qui implique que seulement les segments « de couleur safrancee » et « couleur de safran » peuvent être entièrement considérés comme des substitutions à polarité identique. Pour la traduction descriptive de

¹ *Inclusion lexicale* implique qu'un segment A est imbriqué dans un segment B, parfois hyperonyme du A (« pays de Thrace », « couleur d'or », etc.).

Rachmuhl2003 (« couleur de safran comme un voile de mariée »), il s'agit d'une substitution analytique, alors que « jaune », « safran » et « de jaune-dorée » sont plutôt des substitutions synthétiques. Par ailleurs, pour « éclatants », aucun lien n'est observé pour « couleur d'or » dans l'onglet « Synonymes » du *Dico en ligne* du Robert, mais *doré* et *éclatant* peuvent tout de même être liés par un synonyme commun, *brillant*². Outre le fait qu'aucun synonyme n'est fourni pour *s.v.* **or** dans le dictionnaire du Robert, nous constatons que, pour pouvoir apprécier pleinement la relation entre deux unités lexicales, il est préférable que leur catégorie grammaticale soit identique.

Pour la substitution « couleur de feu » (Banier1732), les ressources dictionnaires consultées ne révèlent pas de relation de paraphrase lexicale avec « couleur d'or ». Dans la première édition du *Dictionnaire de l'Académie* (1694), la définition *s.v.* **couleur de feu** est « Couleur rouge & fort vive, approchant de celle de feu ». Le rouge vif est bien présent sur les pistils de safran et la flamme comporte aussi les tons jaune et orange, ce qui nous permet de lier sémantiquement les deux couleurs malgré la différence des objets dont les noms ont servi à les désigner (autrement dit, les couleurs d'or, de safran et de feu sont comparables, mais aucun autre trait sémantique ne permet de lier l'or, le safran et le feu). Néanmoins, le choix traductif paraît ici influencé également par des références extralinguistiques. Dans *Mythologie grecque et romaine*, Pierre Commelin écrit *a propos* du dieu Hyménée :

On le représentait sous la figure d'un jeune homme blond couronné de fleurs, surtout de marjolaine, tenant de la main droite un flambeau, et de la gauche un voile de couleur jaune, cette couleur étant même, à Rome, particulièrement affectée aux noces. Ainsi, dans les mariages romains, le voile de la jeune épouse était d'un jaune éclatant. Parfois ce dieu, couronné de roses, porte un vêtement blanc et brodé de fleurs ; certains mythologues lui donnent un anneau d'or, un joug et des entraves aux pieds, allégorie rendue plus transparente encore par deux flambeaux qui n'ont qu'une même flamme et que l'on place dans ses mains ou auprès de lui.³

Si le voile jaune est donc un des attributs de ce dieu, tenu dans la main gauche, Ovide l'associe plutôt à son vêtement. Par ailleurs, le feu est bien présent dans les représentations, notamment par le flambeau qu'Hyménée tient dans sa main droite, mentionné dans la suite du livre X (« Fax quoque, quam tenuit, lacrimoso stridula fumo / usque fuit, nullosque invenit motibus ignes »⁴). Enfin, comme nous le lisons dans le *Dictionnaire des Antiquités romaines et grecques* d'Anthony Rich (1861), le voile nuptial des mariées romaines s'appelait *flammeum*, car « [i]l était d'un jaune foncé et brillant (Plin. *H. N.* XXI, 22), comme une flamme, particularité d'où lui vint son

² *s.v.* **doré**, 1^{ère} ● « **brillant**, ambré, cuivré, mordoré » et *s.v.* **éclatant** I, 1^{ère} ● « **brillant**, ardent, éblouissant, étincelant, flamboyant, rutilant, vif, voyant », *Le Dico*, *op. cit.*

³ Pierre Commelin, *Mythologie grecque et romaine*, Paris, Garnier frères, coll. « Classiques Garnier », 1960, p. 213-214.

⁴ « La torche même qu'il tient ne cesse de siffler en répandant une fumée qui provoque les larmes ; il a beau l'agiter, il n'en peut faire jaillir la flamme », voir Ovide, *op. cit.*, p. 320. Pour les analyses critiques, nous utilisons cette traduction moderne, destinée au public des chercheurs, et qui n'appartient pas à notre corpus.

nom, et de dimensions considérables, suffisantes pour couvrir toute la personne de la tête aux pieds »⁵.

Alors que ce choix de traduction de Banier¹⁷³² peut être aisément compris et argumenté, les trois traductions suivantes, de Fontanelle¹⁷⁸⁹, Villenave¹⁸⁰⁶ et Gros¹⁸³⁴ proposent une autre couleur, et traduisent tout l'extrait de la même manière : « vêtu d'une robe de pourpre »⁶. Plusieurs hypothèses peuvent expliquer cette adaptation, mais aucune affirmation définitive ne semble admise.

En effet, le fait que les trois textes qui se succèdent chronologiquement fassent le même choix d'adaptation peut impliquer que Fontanelle¹⁷⁸⁹ ait servi de version de référence à Villenave¹⁸⁰⁶ et, directement ou pas, à Gros¹⁸³⁴. Si dans l'introduction de sa traduction, Villenave cite son prédécesseur, ce n'est pas pour louer la qualité de son travail : « Fontanelle, avec de l'esprit et du talent, s'est trompé en voulant être trop servilement littéral, et il a paru oublier que le génie de la langue latine adoptoit des expressions et des tours qui deviendroient dans la langue française une importation dangereuse »⁷. Gros, quant à lui, renvoie très régulièrement à la traduction de Villenave, autant pour relever les fautes de son prédécesseur, que pour argumenter ses propres choix traductionnels. Tant pour Villenave¹⁸⁰⁶ que pour Gros¹⁸³⁴, les notes sont très riches. 16 notes pour ce premier et 23 pour le deuxième sont proposées au sein du livre X. Elles sont composées sur cinq pages pour les deux traductions et mélangent les notes explicatives, philologiques et intertextuelles. Par ailleurs, le nom de Villenave est mentionné dans quatre des 23 notes de Gros.

Les trois textes proposent le latin en miroir et présentent un appareil critique exhaustif, ce qui semble exclure l'hypothèse d'une erreur traductionnelle perpétuée d'une version à l'autre (il est d'ailleurs difficile de déterminer les causes possibles d'une confusion entre *croceo* et *purpura*). Dans leur appareil critique, Villenave et Gros se positionnent en tant que correcteurs des traductions précédentes (la mention « traduction nouvelle » est d'ailleurs présente dans les titres de ces œuvres). Le fait qu'aucune note ne précise ce choix peut impliquer que, pour les trois auteurs, il paraissait

⁵ Anthony Rich, *Dictionnaire des Antiquités romaines et grecques*, trad. Adolphe Chéruel, Paris, Firmin Didot Frères, 1861, p. 274.

⁶ Mise à part ce segment, les trois traductions sont relativement différentes :

- Fontanelle¹⁷⁸⁹ : « L'Hymen, vêtu d'une robe de pourpre, quittant la Crète, traversant les airs, se rendit dans la Thrace, attiré par la voix d'Orphée ; »
- Villenave¹⁸⁰⁶ : « L'Hymen, vêtu d'une robe de pourpre, s'élève des champs de Crète dans les airs, et vole vers la Thrace, attiré par la voix d'Orphée »
- Gros¹⁸³⁴ : « Des champs de la Crète l'Hymen, vêtu d'une robe de pourpre, s'élance dans les vastes plaines de l'air et dirige son vol vers la Thrace, où d'une voix impuissante, Orphée invoque son appui »

⁷ Mathieu-Guillaume-Thérèse Villenave, « Avertissement », *Les Métamorphoses d'Ovide, traduction nouvelle, avec le texte latin, suivie d'une analyse, de l'explication des fables, de notes géographiques, historiques, mythologiques et critiques*, vol. 1, Paris, F. Gay, Ch. Guestard, 1806, p. V, en ligne, <<https://gallica.bnf.fr/ark:/12148/bpt6k15251514>>.

évident. Cependant, en 1866, la traduction de Gros connaît une nouvelle édition, « refondue par M. Cabaret-Dupaty »⁸, qui propose deux modifications de la première phrase du livre X :

Des champs de la Crète l'Hymen, vêtu d'une robe ([de pourpre]^{Gros} | [d'or]^{Cabaret-Dupaty}),
s'élance dans les vastes plaines de l'air et dirige son vol vers la Thrace où Orphée in-
voque [inutilement]^{Cabaret-Dupaty} son appui.

La raison du changement de la couleur n'est pas commentée ni à l'aide d'une note ni au sein de la préface qui, écrite par Jean-Pierre Charpentier, est en réalité une notice sur la vie d'Ovide et ne propose pas d'introduction philologique de la nouvelle traduction. Par ailleurs, la couleur pourpre apparaît également dans une traduction antérieure, Habert1557 (« vestu de Pourpre »), ce qui rend l'analyse de cette adaptation d'autant plus difficile.

S'agit-il d'une appréciation particulière de la couleur de safran ? En effet, si, dans la colorimétrie, elle se rapproche d'un jaune orangé, les pétales de la fleur sont violacés. À partir de la 4^e édition du *Dictionnaire de l'Académie française* (1762), *safran* est défini comme « Plante bulbeuse qui fleurit au commencement de l'automne, et qui porte une fleur bleue mêlée de rouge et de purpurin », alors que les éditions précédentes définissent la couleur plutôt comme « bleuâtre en dehors, & jaunâtre dans le fond ». Cependant, toutes les éditions de ce dictionnaire insistent bien sur l'appartenance de la couleur de safran aux nuances du jaune, ce qui nous invite à considérer que le glissement vers le pourpre a aussi des motivations sémantico-pragmatiques.

En effet, depuis le Moyen Âge, la symbolique du jaune est plutôt négative et son rôle symbolique ne cesse de décroître. Comme le constate Michel Pastoureau,

Le déclin du jaune date du Moyen Âge qui en fait une couleur ambivalente. D'un côté, le mauvais jaune, celui de la bile amère et du soufre démoniaque : il est signe de mensonge, d'avarice, de félonie, parfois de maladie ou de folie. De l'autre, le bon jaune, celui de l'or, du miel et des blés mûrs : il est signe de pouvoir, de joie, d'abondance. Toutefois, à partir du XVI^e siècle, la place du jaune dans la culture matérielle ne cesse de reculer. La Réforme protestante, la Contre-Réforme catholique, plus tard les « valeurs bourgeoises » du XIX^e siècle le tiennent en peu d'estime.⁹

Des expressions familières comme « faire des contes jaunes », « faire voir son bec-jaune », ou encore « être peint en jaune » témoignent de cette appréciation négative à travers les époques. De même, dans les définitions de *safran*, des éditions 5 à 7 du *Dictionnaire de l'Académie française* (1798-1878), nous lisons : « Fam., *Être jaune comme du safran, avoir le teint jaune comme du safran, comme safran, Avoir la maladie ictérique, la jaunisse* ». Le manque de ce mot dans les traductions publiées entre 1732 et 1869 témoigne peut-être de l'empreinte de l'épidémie de l'ictère catarrhale qui sévit lors des guerres napoléoniennes dans la langue et la société de cette époque¹⁰.

⁸ Ovide, *Les Métamorphoses d'Ovide*, trad. J.-R.-T. Cabaret-Dupaty et Étienne Gros, 2^e éd., Paris, Garnier frères, coll. « Bibliothèque latine-française », 1866.

⁹ Michel Pastoureau, *Jaune : histoire d'une couleur*, Paris, Seuil, 2019, 4^e de couverture.

¹⁰ Voir par exemple Roberto Zaugg et Andrea Graf, « Guerres napoléoniennes, savoirs médicaux, anthropologie raciale », *Histoire, médecine et santé*, n° 10, 2016, p. 17-43, en ligne, <<https://journals.openedition.org/hms/1014>>.

D'un autre côté, la symbolique du pourpre est, elle aussi, ambiguë. C'est d'un manteau de pourpre, en plus d'une couronne d'épines, que Jésus-Christ a été affublé pendant sa Passion. En ancienne Rome, cette couleur a été portée exclusivement par les empereurs pour devenir ensuite un symbole de dignité, de pouvoir et de puissance (des rois, mais aussi des dignitaires et des cardinaux catholiques). En outre, dans *Les Épithètes* de Maurice de La Porte (1571), nous lisons « Entre tous les Pourpres la Tyrienne a emporté le bruit, car elle a une couleur de sang figé, et noirâtre, et a d'ailleurs un lustre fort excellent. C'est pourquoi Homère appelle le sang purpurin »¹¹.

Par ailleurs, nous pouvons nous demander si la modification symbolique concerne le voile de mariée ou l'habit d'Hyménée. Les traducteurs, qui soulignent à plusieurs reprises l'importance d'adapter l'œuvre d'Ovide à leur public¹², ont-ils voulu rapprocher la représentation d'Hyménée à celle d'un prêtre catholique ? Ou s'agit-il plutôt d'un présage de la fin tragique des noces, par le rapprochement avec la couleur du sang ? S'il paraît plausible d'expliquer le remplacement des couleurs safran et jaune par une couleur moins négativement connotée, le choix du pourpre, plutôt que d'une couleur d'or ou de feu, voire d'un adjectif plus vague, comme *éclatant* ou *brillant*, est difficile à expliquer de manière univoque.

En définitive, nous constatons que, si l'analyse du segment « croceo velatus amictu » à l'aide d'une grille paraphrastique spécifique nous a amenée à des observations intéressantes, une étude aussi exhaustive n'est pas envisageable pour un corpus comme le nôtre. Si, pour certains points d'analyse, il est pertinent de s'intéresser en détail aux relations de paraphrase et de reformulation, nous avons besoin dans un premier temps d'un système plus général, qui nous permettrait de localiser des passages potentiellement intéressants à comparer et à analyser avec minutie.

Néanmoins, nous observons tout de même quelques pistes d'exploitation d'un classement des paraphrases pour la détection (semi-)automatique des relations intertextuelles. Tout d'abord, certains types de paraphrases paraissent plus déterminants que d'autres pour juger de l'effectivité d'une relation. Il s'agit notamment des changements morphologiques et lexicaux, ainsi que des modifications de graphie. Si les relations syntaxiques sont relativement peu pertinentes et le classement exhaustif reste difficile à établir, elles ne doivent pour autant pas être ignorées. En réalité, la typologie des paraphrases est exploitable pour trois objectifs distincts : détermination des procédés linguistiques (i) permettant de considérer que deux extraits entretiennent une relation intertextuelle ;

¹¹ Maurice de La Porte, *Les Épithètes*, Paris, Gabriel Buon, 1571, p. 215r-215v.

¹² C'est notamment le cas de Villenave qui, par exemple, écrit dans la note 68 du livre IX : « Ovide, selon la remarque de Banier, a écrit cette fable avec tout l'art d'un homme qui connoissoit parfaitement les foiblesses du cœur humain. Mais il entre dans des détails trop délicats pour des oreilles chastes » (p. 288), ou encore, dans la note 89 (p. 293) : « Ovide a réduit presque tous ses traducteurs à omettre plusieurs détails trop libres qui se trouvent dans la fable de Byblis et Caunus, et dans celle d'Iphis. ». Pour conclure cette même note, le traducteur cite ces deux vers de l'*Art poétique* de Boileau : « Le latin dans les mots brave l'honnêteté ; / Mais le lecteur français veut être respecté ».

(ii) qui peuvent rendre la détection de cette relation plus difficile ; (iii) permettant d'évaluer le degré de proximité entre deux extraits dont la relation intertextuelle est avérée. Ainsi, il nous paraît nécessaire que les paraphrases minimales, comme changements de graphie, la flexions et dérivations puissent être prises en compte pour procéder à l'établissement des relations entre extraits textuels.

Pour déterminer si une relation est effective, il semble suffisant de comparer uniquement les mots forts de chaque extrait. Par exemple, pour 15 mots forts décomptés respectivement dans les traductions de Bellegarde1701 et de Duryer1702 (Tableau 21), 10 peuvent être appariés grâce aux paraphrases minimales : six mots sont identiques (*Dieu, mariage, air, Thrace, voix, Orphée*) et deux couples se différencient seulement par la graphie (*vêtu / vêtu, robe / robe*). Nous observons également un changement de flexion (*élevant / éleva*) et une paraphrase dérivationnelle (*or / jaune-doré*).

Tableau 21. Relations lexicales entre les mots forts des traductions de Bellegarde1701 et de Duryer1702

Bellegarde1701	Duryer1702
Le <u>[Dieu]¹</u> du <u>[mariage]²</u> <u>[vêtu]³</u> d'une <u>[robe]⁴</u> <u>[couleur]⁵</u> d' <u>[or]⁶</u> s' <u>[élevant]⁷</u> en l' <u>[air]⁸</u> , <u>[prit]⁹</u> le <u>[chemin]¹⁰</u> de la <u>[Thrace]¹¹</u> , <u>[attiré]¹²</u> par les <u>[charmes]¹³</u> de la <u>[voix]¹⁴</u> d' <u>[Orphée]¹⁵</u> .	De là <u>[Hymen]</u> le <u>[Dieu]₁</u> des <u>[noces]</u> , <u>[vêtu]₃</u> d'une <u>[robe]₄</u> de <u>[jaune-doré]₆</u> , s' <u>[éleva]₇</u> en l' <u>[air]₈</u> pour <u>[aller]</u> en <u>[Thrace]₁₁</u> , où l' <u>[appelloit]</u> la <u>[voix]₁₄</u> d' <u>[Orphée]₁₅</u> pour <u>[assister]</u> à son <u>[mariage]₂</u> .

Néanmoins, l'alignement des deux occurrences du *mariage* n'est pas tout à fait pertinent, car les deux mots n'apparaissent pas dans les syntagmes comparables (« Le Dieu du **mariage** » / « pour assister à son **mariage** »). Le syntagme correspondant chez Duryer1702 étant plutôt « le Dieu des noces », l'appariement optimal pour *mariage* serait son synonyme, *noces*. Le contexte de chaque occurrence doit donc être pris en compte pour optimiser l'appariement, mais la recherche des mots identiques et des paraphrases minimales, comme des changements flexionnels et dérivationnels, parmi les mots forts de chaque extrait semble déjà suffisante pour évaluer la proximité entre eux. Il reste toutefois à déterminer la taille optimale des extraits à comparer pour que l'évaluation de proximité soit la plus efficace et à trouver les critères de découpage (syntaxiques, sémantiques, etc.) les plus pertinentes.

Les paraphrases syntaxiques paraissent utiles surtout pour déterminer ce qui rend la mise en relation difficile. Si la comparaison des syntagmes du même type (nominales, adjectivales, verbales) paraît la plus pertinente dans un premier temps, nous avons vu que des changements d'ordre des mots, des ellipses, des suppressions ou, au contraire, des paraphrases analytiques, doivent également être pris en compte. Nous le voyons déjà pour le couple Bellegarde1701 / Duryer1702, très proche, mais au sein duquel les mots communs n'apparaissent pas dans le même ordre.

La comparaison de l'extrait de Bellegarde1701 avec la traduction analytique de Rachmuhl2003 révèle à son tour (Tableau 22) que seulement cinq mots forts des deux extraits sont identiques, alors que ce deuxième contient plus de mots forts au total (18).

Tableau 22. Relations lexicales entre les mots forts des traductions de Bellegarde1701 et de Rachmuhl2003

Bellegarde1701	Rachmuhl2003
Le [Dieu] ¹ du [mariage] ² vêtu ³ d'une robe ⁴ [couleur] ⁵ d'or ⁶ s'élevant ⁷ en l'air ⁸ , prit ⁹ le chemin ¹⁰ de la [Thrace] ¹¹ , attiré ¹² par les charmes ¹³ de la voix ¹⁴ d'[Orphée] ¹⁵ .	Hyménée, le [dieu] ₁ du [mariage] ₂ , se dirigeait à travers le ciel immense vers la [Thrace] ₁₁ . Son manteau, [couleur] ₅ de safran comme un voile de mariée, flottait au vent. Il se rendait aux noces du poète [Orphée] ₁₅ .

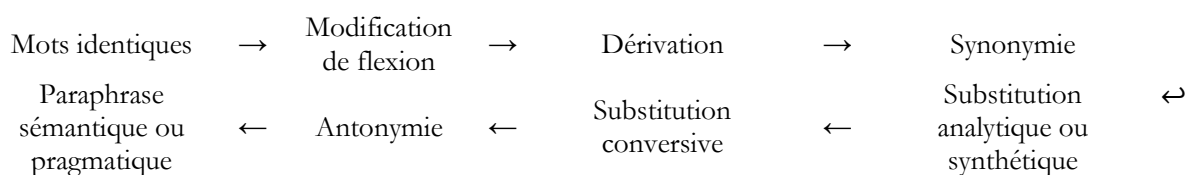
Alors que, pour le couple Bellegarde1701 / Duryer1702, chaque mot fort apparié est précédé ou suivi d'un mot fort non apparié au maximum, pour Rachmuhl2003, nous comptons jusqu'à huit mots forts non appariés à la suite (par exemple entre *couleur* et *Orphée*). Cette variation d'espace entre les mots communs doit être prise en compte dans la recherche des relations intertextuelles notamment pour permettre la comparaison efficace des extraits de longueurs différentes.

Par ailleurs, les liens entre deux couples ne se limitent pas aux changements minimaux, mais incluent aussi les paraphrases lexicales, comme substitutions synonymiques ou antonymiques et substitutions analytiques, synthétiques ou conversives. Pour le couple Bellegarde1701 / Duryer1702, une relation de synonymie est à établir entre *mariage* et *noces*, alors que l'expression polylexicale *prendre le chemin* peut être considérée comme une substitution analytique du verbe *aller*. Pour Bellegarde1701 / Rachmuhl2003 nous avons révélé des relations synonymiques entre *robe* et *manteau* et entre *or* et *safran*. Nous pouvons ajouter à cela des paraphrases conversives, voire sémantiques, comme la relation métaphorique entre « air » et « ciel immense » ou des relations plus distendues entre *attirer* et *appeler*, ou encore *s'élever* et *se diriger*.

Conclusion du chapitre

Comme nous l'avons montré, la détection des paraphrases lexicales pour un corpus diachronique est difficile. Dans l'idéal, il serait nécessaire de déterminer, pour chaque mot, le sens concerné et conforme à l'époque de publication de chaque œuvre et de restreindre la recherche aux synonymes concernés par ce sens uniquement. Mais cela implique un travail fastidieux, à effectuer manuellement et conditionné par le contenu, parfois restreint, des ressources existantes. Les paraphrases sémantiques et pragmatiques sont, quant à elles, extrêmement difficiles à déterminer de manière objective, car elles sont souvent subtiles et renvoient à des compétences extralinguistiques. Enfin, l'appariement doit prendre en compte non seulement les substitutions polylexicales, mais aussi le fait que, parfois, plusieurs appariements sont possibles. Par exemple, pour le couple « voilé d'un manteau » et « couvert d'un voile », nous observons une relation de dérivation entre *voilé* et *voile* et deux substitutions à polarité identique, une entre *voilé* et *couvert* et une entre *manteau* et *voile*. Alors que toutes ces relations sont pertinentes, pour d'autres cas de figure, le choix d'un appariement optimal peut s'avérer nécessaire : pour le couple « à travers l'immensité de l'éther [...] Hyménée s'éloigne, se dirige vers le territoire des Ciconiens » et « Hyménée [...] s'éloigne à travers les plaines immenses de l'air et se dirige vers les rivages des Ciconiens », *territoire* doit être apparié à *rivages* qui a le même complément du nom et non pas à *plaines*, qui est pourtant sémantiquement plus proche (*s.v.* **territoire 1**. « **étendue** de la surface terrestre sur laquelle vit un groupe humain » / *s.v.* **plaine**, n. f. « vaste **étendue** du pays plat ou faiblement ondulé »¹³ *vs* *s.v.* **rivage 1**. « Partie de la terre qui borde une mer »). En même temps, la prise en compte des relations trop incertaines, sémantiques, voire même conversives, surtout pour des verbes très polysémiques, risque d'inclure une quantité trop importante des non-paraphrases si elle n'est pas accompagnée d'une supervision humaine très attentive.

Un système d'évaluation de la force des relations intertextuelles fournirait potentiellement un moyen d'écarter des couples peu pertinents, sans pour autant perdre ceux les plus subtiles et les plus allusifs. Il faciliterait également la détermination des résultats importants, à observer de manière plus détaillée. Selon nous, certains types de paraphrases peuvent être gradués : si la relation entre les mots identiques est la plus forte, celle entre les paraphrases sémantiques ou pragmatiques est principalement allusive.



¹³ Voir l'onglet « Définition » du *Dico en ligne* (Le Robert). Aucun synonyme n'est proposé pour *s.v.* **plaine**.

Si nous constatons donc plusieurs difficultés de l'analyse des paraphrases à l'aide des classements existants, notamment pour des corpus riches en relations intertextuelles, notre investigation nous permet de déterminer nos besoins préliminaires concernant la détection (semi-)automatique des correspondances. Nous ne nous attendons pas à trouver un outil informatique capable de répondre à toutes nos attentes. Notre but n'est pas non plus d'écarter toute intervention humaine. Mais, parmi les techniques et méthodes existantes pour la détection automatique des correspondances textuelles, nous cherchons :

1. des outils capables de travailler avec plusieurs textes à la fois, et adaptés autant que possible à un corpus littéraire ;
2. des paramétrages modulables permettant d'adapter l'outil aux caractéristiques linguistiques de notre corpus, et notamment l'évolution diachronique du français ;
3. la possibilité de prendre en compte des relations morphologiques, dérivationnelles et lexicales, voire d'injecter des ressources nécessaires au logiciel ;
4. les résultats de la détection faciles à extraire et à enrichir selon nos besoins pour produire des visualisations modulables.

Pour déterminer si ces attentes peuvent être satisfaites, il convient à présent de nous intéresser aux traitements computationnels dédiés à la détection de l'intertextualité.

3. Prolégomènes des approches computationnelles de l'intertextualité

L'intertextualité, la notion littéraire qui paraissait révolutionnaire au XX^e s., fait désormais partie des théories critiques traditionnelles. Depuis 1980 environ, les critiques et chercheurs s'intéressent davantage à d'autres méthodes d'analyse, comme les études de réception¹ et de la fiction² ou, plus récemment, les tournants spatial³ et culturel⁴. En même temps, l'intertextualité est généralement considérée comme une école de pensée stable et bien connue, qui occupe une place importante dans l'enseignement littéraire. Elle semble également s'être intégrée dans le vocabulaire non spécialiste, voire interdisciplinaire, en tant que simple relation entre deux textes, établie selon des critères variables⁵.

Toutefois, comme nous nous efforcerons de le démontrer, l'intertextualité se réactualise et reste innovante dans le contexte numérique. La mémoire et la reconnaissance comprises dans leur dimension informatique promettent d'ouvrir des perspectives encore peu explorées pour l'analyse des relations intertextuelles à une échelle plus étendue. Dans l'ouvrage *Quantitative Intertextuality: Analyzing the Markers of Information Reuse*, Christopher Forstall et Walter Scheirer, prônent « l'unification d'un vaste ensemble des phénomènes apparemment disparates en un seul modèle »⁶. Selon les auteurs,

[c]e modèle unificateur est en grande partie construit sur la compréhension structuraliste et poststructuraliste des œuvres littéraires, mais est soutenu et renforcé tout au long de la pensée plus récente dans la science cognitive. Non moins important, le modèle est également basé sur les résultats pratiques de la recherche dans les humanités numériques. Nous appelons cette approche *l'intertextualité quantitative*.⁷

Même si les techniques informatiques occupent une place importante dans cette nouvelle méthodologie de recherche, l'intertextualité quantitative repose surtout sur la curiosité interdisciplinaire, et invite les chercheurs à se demander comment les différentes disciplines peuvent échanger et

¹ Voir Yves Chevrel, « Les études de réception », dans Pierre Brunel (dir.), *Précis de littérature comparée*, 1^{re} éd., Paris, Presses Universitaires de France, 1989, p. 177-214.

² Voir par exemple Thomas Pavel, *Univers de la fiction*, trad. Thomas Pavel, Paris, Seuil, coll. « Poétique », 1988.; Françoise Lavocat, *Fait et fiction : pour une frontière*, Paris, Seuil, coll. « Poétique », 2016.

³ Voir notamment le dossier « Nouvelles géographies » dirigé par Jacques Lévy (*Le Débat*, n° 92, 1996).

⁴ Voir par exemple Alexandre Gefen, *L'idée de littérature : de l'art pour l'art aux écritures d'intervention*, Paris, Éditions Corti, coll. « Les essais », 2021.

⁵ Dans le *Dico en ligne* (Le Robert) notamment, *intertextualité* est définie comme « [e]nsemble des relations existant entre plusieurs textes (notamment littéraires) ».

⁶ « The first part is an argument for uniting a broad set of seemingly-disparate phenomena in a single model. » (nous traduisons). Voir Christopher Forstall et Walter Scheirer, « What is Quantitative Intertextuality? », *Quantitative Intertextuality: Analyzing the Markers of Information Reuse*, Cham, Springer International Publishing, 2019, p. 4, en ligne, <https://doi.org/10.1007/978-3-030-23415-7_3>.

⁷ « That unifying model is in large part built on structuralist and post-structuralist understanding of literary works, but is supported and strengthened throughout by more recent thought in cognitive science. No less important, the model is also grounded in the practical results of research in the Digital Humanities. We call this approach *quantitative intertextuality* » (nous traduisons). Voir *Ibid.*, p. 4.

enrichir mutuellement leurs outils et méthodes. Malgré le qualitatif « quantitative », il ne s'agit pas de compter le nombre de relations intertextuelles que chaque texte contient, mais de comprendre de quelle manière un texte est intertextuel. En cela, l'intertextualité quantitative ne fait pas abstraction de la réception du texte littéraire. Au contraire, comme le soulignent Forstall et Scheirer, elle en est indissociable :

[p]our nous, l'expérience du lecteur (et, dans la mesure où elle est accessible, celle de l'écrivain aussi), est au centre de tout cela. Dans l'immédiat, il s'agit d'une simple question d'opportunité pratique : nous construisons un modèle qui ne peut, pour l'instant, traiter que des caractéristiques mesurables du texte lui-même et ne peut, en soi, rien savoir de l'expérience humaine. Cependant, nous souscrivons au pluralisme des modèles et nous ne craignons pas qu'une théorie de l'intertextualité centrée sur l'expérience humaine soit en quelque sorte diminuée ou "expurgée" par un modèle de calcul. Bien au contraire, chacun, croyons-nous, éclaire et ajoute de la profondeur à l'autre.⁸

Alors que le structuralisme et le poststructuralisme ont été écartés par la critique moderne au profit des théories plus cognitives, la lecture distante peut permettre, paradoxalement, de réconcilier les deux : « nous sommes enfin en mesure de tester scientifiquement quelques théories et modèles ambitieux proposés par les penseurs antérieurs dont la vision a dépassé les outils à leur disposition »⁹. Toutefois, pour surmonter l'aspect purement quantitatif des modèles informatiques et exploiter pleinement le potentiel de la lecture distante dans l'analyse des relations intertextuelles, il est indispensable de faire coopérer les méthodes informatiques, statistiques et mathématiques avec les théories littéraires, les méthodes linguistiques et les modèles cognitifs.

Différentes méthodes, plus ou moins propices à une telle coopération, existent. Quels que soient leurs objectifs, selon Forstall et Scheirer, elles se composent toutes de trois étapes principales :

1. L'extraction des caractéristiques textuelles mesurables et quantifiables ;
2. La reconnaissance des patrons à travers les procédures mathématiques ;
3. L'interprétation humaine et subjective des résultats.¹⁰

Alors que dans la partie suivante¹¹, nous nous intéresserons spécifiquement aux logiciels dédiés à la détection des réutilisations textuelles, dans ce chapitre, nous présentons les principales mé-

⁸ « For us, the reader's experience (and to the degree that it is accessible, that of the writer as well) is at the centre of this. For the present, it is a simple matter of practical expediency: we are building a model that can, for now, treat only measurable features of the text itself and can, in itself, know nothing of the human experience. We subscribe to a pluralism of models, however, and have no fear that a theory of intertextuality centred on the human experience is somehow diminished or "explained away" by a computational one. Very much to the contrary, each, we believe, illuminates and adds depth to the other » (nous traduisons). Voir Christopher Forstall et Walter Scheirer, « Statistical Learning as a Model for Intertextuality », *Quantitative Intertextuality: Analyzing the Markers of Information Reuse*, Cham, Springer International Publishing, 2019, p. 29, en ligne, <https://doi.org/10.1007/978-3-030-23415-7_3>.

⁹ « In many cases, we are in positions to finally scientifically test some of the ambitious theories and models proposed by earlier thinkers whose vision exceeded the tools at their disposal » (nous traduisons). *Ibid.*, p. 25.

¹⁰ Christopher Forstall et Walter Scheirer, « What is Quantitative Intertextuality ? », *op. cit.*, p. 6.

¹¹ Voir partie 3 « Logiciels dédiés à la détection automatique des réemplois textuels », p. 244.

thodes et algorithmes ayant été ou pouvant être exploités pour l'interrogation d'un corpus de réécritures en vue de la détection (semi-)automatique des relations intertextuelles. Cette manière de procéder nous permettra de présenter quelques méthodes-phares utilisées dans le traitement computationnel de textes littéraires, avant de nous pencher plus spécifiquement sur les logiciels dédiés à la détection des réutilisations textuelles et ainsi de graduer de manière plus progressive la technicité de nos propos. Comme nous le verrons en expérimentant des méthodes comme la stylométrie, la modélisation thématique ou l'alignement des séquences sur notre corpus, certains outils et certaines méthodes semblent plus aptes que d'autres à répondre à nos besoins.

En effet, l'interrogation de notre corpus avec ces méthodes, pour la plupart intégrés dans les logiciels de détection automatique des réutilisations textuelles, est indispensable pour ensuite mieux apprécier la performance de ces logiciels et choisir ceux qui semblent répondre le mieux à nos besoins. Notre but n'est donc pas d'expliquer en détail le fonctionnement des algorithmes impliqués dans ces technologies, mais d'interroger leur performance pour le traitement des corpus littéraires, et notamment pour l'analyse des relations intertextuelles au sein d'un corpus des réécritures d'un mythe. Aussi, pour que le lecteur puisse apprécier pleinement les résultats obtenus avec chaque méthode, nous nous efforçons d'accompagner nos expérimentations d'annexes complets permettant de les reproduire.

a) Stylométrie

La stylométrie est une méthode statistique particulière, évoquée régulièrement dans le contexte de la détection du plagiat. Cette approche part du principe que la fréquence d'utilisation des mots ou des séries des mots constitue un des facteurs caractéristiques du style d'un auteur, ce qui permet, notamment, de cibler des passages qui s'éloignent des caractéristiques du reste du document¹². Appliquée aux études littéraires, la stylométrie est exploitée surtout pour l'attribution de l'auctorialité des textes anonymes ou ceux écrits sous un pseudonyme. Elle permet également de mesurer le style des auteurs ou des œuvres.

Déjà en 1887, bien avant l'avènement du numérique, T. C Mendenhall a compté la longueur des mots dans les différents extraits des œuvres de Dickens pour constater que les courbes ainsi obtenues sont très semblables, alors qu'elles diffèrent nettement de celles des autres auteurs¹³. Cette approche pionnière s'est avérée trop basique pour pouvoir s'appliquer à tous les cas de figure : on

¹² Voir par exemple Parvati Iyer et Abhispita Singh, *Document Similarity Analysis for a Plagiarism Detection System*, Pune, India, 2005 ; Gabriel Oberreuter et Juan D. Velásquez, « Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style », *Expert Systems with Applications*, vol. 40, n° 9, juillet 2013, p. 3756-3763, en ligne, <<https://www.sciencedirect.com/science/article/pii/S0957417412013231>>.

¹³ T. C. Mendenhall, « The Characteristic Curves of Composition », *Science*, vol. 9, n° 214, 1887, p. 237-249, en ligne, <<https://www.jstor.org/stable/pdf/1764604>>.

constate rapidement qu'outre le décompte de la longueur des unités lexicales, pour arriver à la fiabilité satisfaisante des résultats, il est nécessaire de prendre en compte la fréquence des mots employés, d'analyser la syntaxe et d'observer l'emploi caractéristique de la ponctuation.

Le développement fulgurant des techniques informatiques a permis d'améliorer de manière significative les performances de la stylométrie et d'étendre ses applications à plusieurs domaines. Mais ce succès a également été source de controverses. De nombreux problèmes éthiques, liés à l'utilisation de la stylométrie pour résoudre des affaires criminelles¹⁴ ou pour démasquer les auteurs écrivant sous pseudonyme¹⁵, font que cette méthode ne fait pas l'unanimité.

Outre les questions éthiques, un autre problème concerne la compréhension des méthodes et algorithmes mis en place. Le danger d'une exploitation erronée d'un outil est particulièrement accru si l'on ne comprend pas entièrement son fonctionnement. Or, s'il est relativement facile d'exploiter les outils stylométriques et d'obtenir les premiers résultats, il est bien plus difficile de comprendre les méthodes statistiques mises en œuvre et, par conséquent, juger si les résultats obtenus ne sont pas biaisés par un paramétrage non adapté de l'outil¹⁶ ou la mauvaise préparation du corpus¹⁷.

Pour mieux comprendre les enjeux et les écueils de la stylométrie, observons concrètement ce que nous obtenons pour notre corpus des réécritures du mythe d'Orphée et Eurydice. De nombreuses méthodes stylométriques existent, dont trois, Delta, Zeta et Iota, sont les plus fréquemment utilisées. La première est privilégiée pour l'attribution de l'auctorialité des œuvres anonymes¹⁸.

¹⁴ À ce sujet, voir par exemple Marie Puren et Simon Gabay, « Affaire Grégory : la stylométrie permettra-t-elle enfin d'identifier le corbeau ? », *The Conversation*, En ligne, 2021, en ligne, <<http://theconversation.com/affaire-gregory-la-stylometrie-permettra-t-elle-enfin-d-identifier-le-corbeau-153899>>.

¹⁵ Voir par exemple Patrick Juola, « How a Computer Program Helped Show J.K. Rowling write A Cuckoo's Calling », *Scientific American*, 20 août 2013, en ligne, <<https://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/>>

¹⁶ Voir notamment le débat autour de la proximité du style de Molière et de Corneille. Une thèse selon laquelle les deux dramaturges étaient la même personne, émise entre autres par Cyril et Dominique Labbé grâce aux approches stylométriques (« La distance intertextuelle et l'attribution d'auteur. Corneille et Molière », *Journal of Quantitative Linguistics*, vol. 8, n° 3, 2000, p. 213-231, en ligne, <<https://halshs.archives-ouvertes.fr/halshs-00137675>>.), a été invalidé 19 ans plus tard par Florian Cafiero et Jean-Baptiste Camps (« Why Molière most likely did write his plays », *Science Advances*, vol. 5, n° 11, American Association for the Advancement of Science, novembre 2019, en ligne, <<https://www.science.org/doi/10.1126/sciadv.aax5489>>.), avec les mêmes méthodes stylométriques. Selon Christopher Schöch, les Labbé auraient artificiellement lissé les différences entre les deux auteurs en privilégiant la mesure qui augmente le poids des lemmes fréquents (la distance euclidienne). Le travail sur les textes lemmatisés aurait réduit la disparité des fréquences de différentes formes des mots privilégiés par l'un ou l'autre auteur. Voir Christof Schöch, « Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik », dans Lars Schneider et Christof Schöch (dir.), *Literaturwissenschaft im digitalen Medienwandel*, PhiN, coll. « Beihefte zu Philologie im Netz », 2014, p. 130-157, en ligne, <<https://hal.archives-ouvertes.fr/hal-00957091>>.

¹⁷ « L'ambition de la société suisse OrphAnalytics, qui espère retrouver le corbeau de l'affaire Grégory avec pour unique preuve cinq courtes lettres, rejoint ces dernières recherches sur les réseaux sociaux, dont les publications ne dépassent parfois pas 140 caractères. [...] Le raccourcissement des textes analysés va cependant de pair avec une diminution de la fiabilité de la stylométrie. Une récente étude avec des textes du XIX^e siècle a ainsi déterminé que 5 000 mots était un prérequis pour garantir la solidité des résultats. », Marie Puren et Simon Gabay, « Affaire Grégory », *loc. cit.*

¹⁸ Voir John Burrows, « 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship », *Literary and Linguistic Computing*, vol. 17, n° 3, septembre 2002, p. 267-287, en ligne, <<https://doi.org/10.1093/lilc/17.3.267>>.

Pour calculer la mesure appelée *la distance Delta*, on mélange un corpus d’auteurs différents et on génère la liste des 150 mots les plus fréquents pour tous les textes. Pour notre corpus de réécritures, nous retrouvons en première position les déterminants, les prépositions, les auxiliaires, etc., considérés comme formes faibles (ou vides). La première forme forte (ou active), *Orphée*, apparaît en 49^e position. Sa fréquence absolue, c’est-à-dire le nombre d’occurrences de cette forme dans le corpus, atteint 940, alors que la forme la plus fréquente, *de*, apparaît 11 028 fois¹⁹. Il est normal que les mots faibles soient les plus fréquents au sein d’un corpus. Dans la méthode Delta, c’est précisément la fréquence d’emploi de ces mots-là qui est considérée comme caractéristique du style de chaque auteur. La fréquence de l’utilisation de chaque mot qui figure sur la liste est comparée pour chaque texte, et ce dernier est représenté comme une série de nombres (vecteurs). Enfin, une métrique des proximités est générée à l’aide d’une formule qui permet de calculer les distances.

Plusieurs formules de calcul sont proposées et de nombreux paramétrages sont à prendre en compte et à adapter aux particularités du corpus exploité. Même si certaines mesures semblent plus performantes que d’autres²⁰, nous verrons que le choix optimal n’est pas aisé et demande de connaître au préalable les composants de son corpus. Nous expérimentons cela à l’aide de la bibliothèque Stylo, conçue en langage R par Maciej Eder, Jan Rybicki et Mike Kestemont²¹. Nous générons quatre métriques de proximité en utilisant deux mesures, la mesure Delta Classique (Figure 27 et Figure 29) et la mesure Delta Cosine (Figure 28 et Figure 30), considérée comme la plus performante par Evert *et al.*²². Pour chaque mesure, nous prenons en compte d’abord les 100 mots les plus fréquents (Figure 27 et Figure 28), puis les 200 mots les plus fréquents (Figure 29 et Figure 30)²³.

Pour évaluer la pertinence de chaque choix des paramètres, nous pouvons par exemple observer la proximité établie automatiquement entre les œuvres du même auteur. Au sein de notre corpus, il s’agit de cinq cas de figure :

¹⁹ Le corpus contient 303 179 mots au total. Pour la liste des 150 formes les plus fréquentes, leur fréquence absolue et relative, voir [1-2-3_150-mots-frequents.xlsx](#).

²⁰ Voir par exemple Stefan Evert et al., « Understanding and explaining Delta measures for authorship attribution », *Digital Scholarship in the Humanities*, vol. 32, décembre 2017, p. 4-16, en ligne, <doi: 10.1093/llc/fqx023>.

²¹ Voir https://computationalstylistics.github.io/stylo_nutshell/# et Maciej Eder, Jan Rybicki et Mike Kestemont, « Stylometry with R: A Package for Computational Text Analysis », *The R Journal*, vol. 8, n° 1, 2016, p. 107-121, en ligne, <<https://journal.r-project.org/archive/2016/RJ-2016-007/index.html>>. Pour reproduire l’exploitation de notre corpus avec cette bibliothèque, voir le contenu du dossier [II_3_stylometrie](#).

²² Stefan Evert et al., « Understanding and explaining Delta measures for authorship attribution », *loc. cit.*

²³ Voir les quatre schémas dans le dossier [II_3_stylometrie](#) :

- [1-ClassicDelta_100MFW](#) : mesure Delta Classique prenant en compte les 100 mots les plus fréquents ;
- [2-CosineDelta_100MFW](#) : mesure Delta Cosine prenant en compte les 100 mots les plus fréquents ;
- [3-ClassicDelta_200MFW](#) : mesure Delta Classique prenant en compte les 200 mots les plus fréquents ;
- [4-CosineDelta_200MFW](#) : mesure Delta Cosine prenant en compte les 200 mots les plus fréquents.

- Deux versions très proches de la cantate *La mort d'Orphée* d'Hector Berlioz (Berlioz1827 et Berlioz1828) ;
- Deux parts de la mazarinade anonyme *L'Orphée grotesque, avec le bal rustique. En vers burlesques* (Martin1649) et *Suite de l'Orphée avec les Bacchantes ou Les Rudes Joyeuses* (Martin1649-2) ;
- La brève nouvelle de Pierre-Simon Ballanche, « Sixième Fragment » (Ballanche1809) et son épopée *Essais de palingénésie sociale. Orphée* (Ballanche1829) ;
- Deux poèmes de Tristan l'Hermite, *La Lyre d'Orphée* (Hermite1641) et *Orphée aux Enfers parlant à Pluton* (Hermite1652)
- Deux œuvres théâtrales de Jean-François Regnard, une comique, *La Descente d'Arlequin aux Enfers* (Regnard1689) et une sérieuse, *Orphée aux enfers, opéra* (Regnard1699).

Nous nous attendons à ce que la stylométrie détecte automatiquement la proximité des œuvres écrites par le même auteur. Le cas échéant, leurs identifiants apparaîtront sur la même branche dans le dendrogramme²⁴ qui permet de visualiser les résultats, de préférence un à côté de l'autre (sans qu'un autre identifiant les sépare).

D'abord, Berlioz1827 / Berlioz1828 et Martin1649 / Martin1649-2 sont effectivement adjacents dans les quatre visualisations, peu importe la mesure prise en compte. En revanche, Ballanche1809 et Ballanche1829 sont voisins directs seulement dans les résultats avec la mesure Delta Cosine qui prend en compte les 100 mots les plus fréquents (ci-après : 100 MFW). Pour deux autres paramétrages, Delta Classique 100 MFW et Delta Cosine 200 MFW, les deux œuvres ne font pas partie de la même branche, mais restent proches, car seulement une œuvre se place entre elles, *L'Histoire des amants célèbres dans l'Antiquité* des Esquiros (Esquiros1849) pour la première mesure et *Orphée d'Afrique* de Werewere Liking (Liking1981) pour la deuxième. Nous avons classé ces œuvres, comme ceux de Ballanche, parmi les réécritures idéologiques qui insistent sur l'aspect initiatique du mythe orphique²⁵, nous pouvons donc considérer ces deux résultats comme plutôt satisfaisants. Pour Delta Classique 200 MFW, en revanche, cinq œuvres séparent les deux textes de Ballanche, ce qui est assez décevant. Ensuite, les deux poèmes de Tristan l'Hermite, Hermite1641 et Hermite1652, sont éloignés dans tous les résultats sauf un, celui obtenu avec Delta Cosine 200 MFW où ils sont séparés uniquement par le texte anonyme *Le Ballet des machines représentant le V^e acte de la Mort d'Orphée et d'Euridice*²⁶. Enfin, les deux pièces de Regnard, Regnard1689 et Regnard1699, sont positionnées très loin l'une de l'autre dans les quatre résultats.

²⁴ Dendrogramme est un type de diagramme qui permet de visualiser l'arborescence complexe des relations hiérarchiques (comme, par exemple, l'arbre généalogique).

²⁵ Voir partie 1, chap. 3, sec. b) « Répartition fonctionnelle », p. 76.

²⁶ Il semble peu probable que ce texte soit écrit par l'Hermite. C'est plutôt la proximité avec *La grande Journée des machines, ou Le mariage d'Orphée et d'Euridice* de Chapoton (1648, Chapoton1648) qui serait à prendre en compte. Voir 1.3a) « Répartition chronologique », p. 68.

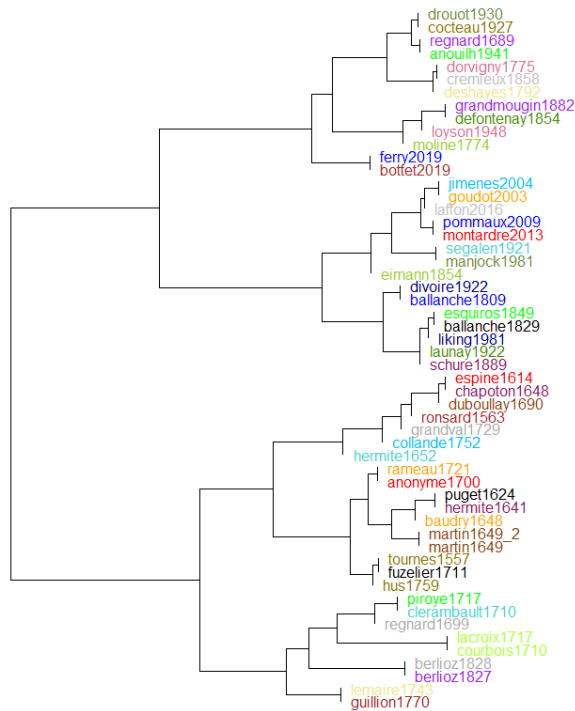


Figure 27. Résultats avec la mesure Delta Classique
prenant en compte les 100 mots les plus fréquents

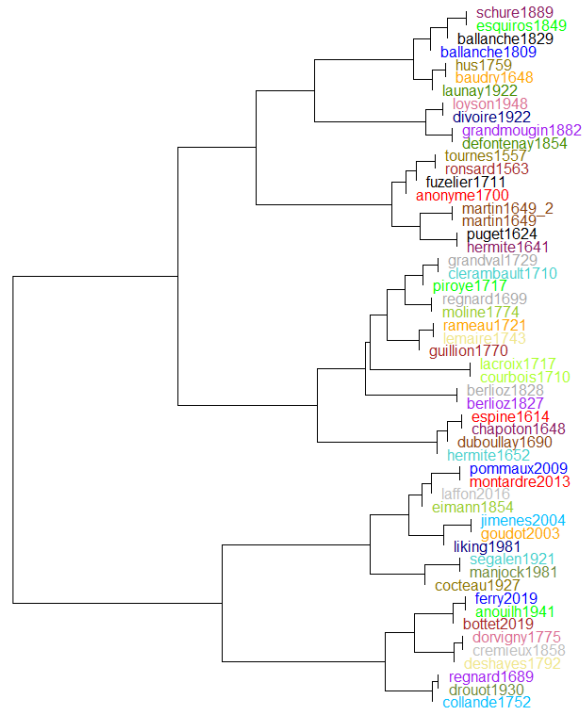


Figure 28. Résultats avec la mesure Delta Cosine
prenant en compte les 100 mots les plus fréquents

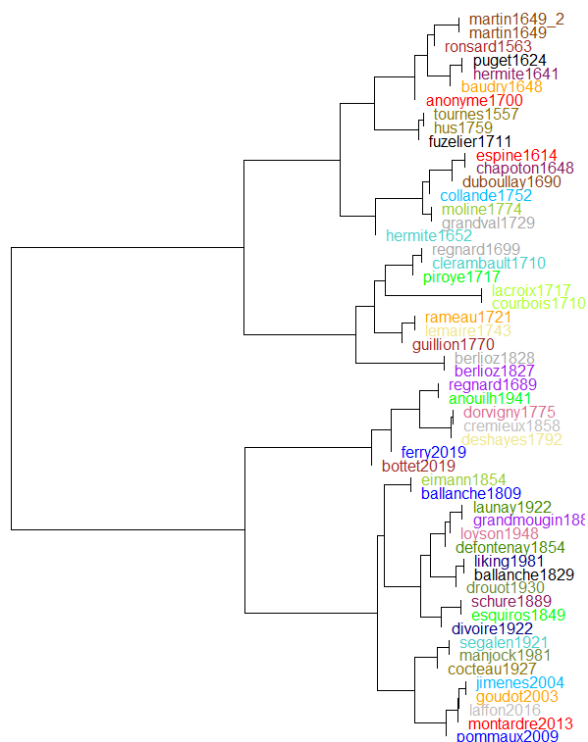


Figure 29. Résultats avec la mesure Delta Classique
prenant en compte les 200 mots les plus fréquents

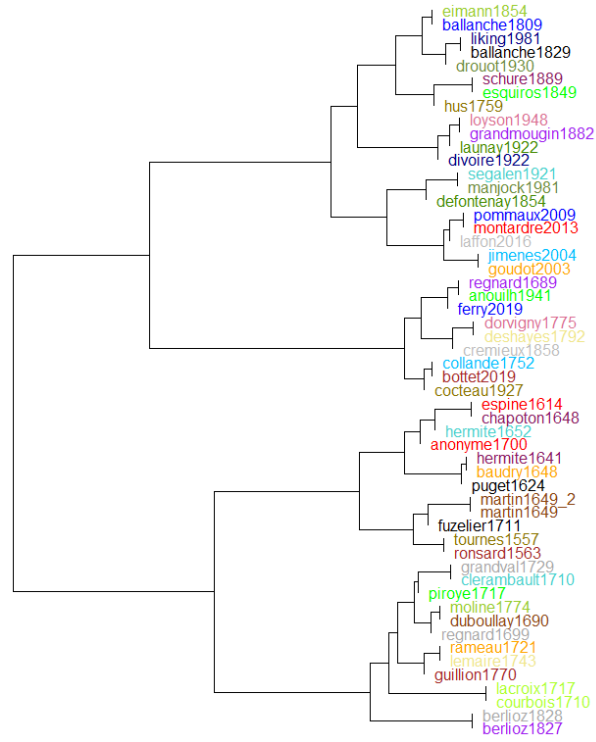


Figure 30. Résultats avec la mesure Delta Cosine
prenant en compte les 200 mots les plus fréquents

Cette observation préliminaire semble privilégier les résultats de Delta Cosine 100 MFW (Figure 28), la seule qui détecte correctement la parenté des œuvres Ballanche1809 et Ballanche1829. En plus, la cohérence thématique est davantage respectée avec cette mesure et les résultats sont plus concordants avec notre classement manuel. Du point de vue générique, les cantates sont généralement regroupées ensemble (de Grandval1729 à Berlioz1827), comme certains romans de jeunesse (de Pommaux2009 à Goudot2003). Les résultats sont également satisfaisants pour les types des réécritures que nous avons déterminés au préalable : la plupart des réécritures idéologiques apparaissent ensemble (de Schure1889 à Defontenay1854), les quatre opéras-comiques (Dorvigny1775, Cremieux1858, Deshayes1792 et Regnard1689) font partie de la même branche et la version française de l'opéra *Orphée et Eurydice* de Gluck (Moline1774) est entourée d'une part par la cantate *Orphée* de Rameau (Rameau1721) qui est intégrée dans la pièce et, d'autre part, par l'opéra *Orphée aux Enfers* de Regnard (Regnard1699) qui propose un traitement comparable du mythe.

La concordance de ces résultats avec notre classement manuel est intéressante, mais plusieurs difficultés persistent. Tout d'abord, alors que la stylométrie nous permet de déterminer automatiquement quelles œuvres sont proches, la nature de cette proximité n'est pas stable. En effet, pour certaines proximités, les ressemblances du style auctorial semblent prévaloir sur d'autres facteurs, alors que, pour d'autres, nous observons plutôt des regroupements génériques, synchroniques (les œuvres parues à la même époque ont tendance à être classées ensemble), voire thématiques. Par ailleurs, alors que la mesure Delta Cosine 100 MFW s'est montrée plus probante que les trois autres, il est difficile de déterminer si le paramétrage optimal a été atteint ou si une autre mesure peut s'avérer plus efficace. Enfin, comme nous souhaitons détecter concrètement les réemplois textuels, et donc pouvoir travailler sur les extraits textuels directement et non seulement sur les statistiques générales de chaque œuvre, la comparaison globale des œuvres, facilitée par la méthode Delta, n'est pas suffisante pour nos besoins.

C'est la méthode Zeta qui s'intéresse davantage au détail de chaque texte afin de faciliter une étude quantitative du style d'un auteur. Elle est initialement proposée par Hugh Craig et Arthur F. Kinney dans une étude consacrée à William Shakespeare²⁷. Les auteurs ont analysé l'emploi du mot *gentle* par le dramaturge en extrayant de chaque texte du corpus 3 000 mots et en décomptant le nombre d'extraits par texte qui contiennent au moins une occurrence du mot recherché. La comparaison des fréquences dans les textes de Shakespeare avec celles des autres auteurs a révélé un emploi particulièrement récurrent du mot *gentle* chez ce premier. Appliquée à tous les mots forts présents dans le corpus (contrairement à Delta, qui privilégie les mots faibles), la méthode Zeta

²⁷ D. H. Craig et Arthur F. Kinney, *Shakespeare, computers, and the mystery of authorship*, New York, Cambridge University Press, 2009.

permet donc de déterminer les mots préférés ou évités par des auteurs et, par la même occasion, d'identifier les textes anonymes qui se rapprochent le plus du style de chacun.

Dans le cadre de notre recherche, nous avons moins besoin d'un lexique spécifique à chaque auteur que du vocabulaire commun à tous les textes, susceptible de nous fournir une liste des mots-clés les plus saillants du mythe. Ce qui nous intéresse, en somme, c'est une approche inverse de la méthode Zeta : en partant de proximités entre les textes, nous souhaitons identifier les mots forts les plus saillants et les plus récurrents dans notre corpus.

b) Détection et modélisation des topiques (*topic modeling*)

Une telle approche existe : il s'agit de la détection et modélisation des topiques (*topic modeling*)²⁸. Appuyée, elle aussi, sur des calculs statistiques, cette technique de l'apprentissage machine non-supervisé part d'un principe que chaque texte possède des thèmes qui le caractérisent et que ces thèmes peuvent être décrits par des mots-clés. Les topiques sont alors compris comme des patrons récurrents des mots en cooccurrence. La difficulté de l'application de cette méthode au corpus littéraire vient du fait que l'algorithme le plus populaire, *Latent Dirichlet Allocation* (LDA) donne des résultats satisfaisants uniquement avec des corpus massifs, bien plus grands que le nôtre. C'est particulièrement visible quand nous demandons les 10 topiques les plus saillants pour notre corpus et les 10 mots-clés les plus caractéristiques de chaque topique (Tableau 23)²⁹. Sur 100 mots-clés trouvés par l'algorithme, 8 seulement sont uniques. Certains, comme *amour*, *Eurydice*, *jour* et *mort* apparaissent dans chaque topique (10 occurrences sur 100 chacun).

Tableau 23. Résultats de la détection des topiques avec l'algorithme LDA

N°	Dix mots-clés les plus saillants
1.	Dieu, vie, homme, mort, jour, amour, cœur, voix, deux, Eurydice
2.	Eurydice, jour, mort, homme, amour, vie, dieu, seul, voix, grand
3.	Eurydice, amours, homme, beau, jour, air, mort, deux, terre, grand
4.	Mort, amour, dieu, jour, homme, Eurydice, vie, grand, œil, aimer
5.	Eurydice, amour, aimer, mort, vie, dieu, cœur, monde, jour, main
6.	Amour, homme, Eurydice, seul, fois, jour, mort, vers, deux, femme
7.	Homme, dieu, Eurydice, amour, vie, mort, temps, seul, jour, grand
8.	Homme, vie, amour, Eurydice, jour, temps, mort, deux, terre, dieu
9.	Eurydice, homme, amour, dieu, jour, mort, vie, seul, grand, cœur
10.	Homme, vie, dieu, seul, grand, Eurydice, mort, amour, terre, jour

²⁸ Pour les articles d'introduction destinés aux chercheurs en Humanités numériques, voir par exemple Megan R. Brett, « Topic Modeling: A Basic Introduction », *Journal of Digital Humanities*, vol. 2, n° 1, 2012, en ligne, <<http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/>> ; Ted Underwood, « Topic modeling made just simple enough », dans *The Stone and the Shell*, 7 avril 2012, en ligne, <<https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/>>. Pour les études plus détaillées, voir par exemple David M. Blei, « Probabilistic topic models », *Communications of the ACM*, vol. 55, n° 4, avril 2012, p. 77-84, en ligne, <<https://dl.acm.org/doi/10.1145/2133806.2133826>>; John W. Mohr et Petko Bogdanov, « Topic models: What they are and why they matter », *Poetics*, vol. 41, n° 6, « Topic Models and the Cultural Sciences », décembre 2013, p. 545-569, en ligne, <<https://www.sciencedirect.com/science/article/pii/S0304422X13000685>>.

²⁹ Pour reproduire la détection des topiques pour notre corpus avec le langage Python, voir [II 3 topic-modeling](#).

Nous recensons également 9 occurrences d'*homme*, 8 de *dieu* et *vie*, 6 de *grand*, 5 de *seul*, 4 de *deux*, 3 de *cœur* et *terre*, puis 2 d'*aimer*, *temps* et *voix*. Enfin les 8 mots uniques sont : *air*, *beau*, *femme*, *fois*, *main*, *monde*, *œil* et *vers*.

Ces résultats ne sont pas incorrects, mais manquent de précision et leur fiabilité reste incertaine. Il est vrai que les mots-clés les plus fréquents *Eurydice*, *amour* et *mort* rendent bien compte des mythes centraux du mythe d'Orphée et Eurydice. Toutefois, il est étonnant de ne constater que deux occurrences relatives au chant d'Orphée (*air* et *voix*) et de ne voir aucun indice (autre *dieu*) de la catabase du poète. Par ailleurs, le topique n° 7 est désigné comme le plus caractéristique pour toutes les œuvres, ce qui empêche par exemple l'exploitation des résultats pour la partition automatique du corpus. Enfin, alors que tous les textes du corpus traitent effectivement du même sujet, nous sommes surprise que les caractéristiques particulières de certains sous-ensembles ne soient pas relevées : nous nous attendions par exemple à retrouver des termes génériques caractéristiques, comme *récitatif* ou *scène*, voire même des indices nous permettant déjà de constater quelques traitements particuliers des mythes traditionnels.

Un autre algorithme exploitable pour la détection des topiques est appuyé sur la méthode de la factorisation matricielle non-négative (*Non-negative Matrix Factorization*, ci-après : NMF)³⁰. Nous constatons que cette méthode s'applique bien mieux à notre petit corpus et propose des résultats plus probants³¹ (Tableau 24).

Tableau 24. Résultats de la détection des topiques avec l'algorithme NMF

N°	Dix mots-clés les plus saillants
1.	Homme, Orphée, âme, cœur, dieu, amour, prêtre, vie, grand, terre
2.	Eurydice, récitatif, Pluton, Orphée, ardeur, écouter, tendre, trahir, arme, toujours
3.	Orphée, Eurydice, Charon, lyre, Hadès, Aristée, Jason, Enfer, tête, Perséphone
4.	Arlequin, ah, scène, air, femme, petit, cela, Javotte, Roger, voilà
5.	Âme, amour, Eurydice, mort, beau, tôt, dieu, toujours, mal, esprit
6.	Soumettre, bienfait, univers, belliqueux, mal, accent, Latone, loi, céleste, dieu
7.	Nyango, Orphée, spectateur, Ndinga, père, salle, image, lumière, maison, ça
8.	Vielle, danser, vielleur, danse, bal, burlesque, veuf, grotesque, jouer, femme
9.	Parmi, chercher, lumière, funeste, dedans, connaître, tribunal, accord, former, redemander
10.	Amour, Eurydice, récitatif, air, tendre, cesser, cantate, sombre, amant, objet

³⁰ Pentti Paatero et Unto Tapper, « Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values », *Environmetrics*, vol. 5, n° 2, 1994, p. 111-126, en ligne, <<https://onlinelibrary.wiley.com/doi/abs/10.1002/env.3170050203>>.

³¹ Les articles d'introduction à la détection des topiques pour les chercheurs SHS que nous avons consultés (voir n. 28) introduisent principalement la méthode LDA, ce qui nous incite à penser qu'il s'agit de l'algorithme le plus populaire, du moins dans le domaine des Humanités numériques. Cependant, l'étude comparative de Shini George et de Srividhya Vasudevan (« Comparison of LDA and NMF Topic Modeling Techniques for Restaurant Reviews », *Indian Journal of Natural Sciences*, vol. 10, n° 62, octobre 2020.), par exemple, révèle que la méthode NMF obtient meilleurs résultats que la LDA (« The Topic Modelling methods LDA and NMF were compared based on the performance metrics such as Mean Coherence Score and execution time. In that, NMF secured more Coherence score as 0.729 while comparing with LD Coherence Score as 0.469 and also NMF has less execution time when compared to LDA. Hence NMF is better while comparing with LDA. », p. 28 216).

Remarquons tout d'abord que le vocabulaire de chaque topique est beaucoup plus spécifique que ce que nous obtenons avec la méthode LDA. Un seul mot-clé apparaît de manière particulièrement récurrente : il s'agit du nom d'Eurydice, qui se retrouve parmi les termes les plus saillants des quatre sur dix topiques (n^{os} 2, 3, 5 et 10). Dans trois autres, elle n'est pas mentionnée directement, mais nous retrouvons des mots du champ lexical de l'amour (*cœur, âme, amour, amant*) et celui du mariage (*femme, veuf*), ainsi que d'autres noms féminins, extérieurs au mythe (*Javotte, Nyango*)³².

Par ailleurs, la coprésence du nom *femme* et de l'adjectif *petit* parmi les mots-clés du topique n° 4 renvoie potentiellement à la périphrase « petite femme ». En effet, ce syntagme apparaît sept fois dans notre corpus et désigne toujours Eurydice :

- cinq fois dans la comédie de Regnard, *Arlequin aux Enfers* (Regnard1689) : « Bonjour, ma petite femme ! », « Ah, ma petite femme ! », « Là, là, là, ma petite femme, ce ne sera rien », « Ma petite femme, je n'ai jamais douté de votre affection » et « mais il n'en coûte pas grand'chose pour faire enterrer une petite femme » ;
- une fois dans l'opéra-comique *Le Petit Orphée* (Deshayes1792) : « Tu fuis ta petite femme / Qui t'aime si tendrement. » ; et
- une fois dans l'opéra burlesque *Orphée aux Enfers* (Crémieux1858) : « Mais avec une jolie petite femme qu'il venait d'enlever à son mari ! ».

Le topique n° 3 est le plus caractéristique de ces trois textes, en plus de trois autres ouvrages que nous avons classés manuellement parmi les détournements comiques du mythe (Tableau 25) : le divertissement *Orphée ou Arlequin aux Enfers* (Fuzelier1711), *Roger-Bontems et Javotte, parodie d'Orphée et Eurydice* (Dorvigny1775) et *Orphée. Tragédie en un acte et un intervalle* (Cocteau1927). Nous y retrouvons également une autre pièce de théâtre, *Eurydice* (Anouilh1941) qui a été classée plutôt comme une modernisation. Nous verrons toutefois que les deux pièces de théâtre modernes, Cocteau1927 et Anouilh1941, approchent le mythe de manière particulière, en oscillant sans cesse entre modernisation sérieuse et détournement humoristique³³.

³² Pour l'analyse de l'émancipation d'Eurydice au sein de notre corpus, voir IV.2c) « Exploitation des données obtenues ».

³³ Voir IV.2c) « Exploitation des données obtenues ».

Tableau 25. Détection des topiques avec l'algorithme NMF : partitionnement automatique du corpus

N°	Dix mots-clés les plus saillants	Identifiants des œuvres ³⁴ pour lesquelles le topique est détecté comme le plus caractéristique
1.	Homme, Orphée, âme, cœur, dieu, amour, prêtre, vie, grand, terre	Ballanche1809, Ballanche1829, Defontenay1854, Divoire1922, Esquiros1849, Grandmougin1882, Launay1922, Loyson1948, Schuré1889, Segalen1921
2.	Eurydice, récitatif, Pluton, Orphée, ardeur, écouter, tendre, trahir, arme, toujours	Courbois1710, Lacroix1717
3.	Orphée, Eurydice, Charon, lyre, Hadès, Aristée, Jason, Enfer, tête, Perséphone	Bottet2019, Eimann1854, Ferry2019, Goudot2003, Jimenes2004, Laffon2019, Montandre2013, Pommaux2009
4.	Arlequin, ah, scène, air, femme, petit, cela, Javotte, Roger, voilà	Anouilh1941, Cocteau1927, Crémieux1858, Deshayes1792, Dorvigny1775, Fuzelier1711, Regnard1689
5.	Âme, amour, Eurydice, mort, beau, tôt, dieu, toujours, mal, esprit	Anonyme1700, Buffequin1648, Chapoton1648, Duboullay1690, Espine1614, Hermite1641, Puget1624, Ronsard1563, Tournes1557
6.	Soumettre, bienfait, univers, belliqueux, mal, accent, Latone, loi, céleste, dieu	Berlioz1827, Berlioz1828, Guillion1770, Hus1759
7.	Nyango, Orphée, spectateur, Ndinga, père, salle, image, lumière, maison, ça	Liking1981, Manjock1981
8.	Vielle, danser, vielleur, danse, bal, burlesque, veuf, grotesque, jouer, femme	Martin1649, Martin1649-2
9.	Parmi, chercher, lumière, funeste, dedans, connaître, tribunal, accord, former, redemander	Hermite1652
10.	Amour, Eurydice, récitatif, air, tendre, cesser, cantate, sombre, amant, objet	Clerembault1710, Collande1752, Grandval1729, Lemaire1743, Moline1774, Piroye1717, Rameau1721, Regnard1699

L'observation des résultats pour les autres topiques confirme la pertinence accrue de ceux-ci, au moins en ce qui concerne la division automatique du corpus, par rapport à la méthode LDA. En effet, le regroupement des textes qui partagent le topique le plus saillant converge en grande partie avec notre classement manuel.

Le topique n° 1 regroupe dix œuvres qui insistent toutes sur l'aspect initiatique de la figure orphique. De genres différents (la répartition est égalitaire : quatre genres narratifs, trois poétiques et trois dramatiques), mais classés tous comme transpositions pour les adultes, ces textes sont publiés entre 1809 et 1922. Pour six d'entre eux, nous considérons que la transformation du mythe est idéologique. *Orphée-roi* (Segalen1921) est classé comme une modernisation et les trois poèmes homonymes, *Orphée* (Divoire1922), *Orphée : poème* (Launay1922) et *Orphée, poème* (Loyson1948) sont décrits comme les transmodalisations. Mais les résultats de la détection des topiques nous invitent à nous pencher davantage sur les liens qui peuvent être faits entre eux et les transformations idéologiques³⁵.

³⁴ Pour les références bibliographiques complètes et la catégorisation manuelle de chaque œuvre, voir I.2, « Établissement du corpus : entre l'exhaustivité quantitative et la représentativité herméneutique » p. 47 et le document « 1-1-3_Corpus-détail ».

³⁵ Voir IV.1c) « Exploitation des résultats enrichis et du réseau global des réécritures ».

Le topique n° 3 est le plus saillant majoritairement pour les œuvres destinées aux jeunes lecteurs (sept œuvres sur huit, soit l'ensemble du sous-corpus des livres de jeunesse : Bottet2019, Ferry2019, Goudot2003, Jimenes2004, Laffon2019, Montardre2013 et Pommaux2009). Le texte restant est la nouvelle « Une histoire renouvelée des Grecs » (Eimann1854), une des trois parodies du corpus pour laquelle le topique n° 4 n'est pas le plus caractéristique. Mises à part les deux parties de *L'Orphée grotesque avec le bal rustique* (Martin1649 et Martin1649-2) qui sont les uniques composants du topique n° 8, et *Orphée. Cantate avec symphonie* (Grandval1729) qui parodie la cantate *Orphée, cantate, à voix seule avec symphonie* (Clérembault1710) et se trouve donc dans le même ensemble que son hypotexte (topique n° 10), tous les autres détournements comiques sont des pièces dramatiques, pour la plupart des opéras, ce qui est visible parmi les mots-clés les plus saillants du topique n° 3 (*scène, air*). En revanche, les œuvres de jeunesse sont majoritairement des romans, et les mots-clés pour le topique n° 4 reflètent surtout la trame classique du mythe : nous y retrouvons des noms propres des protagonistes (*Orphée, Eurydice, Charon, Hadès, Aristée, Perséphone*) et des lieux où se déroule l'action (*Enfers*), des attributs principaux des personnages (*lyre*), voire des mots-clés de la trame mythique (*tête*). Dans Eimann1854, une traduction des *Géorgiques* de Virgile est longuement citée pour raconter le mythe. Il est intéressant de constater que le vocabulaire de la traduction en vers mêlé au style plutôt relâché de cette nouvelle humoristique a été rapproché par l'automate aux réécritures destinées au jeune public³⁶.

Le topique n° 5 regroupe les transmodalisations, tant poétiques que dramatiques, en moyen français. Le rapprochement entre ces œuvres est-il dû uniquement à la proximité de l'état de la langue ? Plusieurs facteurs nous incitent à considérer que la similarité entre les œuvres regroupées par le topique n° 5 concerne également le contenu. Tout d'abord, nous y retrouvons les premières pièces baroques qui inscrivent le mythe dans le registre tragique, comme *Le Mariage d'Orphée, sa descente aux Enfers* (Espine1623), la pièce à machines *La Grande journée des machines, ou le Mariage d'Orphée et d'Eurydice* (Chapoton1648) et son *Dessein du poème et des superbes machines du Mariage d'Orphée et d'Eurydice* (Buffequin1648) et, plus tardivement, l'opéra *Orphée, tragédie en musique* (Duboullay1690) et le *Ballet des machines représentant le cinquième acte de la mort d'Orphée et d'Euridice* (Anonyme1700). Ensuite, l'ensemble regroupe également trois œuvres poétiques : les quatre poèmes des *Métamorphoses d'Ovide figurée* (Tournes1557), *Orphée* (Ronsard1563) et *La lyre d'Orphée* (Hermite1641). Les huitains décasyllabes de Tournes1557 suivent le schéma des rimes très classique ABABBCBC. Les deux autres poèmes ont une structure comparable : longs, écrits en alexandrin et avec des rimes croisées, ils sont tous

³⁶ Nous abordons rapidement ce sujet au II.1d) « Théories intertextuelles et observation du corpus, p. 139.

les deux dédiés à un ami illustre (le magistrat Jacques Auguste de Thou pour Ronsard¹⁵⁶³, le chanteur castrat Blaise Berthod pour Hermite¹⁶⁴¹). Par ailleurs, plusieurs travaux académiques relèvent l'inspiration très probable de la poésie de l'Hermite par celle de Ronsard³⁷.

Deux autres topiques sont les plus caractéristiques pour d'autres œuvres de la même période : le topique n° 8 regroupe les deux parties de la parodie *L'Orphée grotesque avec le bal rustique* (Martin¹⁶⁴⁹ et Martin¹⁶⁴⁹⁻²) et le n° 9 est le plus caractéristique seulement de l'autre poème de Tristan l'Hermite *Orphée aux Enfers parlant à Pluton* (Hermite¹⁶⁵²). Ce dernier poème est un chant d'Orphée devant le roi des Enfers, le mythe n'est donc pas raconté dans sa totalité et l'accent est mis surtout sur la souffrance amoureuse d'Orphée, qui réclame son épouse. Les mots-clés les plus saillants du topique n° 8, comme *connaître*, *tribunal*, *accord*, *former*, *redemander*, montrent d'ailleurs une forte inspiration par la source ovidienne : comme dans *Les Métamorphoses*, *Orphée aux Enfers parlant à Pluton* prend, certes, pour motif dominant l'amour, mais se présente surtout sous forme d'un plaidoyer rhétorique et la demande est appuyée sur l'argumentation quasi-juridique. Tout ceci prouve que le regroupement ne dépend pas uniquement de l'état de la langue. Toutefois, pour améliorer davantage la pertinence du traitement effectué avec la méthode NMF, il paraît nécessaire de le refaire avec le corpus entièrement modernisé³⁸.

Enfin, le topique n° 10 regroupe les œuvres sentimentales, focalisées sur le thème de l'amour, à commencer par les opéras *Orphée aux enfers* (Regnard¹⁶⁹⁹) et *Orphée et Eurydice* (Moline¹⁷⁷⁴). Ce topique est également le plus caractéristique des cinq cantates sérieuses (Clérembault¹⁷¹⁰, Lemaire¹⁷⁴³, Piroye¹⁷¹⁷ et Rameau¹⁷²¹) et une parodique (Grandval¹⁷²⁹). Enfin, *Opuscule d'un célèbre auteur égyptien. Contenant l'histoire d'Orphée, par laquelle on pourrait soupçonner qu'il est peu de femmes fidèles* (Collande¹⁷⁵²) a été classé comme une transformation idéologique. Cependant, dans ce texte particulier, il ne s'agit pas tant d'insister sur l'aspect initiatique du mythe orphique, mais de se servir de la figure d'Eurydice et du thème de l'Amoureux dévoué pour démontrer l'infidélité des femmes.

Il est intéressant de constater d'ailleurs que, pour trois autres cantates, ainsi que pour le ballet héroïque *La Mort d'Orphée, ou les Fêtes de Bacchus* (Hus¹⁷⁵⁹) le topique le plus saillant est le n° 6. La proximité entre les trois cantates, au sein desquelles nous retrouvons des passages quasi identiques, prouve la pertinence de ce regroupement. De plus, les quatre ouvrages de cet ensemble abordent

³⁷ Voir par exemple Anne-Pascale Pouey-Mounou, « Du Bellay, Ronsard, Tristan : autour de quelques sonnets des *Amours* », *Cahiers Tristan L'Hermite*, n° XLIII, Paris, Classiques Garnier, « Tristan et la poésie du XVI^e siècle », 2021, p. 35-58, en ligne, <<https://classiques-garnier.com/cahiers-tristan-l-hermite-2021-n-xliii-tristan-et-la-poesie-du-xvie-siecle-du-bellay-ronsard-tristan-autour-de-quelques-sonnets-des-amours.html>>; François Rouget, « *Ordo neglectus* : remarques sur la disposition des recueils poétiques de Tristan L'Hermite », *Cahiers Tristan L'Hermite*, n° XLIII, Paris, Classiques Garnier, « Tristan et la poésie du XVI^e siècle », 2021, p. 151-181, en ligne, <https://classiques-garnier.com/doi/garnier?article=CatMS42_151>.

³⁸ Nous traitons des difficultés de la modernisation des graphies au sein de notre corpus au III. « Logiciels dédiés à la détection automatique des réemplois textuels ».

la question de la mort d'Orphée, contrairement à toutes les œuvres pour lesquelles le topique n° 10 est le plus saillant. Enfin, notons également que les deux œuvres du corpus qui se focalisent sur le mythe de l'Orphée noir, Liking1981 et Manjock1981, ont également été isolées des autres composants du corpus : pour elles, le topique le plus caractéristique est le n° 7.

c) Alignement automatique des corpus parallèles et la comparaison des versions³⁹

Des techniques et outils de plus en plus récents, de plus en plus performants ne cessent d'émerger, notamment en ce qui concerne l'alignement des corpus comparables ou parallèles et la traduction automatique. Cependant, ils demandent souvent des compétences techniques et informatiques qui les rendent difficilement exploitables par les chercheurs littéraires non informaticiens et ils s'appuient sur des corpus d'apprentissage très volumineux, dont nous ne disposons pas. Par exemple, pour aligner des corpus parallèles, comme *Digital Corpus of the European Parliament* (DCEP)⁴⁰, il est nécessaire de disposer des lexiques bilingues et d'utiliser de systèmes de traduction automatique⁴¹. Ce processus se compose généralement de trois étapes : (i) utilisation des métriques de similarité pour détecter des documents comparables au sein d'un corpus plus large ; (ii) utilisation des ressources existantes pour produire la liste des phrases ou segments candidats et (iii) sélection des propositions pertinentes en exploitant différentes mesures, modèles ou règles d'alignement⁴².

Méthode de Barzilay et McKeown

Cependant, alors que ces méthodes se montrent probantes dans de nombreux domaines du traitement automatique des langues⁴³, l'exploitation des outils d'alignement des traductions pour traiter un corpus monolingue (notamment les séries traductives) a été critiquée par exemple par Regina Barzilay et Kathleen McKeown, qui considèrent que

³⁹ Pour un autre exemple d'exploitation des logiciels de comparaison de versions, voir Karolina Suchecka, Nathalie Gasiglia et Karl Zieger, « Édition comparative intermédiaire de séries traductives : exploiter les homologies pour créer des visualisations modulables », *Traitement Automatique des Langues*, vol. 60, n° 3, « TAL et humanités numériques », décembre 2019, p. 37-62, en ligne, <https://www.atala.org/sites/default/files/TAL-60-3_SucheckaEtAl_Edition-ComparativeIntermedialeDeSeriesTraductives-final.pdf>. Certains développements proposés dans cette section sont ébauchés dans cet article.

⁴⁰ Voir Najeh Hajlaoui et al., « DCEP -Digital Corpus of the European Parliament », *Proc. LREC 2014*, Reykjavik, Iceland, 2014, p. 3164-3171.

⁴¹ Dans le cas du projet DCEP, il s'agit du logiciel d'alignement des sentences appelé Hunalign (<https://github.com/danielvarga/hunalign>). Voir Dániel Varga et al., « Parallel corpora for medium density languages », dans Nicolas Nicolov et al. (dir.), *Current Issues in Linguistic Theory*, vol. 292, Amsterdam, John Benjamins Publishing Company, 2007, p. 247-258, en ligne, <<https://benjamins.com/catalog/cilt.292.32var>>.

⁴² Voir Rémi Cardon et Natalia Grabar, « Construction d'un corpus parallèle à partir de corpus comparables pour la simplification de textes médicaux en français », *Traitement Automatique des Langues*, vol. 61, n° 2, 2020, p. 15-39.

⁴³ Voir notamment l'histoire de la traduction automatique très complète de Marianne Reboul (*Comparaison semi-automatique des traductions françaises de l'Odyssée d'Homère (1547-1955)*, Paris, Classiques Garnier, coll. « Cultures et pratiques savantes du numérique », 2022, p. 23-37.).

les corpus parallèles [des séries traductives de textes littéraires] sont très éloignés des corpus parallèles très propres habituellement utilisés dans le traitement automatique. La transposition d'un texte littéraire dans une autre langue n'inclut pas seulement la traduction, mais aussi la restructuration de cette dernière pour qu'elle soit conforme au style littéraire concerné.⁴⁴

Dans leur article, Barzilay et McKeown travaillent sur un corpus de onze traductions anglaises de *Madame Bovary* en s'appuyant effectivement sur la technologie de la traduction automatique (les phrases sont alignées et les schémas des occurrences sont extraits et liés à l'aide de mesures de corrélation), mais en adaptant ces méthodes à la spécificité de leur corpus. Dans un premier temps, les phrases correspondantes sont alignées en présumant que tout ce qui apparaît dans un contexte similaire serait une paraphrase (« il a essayé de la consoler » / « il a essayé de la réconforter »). Les contextes qui entourent les mêmes mots sont ensuite extraits et utilisés pour trouver de nouvelles paraphrases (« il a essayé de la... » suivi d'un verbe à l'infinitif). Enfin, les nouvelles paraphrases sont utilisées pour apprendre de nouvelles règles concernant le contexte (« il a tenté de la... », « il a fait de son mieux de la... »), et ainsi de suite.

Performante pour le corpus en anglais des traductions publiées aux intervalles proches, cette méthode est moins probante pour les séries traductives plus hétérogènes, comme notre sous-corpus des traductions d'Ovide et de Virgile. La syntaxe du moyen français diffère de celle du français moderne, comme les textes en prose sont difficiles à aligner avec les textes en vers. Par ailleurs, l'approche nécessite d'avoir beaucoup de mots en commun ou de préparer le corpus en amont, notamment en alignant manuellement les phrases. Alors que cela reste faisable pour une phrase de chaque texte⁴⁵, le traitement de la totalité des traductions d'Ovide et de Virgile qui composent notre corpus serait bien trop chronophage, sans pour autant garantir une amélioration significative des résultats d'identification des paraphrases. Enfin, dans beaucoup de cas, les mots ou suites de mots qui partagent les contextes identiques ne sont pas pour autant des paraphrases.

Prenons comme exemple deux phrases des traductions de Virgile en vers, celle de Delille en 1770 (Delille1770⁴⁶) d'une part et celle d'un « Naturaliste » en 1863 (Guillois1863⁴⁷) d'autre part (Tableau 26) :

⁴⁴ « [...] parallel corpus is far from clean parallel corpora used in MT. The rendition of a literary text into another language not only includes the translation, but also restructuring of the translation to fit the appropriate literary style » (nous traduisons). Voir Regina Barzilay et Kathleen R. McKeown, « Extracting Paraphrases from a Parallel Corpus », *Proceedings of the 39th Annual Meeting of the ACL*, Toulouse, Association for Computational Linguistics, 2001, p. 50-57, en ligne, <<https://www.aclweb.org/anthology/P01-1008>>.

⁴⁵ Voir chap. 2 « Procédés linguistiques propres des relations intertextuelles », p. 149.

⁴⁶ Virgile, *Les Géorgiques de Virgile, traduction nouvelle en vers françois*, trad. Jacques Delille, Paris, C. Bleuet, 1770, p. 233-239.

⁴⁷ Virgile, *Les Géorgiques, traduction vraie*, trad. Un naturaliste, Paris, Veuve Guillois, 1863, p. 65-68.

Tableau 26. Deux phrases correspondantes dans les traductions de Delille1770 et Guillois1863 : [blocs communs]^{2,4-6} et déplacements^{1,3}

Delille1770	Guillois1863
Près du <u>Strymon</u> ¹ glacé, / [dans] ² les <u>antres</u> ³ de [Thrace] ⁴ / durant sept [mois entiers] ⁵ / [il pleura] ⁶ sa disgrâce.	On dit que [dans] ² la [Thrace] ⁴ et le <u>Strimon</u> ¹ désert, / pendant neuf [mois entiers] ⁵ sous un <u>antre</u> ³ en hiver / [il pleura] ⁶ ; de son chant apprivoisant l'hyène !

Plusieurs difficultés peuvent être observées ici. Si nous suivons la méthode de Barzilay et McKeown, les mots qui partagent les mêmes contextes peuvent être considérés comme périphrases, ce qui fonctionne effectivement pour « dans les antres de Thrace » et « dans la Thrace », sauf que le nom « antre » est également présent dans la phrase de Guillois1863, mais avec les différences du contexte (« dans les antres de Thrace » / « sous un antre en hiver ») et de la flexion (singulier / pluriel). Par ailleurs, alors que, dans les segments « durant sept mois entiers » et « pendant neuf mois entiers » la synonymie entre *durant* et *pendant* devrait pouvoir être identifiée en traitant un corpus plus large, on ne peut pas considérer que *sept* et *neuf* sont des paraphrases⁴⁸.

Par ailleurs, nous observons également des modifications de structure plus générales : alors que Delille1770 précise qu'Orphée pleura « sa disgrâce », le segment qui suit « il pleura » chez Guillois1863 (« de son chant apprivoisant l'hyène ») n'est aucunement une paraphrase de « sa disgrâce », mais doit être lié plutôt avec la phrase suivante, « Adoucissant le tigre, attendrissant le chêne ; / tous en cercle accourus au charme de sa voix », qui correspondrait alors chez Delille1770 à la phrase « Sa voix adoucissoit les tigres des déserts, / Et les chênes émus s'inclinoient dans les airs ».

Enfin, mise à part la différence de graphie du nom du fleuve d'une traduction à l'autre (*Strymon* / *Strimon*), la position différente du mot au sein de la phrase empêche la mise en relation des syntagmes nominaux « Strymon glacé » et « Strimon désert ». En outre, la relation entre « glacé » et « désert » est assez complexe, car Delille1770, contrairement à Guillois1863, omet le qualificatif *deserti* épithète de *Strymon* dans le texte latin⁴⁹, qu'il remplace par « glacé » (qui décrit les antres chez Virgile) alors que le second traduit bien *deserti* par « désert » et n'introduit pas la notion de froid de *gelidis* qu'avec « en hiver ». Si nous voulions proposer un alignement complet des deux traductions, il nous faudrait donc pouvoir lier d'un côté *glacé* et, de l'autre, *désert* et *en hiver*.

⁴⁸ Le deuil d'Orphée dure sept mois dans la version originale : « Septem illum totos perhibent ex ordine menses, / Rupe sub aëria, deserti ad Strymonis undam, / Flevisse et gelidis hæc evolvisse sub antris » (Virgile, *Les Géorgiques*, IV, v. 507-509).

⁴⁹ Voir n. 48.

Par conséquent, alors que nous verrons que la méthode de Barzilay et McKeown s'avère exploitable pour une partie spécifique de notre recherche⁵⁰, nous constatons que, si elle est déjà problématique pour un couple de phrases pourtant assez proches, elle devient impossible pour des réécritures. En effet, la plupart des paraphrases que nous observons, que ce soit dans le corpus de traductions ou dans celui de réécritures, ne se déploient pas de manière linéaire et les déplacements sont nombreux. Nous avons besoin plutôt d'une méthode qui prend en compte les déplacements, suppressions ou insertions des segments pour proposer un alignement optimal des segments communs, comme celle implémentée dans les logiciels utilisés dans les études génétiques pour aligner les différentes versions d'une même œuvre.

MEDITE : logiciel d'alignement de versions textuelles

MEDITE⁵¹ est un des outils de ce type⁵². Il puise entre autres dans l'algorithme d'alignement par fragments grâce à la détection des homologues, une méthode de détection des séquences (utilisée initialement pour l'alignement des macromolécules – ADN ou protéines) afin de faire ressortir leurs régions homologues⁵³. Actuellement, il propose une comparaison de deux textes bruts : les blocs communs sont analysés et les différentes variantes sont signalées grâce à des codes de couleur. Les remplacements sont marqués en bleu, les insertions en vert, les suppressions en rouge et les déplacements en gris. Par défaut, le traitement est sensible à la casse, aux séparateurs et aux signes diacritiques. La longueur minimale des blocs communs est de cinq caractères. Pour que deux variantes soient considérées comme un remplacement, le ratio de la longueur des deux chaînes repérées doit être supérieur ou égal à 50 % (« abcd » et remplacé par « efgh », mais « a » est supprimé et « efgh » est inséré). Enfin, dans le cas de fortes densités des blocs communs et des variantes, les premiers sont insérés dans la variante si la différence de leur longueur par rapport à la longueur des variantes est supérieure ou égale à 50 %. Dans notre expérimentation, nous modifions uniquement le ratio des remplacements à 1 %, en considérant que les suppressions et les insertions sont des blocs qui n'ont pas leurs homologues dans l'autre texte. Les autres paramètres gardent les valeurs par défaut.

⁵⁰ Voir IV.2 « Annotation des entités nommées et leurs périphrases ».

⁵¹ Voir obvil.lip6.fr/medite/.

⁵² Nous privilégions MEDITE car il s'agit d'un développement français et qu'il est doté d'un interface graphique facilitant son exploitation. D'autres logiciels existent, par exemple Collatex (<https://collatex.net/demo/>). Voir aussi Gregor Middell, « Review of “Juxta Web Service, LERA, and Variance Viewer. Web based collation tools for TEF” », *RIDE*, vol. 11, 2020, en ligne, <<https://ride.i-d-e.de/issues/issue-11/web-based-collation-tools/>>.

⁵³ Voir Jean-Gabriel Ganascia et Jean Bourdaillet, « Alignements unilingues avec MEDITE », *Actes des Huitièmes Journées Internationales d'Analyse Statistique des Données Textuelles*, Paris, France, 2006, p. 427-437; Irène Fenoglio et Jean-Gabriel Ganascia, « Le logiciel MEDITE: approche comparative de documents de genèse », dans Aurèle Crasson (dir.), *L'édition du manuscrit - De l'archive de création au scriptorium électronique*, vol. 10, 2008, p. 209-228; Jean-Gabriel Ganascia, « MEDITE - A Unilingual Text Aligner for Humanities. Application to Textual Genetics and to the Edition of Text Variants », *Supporting Digital Humanities (SDH 2011)*, Copenhague, 2011.

Très pertinent pour la recherche génétique et l'alignement des différentes versions du même texte et se différenciant des autres outils de ce type notamment par sa capacité à détecter des déplacements, MEDITE est largement exploité dans de nombreux projets de recherche aux objectifs très variés, comme, notamment, l'ANR Phoebus – eBalzac⁵⁴ ou HyperApollinaire⁵⁵. Outre les projets de génétique littéraire, MEDITE est également exploité pour l'enseignement, par exemple dans le cadre du projet ANR JCJC « Compétences et difficultés des élèves en matière d'écriture à l'entrée du collège (ECRICOL) »⁵⁶. François-Xavier Guerry⁵⁷, quant à lui, diversifie ainsi les attendus : à partir des résultats MEDITE, il élabore une liste des mots-clés présents dans les variantes détectées entre les textes et constate, par exemple, que « [...]l'intérêt de MEDITE ne consiste pas tant dans sa capacité de mettre à jour une homologie isolée, qu'à fournir une sorte d'index exhaustif et facilement lisible de toutes les récurrences lexicales entre deux textes »⁵⁸.

Deux textes de notre corpus de réécritures se prêtent particulièrement bien à l'expérimentation avec MEDITE. Il s'agit des deux versions de la cantate *Orphée* d'Hector Berlioz, l'une publiée en 1827 dans *Séance publique de l'Académie Royale des Beaux-Arts, du 6 octobre 1827* (Didot, 1827, Berlioz1827), et l'autre tirée du manuscrit en partie autographe conservé au département de Musique de la BnF sous la cote RES VMA MS-1. Le catalogue de la BnF établit 1827 comme la date de production du manuscrit, mais l'indication « ouvrage déclaré inexécutable par la Section de musique de l'institut et exécuté à l'école royale de musique le 22 Juillet 1828 » à la page du titre⁵⁹ montre que cette version est postérieure à celle exécutée à la Séance publique de l'Académie Royale des Beaux-Arts. Pour marquer cette différence, nous l'identifions en tant que Berlioz1828.

Les résultats de la comparaison de ces deux versions avec MEDITE sont satisfaisants⁶⁰. Les blocs communs sont nombreux et les variantes sont faciles à identifier. Nous retrouvons principalement des différences formelles, comme :

⁵⁴ Voir notamment <https://ebalzac.com/genetique> et Karolina Suchecka, Victoria Le Fournier et Andrea Del Lungo, « Phoebus e-Balzac : édition numérique exhaustive d'un monument littéraire », dans Clarisse Bardiot, Esther Dehoux et Émilien Ruiz (dir.), *La fabrique numérique des corpus en sciences humaines et sociales*, Villeneuve d'Ascq, Presses universitaires du Septentrion, coll. « Humanités numériques et science ouverte », 2022, p. 278-296, en ligne, <http://www.septentrion.com/html/WYSIWYGfiles/files/CONTENEUR2108_Bardiot_Dehoux_Ruiz.pdf>.

⁵⁵ Voir Olivier Gallet et al., « Apollinaire numérique », *Revue d'histoire littéraire de la France*, vol. 116, n° 3, 2016, p. 533-546.

⁵⁶ Jacqueline Lafont-Terranova et al., « Modéliser le processus d'écriture d'un scripteur de haut niveau : intérêt et limites du repérage automatique des opérations de réécriture à l'aide du logiciel MEDITE », *L'écriture, ses supports, ses archives*, 2017.

⁵⁷ François-Xavier Guerry, « Góngora et ses premiers biographes : une analyse comparative moyennant des outils numériques », *e-Spania. Revue interdisciplinaire d'études hispaniques médiévales et modernes*, n° 29, 2018, en ligne, <<https://journals.openedition.org/e-spania/27502>>.

⁵⁸ *Ibid.*, paragr. 17.

⁵⁹ Voir <https://gallica.bnf.fr/ark:/12148/btv1b55006377x/f9.item>.

⁶⁰ Voir le contenu du dossier [II_3_medite](#) et, plus particulièrement, le résultat obtenu avec l'interface Web de MEDITE ([résultats/berlioz1827_berlioz1828.html](#)) et la transcription tabulée de ce résultat ([medite_resultats-alignes.pdf](#)). Ce traitement peut être reproduit en utilisant les fichiers texte du dossier [versions_texte](#).

- des différences de ponctuation : suppression de virgules et de points de suspension (« Eurydice, » / « Euridice » ; « Ô ciel !... » / « Ô ciel ! »), remplacement de point-virgule par une virgule (« délire ; » / « délire ») ou par des points de suspension (« arrêtez » / « arrêtez... »), etc., et
- des modifications de graphies⁶¹ (« Eurydice » / « Euridice », « accents » / « accens », « brillants » / « brillans », « secrette » / « secrète », « cymbales » / « cimballes », « menaçants » / « menaçans » et « Apollon » / « Appollon »).

Par ailleurs, des didascalies sont ajoutées dans la version Berlioz1828, tant en ce qui concerne les personnages qui chantent (« ORPHÉE », « BACCHANTES ») que la manière dont les répliques sont performées (« *(ensemble)* », « *(à voix étouffée)* », « *La voix doit s'éteindre presque entièrement sur la troisième syllabe du mot Euridice* »). Un seul remplacement est lexical (et syntaxique) : « Quel bruit affreux se fait entendre » est remplacé par « Quels cris affreux se font entendre ». Nous notons également une permutation de deux vers : « Apollon du haut de ton trône / Lance tes feux, je n'ai d'espoir qu'en toi » devient « Entends-moi, je n'ai d'espoir, je n'ai d'espoir qu'en toi. / Lance tes feux du haut de ton trône, Apollon ». Cette modification formelle n'est pas bien reconnue par MEDITE, qui trouve un seul bloc commun pour ces deux vers : « du haut de ton ». Ainsi, « Apollon » devient le remplacement de « Entends-moi, je n'ai d'espoir, je n'ai d'espoir qu'en toi. / Lance tes feux » et « trône, / Lance tes feux je n'ai d'espoir qu'en toi » remplacerait « trône, Apollon » (Tableau 27).

Tableau 27. Comparaison avec MEDITE : Berlioz1827 vs Berlioz1828.

Berlioz1827	Berlioz1828
Apollon du haut de ton trône, Lance tes feux, je n'ai d'espoir qu'en toi.	Entends-moi, je n'ai d'espoir, je n'ai d'espoir qu'en toi. Lance tes feux du haut de ton trône, Apollon.

Enfin, deux variantes sont bien plus substantielles :

- 33 vers de la version Berlioz1827 sont supprimés dans la version Berlioz1828 (de « Art séduisant, plaisir d'innocence, » à « Tout cède aux accents de ma voix ») et
- 18 vers du chant des Bacchantes sont insérés à la fin de la version Berlioz1828.

L'alignement automatique de ces deux versions produit par MEDITE est plutôt satisfaisant, sauf pour le cas déjà commenté de permutation des vers et en ce qui concerne le traitement des déplacements, que le logiciel détecte principalement au sein des longs blocs de suppression et d'insertion. Il s'agit tant des mots faibles (« et », « son » et « les »), que des mots forts « gloire » et « victoire ».

⁶¹ Remarquons que, malgré la proximité présumée de la création des deux versions, ces variantes peuvent être expliquées par de nombreuses réformes d'orthographe ayant lieu en cours du XIX^e siècle et le fait que nous comparons un imprimé avec un manuscrit. Nous n'excluons pas cependant la mise en cause de la chronologie de l'œuvre que nous établissons.

La présence des mots « gloire » et « victoire » tant dans l'extrait supprimé que celui inséré est intéressant étant donné que ce premier décrit les pouvoirs pacificateurs du chant orphique⁶² et le deuxième parle de la mort d'Orphée⁶³, mais il n'en reste pas moins que la détection des déplacements n'est pas optimale, alors que cette fonctionnalité est l'une des particularités du logiciel.

La pertinence des résultats tend à décroître rapidement si la différence entre deux textes est plus importante. Voyons cela en comparant la première version de la cantate de Berlioz (Berlioz1827) avec celle d'Albert Guillion, intitulée *Orphée. Cantate composée à l'occasion de la fête de M. Henry Monton Berton, membre de l'Institut et chevalier de la Légion d'honneur*. Le manuscrit à partir duquel nous transcrivons le texte se trouve au département de Musique de la BnF sous la cote D-5210 et sa date de production est estimée entre 1770 et 1799 (Guillion1770). Les paroles de la cantate de Guillion1770 sont très similaires de celles du segment supprimé de Berlioz1827, mais nous ne pouvons plus considérer qu'il s'agisse de deux versions d'une même œuvre : les paraphrases et remplacements lexicaux sont bien plus nombreux (« divins accords » / « brillants accords », « tous animés par tes sons belliqueux » / « Sont animés par tes chants belliqueux ») et plus complexes (« Sans toi, sans toi, les mortels vertueux qui consacrerait la mémoire » devient « Sans ton pouvoir, qui pourrait du génie, / Des mortels vertueux, / Consacrer la mémoire »). Tout cela pose davantage de problèmes à MEDITE. Voyons deux exemples.

Dans l'exemple de l'alignement présenté au Tableau 28, la proximité entre les deux extraits est avérée, mais la réécriture reste importante. MEDITE détecte correctement cinq blocs communs : « Retentis dans ces bois, », « mes vers. », « seconde »⁶⁴, « que tes » et « accords étonnent l'univers ». En revanche, l'identification et la classification des variantes posent quelques problèmes.

Tableau 28. Comparaison avec MEDITE : Guillion1770 vs Berlioz1827. Premier exemple

Guillion1770	Berlioz1827
	Ô seul bien qui ³ me reste ! ô ma céleste lyre !
Retentis dans ces bois, que tes sons, ô ¹ ma lyre, Unis aux chants d'Orphée embellissent mes vers.	Retentis dans ces bois, viens embellir ¹ mes vers.
Seconde l'ardeur ² qui ³ m'inspire.	Ranime mes accents, seconde mon délire ; ²
Que tes divins ⁴ accords étonnent l'univers.	Que tes brillants ⁴ accords étonnent l'univers.

Tout d'abord, la segmentation des variantes faite par MEDITE est surprenante : il considère qu'il y a un remplacement entre « que tes sons, ô » et « viens embellir », puis le déplacement de « ma lyre », suivie de la suppression « Unis aux [chants] d'Orphée embellissent ». Ainsi, le déplacement « ma lyre » renvoie à un vers de Berlioz1827 situé bien plus loin dans le texte (« Mais tout s'émeut

⁶² « Les guerriers avides de gloire / Sont animés par tes chants belliqueux, / Et les chants de victoire / Sont le prix du courage et des cœurs généreux »

⁶³ « C'est un hymne à ta gloire ! / Ô Bacchus ! Évoé / [...] / Meurs ! Meurs ! Meurs ! / Victoire ! »

⁶⁴ Les différences de casse ne sont pas prises en compte, ce qui permet d'aligner « Seconde » et « seconde ».

aux accords de ma lyre »), alors que nous aurions préféré que la proximité entre « ô ma lyre » et « ô ma céleste lyre » puisse être prise en compte. Plus généralement, les deux déplacements qui pointent vers d'autres extraits (« ma lyre » et « chants ») rendent difficile la compréhension de l'alignement proposé. Notons enfin que la proximité entre « embellissent » et « embellir » est ignorée. Par ailleurs, le découpage de « l'ardeur qui m'inspire » en trois variantes différentes – remplacement, déplacement, puis suppression – rend les résultats davantage bruyants : au lieu de noter la répétition peu significative de « qui » dans les contextes très différents (« l'ardeur qui m'inspire » / « seul bien qui me reste »), nous aurions préféré avoir un remplacement entre « l'ardeur qui m'inspire » et « mon délire », qui mettrait mieux en lumière cette simplification intéressante de Berlioz1827.

Dans le couple suivant (Tableau 29), nous observons que la réécriture de Berlioz1827 a introduit un changement d'ordre de mention de deux animaux, le tigre et le lion, en plus d'autres modifications assez importantes.

Tableau 29. Comparaison avec MEDITE : Guillion1770 vs Berlioz1827. Second exemple

Guillion1770	Berlioz1827
Tigre a perdu sa fureur, Il cède au charme ¹ qui l'entraîne. Le lion ² des agneaux a toute la douceur. Prodige étonnant et ³ céleste harmonie, ⁴	Le lion ² à mes pieds reconnaît un vainqueur, Et le tigre indomptable a perdu sa fureur. Ô ¹ prodige étonnant ! ô ³ céleste harmonie ! ⁴

Cela perturbe fortement MEDITE, même s'il détecte bien le déplacement du syntagme « Le lion ». En effet, si nous observons la correspondance au niveau des remplacements, nous réalisons que « Il cède au charme » de Guillion1770 a été considéré comme remplacement du « ô » de l'exclamation « Ô prodige étonnant ! » chez Berlioz1827. Idéalement, nous aurions aimé aligner « des agneaux a toute la douceur » avec « à mes pieds reconnaît un vainqueur », mais cela est ambitieux pour un traitement automatique, car aucune relation syntaxique, lexicale ou sémantique ne peut être observée ici. *A minima*, il faudrait que le segment « Il cède au charme qui l'entraîne » soit considéré comme suppression chez Guillion1770, et que l'interjection « ô » soit marquée comme une insertion chez Berlioz1827.

En somme, le traitement de variantes proposé par MEDITE est intéressant, mais n'est pas adapté au corpus de réécritures, au sein duquel les textes entretiennent des relations plutôt distendues. Son fonctionnement correspond toutefois partiellement à nos attentes. Tout d'abord, contrairement aux outils d'aide à la traduction, MEDITE ne se restreint pas à l'alignement des phrases, mais détecte et aligne les blocs communs. Ensuite, les différences détectées entre deux textes sont typées, même si seulement quatre types de modifications sont proposés (suppression, insertion, remplacement, déplacement), et que la détection n'est pas toujours optimale. À condition d'avoir

de résultats satisfaisants, ce classement permet d'accéder rapidement aux informations détaillées concernant la relation entre deux textes. Enfin, l'interface graphique facilite l'exploration et l'analyse des résultats.

En revanche, certaines fonctionnalités supplémentaires semblent nécessaires pour mieux comparer deux textes plus différents que deux versions d'une même œuvre. Tout d'abord, il serait utile de pouvoir effectuer la comparaison sur les lemmes (en prenant en compte, par exemple, la proximité entre « embellissent » et « embellir »). Par ailleurs, la prise en compte des catégories grammaticales et la flexion rendrait possible une spécification encore plus poussée (comme « remplacement flexionnel présent, 3^e pers. pl. / infinitif »). Plus encore, pour de textes comparables, mais très différents en surface, comme ceux de notre corpus, il importe de traiter le contenu, mais aussi d'autres indices qui permettraient d'intégrer dans la comparaison :

- les entités nommées et leurs périphrases (« ô ma lyre » / « ô seul bien qui me reste »),
- les interprétations du récit source qui affectent tant le contenu que les protagonistes invoqués (par exemple, pour ne pas chercher des proximités entre l'épisode de l'enchantement des arbres et des animaux et celui du meurtre d'Orphée par les Bacchantes),
- la synonymie (« ardeur » / « délire ») et l'hyperonymie (« sons » / « chants »),
- la substitution synthétique (« rendre plus doux » / « adoucir »),
- ou encore la métaphorisation (« Il cède au charme qui l'entraîne » / « à mes pieds [il] reconnaît un vainqueur »), etc.

Enfin, nous avons vu que le texte de Berlioz¹⁸²⁷ peut être mis en relation au moins avec deux autres textes de notre corpus (Berlioz¹⁸²⁸ et Guillion¹⁷⁷⁰). Ainsi, nous avons besoin d'un outil capable de comparer, voire d'aligner plusieurs textes à la fois, quand c'est pertinent.

Aligneur Odysseus

Le projet qui a réussi à atteindre l'objectif ambitieux de relier efficacement les méthodes de l'alignement automatique des traductions et celles de comparaison automatique des versions d'une même œuvre a été mené par Marianne Reboul dans le cadre de sa thèse *Comparaison semi-automatique des traductions en langue française de l'Odyssée d'Homère (1547-1955)*⁶⁵. Il s'agit, à notre sens, d'un des développements majeurs dans le domaine des humanités numériques littéraires en France. En effet,

⁶⁵ Marianne Reboul, *Comparaison semi-automatique des traductions françaises de l'Odyssée d'Homère (1547-1955)*, op. cit.

l'aligneur Odyseus⁶⁶ conçu par Reboul permet de comparer automatiquement 26 traductions françaises des 24 premiers chants de *L'Odyssée* d'Homère et d'afficher ces résultats au sein d'une interface graphique interactive⁶⁷.

Après l'étape de prétraitement du corpus (tokenisation de textes, lemmatisation, étiquetage, puis export en XML), Reboul segmente son corpus selon un critère assez intéressant. En effet, cette dernière a réalisé que « deux éléments étaient, lorsqu'ils se trouvaient dans le texte-source, très souvent reproduits par les traducteurs dans leurs traductions : il s'agit de noms propres, et, dans une moindre mesure, des pronoms »⁶⁸. Ainsi, chaque nouveau segment commence avec un pronom ou un nom propre (« [1] Muse, dis-nous les maux nombreux, les longues souffrances du divin [2] Ulysse au retour de [3] Troie la ville sacrée »). Reboul applique ensuite l'algorithme de Needleman-Wunsch (qui est conçu initialement pour aligner des séquences de protéines et de nucléotides) pour aligner les différents segments de chaque traduction entre eux. Peu complexe à programmer, cet algorithme permet de faire des calculs rapidement et de prendre en compte “des trous”, c'est-à-dire des séquences qui n'ont de correspondance ni dans la source ni dans la cible. Nous voyons par exemple en Tableau 30 que le nombre de segments et la manière dont ils sont divisés peuvent diverger. C'est grâce à l'algorithme de Needleman-Wunsch qu'il est possible de trouver l'alignement optimal, tout d'abord des segments qui commencent par *Muse* dans les trois traductions, puis du segment qui commence par *Troie* ou *Ilion*, son autre nom.

Tableau 30. Segmentation et alignement de trois traductions de L'Odyssée

1897, Calbet Rosny	1955, Philippe Jaccottet	1991, Frédéric Mugler
	[1] <u>Ô</u>	
[1] <u>Muse</u> , dis-nous les maux nombreux, les longues souffrances du divin	[2] <u>Muse</u> , conte moi l'aventure de l'inventif, celui qui pillà	[1] <u>Muse</u> , dis-moi l'homme inventif, qui erra si longtemps, Lorsqu'il eut renversé les murs de la sainte
[2] <u>Ulysse</u> au retour de		
[3] <u>Troie</u> la ville sacrée [...]	[3] <u>Troie</u> [...]	[2] <u>Ilion</u>

Grâce aux dictionnaires spécifiques, le développement de Reboul prend en compte différentes graphies des noms propres (« Telemache » / « Télémaque ») et les équivalents dans les mythologies grecque et romaine (« Athéna » / « Minerve »⁶⁹). Par ailleurs, une base de données lexicale conçue pour le grec (Greek WordNet⁷⁰) permet de trouver l'équivalence française d'un nom propre en grec

⁶⁶ Marianne Reboul, *OdyseusPolymetis/odyseus_interface: open-source code for publication* (PhD), Zenodo, 2021, en ligne, <<https://zenodo.org/record/5088159>>.

⁶⁷ Voir https://odyseuspolymetis.github.io/odyseus_interface/.

⁶⁸ Marianne Reboul, *Comparaison semi-automatique des traductions françaises de l'Odyssée d'Homère (1547-1955)*, op. cit., p. 62.

⁶⁹ Pour concevoir cette lexique spécifique, Reboul recourt aux méthodes basées sur la sémantique distributionnelle, comme la vectorisation sémantique. Nous présentons davantage cette méthode au III « Logiciels dédiés à la détection automatique des réemplois textuels ».

⁷⁰ Voir <https://greekwordnet.chs.harvard.edu/>.

ancien (« Ζεὺς » / « Zeus »). Nous voyons en revanche, toujours grâce à l'exemple du Tableau 30, que les périphrases des noms propres ne sont pas prises en compte : le segment qui commence par « Ulysse au retour de » chez Calbet Rosny aurait pu être aligné avec « l'inventif, celui qui pilla » chez Jaccottet et « l'homme inventif, qui erra si longtemps » de Mugler.

Pour examiner l'application de cette méthode à notre projet, prenons comme exemple, cette fois-ci, notre sous-corpus des traductions françaises des *Géorgiques* de Virgile, et plus particulièrement une phrase source latine riche en noms propres et pronoms, « Taenarias etiam fauces, alta ostia Ditis, et caligantem nigra formidine lucum ingressus est, Manesque adiit regemque tremendum nesciaque humanis precibus mansuescere corda » (4,460). Cette phrase contient trois noms propres : **Tāenāriūs**, *a, um* (syntagme nominal « Taenarias fauces », litt. « gorge de Ténare »), **Dīs**, *Ditis* (syntagme nominal « ostia Ditis », litt. « porte de Dis ») et **Mānēs**, *ium* (« Manesque », accusatif pluriel). Les traductions des verbe « adiit » (**āđēō**, *is, ire, ii, itum*, « aller ») et « ingressus est » (**īngřēđiōr**, *eris, i, gressus sum*, « entrer ») sont également susceptibles d'introduire le pronom personnel « il ». Nous pouvons donc prévoir que les traductions les plus fidèles de cette phrase seront partagées en 5 segments, correspondant aux segments latins suivants : (i) « **Teanarias** etiam fauces, alta ostia » ; (ii) « **Ditis**, et caligantem nigra formidine lucum » ; (iii) « **ingressus est**, » ; (iv) « **Manesque** » ; « **adiit** regemque tremendum nesciaque humanis precibus mansuescere corda ».

Nous constatons que, effectivement, 5 segments en moyenne ont été détectés pour les 14 traductions prises en compte⁷¹ (7 maximum et 3 minimum). 6 traductions sur 14 sont divisées en 5 segments (sans obligatoirement respecter la même segmentation que celle de l'original). Nous présentons l'alignement obtenu pour notre corpus dans le Tableau 31 où les segments correspondants sont mis en valeur à l'aide des codes-couleur.

⁷¹ Nous excluons de cette expérimentation particulière la traduction en ancien français, Tours1540 (Virgile, « Géorgiques », dans Michel de Tours (dir.), *Les oeuvres de Virgile. Translatees de latin en [vers] francoyses*, Paris, Jean Petit, Pierre Vidoue, 1540, p. 19-29.) : « Jusques aux lieux et obscures cavernes de Tenarus et trop tenebreux cernes il descendit où il trouva les portes du roy Dyty et Pluto assez fortes que dieu d'enfer est nommé, si fort noires que de clarté bien se monstroient contraires. Par les forests paoureuses il passa, pleines d'horreur et umbres. Tant trassa son bas chemin et dur pelerinage, les lieux trouva où estoit le mesnage des esperitz et ames infernales qui la estoient en leurs chambres trop palles. Le roy trouva de celle region qu'onques vivant par sa precaton n'a sceu freyer, et humaines prières tant fussent or dignes et singulieres. ». Cette phrase étant très différente des autres traductions, notamment du point de vue de la syntaxe, et peu fidèle à la phrase source, il nous a paru erroné de l'inclure dans une expérimentation portant sur des tendances globales observables au sein du sous-corpus des traductions publiées majoritairement à partir du XVIII^e s.

Tableau 31. Alignement des 14 traductions d'une phrase de Virgile selon la méthode Odysseus

Identifiant	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5	Segment 6	Segment 7
Cogolin1750	Aux gouffres des	Enfers	il dirige ses pas, affronte sans effroi l'empire du trépas,	Il ose pénétrer ces forêts ténébreuses, qu'une éternelle horreur rend encor plus affreuses	Il ose avec transport s'approcher de	Pluton , dans son antre aborder la barbare	Alecton
Delille1770	C'est peu : malgré l'horreur de ces profondes voûtes,	Il franchit de l'	Enfer les formidables routes, et perçant ces forêts où règne un morne effroi,	Il aborda des Morts l'impitoyable Roi.			
Cournand1805	Il affronta les antres du	Ténare , les	Mânes , le couroux de leur maître barbare, l'épouvantable nuit des	Enfers et ces cœurs qui ne sauraient fléchir la prière et les pleurs.			
Desfontaines1810	Il osa même descendre dans les gouffres du	Ténare , pénétrer dans le royaume profond de	Pluton , y traverser ces forêts ténébreuses où règne un éternel effroi, s'approcher du terrible monarque des morts, et aborder ces lugubres divinités, que les prières des mortels n'ont jamais fléchies.				
Heguïn1827	Il fit plus :	il osa franchir les gouffres du	Ténare , ses forêts noires d'épouvante, et le seuil du palais de	Pluton ; aborder les	Mânes et leur formidable roi, et tenter de fléchir ces cœurs impitoyables que des prières		

Intertextualité : de la théorie littéraire aux méthodes computationnelles

					humaines n'ont jamais fléchis.		
Charpentier1831	Les gouffres du	Ténare , les abîmes de	Pluton , et ces bois remplis d'une sombre horreur	il osa tout affronter	il aborda les	Mânes et leur affreux monarque,	il parut devant ces divinités qui ne savent point s'attendrir aux prières des mortels.
Duchemin1837	Il osa de	Pluton affronter le séjour, Les gouffres infernoux, leurs bois noirs d'épouvante, des	mânes aborder la foule pâlissante. Leur tyran formidable, et supplier des cœurs, Que n'ont jamais touchés ni prières, ni pleurs.				
Desportes1846-2	Il osa même affronter les gouffres du	Ténare .	Il descendit dans les abîmes de	Pluton ; et, traversant de vastes forêts, noir séjour de la crainte,	il aborda les	Mânes , et parut devant l'affreux monarque, devant ces fières divinités que n'attendrissent jamais les prières des mortels	
Desportes1846-1	Entrant même dans la gorge du	Ténare , porte profonde du royaume de	Pluton , et dans le bois obscurci par une crainte ténébreuse (des ténèbres effrayantes)	Il alla trouver les	Mânes , et leur roi redoutable, et des cœurs qui ne savent pas s'adoucir aux prières humaines.		
Guillois1863	Il fait plus,	il s'enfonce en l'horreur des forêts, descend droit au	Ténare , et dans son noir palais aborde le tyran, les	mânes redoutables ; Tous cœurs durs et fermés aux pleurs des misérables !			
Nisard1868	Il osa même descendre	Ténare ;	il vit le profond royaume de	Pluton , et ses bois que remplissent	enfers , et ces cœurs que les prières des		

	dans les gouffres du			l'horreur et les ténèbres, et les	humains n'ont jamais attendris.		
Fournier1876	Sous le	Ténare même, affrontant des ténèbres affreuses et sans lendemain, du séjour de	Pluton , parmi des bois funèbres,	Il osa franchir les chemins ;	il parut devant le roi terrible et vit l'assemblée impassible des dieux sourds aux vœux des humains.		
Cabaret- dupaty1897	Il ose pénétrer jusqu'aux portes du	Ténare , dans les sombres demeures de	Pluton , et, traversant ces bois où règne une ténébreuse horreur,	il aborda les	Mânes et leur roi formidable, et ces divinités que les pleurs des humains n'ont jamais attendries		
Rat1932	Il entra même aux gorges du	Ténare , portes profondes de	Dis , et dans le bois obscur à la noire épouvante, et	il aborda les	Mânes , leur roi redoutable, et ces cœurs qui ne savent pas s'attendrir aux prières humaines.		

Alors que nous retrouvons effectivement dans les traductions les noms propres présents dans la phrase originale, leur fréquence est très variable : le segment commençant par « Ténare » (extraits soulignés en orange) est bien présent dans 11 traductions sur 14, « Mânes » (en jaune) apparaissent dans 9 sur 14 textes et seulement une traduction contient le segment débutant par « Dis » (Rat1932, en gris). La plupart des traducteurs omettent, tout simplement, l'incise « *alta ostia Ditis* » de l'original, mais d'autres ont recours à une périphrase sans doute plus explicite pour leurs lecteurs. Ainsi, la porte de Dis devient « le royaume profond », « les abîmes », « le séjour » ou encore « les sombres demeures » de « Pluton ». Il s'agit ici d'une dénomination à peu près équivalente, Dis étant en quelque sorte le prédécesseur moins connu de Pluton¹. Nous observons par ailleurs que la fréquence des segments commençant dans les traductions par « Pluton » est très haute (10/14). En revanche, dans une traduction, la périphrase « séjour de Pluton » remplace la mention de Ténare (Duchemin1837, « Il osa de Pluton affronter le séjour, / Les gouffres infernaux... ») alors qu'une autre (Cogolin1750) fait apparaître son nom pour transposer la périphrase « *regemque tremendum* » (litt. « roi redoutable » / « roi effroyable »). Cette traduction en vers propose d'ailleurs toute une série de choix intéressants :

Aux [gouffres des Enfers]^{Taenarias fauces} il dirige ses pas,
 Affronte sans effroi [l'empire du trépas]^{ostia Ditis},
 Il ose pénétrer ces forêts ténébreuses
 Qu'une éternelle horreur rend encore plus affreuses
 Il ose avec transport s'approcher de [Pluton]^{regemque tremendum},
 Dans son antre aborder [la barbare Alec-ton]^{Manesque ?}.

Pour chaque nom propre ou périphrase de l'original, Cogolin1750 nous donne une traduction alternative, généralement conforme au lieu ou à la personne initialement désignée dans l'original, sauf pour « la barbare Alec-ton » pour laquelle la mise en parallèle avec « Mânes » semble peu probable. Ainsi, Ténare est remplacé par les Enfers (deux autres traductions font le même choix), la porte de Dis devient la périphrase plus générale « l'empire du trépas » qui renvoie désormais plutôt à un lieu qu'à une personne. Enfin, Pluton est mentionné directement et non plus par périphrase (qui désignait « Dis » dans le texte source).

La complexité des dénominations et leurs référents dans cet extrait est transposable, même si ce n'est qu'à moindre mesure, à la totalité des traductions prises en compte pour cette expérimentation. Ce flottement dénomiatif fait que la segmentation appuyée sur les noms propres ne semble

¹ « *Dis* ou *Ditis*. [...] De tous les grands dieux de Rome, *Dis pater* est peut-être celui dont il est le moins parlé, et auquel on a élevé le moins d'autels et dédié le moins d'inscriptions : soit que les Romains se soient représenté son existence comme immobile et immuable, soit qu'ils aient préféré s'adresser aux dieux Mânes plutôt qu'à leur chef, soit encore que les prêtres aient de bonne heure abandonné volontiers le vieux nom de *Dis pater* pour lui substituer celui de Pluton, la divinité similaire importée de Grèce. » (Charles Victor Daremberg et Edmond Saglio, « *Dis pater* », dans *Dictionnaire des antiquités grecques et romaines*, 1892.).

pas optimale : en effet, nous constatons que la tâche d'alignement des segments correspondants selon ces critères, déjà assez complexe pour un lecteur humain, serait très certainement impossible pour la machine. Il faudrait en effet concevoir des lexiques d'équivalences dénominatives extrêmement détaillées, et arriver à apprendre au logiciel de détecter les périphrases de ces noms propres et d'identifier leur référent. Mises à part les questions techniques, cela poserait également des questions épistémologiques, par exemple : Dis doit-il être considéré comme dénomination équivalente à Pluton, au même titre que Hadès ? Faut-il prévoir un classement plus complexe qui permet de nuancer ces équivalences ? Le cas échéant, selon quels critères ?

À ces problèmes s'ajoutent ceux, encore plus nombreux, de la prise en compte des pronoms personnels. Les choix traductionnels sont extrêmement hétérogènes et il est souvent difficile d'établir le lien exact avec l'original. Ainsi, 26 segments au total commencent par le pronom personnel « il ». Dans la phrase source, deux verbes sont conjugués à la 3^e personne du singulier, « adiit » susceptible d'être traduit par « il alla vers », et « ingressus est », « il entra ». Or, pour « adiit », une seule traduction a recours au verbe « aller », et seulement en tant que semi-auxiliaire, dans la construction « il alla trouver » (Desportes1846-vers). Pour la plupart, les traducteurs préfèrent « il aborda » (5 segments équivalents sur 9), mais trois d'entre eux font des choix plus originaux : « il ose avec transport s'approcher de... » (Cogolin1750), « il vit... » (Nisard1868) et « il parut devant [...] et vit » (Fournier1876).

Les choix traductionnels sont encore plus hétérogènes pour « ingressus est », qui apparaît dans la phrase originale après trois syntagmes nominaux. La plupart de traducteurs choisissent de commencer la phrase par le pronom personnel « il » suivi d'un verbe conjugué, même si 5 d'entre elles s'efforcent à suivre davantage l'ordre syntaxique de l'original (en particulier Charpentier1831, Desportes1846-vers et Fournier1876, à moindre mesure Cogolin1750 et Delille1770). En revanche, seulement deux traducteurs choisissent l'équivalent « entrer » : Rat1932 emploie la forme conjuguée « il entra même » et Desportes1846-vers a recours à un participe présent « entrant même ». Les autres traductions insistent davantage sur l'audace du poète, qui n'est présente dans l'original que par l'adverbe *etiam* (« même », traduit dans 5 extraits sur 14). Il s'agit, pour la plupart, du choix de la structure « il osa » suivie de différents verbes à l'infinitif : nous recensons trois occurrences d'« affronter » et deux de « franchir », de « descendre » et de « pénétrer ». Mais nous observons aussi que, dans quatre textes, plusieurs segments doivent être pris en compte : les traductions de Heguin1827 et Guillois1863 commencent par « Il f(a)it plus ». Dans le cas du premier, le segment est suivi par la construction « il osa franchir », alors que le deuxième redouble les verbes conjugués, sans pour autant répéter le pronom « il » : « il s'enfonce [...], descend ». Enfin, chez deux traducteurs, plusieurs segments peuvent être considérés comme traductions de « etiam [...] ingressus est ». Pour le texte déjà commenté

de Cogolin1750, il s'agit des segments 3 (« il dirige ses pas ; affronte sans effroi... » et 4 (« il ose pénétrer... »). Pour Desportes1846_prose, ce sont les segments 1 (« il osa même affronter ») et 3 (« il descendit dans les abîmes »).

Cette expérimentation, menée sur les différentes traductions de Virgile montre donc quelques problèmes importants :

- Les périphrases des noms propres n'étant pas prises en compte, il n'est pas aisé d'aligner les segments qui réfèrent à Pluton par périphrase (« roi redoutable », « affreux monarque »...) et ceux qui le mentionnent directement. Par ailleurs, nous risquons d'aligner les segments qui commencent par la même entité nommée, mais qui ne traduisent pas le même segment (« Pluton ; aborder les » de Hégouin1827 / « Pluton, et ces bois remplis d'une sombre horreur » de Charpentier1831).
- Il est également dommage que les syntagmes nominaux ayant un nom propre comme noyau ne fassent pas partie d'un même segment (voir par exemple « Les [gouffres du », « Ténare], [les abîmes de », « Pluton]... » de Charpentier1831).
- Pour les segments commençant par un pronom, l'alignement des blocs correspondants paraît particulièrement difficile, surtout en ce qui concerne les traductions en vers qui remanient beaucoup l'original et tendent à être plus descriptives (voir par exemple le segment « C'est peu : malgré l'horreur de ces profondes voûtes » qui ouvre la traduction de Delille1770).
- Alors que le développement de Reboul prend en compte la possibilité qu'il existe des blocs non alignés, le fait que plusieurs segments d'un texte puissent être alignés à un seul segment d'un autre ne semble pas être pris en charge (« Il osa même affronter les gouffres du... » et « Il descendit dans les abîmes de » de Desportes1846_prose correspondent tous deux à « il s'enfonce en l'horreur des forêts, descend droit au » de Guillois1863).
- Par ailleurs, la taille des segments reste arbitraire et, pour notre corpus, assez inégale (« Sous le » de Fournier1876 / « Pluton, y traverser ces forêts ténébreuses où règne un éternel effroi, s'approcher du terrible monarque des morts, et aborder ces lugubres divinités, que les prières des mortels n'ont jamais fléchies » de Desfontaines1810). Nous verrons que ce type de disproportion nuit à l'efficacité des algorithmes d'alignement².

Ces difficultés ne seront que plus prononcées dans le corpus des réécritures, ce qui fait que la méthode de Reboul, extrêmement intéressante et novatrice, ne peut pas répondre aux besoins de

² Voir III. « Logiciels dédiés à la détection automatique des réemplois textuels ».

la recherche qui nous occupe. Nous verrons toutefois, tout au long de notre travail, certaines proximités avec notre projet, par exemple en ce qui concerne le prétraitement du corpus afin d’optimiser les résultats obtenus avec les logiciels de détection automatique des relations intertextuelles, ou encore l’importance que nous accordons aux entités nommées. En effet, il est intéressant de constater que cette méthode permet d’apprécier les degrés de fidélité au texte source (en observant le nombre des blocs non alignés, par exemple). Nous verrons qu’elle s’avère extrêmement précieuse lors de l’étape de l’amélioration des résultats, déjà alignés et segmentés de manière plus stable (en phrases et syntagmes) et annotés (semi-)automatiquement pour la meilleure prise en charge des noms propres et leurs périphrases³.

Conclusion du chapitre

Dans le cadre du projet HyperApollinaire et l’exploitation du logiciel MEDITE en son sein, Olivier Gallet, Laure Michel, Michel Murat *et al.* insistent sur la nécessité de

concevoir des outils capables [de] reconnaître des formes [de l’intertextualité] élargies, ou plus allusives. Si nous disposions sous forme numérisée de la bibliothèque d’Apollinaire, enrichie de ses lectures connues, aussi érudites qu’erratiques, dans des bibliothèques comme la Mazarine, les outils d’alignement révéleraient sans doute bien des réappropriations, et enrichiraient la liste d’exemples que la mémoire, le flair et la ténacité de grands chercheurs apollinariens ont pu faire émerger.⁴

Si ces outils existent et si, comme nous le verrons, leur performance pour la reconnaissance des formes de l’intertextualité est bien plus grande que ce n’est le cas des outils d’alignement comme MEDITE ou Odysseus, plusieurs méthodes et techniques automatiques et semi-automatiques, comme celles que nous avons présentées dans ce chapitre, permettent d’améliorer encore ces résultats. Que ce soit dans la phase de prétraitement ou celle de post-traitement, ils nous donnent la possibilité d’en apprendre toujours davantage sur notre corpus et de l’observer sous des angles très différents. La stylométrie et la modélisation thématique, par exemple, permettent de se rendre compte des tendances particulières observables au sein de notre corpus. Ces techniques, ainsi que d’autres, s’avéreront très précieuses pour l’analyse et l’approfondissement des résultats de la détection automatique des relations intertextuelles, dont l’alignement des séquences peut être considéré comme l’ancêtre, et nous amèneront à quelques nouvelles interprétations concernant les réécritures françaises du mythe d’Orphée et Eurydice.

³ Voir IV.2 « Annotation des entités nommées et leurs périphrases ».

⁴ Olivier Gallet et al., « Apollinaire numérique », *loc. cit.*

Conclusion de la partie

Dans cette partie, nous avons rappelé les théories et critiques littéraires relatives à la notion de l'intertextualité. Nous avons également proposé une synthèse des procédés linguistiques potentiellement impliqués dans ce phénomène, comme les paraphrases ou les reformulations. Enfin, quelques méthodes-phares des approches computationnelles du corpus littéraire, comme la stylo-métrie, la modélisation des topiques et l'alignement des corpus parallèles ont été présentés. Cependant, le principal objectif de cette partie n'est pas théorique : ce sont les quelques expérimentations pratiques que nous avons proposées pour illustrer différents postulats, théories ou classements qui ont été au centre de nos préoccupations.

Tout d'abord, en partant d'une nouvelle parodique de D.-L. Eimann « Une histoire renouvelée des Grecs », nous avons démontré l'hétérogénéité des relations intertextuelles au sein de notre corpus, ainsi que la difficulté de les assigner sans ambiguïté à un concept théorique spécifique. Nous avons montré qu'il est impératif d'apprécier cette richesse de phénomènes observés sans se contraindre à les cloisonner dans une définition particulière et que les approches dualistes de l'intertextualité s'appliquent mal à notre corpus des réécritures. En effet, nous préférons conceptualiser les relations intertextuelles comme un rhizome, où de multiples composants peuvent être liés selon différents traits, que ces liens soient directs ou indirects, et indépendamment de la chronologie. Un tel traitement, qui admet que les textes du corpus forment un réseau de relations non-ordonné ouvre l'analyse vers des horizons encore inexplorés, que ce soit empiriquement ou à l'aide des outils computationnels. En effet, nous ne nous limitons pas à chercher des plagats, citations ou reformulations d'un texte A dans un texte B, mais nous nous demandons avant tout comment les relations intertextuelles se réalisent concrètement au sein d'un corpus dont la proximité des composants est déjà attestée et quels indices permettent de les identifier.

C'est pour cela que nous avons consacré le deuxième chapitre à l'analyse et la comparaison des relations linguistiques potentiellement indices des relations intertextuelles, comme la paraphrase et la reformulation. À partir d'une investigation menée sur une phrase de 21 traductions du livre X des *Métamorphoses* d'Ovide, nous avons montré l'extrême difficulté d'une analyse linguistiquement exhaustive et objectivable de ce type de phénomènes. Nous avons fait face à plusieurs difficultés. Premièrement, nous avons réalisé que l'évolution du sens des mots rend la comparaison des textes au sein d'un corpus diachronique ardue. Ensuite, nous avons constaté que certains types de paraphrases, notamment sémantiques et pragmatiques, sont difficilement objectivables, car elles renvoient à des compétences extralinguistiques. Enfin, il ne nous a pas toujours été possible de déterminer sans aucune ambiguïté qu'il s'agissait d'une telle ou telle relation, voire même si celle-ci était

effective : nous en avons conclu que la prise en compte des relations sémantiques et conversives reste incertaine, notamment pour des verbes très polysémiques.

Cela nous a permis de délimiter des garde-fous méthodologiques avant de procéder à la détection (semi-)automatique des relations intertextuelles. En effet, il serait erroné de prétendre pouvoir automatiser la détection et le classement exhaustif des procédés linguistiques présents au sein de différentes relations intertextuelles. Alors que des outils informatiques sont capables de détecter une similarité formelle entre des chaînes de caractères, trouver des mots identiques, voire lisser les différences de flexion et de dérivation, l'identification des relations plus subtiles ou plus complexes, comme la métaphorisation ou la modernisation, reste un défi même pour un lecteur humain.

Les approches computationnelles qui nous ont intéressées dans le dernier chapitre, la stylométrie et la détection des topiques, ont permis de révéler les tendances particulières observables au sein de notre corpus et de les corrélérer avec son classement manuel. L'alignement automatique des corpus parallèles est, quant à lui, une technique qu'il a été important de présenter, car elle donne les fondations aux algorithmes utilisés dans les logiciels de la détection automatique des réutilisations textuelles que nous présentons dans la partie suivante. En revanche, même si les outils MEDITE ou Odysseus sont performants pour la comparaison de textes très proches (différentes versions d'une même œuvre, séries traductives...), ils ne sont pas en mesure de reconnaître des formes d'intertextualité plus subtiles.

Au fil de différentes expérimentations, nous avons esquissé nos besoins concernant les logiciels de la détection automatique des relations intertextuelles (capables de travailler avec plusieurs textes à la fois, modulables, paramétrables, proposant des résultats faciles à extraire et à enrichir...). Nous avons aussi déterminé les difficultés que nous risquons de rencontrer au cours du traitement automatique et nous avons proposé quelques principes de traitement permettant de les éviter. Enfin, nous avons donné un aperçu des traitements computationnels des corpus littéraires et avons montré différents niveaux de technicité qu'elles requièrent. À présent que nous sommes consciente de difficultés et écueils que nous risquons de rencontrer et que le cadre théorique est bien posé, nous pouvons poursuivre nos expérimentations en confrontant notre corpus aux différents logiciels spécifiquement dédiés à la détection automatique des réutilisations textuelles. C'est ce à quoi nous nous attaquons dans la partie suivante.

III. Logiciels dédiés à la détection automatique des ré- emplois textuels

De nombreuses disciplines s'intéressent, comme nous, à la détection automatique des réutilisations et reformulations dans des textes, que ce soit pour l'identification du plagiat¹, l'attribution de l'auctorialité d'œuvres anonymes² ou encore la génération de résumés³. Ces travaux révèlent que certaines méthodes informatiques et statistiques sont probantes pour la détection automatique des paraphrases et des reformulations⁴. Alors que certaines analyses se focalisent surtout sur les paraphrases⁵, souvent en s'appuyant sur les méthodes d'alignement⁶, d'autres travaux portent sur les reformulations ponctuelles en observant la distribution des mots dans les textes. Ferrero et Simac-Lejeune⁷, notamment, procèdent, après la segmentation du corpus en unités de sens (en considérant comme délimiteurs la ponctuation forte, mais aussi les virgules, les conjonctions de coordination, des mots de liaison, etc.) à l'extraction des mots porteurs de sens (obligatoires à conserver pour garder le sens de la phrase-source). Grâce au thésaurus contenant les mots de substitution possibles pour chaque mot-clé, les auteurs démontrent qu'il est possible d'établir des correspondances en faisant abstraction de l'ordre de mots, des variantes typographiques, de la dérivation ou de la flexion.

Alors que certaines de ces approches puisent dans des méthodes de la linguistique du corpus, de la statistique ou du traitement automatique des langues, d'autres, plus récentes, s'appuient sur les technologies de l'intelligence artificielle, comme l'apprentissage automatique (*machine learning*), voire l'apprentissage profond (*deep learning*). Par exemple, les outils d'aide à la détection du plagiat, comme Compilatio⁸, sont utilisés par les enseignants pour s'assurer que le travail présenté par les élèves et les étudiants est bien le leur. Les logiciels de ce type recherchent principalement la reprise des passages identiques, autrement dit du « copier-coller », dont l'auteur ne serait pas référencé.

¹ Jérémy Ferrero et Alain Simac-Lejeune, « Détection automatique de reformulations – Correspondance de concepts appliquée à la détection du plagiat », *Actes de la 15^e conférence internationale sur l'extraction et la gestion des connaissances*, Luxembourg, 2015.

² Jan Rybicki, « Vive la différence: Tracing the (authorial) gender signal by multivariate analysis of word frequencies », *Digital Scholarship in the Humanities*, vol. 31, n° 4, 2016, p. 746-761.

³ Regina Barzilay et Michael Elhadad, « Using Lexical Chains for Text Summarization », *Intelligent Scalable Text Summarization*, 1997.

⁴ Voir par exemple Dana Shapira et James A. Storer, « Edit Distance with Move Operations », *Proceedings of the 13th Annual Symposium on Combinatorial Pattern Matching*, London, UK, Springer-Verlag, coll. « CPM '02 », 2002, p. 85-98, en ligne, <<http://dl.acm.org.ressources-electroniques.univ-lille.fr/citation.cfm?id=647821.736368>>; Patrick J. Burns et al., « Profiling of Intertextuality in Latin Literature Using Word Embeddings », *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online, Association for Computational Linguistics, 2021, p. 4900-4907, en ligne, <<https://aclanthology.org/2021.naacl-main.389>>.

⁵ Iris Eshkol-Taravella et Natalia Grabar, « Repérage et analyse de la reformulation paraphrastique dans les corpus oraux », *Actes de la 21^{ème} conférence sur le Traitement Automatique des Langues Naturelles*, s. l., 2014, p. 304-315.

⁶ Regina Barzilay et Noemie Elhadad, « Sentence Alignment for Monolingual Comparable Corpora », *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 2003, p. 25-32, en ligne, <<https://www.aclweb.org/anthology/W03-1004>>.

⁷ Jérémy Ferrero et Alain Simac-Lejeune, « Détection automatique de reformulations – Correspondance de concepts appliquée à la détection du plagiat », *loc. cit.*

⁸ <https://www.compilatio.net/>.

Le développement informatique, tant en ce qui concerne la conception d'outils que l'accès aux corpus numériques, marque donc une étape importante pour l'étude de l'intertextualité et le repérage des réutilisations au sein de textes littéraires, notamment en permettant d'accélérer la recherche et d'offrir une vision plus large de ces phénomènes⁹. Toutefois, dans la partie précédente¹⁰, nous avons constaté que les modifications linguistiques qu'il nous faudra prendre en compte pour la détection des relations intertextuelles au sein de notre corpus sont beaucoup plus complexes qu'un simple copier-coller.

Jérémy Ferrero et Alain Simac-Lejeune différencient trois types de similitudes textuelles, dont la *copie*, « qui consiste à copier mot à mot tout ou partie d'un texte dans un autre »¹¹, est la plus facile à détecter automatiquement. Les deux autres, la *paraphrase* et la *reformulation*, continuent à poser problème aux méthodes informatiques. Alors que, selon les auteurs, la paraphrase vise avant tout à détailler ou à expliciter la phrase reprise, au sein de la reformulation, chaque type de modification peut être rencontré, tant que le sens de la phrase-source est conservé. Dans la partie précédente, nous avons pu observer ces différentes similitudes textuelles, tantôt à travers des analyses littéraires et linguistiques traditionnelles, tantôt en nous basant sur des méthodes informatiques pouvant être considérées comme prolégomènes de la détection automatique des relations intertextuelles.

Comme le constatent Enrique Manjavacas *et al.* dans leur article intitulé « On the Feasibility of Automated Detection of Allusive Text Reuse »,

[a]vec l'émergence des méthodes informatiques dans les études littéraires au cours des dernières décennies, l'intertextualité a souvent été présentée comme une application prometteuse, aidant les chercheurs à identifier de potentiels liens intertextuels jusqu'alors passés inaperçus. Beaucoup de progrès ont été accomplis dans ce domaine et un bon nombre de logiciels extrêmement utiles sont désormais disponibles.¹²

Néanmoins, alors qu'il est stimulant de pouvoir détecter des relations intertextuelles au moyen d'outils informatiques, les performances de ces derniers varient. Comme nous le verrons, les logiciels dédiés exploitent relativement peu de connaissances linguistiques, mais des techniques comme

⁹ Neil Coffee et al., « Intertextuality in the Digital Age », *Transactions of the American Philological Association*, vol. 142, n° 2, 2012, p. 383-422, en ligne, <DOI : 10.1353/apa.2012.0010>.

¹⁰ Voir II.2 « Procédés linguistiques propres des relations intertextuelles ».

¹¹ Jérémy Ferrero et Alain Simac-Lejeune, « Détection automatique de reformulations – Correspondance de concepts appliquée à la détection du plagiat », *loc. cit.*

¹² « With the emergence of computational methods in literary studies over the past decades, intertextuality has often been presented as a promising application, helping scholars identifying potential intertextual links that had previously gone unnoticed. Much progress has been made in this area and the number of highly useful tools are now available » (nous traduisons). Voir Enrique Manjavacas, Brian Long et Mike Kestemont, « On the Feasibility of Automated Detection of Allusive Text Reuse », *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, Minneapolis, USA, Association for Computational Linguistics, 2019, p. 1-2, en ligne, <<https://www.aclweb.org/anthology/W19-2514>>.

le plongement lexical¹³ ou l'alignement séquentiel¹⁴ les rendent performants surtout pour des relations proches (majoritairement des citations, plus rarement des reformulations simples ou paraphrases lexicales). Ainsi, il nous paraît excessivement optimiste de parler dans ce contexte d'une détection automatique des relations intertextuelles, dont la complexité, tant du point de vue littéraire que linguistique, a été longuement discutée dans la partie précédente.

Dans cette partie, il est plutôt question de la détection automatique des réemplois textuels (*text reuse*), que Romanello *et al.*¹⁵ définissent comme « une réitération significative d'un texte, généralement au-delà de la simple répétition du langage courant »¹⁶. Dans un premier temps, nous présentons une expérimentation initiale avec le logiciel Phœbus, à ce jour l'unique outil de ce type conçu avec une interface graphique¹⁷, et donc facilement manipulable par des utilisateurs peu familiers d'outils informatiques. Alors que nous nous sommes rapidement rendu compte qu'il n'était pas suffisant pour nos besoins, les expérimentations effectuées avec ce logiciel nous ont permis de déterminer nos attentes et d'établir une liste de critères pour un choix optimal des logiciels à employer au sein de notre projet. Pour cette sélection, nous confrontons six logiciels, pour la plupart plus récents que Phœbus, et surtout toujours accessibles : BLAST¹⁸, Passim¹⁹, Tesserae²⁰, TextPAIR²¹, Textreuse²² et Tracer²³ (nous établissons cette liste restreinte en suivant l'article « Detecting Text Reuse with Passim » de Matteo Romanello et Simon Hengchen²⁴). Nous présentons d'abord les logiciels que nous ne retenons pas, avant de nous pencher davantage sur les deux outils qui se sont révélés le plus adaptés à nos besoins, Tracer et TextPAIR.

¹³ Tomas Mikolov et al., *Efficient Estimation of Word Representations in Vector Space*, arXiv:1301.3781, 6 septembre 2013.

¹⁴ Michael J Wise, « Neweyes: A System for Comparing Biological Sequences Using the Running Karp-Rabin Greedy String-Tiling Algorithm », *Unpublished Basser Department of Computer Science*, 1993; Lasse Bergroth, Harri Hakonen et Timo Raita, « A Survey of Longest Common Subsequence Algorithms », *Proceedings of the Seventh International Symposium on String Processing Information Retrieval*, Corogne, 2000, p. 39-48, en ligne, <doi: 10.1109/SPIRE.2000.878178>.

¹⁵ Matteo Romanello, Aurélien Berra et Alexandra Trachsel, « Rethinking Text Reuse as Digital Classicists », *Digital Humanities conference*, Lausanne, 2014, en ligne, <https://wiki.digitalclassicist.org/Text_Reuse>.

¹⁶ « The meaningful reiteration of text, usually beyond the simple repetition of common language » (nous traduisons). *Ibid.*

¹⁷ Nous verrons par la suite que d'autres logiciels proposent une interface graphique, mais celle-ci ne permet que d'afficher les résultats de la détection, le traitement en soi devant être effectué tantôt en ligne de commande, tantôt à l'aide d'un langage de programmation, comme Python ou R.

¹⁸ Voir <https://blast.ncbi.nlm.nih.gov/Blast.cgi> et § « BLAST (Basic Local Alignment Search Tool) ».

¹⁹ Voir <https://github.com/dasmiq/passim> et § « Passim ».

²⁰ Voir <https://github.com/tesseract/tesseract> et § « Tesserae ».

²¹ Voir <https://github.com/ARTFL-Project/text-pair> et § « Présentation des logiciels retenus, TextPAIR et Tracer ».

²² Voir <https://docs.ropensci.org/textreuse/> et § « Textreuse R ».

²³ Voir <https://www.etrapp.eu/research/tracer/> et § « Présentation des logiciels retenus, TextPAIR et Tracer ».

²⁴ Matteo Romanello et Simon Hengchen, « Detecting Text Reuse with Passim », *Programming Historian*, mai 2021, en ligne, <<https://programminghistorian.org/en/lessons/detecting-text-reuse-with-passim>>. À notre connaissance, il s'agit du recensement le plus récent au moment de la rédaction de ce chapitre.

1. Établissement des critères de l'évaluation avec le logiciel Phœbus

Le logiciel Phœbus¹ a été conçu sous la direction de Jean-Gabriel Ganascia au sein de l'équipe ACASA du LIP6 en collaboration avec l'ITEM (Institut des textes et manuscrits modernes) et des chercheurs en humanités numériques littéraires du laboratoire d'excellence ObViL (Observatoire de la vie littéraire, actuellement ObTic). Avant la désactivation du serveur hébergeant Phœbus dans le courant de l'année 2022², ce logiciel a été le seul parmi ceux que nous présentons à disposer d'une interface graphique, et donc facilement manipulable par des utilisateurs peu familiers des outils informatiques. Nous espérons qu'une version mise à jour sera remise en ligne dans le futur³. En attendant, nous sommes contrainte d'appuyer notre description sur des exemples restreints, issus d'une expérimentation que nous avons menée au début de notre recherche doctorale⁴.

Pour évaluer la performance du logiciel à l'époque, nous avons isolé un petit sous-corpus de six traductions : (i) en ancien français, en vers (Legouais1301⁵) et en prose (Walley1493⁶) ; (ii) juxtalinéaire, en vers (Parnajon1880-1⁷) et en prose (Parnajon1880-2⁸) ; littéraire, moderne ou contemporaine, en vers (Cosnay2017⁹) et en prose (Villenave1806¹⁰).

a) Fonctionnement du logiciel

Phœbus permet la comparaison simultanée de deux textes qu'il prétraite grâce à des techniques talistes. Il élimine notamment les mots faibles, comme les articles, les auxiliaires ou les prépositions,

¹ Mohamed-Amine Boukhaled, Zied Sellami et Jean-Gabriel Ganascia, « Phœbus : un Logiciel d'Extraction de Réutilisations dans des Textes Littéraires », *22ème Conférence sur le Traitement Automatique des Langues Naturelles*, Caen, France, 2015, en ligne, <<https://hal.sorbonne-universite.fr/hal-01198411>>; Jean-Gabriel Ganascia, Pierre Glaudes et Andrea Del Lungo, « Automatic Detection of Reuses and Citations in Literary Texts », *Digital Scholarship in the Humanities*, vol. 29, n° 3, 2014, p. 412-421.

² Jusqu'à 2022, l'interface a été disponible sous cette adresse : <http://obvil-dev.paris-sorbonne.fr/phoebus/>. Actuellement, il est encore possible de le consulter l'interface sous <https://web.archive.org/web/20220121001231/http://obvil-dev.paris-sorbonne.fr/phoebus/>, sans pour autant pouvoir effectuer des traitements.

³ Le directeur du projet, Jean-Gabriel Ganascia, a été interrogé à ce sujet, mais, au moment de la rédaction de ce mémoire, aucune information quant au futur de Phœbus n'a pu nous être fournie.

⁴ Voir Karolina Suchecka, Nathalie Gasiglia et Karl Zieger, « Édition comparative intermédiaire de séries traductives », *loc. cit.* Cette section reprend et développe les expérimentations décrites dans cet article.

⁵ Chrétien Legouais, *Ovide moralisé*, Rouen, Manuscrits de la BM de Rouen., 1301, 246v-249r, 250r-251r, 261v-261v, 271r-272r.

⁶ Thomas Walley, *La Bible des poètes. Métamorphose d'Ovide moralisée par Thomas Walley et traduite par Colard Mansion*, trad. Colard Mansion, Paris, 1493, p. 107v-109v.

⁷ Ovide, « Choix des Métamorphoses », dans Société de professeurs et de latinistes (dir.), *Les auteurs latins expliqués d'après une méthode nouvelle par deux traductions françaises*, trad. Félix de Parnajon, Paris, Hachette, 1880, p. 418-429, 440-447.

⁸ *Ibid.*

⁹ Ovide, *Les métamorphoses*, trad. Marie Cosnay, Paris, Éditions de l'Ogre, 2017. Pour les questions des droits à la traduction, cette œuvre ne fait pas partie de notre corpus.

¹⁰ Ovide, *Les Métamorphoses d'Ovide, traduction nouvelle, avec le texte latin, suivie d'une analyse, de l'explication des fables, de notes géographiques, historiques, mythologiques et critiques*, trad. Mathieu-Guillaume-Thérèse Villenave, vol. 3, Paris, F. Gay, Ch. Guestard, 1806, 296-306, 380-386 p. La version utilisée pour cette expérimentation était celle disponible sur *Biblioteca classica selecta* (<http://bcs.fltr.ucl.ac.be/meta/10.htm>). Elle est légèrement adaptée, et surtout modernisée. Au sein de notre corpus doctoral, nous revenons à la version de 1806.

et procède à la racinisation (*stemming*) des mots conservés, à l'aide de l'algorithme Snowball¹¹ afin de ne garder que les racines des mots retenus ce qui fait, par exemple, que *aimer*, *aimaient*, *aimerai*, mais aussi *aimant*, sont réduits à la racine *aim*. Ensuite, le logiciel détecte les mots racinisés qui apparaissent dans les deux textes selon les paramètres définis par l'utilisateur. Les paramètres par défaut fixent à trois le nombre maximal de “trous” autorisés entre deux occurrences et le nombre minimal de mots communs. L'ordre des mots peut être pris en compte ou ignoré.

Nous présentons à la Figure 31 le produit de la comparaison d'une traduction en prose et en ancien français (Walleys1493) et d'une traduction moderne en vers (Cosnay2017).

```
<phoebus file_1_path="path="/var/www/obvil/phoebus/tmp/phoebus/TXToZGFsb" file_2_path="/var/
www/obvil/phoebus/tmp/phoebus/TXTI5LFOA" gapSize="3" patternSize="3" respectWordOrder=
>false">
  <reuse id="0">
    <pattern size="4">@plaindr@pein@derni@aim</pattern>
    <correctness>0</correctness><precision>H</precision>
    <text_1_char_start_index>4717</text_1_char_start_index>
    <text_1_char_end_index>4827</text_1_char_end_index>
    <text_2_char_start_index>3370</text_2_char_start_index>
    <text_2_char_end_index>3498</text_2_char_end_index>
    <text_1_word_start_index>994</text_1_word_start_index>
    <text_1_word_end_index>1021</text_1_word_end_index>
    <text_2_word_start_index>840</text_2_word_start_index>
    <text_2_word_end_index>872</text_2_word_end_index>
    <text_1_reuse>seconde mort. Mais de lui ne se peut plaindre, fors de trop aimer. Le
    dernier salut luy rendi que a peines</text_1_reuse>
    <text_2_reuse>plaint pas du tout (de quoi se plaindre, si ce n'est d'être aimée ?), elle
    dit un dernier « Adieu », qu'il peut à peine entendre</text_2_reuse>
  </reuse>
</phoebus>
```

Figure 31. Extrait du produit (en XML) du traitement par Phœbus de deux traductions d'Ovide

Les attributs de l'élément XML <phoebus> indiquent les paramètres du traitement (@gapSize : nombre maximal des trous ; @patternSize : nombre minimal de correspondances à trouver ; @respectWordOrder : respect, ou non, de l'ordre des mots). Chaque correspondance repérée est dans un élément <reuse> porteur d'un identifiant unique (@id). Son fils <pattern> fournit les mots racinisés (stemmes) qui ont permis de repérer la correspondance – ici : *plaindr* (*plaindre* dans les deux textes), *pein* (*peines* dans le premier et *peine* dans le deuxième), *derni* (*dernier* dans les deux textes) et *aim* (*aimer* vs *aimée*). Les éléments frères postposés permettent de localiser les extraits repérés et de calculer leurs longueurs en nombre de caractères (de <text_1_char_start_index> à <text_2_char_start_index>) et de mots (de <text_1_word_start_index> à <text_2_word_start_index>). Enfin, les deux segments de textes mis en correspondance sont reproduits au sein des éléments <text_1_reuse> et <text_2_reuse>.

¹¹ Martin F. Porter, « Snowball: A language for stemming algorithms », *Retrieved March*, vol. 1, 2001; Stephen Tomlinson, *Lexical and Algorithmic Stemming Compared for 9 European Languages with Hummingbird SearchServer at CLEF 2003*, dans Peters C et al. (dir.), vol. 3237, 2004, p. 286-300.

b) Analyse des correspondances détectées

Nous avons comparé, une par une, les six traductions entre elles. Les textes n'ont subi aucun prétraitement (ni modernisation ni lemmatisation) et ils ont été soumis au logiciel en format brut (le texte n'a pas été enrichi par des styles ou des mises en forme). Pour Legouais1301, 11 correspondances ont été trouvées avec Walleys1493 et une avec un seul des textes postérieurs, Parnajon1880-1 (« doloit pour sa double mort » / « stupéfait de la double mort »). Pour Walleys1493, les relations avec les textes juxta-linéaires et modernes sont un peu plus nombreuses. Nous en recensons huit : une avec Parnajon1880-1, une avec Parnajon1880-2, deux avec Cosnay2017 et quatre avec Villenave1806. La comparaison des résultats montre la récurrence d'une correspondance, présentée dans le Tableau 32 entre le texte de Walleys1493, à gauche, et les autres textes, à droite. L'étendue de chaque correspondance est marquée par des crochets suivis de l'identifiant du texte lié en indice. Les stemmes cooccurents sont soulignés et suivis par un numéro permettant de faire un lien entre l'extrait de gauche et les extraits correspondants à droite. Par exemple, Walleys1493 et Villenave1806 ne partagent que deux stemmes : seconde⁵ et plaindre⁸.

Tableau 32. Correspondances entre Walleys1493 et les autres textes

Walleys1493 « [<u>mains</u>¹ qui retenir la <u>cuid</u>^{2a}. Mais riens ne print <u>fors</u>³ <u>vent</u>⁴ et ainsi se <u>partit</u> Erudice de son amy et mourut de [<u>seconde</u>⁵ <u>mort</u>⁶. Mais de lui ne se peut <u>plaindre</u>^{7e}]⁸ Villenave1806 fors de trop <u>aim</u>⁹er. Le <u>derni</u>¹⁰ <u>er</u>¹¹ <u>salut</u>¹²]¹³ Parnajon1880-2 luy <u>rendi</u>¹³t que a <u>peine</u>¹⁴ s]¹⁵ Parnajon1880-1, Cosnay2017 <u>entendi</u>¹⁵t Orpheus. Forment se plaingnoit de la seconde mort de s'amie et voulut retourner pour trouver la morte mais la <u>porte</u>¹⁶ <u>trouva</u>¹⁷ <u>ferm</u>¹⁸ee. Et le <u>portier</u>¹⁹ qui la <u>gardoi</u>²⁰t]²¹ Legouais1301 »	Legouais1301
	« tent ses <u>mains</u> ¹ et prendre <u>cuid</u> ^{2e} , mes ne print <u>fors</u> ³ <u>vent</u> ⁴ vain et vuide. Cele se <u>part</u> de son mari, qui de <u>seconde</u> ⁵ <u>mort</u> ⁶ mori. Mes ne se puet de lui blasmer se ne se plaint de trop amer. Le desrain <u>salut</u> ¹² li <u>rendi</u> ¹³ , que cil a paines <u>entendi</u> ¹⁵ . Orpheüs forment se doloit pour sa double mort et voloit retourner pour querre la morte, mes il <u>trouva</u> ¹⁷ <u>fermé</u> la <u>porte</u> ¹⁶ et le <u>portier</u> ¹⁹ qui le <u>gardoi</u> ²⁰ t, »
	Parnajon1880-1
	« quoi en effet se <u>plaindre</u> ⁷ ait-elle sinon d'avoir été <u>aim</u> ⁹ ée ? Et elle dit pour la <u>derni</u> ¹⁰ ère fois un adieu, tel que celui-ci pût le recevoir à <u>peine</u> ¹⁴ de »
	Parnajon1880-2
	« meurt une <u>seconde</u> ⁵ fois, mais sans se <u>plaindre</u> ⁸ de son époux ; de quoi en effet se plaindrait-elle sinon d'être <u>aim</u> ⁹ ée ? Elle lui adresse un <u>dernier</u> ¹¹ adieu »
	Cosnay2017
	« plaint pas du tout (de quoi se <u>plaindre</u> ⁸ , si ce n'est d'être <u>aim</u> ⁹ ée ?), elle dit un <u>dernier</u> ¹¹ « Adieu », qu'il peut à <u>peine</u> ¹⁴ entendre » ¹²
	Villenave1806
	« Eurydice meurt une <u>seconde</u> ⁵ fois, mais sans se <u>plaindre</u> ⁸ ; » ¹³

¹² L'autre relation détectée avec Walleys1493 est « touchoit les cordes et de sa bouche se print a chanter telle chançon »/ « pleuraient sur lui disant de tels chants, et touchant ses cordes selon ».

¹³ Les trois autres relations détectées avec Walleys1493 sont : « serpent tellement la blessa qu' »/ « fleurie, un serpent la blesse au » ; « rive du fleuve infernal fut »/ « assis sur la rive infernale, » ; « meschine, fuyant tout amour feminine »/ « fuyait les femmes et l'amour ».

La correspondance trouvée entre Walleys1493 et Legouais1301 est beaucoup plus longue que celles repérées avec les autres textes. Seize stemmes communs ont été trouvés entre eux, dont *second*, qui est aussi commun avec d'autres textes du corpus (Parnajon1880-2 et Villenave1806), mais ce n'est pas suffisant pour qu'une correspondance puisse être détectée entre l'extrait de Legouais1301 et ceux des traductions plus modernes (nous demandons deux stemmes communs minimum). Au contraire, les autres correspondances repérées regroupent un nombre restreint de stemmes assez récurrents. Ce nombre peut être réduit à quatre lemmes au total : *plaindre*, *aimer*, *dernier* et *peine*.

Notons toutefois quelques difficultés. Du point de vue pratique, le fait de ne pouvoir traiter que des couples de textes à chaque fois rend le traitement d'un grand corpus particulièrement chronophage, même si la possibilité de télécharger les résultats en format XML permet d'automatiser la mise en commun des résultats pour chaque couple. En ce qui concerne la méthode de mise en correspondance par stemmes communs, elle n'est pas optimale. D'abord, la racinisation du mot *dernier* et de sa forme féminine *dernière* n'est pas la même (*dernier* / *derni*), ce qui est problématique, comme la graphie ancienne est un obstacle pour la mise en correspondance de *aimer* (*amer*), *peine* (*paines*) et *dernier* (*desrain*). Par ailleurs, la racinisation ne permet pas toujours de mettre en correspondance les formes conjuguées des mêmes verbes (*plaignoit* / *plaindrait* / *plaint* ou *mourut* / *mori* / *meurt*). Enfin, les emplois des synonymes (*blasmer* chez Legouais1301) et *plaindre* dans les autres textes bloquent les repérages de correspondances.

En revanche, l'avantage de Phœbus réside dans le fait que les résultats contiennent des informations utiles pour faciliter leur analyse. Outre le fait que les fichiers XML sont facilement interrogeables, les informations concernant la nature de chaque correspondance, notamment les stemmes communs et la localisation exacte de chaque extrait, sont extrêmement précieuses notamment pour l'observation des correspondances croisées (où le même fragment d'un texte correspond, au moins partiellement, à d'autres entités du corpus). Il est par exemple possible d'établir un graphe de visualisation qui donne un premier aperçu des relations repérées entre les textes. Dans ce cadre, nous avons créé deux tables des correspondances : celle des nœuds et celle des liens. La table de nœuds (Tableau 33) recense les extraits trouvés les plus complets, leurs identifiants et les nombres totaux de stemmes communs trouvées avec d'autres textes. Ces nœuds correspondent à la totalité des résultats trouvés pour l'épisode de la deuxième mort d'Eurydice dans tous les textes.

Tableau 33. Table des nœuds

ID de l'extrait	Extrait	Stemmes
Legouais1301_9	« tent ses mains et prendre cuide, mes ne prent fors vent vain et vuide. Cele se part de son mari, qui de seconde mort mori. Mes ne se puet de lui blasmer se ne se plaint de trop amer. Le desrain salut li rendi, que cil a paines entendu. Orpheüs forment se doloit pour sa double mort et voloit retourner pour querre la morte, mes il trouva fermé la porte et le portier qui le gardoit, »	17
Walleys1493_10	« qui retenir la cuida. Mais riens ne print fors vent et ainsi se partit Erudice de son amy et mourut de seconde mort. Mais de lui ne se peut plaindre, fors de trop aimer. Le dernier salut luy rendit que a peines l'entendit Orpheus. Forment se plaingnoit de la seconde mort de s'amie et voulut retourner pour trouver la mort mais la porte trouva fermee. Et le portier qui la gardoit »	30
Parnajon1880-1_12	« retirent. Et déjà mourant pour la seconde fois, elle se plaint en quoi que ce soit de son époux : de quoi en effet se plaindrait-elle sinon soi avoir été aimée ? Et elle dit pour la dernière fois un adieu, tel que celui-ci pût le recevoir à peine de tel que celui-ci pût le recevoir à peine de ses oreilles ; et elle fut replongée de nouveau au même lieu. Orphée resta stupéfait de la double mort de son épouse, »	36
Parnajon1880-2_14	« dissipe. Déjà elle meurt une seconde fois, mais sans se plaindre de son époux ; de quoi en effet se plaindrait-elle sinon d'être aimée ? Elle lui adresse un dernier adieu qui parvient à peine à ses oreilles, et elle est de nouveau replongée dans le même gouffre. Orphée, qui voit la mort lui ravir une seconde fois son épouse, »	34
Cosnay2017_11	« Mourant une deuxième fois, de son époux elle ne se plaint pas du tout (de quoi se plaindre, si ce n'est d'être aimé ?), elle dit un dernier « Adieu », qu'il peut à peine entendre et elle roule au lieu où elle était avant. Devant la mort double de sa femme, Orphée resta immobile »	24
Villenave1806_18	« Eurydice meurt une seconde fois, mais sans se plaindre ; »	7
Villenave1806_19	« crime de l'avoir trop aimée ! Adieu, lui dit-elle d'une voix faible qui fut à peine entendue »	6

Remarquons que le nœud Legouais1301_9 pour lequel 15 stemmes communs ont été trouvés avec Walleys1493_10, regroupe également « doloit pour sa double mort », la seule correspondance de Legouais1301_9 avec un texte en français moderne (Parnajon1880-2_12, deux stemmes communs) : la taille de nœud est donc la somme des patterns pour ces deux correspondances (17). Les correspondances de la plus petite taille ont été trouvées pour Villenave1806 avec deux extraits pour un total de 26 mots :

Le malheureux Orphée lui tend les bras, il veut se jeter dans les siens : il n'embrasse qu'une vapeur légère. [Eurydice meurt une seconde fois, mais sans de plaindre ;]¹⁸ et quelle plainte eût-elle pu former ? Était-ce pour Orphée [un crime de l'avoir trop aimée ! Adieu, lui dit-elle d'une voix faible qui fut à peine entendue]¹⁹ ; et elle rentre dans les abîmes.

La métaphorisation et la synonymisation, présentes dans cet extrait, empêchent le repérage de correspondances plus étendues. Les stemmes communs repérés sont les mêmes que dans les autres

textes (*mourir, seconde, plaindre, fois* pour l'extrait 18 et *adieu, aimer, peine* pour l'extrait 19¹⁴), mais les réécritures poétisées (« les abîmes du trépas ») et les reformulations synonymiques (« deux fois ravie » au lieu de « double mort », « seconde mort », etc.), empêchent la réunion des deux correspondances courtes en une longue, du moins en continuant à limiter la taille des “trous” à trois mots.

Nous remarquons également que la correspondance trouvée entre Walleys1493 et Cosnay2017 (« plaint pas du tout (de quoi se plaindre, si ce n'est d'être aimée ?), elle dit un dernier « Adieu », qu'il peut à peine entendre ») fait partie d'un nœud plus large, Cosnay2017_11, pour lequel 24 stemmes communs ont été trouvés au total avec d'autres extraits. Ces relations spécifiques sont décrites dans la table des liens (Tableau 34) où sont renseignés les nœuds sources et cibles, les lemmes des mots¹⁵ qui ont permis d'établir les correspondances et leurs nombres.

Tableau 34. Table des liens

Nœuds sources	Nœuds cibles	Lemmes des mots qui ont permis d'identifier les correspondances	Nb
Legouais1301_9	Walleys1493_10	cuidier, entendre, fermer, fors, garder, main, mort, plaindre, porte, portier, rendre, salut, second, trouver, vent	15
Walleys1493_10	Villenave1806_18	mais, plaindre, second	3
Parnajon1880-2_14	Villenave1806_18	fois, mourir, plaindre, second	4
Parnajon1880-1_12	Cosnay2017_11	adieu, dernier, époux, lieu, Orphée, plaindre, peine, quoi, rester	9
Parnajon1880-1_12	Parnajon1880-2_14	adieu, aimer, déjà, dernier, effet, épouse, époux, même, mort, nouveau, oreille, Orphée, peine, plaindre, quoi, replonger, second, sinon	18
Legouais1301_9	Parnajon1880-1_12	double, mort	2
Walleys1493_10	Parnajon1880-2_14	aimer, dernier, plaindre, seconde	4
Cosnay2017_11	Villenave1806_19	adieu, aimer, peine	3
Walleys1493_10	Parnajon1880-1_12	aimer, dernier, plaindre, peine	4
Walleys1493_10	Cosnay2017_11	aimer, dernier, plaindre, peine	4
Parnajon1880-2_14	Cosnay2017_11	adieu, aimer, dernier, époux, être, plaindre, peine, quoi	8
Parnajon1880-1_12	Villenave1806_19	adieu, aimer, peine	3

Nous pouvons constater que 4 correspondances sont recensées pour le nœud Cosnay2017_11 : avec Parnajon1880-1_12 (9 mots communs), Parnajon1880-2_14 (8), Walleys1493_10 (4) et Villenave1806_19 (3). Plus généralement, pour les 7 nœuds présentés dans le Tableau 33, nous trouvons 12 liens au total, dont le plus grand compte 18 stemmes (Parnajon1880-1_12 / Parnajon1880-2_14). Parmi les 38 lemmes recensés, nous relevons 6 occurrences de *plaindre* et de *aimer*, 5 de *dernier*, *second*

¹⁴ Pour faciliter la lecture, nous présentons les lemmes communs plutôt que de conserver les stemmes : ainsi par exemples les stemmes *derni* pour les correspondances *dernier~dernière* et *dernier* pour les correspondances *dernier~dernier* sont regroupés sous le lemme commun *dernier*.

¹⁵ Voir n. 14.

et *peine* (qui fait partie de la locution adverbiale *à peine*), 4 de *adieu* (contre une seulement de *salut*), 3 de *mort*, etc. Le nom propre *Orphée* n'a été détecté que pour deux comparaisons (Parnajon1880-1_12 / Parnajon1880-2_14 et Parnajon1880-1_12 / Cosnay2017_11).

c) Établissement des graphes de relations

Les Tableau 33 et Tableau 34, fournis à un logiciel comme Gephi, permettent de générer une visualisation des correspondances à l'aide d'un graphe comme celui que nous proposons à la Figure 32. Chaque nœud est représenté par un disque. Les disques de la même couleur représentent les extraits issus de la même traduction (par exemple le violet pour Villenave1806). La taille de chaque nœud dépend du nombre des stemmes communs détectés au total (voir la dernière colonne du Tableau 33) et l'épaisseur des liens qui relient chaque nœud correspond au nombre des stemmes communs pour chaque couple (voir la dernière colonne du Tableau 34). Chaque nœud établit entre deux (Villenave1806_18, Villenave1806_19 et Legouais1301_9) et cinq (Walleys1493_10) liens. Curieusement, pour ceux de Villenave1806, les deux nœuds présents ne sont pas en relation avec les mêmes textes : Villenave1806_18 établit des liens avec Walleys1493 et Parnajon1880-2, Villenave1806_19 avec Parnajon1880-1 et Cosnay2017. Legouais1301_9 est lié uniquement à Parnajon1880-1 et Walleys1493, mais ce dernier lien est d'une taille très importante (15 patterns communs).

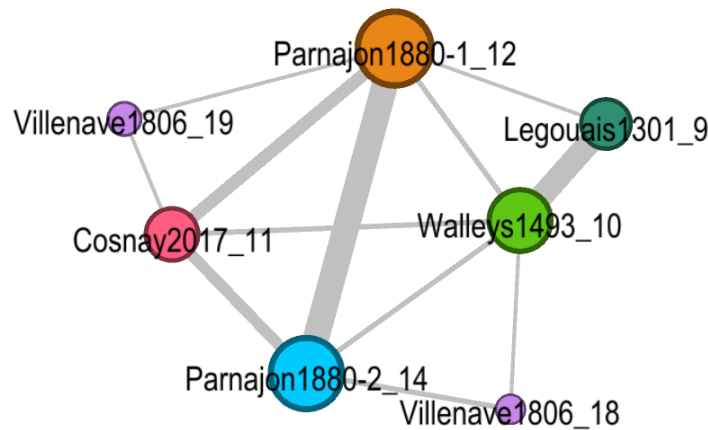


Figure 32. Graphe de l'épisode de la deuxième mort d'Enrydice

Isolé, ce graphe n'apporte peut-être pas d'informations supplémentaires à celles déjà exposées *supra*. En revanche, la visualisation de la totalité des relations, présentée à la Figure 33, permet quelques constats intéressants. Elle montre notamment les groupes des nœuds les plus larges, que nous plaçons aux périphéries afin d'optimiser la visualisation et de faciliter l'analyse plus détaillée de chaque groupe. Comme à la Figure 32, une couleur spécifique est attribuée à chaque texte afin de permettre un aperçu général de la structure des correspondances détectées. En revanche, les

étiquettes ne sont pas affichées, car, pour cette visualisation globale, elles nuisent à la lisibilité des nœuds. Beaucoup de relations ont été repérées uniquement entre deux textes : il s'agit majoritairement des correspondances entre Walleys1493 (vert clair) et Legouais1301 (vert foncé), mais nous recensons également deux relations entre Villenave1806 (violet) et Parnajon1880-2 (bleu) et une entre Parnajon1880-1 (orange) et Cosnay2017 (rose). L'épaisseur des liens est, comme à la Figure 32, dépendante du nombre de patterns impliqués dans la mise en correspondance.

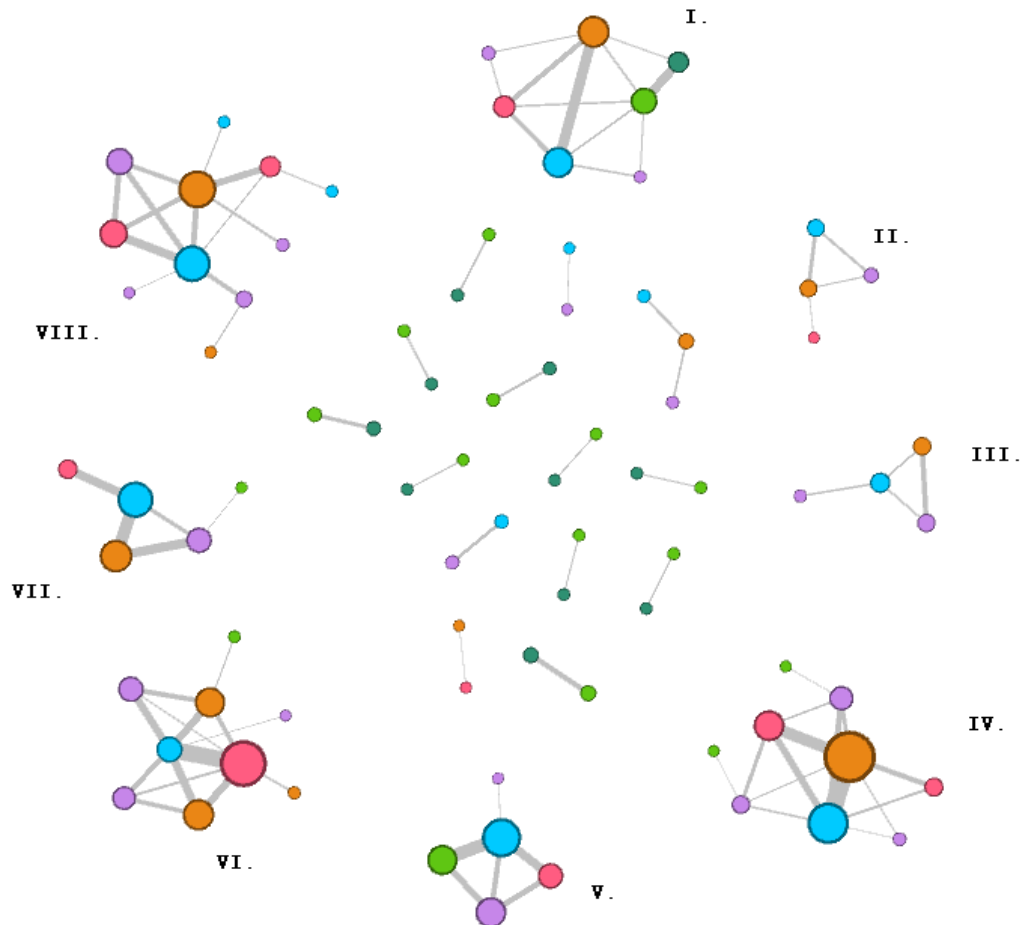


Figure 33. Visualisation spatialisée de la totalité des correspondances (Legouais1301 : vert clair ; Walleys1493 : vert foncé ; Parnajon1880-1 : orange ; Parnajon1880-2 : bleu ; Cosnay2017 : rose, Villenave1806 : violet)

Six graphes ont une taille supérieure à quatre nœuds, mais celui qui concerne l'épisode de la deuxième mort d'Eurydice (présenté en détail à la Figure 32 et marqué par **I.** à la Figure 33) est le seul qui regroupe tous les textes du corpus. Le graphe le plus large (**VIII.**) compte 11 nœuds : il s'agit du chant d'Orphée implorant les dieux des Enfers de lui rendre Eurydice. Parmi ces nœuds, 4 appartiennent à Villenave1806, 3 à Parnajon1880-2 et deux à Parnajon1880-1 et Cosnay2017, aucune correspondance n'a donc permis d'alignement avec les textes en ancien français. 14 liens au total sont établis entre les nœuds : moins que pour le graphe de l'épilogue (**VI.**), qui en compte 15 (mais

9 nœuds seulement) et qui intègre également deux extraits de Walleys¹⁴⁹³. Le troisième graphe le plus large correspond à la réaction des âmes résidant aux Enfers au chant d'Orphée (IV.), il compte 9 nœuds et 14 liens.

Ainsi, la visualisation spatialisée nous semble propice à l'analyse orientée vers un épisode ou un champ lexical en particulier, en permettant l'observation de représentations des relations détectées. Nous la retenons, dans une forme plus élaborée, pour notre prototype de l'édition numérique comparative et la présenterons de manière détaillée ultérieurement¹⁶. À ce stade, soulignons avant tout que le logiciel Phœbus présentait au moins trois avantages :

1. L'interface en ligne rendait ce type de traitement accessible à tous les chercheurs, sans qu'ils aient besoin d'acquérir des compétences informatiques spécifiques.
2. La possibilité de télécharger les résultats en format XML facilitait l'extraction des informations qui nous intéressaient et la compilation des résultats entre plusieurs couples de textes.
3. La précision de la nature des relations, notamment en ce qui concerne les stemmes communs permettaient d'en apprendre davantage sur chaque couple de textes, de se focaliser sur les champs lexicaux communs, etc.

Cependant, nous avons également vu quelques difficultés particulières (mis à part, bien sûr, le problème le plus important de l'indisponibilité du logiciel). Nous réalisons en effet que très peu de correspondances ont été détectées entre les textes en ancien français et les textes modernes, ainsi qu'au sein de certains épisodes, comme le début du livre X et la description du mariage d'Orphée et Eurydice ou le deuil et la mort d'Orphée. Nous pensons que les repérages des correspondances seraient nettement améliorés si, au lieu de détecter les stemmes communs, le logiciel proposait la modernisation et la lemmatisation des mots, voire la prise en compte de la synonymie.

Conclusion du chapitre

Toutes ces observations nous amènent à établir une liste des critères que nous présentons au Tableau 35. Elle se divise en cinq thèmes qui nous semblent principaux pour le choix optimal du logiciel à retenir pour notre projet : (i) la nature des recherches menées à l'aide du logiciel, (ii) les critères techniques nécessaires pour pouvoir l'utiliser (installation, équipement spécifique, etc.), (iii) le format attendu pour les données, (iv) les possibilités de paramétrage et (v) la qualité des résultats obtenus. Pour chacun de ces cinq aspects, nous nous posons des questions précises, que nous indiquons dans la deuxième colonne du Tableau 35.

¹⁶ Voir IV.3 « Présentation de prototype d'une édition comparative numérique ».

Dans la suite de ce chapitre, nous nous demandons lesquels des six principaux logiciels de détection des réutilisations, BLAST, Passim, Tesserae, TextPAIR, Textreuse et Tracer, répondent au mieux à ces critères. Pour faciliter la comparaison des résultats, nous utilisons le même petit corpus de six traductions d'Ovide.

Tableau 35. Critères de l'évaluation des logiciels de détection automatique des réemplois textuels

Recherches existantes	Le logiciel est-il adapté au public des chercheurs en humanités (numériques ou non) ?
	Des projets issus des humanités numériques utilisent-ils le logiciel ?
	Des projets utilisent-ils le logiciel pour exploiter un corpus littéraire ?
	Des projets utilisent-ils le logiciel pour exploiter un corpus en français ?
Critères techniques	Sur quel(s) système(s) d'exploitation le logiciel peut-il être installé ?
	Quelle est la difficulté d'installation et d'utilisation ?
	Existe-t-il une documentation dédiée (mode d'emploi, tutoriel, forum) ?
	Le logiciel continue-t-il d'être maintenu ? Peut-on échanger avec son auteur ou ses auteurs ? Quelle est la fréquence des mises à jour ?
Format	Un formatage spécifique du corpus à l'entrée est-il nécessaire ?
	Le format des résultats à la sortie est-il aisément exploitable ?
Paramétrage	Le logiciel propose-t-il de prendre en compte différentes formes lexicales (mots, lemmes, stemmes) voire de normaliser (ou moderniser) les graphies ?
	Le logiciel propose-t-il le choix entre différentes méthodes de comparaison (algorithmes d'alignement, calculs de similarité, vectorisation, etc.) ?
	Le logiciel propose-t-il le choix du degré de similarité minimum recherché (la taille des "trous", ordre des mots ignoré) ?
	Le corpus soumis à la comparaison peut-il être modulé (est-il possible de partitionner le corpus pour comparer un sous-ensemble avec un autre) ?
	Selon quels critères (métadonnées, identifiants...) ?
Qualité des résultats	Le nombre des relations détectées est-il satisfaisant ?
	Les relations détectées sont-elles pertinentes ?
	Le logiciel fournit-il les détails concernant la mise en relation (mots communs, score...) ?
	Le logiciel fournit-il des informations permettant l'identification des extraits dans le corpus plus globalement et au sein des textes auxquels ils appartiennent plus spécifiquement ?

2. Expérimentation avec BLAST, Tesseract, Textreuse et Passim

Nous présentons dans un premier temps les quatre logiciels que nous ne retenons pas (BLAST, Passim, Tesseract et Textreuse). Nous verrons que certains d'entre eux ne répondent pas du tout aux critères que nous nous sommes fixés, alors que d'autres s'avèrent assez performants, mais semblent pouvoir s'adapter mieux à d'autres types de corpus et à d'autres questions de recherche. Nous utilisons toujours notre petit corpus de six traductions d'Ovide : (i) en ancien français, en vers (Legouais1301¹⁷) et en prose (Walley1493¹⁸) ; (ii) juxtalinéaire, en vers (Parnajon1880-1¹⁹) et en prose (Parnajon1880-2²⁰) ; littéraire, moderne ou contemporaine, en vers (Cosnay2017²¹) et en prose (Ville-nave1806²²).

a) BLAST (Basic Local Alignment Search Tool)²³

BLAST est un logiciel spécifiquement dédié à la recherche des similarités entre des séquences biologiques (nucléotides ou protéines), ce qui permet d'établir des relations fonctionnelles et évolutives entre elles et d'aider, potentiellement, à identifier des gènes qui sont membres de mêmes familles. Il s'agit en réalité d'un ensemble de programmes, dont le premier a été publié en 1990 par des chercheurs du NCBI (*National Center for Biotechnology Information*), qui héberge toujours le projet. C'est donc l'outil le plus ancien parmi les six que nous avons retenus et, sans aucun doute, le plus connu : l'article fondateur d'Altschul *et al.*, « Basic local alignment search tool »²⁴ a été cité 72 978 fois selon le portail Scopus²⁵ et les versions les plus récentes de BLAST sont utilisées quotidiennement par les biotechnologues.

Malgré le fait que le logiciel n'est pas initialement dédié à la détection des similarités textuelles, nous trouvons tout de même quelques projets qui l'appliquent à un corpus textuel. Une de ces

¹⁷ Chrétien Legouais, *Ovide moralisé*, Rouen, Manuscrits de la BM de Rouen., 1301, 246v-249r, 250r-251r, 261v-261v, 271r-272r.

¹⁸ Thomas Walley, *La Bible des poètes. Métamorphose d'Ovide moralisée par Thomas Walley et traduite par Colard Mansion*, *op. cit.*

¹⁹ Ovide, *op. cit.*

²⁰ *Ibid.*

²¹ Ovide, *Les métamorphoses*, *op. cit.* Pour les questions des droits à la traduction, cette œuvre ne fait pas partie de notre corpus.

²² Ovide, *Les métamorphoses*, *op. cit.* La version utilisée pour cette expérimentation était celle disponible sur *Bibliotheca classica selecta* (<http://bcs.fltr.ucl.ac.be/meta/10.htm>). Elle est légèrement adaptée, et surtout modernisée. Au sein de notre corpus doctoral, nous revenons à la version de 1806.

²³ Disponible sur <https://blast.ncbi.nlm.nih.gov/Blast.cgi>. Voir par exemple S. F. Altschul *et al.*, « Basic local alignment search tool », *Journal of Molecular Biology*, vol. 215, n° 3, octobre 1990, p. 403-410, en ligne, <DOI : 10.1016/S0022-2836(05)80360-2> et Ian Korf, Mark Yandell et Joseph Bedell, *BLAST. An Essential Guide to the Basic Local Alignment Search Tool*, 1^{re} éd., Sebastopol, CA, O'Reilly & Associates, 2003.

²⁴ S. F. Altschul *et al.*, « Basic local alignment search tool », *loc. cit.*

²⁵ Voir <https://www.scopus.com/results/citedbyresults.uri?sort=plf-f&cite=2-s2.0-0025183708&src=s&imp=t&sid=11000707c77629dd0d137859ba952a11&sot=cite&sdt=a&sl=0&origin=inward&editSaveSearch=&txGid=a45495b85c031b94de6072bfc2df2374>.

recherches a été menée en 2018 par Aleksí Vesanto²⁶ qui s'intéresse à la détection des réemplois textuels au sein d'un corpus des journaux finlandais (1771-1910)²⁷. Pour pouvoir appliquer BLAST à un corpus textuel, il a été nécessaire de remplacer chacune des 23 lettres minuscules les plus fréquentes par une séquence d'acides aminés pour la reconnaissance desquelles BLAST a été conçu. Confronté au problème de mauvaise qualité de l'océrisation du corpus, principalement due à l'utilisation des lettres gothiques et la qualité du papier et de l'encre, Vesanto a constaté que BLAST est mieux adapté à son projet que les logiciels spécifiquement dédiés aux corpus textuels²⁸. Néanmoins, des modifications du logiciel initial, ainsi que le développement des ressources complémentaires, se sont avérées nécessaires.

Les ressources développées par Vesanto dans le cadre de sa recherche sont librement disponibles²⁹ et leur fonctionnement est présenté par exemple lors du 21st Nordic Conference of Computational Linguistics³⁰. Cependant, en suivant les étapes d'installation du logiciel³¹, nous rencontrons un problème de taille : il n'est plus possible d'accéder au lien de téléchargement de la version adaptée de BLAST³² et le tutoriel ne précise pas quelles modifications ont été apportées au logiciel initial. Le nom du fichier (« ncbi-blast-2.5.0+-src ») nous indique qu'il s'agisse de la version 2.5.0 de NCBI BLAST+ et que c'est une version qui nécessite une compilation à partir des sources (« -src »)³³. Nous pouvons la récupérer directement du site du porteur principal du projet BLAST³⁴ (en fouillant ses archives, car la dernière version disponible au moment de la rédaction de ce chapitre est 2.14.0), mais, dans ses instructions, Vesanto semble indiquer que ce fichier ait été modifié par ses soins (« A *modified* version of BLAST is also needed to run the software »). C'est probablement pour cela que nous n'avons initialement pas réussi à obtenir des résultats.

En suivant les instructions de Vesanto, nous avons structuré notre petit corpus d'essai dans un format demandé (fichier JSON compressé en format GZIP). Chaque phrase de chaque traduction

²⁶ Aleksí Vesanto, « Detecting and Analyzing Text Reuse with BLAST », Master, University of Turku, 2019.

²⁷ Outre ce projet, nous abordons plus en détail, voir aussi Paul Vierthaler et Mees Gelein, « A BLAST-based, Language-agnostic Text Reuse Algorithm with a MARKUS Implementation and Sequence Alignment Optimized for Large Chinese Corpora », *Journal of Cultural Analytics*, vol. 4, n° 2, 2019, en ligne, <<https://doi.org/10.22148/16.034>>.

²⁸ Pour la comparaison des résultats obtenus avec BLAST avec ceux proposés par Passim, voir Aleksí Vesanto et al., « Applying BLAST to Text Reuse Detection in Finnish Newspapers and Journals, 1771-1910 », *Proceedings of the NoDa-LiDa 2017 Workshop on Processing Historical Language*, Gothenburg, Linköping University Electronic Press, 2017, p. 54-58, en ligne, <<https://aclanthology.org/W17-0510>>.

²⁹ Voir <https://github.com/avjves/textreuse-blast>.

³⁰ Aleksí Vesanto et al., « A System for Identifying and Exploring Text Repetition in Large Historical Document Corpora », *Proceedings of the 21st Nordic Conference on Computational Linguistics*, Gothenburg, Sweden, Association for Computational Linguistics, 2017, p. 330-333, en ligne, <<https://aclanthology.org/W17-0249>>.

³¹ Voir <https://github.com/avjves/textreuse-blast#text-reuse-detection-with-blast>.

³² http://bionlp-www.utu.fi/blast_hum/ncbi-blast-2.5.0+-src.tar.gz.

³³ C'est également indiqué dans les instructions d'installation : « After decompressing the source files, BLAST must be compiled », voir <https://github.com/avjves/textreuse-blast#installation>.

³⁴ <https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.5.0/ncbi-blast-2.5.0+-src.tar.gz>.

est inscrite sur une ligne, accompagnée de son identifiant unique et de l'identifiant du texte dont elle provient :

```
{"doc-id": "cosnay2017_1", "title": "cosnay2017", "text": "couvert d'un manteau de safran à travers  
l'air immense s'en va hyménée et vers les rives thraces il avance."}
```

Le texte est mis en minuscules pour faciliter sa traduction en séquences d'acides-amino. Comme le français n'est pas pris en compte par défaut, nous ajoutons manuellement les 23 chaînes de caractères les plus fréquentes pour notre corpus³⁵. Nous lançons ensuite le traitement³⁶ et obtenons le dossier censé contenir les résultats d'alignement. Cependant, initialement, le dossier indiqué (« clusters/filled ») est vide et le traitement nous informe d'une erreur de paramétrage du logiciel BLAST³⁷, que nous avons résolu en modifiant le fichier « blast.py »³⁸.

Nous arrivons donc à obtenir les premiers résultats après environ 20 heures de manipulations informatiques, effectuées uniquement sur le système Linux (l'installation est *apriori* possible sur macOS également). Nous avons dû structurer notre corpus de manière spécifique, remplacer les majuscules par les minuscules et effectuer des traitements statistiques supplémentaires pour obtenir les 23 caractères les plus fréquents. Nous avons également modifié directement le code fourni par le développeur, en cherchant par nous-mêmes où ces modifications doivent être effectuées. Par ailleurs, nous n'avons pas pu utiliser la version de BLAST indiquée par le concepteur du logiciel. Enfin, les informations nécessaires pour l'installation ont dû être recherchées dans plusieurs ressources différentes : instructions d'installation, articles scientifiques, mémoire de recherche... Néanmoins, sans ce développement supplémentaire de Vesanto, nous n'aurions pas été en mesure d'expérimenter BLAST directement sur notre corpus, les tutoriels existants n'étant dédiés qu'au public des biotechnologues et la technicité du traitement étant trop élevée pour nos compétences.

Les résultats obtenus ne justifient pas le temps consacré à l'installation et le paramétrage du logiciel. Les données fournies prennent la forme d'un fichier en format JSON qui donne plusieurs informations sur les similarités détectées (voir l'extrait à la Figure 34). Avant d'expliquer cette structuration, remarquons que, dans les cas où des extraits similaires sont trouvés entre plus que deux textes, les correspondances sont regroupées en *clusters*, c'est-à-dire des groupes ou des amas. Sur

³⁵ Dans les instructions d'installation (voir <https://github.com/avjves/textreuse-blast#running>), il est seulement précisé qu'une autre langue qu'anglais ou finnois doit être ajoutée manuellement. Nous précisons que ce paramètre est pris en compte au sein du fichier « text_encoder.py », au sein de la fonction « make_mapping ». Pour obtenir la liste des caractères d'un document texte et leur fréquence avec un terminal Linux, nous utilisons la commande « cat corpus.txt | tr '[:upper:]' '[:lower:]' | sed 's/(.)/\n1/g' | sort | uniq -c ».

³⁶ La commande minimale, laissant la plupart des paramètres par défaut, est « python3 run_full.py language fr data_folder corpus output_folder resultats ».

³⁷ Voici la description de l'erreur fournie par le terminal : « BLAST query/options error: gap existence and extension values or 3 and 11 not supported for BLOSUM62. Supported values are: 32767, 32767 ».

³⁸ À la ligne 69, nous modifions les paramètres « -gapopen » et « gapextend » en fixant leurs valeurs à 32767.

166 amas de similarités détectées, 26 regroupent plus qu'un couple : nous décomptons 18 amas de trois extraits, sept de quatre et un de cinq.

```
{
  "cluster_0": {
    "length": 24,
    "hits": [
      {
        "title": "cosnay2017",
        "text": "ortel, s'il est possible",
        "node": "cosnay2017_11__47_69",
        "doc_id": "cosnay2017_11",
        "original_indices": [64,88],
        "encoded_indices": ["47", "69"]
      },
      {
        "title": "parmajon1880-2",
        "text": " ortel, si il est loisible",
        "node": " parmajon1880-2_5__82_104",
        "doc_id": "parmajon1880-2_5",
        "original_indices": [101, 126],
        "encoded_indices": ["82","104"]
      }
    ],
    "skips": []
  }
}
```

Figure 34. Exemple d'une correspondance détectée par BLAST

Au sein du fichier JSON, nous disposons notamment des informations concernant l'emplacement des extraits détectés (l'attribut « encoded_indices »). Par exemple, pour la phrase dotée de l'identifiant Cosnay2017_11, « Oh dieux du monde posé sous terre où l'on tombe quand on est mortel, s'il est possible, ne nous compliquons pas de paroles fausses [...] », l'extrait présenté à la Figure 34, pour lequel la similarité a été détectée avec un extrait de la phrase Parmajon1880-2_5, « ortel, s'il est possible » commence au caractère 47 et se termine au caractère 69 (l'extrait comporte alors 24 caractères, comme l'indique l'attribut « length »). Les résultats nous permettent donc de relier l'extrait à son ensemble, ce qui est un point positif. En revanche, plusieurs problèmes sont à noter au niveau de l'alignement des extraits textuels.

Tout d'abord, comme le logiciel ne compare que les chaînes de caractères, les mots au début ou à la fin de l'extrait sont souvent tronqués (« ortel » au lieu de « mortel »). Par ailleurs, les lettres accentuées caractéristiques du français ne sont pas bien encodées (« vip\u00e8re » au lieu de « vipère »).

Nous constatons également qu'il est impossible de prévenir l'alignement des extraits provenant du même texte, ce qui est dommage. Nous retrouvons par exemple un amas de quatre extraits issus du texte Walleys1493 : « rté, erudice demande seulement », « a tous, erudice demande seulement », « outa, erudice demande seulement » et « ours, erudice demande seulement ». Il s'agit du refrain de la chanson d'Orphée devant les Infernaux, répété quatre fois dans le texte. Sans que cela soit dépourvu d'intérêt, par exemple pour analyser le style d'un auteur, la détection des similarités au sein d'un même texte n'est pas souhaitable dans le cadre de notre recherche.

Enfin, le problème le plus important réside dans le fait que BLAST aligne uniquement des extraits quasiment identiques, pour la plupart de taille restreinte, ce qui rend impossible la détection de reformulations ou de paraphrases. Par exemple, au sein du plus grand amas, les cinq extraits suivants sont rapprochés : « our la première fois, vai » (Cosnay2017_30), « our la première fois leurs joue » (Parnajon1880-1_22), « our la première fois » (Parnajon1880-2_16), « pour la dernière fois » (Parnajon1880-2_24) et « our la première fois » (Villenave1806_28). Il est difficile de considérer qu'il s'agit ici d'une réutilisation textuelle significative, et non pas une répétition simple de la langue courante. Nous observons également que l'alignement des chaînes de caractères seules, sans prendre en compte des informations linguistiques, est problématique : nous trouvons au sein de l'amas de cinq similarités « pour la dernière fois », qui n'a que deux caractères de différence avec « pour la première fois » (d/p et n/m), alors que *dernier* et *premier* sont des antonymes.

Sur 166 similarités détectées, 38 ont la taille supérieure à 30 caractères, 16 sont supérieures à 40 et dix seulement dépassent 50 caractères. Celle la plus longue, incluse dans l'amas 144, en compte 109 : l'extrait « accuser de cruauté les dieux de l'érebe, il se retire sur le rhodope et sur l'hémus battu des aquilons » issu de la phrase 39 de Parnajon1880-1 est mis en correspondance avec celui, presque identique, de la phrase 46 de Villenave1806 : « accuser la cruauté des dieux de l'érebe, il se retire sur le mont rhodope, et sur l'hémus battu des aquilons ». Des extraits de cette phrase de Villenave1806 apparaissent d'ailleurs au sein de quatre autres amas et ils correspondent partiellement à celui relevé comme similaire à Parnajon1880-1_46 :

- (i) [142] « cruauté des dieux de l'éreb » / « t plaint les dieux de l'éreb » (Parnajon1880-2_29) / « ruauté les dieux e l'éreb » (Parnajon1880-1_39) ;
- (ii) [153] « il se retire sur le mont rhodop » / « il se retire sur le haut rhodop » (Parnajon1880-2_29) ;
- (iii) [50] « ont rhodope, et sur l'hémus battu de » / « ont rhodope, et sur l'hémus battu de » (Cosnay2017_47) ;
- (iv) [51] « t sur l'hémus batt » / « t sur l'hémus batt » (Parnajon1880-2_29) / « t sur l'hémus battu » (Parnajon1880-1_39) / « t sur l'hémus battu » (Cosnay2017_47).

Nous avons donc affaire à des résultats extrêmement morcelés et répétitifs, ce qui est particulièrement frappant quand nous observons les extraits dans leur contexte (Tableau 36).

Tableau 36. Similarités détectées par BLAST avec la phrase Villenave1806_46

Villenave1806_46 « enfin, las d'accuser la [cruauté des dieux de l'éreb] ¹⁴² e, [il se retire sur le m[ont rhodop] ¹⁵³ e, e[t sur l'hémus batt] ⁵¹ u de] ⁵⁰ s aquilons] ¹⁴⁴ . »	Cosnay2017_47 « il se retrouve en haut du m[ont rhodope, e[t sur l'hémus battu] ⁵¹ de] ⁵⁰ s vents. »
	Parnajon1880-1_39 « las enfin d'accuser de c[ruauté les dieux de l'éreb] ¹⁴² e, il se retire sur le rhodope élevé e[t sur l'hémus batt] ⁵¹ u des aquilons] ¹⁴⁴ . »
	Parnajon1880-2_29 « s'étan[t plaint les dieux de l'éreb] ¹⁴² e être cruels, [il se retire sur le haut rhodop] ¹⁵³ e e[t sur l'hémus batt] ⁵¹ u par les aquilons. »

La phrase 46 de Villenave1806 a été mise en correspondance trois fois avec la 39 de Parnajon1880, trois fois également avec la 29 de Parnajon1880-2 et deux fois avec la 47 de Cosnay2017. La segmentation des extraits est surprenante, par exemple dans l'amas 51, où le logiciel propose d'aligner « t sur l'hémus battu » de Cosnay2017 avec « t sur l'hémus batt » des trois autres textes. C'est d'autant plus surprenant que le segment « et sur l'hémus battu » est identique dans les quatre phrases.

Globalement, les extraits communs se suivent de très près (Parnajon1880-2_29) ou se recoupent (dans les trois autres phrases). Les résultats pourraient alors être simplifiés : au lieu de cinq amas présentant 13 extraits similaires pour quatre phrases, nous pourrions juste considérer que ces quatre phrases contiennent des similarités et doivent être alignées. Cependant, alors qu'un tel regroupement peut être obtenu informatiquement *a posteriori*, l'irrégularité et la répétitivité des résultats de BLAST complexifient cette tâche, sans que ses performances soient suffisantes pour nos besoins.

Nous écartons donc ce logiciel qui ne répond pas aux critères initialement fixés (Tableau 35). Les recherches portant sur des corpus textuels sont très limitées, et nous n'en avons recensée aucune qui utiliserait un corpus littéraire en français. La prise en main du logiciel est relativement difficile et il est possible que la qualité mitigée des résultats soit due à une erreur d'installation de notre part. Par ailleurs, il est nécessaire de formater le corpus de manière spécifique et, même si le format des résultats est plutôt facilement exploitable, leur structuration, et surtout leur répétitivité pose problème. Trop peu d'options de paramétrage concernent la détection des similarités en soi : on ne peut que fixer la longueur minimale et maximale des extraits détectés et préciser la taille des mots-clés à utiliser pour trouver les passages identiques³⁹. Enfin, la qualité des résultats est peu satisfaisante : le logiciel repère des passages identiques, qui sont pour la plupart des simples répétitions de la langue courante (« il reste assis », « il se retire ») et leur segmentation est problématique (mots coupés, quasiment les mêmes extraits sont mis en correspondance plusieurs fois).

b) Textreuse R⁴⁰

Textreuse est conçu par l'historien Lincoln Mullen avec le langage R⁴¹. Ce n'est pas un logiciel, mais un *package* (« un regroupement de fonctions et de données documentées »⁴²). Il ne peut donc

³⁹ Nous ne sommes pas certaines de bien comprendre ce paramètre. Dans les instructions d'installation, nous lisons « Setting for BLAST. This is the size of word seeds to use when finding repeated passages. Lowering will increase computation time, but if the data quality is bad, it might be necessary. Default=6. Range=2-7 », voir <https://github.com/avjves/textreuse-blast#running>.

⁴⁰ Voir <https://github.com/ropensci/textreuse> et Lincoln A. Mullen, *Textreuse: Detect Text Reuse and Document Similarity*, R, 2020.

⁴¹ « R est un environnement intégré de manipulation de données, de calcul et de préparation de graphiques [...]. Le R est un langage particulièrement puissant pour les applications mathématiques et statistiques (et donc actuarielles) puisque précisément développé dans ce but », voir Vincent Goulet, *Introduction à la programmation en R*, 5^e éd., Canada, Bibliothèque et Archives nationales du Québec, 2016, p. 2-3.

⁴² Sophie Baillargeon, *Utilisation de packages R*, Université de Laval, 2021, p. 1.

être utilisé qu'à condition de connaître le langage dans lequel il est codé et de disposer d'un mode d'emploi suffisamment détaillé pour comprendre ses différentes fonctionnalités.

D'après nos recherches, les projets exploitant Textreuse s'intéressent principalement aux domaines de l'histoire du droit⁴³ et de la sociologie des médias⁴⁴. Les auteurs de ces études soulignent avant tout l'efficacité de Textreuse pour le traitement des corpus très volumineux et apprécient le fait qu'il ne propose que les résultats les plus pertinents⁴⁵. Les instructions d'utilisation du logiciel, accompagnées des exemples du code, sont très accessibles, à condition de connaître les bases du langage R. Le traitement est facilement reproductible et adaptable à un autre corpus⁴⁶. Par ailleurs, aucune structuration spécifique de ce dernier n'est demandée, car Textreuse travaille sur le format du texte brut. En somme, il remplit de manière exemplaire les critères techniques que nous nous sommes fixés, ainsi que celles liées au formatage.

Notre expérience est bien plus mitigée du point de vue du paramétrage et de la qualité des résultats. Le package dispose principalement de deux fonctions. La première consiste en calcul de l'indice de similarité de Jaccard⁴⁷ pour chaque couple de textes, ce qui permet, en théorie, d'avoir un premier aperçu des relations qui existent au sein du corpus. Cependant, nous réalisons que les scores obtenus pour notre petit corpus des traductions sont assez décevants (Tableau 37).

Tableau 37. Résultats du calcul de l'indice de similarité de Jaccard avec Textreuse

Cosnay2017	Legouais1301	0,0000000000	Legouais1301	Walleys1493	0,0000000000
Cosnay2017	Parnajon1880-1	0,0017678256	Parnajon1880-1	Parnajon1880-2	0,0066425121
Cosnay2017	Parnajon1880-2	0,0000000000	Parnajon1880-1	Villenave1806	0,0000000000
Cosnay2017	Villenave1806	0,0005524862	Parnajon1880-1	Walleys1493	0,0000000000
Cosnay2017	Walleys1493	0,0000000000	Parnajon1880-2	Villenave1806	0,0022547914
Legouais1301	Parnajon1880-1	0,0000000000	Parnajon1880-2	Walleys1493	0,0000000000
Legouais1301	Parnajon1880-2	0,0000000000	Villenave1806	Walleys1493	0,0000000000
Legouais1301	Villenave1806	0,0000000000			

⁴³ Kellen Funk et Lincoln A. Mullen, « A Servile Copy: Text Reuse and Medium Data in American Civil Procedure », *Rechtsgeschichte - Legal History*, n° 24, Max-Planck-Institut fuer europäische Rechtsgeschichte, Klostermann Verlag, 2016, en ligne, <<http://dx.doi.org/10.12946/rg24/341-343>>.

⁴⁴ Damien Huffer et Shawn Graham, « The Insta-Dead: The rhetoric of the human remains trade on Instagram », *Internet Archaeology*, n° 45, 2017, en ligne, <<http://intarch.ac.uk/journal/issue45/5/index.html>>.

⁴⁵ « Dans notre cas, le traitement de 13 410 documents générerait 89 907 345 comparaisons uniques. L'avantage du package de Mullen est qu'il utilise des raccourcis computationnels (via un processus appelé le « hachage ») pour réduire le nombre de comparaisons à faire. Dans notre cas, le package a identifié 3 438 messages Instagram qui utilisaient les mêmes extraits de 5 mots » (« In our case, 13,410 documents so compared would result in 89,907,345 unique comparisons. The advantage of Mullen's package is that it uses computational shortcuts (via a process of 'hashing') to reduce the number of comparisons to be made. In our case, the package identified 3438 posts that used the same snippets of 5gram text », nous traduisons), *Ibid.*, p. 14.

⁴⁶ Nous n'avons pas réussi à reproduire la fonctionnalité « Minhashing and locality sensitive hashing » (voir <https://docs.ropensci.org/textreuse/#minhashing-and-locality-sensitive-hashing>) qui permet de réduire le nombre de comparaisons à effectuer. Cela est probablement dû au fait que notre corpus est bien trop petit pour que cela soit réellement utile. Pour reproduire notre traitement et regarder plus en détail l'algorithme que nous avons conçu dans le cadre de cette expérimentation, voir [III 2 textreuseR](#).

⁴⁷ Voir par exemple Pang-Ning Tan, Michael Steinbach et Vipin Kumar, *Introduction to data mining*, 2^e éd., Boston, Pearson Addison Wesley, 2018, p. 71-74.

La méthode de calcul Jaccard se base, comme celle de BLAST, uniquement sur la comparaison de chaînes de caractères (similarité syntaxique). Ce n'est pas suffisant, car le score est supérieur à zéro seulement pour quatre couples sur 15, et uniquement pour les textes en français moderne. Or, nous savons que des similarités existent entre tous les textes, mais elles sont surtout sémantiques (ressemblance de signification et non pas de forme). Le classement obtenu pour quatre couples sur huit semble toutefois pertinent : la traduction juxtalinéaire de Parnajon (Parnajon1880-1) est très proche de celle en prose (Parnajon1880-2), du même auteur et parue au sein du même ouvrage. La correspondance est également avérée entre deux traductions en prose fidèles au texte source, Parnajon1880-2 et Villenave1806, mais à moindre échelle (le score est divisé par trois). Enfin, la traduction moderne en vers, Cosnay2017, se rapproche davantage de celle juxtalinéaire Parnajon1880-1, alors que les relations qu'elle entretient avec celle en prose de Villenave1806 sont plutôt distendues. Ces résultats sont représentés à l'aide du graphe à la Figure 35.

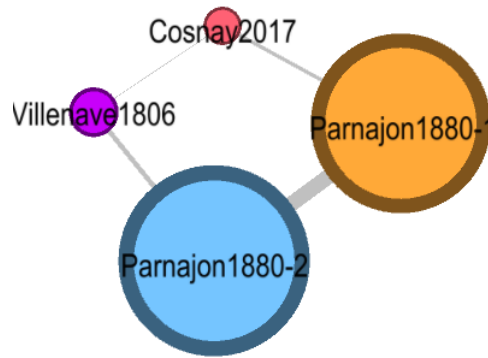


Figure 35. Visualisation des résultats du calcul de l'indice de similarité de Jaccard avec Textreuse

Alors que chacun de ces textes entretient une relation directe avec deux autres, leur force varie. La somme des relations est la plus élevée pour Parnajon1880-2 qui entretient une relation très forte avec Parnajon1880-1, et une plutôt modérée avec Villenave1806. La traduction Cosnay2017, quant à elle, est la moins proche des autres textes. Enfin, les nœuds qui se trouvent des côtés opposés du graphe, Villenave1806/Parnajon1880-1 et Cosnay2017/Parnajon1880-2, représentent les couples le plus éloignés : ils n'ont pas de liens directs, mais partagent des extraits communs avec deux autres textes.

Si nous comparons cette visualisation avec celles que nous avons présentées dans le chapitre précédent (voir Figure 32 et Figure 33), nous réalisons que les résultats obtenus ne sont pas suffisants. Le manque de deux textes en ancien français est marquant, d'autant plus que ces traductions devraient au moins être liées entre elles, étant deux versions différentes du même texte, *Ovide moralisé*. Pour mieux comprendre ce résultat, il convient de s'intéresser à la deuxième fonctionnalité de Textreuse, l'alignement local (*local alignment*).

L'alignement local permet de trouver, pour chaque couple de textes, les deux extraits présentant le plus de similarités. Ainsi, pour les 15 comparaisons effectuées entre nos six traductions, Textreuse ne nous présente que 15 couples d'extraits, alors que nous aurions voulu disposer de tous les extraits similaires et de pouvoir, éventuellement, en écarter certains grâce à un paramétrage spécifique. L'algorithme utilisé pour détecter les similarités, Smith-Waterman, est une autre méthode de calcul utilisée en bioinformatique pour aligner des séquences⁴⁸. Mullen indique que Textreuse propose une version de cet algorithme adaptée au traitement du langage naturel, mais nous n'avons pas trouvé d'informations supplémentaires quant à la nature de ces adaptations.

Pour chaque couple de correspondances, Textreuse fournit le score d'alignement Smith-Waterman, ainsi que les deux extraits correspondants. Les différences entre ces extraits sont marquées avec la suite des caractères « # », comme pour ce couple entre Cosnay2017 et Parnajon1880-2 :

Alignment score: 16

Document A : par ces lieux pleins d'effroi ## ##### par cet immense chaos par

Document B : par ces lieux pleins ##### de terreur par cet immense chaos par

L'alignement de deux textes à l'aide du caractère dièse montre que la différence consiste ici en remplacement « d'effroi » par « de terreur ». Cependant, cette méthode d'encodage de modifications ne reste lisible que pour les couples de taille relativement restreinte, et, rapidement, elle devient difficilement déchiffrable, comme pour cette relation entre Cosnay2017 et Parnajon1880-1 :

Document A : je ne suis pas ##### descendu ici pour voir ### ## ##### Le ##### Tartare ni pour vaincre ### ##### les trois gorges ##### poilues de serpents du monstre fils ### de Méduse La ##### cause ##### de ##### ## ##### ### ##### mon voyage est ##### ma ##### femme ##### Une vipère foulée au pied a répandu en elle son poison ##### et lui a pris ##### les années à ##### venir J'ai voulu ##### supporter ## je ne dirai ##### pas que ### je ##### n'ai ##### pas essayé L'amour a vaincu

Document B : je ne suis ### point descendu ici pour ##### que je visse le sombre Tartare ni pour ##### que j'enchaînasse les trois ##### gosiers ##### ## ##### du monstre ##### issu de Méduse ## gosiers ##### hérissés de serpents mon épouse est cause de mon voyage ### dans ## laquelle ##### épouse une vipère foulée ## ##### a répandu ## ##### son ##### venin et lui a ##### enlevé les années # croissantes ##### J'ai voulu pouvoir supporter et je ne ##### nierai pas ### moi ## l'avoir ##### tenté ### ##### l'amour a vaincu

Nous remarquons par exemple que le suivi des modifications devient particulièrement difficile si une de deux traductions est plus descriptive (« pour vaincre » / « pour que j'enchaînasse ») ou

⁴⁸ Voir Temple F. Smith et Michael S. Waterman, « Identification of common molecular subsequences », *Journal of Molecular Biology*, vol. 147, n° 1, mars 1981, p. 195-197, en ligne, <<https://www.sciencedirect.com/science/article/pii/0022283681900875>>. Outre l'algorithme implémenté dans le logiciel BLAST, dans la partie précédente nous avons également présenté celui de Needleman-Wunsch (Voir II3c), § « Alineur Odysseus », apparenté au Smith-Waterman. Alors que ce premier ne permet que d'alignements des séquences entières, Smith-Waterman est capable d'isoler segments les plus similaires entre deux séquences. Concrètement, alors qu'au sein de l'aligneur Odysseus, Reboul devait segmenter son corpus au préalable, Textreuse isole lui-même les extraits à aligner.

quand les segments de l'original ne sont pas traduits dans le même ordre (« les trois gorges poilues de serpents du monstre fils de Méduse » / « les trois gosiers du monstre issu de Méduse, gosiers hérissés de serpents »). Cette méthode d'encoder les modifications est aussi difficilement exploitable avec des méthodes computationnelles, par exemple si l'on voudrait récupérer la liste des remplacements, suppressions et insertions ou, tout simplement, celle des mots communs. Notons toutefois que le logiciel est capable de détecter les similarités dans de longs extraits, malgré les différences parfois assez importantes. En cela, Textreuse dépasse largement les résultats obtenus avec BLAST, mais aussi ceux de Phœbus, qui détecte la similarité entre les deux extraits seulement à partir de « monstre issu/fils de Méduse ».

La totalité des résultats fournis par Textreuse est présentée en Tableau 38. L'affichage initial de Textreuse est modifié pour mieux montrer les différences et similarités observables entre chaque couple. Ainsi, les extraits correspondants sont mis sur la même ligne et les différences entre eux sont soulignées en gris. Ces résultats sont classés par l'ordre décroissant du score de similarité Smith-Waterman. La ponctuation a été supprimée lors du traitement par Textreuse et les différences de casse ont été ignorées.

Tableau 38. Résultats (remaniés) du traitement avec Textreuse

Source	Cible	Score	Extrait source	Extrait cible
Parnajon1880-1	Parnajon1880-2	55	De là Hyménée voilé d'un manteau couleur de safran s'éloigne à travers l'air immense et il se dirige vers les rivages des Ciconiens et il est appelé en vain par la voix d'Orphée Il fut présent à la vérité mais il n'apporta ni paroles solennelles ni visages joyeux ni présage heureux La torche aussi qu'il tint fut sans cesse crierde par la fumée qui fait pleurer et elle ne trouva aucuns feux par les mouvements L'issue fut plus funeste que l'augure car tandis que la nouvelle mariée se promène à travers les herbes accompagnée d'une troupe de naiades elle tombe la dent d'un serpent ayant été reçue dans son talon Laquelle après que le chantré du Rhodope eut pleurée suffisamment vers les airs supérieurs afin qu'il	De là Hyménée vêtu d'une robe couleur de safran s'éloigne à travers les plaines immenses de l'air et se dirige vers les rivages des Ciconiens où l'appelle en vain la voix d'Orphée Il vient il est vrai mais sans proférer les paroles solennelles sans apporter un front joyeux ni d'heureux présages La torche même qu'il tient ne cesse de jeter en pétillant une fumée qui remplit les yeux des larmes et elle ne peut s'allumer bien qu'il l'agite L'événement fut encore plus triste que les présages Un jour que la nouvelle épouse se promenait dans les prairies accompagnée d'une troupe de naiades elle tombe mordue au talon par un serpent Longtemps le chantré du Rhodope l'a pleuré sur la terre Enfin il veut aussi tenter de fléchir les ombres et il ose

Source	Cible	Score	Extrait source	Extrait cible
			tentât aussi les ombres il osa descendre vers le Styx par la porte du Ténare et il	descendre vers le Styx par la porte du Ténare Il
Parnajon1880-1	Cosnay2017	42	je ne suis point descendu ici pour que je visse le sombre Tartare ni pour que j'enchaînasse les trois gosiers du monstre issu de Méduse gosiers hérissés de serpents mon épouse est cause de mon voyage dans laquelle épouse une vipère foulée a répandu son venin et lui a enlevé les années croissantes J'ai voulu pouvoir supporter et je ne nierai pas moi l'avoir tenté l'amour a vaincu	je ne suis pas descendu ici pour voir Le Tartare ni pour vaincre les trois gorges poilues de serpents du monstre fils de Méduse La cause de mon voyage est ma femme Une vipère foulée au pied a répandu en elle son poison et lui a pris les années à venir J'ai voulu supporter je ne dirai pas que je n'ai pas essayé L'amour a vaincu
Villenave1806	Parnajon1880-2	31	Enfin las d'accuser la cruauté des dieux de l'Érèbe il se retire sur le mont Rhodope et sur l'Hémus battu des Aquilons	enfin d'accuser de cruauté les dieux de l'Érèbe il se retire sur le Rhodope élevé et sur l'Hémus battu des Aquilons
Legouais1301	Walleys1493	24	ne prent fors vent vain et vuide Cele se part de son mari qui de seconde mort mori Mes ne se puet de lui blasmer se ne se plaint de trop amer Le desrain salut li rendi que cil a paines entendi Orpheüs forment se doloit pour sa double mort et voloît Retorner pour querre la morte mes il trouva fermé la porte et le portier qui le gardoit qui son oirre li retardoit C'est noient que jamais le vait quant vit que plus	ne print fors vent et ainsi se partit Erudice de son amy et mourut de seconde mort Mais de lui ne se peut plaindre fors de trop aimer Le dernier salut luy rendit que a peines l'entendit Orpheus Forment se plaignoit de la seconde mort de s'amie et voulut retourner pour trouver la mort mais la porte trouva fermee Et le portier qui la gardoit lui retarda son chemin et si lui dist que jamais recouvrer ne la pourroit Quant Orpheus vit que plus
Villenave1806	Parnajon1880-1	23	dieux de l'Érèbe il se retire sur le mont Rhodope et sur l'Hémus battu	dieux de l'Érèbe être cruels il se retire sur le haut Rhodope et sur l'Hémus battu
Parnajon1880-2	Cosnay2017	16	par ces lieux pleins de terreur par cet immense Chaos	par ces lieux pleins d'effroi par cet immense chaos par
Villenave1806	Cosnay2017	16	dieux de l'Érèbe il se retire sur le mont Rhodope et sur l'Hémus battu des Aquilons Trois fois le soleil avait	dieux de l'Érèbe sont cruels Il se retrouve en haut du mont Rhodope et sur l'Hémus battu des vents Pour la troisième fois le Titan avait
Walleys1493	Parnajon1880-1	7	de la seconde mort de	de la mort de
Walleys1493	Villenave1806	7	de sa voix ensemble de sa herpe les	de sa voix accompagnés des sons plaintifs de sa lyre les

Source	Cible	Score	Extrait source	Extrait cible
Legouais1301	Cosnay2017	6	dire je ne vieng pas	dire vrai je ne suis pas
Walleys1493	Cosnay2017	6	sur la rive	sur la rive
Walleys1493	Parnajon1880-2	6	sur la rive	sur la rive
Legouais1301	Parnajon1880-1	6	est cause de	est cause de
Legouais1301	Parnajon1880-2	5	se part de son	se plaindre de son
Legouais1301	Villenave1806	5	de dolour et de	de Proserpine et de

Tout d'abord, notons qu'une longue correspondance a été détectée entre les deux textes en ancien français, Legouais1301 et Walleys1493, à la quatrième place (avec le score de 24). C'est curieux, car le calcul de Jaccard ne trouvait pas de similarités entre ces deux textes (voir Tableau 37). Par ailleurs, les différentes graphies des mêmes mots (« print » / « prent », « mori » / « mourut », « puet » / « peut », etc.) ne sont pas prises en compte par le logiciel. Ainsi, le couple d'extraits à la ligne suivante, celui présentant la relation la plus forte pour le couple Villenave1806 / Parnajon1880-1, a obtenu presque le même score (23), alors que les extraits sont beaucoup plus courts (168 vs 30 mots). En effet, ces deux passages sont très similaires : nous n'observons qu'un ajout de « être cruels » chez Villenave1806⁴⁹ et le remplacement de « mont » chez Villenave1806⁵⁰ par « haut » chez Parnajon1880-1.

En outre, les résultats dont le score est égal ou inférieur à six sont extrêmement courts et s'apparentent plutôt à des constructions fréquentes (« sur la rive », « est cause de ») ou des syntagmes aux constructions proches syntaxiquement, mais dont le sens est très différent (« se part de son » / « se plaindre de son », « de doulour et de » / « de Proserpine et de »). Curieusement, cela n'est pas tout à fait le cas pour les deux couples dont le score est égal à sept. Pour le premier, Walleys1493 et Parnajon1880-1, les extraits restent courts, mais portent sur un des mots-clés du mythe (*mort*), ce qui permet d'identifier rapidement la partie du récit qui est concernée par la correspondance. Pour le deuxième, entre Villenave1806 et Walleys1493, nous avons l'impression que le logiciel a réussi à détecter une similarité sémantique : les mots communs sont majoritairement des mots faibles (« de sa voix # de sa # les »), mais on constate une relation synonymique entre « accompagné de » et « ensemble », ainsi qu'une relation métonymique entre « lyre » et « harpe ».

Il est regrettable que l'on ne dispose pas de plus d'informations sur la méthode de l'alignement, et notamment sur l'adaptation de l'algorithme de Smith-Waterman au traitement de la langue naturelle. La fréquence des mots dans le corpus est-elle prise en compte ? Les mots faibles ont-ils le même poids que les mots forts ? Des méthodes permettant d'améliorer la détection des relations

⁴⁹ Il ne s'agit pas pour autant d'une omission de Villenave1806. La traduction juxtalinéaire Parnajon1880-1 vise à coller au plus près à la syntaxe de l'original, ce qui résulte parfois en constructions assez surprenantes. Pour cet extrait, la phrase débute avec un gérondif, puis se poursuit avec un syntagme nominal et l'infinitif : « S'étant plaint les dieux de l'Érèbe être cruels ». Dans Villenave1806, la phrase débute par « Enfin, las d'accuser la cruauté des dieux de l'Érèbe ».

⁵⁰ Le logiciel présente les résultats dans l'ordre alphabétique des identifiants, que nous suivons ici, et non pas dans celui chronologique de la publication des traductions, qui serait le plus pertinent pour l'analyse des correspondances.

sémantiques ont-elles été implémentées ? Le mode d'emploi ne permet pas de répondre à ces questions et aucun article scientifique publié par l'auteur ne rentre dans le détail du fonctionnement du logiciel, ce qui est dommage.

En effet, Textreuse aurait été un candidat sérieux pour notre projet s'il était davantage documenté, et à condition de fournir tous les alignements détectés et non seulement un par couple de textes. Il serait également utile de pouvoir fixer le score minimum et de signaler les différences et les similarités d'une manière plus structurée que ce que l'on nous propose actuellement. Autrement, nous apprécions la qualité des alignements locaux dont nous disposons et la facilité de l'installation du logiciel. Textreuse nous paraît particulièrement adapté aux corpus de grande taille. Il pourrait constituer une première étape de processus de la détection des réutilisations textuelles, en permettant de restreindre les recherches uniquement aux textes entre lesquels des relations significatives ont été détectées.

c) Tesseract⁵¹

Le projet Tesseract est le fruit de collaboration internationale, mené par le Département des Lettres classiques de l'Université de Buffalo, le Département de l'Informatique et de l'ingénierie de l'Université de Notre-Dame (États-Unis), ainsi que le Département des Sciences de l'Antiquité de l'Université de Genève. Ses porteurs principaux, Neil Coffee, Walter J. Scheirer et Jean-Pierre Koenig ont publié un bon nombre de ressources extrêmement utiles tant pour la compréhension du fonctionnement du logiciel⁵² que pour l'appréciation des résultats obtenus⁵³. D'autres chercheurs, en particulier spécialistes des Lettres classiques, ont utilisé Tesseract par exemple pour présenter des interactions entre des poètes écrivant en hexamètre latin⁵⁴, comparer l'influence relative des auteurs

⁵¹ Disponible sur <https://tesseract.nd.edu/>. Voir par exemple Neil Coffee et al., « The Tesseract Project: intertextual analysis of Latin poetry », *Literary and Linguistic Computing*, vol. 28, n° 2, 2012, p. 221-228, en ligne, <https://www.academia.edu/3830928/The_Tesseract_Project_Intertextual_Analysis_of_Latin_Poetry>.

⁵² Voir par exemple Christopher Forstall et al., « Modeling the scholars: Detecting intertextuality through enhanced word-level n-gram matching », *Digital Scholarship in the Humanities*, vol. 30, n° 4, 2015, p. 503-515, en ligne, <doi: 10.1093/llc/fqu014>; Christopher W. Forstall et Walter J. Scheirer, *Quantitative Intertextuality: Analyzing the Markers of Information Reuse*, Cham, Springer International Publishing, 2019.

⁵³ Voir par exemple Christopher W. Forstall et al., « Evidence of Intertextuality: Investigating Paul the Deacon's *Augustae Vitae* », *Literary & Linguistic Computing (LLC)*, 2011, p. 285-296; Neil Coffee et al., « Intertextuality in the Digital Age », *loc. cit.*; Neil Coffee et Christopher W. Forstall, « Claudian's engagement with Lucan in his historical and mythological hexameters », dans Valéry Berliacourt, Lavinia Galli Milić et Damien Nelis (dir.), *Lucan and Claudian: context and intertext*, Winter Verlag, Heidelberg, 2016, p. 255-284; Neil Coffee, « Intertextuality as Viral Phrases: Roses and Lilies », De Gruyter Saur, 2019.

⁵⁴ Neil Bernstein, Kyle Gervais et Wei Lin, « Comparative rates of text reuse in classical Latin hexameter poetry », *Digital Humanities Quarterly*, vol. 9, n° 3, décembre 2015.

classiques⁵⁵, ou encore pour investiguer sur des allusions intergénériques détectables entre l'épopée de Lucan et les élégies de Tibulle, Propertius et Ovide⁵⁶.

Toutes ces études ont un point commun : elles portent exclusivement sur le corpus littéraire composé d'auteurs classiques, écrivant en latin ou en grec, en vers ou en prose, et dont les textes ont été pour la plupart écrits entre 207 avant J.-C et 600 après J.-C. En effet, alors que le développement dispose d'une interface Web⁵⁷, ainsi que d'une version indépendante pouvant être installée localement⁵⁸, il n'est pas possible de l'appliquer à un autre corpus que celui déjà disponible. Les utilisateurs disposent donc d'une base des résultats obtenus avec le logiciel conçue dans le cadre du projet, mais n'ont pas d'accès direct à celle-ci.

Pour cette raison, malgré la richesse des recherches existantes et la description précise du fonctionnement du logiciel de détection des réutilisations textuelles, nous sommes dans l'impossibilité d'évaluer Tesseract selon les critères que nous avons établis et l'écartons donc de notre sélection. Notons toutefois que l'excellent ouvrage *Quantitative Intertextuality : Analyzing the Markers of Information Reuse* de Forstall et Scheirer⁵⁹, que nous avons déjà cité à plusieurs reprises, constitue une ressource extrêmement précieuse, notamment pour l'établissement d'un système d'évaluation des résultats obtenus et pour l'amélioration de ces résultats. Nous y reviendrons de manière détaillée à la fin de cette partie⁶⁰.

d) Passim⁶¹

Passim est un logiciel conçu et développé depuis 2012 par David A. Smith de l'Université de Northeastern (Boston) grâce au soutien d'Andrew W. Mellon Foundation⁶² et de la Fondation nationale pour les sciences humaines⁶³. Il est principalement utilisé par des historiens pour traiter des corpus massifs.

⁵⁵ James O Gawley et A Caitlin Diddams, « Comparing the intertextuality of multiple authors using Tesseract: A new technique for normalization », *Digital Scholarship in the Humanities*, vol. 32, n° suppl_2, décembre 2017, p. ii53-ii59, en ligne, <<https://doi.org/10.1093/lc/fqx038>>.

⁵⁶ Patrick J. Burns, « Measuring and Mapping Intergeneric Allusion in Latin Poetry using Tesseract », *Journal of Data Mining & Digital Humanities*, vol. Special Issue on..., n° Towards a Digital Ecosystem:..., août 2017, p. 3821, en ligne, <<https://jdm.dh.episciences.org/3821>>.

⁵⁷ Voir <https://tesseract.caset.buffalo.edu/> [archivé sur <https://web.archive.org/web/20250525083114/https://tesseract.caset.buffalo.edu/>].

⁵⁸ Voir <https://github.com/tesseract/tesseract>.

⁵⁹ Christopher Forstall et Walter Scheirer, « What is Quantitative Intertextuality ? », *op. cit.*

⁶⁰ Voir III.4a) « Évaluation des résultats ».

⁶¹ Disponible sur <https://github.com/dasmiq/passim>. Voir par exemple David A. Smith, Ryan Cordell et Elizabeth Maddock Dillon, « Infectious texts: Modeling text reuse in nineteenth-century newspapers », *2013 IEEE International Conference on Big Data*, IEEE, 2013, p. 86-94; David A. Smith et al., « Detecting and modeling local text reuse », *IEEE/ACM Joint Conference on Digital Libraries*, IEEE, 2014, p. 183-192.

⁶² Voir <https://www.mellon.org/>.

⁶³ Voir <https://www.neh.gov/>.

Le projet *impresso*⁶⁴, par exemple, met à disposition des chercheurs un corpus des numéros des 76 journaux suisses et luxembourgeois parus entre 1738 et 2017. Ce grand corpus est interrogeable au sein d'une interface Web⁶⁵ qui permet, notamment, d'explorer les relations intertextuelles détectées entre différents articles grâce à Passim. Marten Düring *et al.* expliquent leur choix par la capacité du logiciel « d'évoluer, grâce à l'architecture de calculs en parallèle »⁶⁶. Par ailleurs, « [d]es tests préliminaires sur le corpus *impresso* ont montré que l'alignement par la recherche approximative permettait à Passim de détecter la réutilisation malgré la présence (modéré) de bruit OCR »⁶⁷.

Le logiciel est également utilisé pour exploiter les archives de l'économiste et philosophe britannique John Stuart Mill à partir de ses emprunts au sein de The London Library⁶⁸, mais conjointement avec un autre logiciel, TextPAIR, que nous présentons dans le chapitre suivant. Helen O'Neill *et al.* ne précisent pas la raison pour laquelle ils décident de recourir aux deux logiciels, mais déclarent que Passim « détecte les couples pertinents là où les variations textuelles et les erreurs OCR rendent les approches plus simples moins efficaces. En revanche, ces méthodes sont plus coûteuses en temps et en mémoire »⁶⁹.

Enfin, Passim a aussi été adapté pour le traitement d'un corpus de textes en arabe au sein du projet Kitab (*Knowledge, Information Technology, & the Arabic Book*)⁷⁰. Alors que le choix du logiciel n'est pas expliqué, les chercheurs engagés au sein du projet précisent que « l'algorithme identifie naïvement la réutilisation [...]. Cela implique que les résultats produits par Passim répondront mieux à certaines questions de recherche plutôt qu'à d'autres. [...] [D]ans les cas où un auteur paraphrase principalement un travail antérieur, Passim est peu susceptible d'être utile »⁷¹.

⁶⁴ Voir <https://impresso-project.ch/> et Maud Ehrmann et al., *Language Resources for Historical Newspapers: the Impresso Collection*, Marseille, European Language Resources Association (ELRA), 2020, p. 958-968, en ligne, <<https://www.aclweb.org/anthology/2020.lrec-1.121.pdf>> ; Marten Düring et al., *Impresso Text Reuse at Scale*, 2023, en ligne, <<https://hal.science/hal-04151808>>.

⁶⁵ Voir <https://impresso-project.ch/app/>.

⁶⁶ « [The reason for choosing Passim over existing alternatives was its ability to] scale up, guaranteed by the software's parallel computing architecture. » (nous traduisons). Voir Marten Düring et al., *Impresso Text Reuse at Scale*, p. 4.

⁶⁷ « Preliminary tests on the *impresso* corpus showed that Passim's fuzzy alignment algorithm was able to detect reuse despite the presence of (moderate) OCR noise. » (nous traduisons). *Ibid.* Curieusement, le même argument est utilisé par Aleks Vesanto pour justifier le choix de BLAST, voir Aleks Vesanto et al., « Applying BLAST to Text Reuse Detection in Finnish Newspapers and Journals, 1771-1910 », paragr. 3.1 BLAST evaluation.

⁶⁸ Helen O'Neill et al., « Text mining Mill: Computationally detecting influence in the writings of John Stuart Mill from library records », *Digital Scholarship in the Humanities*, vol. 36, n° 4, décembre 2021, p. 1013-1029, en ligne, <<https://doi.org/10.1093/lhc/fqab010>>.

⁶⁹ « Passim detects pairs worth aligning where textual variation and OCR errors mean that more straightforward approaches are less robust, but it is therefore more computationally expensive in both memory and time » (nous traduisons). Voir Matteo Romanello et Simon Hengchen, « Detecting Text Reuse with Passim », *loc. cit.*

⁷⁰ Voir <https://kitab-project.org/> et Mathew Barber, « Adventures in Alignments: Training an Algorithm to Recognise Text Reuse », dans *KITAB*, 7 août 2020, en ligne, <<https://kitab-project.org/Adventures-in-Alignments-Training-an-Algorithm-to-Recognise-Text-Reuse/>>.

⁷¹ « The algorithm naively identifies reuse by looking for instances of shared text [...]. This means that the text reuse data produced by Passim suits certain research questions better than other. [...] [I]n cases in which an author mostly paraphrases an earlier work, Passim is unlikely to be helpful » (nous traduisons). Voir <https://kitab-project.org/methods/text-reuse#introducing-passim>.

En plus des articles scientifiques qui décrivent le fonctionnement du logiciel et son emploi au sein des projets de recherche, l'article « Détecter la réutilisation de texte avec Passim »⁷² de Romanello et Hengchen propose un guide exhaustif pour l'installation, la préparation des données, l'exécution et l'utilisation des résultats de Passim. Ce tutoriel pas-à-pas, destiné au public des chercheurs des humanités numériques, notamment des historiens, ne nécessite que quelques notions informatiques de base, comme l'utilisation de l'environnement Bash et du langage Python.

Romanello et Hengchen argumentent leur choix de Passim plutôt qu'un autre logiciel de détection des réutilisations de par son adaptation aux différentes tailles de corpus et différents types de projets, comme :

- (i) la recherche de plusieurs éditions d'une même œuvre au sein d'une bibliothèque numérique ;
- (ii) la recherche des citations dans un corpus textuel ;
- (iii) l'étude de la diffusion des textes et de leur viralité (comme dans le projet The Viral Texts⁷³) ;
- (iv) l'identification et le filtrage des documents en doublon.

Il s'agit donc principalement de la recherche des passages ou textes identiques présents au sein d'un corpus (copies, répétitions, plagats, citations, etc.), même si les auteurs mentionnent également que « Passim peut [...] regrouper des copies d'un même article qui ne diffèrent que par des erreurs de reconnaissance optique de caractères (OCR), mais il peut aussi aider à retrouver des textes qui partagent le même modèle journalistique, comme des horoscopes ou des publicités »⁷⁴.

Plusieurs articles mentionnent donc l'insuffisance de Passim pour la détection des relations plus complexes qu'un simple copier-coller. Nous verrons si notre expérimentation nous mène à la même conclusion.

Cependant, il est d'abord nécessaire de faire fonctionner le logiciel, ce qui s'avère difficile malgré le déroulé pas-à-pas de Romanello et Hengchen. D'après les recommandations des auteurs, nous procédons à l'installation sur le système Linux⁷⁵. Nous vérifions ensuite l'installation⁷⁶ et obtenons la même sortie que les auteurs du tutoriel. Nous lançons le logiciel et le traitement s'achève sans qu'aucune erreur soit signalée. Néanmoins, le dossier censé contenir les résultats du traitement reste

⁷² Matteo Romanello et Simon Hengchen, « Detecting Text Reuse with Passim », *loc. cit.*

⁷³ Voir notamment <https://viraltxts.org/> et l'article déjà mentionné de David A. Smith, Ryan Cordell et Elizabeth Maddock Dillon, « Infectious texts ».

⁷⁴ Matteo Romanello et Simon Hengchen, « Detecting Text Reuse with Passim », *loc. cit.*, part. Introduction à la réutilisation de texte.

⁷⁵ « Notez que l'installation de Passim sur Windows est plus difficile que celle pour macOS ou pour Linux. Par conséquent, nous vous recommandons d'utiliser macOS ou Linux (ou une machine virtuelle) pour cette leçon », voir *Ibid.* Nous suivons la partie « Instructions pour Linux ».

⁷⁶ Voir partie « Vérifiez l'installation ».

désespérément vide, que ce soit pour les tentatives effectuées avec notre petit corpus des traductions ou celles avec le fichier de test fourni dans l'article⁷⁷.

En cherchant la solution à notre problème, nous réalisons que la même difficulté a été signalée en 2018 par un autre utilisateur sur le site GitHub du logiciel⁷⁸, mais elle est restée sans réponse. En revanche, nous trouvons sur ce même site une nouvelle version du tutoriel⁷⁹ proposée par le développeur du logiciel, David Smith, et un Jupyter Notebook⁸⁰ qui remplit la même fonction, mais qui facilite davantage l'installation et la prise en main du logiciel.

Jupyter Notebook (ci-après : calepin) est un programme qui contient à la fois le code et le texte explicatif. Le code peut être exécuté morceau par morceau pour afficher progressivement les résultats et l'utilisateur peut modifier le code pour l'adapter à ses besoins. C'est donc un excellent outil pédagogique qui ne nécessite pas de connaissances informatiques très pointues et où l'utilisateur est très accompagné dans sa découverte du fonctionnement d'un logiciel. Alors que lors d'un tutoriel, l'utilisateur doit recopier les commandes dans les environnements dédiés (exécuteur des commandes), parfois en modifiant certaines valeurs, dans un calepin, il suffit d'un clic. Cela permet d'éviter certaines erreurs humaines et les difficultés liées au fait que l'utilisateur ne travaille pas dans le même environnement que l'auteur du tutoriel.

L'autre avantage du calepin proposé par Smith est qu'il utilise une bibliothèque Python de *Passim*. L'outil est donc bien plus facile à installer : une seule commande⁸¹ suffit pour remplacer une dizaine d'étapes détaillées dans le tutoriel de Romanello et Hengchen (télécharger la version indépendante du logiciel, installer les dépendances comme Java et Spark, puis compiler le code-source).

Par ailleurs, Smith présente les fonctionnalités du logiciel avec des exemples concrets des textes littéraires (deux poèmes de Longfellow), commente les résultats et explique les aspects techniques, comme la structure des données et les paramétrages du logiciel. En somme, ce calepin répond de manière exemplaire au besoin de documentation « clés en main », explicite et dédiée aux chercheurs non-experts en informatique. Nous parvenons à le faire fonctionner et à l'adapter à nos besoins sans encombre⁸². En suivant les instructions de Smith, nous procédons à deux types de traitements.

⁷⁷ Voir partie « Format JSON de base ». Le fichier `test.json` contient deux lignes et vise à comparer les phrases « *This is the text of a document* » et « *This is the text of another document.* ».

⁷⁸ Voir <https://github.com/dasmiq/passim/issues/4>.

⁷⁹ Voir <https://github.com/dasmiq/ph-submissions/blob/passim2-replace/en/published/originals/detecting-text-reuse-with-passim.md>.

⁸⁰ Voir https://github.com/dasmiq/passim/blob/seriatim/docs/passim_quickstart.ipynb.

⁸¹ Il s'agit de l'instruction permettant d'installer la bibliothèque *Passim* à l'aide de *Pip*, un utilitaire fourni avec Python, : « `pip install git+https://github.com/dasmiq/passim.git@seriatim` ».

⁸² Le Jupyter Notebook de David Smith traduit, complété et adapté à notre corpus de test, se trouve dans le dossier [III 2 *Passim*](#).

Le premier traitement consiste en la détection des réutilisations textuelles par défaut. Comme pour BLAST⁸³, le résultat est une liste des amas (*clusters*) regroupant au moins deux extraits considérés comme correspondants. Nous obtenons trois clusters : le premier regroupe les extraits des trois textes et les deux autres présentent les couples des correspondances (voir Tableau 39).

Tableau 39. Résultats du traitement de base avec *Passim*

Id. amas	Id. document	Texte de la correspondance
17179869184	Parnajon1880-1	De là Hyménée, voilé d'un manteau couleur de safran, s'éloigne à travers l'air immense, et il se dirige vers les rivages des Ciconiens, et il est appelé en vain par la voix d'Orphée. Il fut présent à la vérité, mais il n'apporta ni paroles solennelles, ni visages joyeux, ni présage heureux. La torche aussi, qu'il tint, fut sans-cesse crier par la fumée qui fait pleurer, et elle ne trouva aucuns feux par les mouvements. L'issue fut plus funeste que l'augure : car tandis que la nouvelle mariée se promène à travers les herbes, accompagnée d'une troupe de naiades, elle tombe, la dent d'un serpent ayant été reçue dans son talon. Laquelle après que le chanter du Rhodope eut pleurée suffisamment vers les airs supérieurs, afin qu'il tentât aussi les ombres, il osa descendre vers le Styx par la porte du Ténare, et il alla trouver à travers les peuples légers, et les fantômes s'étant acquittés des sépultures, Proserpine, et le maître des ombres occupant des royaumes désagréables ; et ses cordes étant frappées selon (en accord avec) ses chants il dit ainsi : « Ô divinités du monde placé sous terre, dans lequel nous retombons, nous tous qui sommes créées de mortel, si il est loisible, et si vous me permettez de dire
	Parnajon1880-2	De là Hyménée, vêtu d'une robe couleur de safran, s'éloigne à travers les plaines immenses de l'air, et se dirige vers les rivages des Ciconiens où l'appelle en vain la voix d'Orphée. Il vient, il est vrai, mais sans proférer les paroles solennelles, sans apporter un front joyeux, ni d'heureux présages. La torche même qu'il tient, ne cesse de jeter en pétillant une fumée qui remplit les yeux des larmes, et elle ne peut s'allumer, bien qu'il l'agite. L'événement fut encore plus triste que les présages. Un jour que la nouvelle épouse se promenait dans les prairies, accompagnée d'une troupe de naiades, elle tombe, mordue au talon par un serpent. Longtemps le chanter du Rhodope l'a pleuré sur la terre. Enfin il veut aussi tenter de fléchir les ombres et il ose descendre vers le Styx par la porte du Ténare. Il passe au milieu d'un léger peuple de fantômes, mortels qui ont reçu les honneurs de la sépulture, et il arrive auprès de Proserpine et du souverain de ce triste royaume. Alors accompagnant sa voix de lyre, il s'exprime ainsi : « Ô divinités du monde souterrain dans lequel nous retombons, nous tous qui naissons mortels, s'il m'est permis de parler, si vous souffrez que, laissant les détours d'un langage artificieux, je dise la
	Cosnay2017	rien de joyeux. Sa torche, qu'il tient, siffle une fumée qui fait mal aux yeux, c'est tout. Il ne trouve pas le feu, même en la remuant. La suite est pire que le présage : pendant que la nouvelle mariée se promène sur les herbes avec la foule des Naïades, elle meurt, d'une dent de serpent reçue dans le talon. Après que le poète du Rhodope dans les ciels d'en haut l'a bien pleurée, il veut essayer les ombres et ose descendre au Styx par la porte de Taenaria. Après les peuples légers et les fantômes passés au tombeau, il approche Perséphone et celui qui tient les royaumes répugnants, le maître des ombres. Il gratte le

⁸³ Voir Figure 34 et paragraphe qui suit la figure.

0	Parnajon1880-1	re prise et de prendre, la malheureuse ne saisit rien sinon les airs qui se retirent. Et déjà mourant pour la seconde fois, elle se plaignit en quoi que ce soit de son époux : de quoi en effet se plaindrait-elle sinon soi avoir été aimée ? Et elle dit pour la dernière fois un adieu, tel que celui-ci pût le recevoir à peine de ses oreilles ; et elle fut replongée de nouveau au même lieu. Orphée resta-stupéfait de la double mort de son épouse, non autrement que celui qui vit avec effroi les trois cous du chie
	Parnajon1880-2	en arrière. Elle lui tend les bras ; elle veut se jeter dans les siens ; elle tâche de le saisir elle-même : l'infortunée n'embrasse que l'air qui se dissipe. Déjà elle meurt une seconde fois, mais sans se plaindre de son époux ; de quoi en effet se plaindrait-elle sinon d'être aimée ? Elle lui adresse un dernier adieu qui parvient à peine à ses oreilles, et elle est de nouveau replongée dans le même gouffre. Orphée, qui voit la mort lui ravir une seconde fois son épouse, reste interdit. Tel fut ce mortel qui vit avec effroi Cerbère dont la tête du milieu était chargée de chaînes ; la crainte ne le quitta qu'avec sa première forme ; son corpus fut changé e
1	Villenave1806	de la Thrace veut repasser le Styx et fléchir l'inflexible Charon. Toujours refusé, il reste assis sur la rive infernale, ne se nourrissant que de ses larmes, du trouble de son âme, et de sa douleur. Enfin, las d'accuser la cruauté des dieux de l'Érèbe, il se retire sur le mont Rhodope, et sur l'Hémus battu des Aquilons
	Parnajon1880-2	r la rive, sans prendre soin de sa personnes, sans toucher aux présents de Cérès. Ses regrets, sa douleur, ses larmes, sont ses seuls aliments. Las enfin d'accuser de cruauté les dieux de l'Érèbe, il se retire sur le Rhodope élevé et sur l'Hémus battu des Aquilons

Il est frappant de constater la longueur des extraits communs détectés : les trois amas couvrent presque la totalité des traductions de Parnajon1880-1 et Parnajon1880-2. Ces deux textes sont, certes, proches, mais pas identiques, ce qui nous amène à considérer que Passim peut tout de même prendre en compte la reformulation. Pour comparer, Textreuse (voir Tableau 38) trouve aussi la proximité entre les débuts des deux traductions de Parnajon (c'est d'ailleurs le couple qui obtient le score le plus élevé), mais les extraits sont plus courts, car ils s'arrêtent au « Ténare ». Par ailleurs, Textreuse n'a pas détecté le rapprochement entre la traduction de Cosnay à celles de Parnajon. En revanche, aucune relation n'a été détectée entre les traductions en ancien français ni entre les traductions de Villenave1806 et Cosnay2017, ce qui est dommage et confirme l'efficacité restreinte de Passim pour la détection des relations plus complexes.

Pour approfondir ces premières observations, il est possible de compléter le premier traitement par défaut par un traitement qui permet de confronter un document ou un sous-corpus (ci-après : document ou corpus source) à un autre document ou sous-corpus (ci-après : document ou corpus cible). Le logiciel procède alors à une comparaison de chaque ligne du document (ou corpus) cible avec les lignes du document (corpus) source. Concrètement, le premier traitement a dévoilé des relations entre quatre textes : Villenave1806, Parnajon1880-1, Parnajon1880-2 et Cosnay2017. Dans le deuxième traitement, nous comparons les couples des documents phrase par phrase. Les caractères

différents des phrases considérées comme correspondantes sont remplacés par « - ». Ainsi, les extraits de Villenave1806 et Parnajon1880-2 de l'amas 1 sont présentés comme suit :

Tableau 40. Détail de la relation détectée par Passim entre Parnajon1880-2 et Villenave1806⁸⁴

Parnajon1880-2	Villenave1806
Cependant il reste assis sept jours sur la rive, sans prendre soin de sa personne, sans toucher aux présents de Cérès.	--refusé, il reste assi-----s sur la rive- -in-fer-nale, --n- -e s-e no-urrisan--t-- qu--e- de --s--
Ses regrets, sa -dou-leur, ses larmes, sont ses seuls aliments.	-es -larmes, du trouble de son --âme-, --et de- s-a doule-ur.
-Las enfin d'accuser de cruauté les dieux de l'Érèbe, il se retire sur le ---Rhodope élevé et sur l'Hémus battu des Aquilons.	Enfin, las d'accuser la cruauté des dieux de l'Érèbe, il se retire sur le mont Rhodop--e,- et sur l'Hémus battu des Aquilons.

Un tel signalement des différences reste assez difficile à comprendre. D'abord, les traits d'union codent tant les suppressions que les insertions. Ensuite, alors que les déplacements semblent visiblement pris en compte, comme c'est le cas de la deuxième ligne du tableau où « sa douleur » et « ses larmes » n'apparaissent pas dans le même ordre, ces modifications ne sont pas signalées. Enfin, le suivi des différences est ardu là où elles sont nombreuses, comme pour la première ligne du tableau.

Néanmoins, la comparaison ligne par ligne reste très intéressante pour observer le déploiement des extraits communs dans le texte. Pour le couple Parnajon1880-1 et Parnajon1880-2, elle permet d'aboutir, après quelques ajustements⁸⁵, à l'alignement suivant :

Tableau 41. Alignements des traductions Parnajon1800-1 et Parnajon1880-2 avec Passim

Parnajon1880-2	Parnajon1880-1
De là Hyménée, vêtu d'une robe couleur de safran, s'éloigne à travers les plaines immenses de l'air, et se dirige vers les rivages des Ciconiens où l'appelle en vain la voix d'Orphée.	De là Hyménée, voilé d'un manteau couleur de safran, s'éloigne à travers l'air immense, et il se dirige vers les rivages des Ciconiens, et il est appelé en vain par la voix d'Orphée.
Il vient, il est vrai, mais sans proférer les paroles solennelles, sans apporter un front joyeux, ni d'heureux présages.	Il fut présent à la vérité, mais il n'apporta ni paroles solennelles, ni visages joyeux, ni présage heureux.
La torche même qu'il tient, ne cesse de jeter en pétillant une fumée qui remplit les yeux des larmes, et elle ne peut s'allumer, bien qu'il l'agite.	La torche aussi, qu'il tint, fut sans-cesse crierde par la fumée qui fait pleurer, et elle ne trouva aucuns feux par les mouvements.
L'événement fut encore plus triste que les présages.	L'issue fut plus funeste que l'augure :
Un jour que la nouvelle épouse se promenait dans les prairies, accompagnée d'une troupe de naiades, elle tombe, mordue au talon par un serpent.	car tandis que la nouvelle mariée se promène à travers les herbes, accompagnée d'une troupe de naiades, elle tombe, la dent d'un serpent ayant été reçue dans son talon.

⁸⁴ Le changement de l'ordre des traductions par rapport au Tableau 39 vient du fait que Passim a découpé le document cible (Parnajon1880-2) en lignes qu'il a comparées au document source (Villenave1806).

⁸⁵ Nous optimisons surtout le découpage des segments de Parnajon1880-1. Pour les résultats bruts, voir le fichier *total-docwise.csv* dans le dossier 1.3.1-Passim.

Longtemps le chancre du Rhodope l'a pleuré sur la terre.	Laquelle après que le chancre du Rhodope eut pleurée suffisamment vers les airs supérieurs,
Enfin il veut aussi tenter de fléchir les ombres et il ose descendre vers le Styx par la porte du Ténare.	afin qu'il tentât aussi les ombres, il osa descendre vers le Styx par la porte du Ténare,
Il passe au milieu d'un léger peuple de fantômes, mortels qui ont reçu les honneurs de la sépulture, et il arrive auprès de Proserpine et du souverain de ce triste royaume.	et il alla trouver à travers les peuples légers, et les fantômes s'étant acquittés des sépultures, Proserpine, et le maître des ombres occupant des royaumes désagréables ;
Alors accompagnant sa voix de lyre, il s'exprime ainsi :	et ses cordes étant frappées selon (en accord avec) ses chants il dit ainsi :
« Ô divinités du monde souterrain dans lequel nous retombons, nous tous qui naissons mortels, s'il m'est permis de parler, si vous souffrez que, laissant les détours d'un langage artificieux, je dise la vérité, ce n'est pas pour voir le sombre Tartare que je suis venu ici, ni pour enchaîner les trois têtes, hérissées de serpents, du monstre qu'enfanta le sang de la Méduse.	« Ô divinités du monde placé sous terre, dans lequel nous retombons, nous tous qui sommes créées de mortel, si il est loisible, et si vous me permettez de dire [...]
19 phrases non-alignées	16 phrases non-alignées
Elle lui tend les bras ; elle veut se jeter dans les siens ; elle tâche de le saisir elle-même : l'infortunée n'embrasse que l'air qui se dissipe.	[...] de prendre, la malheureuse ne saisit rien sinon les airs qui se retirent.
Déjà elle meurt une seconde fois, mais sans se plaindre de son époux ; de quoi en effet se plaindrait-elle sinon d'être aimée ?	Et déjà mourant pour la seconde fois, elle se plaint en quoi que ce soit de son époux : de quoi en effet se plaindrait-elle sinon soi avoir été aimée ?
Elle lui adresse un dernier adieu qui parvient à peine à ses oreilles, et elle est de nouveau replongée dans le même gouffre.	Et elle dit pour la dernière fois un adieu, tel que celui-ci pût le recevoir à peine de ses oreilles et elle fut replongée de nouveau au même lieu.
Orphée, qui voit la mort lui ravir une seconde fois son épouse, reste interdit.	Orphée resta-stupéfait de la double mort de son épouse,
Tel fut ce mortel qui vit avec effroi Cerbère dont la tête du milieu était chargée de chaînes ; la crainte ne le quitta qu'avec sa première forme ; son corpus fut changé en pierre.	non autrement que celui qui vit avec effroi les trois cous du chien [...]
6 phrases non-alignées	3 phrases non-alignées

Nous constatons ainsi que les deux traductions s'éloignent l'une de l'autre à partir du début du chant d'Orphée devant les infernaux et jusqu'à la sortie des Enfers. Elles se rejoignent au moment où le poète se retourne pour regarder Eurydice. Un extrait correspondant a été donc trouvé pour 15 sur 40 phrases de Parnajon1880-2 (37,5 %). Nous observons aussi que Passim est capable de découper de longs segments du texte source pour les aligner au mieux au segment cible, comme c'est le cas des lignes 6-9 où quatre phrases de Parnajon1880-2 correspondent à une seule phrase très longue de Parnajon1880-1, ou, inversement, regrouper plusieurs phrases sources pour les faire correspondre à un long segment cible. Cette flexibilité peut s'avérer très utile notamment pour la comparaison des textes de genres différents, comme poèmes et pièces de théâtre.

Suite à cette expérimentation, nous considérons que Passim est un bon candidat et qu'il répond bien à la plupart des critères que nous avons fixés. Adapté au public des chercheurs en humanités,

il est richement documenté, même s'il ne semble pas avoir été beaucoup exploité par des littéraires. Il continue à être financé et maintenu et, malgré des difficultés d'installation que nous rencontrions initialement, nous avons trouvé des ressources supplémentaires nous permettant de recourir à une installation alternative (bibliothèque Python). Le corpus ne nécessite pas de préparation particulière et le format de sortie (JSON) est plutôt facile à explorer. C'est aussi le premier logiciel parmi ceux déjà présentés qui permet de comparer des sous-ensembles du corpus. Néanmoins, il ne semble pas y avoir beaucoup de possibilités de paramétrage et les paramètres servent avant tout à s'ajuster aux différentes tailles de corpus. Par ailleurs, pour les textes proches, les résultats sont satisfaisants : Passim détecte de longs segments communs, même là où on observe quelques déplacements, suppressions et insertions. Mais il est dommage que les résultats manquent pour les autres traductions de ce corpus, non seulement celles en ancien français, mais aussi les deux traductions plus modernes. Par conséquent, nous pensons que Passim s'adapte mieux aux recherches portées sur des corpus plus grands et plus homogènes que le nôtre, qui visent plus à révéler les liens jusqu'alors inconnus entre les documents qu'à mettre à jour les liens avérés, mais plus subtils, comme c'est le cas des réécritures du mythe d'Orphée.

Conclusion du chapitre

Les quatre logiciels testés présentent donc différents avantages et inconvénients. Leurs résultats divergent, malgré les modes de fonctionnement plutôt comparables. Ce que nous observons en particulier, ce qu'ils ont tous une capacité très restreinte à détecter des reformulations allant au-delà d'un simple copier-coller. En effet, ils s'appuient avant tout sur la comparaison des chaînes de caractères, même si certains, comme Passim, sont capables de prendre en compte le changement de l'ordre des mots, voire quelques variations graphiques. Contrairement à Phœbus, aucun de ces logiciels ne propose d'ignorer les mots faibles et de procéder à la racinisation des unités lexicales conservées. Nous trouvons également peu de projets qui exploitent un corpus littéraire ou, *a minima*, un corpus français.

En reprenant au Tableau 42 les critères de l'évaluation du Tableau 35, nous constatons que seulement Textreuse et Passim répondent positivement à au moins la moitié des objectifs fixés. Ils sont satisfaisants notamment en ce qui concerne les recherches existantes, les critères techniques et le format, mais restent incomplets quant aux questions du paramétrage et de la qualité des résultats. Voyons à présent ce que nous obtenons avec les deux logiciels retenus, TextPAIR et Tracer.

Tableau 42. Résultats de l'évaluation de BLAST, Textreuse, Tesseract et Passim

		BLAST	Textreuse	Tesseract	Passim
Recherches existantes	Le logiciel est-il adapté au public des chercheurs en humanités (numériques ou non) ?	Non	Oui	Oui	Oui
	Des projets issus des humanités numériques utilisent-ils le logiciel ?	Oui	Oui	Oui	Oui
	Des projets utilisent-ils le logiciel pour exploiter un corpus littéraire ?	Non	Non	Oui	Non
	Des projets utilisent-ils le logiciel pour exploiter un corpus en français ?	Non	Non	Non	Non
Critères techniques	Sur quel(s) système(s) d'exploitation le logiciel peut-il être installé ?	Linux macOS	Linux macOS Windows	Aucun	Linux macOS (Windows)
	Quelle est la difficulté d'installation et d'utilisation ?	Difficile	Facile avec les bases du R	N/A	Facile avec les bases du Python
	Existe-t-il une documentation dédiée (mode d'emploi, tutoriel, forum) ?	Oui, mais peu efficace	Oui, très efficace	Non	Oui, très efficace
	Le logiciel continue-t-il d'être maintenu ? Peut-on échanger avec son auteur ou ses auteurs ? Quelle est la fréquence des mises à jour ?	Non, depuis 2018	Oui	La version 5 en 2021	Oui
Format	Un formatage spécifique du corpus à l'entrée est-il nécessaire ?	JSON	Texte brut	N/A	Texte brut
	Le format des résultats à la sortie est-il aisément exploitable ?	JSON	Choix d'utilisateur	N/A	JSON
Paramétrage	Le logiciel propose-t-il de prendre en compte différentes formes lexicales (mots, lemmes, stemmes) voire de normaliser (ou moderniser) les graphies ?	Non	Non	N/A	Non
	Le logiciel propose-t-il le choix entre différentes méthodes de comparaison (algorithmes d'alignement, calculs de similarité, vectorisation, etc.) ?	Non	Non	N/A	Non
	Le logiciel propose-t-il le choix du degré de similarité minimum recherché (la taille des "trous", ordre des mots ignoré) ?	Non	Oui, la taille des n-grammes	N/A	Oui, la taille des n-grammes
	Le corpus soumis à la comparaison peut-il être modulé (est-il possible de partitionner le corpus pour comparer un sous-ensemble avec un autre) ? Selon quels critères (métadonnées, identifiants...) ?	Non	Non	N/A	Oui
Qualité des résultats	Le nombre des relations détectées est-il satisfaisant ?	Non	Non (un couple par texte)	N/A	Non (seulement les textes les plus proches)
	Les relations détectées sont-elles pertinentes ?	Non	Partiellement	N/A	Oui
	Le logiciel fournit-il les détails concernant la mise en relation (mots communs, score...) ?	Non	Oui, score et différences	N/A	Oui, différences
	Le logiciel fournit-il des informations permettant l'identification des extraits dans le corpus plus globalement et au sein des textes auxquels ils appartiennent plus spécifiquement ?	Oui, identifiant et localisation	Non	N/A	Oui, localisation

3. Présentation des logiciels retenus, TextPAIR¹ et Tracer²

Alors que dans le chapitre précédent, nous avons présenté un par un les quatre logiciels non-retenus, dans ce chapitre nous proposons une analyse conjointe de TextPAIR et Tracer en suivant les cinq critères de comparaison : (i) recherches existantes, (ii) critères techniques, (iii) format, (iv) paramétrage et (v) qualité des résultats. Cela nous permet de mieux exposer les différences et proximités entre les deux logiciels et, quand c'est le cas, de montrer la complémentarité de leurs traitements. Nous constatons aussi que TextPAIR et Tracer ne sont pas toujours supérieurs aux logiciels présentés précédemment : moins récents, plus difficiles à installer ou demandant un formatage spécifique du corpus, ils ne peuvent pas être exploités aussi facilement que Passim ou Textreuse. En revanche, ils présentent des avantages considérables, notamment en ce qui concerne les paramétrages possibles et la qualité des résultats.

a) Recherches existantes

Tracer a été conçu au sein du projet eTRAP (*Electronic Text Reuse Acquisition Project*) de l'université de Göttingen. Le directeur du projet, Marco Büchler, est spécialisé en sciences computationnelles et le traitement automatique des langues appliqué aux données massives. Ses deux collègues, Maria Morits et Eleonora Litta sont spécialisées en linguistique computationnelle et historique, ainsi qu'en philologie grecque et latine. Le projet a été financé entre 2015 et 2018. La dernière mise à jour du projet date de 10 juin 2018.

Le logiciel est accompagné d'une documentation très riche, qui inclut le manuel d'utilisation³, la bibliographie des publications⁴ et un site de téléchargement des ressources (disponible à l'inscription)⁵. Par ailleurs, entre 2015 et 2018, douze formations ont été proposées dans différents pays de l'Europe (Allemagne, Grande Bretagne, Irlande, Pologne, Italie, Estonie et Pays-Bas). En 2016, Tracer a même été nommé aux *DH Awards 2016* dans la catégorie « Le meilleur outil ou suite d'outils pour les humanités numériques ».

Même si les projets qui exploitent ce logiciel viennent principalement de son laboratoire d'origine, ils sont d'une grande richesse. En effet, Tracer a été appliqué aux langues anciennes, comme

¹ Disponible sur <https://artfl-project.uchicago.edu/text-pair>. Voir par exemple Russell Horton, Mark Olsen et Glenn Roe, « Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections », *Digital Studies / Le Champ numérique*, vol. 2, n° 1, 2010, en ligne, <doi: 10.16995/DSCN.258>.

² Disponible sur <https://www.etrapp.eu/research/tracer/>. Voir notamment Marco Büchler, « Informationstechnische Aspekte des Historical Text Re-use », Université de Leipzig, 2013.

³ Voir <https://gfranzini.gitbooks.io/tracer/content/>.

⁴ Voir <https://www.etrapp.eu/academic-output/>.

⁵ Voir <https://vcs.etrapp.eu/tracer-framework/tracer/-/tree/master>.

le grec⁶ et le latin⁷, le tibétain⁸ ou le copte⁹, mais aussi aux langues vivantes, comme l'anglais¹⁰, l'allemand¹¹, le suédois¹² ou le polonais¹³. La majorité de ces recherches ont été menées sur un corpus littéraire, comme des traductions anglaises ou suédoises de la Bible ou différentes versions des contes des Grimm, mais d'autres types de corpus ont aussi été exploités, comme celui regroupant des requêtes utilisateurs pour des services linguistiques¹⁴. Nous ne trouvons pas de recherches sur des corpus francophones, mais une telle diversité linguistique prouve déjà l'adaptabilité du logiciel aux différents langages et alphabets.

Plusieurs articles cités proposent des méthodes d'évaluation du logiciel Tracer en comparant ses performances en fonction des paramètres choisis par l'utilisateur. Nous ne trouvons pas d'étude qui compare son fonctionnement avec celui d'un autre logiciel dédié à la détection des réutilisations textuelles. En revanche, dans « Using and evaluating TRACER for an *Index fontium computatus* of the *Summa contra Gentiles* of Thomas Aquinas » de Franzini *et al.*, nous lisons que

Tracer [est] un outil indépendant de la langue qui comprend des centaines d'algorithmes de la recherche d'information (RI) conçu pour fonctionner avec les langues historiques et modernes. Tracer est un outil en ligne de commande Java piloté par un fichier de configuration XML, que les utilisateurs peuvent modifier pour répondre à leurs besoins de détection. Tracer suit une architecture en six étapes, qui démystifie le processus de détection en stockant sur le disque la sortie de chaque étape de traitement afin que les utilisateurs puissent suivre et localiser plus facilement d'éventuelles erreurs dans la chaîne du traitement. Tracer est résistant au bruit de l'OCR et est capable de détecter tant les citations (quasi-)identiques que des reformulations. La détection de la paraphrase est possible, mais nécessite l'utilisation des ressources linguistiques pour aider Tracer à faire correspondre un mot à ses synonymes et une forme infléchie à sa forme de base¹⁵.

⁶ Marco Büchler et al., « Detection of citations and text reuse on ancient greek texts and its applications in the classical studies: eAQUA project ».

⁷ Greta Franzini et al., « Using and evaluating TRACER for an *Index fontium computatus* of the *Summa contra Gentiles* of Thomas Aquinas », dans Elena Cabrio, Alessandro Mazzei et Fabio Tamburini (dir.), *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018 : 10-12 December 2018, Torino*, Torino, Accademia University Press, coll. « Collana dell'Associazione Italiana di Linguistica Computazionale », 2018, p. 199-205, en ligne, <<https://books.openedition.org/aaccademia/3354>>.

⁸ Orna Almogi et al., « A Hackathon for Classical Tibetan », *Journal of Data Mining & Digital Humanities*, vol. Special Issue on..., n° Towards a Digital Ecosystem..., janvier 2019, p. 2047, en ligne, <<https://jdmhd.episciences.org/2047>>.

⁹ So Miyagawa et al., « Building Linguistically and Intertextually Tagged Coptic Corpora with Open Source Tools », *Leveraging Open Data*, Tokyo, 2018, p. 139-141.

¹⁰ Greta Franzini et al., « Towards a Historical Text Re-use Detection », *op. cit.*

¹¹ Greta Franzini et al., « Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm », *Frontiers in Digital Humanities*, vol. 5, Frontiers, avril 2018, en ligne, <<https://www.frontiersin.org/journals/digital-humanities/articles/10.3389/fdigh.2018.00004/full>>.

¹² Dimitrios Kokkinakis et Mats Malm, « Detecting Reuse of Biblical Quotes in Swedish 19th Century Fiction using Sequence Alignment », *Corpus-based Research in the Humanities workshop (CRH)*, Varsovie, Pologne, 2016, p. 79-86.

¹³ Greta Franzini, « English translations of Pan Tadeusz: a comparison with TRACER », *Corpus-based Research in the Humanities workshop (CRH)*, Varsovie, Pologne, 2016.

¹⁴ Marco Büchler et al., « Mining and analysing one billion requests to linguistic services », *2016 IEEE International Conference on Big Data (Big Data)*, 2016, p. 3230-3239, en ligne, <<https://ieeexplore.ieee.org/document/7840979>>.

¹⁵ « The HTRD on this corpus was performed (server-side) with TRACER, a language-agnostic framework comprising hundreds of information retrieval (IR) algorithms designed to work with historical and modern languages alike.

La capacité du logiciel à détecter des réutilisations moins évidentes que copier-coller semble se confirmer aussi dans les résultats présentés dans la totalité des articles dédiés à Tracer¹⁶. Ils dépendent toutefois des paramétrages du logiciel et l'utilisation des ressources supplémentaires.

Les recherches autour de TextPAIR sont aussi abondantes, mais portent davantage sur un corpus en français. Cela est en partie dû au fait que le logiciel est conçu au sein du projet ARTFL (*American and French Research on the Treasury of the French Language*), dont les recherches se penchent principalement sur la littérature française du XVIII^e au XX^e siècle. La première version, appelée alors Philoline, est présentée en 2008¹⁷ par Mark Olsen and Russell Horton¹⁸ et la version la plus récente, TextPAIR 2.2.0.1, conçue sous la supervision de Clovis Gladstone, date de juillet 2024¹⁹. Cette longévité extraordinaire pour un projet des humanités numériques est le résultat d'une collaboration étroite entre l'université de Chicago, qui dispose du personnel employé à plein temps et de longue date au sein du projet, et les institutions françaises, comme le laboratoire ATILF (Analyse et Traitement Informatique de la Langue Française) de l'université de Lorraine ou le laboratoire OB TIC (Observatoire des textes, des idées et des corpus) de Sorbonne Université.

Plus récemment, grâce au financement de *National Endowment for the Humanities*, le projet ARTFL a mis à disposition des chercheurs le site *Intertextual Hub*²⁰. Il s'agit d'un environnement de lecture numérique qui permet d'accéder aux résultats de la détection des réutilisations avec TextPAIR au sein d'un grand corpus composé des textes littéraires (Rousseau, Beaumarchais, Voltaire...), historiques (pamphlets, journaux de Marat) et juridiques (débats parlementaires de la Révolution française) du XVIII^e siècle. L'interface permet de faire une recherche par auteur ou par œuvre et d'afficher les extraits communs dans leurs contextes. À la Figure 36, nous présentons une correspondance trouvée entre Rousseau et Robespierre. Les deux textes sont alignés et les extraits communs sont mis en rouge. Nous disposons également d'une référence bibliographique de chaque texte.

TRACER is a Java command-line tool driven by an XML configuration file, which users can modify to fit their detection needs. TRACER follows a six-step architecture, which demystifies the detection process by storing the computed output of each step on the disk so that users can more easily follow and locate errors in the processing chain, if any. TRACER is resilient to OCR-noise and capable of detecting both (near-)verbatim quotations and looser forms of TR. The detection of paraphrase requires the use of linguistic resources to help TRACER match a word against its synsets and an inflected form against its base-form. » (nous traduisons). Voir Greta Franzini et al., « Using and evaluating TRACER for an Index fontium computatus of the Summa contra Gentiles of Thomas Aquinas », *op. cit.*, chap. 3.3 Text reuse detection with TRACER.

¹⁶ Nous gardons tout de même en tête que la plupart des études ont été menées par les membres du projet. Par ailleurs, Franzini *et al.*, observent un « échec général de la publication des résultats négatifs des études de la détection automatique des réutilisations textuelles » (« the general failure of HTRD studies to publish negative results », nous traduisons). Greta Franzini et al., « Using and evaluating TRACER for an Index fontium computatus of the Summa contra Gentiles of Thomas Aquinas », *op. cit.*

¹⁷ Voir <https://code.google.com/archive/p/text-pair/>.

¹⁸ Voir notamment Russell Horton, Mark Olsen et Glenn Roe, « Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections », *loc. cit.*

¹⁹ Voir <https://github.com/ARTFL-Project/text-pair/releases/tag/v2.2.0.1>. 33 nouvelles versions ont été proposées depuis septembre 2018.

²⁰ Voir <https://intertextual-hub.uchicago.edu/>.

L'accès aux contextes des extraits communs est une fonctionnalité extrêmement importante. Ici, par exemple, il permet de réaliser qu'il s'agit d'une citation : Robespierre dit explicitement « [s]e servir de l'expression de J. J. Rousseau ». En somme, il s'agit d'une interface comparable à celle proposée par Tesserae, mais pour un corpus en français.

ROUSSEAU, JEAN-JACQUES, 1712-1778. • **CONSID.**
SUR GOUV. DE POLOGNE • *Considérations sur le*
gouvernement de Pologne • 1771

See this passage in document

ensuite par compagnies, puis par régimens ; jusqu'à ce que leur tour étant venu ils se rassemblent en campagne et forment successivement de petits camps dans lesquels on les exerce à toutes les manoeuvres qui conviennent à l'infanterie. Tant qu'ils ne sortent pas du lieu de leur **demeure, peu ou point détournés de leurs travaux**, ils n'ont **aucune paye, mais sitôt qu'ils marchent en campagne**, ils ont le pain de munition et sont à la solde de l'état, et il n'est permis à personne d'envoyer un autre homme à sa place afin que chacun soit exercé lui-même et que tous fassent le service. Dans un état tel que la Pologne on peut tirer de ses vastes provinces de quoi

ROBESPIERRE, MAXIMILIEN, 1758-1794. • **Discours**
sur l'organisation des Gardes nationales / • 1790

See this passage in document

pour combattre les ennemis du dehors, ce Là tout habitant est soldat , mais seulement quand il faut l'être y pour me servir de l'expression de J. J. Rousseau . Les jours de dimanche et de fête, on exerce ces milices selon l'ordre de leurs rôles. Tant qu'ils ne sortent point de leurs **demeures , -peu ou point détournés de leurs travaux** , ils n'ont **aucune paie ; mais sitôt qu'ils marchent en campagne** ^ ils sont à la odde de l'état». Quels qu'aient été nos moeurs et nos idées, avant la révolution, il est peu de françois , même parmi les moins fortunés , qui ne pussent ou qui ne voulussent se prêter à un service de cette espece , que l'on pourroit rendre, parmi nous y encore

Figure 36. Interface de visualisation du site Intertextual Hub

D'autres projets utilisant TextPAIR se penchent sur les corpus littéraires en français. Parmi eux, il convient notamment de mentionner le projet ERC ModERN (*Modelling Enlightenment. Reqssembling Networks of Modernity through data-driven research*) hébergé par Sorbonne Université sous la direction de Glenn Roe et Motasem Alrahabi²¹. Ce projet s'intéresse à un corpus textuel du XVIII^e siècle, mais il y ajoute également des textes critiques du XIX^e s et certaines œuvres majeures du XVII^e s. Il s'agit d'un corpus très massif et très hétérogène, ce qui implique la présence des doublons, voire différentes versions du même texte, de potentiels biais de la reconnaissance optique des caractères et la répartition parfois inégale des textes²². Selon les auteurs de l'article « À la recherche des réseaux

²¹ Voir notamment Valentina Fedchenko, Dario Maria Nicolosi et Glenn Roe, « À la recherche des réseaux intertextuels : défis de la recherche littéraire à grande échelle », *Humanités numériques*, n° 9, Humanistica, juin 2024, en ligne, <<https://journals.openedition.org/revuehn/3940#bodyftn2>>.

²² Nous ne disposons pas d'informations sur la totalité du corpus, mais l'échantillon présenté dans l'article « À la recherche des réseaux intertextuels : défis de la recherche littéraire à grande échelle » de Fedchenko *et al.* contient

intertextuels : défis de la recherche littéraire à grande échelle », TextPAIR permet de répondre à une partie de ces difficultés :

[N]ous avons opté pour Text-PAIR, car il a été spécialement conçu pour répondre aux besoins de la recherche en études littéraires et s'appuie sur l'environnement PhiloLogic. Celui-ci indexe et harmonise les métadonnées des fichiers XML-TEI, les organisant dans une base de données facilement interrogeable à l'aide d'un moteur de recherche. En sus, Text-PAIR est relativement facile à paramétrer et bien adapté à l'extraction des réemplois textuels partiels, ce qui est important pour le traitement des données océrisées.

Finalement, les défis liés à l'installation d'un environnement de travail auxiliaire nous ont semblé acceptables, car ils sont compensés par la capacité à lire et à extraire des informations, même lorsque la structure des fichiers est hétérogène ou présente des anomalies²³.

Nous reviendrons aux défis liés à l'installation du logiciel. Retenons pour l'instant les avantages mentionnés par les auteurs de l'article, à savoir la facilité de paramétrage de traitement avec TextPAIR et l'adaptation du logiciel à l'extraction des réemplois textuels partiels (au-delà d'un simple copier-coller).

Nous avons également eu l'occasion de travailler avec TextPAIR au sein du projet eBalzac, financé par l'Agence nationale de la recherche pour la période 2015-2019 et porté par Andrea Del Lungo, Pierre Glaudes et Jean-Gabriel Ganascia des laboratoires CELLF et LIP6 de Sorbonne Université. Les travaux menés au sein du projet ont abouti à la création du site ebalzac.com, qui propose une édition électronique de *La Comédie humaine*²⁴. Des travaux continuent pour le développement de l'axe hypertexte du projet, qui vise à proposer des recherches et des comparaisons intertextuelles entre l'œuvre de Balzac et un vaste corpus de textes littéraires et scientifiques de l'époque, publiés entre 1800 et 1850. Pour l'instant, le corpus balzacien, composé de 95 textes de *La Comédie humaine*, a été confronté au corpus romanesque de 175 textes d'auteurs contemporains ou antérieurs à Balzac (George Sand, Chateaubriand, Gautier, Sue, etc.)²⁵, ainsi qu'au corpus des 6 ouvrages scientifiques, issus notamment des sciences naturelles, de la médecine et de la physiologie²⁶. Ces expérimentations ont permis de nombreuses découvertes, comme des écritures à quatre mains de Balzac (*Une fille d'Ève*, *Madame de Firmiani*) et Théophile Gautier (*Mademoiselle de Maupin*),

1 054 textes de l'auteur le plus présent, Voltaire, alors que le deuxième auteur le plus présent est Carmontelle, avec 73 textes (voir § 11).

²³ Valentina Fedchenko, Dario Maria Nicolosi et Glenn Roe, « À la recherche des réseaux intertextuels », *loc. cit.*, paragr. 15.

²⁴ Pour la présentation détaillée du site ebalzac.com et ses enjeux éditoriaux, voir Andrea Del Lungo, « L'édition numérique eBalzac. Une nouvelle lecture de *La Comédie humaine* », *The Balzac Review / Revue Balzac*, n° 4, « L'édition / Publishing », 2021, p. 119-140, en ligne, <https://classiques-garnier.com/doi/garnier?article=RbzMS04_119>.

²⁵ Andrea Del Lungo et Karolina Suchecka, « Projet eBalzac : construire une bibliothèque hypertextuelle des sources intellectuelles », dans Clarisse Bardiot, Émilien Ruiz et Esther Dehoux (dir.), *La Fabrique numérique des corpus en sciences humaines et sociales*, Presses universitaires du Septentrion, 2022.

²⁶ Andrea Del Lungo et Karolina Suchecka, « Le projet eBalzac : un hypertexte des sources scientifiques », *Balzac et les disciplines du savoir*, Classiques Garnier, à paraître.

qui ont été détectées par TextPAIR malgré quelques ajouts et reformulations de Gautier. Des proximités sémantiques assez surprenantes ont également été relevées par exemple en ce qui concerne les descriptions des lieux et des personnages balzaciens très proches de celles d'Eugène Sue. Enfin, le logiciel a également permis de formuler ou de confirmer plusieurs hypothèses, et notamment celle de l'influence des théories scientifiques de l'époque – par exemple celles de la physiognomonie et la phrénologie – sur l'écriture balzacienne, surtout dans la description physique des personnages.

TextPAIR s'est également montré efficace pour la détection des réutilisations au sein des corpus volumineux. Comme le constate Jean-Gabriel Ganascia,

Textpair [...] permet de faire varier de nombreux paramètres, par exemple le nombre k de sauts, la taille n des n -grams, l'étiquette – nom, adjectif, verbe – et la nature des termes – lemmes, racines, étiquettes syntaxiques, etc. – que l'on retient dans les k -skip- n -grams.

Toutes ces techniques se sont avérées très efficaces dans le sens où il est maintenant possible de détecter de minuscules fragments de textes réutilisés dans de très grands corpus comprenant des centaines de milliers de livres.²⁷

Néanmoins, cette efficacité du logiciel à traiter des corpus volumineux n'est pas sans un contre-côté. En effet, TextPAIR génère un nombre très important de résultats, de tailles variables et dont une partie ne serait due qu'à une « multiplication artificielle lorsqu'un même fragment est souvent repris »²⁸.

D'après ces retours d'expériences, les deux logiciels ne présentent pas les mêmes avantages. Alors que des chercheurs soulignent la capacité de TextPAIR de traiter efficacement des corpus volumineux, pour Tracer c'est plutôt la transparence des différentes étapes du traitement et son aptitude à détecter des relations plus complexes qui sont mises en avant. En parallèle, les deux logiciels offrent un grand choix de paramétrages et sont tout à fait adaptés au public des chercheurs en humanités. Ils ont été utilisés par de nombreux projets issus des humanités numériques, exploitant principalement un corpus littéraire. Alors que TextPAIR a été utilisé surtout pour les corpus en français²⁹, Tracer a montré son efficacité pour différentes langues et des questions de recherche

²⁷ Jean-Gabriel Ganascia, « Détection automatique de phénomènes intertextuels », *Genesis. Manuscripts – Recherche – Intervention*, n° 51, Sigales, décembre 2020, p. 63-77, en ligne, <<https://journals.openedition.org/genesis/5671>>.

²⁸ *Ibid.*, p. 19.

²⁹ Il convient tout de même de mentionner des recherches menées :

- en anglais (voir notamment Charles Cooney et Clovis Gladstone, « Opening New Paths for Scholarship: Algorithms to Track Text Reuse in ECCO », dans Simon Burrows et Glenn Roe (dir.), *Digitizing Enlightenment: Digital Humanities and the Transformation of Eighteenth-Century Studies*, Oxford University Studies in the Enlightenment, Voltaire Foundation in association with Liverpool University Press, 2020.) ;
- en latin (voir notamment Charles Cooney et Clovis Gladstone, *Tracing Swerves of Influence : text reuse and the reception of Lucretius in 18th-century England*, DH2017, Montréal, 10 août 2017.) ;
- et en chinois (Jeffrey Tharsen et Clovis Gladstone, « Using Philologic For Digital Textual and Intertextual Analyses of the Twenty-Four Chinese Histories 二十四史 », *Journal of Chinese History 中國歷史學刊*, vol. 4, n°

très diverses. Il a également pu être pris en main *ad hoc*, lors des conférences et formations, ce qui prouve la facilité relative de son fonctionnement.

b) Critères techniques

En ce qui concerne les critères techniques, force est de constater que les deux logiciels ne présentent pas les mêmes défis. Pour Tracer, nous disposons d'un manuel d'utilisation très détaillé³⁰, incluant une explication des prérequis système (logiciels à installer au préalable)³¹, la partie « Téléchargement et installation »³², ainsi que toute une section consacrée à la résolution des problèmes³³. Il est également possible de signaler tout problème via un formulaire dédié³⁴.

Tracer peut être installé sur tous les systèmes d'opération (macOS, Linux ou Windows), tant que l'utilisateur ait installé au préalable la bonne version de *Java Software Development Kit*³⁵ et qu'il dispose d'un terminal³⁶. Comme l'indiquent les auteurs du manuel, « [i]l s'agit de l'outil le plus complet à ce jour et il est continuellement amélioré grâce à des retours d'expérience recueillis lors des nombreux tutoriels et ateliers donnés par l'équipe eTRAP lors de conférences et événements internationaux »³⁷. Notre expérience confirme la facilité d'installation de Tracer, que ce soit sur Windows ou sur Linux. Cependant, le logiciel n'est plus maintenu depuis 2018, alors que l'accès au code source doit être accordé par le (ou les) propriétaire(s) du compte GitLab qui l'héberge³⁸. Nous disposons de cet accès depuis 2019, mais notre demande plus récente (2024) est restée sans réponse. Cela implique que les nouveaux utilisateurs ne peuvent potentiellement plus installer le logiciel. En revanche, Tracer est mis à disposition de la communauté sous la licence très permissive, AFL (*Academic Free Licence*)³⁹, les utilisateurs qui possèdent le code peuvent donc le partager à leur tour.

2, 2020, p. 558-563, en ligne, <<https://www.cambridge.org/core/journals/journal-of-chinese-history/article/abs/using-philologic-for-digital-textual-and-intertextual-analyses-of-the-twentyfour-chinese-histories/E390EDB0BEFC2F96F1B16265BF8A4395>>).

³⁰ Voir <https://gfranzini.gitbooks.io/tracer/content/>.

³¹ Voir <https://gfranzini.gitbooks.io/tracer/content/manual/system-prerequisites.html>.

³² Voir <https://gfranzini.gitbooks.io/tracer/content/manual/download-and-installation.html>.

³³ Voir <https://gfranzini.gitbooks.io/tracer/content/support/troubleshooting/>.

³⁴ Malheureusement, le lien indiqué renvoie vers une page réservée aux utilisateurs inscrits : http://www.etrapp.eu/redmine/login?back_url=http%3A%2F%2Fwww.etrapp.eu%2Fredmine%2Fprojects%2Ftracer.

³⁵ Voir <https://www.oracle.com/java/technologies/downloads/>.

³⁶ Pour les utilisateurs Windows, il est nécessaire d'installer un autre logiciel que celui installé de base (Command Line au lieu de PowerShell).

³⁷ « It is the most comprehensive tool yet and it is continuously improved thanks to the feedback gathered by the numerous tutorials and workshops given by the eTRAP (Electronic Text Reuse Acquisition Project) team at international conferences and events » (nous traduisons). Voir <https://gfranzini.gitbooks.io/tracer/content/manual/introduction/>.

³⁸ Voir <https://vcs.etrapp.eu/tracer-framework/tracer>.

³⁹ Voir <https://opensource.org/license/afl-3.0.php>.

Pour TextPAIR, nous ne disposons pas de manuel aussi détaillé, mais nous trouvons les instructions d'installation pas à pas sur le compte GitHub hébergeant le projet⁴⁰. Elles sont destinées à un public plutôt expert et ne visent pas, autant que celles de Tracer, à expliciter les différentes étapes de l'installation. Par ailleurs, pour installer le logiciel localement, il est nécessaire de disposer du même système d'opération que le développeur (ici : Linux 64-bit) et de réussir à reproduire le même environnement de travail. Certaines étapes sont difficiles, d'autant plus que les instructions ne précisent pas la fonctionnalité de certains outils à installer (notamment les manipulations liées à PostgreSQL) et qu'ils impliquent des modifications avancées des fichiers-système. Alors que nous arrivons à installer la version 1.0.11 sans difficultés majeures, nous n'arrivons pas à visualiser les résultats au sein d'un serveur Web. Nous échouons l'installation de la version la plus récente (2.2.0.1), mais cela peut être dû à l'installation déjà existante de la version antérieure.

En revanche, les récentes mises à jour de TextPAIR permettent de l'utiliser sur n'importe quel système via un outil dédié au partage des applications, Docker. Il s'agit de ce que l'on appelle une « plateforme de conteneurisation »⁴¹ qui permet de reproduire son environnement de travail sur une machine virtuelle et de le partager avec d'autres utilisateurs sans qu'ils aient besoin de remplir des prérequis d'installation particuliers. Concrètement, nous allons pouvoir exécuter TextPAIR « à distance », sans avoir à nous préoccuper des difficultés d'installation manuelle. Il est tout de même nécessaire de se former au logiciel Docker qui reste un outil dédié aux informaticiens.

Nous avons donc affaire à deux logiciels plus ou moins faciles à installer et plus ou moins documentés, mais qui sont susceptibles de fonctionner sur tous les systèmes d'opération. En revanche, alors que Tracer n'est plus maintenu depuis 2019, les nouvelles versions de TextPAIR continuent à apparaître et le développement est maintenu à un rythme relativement soutenu. Les efforts considérables sont fournis pour améliorer le fonctionnement du logiciel et faciliter son accès, malgré les difficultés d'installation qui persistent. Nous regrettons seulement que l'utilisateur ne soit pas davantage accompagné dans le processus d'installation et que le guide d'utilisation reste tant technique que sommaire.

Le lancement du traitement se fait pour les deux logiciels à l'aide d'une ligne de commande, après la modification au préalable des fichiers de paramétrage auxquels nous reviendrons ci-dessous

⁴⁰ Voir <https://github.com/ARTFL-Project/text-pair?tab=readme-ov-file#installation> et https://github.com/ARTFL-Project/text-pair/blob/master/docs/ubuntu_installation.md.

⁴¹ Dan Cohen, « Docker : qu'est-ce que c'est et comment l'utiliser ? », dans *DataScientest*, 23 décembre 2020, en ligne, <<https://datascientest.com/docker-guide-complet>>.

(§ d)). Tant pour Tracer⁴² que pour TextPAIR⁴³, la commande se compose des arguments obligatoires et optionnels et son fonctionnement est décrit de manière détaillée dans les instructions d'utilisation. Même si la nécessité d'exécuter le traitement en ligne de commande peut paraître technique aux utilisateurs habitués aux interfaces graphiques, cette méthode est très fréquente dans les développements conçus au sein de projets des humanités numériques et, à l'usage, ne présente pas de difficultés particulières, d'autant plus que son fonctionnement a été clairement explicité par les développeurs.

Pour Tracer, nous disposons également d'une description très détaillée de son traitement, qui est divisé en six étapes : pré-traitement, entraînement, sélection, appariement, évaluation et post-traitement⁴⁴. La description du fonctionnement pour TextPAIR est un peu plus sommaire⁴⁵, mais suffisante pour rendre compte de la proximité du traitement effectué par les deux logiciels. Comme les autres outils dédiés à la détection des réutilisations textuelles, TextPAIR et Tracer s'appuient sur les méthodes d'analyse de séquences issues de la bio-informatique pour générer un ensemble de séquences des mots (des *n-grammes*) qui se chevauchent pour chaque texte du corpus, puis stockent et indexent les informations à analyser par rapport aux séquences d'autres textes. Nous détaillons davantage ce fonctionnement dans les sections suivantes, consacrées au format des données et le paramétrage des deux logiciels.

En somme, les critères techniques pour les deux logiciels paraissent à premier temps hors de la portée des chercheurs non-informaticiens. Mais, alors qu'un accompagnement d'une personne plus expérimentée peut s'avérer nécessaire pour l'installation des deux logiciels, leur fonctionnement et prise en main sont facilités, notamment pour Tracer, par la documentation exhaustive. Autant TextPAIR que Tracer peuvent être installés sur différents systèmes, mais ils nécessitent quelques manipulations au préalable et, dans le cas de TextPAIR, une formation rapide à l'outil Docker.

c) Format

En ce qui concerne le format des données à l'entrée, nous observons des différences majeures dans le fonctionnement des deux logiciels. Pour TextPAIR, le corpus d'entrée peut être chargé en format XML-TEI. Dans le cadre de notre projet, nous n'avons donc aucune préparation du corpus à faire au préalable, ce qui est très commode, mais induit quelques problèmes auxquels nous reviendrons⁴⁶. Alors que le logiciel fait généralement abstraction du balisage lors de la recherche des

⁴² Voir <https://gfranzini.gitbooks.io/tracer/content/manual/execution-of-tracer.html>.

⁴³ Voir <https://gfranzini.gitbooks.io/tracer/content/manual/execution-of-tracer.html>.

⁴⁴ Voir <https://gfranzini.gitbooks.io/tracer/content/manual/configuration/>.

⁴⁵ Voir <https://github.com/ARTFL-Project/text-pair/tree/master?tab=readme-ov-file#textpair-pairwise-alignment-for-intertextual-relations>.

⁴⁶ Voir III.4b) « Problèmes de préparation du corpus et du paramétrage des logiciels ».

correspondances, certaines métadonnées renseignées dans le fichier XML vont être retranscrites dans le fichier des résultats (l’auteur, le titre, etc.). Le logiciel va également chercher la date de création de l’œuvre pour comparer les textes sources seulement aux textes cibles publiés ultérieurement.

TextPAIR peut aussi traiter les fichiers en format texte, mais il est alors nécessaire de joindre au traitement un fichier en format CSV contenant les métadonnées des textes composant le corpus (avec *a minima* le nom de chaque fichier inscrit sur une ligne). Enfin, le projet ARTFL propose aussi le logiciel Philologic⁴⁷ qui est un outil dédié à la publication d’un corpus structuré en XML-TEI et à la recherche plein texte au sein de ce corpus (recherche de la fréquence et du contexte d’un mot, collocation, etc.). Ce logiciel peut être couplé avec TextPAIR. Dans ce cas-là, il faut renseigner au logiciel le lien vers la base de données créée avec Philologic.

Tracer, en revanche, nécessite une préparation du corpus au préalable afin de lui soumettre un seul fichier tabulé au format texte⁴⁸. Chaque ligne du fichier contient quatre champs :

- (i) l’identifiant unique du segment textuel (phrase, paragraphe, etc.), composé de sept chiffres et permettant de regrouper les segments appartenant au même texte (0100001 pour le premier segment du premier texte, 0200002 pour le deuxième segment du deuxième texte, etc.) ;
- (ii) le segment textuel à comparer aux autres segments (phrase, paragraphe, chapitre, etc.) ;
- (iii) la date de création (optionnelle, peut être remplacée par la valeur « NULL ») ;
- (iv) le titre du texte ou une autre information qui sera exploitée pour la visualisation des résultats.

La préparation manuelle d’un fichier de ce type est chronophage, surtout pour un corpus volumineux. Cependant, certaines ressources peuvent faciliter le pré-traitement automatique des textes du corpus, qu’elles soient conçues spécifiquement pour Tracer⁴⁹ ou dédiées plus largement au traitement automatique des langues⁵⁰. Quoi qu’il en soit, ces ressources doivent être adaptées à la spécificité de chaque recherche (la langue du corpus, le genre des textes qui le composent, etc.), ce qui demande des compétences techniques plus ou moins importantes.

Par ailleurs, il est nécessaire de choisir la segmentation la plus adaptée, ce qui peut s’avérer difficile. D’une part, on peut opter pour l’appariement des extraits de taille réduite, comme une phrase,

⁴⁷ Voir <https://artfl-project.uchicago.edu/philologic4>.

⁴⁸ Voir <https://gfranzini.gitbooks.io/tracer/content/manual/corpus-preparation.html>.

⁴⁹ Comme par exemple les scripts Python créés par Mike Kestemont dans le cadre d’une recherche sur la série *Harry Potter* de J.K. Rowling (voir <https://github.com/mikekestemont/potter/blob/master/harry/intertextuality/intertextuality.ipynb>) ou les scripts R mis à disposition par José Luis Losada pour les pièces de théâtre en vers (voir <https://github.com/editio/tracer-scripts>).

⁵⁰ Voir par exemple *NLTK Sentence Segmentation* (<https://textanalysisonline.com/nltk-sentence-segmentation>) qui permet de découper un texte en phrases.

ce qui ne mettra pas en valeur les correspondances plus longues, comme celles détectées par Passim (voir Tableau 39). Au contraire, comparer de longs extraits, par exemple les paragraphes, risque d'entraver la détection des correspondances plus courtes. La segmentation peut aussi faire abstraction de la structure des textes en les découpant en extraits de taille égale (par exemple de 200 caractères) ou se conformer plutôt à la spécificité de chaque texte (découpage en vers pour les textes lyriques, etc.). Les résultats peuvent s'avérer plus ou moins pertinents en fonction de la solution choisie. Nous reviendrons plus en détail sur ces difficultés dans la suite de notre travail⁵¹. Pour notre petit corpus des traductions, nous décidons de suivre le manuel d'utilisation du logiciel et segmentons les textes en phrases.

Les divergences entre TextPAIR et Tracer concernent également le format des résultats à la sortie. Le premier présente les données en format JSON, soit comme une liste de propriétés associées à une valeur. Nous y retrouvons des informations bibliographiques (l'auteur, le titre, le lieu de publication) des textes dont proviennent les extraits source et cible, ainsi que les extraits communs, leurs contextes et leurs emplacements respectifs dans les textes⁵². Ce fichier n'est pas destiné à la lecture humaine (mais peut être réutilisé pour la suite du traitement), car TextPAIR intègre une interface de visualisation identique à celui déjà présenté à la Figure 36. En son sein, les résultats peuvent être recherchés selon plusieurs critères ou affichés dans leur totalité. Voici par exemple ce que nous obtenons pour la recherche des correspondances contenant le mot « empire » et dont le texte-source est celui de Villenave1806 :

1

Source

Ovide • *Les Métamorphoses d'Ovide, traduction nouvelle, avec le texte latin, suivie d'une analyse, de l'explication des fables, de notes géographiques, historiques, mythologiques et critiques* • 1806

tous soumis. Après un court séjour sur la terre, un peu plus tôt ou un peu plus tard, nous arrivons dans cet asile ténébreux ; nous y tendons tous également ; c'est ici notre dernière demeure. **Vous tenez sous vos lois le vaste empire du genre humain** .

Lorsqu'Eurydice aura rempli la mesure ordinaire de la vie, elle rentrera sous votre puissance. Hélas ! c'est un simple délai que je demande ; et si les Destins s'opposent à mes vœux, je renonce

Target

Ovide • *Choix des Métamorphoses. Les auteurs latins expliqués d'après une méthode nouvelle par deux traductions françaises. L'une littérale et juxtaposée au texte latin, l'autre correcte et précédée du texte latin* • 1880

Tout vous appartient : après être demeurés quelque temps sur la terre, tôt ou tard nous nous hâtons d'arriver à la même demeure ; c'est ici que nous nous rendons tous. C'est notre dernier séjour ; et **vous tenez sous vos lois l'empire le plus vaste du genre humain** .

Elle aussi, quand, parvenue à la vieillesse, elle aura accompli le cours de sa destinée, elle vous appartiendra. Ce n'est pas un don, c'est un prêt que je vous demande.

Show differences

Figure 37. Interface TextPAIR : affichage d'un couple des correspondances

⁵¹ Voir III.4b) « Problèmes de préparation du corpus et du paramétrage des logiciels ».

⁵² Pour consulter le fichier des résultats en format JSON, voir les annexes, dossier [III 3 textpair](#), fichier [alignments_textpair.jsonl](#).

Le texte cible est celui de Parnajon1880-2. Les extraits correspondants, mis dans leur contexte, sont affichés en gras et en couleur et alignés l'un à côté de l'autre. Nous disposons également des informations bibliographiques (auteur, titre et date de publication) des deux textes. Un bouton en bas de chaque couple des relations permet d'afficher les différences entre les extraits détectés comme correspondants :

tous soumis. Après un court séjour sur la terre, un peu plus tôt ou un peu plus tard, nous arrivons dans cet asile ténébreux ; nous y tendons tous également ; c'est ici notre dernière demeure. **Vous tenez sous vos lois le vaste empire du genre humain** .
Lorsqu'Eurydice aura rempli la mesure ordinaire de la vie, elle rentrera sous votre puissance. Hélas ! c'est un simple délai que je demande ; et si les Destins s'opposent à mes vœux, je renonce

Tout vous appartient : après être demeurés quelque temps sur la terre, tôt ou tard nous nous hâtons d'arriver à la même demeure ; c'est ici que nous nous rendons tous. C'est notre dernier séjour ; et **vous tenez sous vos lois l'empire le plus vaste du genre humain** .
Elle aussi, quand, parvenue à la vieillesse, elle aura accompli le cours de sa destinée, elle vous appartiendra. Ce n'est pas un don, c'est un prêt que je vous demande.

Figure 38. Interface TextPAIR : affichage incluant les différences entre les extraits alignés

Il s'agit d'une simple comparaison au niveau des chaînes de caractères où les suppressions sont barrées et mis en vert et les insertions sont en bleu. Même s'il est dommage de ne pas disposer d'une détection des variantes plus complexe, cela facilite tout de même l'analyse manuelle. Pour notre exemple, nous constatons rapidement que ce couple des correspondances est proche, car elle ne diffère que par une mise en minuscule du pronom personnel (« Vous » / « vous ») et une modification du syntagme adjectival (« le vaste empire ») par le déplacement de l'adjectif et l'ajout du superlatif relatif de supériorité (« l'empire le plus vaste »).

L'interface graphique proposée par TextPAIR est extrêmement utile pour l'exploration des résultats. La possibilité de filtrage permet par exemple de cibler la recherche sur un auteur en particulier ou de s'intéresser aux certains mots-clés très récurrents dans le mythe (mort, amour, lyre, etc.) en observant leur présence au sein des couples de réutilisation. Elle permet également de parcourir rapidement les résultats pour mesurer leur taille, évaluer leur pertinence, etc. Il est cependant dommage que l'interface soit créée sur le serveur local, ce qui rend impossible son téléchargement et le partage avec les personnes qui ne disposent pas de l'installation de TextPAIR⁵³. Les résultats peuvent donc être transmis uniquement en format JSON, difficilement lisible par un humain.

Tracer, quant à lui, présente les résultats au sein d'un tableau de six colonnes (Tableau 43) qui renseigne les identifiants du segment source et du segment cible, le nombre des unités communes et le degré de similarité, puis le contenu des deux segments mis en correspondance.

⁵³ C'est la raison pour laquelle nous ne pouvons pas joindre l'interface décrite aux annexes de ce travail de thèse.

Tableau 43. Exemple des résultats Tracer

Id A	Id B	Similarité		Segment A	Segment B
		Nombre	Score		
0300020	0400015	5	0,83	Vous tenez sous vos lois le vaste empire du genre humain.	C'est notre dernier séjour ; et vous tenez sous vos lois l'empire le plus vaste du genre humain

Le couple présenté au Tableau 43 correspond à celui détecté par TextPAIR (Figure 37 et Figure 38). Nous voyons que les segments alignés sont des phrases, dépourvues de leur contexte immédiat. Leurs identifiants correspondent à ceux qui leur ont été attribués dans le fichier soumis à l'entrée : 0300020 pour la vingtième phrase de Villenave1806 et 0400015 pour la quinzième phrase de Parnajon1880-2. Ce couple partage cinq unités communes : « lois », « vaste », « empire », « genre » et « humain » (les autres mots identiques, comme « vous » et « tenez » figurent dans la liste des mots faibles du logiciel et ne sont donc pas pris en compte⁵⁴). Enfin, le score de similarité des segments 0300020 et 0400015 est de 0,83⁵⁵. Nous disposons donc de quelques informations concernant la relation détectée, mais elles restent sommaires, notamment en ce qui concerne la similarité. Nous aurions aimé, par exemple, accéder au détail des entités qui motivent la mise en correspondance (notamment une liste des unités communes, en plus de leur nombre).

Comme TextPAIR, Tracer intègre une interface d'exploration des résultats, TRAViz⁵⁶, qui propose trois types de visualisations. La première, appelée « la grille des réutilisations textuelles » (*Text Re-use Grid*) vise à donner une vue d'ensemble des relations détectées au sein du corpus. Malheureusement, elle ne semble pas fonctionner pour le corpus de test, car la grille censée visualiser les résultats est vide⁵⁷. Nous arrivons au même constat pour le troisième type de visualisation, « visualisation de l'alignement des réutilisations textuelles » (*Text Re-use Browser*)⁵⁸. La deuxième visualisation, en revanche, est bien fonctionnelle. Elle permet d'aligner deux textes et d'examiner le déploiement linéaire des correspondances détectées entre eux. Trois diagrammes sont générés que nous présentons sur l'exemple de comparaison effectuée entre Villenave1806 et Parnajon1880-2.

Le premier est un diagramme par point (Figure 39). Sur l'axe vertical, nous avons le texte de Villenave1806 et sur l'axe horizontal celui de Parnajon1880-2. La longueur des axes est proportionnelle

⁵⁴ Nous le précisons davantage *infra*, § Paramétrage.

⁵⁵ La valeur du score oscille entre 0 (rien en commun) et 1. Nous expliquons davantage ce calcul au chapitre suivant (III.4c), sous-section « Manque de mesure du degré de similarité ».

⁵⁶ Voir <http://www.traviz.vizcovery.org/> ainsi que Stefan Jänicke et al., « Visualizations for Text Re-use », *Proceedings of the 5th International Conference on Information Visualization Theory and Applications*, 2014, p. 59-70. L'interface est également présentée dans le manuel du logiciel Tracer : <https://gfranzini.gitbooks.io/tracer/content/manual/postprocessing.html>.

⁵⁷ L'interface peut être consultée dans les annexes, dans le dossier 1-3-3_textpair, puis TraVIZ, puis html. Il faut ensuite ouvrir le fichier index.html dans un navigateur Web. Pour l'instant, nous ne trouvons pas la raison de ce dysfonctionnement.

⁵⁸ Dans ce cas précis, il semblerait qu'un des fichiers nécessaires pour le fonctionnement de cette visualisation, *sims.js*, est manquant.

à la longueur des œuvres. Nous constatons donc immédiatement que le premier texte est relativement plus long. Les nombres inscrits au début et à la fin de chaque axe correspondent aux identifiants de la première et de la dernière phrase de chaque texte. Les points de couleur dans la zone blanche représentent les similitudes détectées entre les deux textes et leur couleur indique le degré de similitude : le vert foncé code les phrases très proches, le jaune indique celles qui le sont beaucoup moins. Dans notre cas, cette vision d'ensemble permet d'évaluer rapidement la pertinence de certaines correspondances. Comme il s'agit des séries traductives, les correspondances des deux textes sont censées se déployer de manière linéaire : nous nous attendons à ce que la première phrase du texte source corresponde à une des phrases initiales du texte source, et non pas à un extrait ultérieur, relatif à un autre épisode du mythe. Nous voyons que c'est généralement le cas pour le couple Villenave1806 et Parnajon1880-2, car la plupart des points du diagramme suivent la courbe croissante. Cependant, deux points jaunes s'écartent de cette tendance. Cela est d'autant plus visible dans le deuxième diagramme généré (Figure 40).

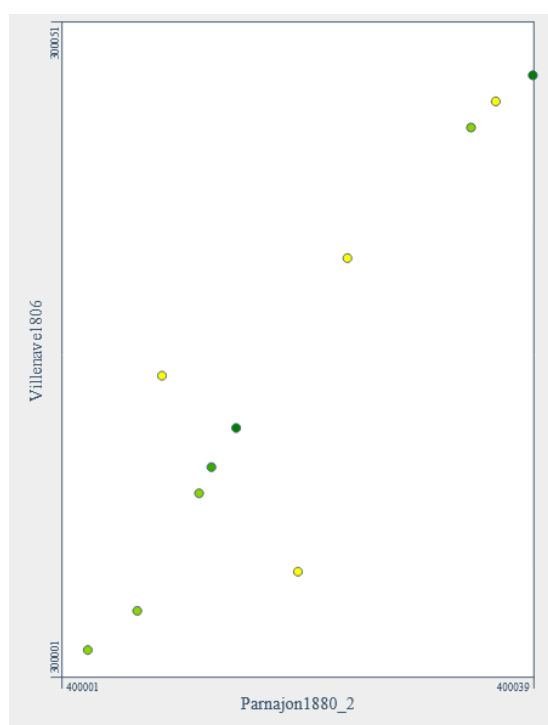


Figure 39. TraVIZ : diagramme par point

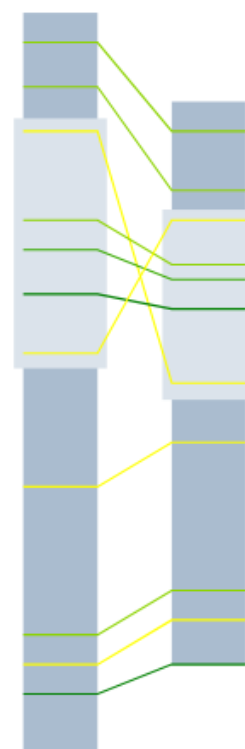


Figure 40. TraVIZ : schéma de l'alignement des réutilisations

En son sein, les deux textes sont représentés par deux rectangles gris. Les lignes de raccordement se rejoignent dans des phrases similaires à leur emplacement exact dans chaque œuvre. Ce schéma de l'alignement des réutilisations confirme de manière encore plus visuelle ce que nous avons observé à grâce au diagramme par point : alors que la plupart des lignes sont quasiment verticales,

deux lignes jaunes se croisent (elles sont mises en avant sur le diagramme à la Figure 40 à l'aide des rectangles gris clair). Avant de nous intéresser au contenu de ces segments, nous pouvons déjà présupposer que les deux couples des correspondances devraient probablement être inversés pour continuer à suivre le déploiement linéaire des correspondances.

La troisième visualisation permet de confirmer cette thèse. Elle offre l'interface de lecture semblable au schéma de l'alignement des réutilisations, mais incluant le texte des deux œuvres, affichées en deux colonnes (Figure 40). Grâce à elle, nous constatons que la phrase 9 de Villenave1806 (« Là, unissant sa voix plaintive aux accords de sa lyre, il fait entendre ces chants : ») est alignée à la phrase 20 de Parnajon1880-2 (« À ces plaintes qu'accompagnent les accords de sa lyre, les ombres glacées versent des larmes »), alors que la phrase 25 de Villenave1806 (« Aux tristes accents de sa voix, accompagnés des sons plaintifs de sa lyre, les ombres et les mânes pleurent attendris ») est, quant à elle, alignée avec la phrase 9 de Parnajon1880-2 (« Alors accompagnant sa voix de lyre, il s'exprime ainsi : »). Il serait plus pertinent de faire correspondre d'une part les phrases 9 et, de l'autre, les phrases 20 et 25. Toutefois, en dehors du cadre de l'alignement des traductions, la proposition de Tracer reste valable, car les deux couples partagent plusieurs mots communs (« lyre », « voix », « accords »).

Villenave1806		Parnajon1880_2
Il arrive au pied du trône de Proserpine et de Pluton, souverains de ce triste et ténébreux empire.	300008	400008 ont reçu les honneurs de la sépulture, et il arrive auprès de Proserpine et du souverain de ce triste royaume.
Là, unissant sa voix plaintive aux accords de sa lyre, il fait entendre ces chants :	300009	400009 Alors accompagnant sa voix de lyre, il s'exprime ainsi :
« Divinités du monde souterrain où descendent successivement tous les mortels, souffrez que je laisse les vains détours d'une éloquence trompeuse.	300010	400010 « Ô divinités du monde souterrain dans lequel nous retombons, nous tous qui naissons mortels, s'il m'est permis de parler, si vous souffrez que, laissant les détours d'un langage artificieux, je dise la vérité, ce n'est pas pour voir le sombre Tartare que je suis venu ici, ni pour enchaîner les trois têtes, hérissées de serpents, du monstre qu'enfanta le sang de la Méduse.
Ce n'est ni pour visiter le sombre Tartare, ni pour enchaîner le monstre à trois têtes, né du sang de Méduse, et gardien des Enfers, que je suis descendu dans votre empire.	300011	400011 C'est mon épouse que je cherche en ces lieux : une vipère sur laquelle elle a marché, a fait couler le venin dans ses veines, et l'a enlevée à la fleur de l'âge.
Je viens chercher mon épouse.	300012	400012 J'ai voulu me résigner ; je l'ai essayé, je l'avoue : l'amour a triomphé.
La dent d'une vipère me l'a ravie au printemps de ses jours.	300013	400013 Je vous conjure donc par ces lieux pleins de terreur, par cet immense Chaos, par ce vaste et silencieux royaume, de renouer la trame, trop tôt coupée, des jours d'Eurydice.
J'ai voulu supporter cette perte ; j'ai voulu, je l'avoue, vaincre ma douleur.	300014	400014 Tout vous appartient : après être demeurés quelque temps sur la terre, tôt ou tard nous nous hâtons d'arriver à la même demeure ; c'est ici que nous nous rendons tous.
L'Amour a triomphé.	300015	400015 C'est notre dernier séjour ; et vous tenez sous vos lois l'empire le plus vaste du genre humain.
La puissance de ce dieu est établie sur la terre et dans le ciel ; je ne sais si elle l'est aux enfers : mais je crois qu'elle n'y est pas inconnue ; et, si la renommée d'un enlèvement antique n'a rien de mensonger, c'est l'amour qui vous a soumis ; c'est lui qui vous unit.	300016	400016 Elle aussi, quand, parvenue à la vieillesse, elle aura accompli le cours de sa destinée, elle vous appartiendra.
Je vous en conjure donc par ces lieux pleins d'effroi, par ce chaos immense, par le vaste silence de ces régions de la nuit, rendez-moi mon Eurydice ; renouez le fil de ses jours trop tôt par la Parque coupé.	300017	400017 Ce n'est pas un don, c'est un prêt que je vous demande.
Les mortels vous sont tous soumis.	300018	400018 Que si les destins me refusent cette faveur pour mon épouse, je suis résolu à ne point retourner sur la terre.
Après un court séjour sur la terre, un peu plus tôt ou un peu plus tard, nous arrivons dans cet asile ténébreux ; nous y tendons tous également ; c'est ici notre dernière demeure.	300019	400019 Réjouissez-vous : vous aurez deux victimes. »

Figure 41. TraVIZ : interface de lecture

Par ailleurs, il est possible de cliquer sur une des lignes de raccordement pour aligner deux phrases et afficher les mots qu'elles ont en commun (Figure 42⁵⁹). Pour le couple déjà présenté au Tableau 43, le trait est vert foncé, ce qui est signe d'une correspondance forte. Nous constatons toutefois que les unités communes soulignées ne concordent pas avec les résultats de Tracer en format tabulé : elles sont au nombre de neuf et incluent les mots faibles (« tenez », « sous », « vos » et « le ») en plus de cinq mots forts qui ont permis la mise en correspondance (« lois », « vaste », « empire », « genre » et « humain »).



Figure 42. TraVIZ : exemple d'alignement des deux phrases

Le clic sur une ligne de raccordement ouvre également une nouvelle fenêtre où le couple analysé est aligné à l'aide d'un graphe orienté acyclique⁶⁰. Cela permet d'examiner les proximités et les différences entre deux (ou plusieurs) phrases similaires et facilite l'analyse de celles qui ne diffèrent que par une flexion (singulier *vs* pluriel) ou l'emploi d'un synonyme. Ainsi, les textes proches sont comparés efficacement, comme c'est le cas de notre couple d'exemple (Figure 43) où le texte de Villenave1806 est représenté en rouge et celui de Pamajon1880-2 en bleu.

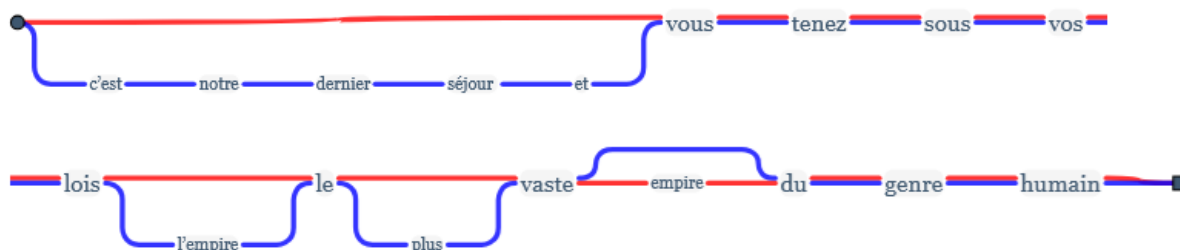


Figure 43. TraVIZ : alignement en un graphe orienté acyclique de deux phrases correspondantes

Les deux lignes se rejoignent à partir du pronom « vous », puis se séparent brièvement pour la modification du syntagme adjectival, que nous avons déjà discuté *supra*. Cependant, les correspondances aussi proches sont relativement rares, surtout dans notre corpus des réécritures, cette visualisation risque donc de se montrer beaucoup moins efficace pour des reformulations qui ne se déploient pas de manière linéaire (les déplacements sont nombreux) et ne se limitent pas à

⁵⁹ Nous sommes consciente que la police de cette figure ne facilite pas la lecture, mais son agrandissement résulte en un dérèglement de l'affichage. Nous décidons donc de laisser la taille originale et de décrire le contenu textuel de manière plus détaillée et renvoyons le lecteur à l'interface inclus dans les annexes (voir n. 57).

⁶⁰ Un graphe orienté acyclique est une structure hiérarchique et orientée, comme un arbre syntaxique.

des modifications flexionnelles ou lexicales. Nous y reviendrons plus en détail dans la suite de notre travail⁶¹.

Ainsi, les deux logiciels proposent, en matière du format des résultats, bien plus que ceux que nous avons essayés auparavant. Outre les fichiers générés à la sortie, plus ou moins lisibles pour un humain, l'utilisateur dispose d'une interface de visualisation des résultats. Certes, nous avons formulé certaines critiques concernant le fonctionnement et l'efficacité de ces interfaces, mais elles restent tout de même une aide inestimable pour l'exploration des couples de correspondances, qui sont nombreux même au sein de notre corpus de test. En effet, nous sommes persuadée qu'une interface d'exploration est indispensable pour notre projet et, même si nous en proposons une faite maison⁶², celles proposées par TextPAIR et Tracer ont grandement alimentées notre réflexion.

d) Paramétrage

Un autre avantage significatif des deux logiciels est qu'ils offrent un large spectre de paramètres, contrairement aux logiciels testés dans le chapitre précédent. Par exemple, pour la phrase « J'ai voulu me résigner à ma perte ; je l'ai tenté, je ne le nierai pas : l'Amour a triomphé », nous pouvons choisir de supprimer des mots très fréquents ou les mots faibles (« ~~J'ai voulu me résigner à ma perte ; je l'ai tenté, je ne le nierai pas : l'Amour a triomphé~~ ») et décider si nous préférons mener une recherche sur les mots (« tenté »), les stemmes (« tent ») ou les lemmes (« tenter »). Tracer permet également d'uniformiser des mots similaires (« essaye » / « essaie »), voire prendre en compte la synonymie (« tenter » / « essayer »). Les logiciels génèrent ensuite les *n*-grammes dont la taille dépend, là encore, des paramètres choisis. Si nous fixons la taille des séquences à trois, nous obtiendrons pour notre exemple cinq trigrammes : vouloir_résigner_perte, résigner_perte_tenter, perte_tenter_nier, tenter_nier_amour, nier_amour_triompher. Il nous reste ensuite à choisir si l'ordre des mots doit être respecté, quel et le nombre minimal des éléments communs pour considérer qu'un couple est en correspondance et quelle est la distance maximale entre chaque mot commun.

Ainsi, notre phrase d'exemple sera appariée à la phrase « J'ai voulu me résigner ; je l'ai essayé, je l'avoue : l'amour a triomphé » tant que la distance de trois mots pleins est permise entre chaque mot commun. En revanche, pour l'appariement de la phrase « J'ai tenté de me résigner, mais l'amour a triomphé », il serait nécessaire d'admettre que l'ordre des mots ne doit pas nécessairement être respecté et que trois mots communs suffisent pour établir la correspondance. Enfin, pour la correspondance avec la phrase « Je voulais me résigner, je ne peux pas le nier, l'amour triompha »,

⁶¹ Voir III.4c) « Problèmes d'exploration et de visualisation ».

⁶² Voir IV.3 « Présentation de prototype d'une édition comparative numérique ».

la recherche doit être menée sur les lemmes (voir Tableau 44 où les mots pleins sont soulignés et les éléments communs sont reliés à l'aide des lettres en orange et en exposant).

Tableau 44. Exemple de la mise en correspondance de TextPAIR et Tracer en fonction des paramètres choisis.

J'ai voulu ^a me résigner ^b à ma perte ; je l'ai tenté ^c , je ne le nierai pas : l'Amour ^d a triomphé ^e .	J'ai voulu ^a me résigner ^b ; je l'ai essayé, je l'avoue : l'amour ^d a triomphé ^e .
	J'ai tenté ^c de me résigner ^b , mais l'amour ^d a gagné.
	Je voulais ^a me résigner ^b , mais l'amour ^d triompha ^e .

L'utilisateur a donc un rôle considérable à jouer pour l'optimisation de la qualité des résultats. Différentes méthodes d'appariement sont également proposées, mais certaines sont déconseillées en première intention : elles alourdissent le traitement de manière importante⁶³, produisent beaucoup de fausses détections⁶⁴ ou sont encore trop expérimentales⁶⁵.

Enfin, pour les deux logiciels, le corpus soumis à la comparaison peut être modulé. TextPAIR permet de renseigner le chemin vers un corpus source et un corpus cible. Il compare alors les textes en respectant leur chronologie : un document source ne sera comparé qu'aux documents cibles (qui lui sont contemporains ou postérieurs). Pour Tracer, il est nécessaire de modifier le fichier texte soumis à l'entrée : pour que les documents soient considérés comme appartenant au même sous-corpus, leurs identifiants doivent commencer par le même chiffre. L'avantage considérable de TextPAIR et Tracer par rapport aux autres logiciels testés est qu'ils prennent nativement en compte le corpus des textes (et non pas des couples de textes à soumettre un par un, comme c'est notamment le cas de Phoebus, BLAST et Passim). Cela rend la recherche d'autant plus modulable.

e) Qualité des résultats

La qualité des résultats des deux logiciels dépend des paramétrages choisis et, pour Tracer, du format des données à l'entrée. Par exemple, si nous effectuons un traitement avec Tracer en gardant tous les paramètres par défaut, même en fournissant des lexiques adaptés à notre corpus (notamment une liste de lemmes et une de mots faibles), nous n'obtenons que huit couples de correspondances (Tableau 45).

⁶³ Dans le fichier de paramétrage, il est conseillé de ne pas trop augmenter la distance admise entre les mots communs, car cela risque d'augmenter l'utilisation du RAM (mémoire vive) de manière importante.

⁶⁴ Surtout l'« appariement des fenêtres mobiles » (*Moving Window Linking*) : « si vous ne connaissez pas bien vos données [...], la recommandation est d'exécuter d'abord une tâche de détection sans l'implémentation de la fenêtre mobile [...]. Utilisez [cette option] uniquement si les résultats ou le rappel de cette première analyse est trop faible et/ou si vous souhaitez trouver une réutilisation plus petite » (nous traduisons, voir § « Moving Window Linking Implementation for short reuse », <https://gfranzini.gitbooks.io/tracer/content/manual/configuration/step-4-linking.html>).

⁶⁵ Comme l'appariement vectoriel de TextPAIR : « il existe deux types d'algorithmes d'appariement, avec des paramètres différents, la recommandée est *sa* (alignement des séquences). L'algorithme *vsa* (alignement vectoriel) est hautement expérimentale, encore en développement intensif et son fonctionnement n'est pas garanti » (nous traduisons, voir <https://github.com/ARTFL-Project/text-pair/tree/master?tab=readme-ov-file#in-the-matching-section>).

Tableau 45. Résultat Tracer avec les paramètres de base

Id. source	Id. cible	Similarité		Texte source	Texte cible
Parnajon1880-1, ph. 1	Parnajon1880-2, ph. 1	9	0,5	De là Hyménée, voilé d'un manteau couleur de safran, s'éloigne à travers l'air immense, et il se dirige vers les rivages des Ciconiens, et il est appelé en vain par la voix d'Orphée.	De là Hyménée, vêtu d'une robe couleur de safran, s'éloigne à travers les plaines immenses de l'air, et se dirige vers les rivages des Ciconiens où l'appelle en vain la voix d'Orphée.
Parnajon1880-1, ph. 19	Parnajon1880-2, ph. 30	6	0,66	Et déjà mourant pour la seconde fois, elle se plaignit en quoi que ce soit de son époux : de quoi en effet se plaindrait-elle sinon soi avoir été aimée ?	Déjà elle meurt une seconde fois, mais sans se plaindre de son époux ; de quoi en effet se plaindrait-elle sinon d'être aimée ?
Parnajon1880-1, ph. 23	Cosnay2017, ph. 44	4	0,5	Il (Orphée) resta assis cependant sur la rive durant sept jours, sale, sans don de Cérès.	Pendant sept jours il reste assis sur la rive, sans don de Cérès.
Villenave1806, ph. 20	Parnajon1880-2, ph. 15	5	0,83	Vous tenez sous vos lois le vaste empire du genre humain.	C'est notre dernier séjour ; et vous tenez sous vos lois l'empire le plus vaste du genre humain.
Villenave1806, ph. 47	Parnajon1880-2, ph. 39	6	0,66	Enfin, las d'accuser la cruauté des Dieux de l'Érèbe, il se retire sur le mont Rhodope, et sur l'Hémus battu des Aquilons.	Les enfin d'accuser de cruauté les dieux de l'Érèbe, il se retire sur le Rhodope élevé et sur l'Hémus battu des Aquilons.
Legouais1301, ph. 28	Walleys1493, ph. 25	9	0,5	Ne Thicius pas ne rebaille aus ostoirs rungier son guisier, et sans la fontaine espuisier ont lor saaz et lor tamis les Belidienes jus mis, et, se la renomée est voire qui ce me fait entendre et croire, les Eumenides ploroient, qui la doucour dou chant ooient, ce qu'ains mes ne pot avenir.	Et Ticius laissa a donner aux vultours son gisier à rongier : et les Belidiennes sans la fontaine espuiser, mirent jus leurs seaulx et leurs tamis.
Legouais1301, ph. 39	Walleys1493, ph. 43	3	0,5	Sans boire et sans manger vivoit, pensans sor dou duel qu'il avoit.	Sans manger et sans boire vivoit luy soustenant de son dueil.
Legouais1301, ph. 47	Walleys1493, ph. 49	7	0,58	Ce fu cil qui premierement a prist ceulz de Trace a retraire d'amour femeline et a faire des joennes malles lor deduit, dont or sont cil de Trace tuit.	Orpheus fut celui qui premier a prist ceulx de Trace a eulx retraire d'amour femenine : et faire leur deduit avec jeunes damoiseaulx, dont ceulx de Trace en sont encores tous entachiez.

La liste des lemmes, incluant la modernisation des textes en ancien français⁶⁶, facilite la détection des relations entre les textes en ancien français (mais pas entre eux et les textes modernes). Seulement les relations les plus fortes sont détectées : la similarité des phrases oscille entre 0,5 et 0,83. Même si ces résultats sont corrects, ils ne sont pas aussi satisfaisants que, par exemple, ceux fournis par Passim.

La recherche limitée aux correspondances très proches est pertinente pour trouver des réutilisations au sein d'un corpus de textes sans lien apparent. Mais dans le cadre de notre recherche, qui regroupe des œuvres dont la proximité thématique est avérée, nous espérons aussi trouver des correspondances moins évidentes. Par ailleurs, nous constatons qu'uniquement les segments de longueurs comparables sont alignés, alors que nous avons vu à plusieurs reprises que les traductions sont plus ou moins descriptives.

En effet, nous réalisons qu'en changeant seulement la méthode par défaut d'évaluation des résultats pour une plus adaptée aux segments de taille inégale⁶⁷, nous obtenons 89 couples de correspondances. Les phrases de tailles différentes ont pu être alignées et une seule phrase d'un texte est parfois alignée à plusieurs phrases d'un autre texte. C'est par exemple le cas du segment 5 de Parnajon1880-1 qui est mis en correspondance avec trois segments de Parnajon1880-2 et un segment de Cosnay2017 (Tableau 46).

Tableau 46. Exemple des résultats avec la méthode d'évaluation adaptée aux segments de tailles inégales

Parnajon1880-1, phrase 5 « Laquelle après que le chantre du Rhodope eut pleurée suffisamment vers les airs supérieurs, afin qu'il tentât aussi les ombres, il osa descendre vers le Styx par la porte du Ténare, et il alla trouver à travers les peuples légers, et les fantômes s'étant acquittés des sépultures, proserpine, et le maître des ombres occupant des royaumes désagréables ; et ses cordes étant frappées selon (en accord avec) ses chants il dit ainsi : »	Parnajon1880-2, - ph. 6 : « Longtemps le chantre du Rhodope l'a pleuré sur la terre. » (2 mots forts en commun et score de similarité de 1) - ph. 7 : « Enfin il veut aussi tenter de fléchir les ombres, et il ose descendre vers le Styx par la porte du Ténare. » (4 et 0,66) - ph. 8 : « Il passe au milieu d'un léger peuple de fantômes, mortels qui ont reçu les honneurs de la sépulture, et il arrive auprès de Proserpine et du souverain de ce triste royaume. » (6 et 0,60)
	Cosnay2017, ph. 9 : « Après les peuples légers et les fantômes passés au tombeau, il approche Perséphone et celui qui tient les royaumes répugnants, le maître des ombres. » (5 et 0,63)

Ainsi, les interrogations liées au choix de segmentation du corpus à l'entrée deviennent moins pré-occupantes, Tracer étant capable de prendre efficacement en charge la différence des tailles des

⁶⁶ Nous détaillons ces lexiques au III.4b) « Problèmes de préparation du corpus et du paramétrage des logiciels ».

⁶⁷ Voir la section « Containment measure for imbalanced length of reuse » du manuel de Tracer (<https://gfranzini.git-books.io/tracer/content/manual/configuration/step-5.-scoring.html>).

extraits et, au moins pour le sous-corpus des traductions, repérer plusieurs extraits courts en lien avec un extrait plus long. Nous regrettons toutefois qu'aucune correspondance n'ait été trouvée avec trois autres textes, ou au moins celui de Villenave1806, ni avec d'autres extraits relatifs à la descente d'Orphée aux Enfers dans la traduction de Cosnay2017.

En revanche, si en plus de modifier la méthode d'évaluation, nous diminuons de moitié le seuil de similarité initialement fixé à 0,5, Tracer détecte 219 segments similaires⁶⁸, donc 130 nouveaux couples pour lesquels la similarité se situe entre 0,25 et 0,49. Ainsi, pour la phrase 5 de Parnajon1880-1, discutée *supra*, cinq nouveaux appariements sont proposés en plus de ceux déjà détectés dans le traitement précédent (Tableau 47).

Tableau 47. Exemple des correspondances supplémentaires obtenues en baissant le seuil de similarité à 25 %

Parnajon1880-1, phrase 5 « Laquelle après que le chantre du Rhodope eut pleurée suffisamment vers les airs supérieurs, afin qu'il tentât aussi les ombres, il osa descendre vers le Styx par la porte du Ténare, et il alla trouver à travers les peuples légers, et les fantômes s'étant acquittés des sépultures, proserpine, et le maître des ombres occupant des royaumes désagréables ; et ses cordes étant frappées selon (en accord avec) ses chants il dit ainsi : »	Cosnay2017, ph. 8 : « Après que le poète du Rhodope dans les ciels d'en haut la bien pleurée, il veut essayer les ombres et ose descendre au Styx par la porte de Taenaria. » (4 mots forts en commun, 33 % de similarité)
	Villenave1806, - ph. 6 : « Après avoir longtemps imploré par ses pleurs les divinités de l'Olympe, le chantre du Rhodope osa franchir les portes du Ténare, et passer les noirs torrents du Styx, pour fléchir les dieux du royaume des morts. » (4, 33 %) - ph. 7 : « Il marche à travers les ombres légères, fantômes errants dont les corps ont reçu les honneurs du tombeau. » (3, 43 %) - ph. 30 : « Ni le dieu de l'empire des morts, ni son épouse ne peuvent résister aux accords puissants du chantre de Thrace. » (2, 33 %)
	Parnajon1880-2, ph. 36 : « Orphée essaie de fléchir Charon ; vainement il veut traverser de nouveau le Styx, le nocher le repousse. » (2, 25 %)

Un nouveau couple a ainsi été trouvé avec le texte de Cosnay2017. Elle précède directement le segment déjà aligné à la phrase 5 de Parnajon1880-1 et permet donc de compléter la correspondance. Trois nouveaux segments ont également été appariés avec la traduction de Villenave1806, mais deux seulement, la phrase 6 et la phrase 7, sont effectivement en relation. Le dernier segment correspond

⁶⁸ Les résultats des trois traitements peuvent être consultés dans le fichier [resultats-differents-parametres.xlsx](#) inclu dans les annexes, au sein du dossier dédié à Tracer (III.3 [tracer](#)). Nous supprimons les doublons, c'est-à-dire les cas de figure où un couple se trouve deux fois dans les résultats (texte source devient texte cible et inversement). Pour les résultats avec la modification de la méthode d'évaluation, 16 doublons sont supprimés (nous passons de 105 à 89 couples). Pour les résultats avec le seuil de similarité à 0,25, nous écartons 139 doublons (nous passons de 339 à 220 couples). Les fichiers [0.25_corpus-corpus.score.expanded](#) et [containment_corpus-corpus.score.expanded](#) contenant les résultats initiaux sont également inclus dans ce dossier.

à la phrase 30, la fin de l'enchantement des infernaux par le chant d'Orphée. Pareillement, la nouvelle correspondance détectée entre Parnajon1880-1 et Parnajon1880-2 n'est pas tout à fait pertinente, car la phrase 36 relate la tentative de retour du poète aux Enfers après avoir perdu Eurydice pour la seconde fois. La similarité peu élevée (25 %) et peu de mots forts communs (2) peuvent donc constituer l'indice d'une relation peu forte et donc, potentiellement, non pertinente. En revanche, la phrase 30 de Villenave1806 obtient le même score de similarité (0,33) que la phrase 6 de Villenave1806 et la phrase 8 de Cosnay2017, toutes les deux considérées comme effectivement correspondantes à la phrase 5 de Parnajon1880-1.

Nous constatons que parmi 12 couples ayant obtenu le score de similarité le plus faible (0,25), deux seulement relatent le même épisode du mythe. Il s'agit de la phrase « Et déjà mourant pour la seconde fois, elle ne se plaignit en quoi que ce soit de son époux : de quoi en effet se plaindrait-elle sinon d'avoir été aimée ? » de Parnajon1880-1 qui est correctement mise en correspondance avec une phrase de Walleys1493 (« Mais riens ne print fors vent et ainsi se partit Erudice de son amy et mourut de seconde mort ») et une phrase de Legouais1301 (« Cele se part de son mari, qui de seconde mort mori. »). Ce segment de Parnajon1880-1 est également aligné avec le même score de similarité à deux autres phrases de Walleys1493, « Mais icelle a lui ne se voulut accorder ne habandonner pour priere ne pour requeste qu'il lui sçeust faire, car trop et de bon cueur aimoit son nouvel espoux Orpheus » et « Forment se plaignoit de la seconde mort de s'amie et voulut retourner pour trouver la mort mais la porte trouva fermee ». Malgré quelques unités communes (« aimer » et « époux » pour la première correspondance ; « second » et « mort » pour la deuxième), ces segments ne se réfèrent pas au même moment du récit mythique.

La précision des résultats à 0,25 de similarité est très faible (2/12), mais elle s'améliore rapidement : parmi 40 couples dont le score oscille entre 0,27 et 0,29, 28 segments sont appariés correctement. Ainsi, une analyse plus approfondie de la pertinence des résultats en fonction des différents paramètres s'impose afin de trouver un seuil optimal. Il est également nécessaire de trouver un moyen de supprimer ceux qui s'avèrent peu pertinents. Mais pour ce faire, il faut établir des critères d'évaluation stricts. Pour le corpus de test, nous avons pu décider si une correspondance est pertinente ou non en nous fiant au texte source que les six traductions partagent. Pour notre corpus des réécritures, il nous faudra trouver une autre méthode qui prend mieux en compte son hétérogénéité et la diversité de ses sources. Nous y reviendrons⁶⁹.

La vue globale des relations détectées entre 219 couples des relations montre (Tableau 48) qu'elles se répartissent de manière assez égalitaire et tous les textes partagent au moins un segment

⁶⁹ Voir III.4a) « Évaluation des résultats ».

correspondant. Les relations sont particulièrement nombreuses entre les deux traductions en ancien français (36 couples), les textes de Parnajon1880-2 et Villenave1806 (29) et les deux traductions de Parnajon1880 (28). Pour ces dernières, il s'agit généralement des relations fortes, car le score moyen de similarité est de 0,455. En revanche, il n'est que de 0,375 pour le couple Parnajon1880-2 et Villenave1806. La relation entre Cosnay2017 et Parnajon1880-2 présente le score moyen très élevé (0,53). D'ailleurs, la traduction moderne de Cosnay2017 entretient des relations fortes avec la plupart des traductions, même pour le peu de relations détectées avec les deux traductions les plus anciennes. Ces dernières sont rarement mises en correspondance avec les textes modernes et nous avons déjà constaté *supra* que certaines d'entre elles ne sont pas pertinentes. Rappelons toutefois que nous n'avons obtenu aucun résultat permettant de lier un texte moderne avec une traduction ancienne avec les autres logiciels de détection des réutilisations.

Tableau 48. Résultats Tracer : nombre des relations détectées pour chaque texte et leur score moyen de similarité

	Legouais1301		Walleys1493		Villenave1806		Parnajon1880-1		Parnajon1880-2		Cosnay2017	
Legouais1301	7	0,5										
Walleys1493	36	0,37	5	0,31								
Villenave1806	5	0,445	6	0,35	7	0,33						
Parnajon1880-1	2	0,27	6	0,295	18	0,37	0	0				
Parnajon1880-2	1	0,285	5	0,31	29	0,375	28	0,455	1	0,33		
Cosnay2017	2	0,355	2	0,365	15	0,465	25	0,44	19	0,53	0	0

Des correspondances intratextuelles sont également présentes, même si elles sont peu nombreuses : sept pour Legouais1301 et Villenave1806, cinq pour Walleys1493 et une pour Parnajon1880-1. Alors qu'elles ne sont pas considérées comme pertinentes pour le corpus des traductions, elles peuvent s'avérer utiles pour certaines réécritures au sein desquelles l'auteur réutilise le *topos* mythique de manière ponctuelle et récurrente. Il nous paraît donc important de garder cette option pour la suite de notre expérimentation. Si toutefois il s'avère qu'elle multiplie les détections fautives, elle pourra être désactivée dans les paramètres du logiciel.

Ainsi, les résultats de Tracer, initialement décevants, ont pu être améliorés de manière très significative grâce aux modifications du paramétrage par défaut du logiciel. Nous avons passé de 8 couples de correspondances à 89 en changeant la méthode d'évaluation, puis à 219 en diminuant le seuil minimal de similarité. La segmentation en phrases se montre utile tant pour l'évaluation de la pertinence des résultats que pour le regroupement des correspondances en extraits plus larges. L'identifiant unique attribué à chaque phrase facilite l'exploration des résultats et la contextualisation d'un couple isolé. Enfin, l'appariement des textes des différents états de la langue est efficace

grâce aux lexiques adaptés fournis par l'utilisateur. Pour ce corpus de test, les résultats semblent généralement pertinents, sauf pour les couples dont la similarité est faible et le nombre des mots communs inférieur à trois. Une révision humaine paraît nécessaire pour mieux évaluer la pertinence des résultats en fonction des différents paramétrages.

En ce qui concerne TextPAIR, le traitement par défaut fournit déjà des résultats relativement satisfaisants, car 106 couples de correspondances sont détectés⁷⁰. Néanmoins, ils présentent un problème de segmentation des extraits communs, ce qui produit des répétitions et fausse le nombre réel des segments appariés. Nous trouvons, par exemple, deux couples de correspondances entre Villenave1806 et Parnajon1880-1 qui concernent pratiquement le même extrait :

- (i) couple n° 18 : « lieux pleins d'effroi, pas ce chaos immense, par le vaste silence » / « lieux, pleins d'effroi, par cet immense Chaos, et par les silences de ce vaste »
- (ii) couple n° 19 : « effroi, pas ce chaos immense, par le vaste silence » / « effroi, par cet immense Chaos, et par les silences de ce vaste »

Dans ce cas précis, nous supprimons le couple contenant les extraits plus courts. Cependant, parfois les extraits communs se chevauchent, mais ne se recouvrent pas complètement, comme pour les couples 139 et 140 détectés entre Parnajon1880-2 et Cosnay2017 :

- (i) couple 139 : « fois, mais sans se plaindre de son époux ; de quoi en effet se plaindrait » / « fois, de son époux elle ne se plaint pas du tout (de quoi se plaindre, si ce n'est d'être aimée »
- (ii) couple 140 : « époux ; de quoi en effet se plaindrait-elle sinon d'être aimée ? Elle lui adresse un dernier adieu qui parvient » / « époux elle ne se plaint pas du tout (de quoi se plaindre, si ce n'est d'être aimée ?), elle dit un dernier « Adieu », qu »

Nous choisissons dans ces cas-là de fusionner les deux couples en un seul. Les correspondances 139 et 140 deviennent alors un seul couple contenant les extraits communs suivants : « fois, mais sans se plaindre de son époux ; de quoi en effet se plaindrait-elle sinon d'être aimée ? Elle lui adresse un dernier adieu qui parvient » et « fois, de son époux elle ne se plaint pas du tout (de quoi se plaindre, si ce n'est d'être aimée ?), elle dit un dernier « Adieu », qu ».

Les résultats sont ainsi diminués de 26 couples pour arriver à 76 relations⁷¹. Cependant, alors que pour ce petit corpus, la suppression et la fusion des extraits en chevauchement ont pu être menées manuellement, ce n'est pas envisageable pour le corpus des réécritures, pour lequel nous

⁷⁰ Comme pour Tracer (voir n. 68), nous présentons les résultats triés en enlevant les doublons (71).

⁷¹ Ces résultats peuvent être consultées dans la première feuille du fichier [resultats-textpair.xlsx](#). Pour le contenu de la deuxième feuille, voir n. 74.

obtenons les résultats bien plus volumineux. Nous aborderons ce problème en détail dans le chapitre suivant⁷².

Une autre difficulté que nous rencontrons concerne les deux traductions anciennes, pour lesquelles seulement sept relations ont été trouvées (Tableau 49). En plus, une seule correspondance est établie entre un texte ancien (Walleys1493) et un texte moderne (Parnajon1880-1) et elle n'est pas considérée comme pertinente, car les deux extraits ne se réfèrent pas au même épisode du mythe⁷³. Les résultats ne contiennent que deux fausses détections, et le deuxième est celui établi entre Legouais1301 et Walleys1493 (« Quant Orpheus la vit venent liez » / « Quant Orpheus vit que plus entrer ne pourroit »). TextPAIR est bien moins efficace que Tracer qui propose 36 relations entre ces deux textes (dont sept non pertinentes).

Tableau 49. Résultats TextPAIR : nombre des couples de relations détectés pour chaque texte et leur longueur moyenne

	Legouais1301	Walleys1493	Villenave1806	Parnajon1880-1	Parnajon1880-2	Cosnay2017
Legouais1301	0					
Walleys1493	7	0				
Villenave1806	0	0	0			
Parnajon1880-1	0	1	7	0		
Parnajon1880-2	0	0	17	14	0	
Cosnay2017	0	0	6	15	9	0

Comme pour Tracer, les trois traductions modernes ont beaucoup d'extraits en commun (17 pour Villenave1806 et Parnajon1880-2, 15 pour Parnajon1880-1 et Cosnay2017 et 14 pour les deux traductions de Parnajon1880). Malheureusement, aucune autre information n'est fournie en ce qui concerne la similarité du couple aligné. Nous ne disposons ni de score ni de nombre d'unités mises en commun, ce qui rend plus difficile l'appréciation globale des résultats.

En revanche, contrairement à Tracer, TextPAIR est capable de détecter des réutilisations plus longues qu'une phrase. Par exemple, la plus longue relation est détectée entre les deux textes de Parnajon1880. Elle englobe tout le début du mythe, de l'arrivée d'Hyménée jusqu'au début du chant d'Orphée devant les infernaux, soit plus de 220 mots. Par ailleurs, les extraits communs peuvent être contextualisés facilement grâce à la présence du texte les précédant et les suivant. Cela facilite l'appréciation de la correspondance pour les extraits plus courts (six mots minimum), comme par exemple « monde placé sous terre, dans lequel » (Parnajon1880-1) / « monde posé sous terre où l'on tombe » (Cosnay2017). La présence du contexte immédiat permet aussi d'observer la distance entre

⁷² Voir III.4c) « Problèmes d'exploration et de visualisation ».

⁷³ « Forment se plaignoit de la seconde mort de s'amie et **voulut retourner pour trouver la mort mais la porte** trouva fermee » / « Que si les destins refusent cette faveur pour mon épouse, il est décidé pour moi de ne pas **vouloir retourner : réjouissez-vous de la mort de deux** victimes ».

les extraits communs des mêmes textes et, éventuellement, leur regroupement. Prenons comme exemple les deux premiers couples détectés entre Villenave1806 et Cosnay2017 (Tableau 50).

Tableau 50. Deux premiers couples détectés entre Villenave1806 et Cosnay2017

Villenave1806	Cosnay2017
L'Hymen, vêtu d'une robe de pourpre, s'élève des champs de Crète dans les airs, et vole vers la Thrace, où la voix d'Orphée l'appelle en vain à ses autels. L'Hymen est présent à son union avec Eurydice, mais il ne profère point les mots sacrés ; il ne porte ni visage serein, ni présages heureux. La torche qu'il tient pétille, répand une fumée humide, et le dieu qui l'agite ne peut ranimer ses	Couvert d'un manteau de safran à travers l'air immense s'en va Hyménée, et vers les rives thraces il avance. L'appel d' Orphée l'appelle en vain. Il est là. Mais il n'apporte ni paroles rituelles ni visage heureux ni rien de joyeux. Sa torche, qu'il tient, siffle une fumée qui fait mal aux yeux, c'est tout. Il ne trouve pas le feu, même en la remuant.
pourpre, s'élève des champs de Crète dans les airs, et vole vers la Thrace, où la voix d'Orphée l'appelle en vain à ses autels. L'Hymen est présent à son union avec Eurydice, mais il ne profère point les mots sacrés ; il ne porte ni visage serein, ni présages heureux. La torche qu'il tient pétille, répand une fumée humide, et le dieu qui l'agite ne peut ranimer ses mourantes clartés. Un affreux événement suit de près cet augure sinistre. Tandis que la nouvelle épouse court sur l'herbe	va Hyménée, et vers les rives thraces il avance. L'appel d'Orphée l'appelle en vain. Il est là. Mais il n'apporte ni paroles rituelles ni visage heureux ni rien de joyeux. Sa torche, qu'il tient, siffle une fumée qui fait mal aux yeux, c'est tout. Il ne trouve pas le feu, même en la remuant. La suite est pire que le présage : pendant que la nouvelle mariée se promène sur les herbes avec la foule des Naiades,

Les extraits correspondants du premier couple font partie du contexte du couple suivant. La distance entre eux est très restreinte chez Cosnay2017 (deux mots : « rituelles ni »). Elle est plus grande pour la traduction plus descriptive de Villenave1806 (19 mots), ce qui empêche le regroupement automatique de ces deux couples par TextPAIR. Or, nous pouvons considérer que l'enchâssement du contexte immédiat prouve une forte proximité entre deux textes, qui justifierait leur regroupement en une correspondance plus large. Idéalement, pour ce corpus qui partage une source commune, on souhaiterait que la totalité des textes de chaque traduction, extraits communs et contextes confondus, soit reliée. TextPAIR serait ainsi aussi performant qu'un outil d'alignement des traductions.

C'est presque le cas pour le couple Parnajon1880-1 et Parnajon1880-2, car les 17 correspondances détectées incluent quasiment la totalité des deux textes⁷⁴. Il manque uniquement la phrase « Celle-ci aussi sera de votre droit (votre propriété), lorsque mûre elle aura accompli les années régulières » pour le premier et « C'est notre dernier séjour ; et vous tenez sous vos lois l'empire le plus vaste du genre humain » pour le deuxième. Pour comparer, 25 phrases de Parnajon1880-1 et 19 de Parnajon1880-2 manquent dans l'alignement avec Passim (Tableau 41), 4 et 12 respectivement sont

⁷⁴ La deuxième feuille du fichier [resultats-textpair.xlsx](#) présente les extraits avec leur contexte. Les segments communs sont mis en gras. Nous y ajoutons également une colonne « amas » qui permet de regrouper les couples en chevauchement.

absentes des résultats les plus complets de Tracer. L'appariement est aussi très satisfaisant pour deux autres couples des traductions modernes : seulement trois phrases de Parnajon1880-1 manquent dans son appariement avec Cosnay2017 et huit phrases de Villenave1806 sont absentes des correspondances détectées avec Parnajon1880-2.

Néanmoins, alors que les contextes facilitent le regroupement des relations établies entre les couples des textes, comme TextPAIR ne traite pas le texte déjà segmenté et ne propose pas le regroupement des amas, il est bien plus difficile de trouver les correspondances croisées où un segment d'un texte est mis en relation avec plusieurs autres segments. Par exemple, nous retrouvons les liens entre toutes les traductions modernes pour l'extrait correspondant à la phrase latine « Esse deos Erebi crudeles questus, in altam se recipit Rhodopen pulsumque aquilonibus Haemum » (X, 76-77) et son contexte immédiat. Nous constatons toutefois que leur segmentation diffère (Tableau 51).

Tableau 51. Correspondances croisées détectées par TextPAIR

Id. source	Id. cible	Texte source	Texte cible
Villenave1806	Paranjon1880_1	Dieux de l'Érèbe, il se retire sur le mont Rhodope, et sur l'Hémus battu des Aquilons. Trois fois	dieux de l'Érèbe être cruels, il se retire sur le haut Rhodope et sur l'Hémus battu par les aquilons
Villenave1806	Paranjon1880_2	Enfin, las d'accuser la cruauté des Dieux de l'Érèbe, il se retire sur le mont Rhodope, et sur l'Hémus battu des Aquilons. Trois fois	enfin d'accuser de cruauté les dieux de l'Érèbe, il se retire sur le Rhodope élevé et sur l'Hémus battu des Aquilons
Villenave1806	Cosnay2017	mont Rhodope, et sur l'Hémus battu des Aquilons. Trois fois	mont Rhodope, et sur l'Hémus battu des vents. Pour la troisième fois
Paranjon1880_1	Paranjon1880_2	Cérès. Le souci, et la douleur de son cœur, et ses larmes furent ses aliments. S'étant plaint les dieux de l'Érèbe être cruels, il se retire sur le haut Rhodope et sur l'Hémus battu par les aquilons	Cérès. Ses regrets, sa douleur, ses larmes, sont ses seuls aliments. Les enfin d'accuser de cruauté les dieux de l'Érèbe, il se retire sur le Rhodope élevé et sur l'Hémus battu des Aquilons
Paranjon1880_1	Cosnay2017	douleur de son cœur, et ses larmes furent ses aliments. S'étant plaint les dieux de l'Érèbe être cruels, il se retire sur le haut Rhodope et sur l'Hémus battu par les aquilons	douleur de l'âme, les larmes le nourrissent. Il se plaint que les dieux de l'Érèbe sont cruels. Il se retrouve en haut du mont Rhodope, et sur l'Hémus battu des vents. Pour la troisième
Paranjon1880_2	Cosnay2017	Rhodope élevé et sur l'Hémus battu des Aquilons	Rhodope, et sur l'Hémus battu des vents. Pour la troisième

En cela, la segmentation imposée par Tracer facilite beaucoup ce type d'investigation, car il suffit de trouver l'identifiant de la phrase concernée pour chaque texte et d'afficher les couples composés avec ces identifiants. Le nombre des mots communs et le degré de similarité constituent aussi des indices bien plus parlants que la différence de la longueur de chaque extrait (Tableau 52).

Tableau 52. Correspondances croisées détectées par Tracer

Id. source	Id. cible	Similarité		Texte source	Texte cible
Villenave1806, ph. 47	Parnajon1880- 1, ph. 25	4	0,57	Enfin, las d'accuser la cruauté des Dieux de l'Érèbe, il se retire sur le mont Rhodope, et sur l'Hémus battu des Aquilons.	S'étant plaint les dieux de l'Érèbe être cruels, il se retire sur le haut Rhodope et sur l'Hémus battu par les aquilons.
Villenave1806, ph. 47	Parnajon1880- 2, ph. 39	6	0,75	Enfin, las d'accuser la cruauté des Dieux de l'Érèbe, il se retire sur le mont Rhodope, et sur l'Hémus battu des Aquilons.	Les enfin d'accuser de cruauté les dieux de l'Érèbe, il se retire sur le Rhodope élevé et sur l'Hémus battu des Aquilons.
Villenave1806, ph. 47	Cosnay2017, ph. 47	3	0,6	Enfin, las d'accuser la cruauté des Dieux de l'Érèbe, il se retire sur le mont Rhodope, et sur l'Hémus battu des Aquilons.	Il se retrouve en haut du mont Rhodope, et sur l'Hémus battu des vents.
Parnajon1880-1, ph. 25	Parnajon1880- 2, ph. 39	4	0,57	S'étant plaint les dieux de l'Érèbe être cruels, il se retire sur le haut Rhodope et sur l'Hémus battu par les aquilons.	Les enfin d'accuser de cruauté les dieux de l'Érèbe, il se retire sur le Rhodope élevé et sur l'Hémus battu des Aquilons.
Parnajon1880-1, ph. 25	Cosnay2017, ph. 47	2	0,29	S'étant plaint les dieux de l'Érèbe être cruels, il se retire sur le haut Rhodope et sur l'Hémus battu par les aquilons.	Il se retrouve en haut du mont Rhodope, et sur l'Hémus battu des vents.
Parnajon1880-2, ph. 39	Cosnay2017, ph. 47	2	0,4	Les enfin d'accuser de cruauté les dieux de l'Érèbe, il se retire sur le Rhodope élevé et sur l'Hémus battu des Aquilons.	Il se retrouve en haut du mont Rhodope, et sur l'Hémus battu des vents.

La confrontation de ce que nous obtenons avec les deux logiciels est intéressante. Nous voyons par exemple que le couple Parnajon1880-1 et Cosnay2017 est celui qui obtient le moindre score de similarité avec Tracer (0,29), mais, parmi ceux proposés par TextPAIR, il fait partie du deuxième extrait le plus long en intégrant une phrase précédente pour Parnajon1880-1 et deux pour Cosnay2017. Celles-ci ne font pas partie des résultats de Tracer.

Conclusion du chapitre

En somme, les résultats pour les deux logiciels sont de qualité très satisfaisante, même si quelques paramétrages sont nécessaires, notamment en ce qui concerne le traitement avec Tracer. Ils sont bien plus nombreux que ce que nous avons observé pour les autres logiciels testés. Pour TextPAIR, nous relevons peu de correspondances fautives, mais les problèmes de segmentation nécessitent le nettoyage des résultats au préalable. Pour Tracer, le dernier traitement, qui baisse le

seuil de similarité de manière importante, multiplie le nombre de fausses détections, mais leur nombre reste tout de même restreint et dépend pour la plupart du degré de similarité. Nous jugeons toutefois nécessaire une évaluation humaine des résultats.

Les deux logiciels sont inégaux en ce qui concerne les informations sur la mise en relation. Tracer donne le nombre des mots communs et le seuil de similarité, ce qui est déjà assez sommaire. Pour TextPAIR, même l'identification de l'emplacement des extraits et le calcul de la longueur des chaînes communes ne sont pas permis, car le début et la fin de chaque segment sont renseignés en octets, unités informatiques appuyées sur le système binaire et permettant d'encoder de caractères ou de valeurs numériques. Comme un caractère peut se composer de différent nombre d'octets, ces informations, utiles pour les machines, ne sont pas exploitables par les humains. Cependant, il ne faut pas oublier que TextPAIR fournit une interface dédiée à l'exploration des résultats. Or le fichier que nous exploitons dans cette section est destiné à « [l']inspection plus approfondie des alignement et [aux] potentielles tâches de post-traitement, telles que le filtrage ou le regroupement des alignements en amas »⁷⁵. Il revient donc à l'utilisateur de le corriger et l'enrichir pour qu'il réponde à ses besoins.

TextPAIR est également moins efficace en ce qui concerne l'identification des extraits au sein des textes. Tracer demande les identifiants uniques de chaque phrase qui permettent à la fois d'identifier le texte (les deux premiers chiffres identiques à tous les segments du même texte) et de localiser l'extrait. Il est alors aisé d'observer le déploiement des correspondances et de trouver les correspondances croisées. Cela est bien plus difficile pour les résultats de TextPAIR qui ne traite pas le texte déjà segmenté et ne propose pas le regroupement des amas. En revanche, la présence du contexte proche facilite l'étude des correspondances de leur contexte et l'analyse de la relation plus globale entre deux textes. Les résultats segmentés de Tracer paraissent plus isolés et sont généralement moins complets, sauf pour les textes anciens.

Si nous reprenons le tableau avec les critères de l'évaluation (Tableau 35) pour résumer nos observations concernant le fonctionnement de Tracer et TextPAIR (Tableau 53), nous constatons que les deux logiciels ne présentent pas les mêmes avantages, mais répondent tous les deux bien mieux à nos besoins que les quatre outils testés dans la section précédente (Tableau 42). Tout d'abord, ils sont tout à fait adaptés au public des chercheurs en humanités numériques et ont été exploités principalement au sein des recherches portant sur des corpus littéraires. Le grand choix de paramétrages permet à l'utilisateur de mieux contrôler la recherche, d'optimiser le nombre des

⁷⁵ « These files are designed to be used for further inspection of the alignments, and potential post processing tasks such as alignment filtering or clustering » (nous traduisons). Voir <https://github.com/ARTFL-Project/text-pair?tab=readme-ov-file#post-processing-alignment-results>.

résultats et d'évaluer leur qualité. On peut également fournir des lexiques (de lemmes, de synonymes, de mots faibles, etc.) caractéristiques du corpus analysé (différents états de la langue, mots-clés, différentes formes des noms propres, etc.). Par ailleurs, contrairement à Phœbus, BLAST et Passim, TextPAIR et Tracer prennent nativement en compte un corpus de textes et non pas des couples à soumettre un par un, ce qui rend la recherche d'autant plus modulable. Enfin, la quantité et la qualité des résultats sont bien supérieures à ce que nous obtenons avec les autres logiciels.

Tableau 53. Résultats de l'évaluation de Tracer et TextPAIR

Critères	Tracer	TextPAIR
Le logiciel est-il adapté au public des chercheurs en humanités (numériques ou non) ?	Oui	Oui
Des projets issus des humanités numériques utilisent-ils le logiciel ?	Oui	Oui
Des projets utilisent-ils le logiciel pour exploiter un corpus littéraire ?	Oui	Oui
Des projets utilisent-ils le logiciel pour exploiter un corpus en français ?	Non	Oui
Sur quel(s) système(s) d'exploitation le logiciel peut-il être installé ?	Linux macOS Windows	Linux macOS* Windows*
Quelle est la difficulté d'installation et d'utilisation ?	Plutôt difficile	Plutôt difficile
Existe-t-il une documentation dédiée (mode d'emploi, tutoriel, forum) ?	Oui, riche et complète	Oui, mais technique et plutôt restreinte
Le logiciel continue-t-il d'être maintenu ? Peut-on échanger avec son auteur ou ses auteurs ? Quelle est la fréquence des mises à jour ?	Non, depuis 2018	Oui
Un formatage spécifique du corpus à l'entrée est-il nécessaire ?	Oui, segmentation et attribution des identifiants	Non, possibilité de traiter directement XML-TEI
Le format des résultats à la sortie est-il aisément exploitable ?	Oui, mais problèmes de fonctionnement du logiciel de visualisation	Oui, JSON de travail et interface de visualisation
Le logiciel propose-t-il de prendre en compte différentes formes lexicales (mots, lemmes, stemmes) voire de normaliser (ou moderniser) les graphies ?	Oui	Oui
Le logiciel propose-t-il le choix entre différentes méthodes de comparaison (algorithmes d'alignement, calculs de similarité, vectorisation, etc.) ?	Oui	Oui
Le logiciel propose-t-il le choix du degré de similarité minimum recherché (la taille des "trous", ordre des mots ignoré) ?	Oui	Oui
Le corpus soumis à la comparaison peut-il être modulé (est-il possible de partitionner le corpus pour comparer un sous-ensemble avec un autre) ? Selon quels critères (métadonnées, identifiants...) ?	Oui, avec une structuration spécifique du corpus à l'entrée	Oui, en mettant les fichiers sources et les fichiers cibles dans des dossiers différents

Le nombre des relations détectées est-il satisfaisant ?	Oui, après la modification des paramètres de base	Oui, même avec les paramètres de base
Les relations détectées sont-elles pertinentes ?	Oui, le nombre de relations peu pertinentes dépend du paramétrage	Oui, après la suppression des doublons (totaux ou partiels)
Le logiciel fournit-il les détails concernant la mise en relation (mots communs, score...) ?	Oui, nombre de mots communs et degré de similarité	Non, mais l'interface permet d'afficher les différences entre deux extraits.
Le logiciel fournit-il des informations permettant l'identification des extraits dans le corpus plus globalement et au sein des textes auxquels ils appartiennent plus spécifiquement ?	Oui, identifiant unique permet d'identifier le texte et localiser le segment en son sein	Oui, les références bibliographiques permettent d'identifier le texte. Le placement de l'extrait est fourni en octets, ce qui rend sa localisation plus difficile.

TextPAIR et Tracer paraissent également complémentaires. TextPAIR traite plus efficacement les corpus volumineux et ne demande pas de formatage spécifique des fichiers à l'entrée, car il travaille nativement avec le XML-TEI. D'autres projets de recherche portant sur un corpus français sont en cours, ce qui facilite potentiellement l'échange des usages et méthodes. Par ailleurs, le logiciel est toujours activement maintenu et mis à jour et son développeur principal, Clovis Gladstone, est accessible et très actif au sein de la recherche francophone. Enfin, les résultats sont facilement exploitables grâce à une interface de visualisation. Ils sont plus fournis pour les textes en français moderne et plus faciles à contextualiser que ceux de Tracer. Tracer, quant à lui, se distingue par une documentation extrêmement fournie et la transparence de différentes étapes du traitement. Avec le paramétrage adapté au corpus, il s'avère efficace à détecter des relations complexes. Nous apprécions en particulier la richesse et l'exactitude des correspondances détectées entre les textes en ancien français.

Nous avons également relevé quelques désavantages. Premièrement, les deux logiciels sont plutôt difficiles à installer et à prendre en main, même si cela est facilité par la présence de la documentation (exhaustive pour Tracer, mais un peu trop technique pour TextPAIR). Tracer nécessite également une segmentation du corpus à l'entrée et l'attribution des identifiants uniques, ce qui peut être chronophage si ce n'est pas fait automatiquement. TextPAIR, quant à lui, montre des problèmes de segmentation des extraits communs au sein des résultats, ce qui augmente le nombre des doublons. Pour les deux logiciels, les informations sur la mise en relation des couples manquent. Enfin, l'interface de visualisation de Tracer, TraVIZ ne semble pas fonctionner correctement et celui de TextPAIR est difficile à partager.

4. Présentation des résultats initiaux et problèmes rencontrés

Dans le chapitre précédent, nous avons présenté les deux logiciels retenus pour la détection des correspondances au sein de notre projet. À présent, notre objectif est de soumettre à Tracer et à TextPAIR notre corpus des réécritures pour qu'il soit confronté à lui-même et au corpus des traductions. Nous ne souhaitons pas chercher des relations entre les traductions de la même source antique ni entre les traductions d'Ovide et de Virgile, pour mieux nous focaliser sur ce que les logiciels trouvent pour les réécritures. Cela est facilement réalisable pour le traitement avec TextPAIR, car il suffit de mettre le corpus source dans un dossier et le corpus cible dans l'autre. La tâche est plus complexe pour le traitement avec Tracer, qui impose de lui soumettre un fichier tabulé où chaque segment de chaque texte est décrit par un identifiant unique à sept chiffres. Les deux premiers chiffres servant à identifier l'œuvre, nous choisissons de regrouper toutes les traductions d'Ovide sous le numéro 54 et toutes celles de Virgile sous le numéro 55. Nous utilisons ensuite les deux chiffres suivants pour identifier chaque traduction séparément. Ainsi, alors que l'identifiant 0100001 correspond à la première phrase de la réécriture Anonyme1700, l'identifiant 5401001 code la première phrase de la traduction d'Ovide (54) par Banier1732 (54-01) et l'identifiant 5501001 renvoie à l'incipit de la traduction de Virgile (55) par Beague1998-1 (55-01)¹.

Nous obtenons 26 859 couples potentiels avec Tracer² et 10 084 avec TextPAIR³. Toutefois, nous avons vu précédemment que ces résultats peuvent contenir des doublons. Nous ne les supprimons pas manuellement, mais retenons qu'un filtrage est nécessaire dans une des étapes initiales du post-traitement des résultats.

Les relations entre les réécritures et les traductions sont bien présentes dans les résultats de Tracer, car 9 693 couples contiennent un segment d'une des traductions (majoritairement d'Ovide, 6 522 couples, contre 3 171 pour Virgile), soit 36 % des relations détectées. Ce chiffre est bien moindre pour TextPAIR, car des traductions sont mises en relation avec des réécritures seulement au sein de 804 couples (soit 8 % des résultats). Parmi eux, 71 seulement concernent les traductions de Virgile. Les textes les plus présents dans les résultats de deux logiciels sont ceux de Ballanche1829 (6 117 pour Tracer et 2 019 pour TextPAIR, soit 8 136 couples au total), de Launay1922 (3 177 et 779, soit 3 956), de Laffon2016 (2 816 et 966 pour un total de 3 782) et d'Anouilh1941 (1 939 et 1 255 respectivement, soit 3 734).

Il faut toutefois rester prudent avec ces nombres qui ne reflètent pas la pertinence effective des résultats. Dans la suite de ce chapitre, nous évaluons un échantillon de ceux-ci pour nous rendre

¹ Voir, dans les annexes de cette partie, dossier [III 4 résultats](#), le fichier [corpus-tracer.xlsx](#).

² Voir le fichier [corpus-corpus.score.expanded](#).

³ Voir le fichier [alignements.tab](#).

compte de ce que nous obtenons réellement pour ce corpus spécifique. Cela nous permet d'identifier les principaux problèmes du traitement et de réfléchir aux solutions pour les pallier dans l'étape du post-traitement. Nous divisons ces problèmes en deux sections et abordons en premier lieu les difficultés liées à la préparation du corpus et au paramétrage des logiciels. Nous réfléchissons ensuite aux questions d'exploration et de visualisation plus efficaces des résultats.

a) Évaluation des résultats

À notre connaissance, très peu de projets de détection automatique des réutilisations textuelles procèdent à une appréciation manuelle d'un échantillon des résultats. Pourtant, la critique littéraire insiste bien sur le fait qu'une relation intertextuelle dépend de la subjectivité du lecteur⁴ qui voit ou non une certaine similarité entre deux phrases et parvient à interpréter le lien qui les unit. Dans l'article « On the Feasibility of Automated Detection of Allusive Text Reuse » Manjavacas *et al.* demandent à quatre annotateurs experts en latin d'évaluer la pertinence des relations détectées entre les textes de Bernard de Clairvaux et la Vulgate. Sur 876 couples, 147 sont considérés comme non pertinents. Pour les couples restants, les auteurs calculent le taux d'accord entre annotateurs et constatent que seulement 22 % d'évaluations sont identiques. Ils en concluent que

[s]i nos résultats restent difficiles à évaluer en l'absence de travaux comparables, ils sont suffisamment faibles pour mettre en doute la faisabilité de la tâche qui est en fait rarement explicitement mise en question. Les annotateurs ont indiqué de façon informelle que, contrairement à leurs attentes, la tâche n'était pas simple et exigeait un niveau considérable de concentration et d'interprétation. [...] Cela suggère qu'il existe des différences claires au niveau de l'appréciation individuelle des allusions.⁵

Une solution pour pallier cette difficulté de l'appréciation objectivable est proposée par Coffee *et al.* dans leur article « Intertextuality in the Digital Age ». Les auteurs conçoivent un schéma de classement en cinq points⁶. Les relations des types 1 et 2 sont considérées comme non pertinentes, c'est-à-dire qu'elles ne « répondent pas aux critères de la recherche »⁷. Cette précision est importante, car nous verrons que l'appréciation objectivable de la pertinence ou non d'un couple des relations n'est pas toujours aisée. Les relations du type 3 à 5, quant à elles, sont considérées comme potentiellement pertinentes, mais le degré de certitude quant à leur relation effective varie. Coffee

⁴ Voir II.1. « Intertextualité dans la théorie littéraire ».

⁵ « While our result remains hard to assess in the absence of comparable work, it is low enough to cast doubts over the feasibility of the task, which is in fact rarely explicitly questioned. The annotators informally reported that, against their expectations, the task was not straightforward and required a considerable level of concentration and interpretation. [...] This suggests that there exist clear differences at the level of individual allusions. » (nous traduisons). Enrique Manjavacas, Brian Long et Mike Kestemont, « On the Feasibility of Automated Detection of Allusive Text Reuse », sect. Inter-annotator agreement results and discussion.

⁶ Neil Coffee *et al.*, « Intertextuality in the Digital Age », *loc. cit.*, p. 392.

⁷ *Ibid.*, p. 393.

et al. considèrent que leur système d'évaluation est propice à une appréciation efficace des résultats, car :

les chercheurs qui étudient les modèles linguistiques pourraient potentiellement examiner ces parallèles de manière rentable [...]. Les critiques qui étudient le style littéraire pourraient évaluer la similitude de deux instances d'anaphorose [...]. Dans le cours normal de l'interprétation littéraire, établir un lien entre les mots communs, couramment exprimés, mettrait, selon Hinds, « des contraintes... sur un décorum philologique d'interprétabilité ». Inversement, pour qu'une phrase soit déclarée significative et qu'on lui attribue un score d'au moins 3, elle doit représenter un idiome cohérent et distinctif.⁸

Pour évaluer la pertinence de chaque couple, nous suivons ce schéma de classement en cinq points proposé par Coffee *et al.*, mais nous l'adaptions aux spécificités de notre projet. En effet, alors que ces derniers établissent des critères généraux liés à la similitude formelle et à la proximité du contexte, nous évaluons également le renvoi aux mêmes épisodes du mythe, en particulier pour les correspondances des types 3 à 5 (Tableau 54).

Tableau 54. Grille d'évaluation de l'échantillon des résultats

Type 1	Relations non pertinentes issues du dysfonctionnement de l'outil, de la mauvaise manipulation des résultats ou des problèmes de segmentation du corpus.
Type 2	Relations non pertinentes entre des mots vides (auxiliaires, articles, etc.) ou des éléments formels des œuvres, des expressions très fréquentes et souvent figées ou des noms propres considérés isolément.
Type 3	Relations entre des mots forts, mais dont la pertinence est difficile à évaluer (insuffisance du contexte, impossibilité d'identifier l'épisode du mythe évoqué, etc.).
Type 4	Relations potentiellement pertinentes, mais qui ont peu de mots identiques ou dont les contextes diffèrent. La parallèle est principalement allusive et/ou la référence au même épisode du mythe n'est pas assurée.
Type 5	Relations pertinentes qui font référence au même épisode du mythe et/ou contiennent des mots et expressions similaires, voire identiques.

Pour évaluer la pertinence des réutilisations détectées par TextPAIR et Tracer, un échantillon des résultats est évalué par un lecteur humain. Le traitement initial est effectué sans amélioration, sauf pour les lemmes extraits de l'annotation du corpus par l'outil TreeTagger. 500 couples sont ainsi extraits de manière aléatoire des résultats de Tracer et de TextPAIR (l'échantillon se compose donc de 1 000 couples de correspondances⁹).

⁸ « Researchers investigating linguistic patterns could potentially examine such parallels profitably [...]. Critics investigating literary style could evaluate the similarity of two instances of anaphora [...]. But in the regular course of literary interpretation, positing any link between the passages based on such common words, commonly expressed, would, in the words of Hinds, put “strains ... upon a philological decorum of interpretability.” Conversely, for a phrase to be declared meaningful and be given a score of at least 3, it needed to represent a coherent, distinctive idiom » (nous traduisons). *Ibid.*, p. 395.

⁹ Voir, dans les annexes de cette partie, dossier [III 4 résultats](#), le fichier [echantillon-evalue.xlsx](#).

Dans notre évaluation, nous nous heurtons aux mêmes difficultés que celles évoquées par Manjavacas *et al.*¹⁰, en ce qui concerne le maintien de la concentration et les variations de l'interprétation, alors même que nous connaissons bien notre corpus. Pour cela, le classement à cinq points qui prend en compte différents niveaux de certitude est très important, car il est parfois impossible de trancher clairement entre une relation pertinente ou pas sans accéder au contexte plus large que celui proposé par les logiciels. Les types 3 à 5 permettent de bien rendre compte de cela. Les types 1 et 2, quant à eux, sont basés sur des critères plutôt formels (dysfonctionnement de l'outil, seulement mots vides en commun, etc.), leur identification univoque ne pose donc pas de problèmes majeurs. S'il y a un flottement d'appréciation, il est ressenti surtout entre les couples du type 3 et du type 4. Une certaine liberté d'interprétation nous paraît nécessaire pour ne pas écarter des résultats potentiellement pertinents à cause de critères d'évaluation trop restreints.

Comme le montre la Figure 44, les correspondances potentiellement pertinentes restent majoritaires (62 %). Mais celles non pertinentes (type 1 et 2) représentent tout de même presque 38 % de l'échantillon et les correspondances du type 2 en forment l'ensemble le plus grand (34,7 %). Elles se retrouvent majoritairement dans l'échantillon des résultats de TextPAIR (20/22 pour le type 1 et 224/347 pour le type 2). La possibilité de les écarter automatiquement dans l'étape du post-traitement permettrait d'améliorer les résultats de manière importante. Toutefois, leur analyse soulève plusieurs difficultés.

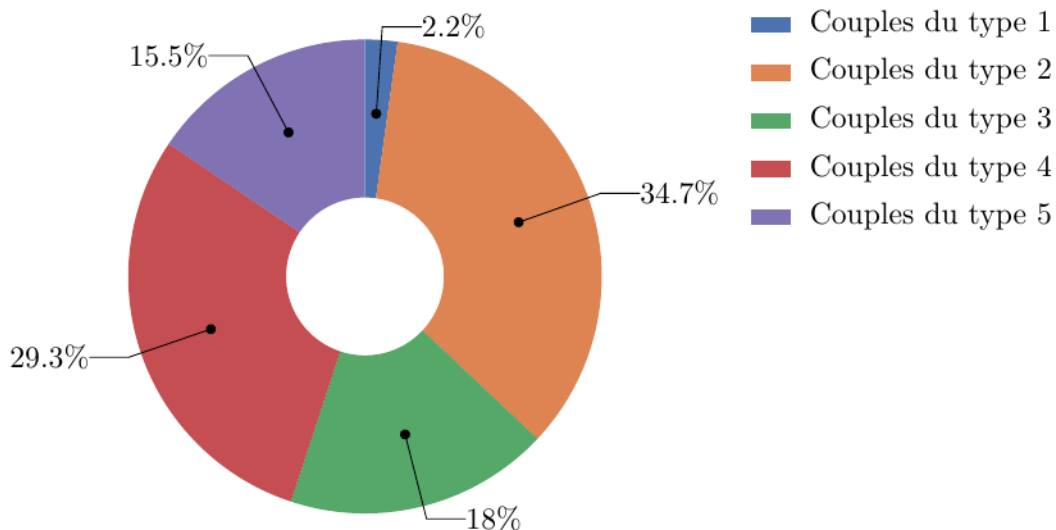


Figure 44. Observation des 1 000 couples choisis aléatoirement parmi les résultats de Tracer (500/22 685) et TextPAIR (500/10 084)

¹⁰ Dans le cadre restreint de ce travail de thèse, nous ne procédons pas à la confrontation des évaluations des différents annotateurs, qui nécessiterait un financement permettant le dédommagement des participants, mais c'est une piste que nous aimerions explorer dans le futur.

Couples du type 1

Premièrement, même si TextPAIR peut détecter de longs segments communs, les résultats sont plus susceptibles d'être dominés par des données bruyantes. En effet, la préparation du corpus pour Tracer nous contraint à éliminer des éléments formels des œuvres, comme titres, désignations de locuteurs dans le dialogue, ou encore des notes de l'éditeur. Mais, comme TextPAIR traite directement les fichiers XML, tous ces éléments sont inclus dans les résultats, ce qui explique le nombre élevé des correspondances du type 1. Nous y trouvons par exemple une relation établie entre le titre de la traduction de Rachmuhl2003 et un extrait du roman de Liking1981, ou encore un couple qui apparie les titres des scènes chez Crémieux1858 et Cocteau1927 (Tableau 55 où les correspondances de TextPAIR sont signalées avec [TP], les extraits communs sont soulignés et le contexte immédiat est fourni).

Tableau 55. Relations du type 1 dues aux problèmes du traitement du corpus

Rachmuhl2003 [TP] : Françoise Rachmuhl Vie et <u>mort d'Orphée Vie et mort d'Orphée</u> Orphée était-il le fils d'Apollon, le dieu des arts et de la musique ?	Liking1981 [TP] : Partir <u>mort de ma vie de mort</u> <u>Marcher et vivre</u> éternellement. Or, je vois les combats...
Crémieux1858 [TP] : Il <u>sort</u> . Scène III Eurydice Que veut-il dire avec son petit nanan semé dans les blonds épis ?...	Cocteau1927 [TP] : (Il <u>sort</u> .) Scène III Eurydice, <u>Heurtebise Eurydice</u> Vous voyez.

18 des 20 correspondances du type 1 de l'échantillon de TextPAIR sont ainsi dues au fait que le logiciel n'exploite pas le balisage XML et ne permet pas l'omission de certains éléments structurels ou éditoriaux peu pertinents pour la détection des similarités et pouvant nuire à la qualité des résultats. Par ailleurs, les deux couples du type 1 restant de l'échantillon de TextPAIR et un autre de Tracer n'ont rien en commun, ce qui peut être dû à des problèmes internes du fonctionnement des logiciels (Tableau 56 où les correspondances de Tracer sont signalées avec [Tr])

Tableau 56. Relations du type 1 dues au dysfonctionnement des logiciels

Duboullay1690 [TP] : Ô ! toy, qu'un charme plein d'horreur/ Vient d'instruire en secret à servir ma <u>fureur</u> ,/ <u>Serpent, que sous ces fleurs cache</u> cette prairie ?	Ballanche1829 [TP] : Dans nos peintures, le vêtement, c'est le corps ; le corps <u>nu</u> , c'est l'ame. <u>L'homme habillé</u> , c'est l'homme revêtu de son corps mortel, de son corps qu'il doit laisser à la terre.
Schuré1889 [TP] : Un sentier fleuri, une <u>pente escarpée</u> , et puis des <u>rochers hantés</u> par la foudre avec l'espace immense autour	Divoire1922 [TP] : Et lourds, et grossiers, il faut qu'on <u>expie</u> / <u>Le malheur d'ignorer ton spectre fraternel</u> / Dont le rythme parfait convie/ À chanter les grands chants du ciel.
Ballanche1829 [Tr] : Eurydice, à ce moment douloureux, était pleine d'une tristesse infinie.	Manjock1981 [Tr] : J'avais au fond de moi cette image d'une lumière et d'une chaleur en pleine tempête.

La dernière relation du type 1 dans l'échantillon de Tracer, en revanche, résulte d'une mauvaise segmentation du corpus : en effet, nous réalisons que le segment 32 de la traduction de Parnajon1880-1 est composé de sept phrases et correspond à la totalité du meurtre d'Orphée par les Bacchantes. Ce long fragment (334 mots) est mis en correspondance avec une phrase de Divoire1922, « Ô pacte harmonieux des êtres,/ Une coupe épanchée en l'honneur des ancêtres/ Ne suffit pas pour que ton règne puisse naître » avec laquelle il partage des mots isolés. Nous corrigeons cette erreur dans les traitements suivants.

Couples du type 2

Pour les couples de type 2, différents facteurs font que les relations sont considérées comme non pertinentes (Tableau 57). Certains sont plutôt formels et concernent les caractéristiques des mots en commun (uniquement des mots faibles, mots forts trop éloignés, uniquement des noms propres). D'autres peuvent dépendre de l'appréciation de l'évaluateur (perception des différents sens des mots communs, jugement de certaines constructions comme banalités, etc.). D'autres encore sont plus spécifiques : au sein d'un couple, une affirmation peut être appariée à une négation (« je t'aime » / « je ne t'aime pas »). Enfin, des ressemblances peuvent aussi résulter des proximités génériques uniquement (didascalies du théâtre du type « Il sort », etc.).

Tableau 57. Facteurs caractéristiques des couples du type 2 et quelques exemples tirés de l'échantillon

Uniquement mots faibles en commun	« Oui, oui, mais <u>il ne faut plus t'en aller</u>. <u>Il ne faut plus</u> me mépriser » / « <u>Il ne faut pas les laisser s'en aller</u> ! <u>il faut</u> leur courir après ! » (Segalen1921 vs Anouilh1941) « <u>Il a fait</u> clair <u>tout</u> d'un <u>coup</u> sur la crête » / « <u>Il fait</u> naître <u>tout</u> à <u>coup</u> un silence » (Segalen1921 vs Puget1624) « Le sol tout entier a dû être créé par l'homme <u>avant qu'il pût s'y établir</u>. » / « Le roi Éétès a fait subir à Jason de terribles épreuves <u>avant qu'il puisse s'emparer</u> de la Toison d'or. » (Ballanche1829 vs Laffon2016)
Constructions semblables, différents sens	« <u>J'ai perdu</u> Eurydice » / « Ah ! je suis d'un tel coup si rudement saisi que ma parole en flamme à mes lèvres expire, et que sur tous mes sens <u>j'ai perdu</u> mon empire ! » (Rachmuhl2003 vs Grandmougin1882) « <u>Je suis</u> désordonnée, c'est vrai, <u>je suis</u> paresseuse, <u>je suis</u> lâche... » / « <u>Je suis</u> un malheureux, <u>je suis</u> fils de la douleur, <u>je suis</u> homme, mère ! » (Anouilh1941 vs Loyson1929) « <u>Tel</u> fils, <u>telle</u> mère... » / « Je veux savoir ; <u>tel</u> est mon grand désir et <u>tel</u> est ton devoir » (Liking1981 vs Grandmougin1882) « Je lui jette ceci en pleine face : <u>tu n'es pas</u> un vrai chef de horde ! <u>tu n'es pas</u> un gardien de troupeau ! » / « Non, <u>tu n'es pas</u> assez sotte ; non, non, ma chère Javotte, non, <u>tu n'es pas</u> assez sotte pour mourir pour tout de bon. » (Segalen1921 vs Dorvigny1775) « <u>Sois béni</u>, ô Vulcain !... » / « Ô Phœbus Apollon, dieu vivant, <u>sois béni</u> ! » (Crémieux1858 vs Grandmougin1882) « Plus loin, les vautours <u>cessent brusquement de</u> déchiqueter de leurs becs voraces les entrailles de Tityos, et ses mains peuvent oublier de les repousser » / « Neuf heures sonnent au loin. Le père <u>cesse brusquement de</u> ronfler et s'éveille avec des borborygmes. » (Goudot2003 vs Anouilh1941)

Banalités	« Chaque nom voulait <u>dire quelque chose</u>, quelque chose qui était en rapport avec les circonstances de la naissance » / « Mais seulement pour en <u>dire quelque chose</u> qui tesmoigne que c'est un ombre de leur corps » (Liking1981 <i>vs</i> Puget1624) « <u>Je t'en prie</u>, n'est-ce pas ! » / « Sois aimable, <u>je t'en prie</u>. » (Cocteau1927 <i>vs</i> Anouilh1941) « Orphée était trop préoccupé par sa marche pour <u>poser la question</u> à son guide. » / « Depuis les soins miraculeux de mon frère, j'avais commencé à chercher je ne sais quoi, à <u>poser des questions</u> » (Goudot2003 <i>vs</i> Liking1981)
Mots forts en commun sont isolés ou appartiennent aux syntagmes différents	« Le <u>seul jeune homme</u> qui s'acharnait à suivre les voies de la tradition, tout en vivant avec son temps. » / « Au fond, un <u>seul</u> client, un <u>jeune homme</u>, chapeau rabattu sur les yeux, en imperméable, l'air absent. » (Liking1981 <i>vs</i> Anouilh1941) « <u>Je</u> suis bien trop <u>triste</u> pour <u>pleurer</u>. » / « Si le ciel est <u>triste</u>, <u>je</u> <u>pleure</u> ! » (Anouilh1941 <i>vs</i> Grandmougin1882) « N'<u>interromps</u> pas une telle paix, <u>supplia</u> une voix féminine. » / « Les <u>supplices</u> des criminels, qui expiaient leurs fautes dans les Enfers, s'<u>interrompirent</u>. » (Goudot2003 <i>vs</i> Rachmuhl2003)
Uniquement noms propres (et mots faibles) en commun	« Mais elle <u>n'est pas</u> morte avec toi, <u>Orphée</u> ! » / « Ciel ! ce <u>n'est pas Orphée</u> !... » (Segalen1921 <i>vs</i> Defontenay1854) « Mais, par un dernier sortilège, sa tête <u>appelait</u> encore : “<u>Eurydice</u> ! <u>Eurydice</u> !” » / « Ils <u>appellent Eurydice</u>. » (Laffon2016 <i>vs</i> Villenave1806) « (Pendant que Jupiter a le dos tourné, <u>Pluton</u> fait des signes à John, comme pour lui demander si <u>Eurydice</u> est dans l'appartement du fond.) » / « Ainsi retourna pour toujours l'infortunée <u>Eurydice</u> dans le Roïaume de <u>Pluton</u>. » (Crémieux1858 <i>vs</i> Banier1732)
Ressemblances génériques	« Orphée <u>regarde autour de lui</u> sans comprendre. » / « Orphée <u>regarde autour de lui</u>, le chœur s'en approche » (Anouilh1941 <i>vs</i> Moline1774) « <u>Le théâtre</u> change et <u>représente</u> le Palais de Pluton. » / « <u>Le Théâtre</u> <u>représente</u> un bois au milieu duquel on a élevé un tombeau rustique » (Fuzelier1711 <i>vs</i> Deshayes1792)
Affirmations appairées avec négations	« Autrement <u>tu perdras</u> pour la dernière fois <u>sans espoir</u> de revoir celle que tu reçois » / « Alors <u>un espoir fou monta</u> en lui : “<u>Rien n'est perdu</u> !”, songea-t-il tout à coup » (Espine1614 <i>vs</i> Jimenes2004) « Aristée, laisse-moi, lui dit sèchement Eurydice. <u>Je ne t'aime pas</u>. <u>Je ne t'aimerais jamais</u>. » / « Que lui as-tu dit ? [Eurydice :] Je lui ai dit que <u>je l'aime</u>... » (Pommaux2009 <i>vs</i> Segalen1921)
Différents sens des mots communs	« Cela fait <u>des jours et des nuits</u> que nous attendons, aux aguets » / « celle qui colore tout sur la terre et dans les airs, celle qui fait <u>nos jours et nos nuits</u>. » (Laffon2016 <i>vs</i> Ballanche1829) « Un monde où la Justice avait un <u>libre</u> cours ? » / « Et que j'ai refusé parce que la place de violon n'était pas <u>libre</u> pour toi ? » (Grandmougin1882 <i>vs</i> Anouilh1941) « Regarde... Ils arrivent eux aussi les temps où le Grand Abîme fait <u>coulisser</u> sa <u>machine</u> et il bée là assoiffé de sang et de pourriture gare ! » / « Voici venir notre Eurydice ; je pars : – mais je suis toujours là, prêt à sortir de la <u>coulisse</u>, comme un <i>Deus ex machina</i> ! » (Liking1981 <i>vs</i> Crémieux1858) « Je <u>cours</u> saluer et embrasser mes frères comme autrefois. » / « Un monde où la Justice avait un libre <u>cours</u> ? » (Defontenay1854 <i>vs</i> Grandmougin1882)

Soulignons d'abord que malgré le fait que nous choisissons d'écarter les mots faibles en paramétrant le traitement des deux logiciels, une partie des couples n'a tout de même que ces mots-là en commun. Nous constatons que ces relations se retrouvent majoritairement parmi les résultats de Tracer, ce qui nous permet de réaliser que, même si nous pouvons lui soumettre une liste des mots faibles dans les paramétrages, ce logiciel ne l'exploite pas, mais procède à un filtrage des mots sur

la base de leurs fréquences. Or, la spécificité de notre corpus fait que parmi les 150 mots les plus fréquents nous retrouvons également les mots-clés du mythe, comme les noms des protagonistes (« Orphée » est le 49^e mot le plus fréquent, et précède les verbes comme « dire », « voir », « aller » ou encore « venir »¹¹) ou les thèmes majeurs (« amour », « mort », « lyre »). Nous ne pouvons donc pas nous permettre d'écarter trop de mots les plus fréquents, puisque, outre les mots faibles, des éléments cruciaux pour notre recherche seraient également ignorés.

Remarquons aussi que la grande majorité des couples du type 2 partage effectivement des mots communs forts, l'appréciation de leur pertinence n'est donc pas en lien avec les performances de logiciels, mais dépend des questions particulières posées par le projet de recherche. Les résultats d'évaluation sont donc susceptibles de varier d'un projet à l'autre, même si le corpus reste le même. Par ailleurs, nous constatons que certains facteurs qui font qu'un couple est considéré comme non pertinent sont également caractéristiques de certains couples classés comme pertinents. C'est par exemple le cas des correspondances liées par la répétition d'apostrophe « Eurydice » :

« Mais, au lieu de me répondre, il appelait : « Eurydice ! Eurydice ! » en sanglotant. Un nom que je ne connaissais pas. » (Laffon2016)
« Eurydice, Eurydice ! / De ce doux nom / Tout retentit, ces bois, ces rochers, ce val-lon. » (Moline1774)

En effet, les entités nommées s'avèrent souvent cruciales pour l'identification de l'épisode du mythe voire la révélation d'une réécriture originale. Il peut par exemple s'agir de l'occurrence d'un nom propre qui n'est pas habituellement présent dans le récit mythique, et qui apparaît dans deux réécritures distinctes.

Dans l'échantillon, nous trouvons notamment le nom de Talaon, un personnage important dans l'épopée de Ballanche¹⁸²⁹ (« Va, ce n'est pas Talaon qui te conseillera une vie oisive »). Ce titan, présenté comme père d'Eurydice, est repris dans le poème de Divoire¹⁹²² (« Fille de Talaon, il fallait que l'élu connût les stations qui marquent leur voie aux fils d'Apollon »). Même si les deux extraits n'ont en commun que ce nom, la relation est significative, car elle peut constituer la preuve de l'influence de l'œuvre de Divoire¹⁹²² par celle de Ballanche¹⁸²⁹.

Inversement, nous relevons aussi un couple où la phrase « Bravo, Aristée, berger des abeilles ! » de Jimenes²⁰⁰⁴ est alignée avec un long segment de Puget¹⁶²⁴ (Tableau 58). Ils partagent deux mots forts en commun « berger » (« Le Berger obstiné ») et « abeilles » (« il s'anime pourtant peu à peu, et après que ses lèvres, comme deux nouvelles Abeilles, eurent couvé sur toutes les fleurs de son visage pour en flairer l'odeur »). Cependant, le deuxième extrait contient également un nom propre,

¹¹ Voir, dans les annexes du chapitre II.3 « Prolégomènes des approches computationnelles de l'intertextualité », le fichier [II_3_150-mots-frequents](#).

qui n'est pas celui d'Aristée mais de Celidor, qu'on retrouve par exemple dans la *Pastorale du berger Célidor et de Flormonde sa bergère* de 1629¹².

Tableau 58. Couple de type 5 entre Jimenes2004 et Puget1624

« Bravo, [Aristée], <u>berger</u> des <u>abeilles</u> ! » (Jimenes2004)	« Sur ces mots, il se met en action pour la baiser, à quoy elle resiste, [Celidor] la presse et la force à luy permettre de prendre quelque baiser qui sert de matiere au feu qui embrase son ame, il s'anime pourtant peu à peu, et apres que ses levres, comme deux nouvelles <u>Abeilles</u>, eurent coureu sur toutes les fleurs de son visage pour en flairer l'odeur, la passion le fait souspirer apres de plus solides alliments, mais il trouve une si forte resistance à ces attaques qu'à mesme temps qu'il conçoit le desir de son triomphe, il en perd l'esperance : car Euridice emprunte tous les traicts de l'Amour, de la Colere, et de la Vengeance, et les décoche contre le <u>Berger</u> obstiné qui en considerant avec admiration le pris du Laurier de sa victoire, ranime de nouveau son courage pour en estre le conquerant. » (Puget1624)

Par ailleurs, certaines relations basées sur des expressions très fréquentes (que nous appelons les « banalités ») sont tout de même classées comme pertinentes. C'est par exemple le cas du rire qui met en commun deux pièces comiques (« Ah ! ah ! ah ! etc., etc., etc. [Le chœur :] Ah ! ah ! ah ! etc., etc., etc. » (Crémieux1858) / « Ah ! ah ! vous avez bon air ! Bon air vous avez ! [Chœur des forgerons :] Ah ! ah ! vous avez bon air ! » (Dorvigny1775)) ou encore du placement de l'action au bord d'un fleuve, qui est un décor central du mythe (« On est tout au bord du Fleuve, en pleine nuit, sous la lune plus haute. » (Segalen1921) / « Il est au bord d'un fleuve » (Montardre2013)).

De même, certaines affirmations appariées avec les négations peuvent être considérées comme indices de réécriture quand elles concernent les épisodes clés du mythe. Dans le couple suivant, c'est le retour d'Eurydice des Enfers qui semble être remis en question :

« Eurydice, l'amour de ton époux m'a convaincu de te laisser repartir parmi les vivants » (Bottet2019) / « – Laisse Eurydice parmi les Ombres, lui ordonne-t-il, et ses yeux phosphorescents jettent d'inquiétantes lueurs. » (Laffon2016)

Dans le roman de Jimenes2004, quant à lui, Orphée fait une déclaration très conforme à la trame principale du mythe : « Moi qui aime Eurydice et qui n'ai jamais douté de mon amour pour elle, même au plus profond des Enfers ». Mais cette affirmation classique est appariée à une autre, plus déconcertante, car issue de la pièce de théâtre de Cocteau1927 : « [Orphée :] Je t'aime. [Eurydice :] Tu aimes le cheval. Je passe en second. ».

¹² Voir Bénédicte Louvat, « Influence italienne et satire médicale dans la *Pastorale du berger Célidor et de Flormonde sa bergère* (Béziers, 1629) », *Quand Minerve passe les monts : modalités littéraires de la circulation des savoirs, Italie-France, Renaissance-XVII^e siècle*, International, Pessac, Presses Universitaires de Bordeaux, 2021, p. 45-55.

Couples du type 3

Les couples pertinents (3-5) sont majoritaires parmi l'échantillon (presque 63 %), mais les résultats de Tracer paraissent bien plus probants que ceux de TextPAIR (75,2 % *vs* 50,6 %). Il convient toutefois de noter que l'échantillon de Tracer contient beaucoup plus de couples du type 3, dont la pertinence effective est incertaine. Cela est dû principalement aux deux facteurs.

Premièrement, le contexte auquel nous accédons est souvent trop restreint pour interpréter la relation. Cela peut aller d'une difficulté relativement simple, comme le besoin d'identifier le référent d'un pronom personnel (« Elle est là !... j'en suis sûr ! » (Crémieux1858) / « Mais Nyango vit, j'en suis sûr. » (Liking1981) ou d'examiner la concordance de la mention des deux personnages habituellement extérieurs au mythe (« Les prêtres d'Isis et d'Osiris m'ont livré leurs secrets » (Schuré1889) / « Isis et Osiris représentent les deux sexes de la nature » (Esquiros1849)). Mais dans beaucoup de cas, il s'agit de deux réécritures symboliques qui nécessitent une analyse plus approfondie (« Je suis comme les animaux sauvages ; je n'ai d'autre jardin que les forêts inhabitées ; la solitude est mon domaine, le ciel est mon pavillon » (Ballanche1829) / « Son regard a cessé de tendre vers le ciel, vers les solitudes du ciel et l'abstrait domaine des nombres » (Divoire1922)).

Ensuite, les appariements d'une phrase courte avec une phrase longue sont souvent difficiles à évaluer, surtout si au moins un des segments est symbolique. Voyons par exemple ce couple trouvé entre Grandmougin1882 et Chapoton1648 :

Tableau 59. Couple du type 3 entre Grandmougin1882 et Chapoton1648

J'attends sans peur, je suis tranquille, je vous fais de moi-même un immense abandon. (Grandmougin1882)	Je veux, je n'ose pas, je cherche qui me nuit ; le repos m'abandonne, et la douleur me suit ; je brûle de vous voir, mais je perds cette envie quand je me ressouviens que vous perdriez la vie, je fais vœu d'obéir, et d'abord dans mon sein un autre sentiment succède à ce dessein, car les tristes pensers où la fureur me plonge, me font croire qu'icy tout passe pour un songe, et je crains justement de me trouver confus, s'il faut qu'à mon réveil je ne vous trouve plus, mais je crains vainement l'autorité fatale du Prince qui regit cette terre infernale, ne sçauroit m'abuser, ce que donnent les Dieux est toujours si parfait et si Religieux qu'il n'en provient jamais une mauvaise issue, quand je craignois pour vous mon ame estoit deceuë achevons cette course, et dans peu le Soleil fera voir à mes yeux un tresor sans pareil ; mais je veux que Pluton me tienne sa promesse, ne vous perdez-vous point dans cette nuit épaisse ? (Chapoton1648)

L'analyse de cette relation demande beaucoup de concentration, d'abord pour savoir quels mots forts du segment plus court se retrouvent dans le plus long, et ensuite pour confronter un texte en français moderne à celui en moyen français. Par conséquent, certains couples sont sans doute classés initialement dans le type 3, alors qu'une relecture plus attentive, voire une visualisation plus parlante, permettrait de trancher plus facilement entre le type 2 et le type 4.

Les résultats du type 3 de TextPAIR, en revanche, sont plus complexes et moins liés au fonctionnement du logiciel. Nous constatons toutefois qu'alors que ce logiciel, non restreint comme Tracer par la segmentation que nous lui imposons, est *a priori* capable de détecter des segments communs plus étendus et des reformulations incluses dans de longs fragments, l'échantillon a révélé que cela n'est vrai que pour des réemplois très proches (type 5). Pour des correspondances plus subtiles, au contraire, TextPAIR trouve des extraits très courts, reliés par des syntagmes ou des constructions polylexicales très proches. C'est surtout ce type de détections que nous retrouvons parmi les couples du type 3 (« Construis-nous quelque dieu que l'on puisse adopter » (Diivoire1922) / « Vous venez sous l'adveu de quelque Dieu Celeste » (Chapoton1648)).

Couples du type 4

Les cas de figure des relations plutôt allusives sont encore plus présents au sein des couples du type 4. Tout d'abord, nous relevons plusieurs relations où un texte qui suit de près la trame classique du mythe est apparié à un autre qui la modifie (Tableau 60).

Tableau 60. *Quelques exemples des métaphorisations, poétisations ou ridiculisations au sein des couples du type 4*

« Tout vous est dû et après nous être un peu attardés, nous nous hâtons tôt ou tard vers le même séjour. » (Beague1998-2)	« À quoi bon le travail, dont, sans peine et sans bruit, un ver doit, tôt ou tard, anéantir le fruit ? » (Launay1922)
« La nuit, sur son rameau, sans fin, la pauvre mère envoie au loin sa plainte amère, et l'écho gémit de ses cris. » (Fournier1876)	« Si seulement – c'est donc un désir bien superbe – quand j'ai dit : « Courbe-toi », je voyais le brin d'herbe s'incliner, comme il fait quand a parlé le vent ; si l'oiselet voulait – ô rêve décevant ! – quand j'appelle, accourir vers ma douceur pareille à celle du nid tiède, où sa mère le veille ; si l'écho répondait, sans répéter, narquois, mes cris désespérés, mes appels, mes pourquoi !... » (Launay1922)
« Il leur tendait ses mains suppliantes, et sa voix, pour la première fois impuissante, leur adressait des prières inutiles. » (Nisard1869)	« À tous les Dieux j'adresse ma priere, hélas ! autant en emporte le vent. » (Deshayes1792)
« Mais..., ajoute aussitôt Hadès sévèrement, à une seule condition, ne te retourne jamais pour voir ta femme, jusqu'à ce que vous ayez franchi la Porte des Enfers. » (Laffon2016)	« [Isabelle :] S'il est étonnant de voir un mari chercher sa femme jusqu'aux enfers, il ne l'est pas moins de voir une femme souhaiter avec empressement de retourner avec son mari, quand une fois elle en a été séparée. [Pluton :] Voilà un petit début qui n'est point sot. » (Regnard1689)

Dans les deux premiers couples, nous observons les liens entre deux traductions d'Ovide et le poème de Launay1922 qui a recours tantôt à une métaphore de la mort (« un ver doit, tôt ou tard, anéantir le fruit »), tantôt à une poétisation de l'épisode de Philomèle métamorphosée en oiseau et pleurant ses petits. Les deux autres couples présentent plutôt des ridiculisations des mythes

originels : chez Deshayes1792, « les prières inutiles » deviennent « autant en emporte le vent », alors qu'un des personnages de Regnard1689 s'étonne qu'un mari veuille chercher sa femme aux Enfers (et que celle-ci souhaite en sortir avec lui).

Les couples du type 4 regroupent aussi des procédés de ridiculisation semblables qui peuvent être observés entre différentes parodies du mythe (Tableau 61).

Tableau 61. *Quelques exemples des procédés de ridiculisation semblables au sein des couples du type 4*

« [Orphée :] Eurydice !... ma femme ?... [Eurydice :] Ah ! mais, c'est qu'il est temps de s'expliquer, à la fin ! » (Crémieux1858)	« [Orphée :] Où est-elle ? Eurydice !... Elle boude. Ah ! ça... Je deviens fou ! » (Cocteau1927)
« [Eurydice :] Je suis laide ? Quelquefois, quand j'ai pleuré ou trop ri, il me vient une petite tache rouge au coin du nez. J'aime mieux vous le dire tout de suite, pour que vous n'ayez pas une mauvaise surprise après. [Orphée :] Je me ferai une raison. » (Anouilh1941)	« [John :] Je n'ai qu'un défaut, madame, j'aime mieux vous le dire tout de suite, pour que vous ne me le reprochiez pas plus tard : je m'enivre quelquefois !... [Eurydice :] Il n'est pas fou, le malheureux, il est gris ! » (Crémieux1858)
« Eurydice s'ennuya ; elle alla au bois, et, rencontrant sur ses pas Aristée, – un élégant de ce temps-là, – elle le regarda longuement. Pardonnez-lui ce regard. Elle était femme, elle s'ennuyait, et, comme toutes les femmes qui s'ennuient, elle voulait un combat – pour avoir un vainqueur. » (Eimann1854)	« [Colombine :] Oui, tu le prends comme cela ? Et moi, je ne veux pas sortir. Une jolie femme comme moi, en tout pays, ne manque point de mari. [Arlequin :] Oh ! Je sais bien qu'il y a partout assez de gens qui se mêlent de ces emplois-là. » (Regnard1689)

Nous voyons que ce type de relations ne constitue pas de preuve d'une inspiration d'un auteur par un autre, mais montre plutôt une proximité générique, une manière semblable de parodier différents épisodes du mythe. Ainsi, le premier couple permet de constater que tant Crémieux1858 que Cocteau1927 ridiculisent l'amour d'Orphée et Eurydice : les deux amants se disputent, Eurydice « boude », etc. Alors que, pour le couple établi entre Crémieux1858 et Anouilh1941, le personnage qui prend la parole n'est pas le même (Eurydice / John Styx), la proximité réside dans les aveux inattendus et plutôt absurdes. Dans les deux cas, l'aveu est accompagné par des affirmations semblables : « J'aime mieux vous le dire tout de suite, pour que vous [n'ayez pas une mauvaise surprise après / ne me le reprochiez pas plus tard] ». Enfin, dans le troisième couple, Eimann1854 et Regnard1689 mettent tous les deux en doute la fidélité d'Eurydice.

Pour d'autres couples, il s'agit plutôt de la modification des épisodes du mythe, que ce soit dans le but de ridiculisation ou de modernisation. Dans le Tableau 62, les textes qui suivent fidèlement la trame originelle (à gauche) sont appariés à ceux qui s'en éloignent (à droite). Nous pouvons alors constater que Grandmougin1882 altère le paysage paisible de la lisière de la forêt où Orphée et Eurydice se sont mariés, Laffon2016 amplifie les échanges entre Orphée et les infernaux en les rendant plus menaçants et Defontenay1854 introduit de nouveaux personnages, Cresphonte et Phaëtuse qui, en plus des bacchantes, sont responsables de la mort d'Orphée.

Tableau 62. Quelques exemples de la modification de la trame originelle du mythe au sein des couples du type 4

« Un cortège joyeux se forma dès l'aube, amenant Eurydice à la lisière de la forêt. » (Jimenes2004)	« Dans le ciel a rougi déjà plus d'une aurore depuis qu'un jour, non loin de l'azur du Bosphore, je vous trouvais sanglants auprès de la forêt où mon hymen paisible et joyeux était prêt ! » (Grandmougin1882)
« Eurydice, l'amour de ton époux m'a convaincu de te laisser repartir parmi les vivants » (Bottet2019)	« – Laisse Eurydice parmi les Ombres, lui ordonne-t-il, et ses yeux phosphorescents jettent d'inquiétantes lueurs. » (Laffon2016)
« Comme les oiseaux s'assemblent à l'entour d'un Hibou, quand ils le rencontrent de jour, comme ce nombre de chiens qu'on voit le matin dans l'amphitheatre, se vont jeter sur le Cerf qui en sera bien-tôt la proie ; Tout de même les Bacchantes se précipitent sur Orphée, et le frappent avec leurs Thyrses qui n'étoient pas faits pour cet usage. » (Duryer1702)	« Orphée, la tête fracassée par Cresphonte, est précipité dans le fleuve par les mains de Phaëtuse et des bacchantes... » (Defontenay1854)

Un peu moins intéressantes du point de vue analytique, certains couples sont appariés, alors qu'elles renvoient aux différents épisodes du mythe (Tableau 63).

Tableau 63. Quelques exemples des couples du type 4 qui renvoient aux différents épisodes du mythe

« Mais la reine des enfers lui avait défendu de tourner la tête, et de jeter les yeux sur son épouse. » (Desfontaines1810)	« Roi et reine des Enfers, je demande humblement votre grâce pour Eurydice. » (Bottet2019)
« Son crime est tel que nous l'égorgerons sur notre antique autel ! » (Grandmougin1882)	« Là les Dieux infernaux sur de sombres autels, écoutent tous les vœux qui leur font les mortels, et s'appaisants souvent par le sang des victimes divertissent les maux et pardonnent les crimes. » (Chapoton1648)
« Si je charme les tigres, pierres et les chênes, se dit-il, pourquoi pas les créatures et les dieux du royaume des Morts ?... » (Pommaux2009)	« On dit que durant sept mois entiers l'inconsolable Orphée pleura au pied d'un rocher, sur les bords déserts du Strymon, et fit retentir du récit de ses douleurs les antres glacés de la Thrace, entraînant sur ses pas les tigres adoucis et les chênes mêmes, charmés de la douceur de ses chants. » (Desportes1846-2)
« Les Bacchantes affolées et les Thraces saisis par l'horreur de leur crime, s'enfuirent dans la nuit en poussant des cris de détresse. » (Schuré1889)	« Comprenant qu'il leur était impossible de rivaliser avec lui, les créatures s'enfuirent en poussant des cris stridents qui achevèrent de dégriser les Argonautes. » (Jimenes2004)

Si nous considérons ces relations comme pertinentes et les classons dans le type 4, c'est parce que certaines réécritures ne se consacrent pas entièrement au mythe. Il est donc intéressant de disposer des extraits qui permettent de localiser les segments spécifiquement dédiés à un épisode ou un autre. Dans le sens inverse, ces détections peuvent également aider à écarter les épisodes qui ne concernent pas l'histoire d'Orphée et Eurydice. C'est par exemple le cas du dernier couple où les

cris des Bacchantes chez Schuré1889 sont appariés à ceux de Sirènes, décrits par Jimenes2004 qui intègre le récit du voyage d'Orphée avec les Argonautes. Par ailleurs, la formulation semblable des épisodes différents constitue également un indice de la manière dont le mythe est réécrit : Pommaux2009, par exemple, modifie la chronologie originelle du mythe en plaçant l'épisode d'enchantement des arbres et d'animaux avant celui de la descente aux Enfers.

Certains couples du type 4 sont bien plus difficiles à évaluer, car tant le contenu des segments que la relation qu'ils entretiennent sont symboliques et/ou très allusifs. En revanche, le contexte est généralement suffisant et, même si les segments sont courts, ils ont tout de même des mots-clés du mythe en commun, ce qui fait que ces couples sont évalués comme du type 4 et non pas du type 3. Leur pertinence effective reste toutefois à confirmer. Dans Tableau 64, nous présentons quelques couples ayant la symbolique ou, *a minima*, la thématique commune.

Tableau 64. Quelques exemples des couples du type 4 ayant la symbolique ou la thématique commune

« Toute idée se dérobaît, en Égypte, sous un symbole ; et le trésor des connaissances humaines était enfoui dans des criptes fermés à la lumière du soleil. » (Esquiros1849)	« Comme tous les peuples antiques qui reçurent leur organisation des Mystères, comme l'Égypte, comme Israël, comme l'Étrurie, la Grèce eut sa géographie sacrée, où chaque contrée devenait le symbole d'une région purement intellectuelle et supraterrrestre de l'esprit. » (Schuré1889)
« Ma bien-aimée, lui disait-il, je me reproche maintenant de t'avoir arrachée à ton sol natal. » (Ballanche1829)	« Ma bouche, avec bonheur, la nomme bien-aimée, mais, vois-tu, ce que j'aime avant tout, c'est mon sol ! » (Grandmougin1882)
« Plus l'Envie à son bras oppose d'Ennemis, et plus grande sera sa Gloire. » (Duboullay1690)	« Mais l'Envie s'oppose à l'accomplissement de ses vœux, car toutes les Nymphes les plus parfaites vont à la chasse des affections de son Orphée, armées des puissants attraites de leurs beauté qui ne trouvent jamais de la résistance. » (Puget1624)
« Les choses que je viens de te dire sont des germes que je dépose dans le champ défriché de ton esprit. Il faut que ces germes se développent eux-mêmes. Maintenant donc nous te livrons à tes propres méditations. » (Ballanche1829)	« Il avait succombé dans sa tâche étourdie./ Moisson de vérité que notre esprit mendie,/ Tu ne peux d'un labour croître au champ défriché./ Il faut l'effort de tous à la glèbe attaché,/ Non l'égoïste ardeur de quelque orgueil superbe ;/ L'homme plante un épi ; Dieu seul groupe la gerbe ! » (Launay1922)
« Hâte-toi, mon bien-aimé, de donner le dernier baiser du frère, le premier et dernier baiser de l'époux à la vierge qui ne pouvait savoir de toi les secrets divins de l'épouse. » (Ballanche1829)	« Tu sçauras encore, mon cher Espoux, luy dit-elle, qu'après que Caron m'eut fait passer le fleuve Stix, il me vouloit baiser pour son droict de passage, ayant pris d'Atropos que j'estois nouvellement mariee, à quoy j'ay resistay, luy répondant que j'avois esté veritablement mariee, mais que pourtant j'estois fille, et qu'ainsi, il n'avoit point droit de me baiser. » (Puget1624)

Chez Esquiros1849 et Schuré1889, par exemple, la mention du nom propre « Égypte », absent du récit classique du mythe, est accompagnée par l'occurrence du mot « symbole » dans les deux segments appariés. Trois relations sont établies avec l'épopée très symbolique de Ballanche1829 : l'attachement

au sol (natal) est le sujet commun de la relation avec Grandmougin1882. Dans celle avec Launay1922, nous retrouvons la métaphore agricole de l'évolution des peuples, alors que pour le couple établi avec Puget1624, c'est la virginité d'Eurydice qui est abordée.

Enfin, nous notons également quelques renvois qui pourraient constituer une preuve d'inspiration d'un texte par un autre, mais qui sont trop allusifs pour que les couples soient classés dans le type 5 (Tableau 65).

Tableau 65. *Quelques exemples des renvois allusifs au sein des couples du type 4*

« Pour la lumière qu'ils apportèrent/ Et pour l'espoir qui en naîtra.../ Orphée, Nyango, Nyango – Orphée./ Deux noms qui sonnaient bien./ Allaient-ils bien ensemble ? » (Liking1981)	« On a célébré les noces d'Orphée et de Nyango ; deux familles se sont réconciliées. Il fallait un peu d'harmonie dans ce monde déchiré ! » (Manjock1981)
« [Le Chœur :] Non, non, nous sommes inflexibles,/ Non, la pitié deux fois n'entre point aux Enfers. » (Duboullay1690)	« Soyez, soyez sensibles/ À l'excès de mes malheurs ! [Furies :] Non ! Non ! Non ! [Orphée :] Laissez-vous toucher par mes pleurs, etc. » (Moline1774)
« Il faut les vaincre, les soumettre au joug des lois, et leur faire aimer l'ordre social que nous leur imposerons. Pourquoi n'auraient-ils pas à leur tour la connaissance d'eux-mêmes, la capacité du bien et du mal ? » (Ballanche1829)	« Taureaux d'airain jetant le feu par les narines,/ Qu'il faut soumettre au joug, ivres d'un fol effroi ;/ C'est le champ labouré sur l'ordre du vieux roi,/ Où des dents de serpents, graine à tous vents semée,/ Pour d'étranges horreurs font germer une armée » (Launay1922)

En effet, certains textes sont plus ou moins susceptibles d'influencer les réécritures postérieures. Il s'agit notamment des relations entre une œuvre et son adaptation, comme c'est le cas du roman de Liking1981, adapté en théâtre-rituel par Manjock1981, des relations entre les œuvres du même genre (et sous-genre), comme la tragédie en musique de Duboullay1690 et l'opéra héroïque de Moline1774, ou encore celle entre les œuvres proposant le même type de transformation du mythe (idéologique pour Ballanche1829 et Launay1922).

Couples du type 5

Enfin, les couples du type 5 constituent un ensemble relativement facile à évaluer. Il s'agit pour la plupart des extraits quasi identiques ou dont la référence au même épisode du mythe paraît de manière évidente. Parmi ces couples, ceux qui présentent des liens entre une traduction d'une source antique et une réécriture sont très présents (39/82 pour Tracer, 27/73 pour TextPAIR). Nous en présentons quelques exemples au Tableau 66.

Tableau 66. Relations du type 5 entre les traductions de Virgile et d'Ovide (à gauche) et les réécritures (à droite)

« Par quels pleurs, par quels chants pourra-t-il toucher encore les divinités des enfers ? déjà la froide Eurydice voguait dans la barque du Styx. » (Nisard1868)	« Pour comble de douleur, il revoit alors son Eurydice qui voguait pâle et froide sur la fatale barque. » (Esquiros1849)
« Et le pacte est rompu, Pluton reprend sa proie ; trois fois l'Averne en pousse un hurlement de joie ! » (Guillois1863)	« Alors le pacte fut rompu, et le tyran des morts reprit ses droits ; alors toutes les voûtes de l'Averne retentirent, dans leurs vastes profondeurs, d'un long gémissement ; alors les supplices des coupables, qui avaient été suspendus, reprirent une activité nouvelle ; alors le terrible gardien des demeures désolées poussa d'affreux rugissements. » (Ballanche1829)
« Tantale oublia sa soif, et cessa de chercher l'eau qui le fuit ; la rouë d'Ixion s'arrêta ; les vautours ne rongèrent plus le cœur de Titye ; les Belides ne songèrent plus à remplir leurs tonneaux ; Sisiphe se reposa sur la pierre qu'il roule éternellement. » (Bellegarde1701)	« Sisiphe en oublia de tenir son rocher, Tantale cette soif qu'il ne peut étancher et les cruelles sœurs, les fiers Danaïdes, ne s'aperçoivent pas que leurs seaux estoient vuides : » (Hermite1641)
« L'ame de Orpheus descendit en enfer sans demourance ou il vit la valee tenebreuse et les regnes que paravant avoit veuz. » (Walleys1493)	« Quand Orphée descendit aux Enfers pour demander sa femme à Pluton ; voici ce que le Monarque Ténébreux lui dit, en la lui rendant. » (Dorvigny1775)

Même si les formulations ne sont pas toujours identiques, la présence des mots-clés saillants de chaque épisode (noms propres, verbes comme « descendre », etc.) facilite l'évaluation de la relation, même sans avoir accès au contexte plus large.

TextPAIR paraît très efficace tant pour repérer des proximités étendues (comme pour les cantates de Guillion1770 et Berlioz1828), que pour trouver des emprunts plus localisés (mais tout de même étendus), soit à la traduction d'une source antique, comme c'est le cas d'Esquiros1849 qui transforme de manière intéressante un extrait de Charpentier1831, soit à une autre réécriture, comme une reprise de l'opéra-comique de Deshayes1792 par celle de Crémieux1858 (Tableau 67).

Tableau 67. Quelques exemples de couples du type 5 de l'échantillon de TextPAIR

« Ô prodige tonnant, ô céleste harmonie ! Tout l'univers est soumis à tes loix, » (Guillion1770)	« Ô prodige étonnant ! ô celeste harmonie ! Tout l'univers est soumis à tes lois. » (Berlioz1827)
« Tout à coup sa tendresse imprudente le trahit : faute bien pardonnable, si l'enfer savait pardonner ; il s'arrête, et déjà aux portes du jour, oubliant sa promesse et vaincu de son amour, il se retourne » (Charpentier1831)	« quand je ne sais quelle démence saisit le trop faible amoureux : faute bien pardonnable sans doute, si les prêtres savaient pardonner ! Arrivé presque à l'embouchure du puits, au moment où une demi-clarté annonçait un retour prochain au monde extérieur, il s'arrêta. » (Esquiros1849)
« [Air :] Va-t'en voir s'ils viennent. [Euridice :] <i>(passant doucement autour d'Orphée)</i> Mais je suis en vérité/ Bien mal avisée. » (Deshayes1792)	« [Jupiter, à part :] Crampon, va !... <i>(À Mercure.)</i> Va-t'en voir s'ils viennent !... <i>(Rêvant.)</i> Cette petite Eurydice est donc bien jolie ?... » (Crémieux1858)

Tracer, quant à lui, regroupe tout de même beaucoup de phrases identiques, mais dont la pertinence analytique reste à évaluer. Dans l'échantillon, la phrase « Je t'en supplie » de Goudot2003 est alignée à deux autres segments, « Grâce, je t'en supplie... » de Crémieux1858 et « Mon chéri, tais-toi, je t'en supplie ! » de Anouilh1941. *Supplier* fait partie d'un des mots-clés du mythe, ce qui fait que ces relations ne sont pas considérées comme banalités, mais le contexte plus large peut nous amener à reconsidérer cette évaluation initiale.

En revanche, le logiciel peut aligner deux segments de taille différente, ce qui facilite par exemple le repérage des reformulations plus poétiques, comme c'est le cas de Puget1624, aligné avec la phrase plus expéditive de la traduction d'Ovide par Nisard1869 (Tableau 68). Il est aussi capable de mettre en relation deux constructions semblables, sans que les deux segments partagent beaucoup de mots communs, comme c'est le cas de trois autres couples présentés au Tableau 68.

Tableau 68. Quelques exemples de couples du type 5 de l'échantillon de Tracer

« Il l'aimait. Adieu ! ce fut le dernier adieu, et à peine parvint-il aux oreilles d'Orphée : déjà l'Enfer a reconquis sa proie » (Nisard1869)	« Cependant, la Mort devient exorable pour la seconde fois, et permet à Eurydice d'employer ces dernières paroles à dire adieu à son Orphée : car elle s'écrie d'une voix plaintive, qui blessait les cœurs par les oreilles : » (Puget1624)
« Adieu charmant Orphée, adieu ma chère vie, c'est enfin pour jamais que je te suis ravie. » (Hermite1641)	« Et déjà (ainsi donc) adieu : je suis emportée entourée d'une nuit épaisse, et tendant vers toi de faibles mains, hélas ! moi qui ne suis plus tienne. » (Desportes1846-1)
« Ô ciel ! Qui pourrait en ce jour retenir le transport de mon âme égarée ? » (Moline1774)	« – Comment un mortel égaré parmi les ombres pourrait-il voir l'invisible ? » (Goudot2003)
« Entre les rochers et la mer, où la Thrace barbare arrête son rivage, Orphée erre, sauvage. » (Divoire1922)	« Oui ! c'est toi que sans fin, sur un affreux rivage, seul, et déplorant son veuvage, il chantait jour et nuit encor ! » (Fournier1876)

Les deux couples établis avec des traductions (Hermite1641 / Desportes1846-1 et Divoire1922 / Fournier1876) permettent surtout de localiser l'endroit dans les réécritures Hermite1641 et Divoire1922 où le mythe est abordé de manière plus allusive, sans que cela implique une relation effective entre les textes appariés (la traduction de Desportes1846-1 est d'ailleurs postérieure à la réécriture de Hermite1641, elle ne peut donc pas être considérée comme source). Pour le couple Moline1774 / Goudot2003, la relation reste à confirmer, mais le ton plaintif, la similarité entre « mortel égaré » et « âme égarée », ainsi que la proximité lexicale entre « ciel » et « âme » d'un côté et « ombres », « invisible » de l'autre constituent autant d'indices renforçant le lien entre deux segments.

b) Problèmes de préparation du corpus et du paramétrage des logiciels

L'observation et l'évaluation du corpus nous permettent ainsi de constater que quelques problèmes du traitement persistent et nuisent à la qualité des résultats ou à leur appréciation. Dans

cette section, nous nous focalisons sur cinq problèmes liés à la préparation du corpus et le paramétrage des logiciels : (i) les problèmes liés au format du corpus et des résultats, (ii) l'appariement des phrases ayant uniquement des mots faibles en commun, (iii) les erreurs de lemmatisation qui nuisent à la recherche des mots communs, (iv) les difficultés pour la prise en compte optimale de la synonymie et des reformulations polylexicales et (v) l'appariement des mots issus des syntagmes différents.

Problèmes liés au format du corpus et des résultats

Travailler simultanément avec deux outils de conception différente crée le besoin d'un traitement qui peut surmonter ces différences afin d'obtenir des résultats comparables et compatibles. Cela implique l'identification et l'unification des couples sous un format commun, ainsi que la compilation des extraits détectés par les deux logiciels.

Par ailleurs, nous avons réalisé que le contenu soumis aux deux logiciels diffère, car alors que Tracer demande de partitionner au préalable le corpus en segments, TextPAIR travaille directement à partir des fichiers XML-TEI, mais il n'exploite pas le balisage initial ni le restitué. Les résultats contiennent alors beaucoup de bruit, en incluant les éléments formels des œuvres, comme notes de bas de page, titres de parties, noms de personnages qui prononcent les répliques, etc. Les couples détectés à la base de ces éléments sont majoritairement considérés comme non pertinents, mis à part quelques cas de figure mineurs au sein de l'échantillon. Il semble donc nécessaire de prétraiter également le corpus soumis à TextPAIR, soit en supprimant les éléments XML-TEI non désirables (<note>, <speaker>, <label>, etc.), soit en soumettant les fichiers en format texte.

Enfin, nous avons déjà constaté dans le chapitre précédent¹³ que les résultats des deux logiciels sont susceptibles de contenir des doublons, qui risquent de nuire au repérage des couples pertinents et de fausser certaines analyses, notamment quantitatives, des résultats. Il faut donc les identifier et les supprimer dès la première étape du post-traitement.

Appariement des phrases n'ayant que des mots faibles en commun

Alors que nous pensions pouvoir écarter les appariements des couples n'ayant que les mots faibles en commun grâce au paramétrage efficace des deux logiciels, l'échantillon dévoile que leur quantité reste relativement importante parmi les résultats. Il paraît donc nécessaire de concevoir une méthode de post-traitement des résultats qui permet d'identifier les mots communs, de préciser leur type (mots forts / faibles) et leur degré de similarité (formes, lemmes ou racines communes), puis d'écarter automatiquement les couples qui semblent liés uniquement par des mots faibles.

¹³ Voir III.3 « Présentation des logiciels retenus, TextPAIR et Tracer ».

Erreurs de lemmatisation qui nuisent à la recherche des mots communs

Pour optimiser la qualité du post-traitement et éviter de potentielles erreurs d'alignement, il paraît indispensable de rendre les outils informatiques sensibles aux relations linguistiques fines. Or, l'analyse de l'échantillon révèle que ceux que nous exploitons sont relativement neutres du point de vue des connaissances linguistiques.

Dans le chapitre précédent, nous avons précisé que les deux logiciels permettent quelques paramètres concernant les formes des mots prises en compte : mots pleins, stemmes (mots racinisés) ou lemmes. Lors de nos expérimentations initiales avec le logiciel Phæbus, nous avons constaté la supériorité des recherches basées sur les lemmes par rapport à celles restreintes aux mots pleins ou stemmes¹⁴. Néanmoins, l'évaluation et l'analyse de l'échantillon montrent tout de même quelques limites de cette méthode, comme la confusion des mots homographes (« dur » / « durer »), manque d'appariement des mots de la même famille (« dieu » / « divin »), etc.

Malheureusement, en cherchant une alternative, nous constatons que l'exploitation de deux outils conçus pour le français, DériF¹⁵, un logiciel qui devine les bases des mots construits à l'aide d'expressions régulières, et Morphonette¹⁶, un réseau morphologique regroupant les mots de la même famille, s'avère peu probante pour un corpus littéraire comme le nôtre. L'analyse constructionnelle mise en œuvre par DériF l'amène à apparier *mauve* et *mauvais* ou *vin* et *divin*. Même en cumulant les résultats obtenus avec les deux outils, relativement peu de lemmes peuvent être liés à leur base morphologique (2 007 sur 11 653), et de nombreuses erreurs persistent. Morphonette, par exemple, indique un lien morphologique pour les couples *noceur* / *nocif* – sans lien puisque le premier vient de *noce* et du latin classique *nuptiæ* alors que le second vient du latin *nocivus* – ou *voyant* et *voyage*, la base du premier étant *voir*, alors que le second est construit sur le nom *voie*). Finalement, dans 13 cas seulement l'appariement de mots de classes grammaticales différentes serait facilité par ces outils : *admirable* / *admiration*, *inconsolable* / *consolation*, *embrassant* / *embrassement*, *étonnant* / *étonnement*, *perdurable* / *perdurablement*, *successif* / *successivement*, *berceuse* / *bercer*, *charmant* / *charmer*, *déplorable* / *déplore*, *entendement* / *entendre*, *flottant* / *flotter*, *touchant* / *toucher* et *sonnant* / *sonnette* / *sonneur*.

Alors que l'annotation des bases morphologiques des unités lexicales attestées dans le corpus serait une solution souhaitable, nous nous tournons vers la lemmatisation, moins efficace pour la recherche de correspondances (car elle lie les graphies d'un paradigme flexionnel sans que l'appariement des mots dérivés soit réalisé) mais pour laquelle nous disposons d'outils informatiques plus performants. Toutefois, la coprésence de différents états de la langue au sein du corpus pose des

¹⁴ Voir III.1 « Établissement des critères de l'évaluation avec le logiciel Phæbus ».

¹⁵ Voir <https://www.cnrtl.fr/outils/DeriF/>.

¹⁶ Voir <http://redac.univ-tlse2.fr/lexiques/morphonette.html>.

problèmes aux analyseurs conçus pour traiter des corpus synchroniques. Plusieurs solutions s’offrent à nous : modernisation des graphies anciennes, combinaison d’outils, entraînement des modèles d’étiquetage, correction manuelle... Quelle que soit la méthode privilégiée, il importe de pouvoir prendre la main sur ce prétraitement. Nous employons TreeTagger¹⁷, un annotateur morpho-syntaxique qui fournit un lemme et une catégorie grammaticale pour chaque mot-occurrence. Il est appliqué avec des paramétrages distincts aux textes des différents états de langue réunis dans notre corpus. Nous procédons ensuite à la correction manuelle des lemmes et des catégorisations¹⁸.

Le Tableau 69 présente quelques résultats chiffrés relatifs à la lemmatisation et à la catégorisation automatiques, et à leur correction manuelle. Alors que moins de 8 % des mots-occurrences nécessitent une correction (du lemme, de la catégorie grammaticale ou des deux), presque 30 % des formes sont concernées. Finalement, pour plus d’un tiers des lemmes recensés dans le corpus, au moins une forme graphique a subi une correction manuelle de son lemme et/ou de sa catégorie.

Tableau 69. Résultats de la lemmatisation et de la catégorisation automatique, et de leur correction manuelle

	Items traités par la lemmatisation et la catégorisation automatiques	Items dont le lemme et/ou la catégorie grammaticale ont été corrigés manuellement
Mots-occurrences	424 729	33 527
Formes graphiques	24 692	7 108
Lemmes	11 653	3 971

Par ailleurs, si nous comparons les proportions d’items de chacune des quatre catégories grammaticales majeures au sein de l’ensemble des lemmes recensés au moins une fois dans le corpus et de celui des lemmes dont au moins une forme graphique a été corrigée (Tableau 70), nous observons des proportions comparables de verbes et d’adjectifs, alors que les proportions des noms communs et des noms propres sont contrastées : les premiers sont fréquents parmi les lemmes, et plus rares parmi les lemmes corrigés, les seconds sont rares parmi les lemmes et nettement plus fréquents parmi les lemmes corrigés.

¹⁷ Voir <https://www.cis.uni-muenchen.de/%7Eeschmid/tools/TreeTagger/>.

¹⁸ 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 (<https://dev.chartes.psl.eu/pyrrha/>).

Tableau 70. Comparaison des proportions d'items des quatre catégories grammaticales majeures

Lemmes de :	Nombre de lemmes distincts produits par l'annotation automatique	Nombre de lemmes distincts produits par l'annotation automatique dont au moins une forme graphique a été corrigée
Verbes	5 020 (43,08 %)	1 630 (41,07 %)
Noms communs	3 837 (32,93 %)	1 069 (26,93 %)
Adjectifs	1 368 (11,74 %)	399 (10,06 %)
Noms propres	481 (4,13 %)	579 (14,57 %)
Autres catégories	947 (8,12 %)	294 (7,37 %)
Total	11 653 (100 %)	3 971 (100 %)

La correction manuelle des résultats de l'annotation morphosyntaxique est une tâche longue et fastidieuse, mais une annotation fautive induit du bruit alors que, corrigée, elle participe, comme escompté, à une détection automatique plus fructueuse des correspondances.

Les corrections incitent en outre à observer des ensembles lexicaux. C'est le cas pour les noms propres, nombreux parmi les entrées à corriger, qui montrent que le réseau des personnages est très restreint dans les sources anciennes du mythe et significativement enrichi dans les réécritures postérieures. Ce constat, ainsi que ce que nous avons observé au sein de l'échantillon, nous invite à accorder une place importante à la détection et l'appariement des entités nommées¹⁹.

Difficultés pour la prise en compte optimale de la synonymie et des reformulations polylexicales

La prise en compte des procédés les plus fréquents de reformulation, comme les relations synonymiques, paraît également cruciale pour améliorer la pertinence de l'appariement. Nous voyons par exemple que le couple suivant, établi entre la traduction d'Ovide de Banier1732 et le roman de jeunesse de Montardre2013 ne partage que quatre mots forts communs, quoique fréquents : « sept » (deux occurrences), « jours » et « nuits » (Tableau 71).

Tableau 71. Relation entre Banier1732 et Montardre2013. *Mots communs*

Il demeura <u>sept jours</u> et <u>sept nuits</u> sur les Rives du Fleuve infernal, sans d'autre nourriture que ses larmes, et sa douleur	Il reste là <u>sept jours</u> et <u>sept nuits</u> , sans rien avaler, sans toucher aux aliments que Perséphone, émue par son chagrin, lui a fait porter.

Cependant, « demeurer » et « rester », ainsi que « nourriture » et « aliments » ou encore « douleur » et « chagrin » sont synonymes directs, alors que « douleur » et « émouvoir » appartiennent au même champ lexical. Les relations synonymiques peuvent donc, entre autres, permettre de trancher entre une relation peu pertinente établie sur des expressions très fréquentes et un appariement effectif.

¹⁹ Nous y reviendrons dans la suite de ce travail. Voir IV.2 « Annotation des entités nommées et leurs périphrases ».

Or, les ressources lexicales talistes utilisées pour des traitements comparables (comme *WOLF*²⁰) et les méthodes statistiques (comme le « plongement lexical »²¹) s'avèrent peu opératoires pour ce petit corpus. En effet, les résultats d'une expérimentation avec Word2Vec²² 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, elle ne semble pas pouvoir convenir à nos besoins.

Par conséquent, la possibilité de soumettre à Tracer la liste des synonymes des mots présents dans le corpus et, ainsi, d'élargir ses capacités de détection, paraît dangereuse. L'analyse de l'échantillon révèle déjà plusieurs couples non pertinents liés par trop peu de mots communs. Pour que le traitement gagne en efficacité, cette fonctionnalité doit donc être implémentée avec prudence.

Pour ce faire, il paraît judicieux d'intégrer en premier lieu la prise en compte de la synonymie dans l'analyse des résultats déjà obtenus. Cela permettrait, potentiellement, d'exploiter au maximum les relations détectées entre textes en complétant la prise en compte des identités de mots, de lemmes ou de racines. Nous nous penchons en détail sur cette question dans la partie suivante²³.

Appariement des mots issus des syntagmes différents

Nous avons également constaté que, dans certains couples, les mots communs sont, certes, les mots forts, mais faisant partie des syntagmes aux sens bien distincts. Nous réalisons alors l'importance de l'analyse des syntagmes dans leur entièreté en accordant davantage d'importance à leurs noyaux. Il est alors possible d'écarter des phrases qui n'ont pas ou peu de noyaux en commun, comme celui établi entre Grandmougin1882 et Ballanche1829 :

Tableau 72. Relation entre Grandmougin1882 et Ballanche1829 : appariement des syntagmes

Dans des [lieux inconnus]lieu, j'ai [flotté longuement]flotter comme [un nuage noir]nuage [qui glisse]glisser au [firmament]firmament !	[Sa longue barbe]barbe [descendait rudement]descendre sur [sa poitrine velue]poitrine, [siège de la [force]force]siège ; [sa chevelure terrible]chevelure [flottait]flotter [au gré des [vents]vent]gré ; [ses yeux noirs]œil [lançaient]lancer des [éclairs]éclair.

²⁰ Voir <http://pauillac.inria.fr/~sagot/index.html#wolf> et Benoît Sagot et Darja Fišer, *WordNet Libre du Français (WOLF)*, Institut national de recherche en sciences et technologies du numérique (Inria), 2012.

²¹ Plongement lexical, ou *word embedding*, est une méthode relativement nouvelle du TAL. Il consiste en la représentation des mots d'un texte à l'aide de vecteurs. La distance et la direction entre les vecteurs sont susceptibles de refléter la similarité des mots ou une relation particulière qu'ils entretiennent. Cette méthode a été prouvée probante pour la classification du texte, la reconnaissance automatique des entités nommées ou encore la traduction automatique, mais elle nécessite de disposer d'un corpus de taille importante.

²² Voir Tomas Mikolov et al., *Efficient Estimation of Word Representations in Vector Space*, op. cit.

²³ Voir IV.1 « Post-traitement des résultats de la détection automatique des correspondances ».

Ce couple partage trois entités communes, dont deux lemmes des mots forts (*flotter* et *noir*) et un appariement des mots de la même famille (*long* pour « longuement » et « longue »). Cependant, seulement « flotter » est le noyau, les deux autres entités n'étant que compléments, attributs ou modificateurs des syntagmes aux noyaux qui ne peuvent pas être appariés (« nuage noir » / « yeux noirs », « flotté longuement » / « sa longue barbe »).

La délimitation des syntagmes peut également améliorer la pertinence de l'appariement des entités communes au sein des couples pertinents. Par exemple, nous trouvons une périphrase de Styx commune au roman de jeunesse de Goudot²⁰⁰³ et à la pièce de théâtre d'Espine¹⁶¹⁴ : « Derrière elles, il ne distinguait qu'à grand peine le fleuve noir dont Chiron lui avait parlé » / « D'aise et d'estonnement je sens mon cœur épris/ D'estre pour ce grand Dieu sur ce fleuve noir pris ». Alors que nous y trouvons la même périphrase du Styx (« le fleuve noir »), un autre mot commun, « grand », n'est pas à appairer, car il fait partie des syntagmes dont les noyaux (« peine » et « Dieu ») ne sont pas en relation.

c) Problèmes d'exploration et de visualisation

Du point de vue de l'exploration et de la visualisation des couples de correspondances, nous remarquons principalement trois problèmes. Premièrement, les entités communes ne sont pas mises en valeur, ce qui rend difficile l'évaluation d'une relation. Ensuite, l'accès au contexte, tant de l'œuvre que des autres relations établies, est nécessaire pour pouvoir procéder aux analyses approfondies des relations intertextuelles. Enfin, nous avons également besoin d'un indice permettant de mesurer le degré de similarité, autrement dit d'un score calculé automatiquement, utile pour estimer la force d'une relation.

Manque de mise en valeur des entités communes

Un des problèmes majeurs d'exploration des résultats réside dans le fait que nous ne disposons pas d'informations visuelles quant aux proximités et différences entre segments détectés comme communs. Parmi les sept logiciels testés²⁴, quatre proposent une méthode de marquage des différences ou proximités au sein des couples détectés (Tableau 73). Textreuse procède à l'alignement des extraits en répétant le caractère dièse (« # »), alors que Passim remplace certains caractères différents par « - ». Nous avons déjà constaté²⁵ qu'un tel signalement est difficile à comprendre et devient rapidement illisible pour des relations plus complexes que celles du type 5.

²⁴ Voir III.2 « Expérimentation avec BLAST, Tesserae, Textreuse et Passim » et III.3 « Présentation des logiciels retenus, TextPAIR et Tracer »

²⁵ *Ibid.*

Tableau 73. Différents marquages des différences entre couples proposés par les logiciels testés

Textreuse	Celle-ci était ##### parmi les ombres nouvelles, et ##### elle s'avança ### d'un pas lent par suite de ##### sa blessure.	##### Elle se tenait parmi les ombres ### nouvellement arrivées ; elle ### s'avance d'un pas ##### que ralentit sa blessure.
Passim	Celle-ci était parmi les ombres nouvelles, et elle s'avança d'un pas lent -- de sa blessure.	--lle- --ait parmi les ombres nouvelle----es ; elle s'avan-- d'un pas --lent- sa blessure.
Tracer (interface)	Celle-ci était parmi les ombres nouvelles, et elle s'avança d'un pas lent par suite de sa blessure.	Elle se tenait parmi les ombres nouvellement arrivées ; elle s'avance d'un pas que ralentit sa blessure.
TextPAIR (interface)	~~Celle-ci~~ était parmi les ombres nouvelles, ~~et~~ elle s'avança d'un pas lent ~~par suite de~~ sa blessure.	Elle ~~se~~ tenait parmi les ombres nouvellement arrivées ; elle s'avance d'un pas ~~que~~ ralentit sa blessure.

Tracer et TextPAIR, quant à eux, disposent des interfaces d'exploration qui proposent tantôt la mise en surbrillance des mots identiques, tantôt le marquage des suppressions et des insertions. Cependant, ces interfaces ne sont pas tout à fait fonctionnelles et restent difficilement réutilisables dans le cadre de notre édition²⁶. Par ailleurs, l'interface de Tracer ne met en valeur que les mots identiques (y compris les mots faibles) et omet les correspondances au niveau des lemmes (« s'avança » / « s'avance ») ou de l'identité des familles lexicales (« lent » / « ralentit », « nouvelles » / « nouvellement »). TextPAIR, quant à lui, se concentre sur les différences et non pas sur les similarités entre couples, ce qui risque de poser problème pour les relations complexes, qui proposent par exemple des déplacements ou des reformulations non linéaires.

L'évaluation de l'échantillon des résultats révèle que la majorité des relations détectées est plutôt allusive. Les différences entre deux couples sont parfois très importantes, ce qui implique que la visualisation focalisée sur les variantes (comme celles proposées par Textreuse, Passim et TextPAIR) n'est pas optimale. Nous préférons que ce soit les entités communes qui soient mises en valeur, comme dans l'interface de Tracer, mais nos besoins sont plus complexes que la mise en surbrillance des mots identiques.

Dans la section précédente, nous avons constaté que pour résoudre certains problèmes d'appariement, il serait propice de développer une méthode de post-traitement des résultats qui identifie automatiquement les mots forts communs et précise leur degré de similarité (formes, lemmes ou racines communes). Nous avons également évoqué la nécessité de chercher les proximités au niveau des syntagmes et leurs noyaux dans un premier temps avant d'apparier les mots communs.

L'annotation de ces éléments pourrait contribuer à la conception d'une visualisation qui met en valeur les mots, lemmes et radicaux identiques, ainsi que les synonymes de chaque couple détecté. La délimitation des syntagmes et l'identification de leur noyau permettraient également de vérifier

²⁶ Voir III.3c) « Format ».

rapidement la pertinence de l'alignement des mots isolés. Enfin, il serait aussi pertinent de pouvoir observer facilement la présence des noms propres et leurs périphrases au sein de chaque couple.

Dans le Tableau 74, nous présentons de manière schématique les trois visualisations envisagées. Pour l'appariement des entités à la première ligne, les liens entre mots sont mis en gras, ceux entre lemmes sont soulignés, ceux entre les radicaux sont mis en italiques et les relations synonymiques sont en petites capitales. Les syntagmes, ainsi que les périphrases des noms propres (lignes 2 et 3) sont délimités entre crochets orange. Les noyaux de syntagmes sont indiqués en indice et les référents des périphrases sont en exposant. Nous les mettons en gras s'ils sont identiques au sein du couple et en petites capitales pour les relations synonymiques (« être » / « se tenir »)

Tableau 74. Marquage de différents types de similarités envisageables pour la visualisation des résultats.

Entités	Celle-ci ÉTAIT parmi les ombres nouvelles , et elle s'avança d'un pas lent par suite de sa blessure .	Elle SE TENAIT parmi les ombres nouvellement arrivées ; elle s'avance d'un pas que <i>ralentit sa blessure</i> .
Noyaux de syntagmes	Celle-ci [était] ^{ÊTRE} parmi [les ombres nouvelles] _{ombre} , et elle [s'avança] ^{s'avancer} d'un pas lent] _{pas} par suite de [sa blessure] _{blessure} .	Elle [se tenait] ^{SE TENIR} parmi [les ombres nouvellement arrivées] _{ombre} ; elle [s'avance] ^{s'avancer} d'un pas [que ralentit] _{ralentir} [sa blessure] _{blessure} pas.
Noms propres et périphrases	[Celle-ci] ^{Eurydice} était parmi [les ombres nouvelles] _{Mânes} , et [elle] ^{Eurydice} s'avança d'un pas lent par suite de sa blessure.	[Elle] ^{Eurydice} se tenait parmi [les ombres nouvellement arrivées] _{Mânes} ; [elle] ^{Eurydice} s'avance d'un pas que ralentit sa blessure.

Ces différents types d'annotations permettent d'une part d'aider à l'évaluation de la pertinence d'un couple et de sa force et, d'autre part, de procéder à l'analyse linguistique ou stylistique de la relation, voire à la correction de l'appariement automatique. Nous constatons par exemple que l'identité des radicaux, qui nous amène à apparier « nouvelles » et « nouvellement » gagnerait à être traitée plus globalement, en appariant plutôt tout le syntagme « les ombres nouvelles » avec « les ombres nouvellement arrivées », car elles sont toutes deux périphrases du référent *Mânes*.

Cependant, pour garantir l'expérience de lecture optimale et répondre à différents objectifs de recherche, il convient qu'une telle visualisation soit modulable : nous voulons pouvoir accéder tant aux segments annotés, qu'à l'annotation partielle (des mots, des noyaux ou des entités nommées), voire même à l'annotation complète (des mots, des noyaux et des entités nommées). Par ailleurs, le marquage choisi doit être facilement lisible et intuitif, ce que nous obtenons difficilement avec des affichages statiques, comme celui proposé en Tableau 74.

Accès au contexte trop restreint

Lors de l'évaluation de l'échantillon, nous nous heurtons fréquemment au problème du contexte qui est trop restreint pour pouvoir juger de la pertinence d'une relation. Il serait donc intéressant de pouvoir naviguer facilement entre le texte complet et les segments annotés, par exemple en

proposant une édition numérique où chaque œuvre sera dotée des notes intertextuelles. Dans ce cas-là, le lecteur focaliserait son analyse sur un auteur en particulier et l'extrait de son œuvre serait toujours affiché en premier (à gauche).

En revanche, si nous souhaitons plutôt analyser la totalité (ou une partie) des relations, il paraît nécessaire de respecter l'ordre chronologique de l'apparition de chaque œuvre en affichant en premier (à gauche) celle publiée avant. TextPAIR respecte généralement la chronologie si la date de publication est renseignée au bon endroit du document XML-TEI, mais Tracer ne semble pas prendre en compte les dates de publication que nous ajoutons au fichier regroupant le corpus segmenté. Par conséquent, l'analyse de l'échantillon et la présentation des résultats sont parfois compliquées à cause des appariements non chronologiques (le texte postérieur apparaît comme source d'un texte antérieur). C'est important par exemple pour les couples du type 5 où un auteur est susceptible d'être directement inspiré par un autre. Pour faciliter l'interprétation, nous avons ainsi dû remanier l'ordre des couples dans le Tableau 65 et le Tableau 67. Il serait préférable de pouvoir ordonner les résultats automatiquement (par exemple grâce aux filtres) soit selon l'auteur, soit selon la date de publication.

Nous voyons également que certains segments sont susceptibles d'être alignés avec plusieurs autres et former ainsi un ensemble interprétatif intéressant. Prenons comme exemple un extrait d'Esquiros1848 qui transpose le mythe à l'ancienne Égypte tout en reprenant presque à l'identique Virgile²⁷ (Tableau 75). La variante substantielle consiste ici en un remplacement du nom propre *Manes* (traduit par « les Mânes » ou « l'enfer » dans le corpus) par « les prêtres ». Le couple avec la traduction de Charpentier1831 présent dans l'échantillon et mentionné *supra* (voir Tableau 67) s'avère être un des six couples détectés par TextPAIR avec les traductions de Virgile, dont trois antérieures et trois postérieures à l'œuvre d'Esquiros1849.

Tableau 75. Correspondances détectées entre un extrait d'Esquiros1848 et six traductions de Virgile

« Les deux amants avaient repassé ces lieux obscurcis par les ténèbres et par l'horreur de la mort ; ils s'acheminaient vers la lumière ; quand je ne sais quelle démence saisit le trop faible amoureux : <u>faute bien pardonnable sans doute, si les prêtres savaient pardonner !</u> Arrivé presque à l'embouchure du puits, au moment où une demi-clarté annonçait un retour prochain au monde extérieur, il s'arrêta. » (Esquiros1849)	« Rendue à son amour, sa chère Eurydice suivait ses pas ; telle était la condition imposée par Proserpine ! Soudain, oubliant la loi fatale, vaincu par son amour, égaré par son délire (<u>faute hélas ! bien pardonnable, si l'enfer savait pardonner !</u>) il se retourne. » (Héguin1827) « il revenait vainqueur de tous les obstacles ; rendue à son amour, Eurydice remontait au séjour de la lumière : elle suivait son époux ; ainsi l'avait ordonné Proserpine. Tout à coup sa tendresse imprudente le trahit : <u>faute bien pardonnable, si l'enfer savait pardonner ;</u> il s'arrête, et déjà aux portes du jour, oubliant sa promesse et vaincu de son amour, il se retourne : » (Charpentier1831)

²⁷ « Jamque, pedem referens, casus evaserat omnes, / Redditaque Eurydice superas veniebat ad auras, / Pone sequens (namque hanc dederat Proserpina legem), / Quum subita incautum dementia cepit amantem, / Ignoscenda quidem, scirent si ignoscere Manes ! / Restitit, Eurydicenque suam jam luce sub ipsa / Immemor, heu ! victusque animi respexit » (*Georgica*, livre IV, l. 485-491, nous soulignons).

« Les deux amants avaient repassé ces lieux obscurcis par les ténèbres et par l'horreur de la mort ; ils s'acheminaient vers la lumière ; quand je ne sais quelle démence saisit le trop faible amoureux : <u>faute bien pardonnable sans doute, si les prêtres savaient pardonner !</u> Arrivé presque à l'embouchure du puits, au moment où une demi-clarté annonçait un retour prochain au monde extérieur, il s'arrêta. » (Esquiros1849)	« Eurydice, rendue à son amour, suivait ses pas (car telle était la loi imposée par Proserpine), quand tout à coup, oubliant la loi fatale, vaincu par son amour, égaré par son délire (<u>faute, hélas ! bien pardonnable, si l'enfer savait pardonner !</u>), il se retourne, il regarde son Eurydice. » (Desportes1846-2) « rendue, marchait vers les régions de la lumière, le suivant sans qu'il la vit ; Proserpine ne la lui rendait qu'à ce prix. Mais, <u>ô délire soudain d'un amant insensé, et bien digne de pardon, si l'enfer savait pardonner !</u> il s'arrête, et presque aux portes du jour, s'oubliant lui-même, hélas ! et vaincu par l'amour, il regarde son Eurydice. » (Nisard1868) « Eurydice, rendue à ses vœux, remontait au séjour de la lumière en suivant son époux (ainsi l'avait ordonné Proserpine), quand un délire soudain s'empara de l'aveugle amant : <u>faute bien pardonnable, si l'enfer savait pardonner.</u> Presque aux portes du jour, il s'arrête, hélas ! il oublie sa promesse, et, vaincu par l'amour, il tourne la tête pour regarder sa chère Eurydice. » (Cabaret-Dupaty1897) « Eurydice lui étant rendue s'en venait aux souffles d'en haut en marchant derrière son mari (car telle était la loi fixée par Proserpine), quand un accès de démence subite s'empara de l'imprudent amant ; <u>démence bien pardonnable, si les Mânes savaient pardonner !</u> Il s'arrêta, et juste au moment où son Eurydice arrivait à la lumière, oubliant tout, hélas ! et vaincu dans son âme, il se tourna pour la regarder. » (Rat1932)

Si nous nous intéressons uniquement au segment souligné « faute bien pardonnable sans doute, si les prêtres savaient pardonner ! », les variantes peuvent être schématisées comme dans le Tableau 76. Nous pouvons isoler trois éléments du segment de la réécriture que nous retrouvons dans toutes les traductions appariées : (i) « faute », présente dans quatre traductions sur six (deux autres proposent des synonymes : « délire » et « démence »), (ii) « bien pardonnable », identique dans tous les extraits correspondants sauf un, qui propose une reformulation flexionnelle « bien digne de pardon » et (iii) « savaient pardonner », identique dans la traduction de Rat1932 et qui correspond au niveau des lemmes pour toutes les autres (pluriel/singulier). La différence principale est donc le remplacement de l'entité *Enfer* (ou *Mânes*) par « les prêtres » et l'ajout de l'adverbe « sans doute » (alors que dans certaines traductions nous trouvons plutôt l'exclamation « hélas ! »).

Tableau 76. Schématisation des proximités entre le segment central des extraits appariés

« [faute] ¹ [bien pardonnable] ² sans doute, si les prêtres [savaient pardonner] ³ ! » (Esquiros1849)	« (faute ¹ hélas ! bien pardonnable ² , si l'enfer <u>savait pardonner</u> ³ !) » (Héguin1827)
	« faute ¹ bien pardonnable ² , si l'enfer <u>savait pardonner</u> ³ » (Charpentier1831)
	« (faute ¹ , hélas ! bien pardonnable ² , si l'enfer <u>savait pardonner</u> ³ !) » (Desportes1846-2)
	« ô DÉLIRE ¹ soudain d'un amant insensé, et <i>bien digne de pardon</i> ² , si l'enfer <u>savait pardonner</u> ³ ! » (Nisard1868)
	« faute ¹ bien pardonnable ² , si l'enfer <u>savait pardonner</u> ³ » (Cabaret-Dupaty1897)
	« DÉMENCE ¹ bien pardonnable ² , si les Mânes <u>savaient pardonner</u> ³ ! » (Rat1932)

Nous constatons qu'Esquiros1849 ne s'est pas servi d'une des traductions du corpus. Trois hypothèses sont plausibles : utilisation directe du texte latin de Virgile, utilisation d'une traduction dont nous ne disposons pas, ou encore combinaison de plusieurs traductions présentes dans le corpus.

Par ailleurs, TextPAIR propose d'apparier ce segment avec une autre réécriture, l'épopée de Ballanche1829 : « j'ai voulu te distraire de tes grandes pensées, au lieu de me borner à les embellir, et à leur donner du charme ; et voilà que les dieux m'en punissent. Le châtement n'a pas tardé de suivre la faute ; et la faute peut-être était bien pardonnable ». Nous y retrouvons la même notion de « la faute bien pardonnable » qui fait preuve davantage d'une source commune (texte de Virgile) que de l'inspiration d'Esquiros1849 par Ballanche1829. Il serait donc intéressant d'examiner à leur tour les liens entre cet extrait de Ballanche1829 et les traductions de Virgile, mais une telle itération dans un fichier tabulé est extrêmement chronophage et difficile, surtout pour les résultats de TextPAIR au sein desquels les segments ne disposent pas d'un identifiant unique, comme c'est le cas pour Tracer.

L'extrait d'Esquiros1849 est également très présent dans les résultats obtenus avec ce logiciel. Il est moins étendu que pour TextPAIR, car il se limite à la phrase 437²⁸. Le segment a été apparié à dix autres textes, dont trois réécritures et sept traductions de Virgile. Les résultats pour les traductions sont très proches de ce que nous avons trouvé avec TextPAIR : la traduction de Nisard1868, plutôt poétique, est remplacée par celle, plus littérale, mais aussi plus ancienne de Desfontaines1810²⁹

²⁸ « Déjà les deux amants avaient repassé ces lieux obscurcis par les ténèbres et par l'horreur de la mort ; ils s'acheminaient vers la lumière ; quand je ne sais quelle démence saisit le trop faible amoureux : faute bien pardonnable sans doute, si les prêtres savaient pardonner ! »

²⁹ « Cependant un mouvement subit, dont il ne fut point le maître, lui fit oublier la loi : faute pardonnable, si les enfers savaient pardonner ! il s'arrêta, et lors qu'il étoit sur le point de revoir la lumière, vaincu par son ardeur, il voulut voir sa chère Eurydice. »

et une relation supplémentaire est trouvée avec la version juxta-linéaire de Desportes1846-1³⁰. Les relations avec les réécritures sont, en revanche, plus riches, mais aussi plus morcelées. Ainsi, trois segments correspondants sont proposés avec le texte de Ballanche1829, deux avec Duboullay1690 (il s'agit de la répétition de la même phrase) et un avec Ronsard1563 (Tableau 77).

Tableau 77. Relations supplémentaires entre Esquiros1849 et autres réécritures détectées par Tracer

« Déjà les deux amants avaient repassé ces lieux obscurcis par les ténèbres et par l'horreur de la mort ; ils s'acheminaient vers la lumière ; quand je ne sais quelle démence saisit le trop faible amoureux : faute bien pardonnable sans doute, si les prêtres savaient pardonner ! » (Esquiros1849, ph. 437)	Ballanche1829 : « Ce voyage silencieux, parmi de muettes ténèbres, devenait trop long pour le faible époux. », ph. 210 ; « Le châtement n'a pas tardé de suivre la faute ; et la faute peut-être était bien pardonnable. », ph. 1 039 ; « C'est à ce point sans doute que l'on découvrirait ce que fut la faute, ce que furent la promesse et le pardon. », ph. 3 275
	« Helas, fût-il jamais faute plus pardonnable, si l'Enfer sçavoit pardonner ? » (Duboullay1690, ph. 341 et 342)
	« Un sentier est là-bas tout obscur et tout sombre, entremêlé de peur et de frayeur et d'ombre : par ce chemin je sors, et ja presque j'avois passé le port d'enfer, les ruës et le bois, quand, las ! vaincu d'amour, je regarde en arrière, et mal-caut je jettay sur elle ma lumière, faute assez pardonnable en amour, si Pluton sçavoit hélas ! que c'est que de faire pardon. » (Ronsard1563, ph. 62).

Pour les deux derniers textes, Tracer détecte effectivement le renvoi au même épisode du mythe, même malgré une reformulation assez conséquente chez Ronsard1563 (« faute assez pardonnable en amour, si Pluton sçavoit hélas ! que c'est que de faire pardon. »). Les relations détectées avec Ballanche1829 sont plus complexes. La phrase 1 039 fait partie de l'extrait détecté aussi par TextPAIR et cité *supra*. La phrase 3 275 présente des similarités plus fragmentaires (« sans doute » et « promesse » font partie des syntagmes éloignés), la pertinence de la relation peut donc être mise en cause sans accéder au contexte plus large. En revanche, la phrase 210 demande à ce que nous nous intéressions à son contexte, ce qui nous permet de réaliser qu'il s'agit effectivement de l'épisode mythique de la sortie des Enfers :

Eurydice suivait son époux ; elle le suivait à pas furtifs [...]. Ce voyage silencieux, parmi de muettes ténèbres, devenait trop long pour le faible époux. Non, il n'accomplira pas sa promesse, et le roi des épouvantes sera inexorable à tenir la sienne. Cependant le couple divin s'avancait toujours dans la route de mystère et d'effroi. Un moment de plus, et Orphée sauvait sa conquête, et il allait être affranchi de la loi cruelle qui lui fut imposée. Mais il ne peut supporter le poids de ce moment trop rempli de délices. Déjà un crépuscule douteux commençait à remplacer l'obscurité immense ; un rayon de pâle lumière, qui vint flotter autour de ses yeux, lui fit oublier le passé et l'avenir : ce n'était pas une lueur trompeuse, c'était bien un rayon détaché de la clarté du jour ; la marque

³⁰ « Et déjà portant en arrière ses pieds (pas) il avait échappé à tous les hasards, et Eurydice lui étant rendue venait aux airs (régions) d'en-haut, le suivant par derrière (car Proserpine avait donné (imposé) cette loi), lorsqu'une démence soudaine s'empara de l'amant imprudent, démence pardonnable à la vérité si les Mânes savaient pardonner. »

de son salut devint le signe de sa perte. Ainsi donc il touchait au seuil qui sépare l'empire des morts du séjour des vivants, lorsqu'il s'arrêta involontairement pour jeter un regard d'amour sur l'épouse qui lui était rendue, hélas ! rendue et ravie à-la-fois. (Ballanche1829, ph. 207-216, nous soulignons).

Alors que, cette fois-ci, le syntagme « faute bien pardonnable » n'est pas présent, l'épisode, même fortement poétisé, est tout de même facilement identifiable. La similarité avec Esquiros1849 est établie sur des entités communes particulières, notamment la périphrase d'Orphée (« faible amoureux » / « faible époux ») et la description du paysage (« ces lieux obscurcis par les ténèbres » / « ces muettes ténèbres »). Cette double détection implique que la seconde mort d'Eurydice est décrite plusieurs fois chez Ballanche1829. Nous y reviendrons³¹.

Cette investigation prouve la nécessité du regroupement de toutes les correspondances détectées pour un segment, tant par Tracer que par TextPAIR, qu'elles soient directes ou indirectes, car les segments liés au même texte sont aussi à apparier, même si les logiciels n'ont pas su détecter cette relation. Voyons par exemple la visualisation à l'aide de graphe d'un réseau dont le segment 437 d'Esquiros1849 est central (Figure 45).

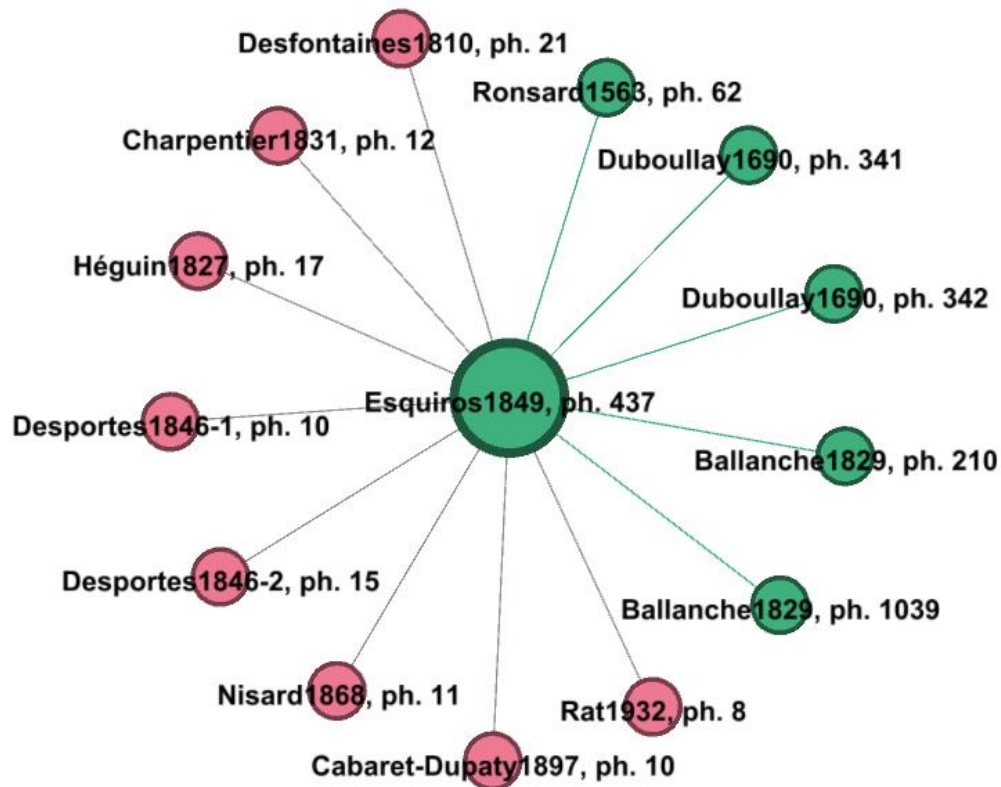


Figure 45. Graphe des relations détectées par Tracer et TextPAIR pour la phrase 437 d'Esquiros1849

³¹ Voir IV.1c) « Exploitation des résultats enrichis et du réseau global des réécritures ».

Dans cette visualisation, les nœuds rouges symbolisent les traductions d'Ovide et les nœuds verts les réécritures. Le nœud représentant la phrase 437 d'Esquiros1849 se situe au centre du graphe. Il est possible de relier indirectement les réécritures correspondantes à Esquiros1849 avec les traductions d'Ovide, voire relier les réécritures entre elles de la même manière. C'est ce que nous appelons les relations indirectes.

Si nous élargissons les relations visualisées à celles des trois autres réécritures (deux phrases de Ballanche1829, deux phrases de Duboullay1690 et une phrase de Ronsard1563), nous obtenons un graphe bien plus complexe (Figure 46). Ce graphe permet d'enrichir davantage nos observations. Nous réalisons par exemple que les deux segments de Duboullay1690 et la phrase 1 039 de Ballanche1829 entretiennent quasiment les mêmes relations que la phrase 437 d'Esquiros1849 en se partageant une relation supplémentaire avec une autre traduction de Virgile. Il s'agit de la phrase 24 de Cogolin1750, « Lorsqu'en proie aux transports du desir qui le presse, par un excès d'amour parjure à sa promesse, il s'arrête, et tourné vers celle qui le fuit regarde en soupirant l'objet qui le séduit, faute aux yeux de l'amour pardonnable et légère. ».



Figure 46. Graphe des relations détectées par Tracer et TextPAIR élargi des relations détectées pour les autres réécritures

En revanche, le texte de Ronsard1563 n'est apparié à aucune traduction, mais uniquement aux réécritures. Ces dernières peuvent être partagées en trois sous-groupes : (i) les textes de genres différents, mais relativement proches chronologiquement du poème de Ronsard d'une part (les tragédies de l'Espine1614³² et de Duboullay1690, le roman de Puget1624³³ et le poème d'Hermite1641³⁴), (ii) les romans symboliques du XIX^e s. (Ballanche1829 et Esquiros1849) et (iii) les œuvres de jeunesse (l'album de Pommaux2009³⁵, le roman de Laffon2016³⁶ et la bande dessinée de Ferry2019³⁷). La relation avec Esquiros1849 (ainsi qu'avec Ballanche1829 et Duboullay1829) permet donc de relier cet extrait de Ronsard1563 à l'épisode correspondant du mythe grâce aux correspondances indirectes avec les traductions d'Ovide.

Le dernier besoin de contextualisation que nous identifions consiste en la possibilité de localiser le placement d'un segment (ou d'une œuvre) dans la totalité des relations détectées. Nous pouvons par exemple nous pencher plus en détail sur la relation entre Esquiros1849 et Ballanche1829. Rien que pour le segment 437 d'Esquiros1849, trois correspondances avec Ballanche1829 ont été proposées. En même temps, ce dernier figure parmi les trois œuvres les plus présentes parmi les résultats. Il est donc intéressant de savoir combien de relations sont établies entre ces deux textes et l'importance de ces proportions par rapport à la totalité des relations détectées pour chacun des deux œuvres, voire pour la totalité des réécritures du corpus.

Voyons par exemple les deux graphes qui représentent le réseau établi entre le texte d'Esquiros1848 et 15 œuvres les plus fréquemment liés avec lui par Tracer (Figure 47) et TextPAIR (Figure 48). Cette fois-ci, nous utilisons les codes-couleurs pour différencier les transformations du mythe. Ainsi, les nœuds représentant les traductions sont en vert, les réécritures idéologiques en rose, ceux pragmatiques (pour la jeunesse) en jaune, les modernisations en bleu et, enfin, les transmodalisations en orange.

Les résultats de Tracer sont bien plus nombreux que ceux de TextPAIR (670 *vs* 156), mais les deux présentent beaucoup de similarités, notamment en ce qui concerne le nombre de réécritures symboliques et la force de la relation avec Ballanche1829 (respectivement 42,2 % et 35,9 % des résultats). Cependant, alors que les deux logiciels relient l'œuvre d'Esquiros1848 avec des traductions, seulement trois d'entre elles figurent sur le graphe de Tracer. Deux sont de Virgile et une d'Ovide (Duryer1702). Cinq traductions sont sur le graphe de TextPAIR, dont deux d'Ovide (celle de Villenave1806 en plus de celle de Duryer1702). Par ailleurs, Tracer détecte plus de relations avec les œuvres

³² « Helas ! je suis contant, revoyant ses beaux yeux, de ne revoir jamais la lumiere des Cieux. »

³³ « Ô faute sans remède ! ô dommageable veuë ! avec trop de travaux tu m'avois obtenuë : »

³⁴ « Tu as donc terminé ta carriere en ton Orient, bel Astre de ma vie, et as estaint pour jamais la lumiere de tes yeux. »

³⁵ « Là-bas sur le chemin, entre deux rochers se profile la terrifiante silhouette du chien Cerbère, gardien des Enfers. »

³⁶ « Dans la caverne obscure où il avance sans repères, Orphée reprend le chemin qui monte vers la lumière. »

³⁷ « Je... j'ai peur de t'avoir perdue en chemin ! »

de jeunesse (Laffon2016, Montardre2013 et Jimenes2004) et les modernisations (Anouilh1941, en plus de celui de Crémieux1858 présent sur les deux graphes), alors que TextPAIR trouve également 5 couples avec la transmodalisation d’Espine1614.

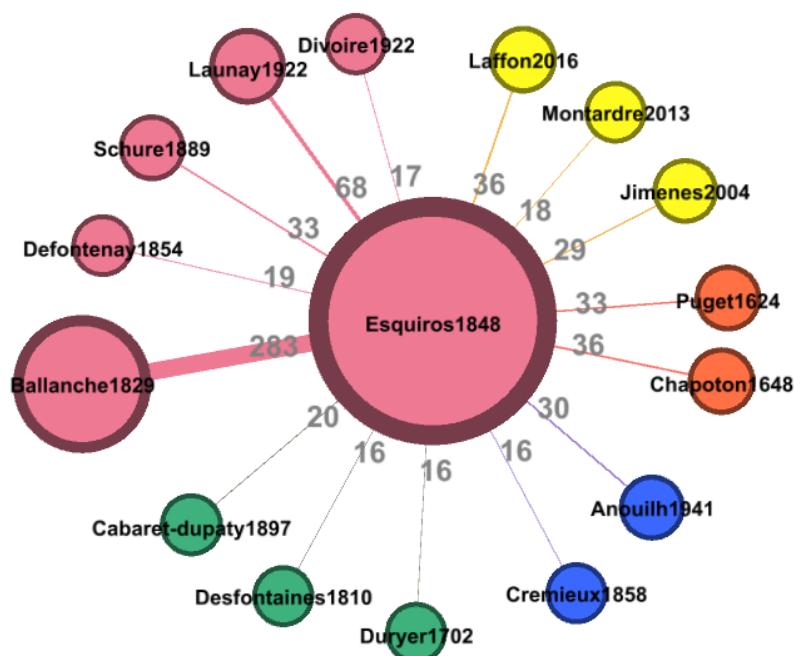


Figure 47. Graphe des relations détectées entre le texte d’Esquiros1848 et les autres œuvres du corpus avec Tracer

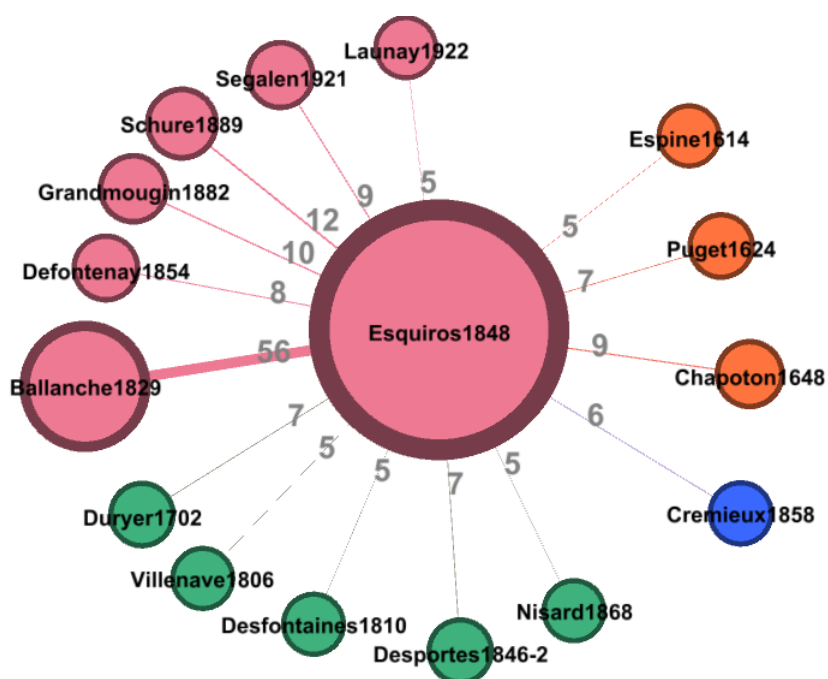


Figure 48. Graphe des relations détectées entre le texte d’Esquiros1848 et les autres œuvres du corpus avec TextPAIR

Comme nous utilisons les résultats bruts, la pertinence effective de ces relations n’est pas confirmée. Cependant, cette visualisation manuelle permet déjà de constater la proximité des œuvres proposant la même transformation du mythe (transformations symboliques) et d’observer la force de la relation entre Esquiros1848 et Ballanche1829. Par ailleurs, la spécificité des relations

établies avec les traductions prouve que le mythe est bien identifiable au sein de la réécriture d'Esquiro1848, mais de manière fragmentaire ou allusive. Les relations détectées avec les traductions les plus représentées sur les deux graphes, à savoir celles de Duryer1702, Desfontaines1810 et Cabaret-Dupaty1897, méritent, quant à elles, une analyse approfondie.

En outre, il serait intéressant de savoir si les relations détectées avec les traductions ont été condensées autour d'un fragment en particulier ou si elles sont plutôt dispersées. Cela est possible grâce à l'identification des segments exigée par Tracer. Si nous trions les résultats pour n'obtenir que les couples de correspondances entre les segments d'Esquiro1848 et ceux des traductions d'Ovide et de Virgile, nous obtenons 283 couples repartis entre 99 phrases de la réécriture. Quinze phrases établissent au moins 5 relations avec les traductions différentes, la première est la phrase 11 (7 relations) et la dernière, la phrase 566. La réécriture est donc entièrement consacrée au mythe, car elle se compose de 586 phrases au total, mais elle amplifie de manière importante la trame originale.

Dans le Tableau 78, nous présentons les 5 phrases d'Esquiro1848 appariées à plus de 10 segments des traductions. Les différents épisodes du mythe y sont clairement identifiables et la réécriture paraît minimale, sauf pour la phrase 566 qui introduit de nouveaux lieux géographiques (Scythie, Tanaïs et monts Riphées) et la phrase 82 dans laquelle Orphée est muni d'une lampe en plus de sa lyre pour descendre « vers le noir séjour des ombres ».

Tableau 78. Cinq segments d'Esquiro1848 appariés à plus de dix traductions

Ph. 380	19 rel.	« Ils ne mettent à cette faveur extraordinaire qu'une condition, c'est qu'Eurydice vous suivra par derrière et que vous ne retournerez point la tête pour la regarder. »
Ph. 486	14 rel.	« Cependant, Orphée ne pouvait ni vaincre ni adoucir son tourment ; il se retira de nouveau en Thrace, sur les bords déserts du Strymon. »
Ph. 566	14 rel.	« Orphée errait seul dans les glaces de la Scythie, le long des bords du Tanaïs, et dans les campagnes voisines des monts Riphées éternellement couverts de neige. »
Ph. 62	13 rel.	« C'est toi, Eurydice ! qu'il appelait au lever du jour : c'est encore toi ! qu'il chantait au tomber de la nuit. »
Ph. 82	12 rel.	« Orphée prit donc une lampe avec sa lyre qu'il laissait depuis longtemps oisive, et s'achemina vers le noir séjour des ombres. »

Nous observons très peu d'appariements significatifs entre la phrase 82 et la phrase 380, sauf la phrase 137 (10 rel.) « Vous voyez d'ici l'origine de votre Cerbère, ce monstrueux gardien du Tartare, sur lequel le poète Virgile m'a lu, dans le temps, de très beaux vers latins. ». Elle est intéressante, car Virgile y est mentionné en tant qu'interlocuteur du narrateur, un docteur mystérieux qui raconte l'histoire d'Orphée et Eurydice lors d'une soirée du Club des Exaltés. L'inspiration par *Les Géorgiques* est donc affirmée et il est intéressant de constater que la détection automatique des correspondances permet de la révéler.

Nous observons aussi que 300 phrases séparent la descente aux Enfers et les retrouvailles avec Eurydice. Esquiros¹⁸⁴⁸ s'éloigne de la trame principale du mythe en remplaçant les Enfers par un temple égyptien et en décrivant les rites des prêtres qui y vivent. L'histoire se rapproche à nouveau du mythe pour décrire la deuxième perte d'Eurydice. Nous pouvons l'observer grâce à l'agglomération des phrases successives reliées avec des traductions entre la phrase 437 et la phrase 453 pour lesquelles 58 relations au total ont été détectées (Tableau 79).

Tableau 79. Deux extraits contenant des segments successifs appariés avec des traductions

Ph. 437-440	26 rel.	« [Déjà les deux amants avaient repassé ces lieux obscurcis par les ténèbres et par l'horreur de la mort ; ils s'acheminaient vers la lumière ; quand je ne sais quelle démence saisit le trop faible amoureux : faute bien pardonnable sans doute, si les prêtres savaient pardonner !] ⁸ [Arrivé presque à l'embouchure du puits, au moment où une demi-clarté annonçait un retour prochain au monde extérieur, il s'arrêta.] ⁰ [Puis, soudain, oublieux des ordres qu'on lui avait donnés, vaincu par ses désirs, impatient de la voir, il se retourne.] ¹⁰ [– Ici, tout s'évanouit, ses travaux, ses efforts ; le pacte juré se rompt, et l'abîme tressaille par trois fois, avec un bruit épouvantable.] ⁸ »
Ph. 441-444	0 rel.	« [Un moment d'oubli avait détruit son bonheur.] ⁰ [Alors Eurydice, tournant vers lui des yeux mourants :] ⁰ [“Quel délire nous a perdu tous deux, cher Orphée ?”] ⁰ [Je te revois, mais pour jamais.] ⁰ »
Ph 445-453	32 rel.	« [Les cruels destins me rappellent au fond de l'abîme.] ⁷ [L'éternelle nuit presse mes yeux noyés par le sommeil.] ⁰ [Adieu. Je suis emportée malgré moi dans les épaisses ténèbres, et je ne puis que te tendre mes mains impuissantes – ces mains qui ne sont déjà plus celles de ton épouse.”] ⁵ [Elle dit ; en vain Orphée veut se jeter sur elle pour la retenir.] ¹ [Il ne la voit plus.] ⁰ [Un horrible artifice l'avait soustraite à ses yeux.] ⁰ [Éploré, il demande à repasser le marais odieux, pour rejoindre Eurydice dans le fond de ces demeures souterraines.] ⁴ [Vaine prière ! l'impitoyable nocher lui défend d'approcher de ces bords.] ⁶ [Pour comble de douleur, il revoit alors son Eurydice qui voguait pâle et froide sur la fatale barque.] ⁹ »

Sur 18 phrases qui composent cet extrait, presque la moitié n'a aucune relation avec une traduction dans les résultats de Tracer, probablement parce que certaines contiennent des reformulations et métaphorisations, comme « mes yeux noyés par le sommeil », « un horrible artifice », « embouchure du puits », etc., alors que d'autres sont très courtes, voire ne contiennent que les mots faibles : « Je te revois, mais pour jamais », « Il ne la voit plus ». Nous les retrouvons cependant parmi les résultats de TextPAIR, par exemple dans une relation avec la traduction de Charpentier¹⁸³¹ et une autre avec celle de Delille¹⁷⁷⁰ (Tableau 80), ce qui prouve la pertinence du regroupement des résultats des deux logiciels, qui s'avèrent souvent complémentaires.

Tableau 80. Résultats complémentaires de TextPAIR

« Quel délire nous a perdu tous deux, cher Orphée ? Je te revois, mais pour jamais. Les cruels destins me rappellent au fond de l'abîme. L'éternelle nuit presse mes yeux noyés par le sommeil. Adieu. Je suis emportée malgré moi dans les épaisses ténèbres, et je ne puis que te tendre mes mains impuissantes – ces mains qui ne sont déjà plus celles de ton épouse. » (Esquiros1848)	« Elle : Quel dieu, quel délire nous a perdus tous deux ? Voici que de nouveau m'entraînent les destins cruels ; le sommeil ferme mes yeux éteints pour jamais. Adieu ! la mort m'enveloppe de ses ombres ; vainement j'étends vers toi mes faibles bras ! Hélas ! ton Eurydice n'est plus ! » (Charpentier1831) « Orphée, ah cher époux ! quel transport malheureux, / Dit-elle ! ton amour nous a perdus tous deux. / Adieu ; l'Enfer se rouvre, et mes yeux s'obscurcissent ; / Mes bras tendus vers toi déjà s'appesantissent ; / Et la Mort déployant son ombre autour de moi / M'entraîne loin du jour, hélas ! et loin de toi. » (Delille1770)

Ainsi, nous voyons que la possibilité d'explorer les résultats de plusieurs points de vue mène à des observations stimulantes, ce qui prouve la nécessité de concevoir des visualisations modulables et personnalisables. Celles-ci peuvent aider à localiser un segment intéressant dans le contexte plus large de son œuvre ou de la totalité des résultats, ou encore à analyser l'ensemble des correspondances qu'une réécriture entretient avec les autres œuvres du corpus. Comme nous l'avons vu avec l'exemple d'Esquiros1848, les résultats obtenus automatiquement avec TextPAIR et Tracer nécessitent d'être améliorés, mais ils se montrent déjà cohérents et pertinents pour l'analyse de la manière dont le mythe est réécrit, à condition de savoir les remanier, les filtrer et les structurer.

Pour cette expérimentation, nous avons conçu nos visualisations en graphes et regroupé les données nécessaires manuellement ou en explorant les résultats en format tabulé (tableaux croisés dynamiques, etc.). Tout cela est chronophage et répétitif, en plus de créer le risque d'erreur et d'omission dû au traitement (semi-)manuel. Il est donc souhaitable que ces manipulations soient faites automatiquement pour chaque œuvre du corpus pour que l'utilisateur puisse s'adonner entièrement à leur exploration et analyse.

Manque de mesure du degré de similarité

Enfin, la dernière difficulté que nous rencontrons consiste en la mesure fiable du degré de similarité. Alors qu'il n'est pas envisageable d'évaluer manuellement la totalité du corpus, un calcul de similarité permettrait potentiellement d'écarter les couples les moins pertinents et de mettre en avant ceux les plus proches.

Alors que TextPAIR ne fournit pas d'information concernant les couples détectées, Tracer indique le nombre des mots communs et un score de similarité. Néanmoins, nous réalisons que ces deux informations ont très peu de concordance avec notre évaluation manuelle (Tableau 81). En effet, le score moyen des couples du type 4 est moins élevé que celui des couples du type 2 et 3 et la différence entre les couples du type 2 et 5 est peu significative. Le nombre de mots communs

est, quant à lui, peu informatif, car presque la totalité des couples de l'échantillon Tracer n'aurait que deux mots en commun.

Tableau 81. Concordance entre l'évaluation manuelle et le score automatique généré par Tracer

	Type 1	Type 2	Type 3	Type 4	Type 5
	Score moyen de similarité entre couples				
	0,375	0,444	0,403	0,388	0,4590
	Nombre de mots communs				
2 mots	2	120	110	177	74
3 mots	0	3	3	3	5
4 mots	0	0	0	0	2
5 mots	0	0	0	0	1

L'équation permettant de calculer le score Tracer est donnée dans le manuel du logiciel³⁸, mais son explication est assez sommaire, ce qui rend notre interprétation incertaine³⁹ :

$$\theta_{\theta}^{R'}(s_i, s_j) = \theta_{\theta}^O(s_i, s_j) \cdot \theta_{\theta}^R(s_i, s_j) = \frac{|\vec{s}_i \cap \vec{s}_j|}{|\vec{s}_i \cup \vec{s}_j|}$$

Nous pouvons tout de même constater que ce calcul est en partie basé sur l'indice de Jaccard. C'est une mesure de similarité qui donne le rapport de la taille de l'intersection des deux ensembles et la taille de leur union en divisant le nombre d'éléments communs aux deux ensembles par le nombre d'éléments distincts⁴⁰. Le résultat est compris entre 0 (deux ensembles qui n'ont rien en commun) et 1 (deux ensembles sont identiques). Le coefficient de Jaccard est insensible aux doublons, car il ne se base pas sur la fréquence des éléments, mais seulement leur présence ou absence. C'est pour cela que la formule de calcul de score implémenté dans Tracer semble ajouter un facteur de pondération ($\theta_{\theta}^O(s_i, s_j)$) qui dépend des paramétrages choisis : l'utilisateur peut par exemple accorder plus de poids à certaines catégories grammaticales (verbes, noms, adjectifs, etc.).

Or, en observant les 33 couples ayant obtenu le score maximal de 1, nous nous rendons compte que le calcul présenté dans le manuel n'est pas forcément celui que nous obtenons dans les résultats. Par exemple, le couple « Eurydice n'est plus ! » / « L'as-tu oubliée aussi ?... je vais être vengé... ce n'est pas à moi que tu rendras Eurydice, c'est à lui !... au petit *trovatore*... » obtient le score Tracer de 1, alors que son indice de Jaccard est de 0,167. Après la suppression des mots faibles, la phrase source compte un seul mot, « Eurydice », et la phrase cible en compte six (« oublier », « venger », « rendre », « Eurydice », « petit », « trovatore »), comme « Eurydice » est le mot commun, nous divisons 1 par 6 pour obtenir le coefficient de 0,167. Il semblerait donc que Tracer calcule plutôt le

³⁸ Voir <https://gfranzini.gitbooks.io/tracer/content/manual/configuration/step-5.-scoring.html/>

³⁹ « The resemblance score θ (theta) is the quotient of the reuse overlap ($\cap$) and the reuse union ($\cup$), where the reuse overlap represents the digital fingerprint of reuse units that appear in two candidate strings/lines, and the reuse union represents the number of all digital fingerprints ».

⁴⁰ Pour deux ensembles A et B la formule de la similarité de Jaccard est la suivante : $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$

pourcentage de similarité des mots forts de la phrase source par rapport aux mots forts de la phrase cible : comme « Eurydice » est le seul mot fort, la similarité est de 1. En suivant la logique de Tracer, ce n'est que si nous considérons la phrase cible comme source que nous arriverons au résultat de 0,167. Cela n'est pas satisfaisant, car les scores sont artificiellement élevés notamment pour les couples composés d'une phrase courte et d'une phrase longue.

D'autres calculs de similarité existent et peuvent s'avérer plus pertinents pour notre corpus. Dans l'article « On the Feasibility of Automated Detection of Allusive Text Reuse », Manjavacas *et al.*⁴¹ proposent trois catégories de calculs de similarité : lexicaux, sémantiques et hybrides. Parmi la première catégorie, nous trouvons le calcul de score utilisé dans le logiciel Tesseract⁴² et destiné à détecter des cas de réutilisation allusive, contenant peu de correspondances lexicales directes. Il est défini de manière suivante :

$$T(s, t) = \ln \left(\frac{\sum_{\omega \in (S \cap T)} \left(\frac{1}{f(\omega, s)} + \frac{1}{f(\omega, t)} \right)}{d_s + d_t} \right)$$

S et T représentent les ensembles de mots présents respectivement dans les documents s et t ; $f(\omega, s)$ désigne la fréquence du mot ω dans le document s ; d_s et d_t correspondent à la distance, en nombre d'unités lexicales, entre les deux mots les moins fréquents dans les documents s et t respectivement. Ainsi, les mots rares partagés entre deux textes sont considérés comme indices d'une réutilisation allusive, même si deux documents partagent très peu d'unités communes. Le logarithme naturel ($\ln$) est appliqué pour atténuer l'impact des valeurs extrêmes et pour normaliser la distribution des scores. Les deux documents doivent partager au moins deux mots pour que le dénominateur puisse être calculé.

Ce calcul s'est montré probant pour de grands corpus d'auteurs, car le fait de donner plus d'importance aux mots que l'auteur n'a pas l'habitude d'employer peut constituer un indice d'emprunt d'un autre auteur. En revanche, dans le cadre de notre recherche, les mots les plus fréquents font régulièrement partie des mots-clés du mythe et doivent donc avoir davantage de poids dans le calcul et non pas le contraire. Par ailleurs, Manjavacas *et al.* constatent que ce calcul est peu efficace pour des corpus de taille restreinte et les textes courts⁴³.

⁴¹ Enrique Manjavacas, Brian Long et Mike Kestemont, « On the Feasibility of Automated Detection of Allusive Text Reuse ».

⁴² Voir III.2c) « Tesseract » et Christopher Forstall et al., « Modeling the scholars: Detecting intertextuality through enhanced word-level n-gram matching », *Digital Scholarship in the Humanities*, vol. 30, n° 4, 2015, p. 503-515, en ligne, <doi: 10.1093/llc/fqu014>.

⁴³ Enrique Manjavacas, Brian Long et Mike Kestemont, « On the Feasibility of Automated Detection of Allusive Text Reuse », paragr. 5.1.a Hand-crafted scoring function.

Les auteurs proposent également parmi les calculs lexicaux l'équation *BOW* & *TF-IDF* qui intègre deux méthodes statistiques utilisées couramment dans le traitement automatique des langues et la recherche d'information. La première, *Bag-of-Words* (*BOW*) ou “sac de mots” est un modèle qui représente un texte en tant que vecteur où chaque dimension représente un mot et sa valeur correspond à la fréquence de ce mot dans le document. Ensuite, le calcul *Tf-IDF*⁴⁴ permet d'affiner ce modèle en pondérant les mots en fonction de leur importance dans le corpus. Ainsi, le poids des mots très fréquents est atténué, mais reste tout de même plus important que dans le calcul Tesseract.

Quant aux méthodes sémantiques et hybrides, elles sont appuyées sur les approches distributionnelles, comme *sentence embedding* (“représentation vectorielle de la phrase”), *Word Mover Distance* (*WMD*, “distance du déplacement des mots”) ou encore *Soft Cosine Similarity*. Elles prennent en compte la similarité sémantique entre les mots et non seulement les correspondances exactes, comme c'est le cas des calculs lexicaux. En introduisant une matrice de similarité, deux mots peuvent être considérés comme similaires, à condition qu'ils soient liés sémantiquement. Toutefois, ces types de calculs nécessitent un modèle pré-entraîné (*Word2Vec*, *FastText*, *BERT*, etc.) proposant une matrice de similarité.

Pour évaluer l'application des calculs de similarité les plus adaptés aux réutilisations textuelles allusives, nous proposons une adaptation de l'expérience proposée par Manjavacas *et al.*⁴⁵ Pour ce faire, nous appliquons six calculs à 500 phrases que nous avons extraites de Tracer et évaluées manuellement⁴⁶ :

- les calculs lexicaux : l'indice de Jaccard et le calcul *BOW* & *TF-IDF*
- les calculs sémantiques avec les modèles *FastText* pour le français⁴⁷ et *CamemBERT*⁴⁸
- les calculs hybrides avec les mêmes modèles et le calcul *Soft Cosine Similarity*, ce qui permet de combiner l'approche lexicale *BOW* & *TF-IDF* et les calculs sémantiques.

Nous lemmatisons les couples de l'échantillon et supprimons les mots faibles pour optimiser les résultats.

Si nous observons les scores moyens obtenus par différents calculs (Tableau 82), nous constatons, comme Manjavacas *et al.*⁴⁹, que les calculs lexicaux et hybrides sont plus efficaces que ceux

⁴⁴ $Tf - IDF(w, s) = Tf(w, s) \times IDF(w)$

⁴⁵ Enrique Manjavacas, Brian Long et Mike Kestemont, « On the Feasibility of Automated Detection of Allusive Text Reuse », tab. 2-4.

⁴⁶ Pour reproduire cette expérimentation et accéder aux résultats détaillés, voir, dans les annexes de cette partie, le dossier [III 4. degre-similarite](#).

⁴⁷ Voir <https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.fr.300.bin.gz>.

⁴⁸ Louis Martin et al., *CamemBERT: a Tasty French Language Model*, 2020, en ligne, <<https://inria.hal.science/hal-02889805>>.

⁴⁹ Enrique Manjavacas, Brian Long et Mike Kestemont, « On the Feasibility of Automated Detection of Allusive Text Reuse », part. 6.

purement sémantiques (peu importe le modèle employé). En effet, les scores moyens pour ces derniers sont comparables entre différents types de couples et le score moyen le plus élevé appartient aux couples du type 1.

Tableau 82. Concordance entre l'évaluation manuelle et différents calculs de similarité

Moyenne calcul	Type 1	Type 2	Type 3	Type 4	Type 5
Jaccard	0,007	0,061	0,069	0,080	0,111
TF-IDF, BOW et <i>Cosine Similarity</i>	0,025	0,121	0,124	0,141	0,215
FastText et <i>Cosine Similarity</i>	0,724	0,615	0,708	0,706	0,701
CamemBERT et <i>Cosine Similarity</i>	0,892	0,858	0,873	0,882	0,881
FastText et <i>Soft Cosine Similarity</i>	0,073	0,137	0,153	0,174	0,238
CamemBERT et <i>Soft Cosine Similarity</i>	0,020	0,107	0,114	0,127	0,187

L'analyse des dix résultats les plus élevés pour chaque calcul (Tableau 83) va dans le même sens. Alors que les calculs sémantiques proposent très peu de couples du type 5 parmi les scores le plus élevés (2 pour FastText et 1 pour CamemBERT), nous arrivons à 6 couples sur 10 du type 5 pour les calculs TF-IDF et les deux calculs hybrides. Alors que pour TF-IDF, les 4 couples restants sont du type 2, ces résultats s'améliorent légèrement pour Soft Cosine avec le modèle CamemBERT (1 couple de type 3 et 3 couples de type 2) et encore davantage pour le calcul effectué avec le modèle FastText (1 couple de type 4, 1 de type 3 et 2 de type 2).

Tableau 83. Types des dix couples ayant obtenu les scores les plus élevés pour chaque calcul

Types parmi les 10 scores max	Type 1	Type 2	Type 3	Type 4	Type 5
Jaccard	0	3	2	1	5
TF-IDF, BOW et <i>Cosine Similarity</i>	0	4	0	0	6
FastText et <i>Cosine Similarity</i>	0	0	3	5	2
CamemBERT et <i>Cosine Similarity</i>	0	0	4	5	1
FastText et <i>Soft Cosine Similarity</i>	0	2	1	1	6
CamemBERT et <i>Soft Cosine Similarity</i>	0	3	1	0	6

Par ailleurs, la valeur du score moyen pour différents calculs (Figure 49 qui n'inclut pas les résultats peu pertinents des calculs sémantiques) est généralement en hausse entre les couples du type 1 et ceux du type 5, la différence étant plus importante entre ceux du type 1 et 2, puis entre 4 et 5.

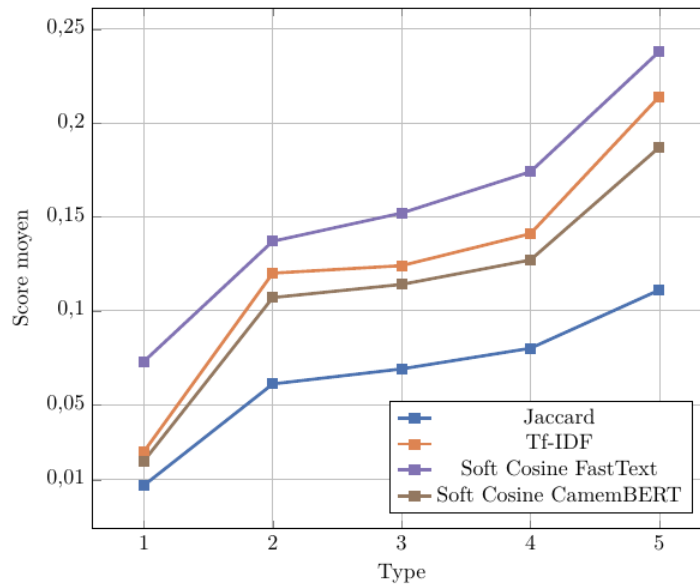


Figure 49. Évolution du score moyen par type de couple pour quatre calculs les plus pertinents.

Cela correspond à notre analyse des résultats de l'évaluation⁵⁰ et reflète la difficulté fréquente de trancher entre les couples des types 2-4, ce qui est plutôt satisfaisant. Le bon score obtenu pour ces trois types des couples prouve la nécessité de prise en compte des mots-clés du mythe qui permettront de trancher plus clairement au moins entre les couples du type 2 et ceux du type 3 et 4.

Cependant, comme Manjavacas *et al.*, nous restons mitigée face à ces résultats. D'une part, le fait que les modèles purement sémantiques sont largement dépassés par des calculs plus basiques, comme Tf-IDF ou Jaccard confirme notre expérience peu probante avec les plongements lexicaux⁵¹. D'autre part, même si les résultats hybrides sont les plus efficaces, ils ne surpassent que de peu les calculs lexicaux, alors qu'ils sont plus lourds en termes de complexité et du temps de traitement. Ce sont aussi des méthodes moins transparentes et donc plus difficiles à comprendre et apprécier par un humain. Par exemple, nous sommes étonnée de voir que le modèle FastText, entraîné sur le corpus de Facebook français, est plus efficace que CamemBERT, conçu spécifiquement pour les corpus littéraires. Cependant, nous n'avons pas les compétences suffisantes pour préciser s'il s'agit d'un problème d'algorithme que nous concevons ou de la non-compatibilité avec notre petit corpus diachronique.

Comme Manjavacas *et al.* nous constatons la difficulté de la tâche de l'évaluation automatique des relations intertextuelles, notamment celles les plus allusives :

⁵⁰ Voir III.4a) « Évaluation des résultats ».

⁵¹ Voir III.4b), § « Difficultés pour la prise en compte optimale de la synonymie et des reformulations polylexicales ».

Nos expériences ont mis en évidence les difficultés de la détection automatique des allusions. Même en supposant des requêtes définies manuellement, le modèle le plus performant ne pouvait trouver la référence correspondante dans les 20 premiers résultats que dans moins de la moitié du jeu de données. [...] Dans l'ensemble, les modèles hybrides [...] ont obtenu de meilleurs résultats, ce qui montre l'efficacité de cette technique [...] et prouve que des améliorations de la détection des réutilisations textuelles allusives sont possibles, quoique limitées, grâce à la sémantique lexicale⁵².

Nous souhaiterions donc disposer plutôt d'un calcul "maison", moins complexe, mais plus transparent en termes d'entités prises en compte dans le score. Il nous semble également nécessaire de l'adapter à la spécificité de notre recherche en accordant plus de poids aux mots-clés du mythe.

Conclusion du chapitre

L'évaluation de l'échantillon des résultats obtenus avec TextPAIR et Tracer pour notre corpus des réécritures (confronté à lui-même et au corpus des traductions) est plutôt satisfaisante. Même si les relations peu pertinentes (notamment du type 2) restent bien présentes, celles pertinentes (type 3-5) montrent plusieurs appariements intéressants et confirment la capacité des deux logiciels d'apparier les relations intertextuelles, même celles allusives.

Néanmoins, plusieurs difficultés persistent, tant en ce qui concerne la préparation du corpus que l'exploration et la visualisation des résultats. En effet, nous avons observé que le format du corpus et des résultats, différent pour les deux logiciels, induit des erreurs et empêche l'exploration conjointe des couples trouvés par TextPAIR et Tracer, ce qui rend nécessaire l'unification des formats à l'entrée et la compilation des résultats à la sortie. Nous souhaitons également écarter les couples qui ne partagent que des mots faibles en commun ou dont les mots communs appartiennent aux syntagmes portant différents sens. Par ailleurs, les erreurs de lemmatisation automatique nous contraignent à corriger manuellement les sorties de TreeTagger. Nous constatons aussi que nombreux couples présentent des relations synonymiques et des reformulations polylexicales qu'il serait intéressant d'aligner.

Du côté de l'exploration et de la visualisation des résultats, les entités communes (mots, lemmes, synonymes...) méritent d'être mises en valeur pour faciliter l'appréciation de la relation observée. Nous souhaitons également avoir accès au contexte plus large de l'extrait et pouvoir moduler certains paramètres (segment à afficher un premier, sous-ensembles à comparer...). Comme nous

⁵² « Our experiments have highlighted the difficulties of automated allusion detection. Even assuming manually defined queries, the best performing model could only find the matching reference within the top 20 hits in less than half of the dataset. [...] Overall, the hybrid models involving soft cosine show best performance, which indicates the effectiveness of such technique to incorporate semantics into BOW-based document retrieval and offers evidence that improvements in allusive reuse detection, however limited, can be gained from lexical semantics. » (nous traduisons). Voir Enrique Manjavacas, Brian Long et Mike Kestemont, « On the Feasibility of Automated Detection of Allusive Text Reuse », part. 6 Conclusions and Future Work.

l'avons montré avec l'exemple d'Esquiros¹⁸⁴⁸, différents types de visualisations permettent des analyses très diverses : de l'observation d'un graphe de relations construit autour d'un segment de texte au placement de toute l'œuvre dans son réseau intertextuel, nous avons besoin de pouvoir choisir les éléments à comparer (tous les segments appariés, les segments appariés avec des traductions, le nombre total des relations détectées avec chaque œuvre...). Enfin, une mesure du degré de similarité peut aider à écarter certains couples peu pertinents, mais elle semble devoir être adaptée à la spécificité du corpus. Nous avons vu que les expérimentations effectuées avec les calculs existants ont été peu concluantes, car toutes les mesures de similarité prise en compte attribuent aux couples du type 2 des scores qui sont supérieurs aux ceux des types 3-4.

Conclusion de la partie

Cette troisième partie, qui s'est intéressée spécifiquement aux méthodes informatiques de la détection automatique des réemplois textuels, nous a d'abord permis de préciser nos besoins en termes d'outillage. En menant une investigation préliminaire avec un outil Phœbus— facile à prendre en main, car disposant d'une interface utilisateur, mais hors service actuellement — nous sommes parvenue à établir une liste de critères pour le choix optimal du logiciel à retenir pour notre projet. Cette liste se divise en cinq thèmes centraux : (i) la nature des recherches menées à l'aide du logiciel, (ii) les critères techniques nécessaires pour pouvoir l'utiliser (installation, équipement spécifique, etc.), (iii) le format attendu pour les données, (iv) les possibilités de paramétrage et (v) la qualité des résultats obtenus.

Nous avons ensuite procédé à l'évaluation des six principaux logiciels de détection des réutilisations, BLAST, Passim, Tesserae, TextPAIR, Textreuse et Tracer. Pour faciliter la comparaison, nous nous sommes laissée guider par la grille d'évaluation et avons utilisé le même petit corpus de six traductions d'Ovide. Le deuxième chapitre nous a permis de présenter les quatre logiciels écartés, BLAST, Textreuse, Tesserae et Passim, et d'argumenter notre choix en exposant les difficultés rencontrées. En effet, nous constatons que malgré les modes de fonctionnement plutôt comparables, les résultats de ces quatre logiciels divergent et ne sont généralement pas très satisfaisants (il s'agit pour la plupart d'un simple copier-coller). Même si certains logiciels, comme Passim, se démarquent par exemple par la facilité d'installation et la rapidité du traitement, aucun ne répond à nos besoins en matière de paramétrage (racinisation ou lemmatisation des mots, suppression des mots faibles, etc.). Par ailleurs, à notre connaissance aucun n'a été exploité avec un corpus français et littéraire.

Le troisième chapitre s'est consacré, quant à lui, à la présentation et la comparaison des deux logiciels retenus, TextPAIR et Tracer. Nous avons décrit les performances et les caractéristiques des deux logiciels selon les cinq thèmes centraux de la grille d'évaluation, ce qui nous a permis d'exposer leur supériorité vis-à-vis des logiciels retenus, ainsi que leur complémentarité, notamment en ce qui concerne les recherches existantes, le paramétrage et les résultats. Nous avons testé plusieurs paramétrages pour aboutir aux résultats qui nous semblent optimaux avant de soumettre aux deux logiciels la totalité de notre corpus des réécritures. Cela nous permet aussi de démontrer que l'année de la conception d'un outil informatique, malgré les idées généralement reçues, ne doit pas être le seul critère conditionnant son emploi. En effet, même si TextPAIR et Tracer ont été conçus bien avant les logiciels comme Textreuse ou Passim, le fait que tous les deux soient initialement

destinés aux corpus littéraires et que TextPAIR soit régulièrement mis à jour, les rend mieux exploitables pour nos besoins. Une relative difficulté technique concernant l'installation et le lancement de ces deux logiciels est palliée par le fait que ses développeurs sont accessibles pour répondre à d'éventuelles questions et que d'autres projets semblables au nôtre sont en train d'y mener des recherches en parallèle.

Enfin, le dernier chapitre a présenté un échantillon des résultats obtenus pour le corpus des réécritures confronté à lui-même et au corpus des traductions. Nous les avons évalués manuellement selon une grille de notation à 5 points. Nous avons présenté nos observations pour chaque type de relation, puis avons présenté les principaux problèmes de traitement, d'exploration et de visualisation.

L'identification de nos besoins en matière d'outillage, la comparaison exhaustive des logiciels existants et l'évaluation des résultats finaux sont indispensables pour optimiser la qualité des résultats obtenus et pour concevoir une méthode de post-traitement et une interface de visualisation adaptées, permettant de pallier la plupart des problèmes identifiés et de procéder à l'analyse des correspondances détectées sans craindre les fausses détections ou les erreurs liées au paramétrage. Nous le montrons dans la partie suivante.

IV. Édition numérique intertextuelle : propositions et perspectives

À présent, nous savons exploiter TextPAIR et Tracer de manière optimale et l'évaluation de l'échantillon des détections montre de bons résultats. Désormais, nous aimerions procéder à l'analyse littéraire et linguistique des correspondances détectées. En effet, notre corpus, en raison de son hétérogénéité, réserve plusieurs défis pour la détection des correspondances. Néanmoins, comme nous l'avons montré à la fin de la partie précédente, certains problèmes persistent, notamment en ce qui concerne l'exploration et la visualisation des couples. En effet, hormis ce qui sert à identifier chaque extrait, aucun détail n'est fourni sur les entités qui motivent la mise en correspondance. Pourtant, notamment pour pouvoir écarter les correspondances fausses ou peu pertinentes, nous devons savoir ce qui est apparié, les motivations des liens (identité de mots, de lemmes ou de racines, ou synonymes communs), le nombre d'entités communes et la distance qui les sépare.

C'est pour cela que nous consacrerons les deux premiers chapitres de cette dernière partie à la présentation de la manière dont nous post-traitons les résultats bruts. De l'unification des formats de TextPAIR et Tracer à l'annotation fine des entités nommées et de leurs périphrases, nous proposons une série d'algorithmes permettant d'enrichir les informations concernant chaque couple aligné, d'écarter les couples peu pertinents restants et de procéder aux premières analyses des réseaux de réécritures ainsi construits.

Ainsi, au premier chapitre, nous nous servons de quelques résultats obtenus pour le texte de Ballanche1809 pour présenter les deux premières étapes de notre traitement, l'unification des formats de TextPAIR et Tracer, l'appariement des entités communes (noyaux des syntagmes, puis unités lexicales) et le calcul du score de ressemblance. Cela nous permet de présenter quelques observations intéressantes pour les œuvres symboliques du corpus, et notamment celles de Ballanche (Ballanche1809 et Ballanche1829), de Schuré (Schuré1889) et de Segalen (Segalen1921).

Au deuxième chapitre, nous constatons l'importance de certains éléments récurrents au sein des couples appariés, comme les périphrases des noms propres (Orphée, Eurydice, les Enfers, Pluton...). Nous montrons que l'intégration de méthodes linguistiques et une analyse préliminaire sélective du corpus améliorent la qualité des résultats de la détection réalisée par les outils informatiques choisis. Après avoir présenté notre développement, nous démontrons son efficacité en proposant une analyse particulière de différentes dénominations d'Eurydice qu'il trouve dans le corpus.

Le troisième chapitre tâche de répondre aux différents problèmes concernant l'exploration et la visualisation des résultats. Nous y présentons notre prototype d'une édition comparative numérique qui propose une interface d'exploration modulable et adaptée aux différents besoins d'analyse. Entre l'édition hyperannotée et l'édition hypertextuelle, il est possible de graduer le contenu présenté et découvrir progressivement les résultats obtenus pour différentes œuvres. Nous tâchons

d'argumenter nos choix éditoriaux et de démontrer qu'il est possible de structurer et présenter des données complexes de manière lisible, sans pour autant être contraint à la simplification.

Enfin, nous consacrons le dernier chapitre à l'introduction de quelques perspectives pour le futur du projet en présentant deux travaux d'amélioration de notre traitement que nous avons engagés sans parvenir à les intégrer au sein du développement : la prise en compte des séries traductives et l'amélioration du traitement des synonymes.

1. Post-traitement des résultats de la détection automatique des correspondances¹

Dans le dernier chapitre de la partie précédente², nous avons soulevé plusieurs problèmes liés à la préparation du corpus et du paramétrage des logiciels et à l'exploration et la visualisation des résultats. Les deux logiciels de détection automatique des correspondances se sont avérés très différents, tout en proposant des paramétrages complémentaires.

Rappelons-nous : TextPAIR détecte des correspondances selon des paramètres modulables (choix du nombre minimal des entités communes à détecter et de leur nature – mots pleins, lemmes ou stemmes (mots racinisés) –, liste des mots à ne pas prendre en compte...) et traite des fichiers XML, mais sans en exploiter le balisage. Dans un premier temps, l'outil génère des *n-grammes* qui se chevauchent pour chaque texte, les stocke et les indexe afin de les confronter à ceux des autres textes. Tracer, lui, traite des textes en format TXT, prédécoupés en segments (des phrases pour notre corpus) dotés d'un identifiant unique. Des paramétrages sont possibles (recherche sur les lemmes ou uniformisation des mots similaires...) et le poids de chaque correspondance est calculé.

Dans ce chapitre, nous montrons que les résultats de ces deux logiciels peuvent être combinés utilement, grâce, notamment, aux différences du traitement des textes soumis. Nous présentons également la manière dont nous post-traitons chacune des sorties, ce qui nous permet de les compiler et d'enrichir les résultats, de filtrer les couples peu ou pas pertinents et de générer une interface d'exploration offrant une vue globale et claire des couples de correspondances détectées³.

La chaîne de post-traitement, qui permet d'intégrer les résultats et les annotations enrichies à des fichiers XML dédiés, comporte trois étapes :

1. Unification des formats des résultats (compilation, filtrage et XMLisation) ;
2. Analyse détaillée de chaque couple de correspondances afin de détecter les noyaux des syntagmes et les unités lexicales liés, leur nombre et leur degré de proximité, de calculer le score de ressemblance, puis d'écarter les couples trop faiblement liés ;
3. Adjunction d'appariements plus fins, notamment entre diverses dénominations d'entités nommées⁴.

¹ Ce chapitre reprend et complète deux de nos articles (co-signés avec N. Gasiglia) : « Réécritures d'un mythe et outils de détection des réutilisations. De l'Orphée de Virgile à celui de Ballanche », *Humanités numériques*, n° 4, Humanistica, 2021, en ligne, <<https://journals.openedition.org/revuehn/2467>> ; « On Digital Comparative Editions and Textual Similarity Detection Tools: Towards a Hypertextual Cartography of a Rewritten Myth », dans Petr Plecháč et al. (dir.), *Tackling the Toolkit: Plotting Poetry through Computational Literary Studies*, Institute of Czech Literature, Czech Academy of Sciences, 2022, p. 167-182, en ligne, <https://www.plottingpoetry.org/books/tackling_toolkit/chapter?chap=11>.

² Voir III.4 « Présentation des résultats initiaux et problèmes rencontrés ».

³ L'interface d'exploration est présentée au IV.3. « Présentation de prototype d'une édition comparative numérique », mais nous l'utilisons dès ce chapitre pour présenter les correspondances analysées.

⁴ Cette étape est décrite de manière détaillée au IV.2 « Annotation des entités nommées et leurs périphrases ».

Nous présentons les deux premières étapes⁵, puis exploitons les résultats enrichis et le réseau des réécritures ainsi construit en nous appuyant principalement sur les réécritures symboliques du corpus qui valorisent un Orphée fondateur des cultes.

a) Unification des formats

Après avoir observé les problèmes liés à la mauvaise interprétation du format XML par TextPAIR, qui résultent de l'accumulation des couples non pertinents, liés par des éléments formels des œuvres⁶, nous décidons de soumettre au logiciel le corpus en format texte, dont le contenu est identique à celui traité par Tracer. Cela facilite la compilation des résultats à l'aide de la segmentation du corpus en phrases associées à un identifiant unique, initialement réalisée pour Tracer. Quelques manipulations s'imposent tout de même pour découper ou réunir les segments détectés par TextPAIR selon le même principe, puis supprimer les doublons et regrouper les correspondances isolées en extraits plus étendus. Par exemple, Tracer propose d'aligner la phrase 18 de Ballanche1809 et la phrase 216 de Ballanche1829 (Tableau 84).

Tableau 84. Extrait d'une correspondance détectée par Tracer

Id A	Id B	Nb	Score	Segment A	Segment B
0300018	0400216	3	0,6	Un instant encore, et elles échappaient à la puissance du dieu des morts : elles touchaient au seuil du séjour des vivants.	Ainsi donc il touchait au seuil qui sépare l'empire des morts du séjour des vivants, lorsqu'il s'arrêta involontairement pour jeter un regard d'amour sur l'épouse qui lui était rendue, hélas ! rendue et ravie à la fois.

Ce couple est également présent parmi les résultats de TextPAIR, mais le segment commun est plus étendu (Figure 50).

```
{
  "source_passage" : "touchaient au seuil du séjour des vivants. Mais, ô faiblesse d'un cœur qui aime ! Orphée s'arrête pour écouter",
  "source_author" : "Pierre-Simon de Ballanche",
  "source_title" : "Sixième Fragment",
  [...]
  "target_passage" : "touchait au seuil qui sépare l'empire des morts du séjour des vivants, lorsqu'il s'arrêta involontairement",
  "target_author" : "Pierre-Simon de Ballanche",
  "target_title" : "Orphée"
}
```

Figure 50. Extrait d'une correspondance similaire détectée par TextPAIR

⁵ Les scripts ayant servi pour le post-traitement décrit dans cette partie se trouvent dans le dossier [IV 1 post-traitement](#). Ils sont également disponibles sur GitHub (<https://github.com/karolinasuchecka/orphidys>) et GitLab (<https://gitlab.univ-lille.fr/karolina.suchecka/>).

⁶ Voir III.4b), § « Problèmes liés au format du corpus et des résultats ».

Dans la première étape de notre traitement, cette correspondance commune aux deux logiciels est révisée et étendue pour arriver au découpage du Tableau 85 (l'extrait détecté par TextPAIR, qui chevauche trois phrases de Ballanche1809, est en italiques) :

Tableau 85 . Identification des phrases correspondantes aux extraits alignés par TextPAIR

Ballanche1809	Ballanche1829
[18] Un instant encore, et elles échappaient à la puissance du dieu des morts : elles <i>touchaient au seuil du séjour des vivants</i> . [19] <i>Mais, ô faiblesse d'un cœur qui aime !</i> [20] <i>Orphée s'arrête</i> pour écouter le soupir qui errait sur les lèvres d'Eurydice, pour prêter l'oreille au léger frôlement de ses vêtements aériens.	[216] Ainsi donc il <i>touchait au seuil qui sépare l'empire des morts du séjour des vivants, lorsqu'il s'arrêta involontairement</i> pour jeter un regard d'amour sur l'épouse qui lui était rendue, hélas ! rendue et ravie à-la-fois.

Après avoir identifié les phrases composant les segments communs de TextPAIR⁷, nous procédons à la comparaison des résultats des deux logiciels et à la suppression des doublons. Nous passons ainsi de 46 314 à 24 289 couples, ce qui prouve la répétitivité des résultats des deux logiciels.

Nous procédons ensuite au regroupement des couples en segments plus étendus. Par exemple, nous retrouvons parmi les résultats de Tracer un couple de correspondances qui précède immédiatement celui présenté au Tableau 85 (Tableau 86).

Tableau 86. Résultats de Tracer qui complètent la correspondance présentée au Tableau 85

Ballanche1809	Ballanche1829
[17] Déjà les ombres devenaient moins opaques ; déjà un faible crépuscule détaché des rayons du soleil, arrivait jusqu'aux yeux des deux nobles créatures qui s'avançaient vers la lumière du jour.	[215] Déjà un crépuscule douteux commençait à remplacer l'obscurité immense : un rayon de pâle lumière, qui vint flotter autour de ses yeux, lui fit oublier le passé et l'avenir : ce n'était pas une lueur trompeuse, c'était bien un rayon détaché de la clarté du jour : la marque de son salut devint le signe de sa perte.

Ce couple est ajouté aux segments précédents afin de faire correspondre les phrases 17-20 de Ballanche1809 et 215-216 de Ballanche1829. Enfin, parmi les résultats TextPAIR, nous trouvons un autre extrait qui permet d'enrichir davantage ce couple (Tableau 87).

Tableau 87. Résultats de TextPAIR qui complètent la correspondance présentée au Tableau 85

Ballanche1809	Ballanche1829
[21] Vaincu par cette puissance contre laquelle l'homme lutte en vain, il se retourne ; et oubliant sa fatale promesse, il permet à son regard d'interroger à la faveur de la clarté naissante, les plis du <i>voile qui lui cachait le visage de son épouse</i> , cette touchante victime qu'il avait tant pleurée.	[217] Si du moins, il eût le temps de soulever le <i>voile qui retombait sur le visage d'Eurydice !</i>

⁷ Voir le script [1_chargement_resultats.py](#).

Ainsi, deux correspondances de Tracer et deux de TextPAIR en deviennent une, mais combinée dans un segment plus étendu (5 phrases de Ballanche1809 et 3 phrases de Ballanche1829). Nous passons ainsi de 24 289 à 24 242 couples, les correspondances se limitant à une phrase restant tout de même majoritaires parmi les résultats⁸.

Les phrases qui composent les correspondances restantes sont enrichies par le balisage en XML afin de fournir explicitement des indices précieux, mais implicites dans les textes⁹. Une annotation à plusieurs niveaux est ainsi associée à chaque phrase, syntagme et mot (Figure 51, pour une partie de la phrase 21 de Ballanche1809).

```
<s id="0300021">[...]
  <phr type="SN" select="visage">
    <w xml:id="0300021_47" n="47" lemma="le" pos="DET:ART">le</w>
    <w xml:id="0300021_48" n="48" lemma="visage" pos="NOM" sameAs=
      "figure physiologie tête air">visage</w>
    <w xml:id="0300021_49" n="49" lemma="de" pos="PRP">de</w>
    <persName xml:id="0300021_50-51" corresp="eurydice" type="périphrase">
      <phr type="SN" select="épouse">
        <w xml:id="0300021_50" n="50" lemma="son"
          pos="DET:POS">son</w>
        <w xml:id="0300021_51" n="51" lemma="épouse" pos="NOM"
          sameAs="conjointe bourgeoise femme régulière">
          épouse</w>
      </phr>
    </persName>
  </phr>
</s>
```

Figure 51. Balisage d'un extrait de la phrase 21 de Ballanche1809

L'élément <s> porte l'attribut @xml:id qui indique son identifiant. Les syntagmes nominaux, verbaux et adjectivaux sont balisés comme <phr> avec un @select qui donne le lemme du noyau lexical. Le syntagme nominal « le visage de son épouse » en contient un autre, « son épouse », aussi annoté comme étant une désignation de l'entité nommée « Eurydice » à l'aide de son élément parent <persName>, qui donne l'identité du référent avec un @corresp et l'identifiant de la périphrase avec un

⁸ Voir le script [2_compilation_resultats.py](#).

⁹ Voir le script [3_xmlisation_phrases.py](#) et son résultat, [3_phrases_xmlises.xml](#). Le format XML n'est pas toujours le plus efficace pour le traitement des données volumineuses. Dans notre développement, nous employons plus régulièrement JSON, mais produisons tout de même des fichiers XML à des étapes-clés de notre post-traitement, notamment afin de les partager avec la communauté via notre dépôt Nakala (<https://nakala.fr/collection/10.34847/nkl.ba7766ru>).

@xml:id¹⁰. Les éléments <w> portent les codes grammaticaux (@pos) et les lemmes (@lemma) – produits par TreeTagger, puis corrigés manuellement¹¹ – et, pour les unités lexicales pleines, une sélection de synonymes (@sameAs) issus du *Dictionnaire électronique des synonymes (DES)*¹².

En ce qui concerne la délimitation des syntagmes, nous souhaitons reconnaître et délimiter trois types de syntagmes : nominaux, verbaux et adjectivaux. Par exemple, le syntagme nominal « son épouse » est en réalité un syntagme prépositionnel introduit par « de », mais notre objectif étant de repérer des dénominations, les syntagmes prépositionnels ne sont pas balisés. Ils peuvent en revanche être restitués en leur associant la préposition antéposée, si un tel besoin se présente.

Les outils du TAL proposant l'extraction automatique des syntagmes se limitent à l'extraction des groupes nominaux (c'est le cas de SpaCy¹³) ou ne peuvent traiter que les textes en anglais (comme Stanza¹⁴). Alors que certains outils du traitement automatique des langues s'emploient à la construction (et la visualisation) des arbres syntaxiques¹⁵, le fait que nous avons affaire à des phrases souvent très complexes et que notre corpus n'est pas synchronique rend ces analyses peu performantes. Nous concevons donc un traitement à base de règles¹⁶. Au Tableau 88, nous comparons ce que nous obtenons par rapport aux résultats de SpaCy pour une des phrases de Ballanche1809.

Tableau 88. Comparaison de l'extraction des syntagmes par SpaCy et par notre développement à base de règles

Extraction des groupes nominaux :	
Avec SpaCy	Avec notre développement
<u>Syntagmes nominaux :</u> « [Un instant] encore, et [elles] échappaient à [la puissance] du [dieu] des [morts] : [elles] touchaient au [seuil] du [séjour des vivants] »	<u>Syntagmes nominaux :</u> « [Un instant encore], et elles échappaient à [la puissance du [dieu des [morts]]] : elles touchaient au [seuil du [séjour des [vivants]]] » <u>Syntagmes verbaux :</u> « Un instant encore, et [elles échappaient à] la puissance du dieu des morts : [elles touchaient au] seuil du séjour des vivants »

¹⁰ Nous décrivons et commentons notre manière d'annotation des périphrases au chapitre suivant, IV.2 « Annotation des entités nommées et leurs périphrases ».

¹¹ Voir III.4b), sous-section « Erreurs de lemmatisation qui nuisent à la recherche des mots communs ».

¹² Il s'agit d'un dictionnaire de synonyme cumulative de langue générale conçu au sein du laboratoire Crisco (université de Caen). Voir <https://crisco4.unicaen.fr/des/> et Laurette Chardon et Jacques François, « Les vedettes du Dictionnaire Électronique des Synonymes et les relations d'adjacence entre leurs synonymes », *Lexique*, n° 27, 2020, p. 21, en ligne, <<https://halshs.archives-ouvertes.fr/halshs-03192815>>. En son sein, les corrélatés sont classés par nombre décroissant de sens élémentaires communs avec la vedette. Les synonymes ainsi extraits ne sont pas toujours pertinents, ce que nous discutons au IV.4b) « Amélioration du traitement des synonymes ».

¹³ SpaCy est un puissant outil de TAL conçu nativement en langage Python. Voir <https://spacy.io/>.

¹⁴ Stanza est un outil créé par Stanford NLP Group (université de Stanford) permettant l'analyse linguistique (de segmentation à la reconnaissance des entités nommées) de 70 langues différentes. Voir <https://stanford-nlp.github.io/stanza/> et Peng Qi et al., *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages*, Association for Computational Linguistics (ACL) System Demonstrations, 2020.

¹⁵ Voir par exemple Sebastian Ruder, « Dependency parsing », dans *NLP-progress*, en ligne, <http://nlp-progress.com/english/dependency_parsing.html>.

¹⁶ Voir la fonction « chunking » du script [3_xmlisation-phrases.py](#).

Premièrement, SpaCy ne prend pas en compte la complexité de certains syntagmes nominaux, pouvant contenir eux-mêmes d'autres syntagmes. Il propose donc trois syntagmes isolés « la puissance », « dieu » et « morts », alors que notre traitement délimite plutôt « la puissance du dieu des morts », « dieu des morts » et « morts ». Ensuite, SpaCy considère les pronoms comme syntagmes nominaux, alors que nous les incluons dans les syntagmes verbaux (« elles échappaient à », « elles touchaient au ». Enfin, la segmentation de SpaCy paraît arbitraire, car rien n'explique pourquoi délimiter le syntagme « séjour des vivants » tout en séparant « dieu des morts » en deux syntagmes isolés. Si des erreurs persistent également dans notre balisage des syntagmes, il nous permet tout de même de réaliser de manière satisfaisante l'étape suivante de notre post-traitement des résultats, à savoir l'appariement des syntagmes correspondants.

b) Détection des noyaux des syntagmes et les unités lexicales liés, puis calcul du score de ressemblance

La recherche de relations lexicales au sein de chaque correspondance implique une comparaison qui exploite les valeurs des attributs en effectuant des appariements sur plusieurs niveaux.

*Alignement des noyaux des syntagmes*¹⁷

Les résultats compilés de Tracer et TextPAIR pour la phrase 21 de Ballanche¹⁸⁰⁹ montrent 15 couples impliquant 14 textes, dont 5 traductions d'Ovide et 3 de Virgile. Une partie de ces résultats se trouve à la Figure 52¹⁸. Tout d'abord, seuls les lemmes noyaux de syntagmes appariés.

☒ Noyaux lemmatisés de syntagmes
 [Noyaux identiques]
 [Noyaux synonymes]
 [Noyaux en relation flexionnelle ou dérivationnelle]

BALLANCHE1809, PHRASE 21

BC[Vaincu]vaincre par D[cette puissance]puissance contre laquelle l'homme lutte en vain, BC[il se retourne]retourner et, B[oublissant]oublier B[sa fatale promesse]promesse, il permet à ABC[son regard]regard d'interroger à la faveur de AB[la clarté naissante]clarté, A[les plis du A[voile]voile]pli qui lui cachait A[le visage de AD[son épouse]épouse]visage, de cette touchante victime qu'il avait tant pleurée.

BALLANCHE1829, PHRASE 1014 [A]

D'une main tremblante, celle qui veut être initiée reçoit des mains du poète [le voile tissu par [l'épouse d'un Titan]épouse]voile : ce voile, emblème d'une science cosmogonique, descend à [longs plis]pli sur [le visage de la vierge]visage : et [ses yeux]œil ont peine encore à soutenir [la clarté du jour]clarté.

CABARET-DUPATY1897, PHRASE 11 [B]

Presque aux portes du [jour]jour, il s'arrête, hélas ! [il oublie]oublier [sa promesse]promesse, et, [vaincu par]vaincre l'amour, [il tourne]tourner la tête [pour regarder]regarder sa chère Eurydice.

COURNAND1805, PHRASE 19 [C]

[Vaincu d']vaincre amour, il [regarde]regarder Eurydice, et perd tout sans retour : les traités sont rompus, l' enfer va la reprendre : trois fois, l' affreux Averse au loin se fait entendre.

VILLENAVE1806, PHRASE 31 [D]

Ni le dieu de [l'empire des morts]empire, ni [son épouse]épouse ne peuvent résister aux accords puissants du chanter de la Thrace.

Figure 52. Quatre phrases en correspondance avec la phrase 21 de Ballanche¹⁸⁰⁹

¹⁷ Voir le script [4 alignement_correspondances.py](#).

¹⁸ Nous utilisons désormais l'affichage des correspondances conçu au sein du projet, que nous présentons en détail dans le chapitre suivant (IV.3 « Présentation de prototype d'une édition comparative numérique »).

La phrase 21 de Ballanche1809 compte 20 syntagmes (parfois enchâssés). Pour Ballanche1829 ([A]), nous trouvons 5 noyaux identiques et une équivalence synonymique. Les deux phrases sont de longueur comparable et le ratio obtenu n'est que de 30 %, ce qui attire l'attention sur la complexité de la reformulation entre les deux extraits. Par ailleurs, le noyau « épouse » est problématique : dans Ballanche1809, il renvoie à Eurydice, alors que dans Ballanche1829 la périphrase « l'épouse d'un Titan » désigne sa mère¹⁹.

La phrase de Cabaret-Dupaty1897 ([B]), une des traductions de Virgile, compte trois noyaux identiques (*oublier*, *promesse* et *vaincre*), une équivalence synonymique (*clarté~jour*) et deux radicaux correspondants (*tourner~retourner* et *regard~regarder*), soit 6 syntagmes sur 11 appariés. Comme cette phrase est relativement courte, le ratio atteint 55 % environ. Mais trois groupes importants de Cabaret-Dupaty1897 restent isolés :

- « il s'arrête », que nous retrouvons dans la phrase 20 de Ballanche1809
- « l'amour », fortement métaphorisé par « cette puissance contre laquelle l'homme lutte en vain » et
- « sa chère Eurydice », périphrasé par « son épouse » et « cette touchante victime qu'il avait tant pleurée ».

Une autre traduction de Virgile présente dans les résultats est celle de Courmand1805 ([C]). Uniquement le début du segment correspond au même épisode du mythe, ce qui explique que seuls deux noyaux sont alignés : *vaincre* et *regard~regarder*. Le ratio n'atteint que 15 % (2 noyaux sur 13 sont alignés). Pourtant, la correspondance entre deux segments, quoique fragmentaire, reste effective.

Au contraire, la traduction d'Ovide de Villenave1806 concerne un épisode antérieur à la seconde mort d'Eurydice, à savoir l'enchantement des Infernaux. Les correspondances nucléaires se limitent à l'identité de la périphrase « son épouse » et la synonymie entre *empire* et *puissance*, mais la phrase est courte, nous arrivons donc tout de même au ratio de 25 % (2 noyaux sur 8 sont appariés).

Ainsi, comme l'appariement des synonymes n'est pas toujours pertinent et qu'une relation peut se limiter à un court extrait d'un long segment, il semble avantageux de fixer un nombre minimal des noyaux identiques, en relation flexionnelle ou en relation dérivationnelle, en plus du ratio de similarité. Par ailleurs, la valeur de l'équivalence (exacte ou au niveau des synonymes) exigée pour que la correspondance nucléaire soit considérée comme effective dépend autant des besoins de recherche que des paramètres utilisés pour le traitement. Si nous fixons le seuil à 20 %, 13 000 couples sont retenus (la moitié des résultats compilés), à 30 %, ce nombre atteint 8 514, à 40 % il

¹⁹ Les sources antiques du mythe ne mentionnent pas les parents d'Eurydice, mais chez Ballanche, elle est la fille de Talaon, le dernier Titan, et de Vola, une nymphe.

n'est que de 5 276, etc. Hélas, aucune solution ne garantit d'écarter toutes les fausses correspondances sans perdre les détections pertinentes.

Nous le constatons en appariant les noyaux de notre échantillon d'évaluation²⁰ (Figure 53). En effet, certains couples de type 2 obtiennent des ratios relativement élevés (supérieurs à 40 %), même si dans la majorité des cas (215/374), moins de 30 % des mots-clés peuvent être appariés. Il est toutefois rassurant de constater que pour les couples du type 5, le ratio de seulement 17 des 155 couples est inférieur à 30 %. La distribution approximativement égale des ratios pour les couples de type 4, en revanche, implique que le traitement des noyaux seuls ne suffit pas. Une adaptation plus spécifique des ressources linguistiques (listes de synonymes adaptées au corpus, mots-clés, entités nommées et les périphrases, etc.) peut améliorer la pertinence de cet appariement.

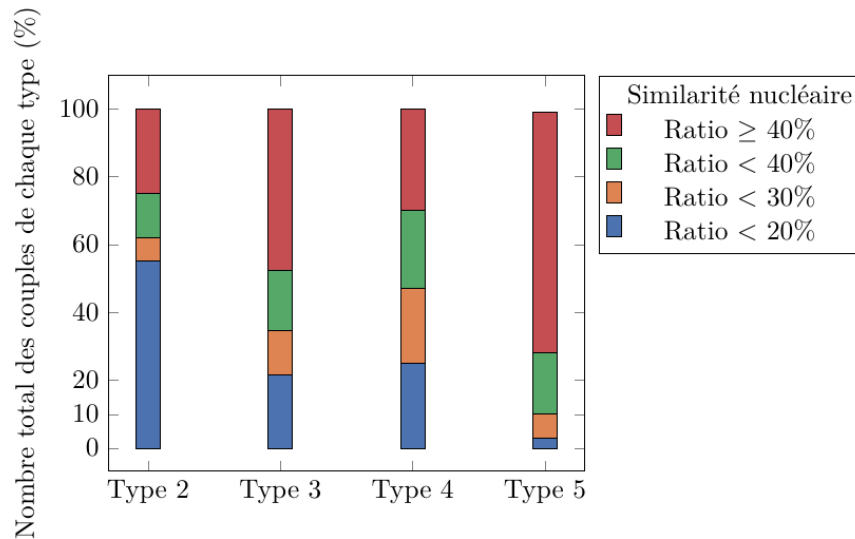


Figure 53. Répartition du ratio de similarité nucléaire au sein de l'échantillon d'évaluation

Pour l'instant, comme nous souhaitons travailler avec les résultats aussi riches que possible, nous n'écarterons que les correspondances dont le seuil de proximité nucléaire ne dépasse pas 20 % ou qui présentent seulement une relation "forte" (non synonymique). La suite du traitement, à savoir la recherche des relations lexicales, permet de peaufiner davantage cette première sélection.

Recherche des relations lexicales²¹

Plusieurs méthodes permettent d'évaluer la proximité lexicale entre les réutilisations d'un texte. Certaines de ces approches montrent la pertinence de la fréquence des mots. Cependant, comme tous les textes de notre corpus partagent le même thème mythologique, les relations pertinentes peuvent être basées sur des mots de fréquences très variables. Par exemple, *époux* et *mari* sont des

²⁰ Voir III.4a) « Évaluation des résultats ».

²¹ Voir le script [5 alignement-mots.py](#).

périphrases récurrentes pour Orphée. Leurs fréquences respectives dans le corpus sont relativement faibles (146 et 94 sur 432 078), mais elles apparaissent régulièrement dans les résultats de la détection. En effet, dans l'échantillon d'évaluation, 45 paires contiennent le mot *époux*, dont 17 sont de type 5 et 10 de type 4. D'autres mots-clés du mythe se retrouvent parmi les paires du type 4-5 tout en faisant partie des 100 mots les plus fréquents au sein du corpus (*amour, dieu, mort, voix, femme...*). Par conséquent, en plus du ratio de proximité des noyaux des syntagmes, nous proposons une méthode d'évaluation de la proximité lexicale simple, qui se concentre principalement sur le type d'équivalence.

Pour tous les couples ayant atteint le seuil de correspondances nucléaires minimal, nous procédons donc à une comparaison lexicale, en écartant les mots faibles et en calculant un score de proximité lexicale. Trois types de relations sont pris en compte :

1. Mot à mot (forme~forme, lemme~lemme, radical~radical) : +5 points
2. Entre synonymes (synonyme~synonyme) : + 0,5 point
3. Mixtes (mot~synonyme) : + 1,5 point

Quand une correspondance de type 1 est trouvée, les unités lexicales du couple ne sont plus prises en compte, mais pour les relations plus faibles la recherche se poursuit afin de trouver le couple lexical obtenant le meilleur score, et les points se cumulent si plusieurs correspondances entre synonymes ou mixtes sont trouvées. Comme quatre synonymes maximum sont associés aux mots pleins, le score lexical maximal (hors relation de type 1) peut atteindre 3 points : une correspondance mixte et trois correspondances entre synonymes.

Cette méthode permet de différencier les niveaux de proximité synonymique. Par exemple, l'équivalence entre *mari* et *époux* obtient 2 points : le lemme *mari* figure parmi les synonymes d'*époux* (+1,5 pts.) et le synonyme *conjoint* est commun aux deux mots (+0,5 pt.). Le couple *mari~homme* reçoit 1,5 point (mot~synonyme), alors que *époux~ami* n'est qu'à 0,5 pt. (synonyme commun *compagnon*). Actuellement, une correspondance synonymique est considérée comme moins importante que celle mot à mot puisque le taux d'erreur pour l'appariement des synonymes reste assez élevé avec la ressource lexicale utilisée²². Les points de chaque couple lexical sont additionnés afin d'obtenir le score total et de décider si le couple peut être retenu dans les résultats finaux.

Pour les couples de la Figure 52, le traitement lexical ne révèle pas d'autres appariements que ceux déjà détectés. Ainsi, la correspondance avec Ballanche¹⁸²⁹ obtient 27 points (six relations forme~forme et une relation mixte²³), celle avec Cabaret-Dupaty¹⁸⁹⁷ obtient 20,5 points (une relation

²² Nous discutons cette question au IV.4b) « Amélioration du traitement des synonymes ».

²³ Il s'agit de la relation *regard~œil* qui obtient 2 points : le lemme *regard* figure parmi les synonymes d'*œil* et le synonyme *prunelle* est commun aux deux mots.

forme~forme, deux relations lemme~lemme, deux radical~radical et une synonyme~synonyme) et enfin celle avec Courmand1805 est à 10 points (une relation forme~forme et une radical~radical).

Nous calculons ensuite un ratio de ressemblance afin de pouvoir comparer les phrases de longueur et de complexité différentes. Le pourcentage du score obtenu est évalué par rapport au score maximal que le couple aurait si tous les mots forts de la phrase la plus courte entraient dans une relation mot à mot. Pour la relation entre Ballanche1809 et Ballanche1829, ce score lexical s'élève à 115 points (soit un ratio de 23,47 %) alors que pour la phrase la plus courte de Cabaret-Dupaty1897 il n'est que de 65 points (soit un ratio de 31,54 %).

L'application de cette méthode à notre échantillon d'évaluation suggère que le ratio de ressemblance lexicale est bien inférieur au score basé sur les noyaux des syntagmes. Si nous fixons le ratio nucléaire à 20 %, alors le seuil du rapport lexical minimal reste supérieur à 10 % ; pour le ratio à 30 %, la similarité lexicale est supérieure à 20 %, etc. Au sein de l'échantillon d'évaluation, le ratio lexical des 58 % des couples du type 2 ne dépasse pas 20 et pour 15 % il se situe entre 20 et 30. Seulement 4 % des couples du type 5 ne dépassent pas le seuil de 20 et 16 % sont entre 20 et 30. À ce stade, nous décidons de fixer un seuil minimal à 10 %.

Cela évite d'écarter des résultats pertinents, mais établis sur plusieurs phrases : les relations entre ce qui correspond à chaque phrase peuvent être plus faibles et risquent donc de ne pas atteindre les seuils imposés, alors que le couple détecté est pertinent dans sa totalité. Pour la correspondance compilée établie entre les phrases 17-21 de Ballanche1809 et 215-217 de Ballanche1829 (Tableau 84-Tableau 87) le ratio nucléaire s'élève à 42,6 % (23 noyaux alignés sur 54) et le ratio lexical à 37 % (106 points sur 285). Mais les entités communes ne se répartissent pas équitablement : la phrase 21, par exemple, a une entité commune avec la phrase 215, trois avec la 216 et deux avec la 217 (Figure 54).

CHOISIR LA
CORRESPONDANCE
ballanche1829, ph. 215-217

☒ Relations lexicales

- forme-forme
- lemme-lemme
- radical-radical
- synonyme-synonyme
- ou mixte

BALLANCHE1809, PHRASES 17-21
BALLANCHE1829, PHRASES 215-217

[17] Déjà¹ les ombres² devenaient³ moins opaques ; déjà un faible⁴ crépuscule⁵ détaché⁶ des rayons⁷ du soleil⁸, arrivait jusqu'aux yeux⁹ des deux nobles créatures qui s'avançaient vers la lumière¹⁰ du jour¹¹.

[18] Un instant encore, et elles échappaient à la puissance¹² du dieu des morts¹³ ; elles touchaient¹⁴ au seuil¹⁵ du séjour¹⁶ des vivants¹⁷.

[19] Mais, ô faiblesse d'un cœur¹⁸ qui aime !

[20] Orphée s'arrête¹⁹ pour écouter le soupir qui errait sur les lèvres d'Eurydice²⁰, pour prêter l'oreille au léger frôlement de ses vêtements aériens.

[21] Vaincu par cette puissance contre laquelle l'homme lutte en vain, il se retourne²¹ ; et, oubliant²² sa fatale promesse, il permet à son regard²³ d'interroger, à la faveur de la clarté naissante, les plis du voile²⁴ qui lui cachait le visage²⁵ de son épouse²⁶, de cette touchante victime qu'il avait tant pleurée.

[215] Déjà¹ un crépuscule⁵ douteux⁴ commençait à remplacer l'obscurité² immense ; un rayon de pâle lumière¹⁰, qui vint flotter autour de ses yeux⁹, lui fit publier²² le passé et l'avenir : ce n'était pas une lueur⁸ trompeuse, c'était bien un rayon⁷ détaché⁶ de la clarté du jour¹¹ ; la marque de son salut devint³ le signe de sa perte.

[216] Ainsi donc il touchait¹⁴ au seuil¹⁵ qui sépare l'empire¹² des morts¹³ du séjour¹⁶ des vivants¹⁷, lorsqu'il s'arrêta¹⁹ involontairement pour jeter un regard²³ d'amour¹⁸ sur l'épouse²⁶ qui lui était rendue²¹, hélas ! rendue et ravie à-la-fois.

[217] Si, du moins, il eût eu le temps de soulever le voile²⁴ qui retombait sur le visage²⁵ d'Eurydice²⁰ !

Figure 54. Correspondances lexicales entre Ballanche1809 (17-21) et Ballanche1829 (215-217)

Ainsi, l'exploration des correspondances, tant nucléaire que lexicale, permet, d'un côté, l'exclusion des relations fausses ou peu pertinentes et, de l'autre, l'intégration et la validation des réécritures effectuées sur de longs fragments, même en cas de reformulations complexes et non linéaires.

Ces manipulations fournissent aussi de précieuses informations pour l'analyse littéraire et linguistique des correspondances validées automatiquement. Les exemples des Figure 53 et Figure 54 révèlent principalement deux types de problèmes :

- Les périphrases des entités nommées ne sont appariées ni entre elles ni avec les noms propres de leur référent, même quand il s'agit d'Eurydice, qui est pourtant souvent évoquée. L'exploitation des attributs de ces éléments très fréquents permet d'éviter les appariements fautifs ou, au contraire, d'en détecter d'autres.
- Certains appariements de synonymes très polysémiques (principalement des verbes fréquents) sont pertinents pour certains de leurs emplois, mais pas pour toutes leurs constructions ni pour les valeurs sémantiques attestées dans notre corpus. Le verbe « retourner », employé pronominalement, par exemple, fait partie des mots-clés saillants de l'épisode de la seconde mort d'Eurydice, mais son sens dans ce contexte est bien précis et ne correspond pas aux synonymes les plus courants du verbe en langue générale (toutes constructions confondues), comme « rendre », « repasser » ou « rappeler ».

Ces difficultés peuvent être surmontées grâce à l'élaboration de lexiques adaptés à la spécificité du corpus couplés à des dictionnaires de langue générale (notamment de synonymes) et à des ressources polyvalentes (utilisées pour le repérage de chaînes de références). Ces lexiques sont progressivement enrichis grâce aux détections de correspondances validées²⁴.

²⁴ Nous discutons cette question au IV.4b) « Amélioration du traitement des synonymes ».

360

Enrichissement de la structure XML initiale²⁵

Pour faciliter l'analyse et la visualisation des résultats enrichis, nous enrichissons la structure XML initiale. Ainsi, les appariements (poly-)lexicaux validés sont intégrés dans le balisage des phrases. Pour l'extrait de la phrase 21 de Ballanche1809 (Figure 51), deux relations exactes sont établies avec le mot « épouse ». Elles sont signalées (Figure 55) à l'aide d'un élément <xr> enfant de <w> qui renseigne l'identifiant du mot correspondant (@corresp), le type de la relation détectée (@type) et le score obtenu par le couple (@cert). L'annotation des périphrases et des noms propres d'entités ajoute, elle, plusieurs correspondances à « son épouse », qui trouve des appariements pertinents dans quatre textes. Le lien fautif avec la périphrase pour la mère d'Eurydice (« l'épouse d'un Titan ») chez Ballanche1829 est remplacé par un lien avec « la vierge »²⁶ et une correspondance est ajoutée avec « sa chère Eurydice » de Cabaret-Dupaty1897.

```
<persName xml:id="0300021_50-51" corresp="eurydice" type="périphrase">
  <xr corresp="0400216_30-35" type="périphrase~périphrase" cert="5"/>
  <xr corresp="0401014_43-44" type="périphrase~périphrase" cert="5"/>
  <xr corresp="5505019_7" type="périphrase~entité-nommée" cert="5"/>
  <xr corresp="5502011_31-33" type="périphrase~entité-nommée" cert="5"/>
  <phr type="SN" select="épouse">
    <w xml:id="0300021_50" n="50" lemma="son"
      pos="DET:POS">son</w>
    <w xml:id="0300021_51" n="51" lemma="épouse" pos="NOM"
      sameAs="conjointe bourgeoise femme régulière">épouse</w>
    <xr corresp="0400216_31" type="forme~forme" cert="5"/>
  </phr>
</persName>
```

Figure 55. Annotation enrichie d'un extrait de la phrase 21 de Ballanche1809

En fin de traitement, le fichier XML de chaque texte est significativement enrichi et une base XML compile les correspondances valides détectées par TextPAIR et Tracer. Ce double enregistrement des résultats permet un large éventail de manipulation et d'analyses, tantôt spécifiques, concentrées sur un texte précis, tantôt transversales, incluant tout ou partie de texte. Avant de passer à l'explication détaillée de la dernière étape de notre traitement, à savoir l'annotation des entités nommées et leurs périphrases, nous proposons d'explorer plus en détail ce que nous obtenons pour les deux œuvres de Ballanche, Ballanche1809 et Ballanche1829, qui nous ont servi jusqu'alors pour la démonstration de notre développement.

²⁵ Voir le script [6_xmlisation_resultats.py](#) et son résultat, [6_resultat_final.xml](#).

²⁶ L'annotation des synonymes ne permet pas d'aligner *femme* et *vierge* au niveau lexical.

c) Exploitation des résultats enrichis et du réseau global des réécritures

Orphée est une figure récurrente chez Pierre-Simon de Ballanche, le mythe se modulant progressivement et devenant une constante créatrice, parfois implicite, parfois centrale. L'auteur se réfère au mythe orphique tantôt pour prôner la force civilisatrice de la musique (*Du sentiment considéré dans des rapports avec la littérature et les arts*, 1801²⁷), pour déplorer l'espoir perdu de la fidélité conjugale (« Sixième Fragment », 1809²⁸) et pour conter les exploits d'un Orphée Argonaute (*Antigone*, 1819²⁹). Enfin, ces motifs, réunis, étoffés et enrichis, illustrent la doctrine sociale palingénésique de l'auteur dans l'épopée *Orphée* (1827-1829³⁰).

Seulement « Sixième Fragment » (Ballanche1809) et *Orphée* (Ballanche1829), dont nous citons plusieurs phrases dans ce chapitre, font partie de notre corpus. Ce premier est un bref essai qui se réfère au mythe sans le parodier ni le moderniser. Malgré des reformulations et une poétisation assez importantes, nous avons vu que plusieurs relations sont détectées entre cette œuvre et des traductions des deux sources antiques (surtout celles de Virgile). Vingt ans plus tard, Ballanche réécrit des extraits de cet essai dans sa longue épopée (Ballanche1829), ce qui l'éloigne du texte de Virgile, et va jusqu'à modifier les emblèmes majeurs du mythe, comme la descente aux Enfers et la mort d'Eurydice³¹. Alors que, dans « Sixième Fragment », Eurydice symbolise, comme dans le mythe, la douceur du mariage et l'inspiration poétique, dans *Orphée*, elle devient beaucoup plus active, initiatrice et elle-même créatrice. L'observation de la place de Ballanche1829 au sein du réseau des correspondances détectées nous permet d'isoler et de relier des œuvres qui s'inscrivent dans le même esprit, notamment en valorisant un Orphée fondateur des cultes.

Par exemple, nous voyons que pour la phrase 21 de Ballanche1809, un nombre assez important de correspondances s'observe avec les différents textes du corpus. Nous avons déjà prôné l'utilité des graphes de correspondances pour mieux exploiter ce type des relations complexes³². Au sein de ceux que nous souhaitons proposer,³³ la taille d'un nœud renvoie au nombre de phrases qui

²⁷ Pierre-Simon Ballanche, *Du sentiment considéré dans ses rapports avec la littérature et les arts*, Lyon, Paris, Ballanche et Barret, Calixte Volland, 1801.

²⁸ Pierre-Simon Ballanche, « Sixième Fragment », *Œuvres de M. Ballanche*, Paris, J. Barbezat, 1809, p. 488-492.

²⁹ Pierre-Simon Ballanche, *Antigone*, Paris, Jules Didot, 1819.

³⁰ Pierre-Simon Ballanche, *Essais de palingénésie sociale. Orphée*, Paris, Jules Didot, 1827-1829.

³¹ La présence du mythe chez Ballanche a été étudiée avec des méthodes empiriques traditionnelles (voir notamment Brian Juden, « Particularités du mythe d'Orphée chez Ballanche », *loc. cit.*; William R. Paulson, « Fragment et autobiographie dans l'œuvre de Ballanche », *loc. cit.*). Notre démonstration s'appuie sur les résultats que nous obtenons pour les œuvres de cet auteur, mais n'aborde pas le sujet dans toute sa complexité.

³² Voir par exemple III.1c) « Établissement des graphes de relations » ou encore III.4c), sous-section « Accès au contexte trop restreint ».

³³ Ci-après, nous utilisons des visualisations générées à l'aide de l'interface conçue au sein du projet que nous présenterons en détail au chapitre IV.3 « Présentation de prototype d'une édition comparative numérique ».

composent l'extrait et l'épaisseur d'un lien reflète le ratio de proximité lexicale entre chaque couple (Figure 56).

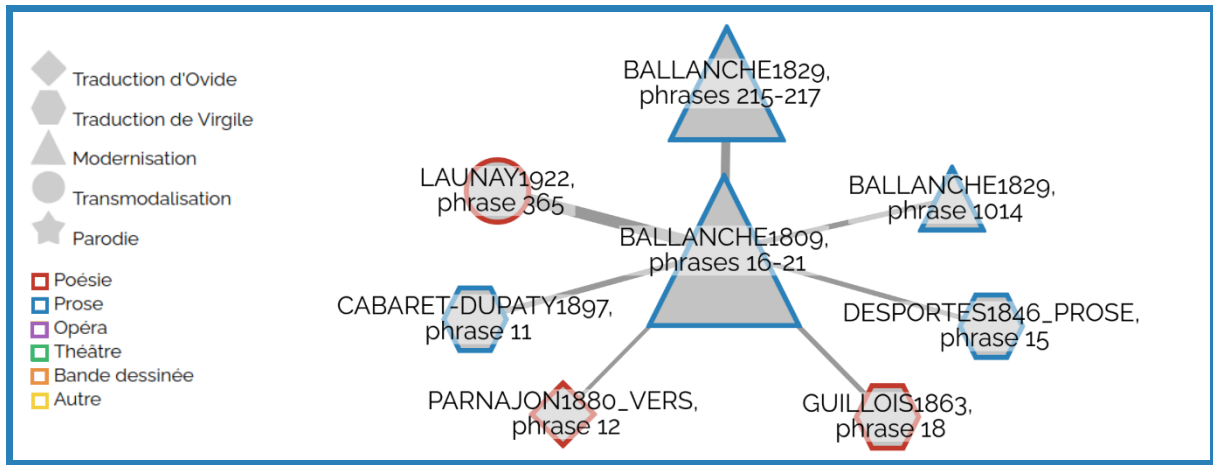


Figure 56. Graphe de correspondances de la phrase 21 de Ballanche1809.

La phrase 21 est intégrée dans un ensemble plus large, qui regroupe 6 phrases (de 16 à 21), grâce, d'une part, à la correspondance avec Ballanche1829 (Figure 54) et, d'autre part, à une correspondance des phrases 16 et 17 avec le poème *Orphée* (Launay1922). Sept relations au total ont été établies :

- trois avec des traductions de Virgile (Desportes1846-2, Guillois1863 et Cabaret-Dupaty1897) ;
- une avec une traduction d'Ovide (Parnajon1880-1) ;
- une avec la transmodalisation de Launay1922 ;
- deux autres avec des extraits très éloignés de Ballanche1829 : les phrases 215-217 d'un côté et la phrase 1014 de l'autre.

En effet, l'extrait 215-217 de Ballanche1829 fait partie d'un chant du fils du roi Évandre, qui raconte la version originelle du mythe, rectifiée ensuite par le poète Thamyris³⁴. Concernant Eurydice, cette rectification se retrouve au sein du second extrait (1 014), où elle meurt pour toujours au moment de l'accomplissement du rite de son mariage avec Orphée. Contrairement au récit de Virgile, cette disparition n'a pas lieu aux Enfers et n'est pas précédée d'une plainte de l'épouse³⁵. À la place, Eurydice s'excuse³⁶ d'avoir distrahit le poète dans sa mission de fondateur des cultes et

³⁴ « [N]ul ne peut mieux que Thamyris lui-même rectifier l'histoire que tu vas dire ; nul ne peut mieux nous instruire de ce que l'on sait sur ce génie sublime, à qui les dieux inspirèrent les lois de la société, celles du bien et du beau ».

³⁵ « Adieu. L'éternelle nuit m'enveloppe et m'entraîne, sans que je puisse te presser de mes mains défaillantes : hélas ! je ne suis plus à toi » (Cabaret-Dupaty1897).

³⁶ « Orphée, [...] j'ai voulu te distraire de tes grandes pensées, au lieu de me borner à les embellir, et à leur donner du charme ; et voilà que les dieux m'en punissent. [...] Continue d'obéir à tes hautes destinées, que je voulais concentrer sur moi seule ; marche à ton noble but, sans te détourner ni à droite ni à gauche, et que la douleur ne fasse pas ce qu'allaient faire les condescendances de l'amour ».

l'encouragement à la poursuivre, malgré la douleur de la perte qui devient alors initiatique³⁷. Si nous nous penchons sur la totalité des correspondances détectées pour Ballanche1829, nous nous rendons compte qu'il s'agit de l'unique relation permettant de lier les deux descriptions de la mort d'Eurydice entre elles et avec les sources anciennes (Figure 57).

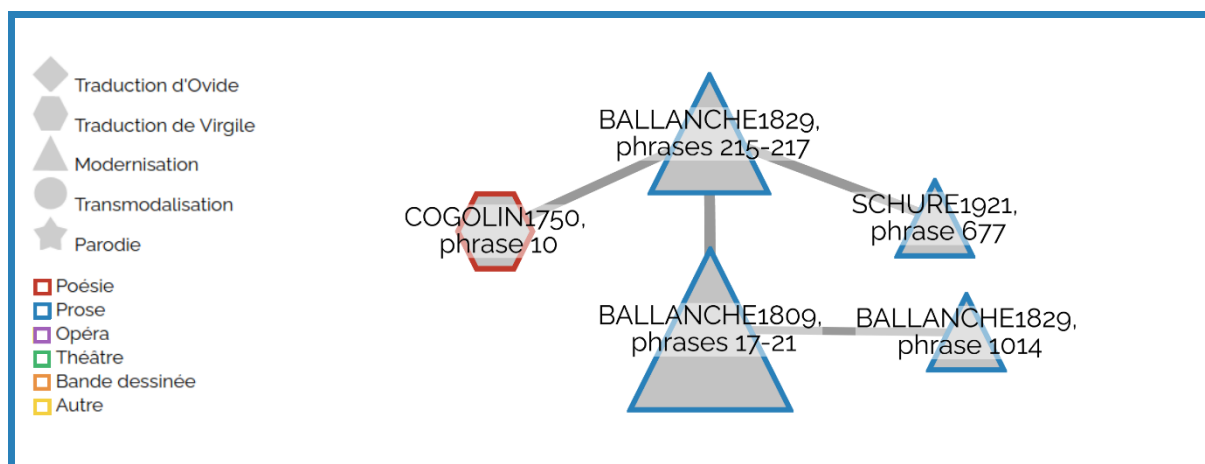


Figure 57. Graphe de correspondances des phrases 215-217 de 1014 de Ballanche1829

Un réseau complexe peut en revanche être observé à proximité de l'extrait 215-217 (Figure 58) : pour les phrases 219-220, 10 correspondances ont été trouvées avec les traductions de Virgile et une avec le poème Launay1922.

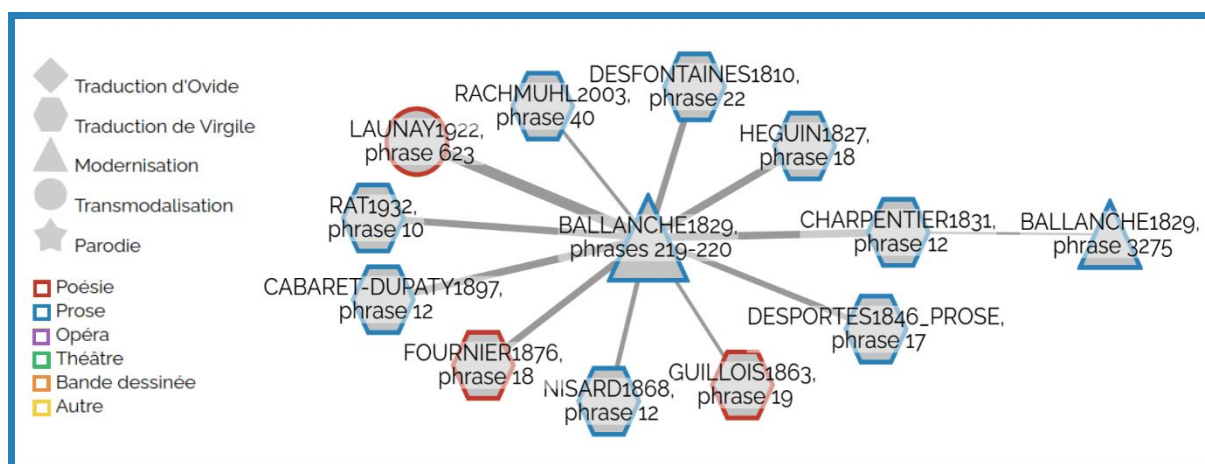


Figure 58. Graphe de correspondances des phrases 219-220 de Ballanche1829

³⁷ Alors que Juden (« Particularités du mythe d'Orphée chez Ballanche », *loc. cit.*) accepte les interprétations découlant du monologue de la femme, il y ajoute une symbolique ésotérique : en accomplissant le rite du mariage, Eurydice s'initie sans le savoir au mystère de l'intelligence divine, mais celui-ci est trop puissant pour qu'elle puisse le supporter. En le plaçant dans le contexte plus large de l'épopée, Juden relève également une parabole de l'ascension pénible des peuples, « sortant de l'enfer de l'ignorance sociale » (p. 148). Notre travail ne dévoile pas encore de relations permettant d'appuyer ces deux axes d'analyse. En revanche, les correspondances détectées avec la partie interprétative de Ballanche1809 apportent de nouvelles pistes sur le rôle de la mort d'Eurydice (voir plus loin).

L'essai Ballanche1809 s'arrête à la seconde mort d'Eurydice, il n'y a donc pas d'autres relations avec cette partie de l'épopée (Ballanche1829). Mais il y a beaucoup de correspondances avec cette dernière et les traductions dans la suite de la description du mythe. Toutefois, l'auteur ne procède pas de la même manière pour la deuxième description de la mort d'Eurydice, beaucoup plus allusive, ce qui empêche la détection de correspondances autres que celle avec Ballanche1809.

Enfin, des éléments de la conclusion de l'essai, dans laquelle l'interprétation du mythe est développée, peuvent également être mis en correspondance avec deux extraits de Ballanche1829 :

- la phrase 28 de Ballanche1809, « Il y a sur la terre comme un long gémissement qui se traîne de génération en génération, depuis les premiers morts jusqu'à nous », est apparée avec la phrase 2 446 de Ballanche1829, « L'homme n'a qu'une voix ; [...] et cette voix qui n'est, le plus souvent, que celle du gémissement, passe de génération en génération, et retentit dans les siècles » ;
- la phrase 34 de Ballanche1809, « Son existence fût demeurée incomplète ; et il aurait, au lieu de la douleur dont il fit une épreuve si cruelle, connu cette autre douleur qui n'a pas moins d'intensité, la solitude de l'âme », est, elle, en lien avec la 1 998 de Ballanche1829, « Cette pensée continue de le suivre dans sa destinée future ; elle accompagne l'âme voyageuse [...] accomplissant ainsi, dans des existences successives, la série d'épreuves par lesquelles il faut qu'elle passe avant de pouvoir être introduite dans son état définitif ».

Pour Ballanche1809, la conclusion renvoie clairement à l'épisode de la mort d'Eurydice, mais, dans l'épopée, les correspondances font partie d'une description de cérémonie funèbre en Égypte ancienne et de la métaphore de la vie de l'homme³⁸. Cette observation permet de relativiser ou d'appuyer quelques analyses traditionnelles du « Sixième Fragment ». Paulson procède à l'interprétation biographique de la rupture amoureuse de Ballanche, en comparant la remontée des Enfers au « temps bref des illusions de l'amour »³⁹. Pour Juden, la douleur de la mort d'Eurydice « résume le sort de l'humanité, séparée de l'unité »⁴⁰. La prise en compte de la partie interprétative de l'essai dans le contexte des correspondances trouvées avec l'épopée confirme que la figure d'Eurydice

³⁸ Il s'agit d'une histoire de Phoenix, racontée à Thamyras par un des prêtres d'Isis. Uranus envoie ses fils trouver l'oiseau qui est en train de construire son tombeau pour renaître de ses cendres. Pour arriver jusqu'à lui, ils doivent passer une série d'épreuves, dont celle de l'énigme : « Y a-t-il un être qui marche tantôt sur quatre pieds, tantôt sur deux, tantôt sur trois, et qui soit d'autant moins assuré dans sa démarche, qu'il semble être mieux appuyé ? [...] » (p. 253). L'extrait de la correspondance fait partie de la réponse formulée. La proximité de cette énigme, et de son contexte, et de celle posée par le Sphinx à Œdipe prouve la complexité des réseaux intertextuels observables au sein de notre corpus.

³⁹ William R. Paulson, « Fragment et autobiographie dans l'œuvre de Ballanche », *loc. cit.*, p. 25.

⁴⁰ Brian Juden, « Particularités du mythe d'Orphée chez Ballanche », *loc. cit.*, p. 142.

comme le deuil de sa mort dépassent la symbolique amoureuse. La conclusion fait écho, dans l'épopée, à la description de la mort et de l'errance de l'âme à la recherche du salut, mais aussi à l'expression de la condition des humains, qui transmettent la souffrance avec le savoir. Le « Sixième Fragment » est majoritairement analysé dans le contexte des autres essais de Ballanche et de sa biographie, ou considéré comme une préfiguration de l'épopée, manquant peut-être encore de la profondeur métaphysique que l'auteur donnera au mythe 20 ans plus tard. Or, comme nous espérons l'avoir démontré, la proximité effective entre les deux textes, matérialisée par des allusions et d'importantes reformulations, prouve que, dans l'essai, la réflexion de Ballanche autour du mythe d'Orphée est déjà bien entamée.

La vue globale des résultats, enregistrés dans la base XML, se révèle, quant à elle, particulièrement utile pour l'observation des relations que les œuvres entretiennent entre elles. Pour Ballanche¹⁸²⁹, par exemple, des correspondances ont été détectées avec 65 textes du corpus, dont 34 réécritures. Toutefois, ce texte est très long et le récit de l'amour avec Eurydice y est ponctuel et rédigé de manière figurée. Ces facteurs importent dans l'analyse de la place de Ballanche¹⁸²⁹ dans le réseau des relations. Ainsi, au lieu de décompter le nombre des correspondances détectées pour chaque œuvre, il est plus pertinent de comparer les différents scores obtenus et leurs moyennes. Par exemple, huit correspondances seulement ont été trouvées avec Ballanche¹⁸⁰⁹, mais la moyenne du ratio lexical s'élève à 27 %, le plaçant ainsi dans les 10 textes les plus proches, alors que celle des 42 correspondances avec Launay¹⁹²² n'est que de 17,45 %, moins que la moyenne générale de la totalité des résultats (19 %).

Parmi les réécritures les plus proches, nous trouvons deux textes en prose, Esquiros¹⁸⁴⁹ (*Histoire des amants célèbres de l'antiquité*) et Schuré¹⁸⁸⁹ (*Les Grands Initiés. Esquisse de l'histoire secrète des religions*). Pour eux, la correspondance avec Ballanche¹⁸²⁹ compte le plus de correspondances (16 et 12) et se place systématiquement dans les 10 textes les plus proches, peu importe le critère de comparaison. Ces œuvres insistent, elles aussi, sur le rôle d'Orphée comme fondateur des cultes. Chez Esquiros¹⁸⁴⁹, la catabase est remplacée par des épreuves initiatiques, auxquelles Orphée est soumis par des prêtres de l'Égypte. Eurydice, mordue par le serpent, est seulement endormie. Schuré¹⁸⁸⁹, lui, met en scène un Orphée déjà initié, transmettant son savoir à des disciples. Dans cette réécriture, Eurydice meurt empoisonnée par Aglaonice de Thessalie, une bacchante, prêtresse d'Hécate et « éprise pour [Eurydice] d'une envie perverse, d'un amour effréné, maléfique ». Dans la pièce de théâtre *Orphée* (Cocteau¹⁹²⁷), Aglaonice endosse un rôle comparable, mais fortement caricaturé⁴¹.

⁴¹ Boulomié relève l'inspiration de Cocteau pour évoquer des traditions qui voient en Eurydice l'alliée involontaire des bacchantes (« La Résurgence du mythe d'Eurydice et ses métamorphoses dans l'œuvre d'Anouilh, de Pascal Quignard, d'Henri Bosco, de Marguerite Yourcenar, de Michèle Sarde, et Jean Loup Trassard », *loc. cit.*), alors que Kushner considère Schuré comme un point de départ des réécritures symboliques du mythe (Eva Kushner, *Le mythe d'Orphée dans la littérature française contemporaine*, *op. cit.*, p. 158.).

Des relations très fortes (dignes d'une reprise : six correspondances avec un ratio moyen de 78 %, Figure 59) ont été détectées entre Schuré1889 et Manjock1981 (*Orphée d'Afrique, théâtre-rituel*), une adaptation du roman *Orphée-Dafric* (Liking1981) qui transfère le mythe sur le continent africain, dans la continuité d'un mythe plus récent, celui de l'« Orphée noir »⁴².

CHOISIR LA CORRESPONDANCE segalen1921.ph.886 v	SCHURE1889	MANJOCK1981
☒ Relations lexicales		
forme-forme		
lemme-lemme		
radical-radical		
synonyme-synonyme		
ou mixte		
	[111] Il s'appelait ¹ maintenant Orphée ² ou Arpha, ce qui veut dire : celui qui guérit ³ par la lumière ⁴ .	[14] Il s'appelle ¹ « la lumière ⁴ qui guérit ³ », il s'appelle Orphée ² .
	[212] Mon enfant, reste dans les sentiers ⁶ fleuris ⁷ de la plaine ⁸ et ne cherche pas au-delà.	[625] Je l'appelle ¹ Orphée ² , la lumière ⁴ qui guérit ³ .
	[226] Car sache que la lumière ⁹ des Dieux ¹⁰ épouvante ¹¹ les faibles ¹² et tue ¹³ les profanateurs ¹⁴ .	[506] Je vis ⁵ dans les sentiers ⁶ fleuris ⁷ de la plaine ⁸ et je ne vois pas les orages sur la montagne.
	[348] Venez ¹⁵ boire ¹⁶ la lumière ¹⁷ du temple, vous qui sortez ¹⁸ de la nuit ¹⁹ , mystes, femmes, initiés.	[255] La lumière ⁹ des dieux ¹⁰ épouvante ¹¹ les faibles ¹² et tue ¹³ les profanateurs ¹⁴ .
	[353] L'étincelle ¹⁹ divine ²⁰ qui nous guide sur terre est en nous ; elle devient ²¹ flambeau dans le temple, étoile ²² dans le ciel ²³ .	[446] Venez ¹⁵ boire ¹⁶ à la lumière ¹⁷ du jour, vous qui sortez ¹⁸ de la nuit ¹⁹ .
	[361] Car dans l'extase ²⁴ des Mystères ²⁵ , chacun ²⁶ voit ²⁷ jusqu'au fond ²⁸ l'âme ²⁹ de l'autre.	[447] L'étincelle ¹⁹ divine ²⁰ devient ²¹ étoile ²² dans le ciel ²³ .
		[448] Dans l'extase ²⁴ des Mystères ²⁵ , chacun ²⁶ voit ²⁷ jusqu'au fond ²⁸ des âmes ²⁹ .

Figure 59. Relations fortes détectées entre Schuré1889 et Manjock1981

Porra constate chez Liking et Ma Njock l'influence conjointe de textes européens et de la réalité africaine, et note l'importance de l'aspect initiatique, focalisé autour d'une lecture mythique égyptienne⁴³, mais ne relève pas de relation forte avec Schuré1889 (d'ailleurs, les correspondances entre Schuré1889 et Liking1981 ne sont pas significatives). Un examen plus approfondi s'impose pour confirmer l'absence du lien dégagé par les recherches traditionnelles, mais elle est plausible compte tenu des convictions de Schuré⁴⁴.

Il est satisfaisant d'observer le réseau de relations, qui se construit à partir de correspondances détectées, le plus souvent en confirmant ce qui a déjà pu être observé, parfois en révélant de nouveaux liens. Par exemple, Brunel relève plusieurs épisodes inspirés par Schuré1889 dans la pièce de

⁴² C'est le titre de la préface de Jean-Paul Sartre à l'*Anthologie de la nouvelle poésie nègre et malgache de langue française* (1948). À ce sujet, voir par exemple János Riesz, « « Orphée Noir » – « Schwarzer Orpheus » – « Black Orpheus » : Quand les marges se mettent en marche vers un centre commun », *Revue de littérature comparée*, vol. 314, n° 2, Klincksieck, 2005, p. 161-177 ; Kathleen Gyssels, « Sartre postcolonial ? Relire Orphée noir plus d'un demi-siècle après », *Cahiers d'études africaines*, vol. 45, n° 179-180, Éditions de l'École des hautes études en sciences sociales, décembre 2005, p. 631-650, en ligne, <<http://journals.openedition.org/etudesafriaines/14952>>.

⁴³ Véronique Porra, « Sur quelques Orphée noirs. Reproduction, adaptation et hybridation du mythe d'Orphée en contexte post(-)colonial », *loc. cit.*, paragr. 19-27.

⁴⁴ « Autour du 1900, une nouvelle phrase de gobinisme, via l'Allemagne, où la *Gobineau-Vereinigung* de Ludwig Schemann (1825-1938) est fondée en 1894, se développe pour des usages d'élitisme aristocratique chez Robert Dreyfus (1873-1939) ou chez le wagnérien Édouard Schuré (1841-1929), et nourrit l'anthropologie raciste de Vacher de Lapouge (1854-1936) » (Dominique Jarrassé, « Trois gouttes d'art nègre. Gobinisme et métissage en histoire de l'art », *Les actes de colloques du musée du quai Branly Jacques Chirac*, n° 1, 2009, en ligne, <<http://journals.openedition.org/actesbranly/96>>).

théâtre symbolique *Orphée-Roi* (Segalen1921)⁴⁵. Au sein de notre corpus, 20 correspondances ont été retenues. La plupart consistent en la cooccurrence des mots *époux*, *épouse* et *homme*⁴⁶ ou en la proximité du décor qui se situe « au bord du fleuve »⁴⁷, mais deux couples concordent avec les analyses de Brunel en dévoilant une réelle inspiration de Segalen1921 par Schuré1889 (Figure 60).

CHOISIR LA CORRESPONDANCE
segalen1921, ph. 886

☒ Relations lexicales

- forme-forme
- lemme-lemme
- radical-radical
- synonyme-synonyme
- ou mixte

SCHURE1889, PHRASE 483
Alors un immense¹ cri d'amour parti² du ciel et de la terre, comme³ si le frisson sacré des Dieux en passant sur la grande lyre⁴ voulait en déchirer⁵ toutes les cordes⁶, en égrener les sons à tous les vents.

SCHURE1889, PHRASE 619
J'entendis seulement⁷ comme un son de corde⁸ qui se déchire⁹ ; puis une voix faible comme¹⁰ un souffle, triste comme un baiser d'adieu, murmura : – Orphée¹¹ !

SEGALEN1921, PHRASE 886
D'un sursaut fulgurant, il déchire⁵ le réseau de la Lyre⁴ ; et le crèvement des cordes⁶ et leurs cinglements trament l'Antre. (qui se fend comme³ un fruit) de rayons faisant au plus profond¹ de l'épaisseur une échappée radieuse, par où, d'un seul bond, s'évade et disparaît² Orphée.

SEGALEN1921, PHRASE 886
D'un sursaut fulgurant, il déchire⁹ le réseau de la Lyre ; et le crèvement des cordes⁸ et leurs cinglements trament l'Antre. (qui se fend comme¹⁰ un fruit) de rayons faisant au plus profond de l'épaisseur une échappée radieuse, par où, d'un seul⁷ bond, s'évade et disparaît Orphée¹¹.

Figure 60. Deux correspondances significatives détectées entre Schuré1889 et Segalen1921

Brunel cite également la phrase 886 de Segalen1921 qui décrit le combat entre Orphée et La Ménade en transe. Orphée utilise sa lyre pour se protéger et la Ménade meurt écrasé par l'Antre. Cependant, le chercheur ne mentionne pas les correspondances détectées, mais met cette phrase en relation avec la précédente œuvre de Segalen, *Dans un monde sonore*⁴⁸, avant de parler du miracle de la lyre, sa résurrection, qu'il met en relation avec le miracle de la tête chez Schuré1886 :

[...] il expira. Penchée sur son cadavre, la magicienne de Thessalie, dont le visage ressemblait maintenant à celui de Tisiphône, épiait avec une joie sauvage le dernier souffle du prophète et s'apprêtait à tirer un oracle de sa victime. Mais quel fut l'effroi de la Thessaliennne, en voyant cette tête cadavéreuse se ranimer à la lueur flottante de la torche, une pâle rougeur se répandre sur le visage du mort, ses yeux se rouvrir tout grands et un regard profond, doux et terrible se fixer sur elle... tandis qu'une voix

⁴⁵ Pierre Brunel, « *Orphée-roi* de Victor Segalen, ou le miracle de la lyre », *Mythocritique : Théorie et parcours*, Grenoble, UGA Éditions, 2016, p. 165-174, en ligne, <<https://doi.org/10.4000/books.ugaeditions.6498>>.

⁴⁶ Comme « Les mystères de l'Époux et de l'Épouse ne sont dévoilés qu'aux hommes divins » (Schuré1889) / « Pourquoi réclamer qu'il te traite en épouse, Lui, qui est autre qu'un homme, et plus Maître que tous les époux ? » (Segalen1921).

⁴⁷ Comme « Tout redevint noir ; les rochers reprirent leurs formes menaçantes : et des lumières errantes brillèrent de tous côtés sous l'épaisseur des arbres, au bord du fleuve et dans les profondeurs de la vallée. » (Schuré1889) est aligné à trois phrases de Segalen1921 : « Il passe là-bas toutes les nuits au bord du fleuve, en chantant, toujours suivi d'elle. », « Plus loin que la forêt hostile, plus loin que les rumeurs des hommes, tout au bord du Fleuve coulant, marche » et « Je l'enviais, là-bas, toutes les nuits au bord du Fleuve, je veux aussi... Chante ! » ?

⁴⁸ « Segalen avait déjà utilisé cette idée pour sa nouvelle *Dans un monde sonore*, et dans un climat de misogynie vaguement analogue. Orphée, déçu par Eurydice, s'apprêtait à fuir sans elle les Enfers : "d'un coup de voix, il déchira la trame de sa lyre : la corne ployée le frappa dans la poitrine, et les fils, en cassant, mordirent ses poignets et ses ongles" » (Pierre Brunel, « *Orphée-roi* de Victor Segalen, ou le miracle de la lyre », *op. cit.*, paragr. 19.).

étrange – la voix d’Orphée – s’échappait une fois encore de ces lèvres frémissantes pour prononcer distinctement ces trois syllabes mélodieuses et vengeresses : Eurydice !⁴⁹

La démonstration analytique de Brunel est bien plus complexe que la mise en relation des extraits correspondants, ce qui montre bien les limites de la détection automatique. En effet, une seule relation a été détectée entre ce long extrait et un segment de Segalen¹⁹²¹ : « Tu prononces des mots sans harmonie... Le son de tes lèvres est vide ». Il fait partie d’une réplique d’Orphée face à la première apparition de La Ménade. Les deux entités communes, *lèvres* et *prononcer~prononces* ne sont pas suffisantes pour considérer cette correspondance comme pertinente.

De l’autre côté, les extraits détectés automatiquement pour la phrase 868 de Segalen¹⁹²¹ ouvrent d’autres pistes d’interprétation. La phrase 619 décrit la deuxième mort d’Eurydice, qui advient chez Schuré¹⁸⁸⁹ dans le rêve d’Orphée⁵⁰ et provoque l’initiation du poète :

Ce nom donné par une âme avait changé mon être. Je sentis passer en moi le frisson sacré d’un immense désir et le pouvoir d’un amour surhumain. Eurydice vivante m’eût donné l’ivresse du bonheur ; Eurydice morte me fit trouver la Vérité.

Cela permet d’orienter l’interprétation non pas vers le miracle de la tête d’Orphée, mais plutôt celui de la (seconde) mort d’Eurydice. Ce n’est d’ailleurs pas la seule parallèle entre Eurydice et la Lyre que nous observons dans ce texte⁵¹.

L’autre extrait de Schuré¹⁸⁸⁹ aligné à la phrase 868 de Segalen¹⁹²¹ (phrase 483) fait, lui aussi, partie d’un songe initiateur. Cette fois-ci, c’est Orphée qui, à l’aide de sa Lyre, plonge son disciple dans un « sommeil magique » :

Et maintenant j’évoquerai *mes* Dieux, afin qu’ils t’apparaissent vivants et qu’ils te montrent, dans une vision prophétique, le mystique hyménée que je prépare au monde et que verront les initiés.

Couche-toi à l’abri de cette roche. Ne crains rien. Un sommeil magique va fermer tes paupières, tu trembleras d’abord et tu verras des choses terribles ; mais ensuite, une lumière délicieuse, une félicité inconnue inondera tes sens et ton être.

Le segment aligné à la destruction de la Lyre chez Segalen¹⁹²¹ correspond à l’union charnelle de Dionysos et Perséphone qui clôt le rêve du disciple d’Orphée. En effet, c’est Dionysos, présenté comme le (deuxième) époux de la déesse qui vient à sa rencontre de la déesse et la réveille à la vie :

O Dionysos ! dit Perséphone, Esprit divin, Verbe de Jupiter, Lumière céleste qui resplendit sous la forme de l’homme ! chaque fois que tu me réveilles, je crois vivre pour

⁴⁹ Édouard Schuré, « Livre V. Orphée (Les mystères de Dionysos) », *Les grands initiés : esquisse de l’histoire secrète des religions*, Paris, Perrin, 1889, p. 314. Nous citons d’après Pierre Brunel, « *Orphée-roi* de Victor Segalen, ou le miracle de la lyre », *op. cit.*, paragr. 25.

⁵⁰ Voici le contexte plus large de l’extrait : « Trois fois je voulus la saisir ; trois fois elle s’évanouit dans mes bras comme une ombre. J’entendis seulement comme un son de corde qui se déchire ; puis une voix faible comme un souffle, triste comme un baiser d’adieu, murmura : – Orphée ! À cette voix je m’éveillai. ».

⁵¹ Voir le prochain chapitre, IV.2 « Annotation des entités nommées et leurs périphrases », et notamment la dernière section, c) « Exploitation des données obtenues ».

la première fois ; les mondes renaissent dans mon ressouvenir ; le passé, le futur redevient l'immortel présent ; et je sens dans mon cœur rayonner l'univers !

Une partie de ce segment est alignée à une autre phrase de Segalen¹⁹²¹, « Il resplendit sous la lumière » par laquelle une Ménade décrit le Roi Orphée et qui précède ces didascalies :

Plus loin que la forêt hostile, plus loin que les rumeurs des hommes, tout au bord du Fleuve coulant marche LE ROI ORPHÉE sous des habits dynastiques. Sa LYRE qu'il dresse haute devant lui, ruisselle de lueurs radieuses, et les cordes nombreuses pleines de son, s'irisent et chatoient.

La proximité entre Orphée et Dionysos-Zagreus, dévoilé par ces deux correspondances, est très présente dans les travaux littéraires sur le mythe orphique, par exemple dans l'ouvrage *Les Visages d'Orphée* de Béague et al. :

Sa mort est décrite sous la forme d'un démembrement du héros (*diasparagmos*). Les différentes parties de son corps sont emportées jusqu'à la mer par le fleuve Hèbre et un sort particulier est généralement réservé à sa tête et à sa lyre. Ce démembrement prend toute sa signification religieuse dans les récits où, présenté comme le fondateur des mystères de Dionysos-Zagreus, il subit le même sort que lui. Selon certaines versions d'ailleurs, il aurait aussi fondé avec Dionysos les mystères d'Éleusis.⁵²

Les auteurs soulignent aussi que les deux sont des héros civilisateurs (initiateurs), transgresseurs (transgression de séparation, plus particulièrement la descente aux Enfers pour ramener une femme), amoureux, tués (démembrés) par des femmes. Enfin, le non-respect de l'opposition théologique entre Apollon et Dionysos permet également de souligner les qualités médiatrices d'Orphée. Il est satisfaisant de constater que les résultats de la détection automatique des correspondances rendent également compte de ce parallèle.

Sans écarter l'interprétation de Brunel, qui rapproche le mystère de la lyre chez Segalen¹⁹²¹ au mystère de la tête d'Orphée chez Schuré¹⁸⁸⁹, notre traitement permet également d'ouvrir et explorer deux autres parallèles : celle entre la lyre et l'Eurydice et celle entre Orphée et Dionysos. Enfin, les correspondances détectées ne laissent aucun doute quant au pouvoir initiatique de la lyre d'Orphée dans les deux œuvres. Nous n'ambitionnons pas de supplanter les analyses empiriques des chercheurs traditionnels qui, comme Brunel, comparent deux textes en prenant en compte la totalité des œuvres d'un auteur, y ajoutent leur sensibilité littéraire et trouvent parfois un sens subtil même à partir d'un mot qui semble, à première vue, insignifiant. La plus-value d'une recherche appuyée sur les outils numériques réside, à notre sens, dans leur capacité de traiter simultanément de larges ensembles de textes et d'attirer l'attention du chercheur sur des phénomènes peu visibles (ou peu attendus, comme le rapprochement entre Schuré¹⁸⁸⁹ et Manjock¹⁹⁸¹).

⁵² Annick Béague et al., *Les visages d'Orphée*, Villeneuve d'Ascq, Presses universitaires du Septentrion, coll. « Savoirs Mieux », 2019, paragr. 25.

Conclusion du chapitre

En partant d'une œuvre particulière, l'essai de Ballanche, et des correspondances trouvées entre elle et d'autres textes du corpus, nous avons analysé les résultats des détections de correspondances opérées par TextPAIR et Tracer, montré la finesse et l'importance des post-traitements et permis de confronter les relations dégagées et interprétées aux analyses de facture classique du *topos* orphique. Comme les recherches traditionnelles, celles soutenues par des outils informatiques exigent un travail minutieux avant d'obtenir des résultats probants. Les deux méthodes ne sont pas concurrentes, elles seront certainement d'autant plus efficaces qu'elles seront combinées.

Dans le cadre de notre projet, nous cherchons à établir des pratiques communes pour la préparation des éditions numériques en sciences humaines. À cette fin, nous nous attachons à produire des données interopérables en libre accès et de la documentation facilitant la reproductibilité de notre développement. De plus, nous proposons une méthode de traitement automatisée qui génère des fichiers enrichis balisés selon la norme XML-TEI. Néanmoins, nous ne pouvons pas prétendre que notre traitement peut être généralisé facilement. Son adaptation à différents corpus peut nécessiter des ajustements pour améliorer la qualité des résultats. Il faudra également beaucoup de travail et de temps pour obtenir les premiers résultats significatifs. Pour certains textes, aucune relation pertinente ne sera trouvée. Néanmoins, tout cela semble inhérent à la recherche littéraire, qu'elle soit traditionnelle et empirique ou soutenue par des techniques computationnelles.

En effet, si la possibilité de détecter automatiquement des relations intertextuelles est passionnante, pour la recherche littéraire et linguistique et pour l'édition (numérique), les détections fautives et les problèmes de reconnaissance inhérents confirment la nécessité de combiner des traitements (semi)-automatique et des méthodes non informatiquement outillées. Une connaissance approfondie du corpus est par ailleurs nécessaire, tant pour analyser les résultats obtenus que pour les améliorer et les enrichir, afin que les traitements réitérés des mêmes textes bénéficient des premières passes et deviennent plus fructueux et plus aptes à dévoiler de nouvelles relations, de plus en plus fines et peu visibles. Nous le constatons davantage de la dernière étape de post-traitement, l'annotation des entités nommées et leurs périphrases, à laquelle nous consacrons le chapitre suivant.

2. Annotation des entités nommées et leurs périphrases¹

Outre la détection des relations syntagmatiques et lexicales, il nous importe en effet d'analyser plus finement la manière dont les reformulations sont construites. Les périphrases² sont de précieux indices au sein du corpus. Dans le sous-corpus comique, par exemple, ils montrent bien la manière dont la langue véhicule l'humour et le rire et permettent de mieux déterminer ce qui caractérise les dépréciations de la trame principale du mythe (« Pluton » / « Prince enfumé », « sa femme » / « cette femelle », « descendre aux Enfers » / « aller au Diable »).

Nous constatons que, alors que le réseau des personnages est très restreint dans les sources anciennes du mythe, il est significativement enrichi dans les réécritures postérieures. Dans ces dernières, nous trouvons des dieux, des héros, des lieux, des mythologies grecques et romaines autant que d'autres horizons culturels (*Babylone, Isis*), auxquels s'ajoutent quelques noms propres réels ou fictifs (*Bonaparte, Coquardeau*) parfois très modernes (*Aspivénin, Mohammed Ali*). Maurice Grevisse définit le nom propre comme « celui qui ne peut s'appliquer qu'à un seul objet ou à une catégorie d'êtres ou d'objets pris en particulier ; il individualise l'être, l'objet ou la catégorie qu'il désigne »³. Mais Roland Eluerd remarque « qu'il est difficile de tenir pour de vrais noms propres les noms de peuples et d'habitants (les entités ne sont pas uniques) ou de voir un nom commun dans *La lune se levait* »⁴. Les deux auteurs s'accordent sur le fait que la frontière entre le nom propre et le nom commun n'est pas fixe et reste souvent difficile à déterminer. Nous le constatons dans notre corpus, au sein duquel les noms propres occupent une place stratégique. Leur stabilité référentielle permet de renvoyer à une connaissance commune partagée, même à travers les époques : les noms des dieux « ne valent, *a priori*, que pour leur référent. Simplement, sur ces référents particuliers sont venues s'ajouter les valeurs de sens léguées par la légende »⁵. Cela implique que le lecteur est capable d'identifier le référent malgré la variation du signe, qui est fréquent dans notre corpus : les variantes de graphies sont nombreuses (*Eurydice, Euridice, Erudice, Eridice*) ; la relation particulière entre les mythologies grecque et romaine fait différer certaines dénominations (*Hadès, Pluton*) ; et l'emploi

¹ Ce chapitre reprend et complète notre article (co-signé avec N. Gasiglia) « Focus sur les traitements linguistiques d'un projet d'humanités numériques. Édition et cartographie hypertextuelle de réécritures », *loc. cit.*

² Nous considérons comme périphrase une « figure dans laquelle on substitue au [...] nom propre [...] une expression imagée ou descriptive qui le définit ou l'évoque » (*s.v. périphrase, Trésor de la langue française*). D'autres termes ont été proposés dans les travaux sur l'onomastique (antonomase, surnom, dénomination propre...). Voir par exemple Michel Murat, « La périphrase : remarques autour d'une figure », *L'information grammaticale*, vol. 13, n° 1, 1982, p. 38-40, en ligne, <https://www.persee.fr/doc/igram_0222-9838_1982_num_13_1_2379> ; Laura Calabrese-Steimberg, « Nom propre et dénomination événementielle : quelles différences en langue et en discours ? », *Corela. Cognition, représentation, langage*, n° 7-1, Université de Poitiers, juin 2009, en ligne, <<http://journals.openedition.org/corela/173>> ; Nicolas Laurent, « Le nom propre dans les Lettres de l'année 1671 de Madame de Sévigné », *Séminaire Madame de Sévigné*, n° 1, « Séminaires de recherche », 2013, p. 13, en ligne, <<http://publis-shs.univ-rouen.fr/ceredi/index.php?id=519>>.

³ Maurice Grevisse, *Le Bon Usage*, Paris, Duculot-Hatier, 1980, p. 224.

⁴ Roland Eluerd, *Grammaire descriptive de la langue française*, Paris, Armand Collin, 2008 [2002], p. 30.

⁵ *Ibid.*

de périphrases descriptives, parfois usuelles, comme les épithètes homériques, se fait souvent sans que le nom propre figure dans le contexte, elles assurent alors seules la dénomination, y compris dans les parodies les plus éloignées des textes sources.

Dans ce chapitre, nous abordons la dernière étape d'enrichissement des résultats des logiciels TextPAIR et Tracer en nous concentrant sur les entités nommées et leurs périphrases. Après un rapide état de lieux de la détection automatique des entités nommées et la description des particularités de ce que nous rencontrons dans notre corpus, nous présentons notre algorithme à base de règles permettant d'extraire et d'identifier ces éléments. Nous proposons enfin un exemple d'exploitation des données ainsi obtenues en analysant différentes occurrences des mentions d'Eurydice dans le corpus.

a) Détection automatique des entités nommées

Si les connaissances culturelles du lecteur lui permettent d'assigner un référent unique aux différentes graphies et périphrases d'une entité, leur appariement (semi-)automatique s'avère être un défi de taille.

Depuis l'apparition du concept d'entité nommée dans les années 1990, la détection et la classification des noms de personnes, organisations ou localisations géographiques font l'objet de nombreux développements informatiques et talistes. Toutefois, relativement peu de tentatives pour définir et pour proposer une typologie des entités nommées⁶ sont linguistiques⁷ : Le Meur, Galliano, et Geoffrois, par exemple, considèrent les entités nommées comme « des types d'unités lexicales particuliers qui font référence à une entité du monde concret dans certains domaines spécifiques, notamment humains, sociaux, politiques, économiques ou géographiques et qui ont un nom (typiquement un nom propre ou un acronyme) »⁸. La définition d'Ehrmann, elle, complète les préoccupations talistes par le point de vue linguistique : « on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus »⁹. Alors que le concept est créé par le TAL pour répondre à ses besoins spécifiques, l'entité nommée n'en reste pas moins, selon la chercheuse, une expression linguistique monoréférentielle (qui renvoie à une entité extralinguistique, voire à un référent unique) et autonome (qui permet à elle seule l'identification du référent). Cette définition, proche de celle du nom propre, mais moins restrictive,

⁶ Le terme *entité nommée* désigne aussi bien les entités que les noms ou les désignations qui permettent d'y référer. Le second emploi dérive du premier, mais la motivation du terme est bien moins évidente pour lui.

⁷ Mohamed Hatmi, « Reconnaissance des entités nommées dans des documents multimodaux », Thèse de doctorat en Informatique et Traitement Automatique des Langues, Université de Nantes, 2014.

⁸ Céline Le Meur, Sylvain Galliano et Edouard Geoffrois, *Conventions d'annotations en Entités Nommées*, ESTER, coll. « Rapport technique de la campagne Ester », 2004, p. 4.

⁹ Maud Ehrmann, « Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation », Thèse de doctorat en Informatique et Traitement Automatique des Langues, Paris 7, 2008, f. 168.

répond à notre besoin de traiter conjointement les noms propres et leurs périphrases. En revanche, les logiciels de reconnaissance d'entités nommées proposent des fonctionnalités qui se prêtent mal à nos besoins : entraînés majoritairement sur des corpus contemporains, journalistiques ou juridiques, ils sont très performants pour la délimitation et la catégorisation des entités nommées présentes dans ces types de corpus, dont les périphrases sont en nombre plus restreint et plus formalisé que ce n'est le cas dans un corpus littéraire.

Par exemple, la boîte à outils pour les humanités numériques mise à disposition par le laboratoire ObTic (Sorbonne Université), Pandore Toolbox¹⁰, inclut la reconnaissance d'entités nommées à l'aide de trois modèles différents : SpaCy, Flair et Bert¹¹. Alors que ce deuxième ne semble pas fonctionner pour le français¹², nous testons les deux autres modèles avec la traduction de Virgile Béague1998-1. Court, écrit en prose et en français moderne, ce texte est un des plus faciles à traiter automatiquement. Manuellement, nous relevons 67 entités nommées (noms propres et périphrases), dont 40 se retrouvent dans les résultats du traitement tant avec SpaCy qu'avec Bert. SpaCy détecte 7 entités supplémentaires que Bert ne trouve pas. Ce dernier comporte quelques détections fautives, principalement dues au mauvais traitement des apostrophes courbes (« l'herbe », « qu'ils s'avançaient » – mais pas seulement, car il considère également « Voilà » comme entité nommée –, et omet quelques occurrences des noms qu'il reconnaît par ailleurs (*Orphée*, *Eurydice*), ce qui est surprenant. SpaCy, quant à lui, détecte uniquement les noms propres seuls, alors que Bert se montre capable de délimiter des syntagmes qui ont le nom propre comme noyaux (une seule occurrence : « pays des Cicones »). En revanche, aucune périphrase n'est proposée par les deux logiciels. La pertinence de SpaCy est donc comparable à ce que nous obtenons déjà grâce à l'annotation morphosyntaxique avec TreeTagger.

Puisque les techniques de reconnaissance automatique d'entités nommées ne nous permettent pas d'aligner les dénominations d'entités correspondantes, nous avons conçu un algorithme capable de les localiser et de les identifier (qu'il s'agisse des noms propres ou des périphrases), ce qui est utile en soi et qui permet de valider la pertinence de la détection. Sa conceptualisation a été facilitée par l'observation des manières dont la reformulation est traitée dans des corpus spécifiques, qu'ils soient littéraires ou terminologiques. De l'apprentissage des relations morphologiques au filtrage et à l'adaptation des synonymes de la langue générale à un corpus spécialisé, nous avons suivi des

¹⁰ Voir <https://obtic-gpu1.mesu.sorbonne-universite.fr:8550/>.

¹¹ Voir https://obtic-gpu1.mesu.sorbonne-universite.fr:8550/entites_nommees.

¹² En passant par Pandore Box, nous rencontrons un message d'erreur et sur le site du développeur (<https://github.com/flairNLP/flair>) le français ne fait pas partie des langues prises en compte.

méthodes appliquées à la terminologie médicale¹³. Nous avons également étudié les méthodes automatiques d'extraction des relations hyperonymiques (patrons lexico-sémantiques, inclusion lexicale et ponctuation¹⁴) et d'extraction des topiques¹⁵, dont nous retenons principalement la méthode d'identification des thèmes selon trois marqueurs : (i) les marqueurs lexicaux de surface (*concernant* X, etc.), (ii) les marqueurs syntaxiques (les positions relative et absolue des SN indiquent où se trouvent le thème et le rhème), (iii) les marqueurs référentiels (des facteurs comme la distance, le degré de saillance ou le facteur positionnel facilitent la détection du référent d'une expression)¹⁶.

Dans le cadre de notre corpus, les marqueurs référentiels et la prise en compte du contexte sont très importants pour la détermination des référents des périphrases, alors que les marqueurs lexicaux de surface sont plutôt rares. Nous constatons, par exemple :

- que la construction de l'adjectif *malheureux* suivi ou précédé d'un nom d'humain masculin singulier désigne Orphée dans 90 % des cas, comme l'adjectif *furieux* apparaît uniquement dans la désignation des Bacchantes ;
- que si nous rencontrons une expression référentielle désignant une entité de sexe masculin et considérée seule, les chances qu'il s'agisse d'Orphée augmentent significativement si le nom ou une périphrase d'Eurydice apparaît à proximité ;
- ou encore, que le nom *épouse* désigne toujours Eurydice quand il est précédé d'un article défini ou qu'il est dépourvu de complément.

Ainsi, nous combinons plusieurs approches pour établir un échantillon d'expansions de noms propres et de périphrases candidates extrait des résultats de la détection automatique des correspondances. Les résultats sont annotés au niveau des mots (lemme et catégorie grammaticale) et des syntagmes grammaticaux (type et noyaux). Nous procédons ensuite à quatre types d'extractions, que nous présentons en Tableau 89, et que nous illustrons avec quelques dénominations de Pluton attestées dans le corpus.

¹³ Voir par exemple Pierre Zweigenbaum, Fadila Hadouche et Natalia Grabar, « Apprentissage de relations morphologiques en corpus », *Actes de TALN 2003 (Traitement automatique des langues naturelles)*, Batz-sur-Mer, ATALA IRIN, 2003, p. 285-294; Natalia Grabar et Pierre Zweigenbaum, « Lexically-Based Terminology Structuring: Some Inherent Limits », *Terminology*, vol. 10, 2004, p. 23-54; Natalia Grabar et Pierre Zweigenbaum, *Utilisation de corpus de spécialité pour le filtrage de synonymes de la langue générale*, Dourdan, 2005, p. 81-90.

¹⁴ Voir par exemple Adel Ghamnia et al., « Extraction de relations : combiner les techniques pour s'adapter à la diversité du texte », *28es Journées Francophones d'Ingénierie des Connaissances (IC 2017)*, Caen, France, 2017.

¹⁵ Voir par exemple Laurence Longo et Amalia Todirascu, « Une étude de corpus pour la détection automatique de thèmes », *Texte et corpus*, vol. 4, janvier 2013, p. 143-155.

¹⁶ Voir Mira Ariel, « Accessibility theory: An overview », dans Ted Sanders, Joost Schilperoord et Wilbert Spooren (dir.), *Human Cognitive Processing*, vol. 8, Amsterdam, John Benjamins Publishing Company, 2001, p. 29, en ligne, <<https://benjamins.com/catalog/hcp.8.04ari>>.

Tableau 89. Différentes dénominations de Pluton

Type	Exemples
Nom propre noyau d'un syntagme (qui est donc un SNp)	« le terrible Hadès » « le farouche Pluton aux cauchemars d'enfant » « Pluton qui domine dans ces tristes lieux »
SNp complément d'un nom commun noyau d'un syntagme (SNC)	« le roi d'Enfer » « le terrible mari de Proserpine » « dieu Pluton »
SNC apposé à un SNp	« Pluton, leur grand roi » « Hadès, l'impitoyable souverain des Enfers que tout le monde hait »
Pronom démonstratif suivi d'une subordonnée relative	« celui que tout le monde hait » « celui qui gouverne les bas lieux »

L'analyse empirique et l'interrogation de l'échantillon ainsi obtenu permettent de réaffirmer quelques connaissances antérieurement établies¹⁷ (l'assignation des qualificatifs spécifiques à certains personnages), de confirmer quelques intuitions (le renvoi systématique à Eurydice comme *épouse*), et de mieux comprendre les procédés linguistiques de la génération des périphrases (marqueurs lexicaux, syntaxiques et référentiels). Grâce aux connaissances enrichies de notre corpus, acquises lors de son exploration manuelle et pendant les étapes d'annotation et de correction des résultats, nous pouvons mieux définir nos besoins pour le traitement des dénominations d'entités.

L'observation des travaux existants et l'analyse de l'échantillon extrait du corpus ont facilité la conception d'un traitement permettant d'extraire et d'identifier une partie des dénominations d'entités attestées dans le corpus. Ces données permettent déjà d'améliorer la pertinence de l'alignement des unités lexicales, d'écarter certains couples non pertinents et d'approfondir notre observation des procédés de réécriture. Mais un inventaire des dénominations d'entités généré à partir du corpus seul est trop pauvre dans le sens où il permet seulement de traiter ce corpus – et encore sans pallier les absences induites par des oublis ou des inattentions de l'analyse humaine – et où il le fait sans anticiper des cas de figure plus larges, potentiellement présents dans d'autres réécritures du mythe, non inclus dans le corpus pour le moment. Par ailleurs, au sein des projets HN, une ligne directrice commune est de produire des outils et des données réexploitables, qui peuvent être adaptés au traitement d'autres corpus. Au-delà de notre projet, un recensement des entités et leurs dénominations constitue une ressource pouvant intéresser non seulement des littéraires, mais aussi des prag-

¹⁷ Nous pouvons exploiter quelques dictionnaires comme celui de Maurice de la Porte (*Les Épithètes*, Paris, Gabriel Buon, 1571.) ou des dictionnaires mythologiques (Pierre Brunel, *Dictionnaire des mythes littéraires*, édition revue et augmentée, Monaco, Éditions du Rocher, 1994; Pierre Grimal, *Dictionnaire de la mythologie grecque et romaine*, Paris, Presses Universitaires de France, coll. « Grands dictionnaires », 1999.). Concernant les études littéraires de la dénomination d'entités, voir par exemple le recueil *Inqualifiables fureurs. Poétique des invocations inspirées aux XVI^e et XVII^e siècles*, Paris, Classiques Garnier, coll. « Études et essais sur la Renaissance », 2019.

maticiens, des onomasticiens, des lexicographes... En outre, une méthode de détection et d'identification adaptable à la spécificité de chaque corpus peut s'avérer mieux exploitable, notamment pour des projets HN littéraires ou historiques, que les outils existants, appuyés sur du français courant et contemporain. Tout cela implique, notamment, que nous devons rendre les ressources produites interoperables et faciles à partager, en les structurant en respect d'un standard existant, documenté et connu de la communauté.

b) Création des lexiques et d'un algorithme à base de règles

C'est pour cela que nous explorons les ressources produites au sein de deux projets existants, Democrat¹⁸ et PROLEX¹⁹, que nous nous en inspirons pour organiser notre inventaire des dénominations d'entités et que nous décidons de construire deux lexiques conjointement nécessaires pour l'extraction et l'identification univoque des référents d'une partie des dénominations d'entités.

Le premier lexique conçu regroupe les noms communs susceptibles d'être employés dans des expressions dénominatives ou des procédés d'inclusion lexicale. Ces noms ont été extraits principalement du corpus annoté au sein du projet Democrat, pour travailler sur les chaînes de (co-)référence²⁰. Le corpus, constitué d'un échantillon d'une soixantaine de textes en français de genres variés (roman, manuel d'éducation, articles encyclopédiques, texte juridique) et d'époques différentes (entre le XI^e et le XXI^e s.) est annoté manuellement pour fournir les référents de toutes les expressions périphrastiques ou pronominales rencontrées. Nous exploitons ces annotations pour extraire 878 des têtes lexicales des périphrases annotées. Nous structurons la liste ainsi obtenue en XML-TEI²¹, en mobilisant un choix d'attributs pour spécifier la nature de chaque nom et ses quelques traits sémantiques. Par exemple (1), *mari* ou *époux* sont annotés comme anthroponymes relatifs au thème de l'amour et décrivant une relation de liaison (a) et *roi* ou *prince* sont considérés comme titres du pouvoir (b), contrairement aux noms *maître* ou *souverain* qui expriment une fonction non formalisée (c). Une annotation spécifique est aussi établie pour les toponymes.

- (1) Noms communs potentiellement porteurs de périphrases (attribut : valeur)
- (a) *mari* → groupe : anthroponyme, thème : amour, relation : liaison
 - (b) *roi* → groupe : anthroponyme, thème : règne, relation : titre
 - (c) *maître* → groupe : anthroponyme, thème : règne, relation : fonction

¹⁸ Voir <https://www.ortolang.fr/market/corpora/democrat/v1.1>.

¹⁹ Voir <https://www.ortolang.fr/market/lexicons/prolex/v3.1>.

²⁰ L'objectif du projet est d'étudier la variation des chaînes de références dans de différents genres discursifs, puis de développer un système de détection automatique des coréférences. Voir <https://www.lattice.cnrs.fr/democrat/motivations.html> et Frédéric Landragin, « Expressions référentielles et chaînes de référence en français : le projet Democrat et son exploration des rapports entre linguistique textuelle et linguistique de corpus », *Écho des études romanes*, vol. 18, n° 1, 2022, p. 49-65, en ligne, <<https://doi.org/10.32725/eer.2022.004>>.

²¹ Voir, dans les annexes de ce chapitre (IV 2 periphrases), le document [introduceurs_noms_propres.xml](#) et son [schéma DTD](#). Pour faciliter leur lecture, dans les exemples (1) et (2), la structure XML-TEI est simplifiée et retranscrite dans un format « attribut : valeur ».

Ce lexique est couplé à un autre, qui est en fait une structure de recensement et de catégorisation des noms propres présents dans notre corpus²². Pour sa modélisation, nous nous appuyons cette fois-ci sur le projet PROLEX et son lexique multilingue des entités nommées extraites d'un corpus journalistique contemporain²³. Nous en retenons le système d'identification des entrées, leurs différentes variantes et leurs dérivés. Selon ce principe, Pluton (2), par exemple, est caractérisé comme *dieu*, mais aussi comme *roi* (a). Il est associé à un identifiant unique spécifique au lexique (*pluton*). Les autres noms propres qui le désignent (*Hadès*, *Aïdonée*) sont regroupés en tant que PIVOTS, associés à la forme *Pluton*, considérée comme NOM (b). Le nom et les pivots peuvent entretenir des relations spécifiques entre eux : *Pluton* et *Hadès* sont des variantes diachroniques et diatopiques, tandis qu'*Aïdonée* est une variante diaphasique de *Pluton*²⁴. Nous renseignons également qu'il s'agit d'un *anthroponyme* désignant une seule personne du genre *masculin* (c) et nous ajoutons des attributs qui permettent d'exprimer des relations entre cette entité et d'autres (parents, membres de la fratrie, etc.²⁵, (d)). Nous ajoutons en outre des alias graphiques (*Pluto*, *Plouton*)²⁶, et des dérivés (*plutonien*, *plutonique*, *plutoniste*).

- (2) Entrée simplifiée Pluton (attribut : *valeur*)
- (a) Groupe : *anthroponymes*, type : *dieu*, subtype : *roi*
 - (b) Identifiant : *pluton*, nom : *Pluton*, pivot₁ : *Hadès*, pivot₂ : *Aïdonée*
 - (c) Collectif : *non*, genre : *masculin*
 - (d) Père : *Saturne*, mère : *Cybèle*, fratrie : *Neptune Jupiter*, liaison : *Proserpine*, règne : *Enfers*
 - (e) Alias₁ : *Pluto*, alias₂ : *Plouton*, dérivé₁ : *plutonien*, dérivé₂ : *plutonique*, dérivé₃ : *plutoniste*

Grâce à la collaboration des deux lexiques, nous concevons un algorithme à base de règles²⁷, permettant d'extraire des syntagmes susceptibles d'être des périphrases et, dans certains cas favorables, d'identifier l'entité à laquelle ils sont référés. Tout d'abord, la totalité du corpus est parcourue à la recherche des syntagmes dont la tête est soit un nom d'entité, soit un nom commun présent

²² Voir le document [lexique_noms_propres.xml](#) et son [schéma DTD](#).

²³ Voir Mickaël Tran et Denis Maurel, « *Prolexbase*. Un dictionnaire relationnel multilingue de noms propres », *TAL*, vol. 47, n° 3, 2006, p. 115-139.

²⁴ Pour la typologie des variantes, voir Eugenio Coseriu, « Le double problème des unités "DIA-S" », *Les cahiers DIA : Études sur la diachronie et la variation linguistique*, 1998, p. 9-16. Au sein du projet PROLEX, différents types de variantes sont relevés, mais la nature des relations entre le nom et le pivot n'est pas précisée. Dans notre lexique, nous le faisons à l'aide d'attributs dédiés qui spécifient tant le type de chaque variante que, dans les cas d'une variante diatopique, la région où elle est employée (Grèce, Égypte, etc.).

²⁵ Au sein du projet PROLEX, elles sont définies comme les relations d'accessibilité, voir <https://repository.ortolang.fr/api/content/prolex/2/Prolexbase.pdf>.

²⁶ Nous suivons ainsi le classement de PROLEX, au sein duquel les alias sont définis comme « des synonymes qui dépendent de la langue. Nous avons regroupé dans le terme d'alias, d'une part, des synonymes exacts (les variantes d'écriture [...], les variantes orthographiques) et d'autre part, des quasi-synonymes, diatopiques ou diastratiques » (p. 5). Par exemple, les alias du pivot « Organisation des nations unies » sont « Nations unies » et « ONU ». Voir <https://repository.ortolang.fr/api/content/prolex/2/Prolexbase.pdf>, § 5.1.3. « Les alias et les dérivés ».

²⁷ Voir le script [traitement_noms_propres.py](#), qui peut être appliqué aux résultats du traitement décrit dans le chapitre précédent ([6 resultat_final.xml](#)). Les entités nommées et périphrases ainsi obtenues sont présentées dans le tableur [resultats_extraction_NP_et_peri phrases.xlsx](#).

dans le premier lexique. Reprenons l'exemple de Pluton, auquel nous voulons assigner la périphrase « roi des Enfers ». Le premier filtrage des dénominations d'entités candidates se fait en fonction de la nature du nom commun. *Roi* étant le titre de pouvoir, nous retenons uniquement des entrées de SOUS-TYPE : *roi*. Pour notre corpus, c'est le cas de 20 entités nommées (16 humains²⁸ et 4 dieux²⁹). Ensuite, les autres composants du syntagme sont pris en compte, notamment les compléments. Pour la tête du complément du nom « Enfers », nous retenons seulement les entrées pour lesquelles l'attribut RÈGNE à la valeur *Enfers*. C'est le cas de Pluton et de Proserpine, mais le nom commun *roi* peut uniquement désigner des entités de sexe masculin, c'est donc la première entité qui est considérée comme étant le bon référent. Dans le cas d'une périphrase imbriquée dans une autre, par exemple pour « le roi du ténébreux empire », c'est la périphrase imbriquée (« ténébreux empire ») qui sera identifiée la première et la reconnaissance de la périphrase englobante se fera à l'aide de son identifiant.

Peu performant au début, notamment à cause d'une détection assez importante de syntagmes non périphrastiques (« une femme volage », « les hommes mariés ») et d'une capacité d'identification limitée (« épouse d'Orphée » est bien identifié comme désignant Eurydice, mais pas « la malheureuse épouse »), cet algorithme est et doit encore être amélioré au fil des traitements, tantôt manuellement, tantôt en exploitant l'appariement des couples de correspondances. Prenons comme exemple deux traductions de Virgile (Tableau 90).

Tableau 90. Alignement des périphrases des entités nommées

Nisard1869	Desportes1863
Il vit [le profond royaume de [Pluton]Pluton]Enfers, et ses bois qui remplissent l'horreur et les ténèbres, et [les mânes]Mânes et [le terrible roi des [enfers]Enfers]Pluton.	Il descendit dans [les abîmes des Enfers]Enfers ; et, traversant de vastes forêts, noir séjour de la crainte, il aborda [les Mânes]Mânes, et parut devant [l'affreux monarque]?

Pour ce couple, la périphrase « le terrible roi des enfers » a été assignée à Pluton (évoqué clairement plus haut). Mais ce n'est pas le cas pour « affreux monarque » de la phrase correspondante qui ne précise pas les Enfers comme complément de *monarque*. La relation synonymique entre *roi* et *monarque*, ainsi que leur repérage comme entités nommées candidates pour l'alignement (avec une identification de Pluton pour la première et sans identification pour la deuxième), permet non seulement la mise en relation, mais aussi d'associer Pluton à la périphrase « affreux monarque », pour cette occurrence et pour toutes celles qui pourraient apparaître dans d'autres textes du corpus.

²⁸ Cègre, Bélus, Danaüs, Éaque, Minos, Rhadamanthe, Rhésus, Sisyphe, Talaon, Tantale, Ascalaphe, Athamas, Josaphat, Polycrate, Salmonée et Ulysse.

²⁹ Jupiter, Junon, Pluton et Proserpine.

Nous constatons par ailleurs que, dans notre corpus, toutes les occurrences de l'adjectif *affreux* au sein des dénominations et périphrases renvoient aux anthroponymes et toponymes liés aux Enfers. S'il y a une confusion possible, c'est entre Pluton, qualifié dans deux textes comme « l'affreux monarque », et Charon, décrit comme « l'affreux nocher des morts » et « ce vieillard tout ensemble affreux et vénérable ». Alors que, pour la première périphrase de Charon, le nom *nocher* nous permet d'identifier celui qui dirige la barque infernale, la deuxième nécessite de prendre en compte le contexte bien plus large : « Ce vieillard tout ensemble affreux et vénérable / fit à ce rare chancre un accueil favorable / et *traversant le fleuve* avec contentement, / pour mieux goûter sa voix, *rama* fort lentement » (Hermite1641).

c) Exploitation des données obtenues

Le regroupement et la schématisation des périphrases déjà identifiées permettent également d'élargir la formalisation des dénominations alternatives, par exemple en ajoutant des mots-clés fréquemment employés pour désigner cette entité nommée. En Figure 61, nous présentons une schématisation simplifiée (sous forme de graphe) d'une partie des périphrases de Pluton extraites du corpus.

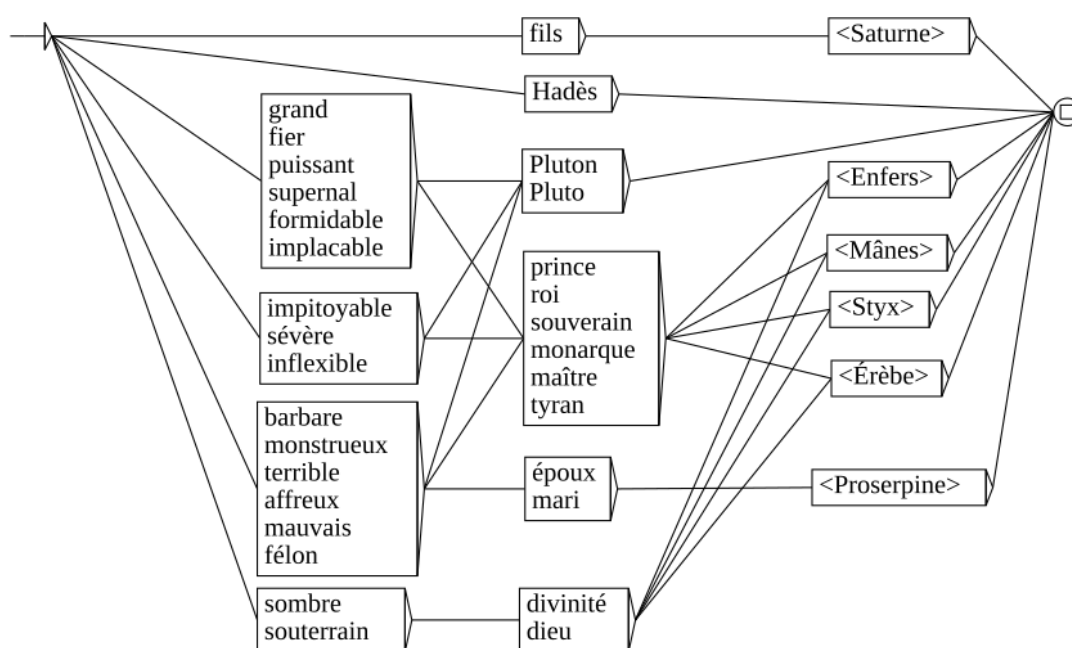


Figure 61. Graphe simplifié d'une partie des périphrases de Pluton attestées dans le corpus

Les périphrases de Pluton insistent régulièrement sur sa divinité ou sa royauté. Il est aussi évoqué comme fils de Saturne et époux de Proserpine. Les noms communs le désignant sont presque toujours accompagnés d'un complément, parfois aussi d'épithètes (antéposées ou postposées). Ces

dernières peuvent être divisées en quatre champs lexicaux. En tant que roi, Pluton est tout aussi loué (*puissant, formidable*) que craint (*impitoyable, terrible*), mais les épithètes qui accompagnent le nom *dieu* sont plutôt neutres et liées aux caractéristiques de son royaume (*souterrain*). Des adjectifs, négatifs uniquement, ont été recensés parmi les constructions avec le nom *époux*.

Ces informations nous permettraient, potentiellement, d'enrichir les points de repère qui permettent de procéder aux extractions et, par exemple, d'assigner à Pluton la périphrase descriptive « souveraine puissance sous qui tout tremble en ce morne royaume », et de prévoir d'autres dénominations possibles, attestées dans le corpus, mais pas encore assignées à un référent, voire absentes des textes, mais exploitables par exemple pour la détection automatique d'allusions au mythe dans d'autres corpus. Ces périphrases « hypothétiques » combinent (comme celles qui sont attestées) différents composants présents dans le graphe (« impitoyable dieu », « sombre fils de Saturne »...), voire en les remplaçant par leurs synonymes (« intraitable despote »).

L'annotation des entités nommées et leurs périphrases offre également quelques nouvelles possibilités d'analyse, par exemple en facilitant l'observation de l'évolution chronologique d'un des personnages. Voyons ce que cela peut révéler concernant la figure d'Eurydice. De nombreux érudits se sont penchés sur l'émancipation de cette figure féminine. Passive et secondaire à l'origine du mythe, Eurydice se voit attribuer progressivement une place majeure dans les textes contemporains. Si nous résumons quelques travaux critiques français, la modification des traits d'Eurydice est radicale : plus active, moins dépendante, elle n'hésite pas à exprimer sa propre parole, et tend même à la confrontation avec Orphée :

Une Eurydice qui ose laisser Orphée se désagréger dans la lumière, au moment où elle atteint le seuil de cette existence, une Eurydice elle-même pulsionnelle, destructrice, et celle qui se donne pour but d'établir le pont entre l'ascèse, manque de communication, et la sexualité, qui espère se guérir en guérissant, concilier le besoin d'être en soi et celui d'être de ce monde.³⁰

Désormais, elle est actrice de sa remontée : elle exige des explications quand Orphée ne la regarde pas, elle est en colère quand il le fait, voire elle refuse de revenir des Enfers :

Eurydice est la source et l'essence de l'œuvre, qu'Orphée saisit et perd en même temps. Orphée, par son regard, incarne l'inspiration, qui annonce l'œuvre et en même temps la voue à l'échec.³¹

D'autres chercheurs ont insisté sur son rapprochement avec des Bacchantes, tantôt pour nuire à Orphée, tantôt parce qu'elle est attirée par les femmes. Enfin, elle est parfois vue comme une jumelle antagoniste d'Orphée, la réécriture proposant un effet miroir entre les deux protagonistes :

³⁰ Metka Zupancic, « Introduction », *Religiologiques*, n° 15, « Orphée et Eurydice : mythes en mutation », 1997, en ligne, <<https://www.religiologiques.uqam.ca/no15/intro.html>>.

³¹ Pierre Brunel, « Orphée moderne », *op. cit.*

Eurydice, c'est plutôt l'aspect noir de la féminité liée au monde nocturne de la terre, du corps, d'une violence dionysiaque incarnée par les Bacchantes qui est, par opposition à un Orphée solaire désormais contesté, menacé ou condamné. Le mythe se met ici au service de revendications féministes à l'existence loin de toutes les représentations masculines narcissiques et aliénantes.³²

Dans « Orphée XIX^e et XX^e s. : interférences littéraires et musicales », Jacqueline Bellas parle d'un conflit de deux trajectoires du mythe, observables à partir du XIX^e s.³³ La première est la matérialisation charnelle concentrée autour d'Eurydice, extrêmement présente notamment à l'opéra et dont l'exemple par excellence est *Orphée et Eurydice* de Gluck (Moline 1774 dans le corpus). Cette trajectoire éclate avec la sortie d'*Orphée aux Enfers* d'Offenbach (Crémieux 1858) qui va la ridiculiser complètement. La deuxième trajectoire, qui montera alors en puissance, est celle de la revalorisation du mythe. Eurydice devient alors qu'une étape d'initiation du poète inspiré, parfois complètement abolie. Comme chez Monteverdi, Orphée est délivré des passions humaines grâce à la mort d'Eurydice, de même, à partir du XIX^e s. l'épisode d'Eurydice semble être un frein à l'élévation du mythe vers un sens nouveau, sans doute plus symbolique.

C'est la mise en marche des mutations successives où l'on progresse vers une épuration des supports littéraires [...]. On ne considère plus Eurydice comme une étape plus ou moins virtuelle de l'initiation, on l'abolit quelquefois, et l'on retourne insensiblement aux sources avec les affabulations mystiques de Ballanche et de Schuré.³⁴

À partir de cet état de lieux de quelques travaux critiques existants, nous pouvons nous demander si l'annotation des occurrences du nom d'Eurydice et de leurs périphrases peut permettre d'observer et de représenter l'évolution de l'importance de la figure d'Eurydice. Ensuite, peut-on, à l'aide de ces mentions, déterminer si elle est dépoétisée ou, au contraire, revalorisée ? Et, enfin, les méthodes informatiques permettent-elles d'observer les liens qui existent entre Eurydice et les Bacchantes dans les réécritures symboliques ?

Parmi les 25 300 phrases composant le corpus, 1 144 contiennent au moins une occurrence du nom d'Eurydice. Leur répartition chronologique est représentée sur le diagramme à la Figure 62. La montée des occurrences entre le XVI^e et le XVIII^e (7 %-7,6 %) semble confirmer la thèse de Bellas : après l'exploitation assez soutenue du motif le plus humain du mythe, on constate une retombée progressive, puis une nouvelle remontée au XXI^e s.

Cependant, si au lieu de regrouper les œuvres par siècles, nous affichons chronologiquement les proportions des phrases contenant le nom d'Eurydice dans chaque œuvre (Figure 63), nous réalisons qu'en réalité il s'agit des vagues très hétérogènes où certaines œuvres, principalement celles non canoniques, présentent des fréquences hautes, supérieures à 20 %, comme le *Dessein du poème*

³² Arlette Bouloumié, « La Résurgence du mythe d'Eurydice et ses métamorphoses dans l'œuvre d'Anouilh, de Pascal Quignard, d'Henri Bosco, de Marguerite Yourcenar, de Michèle Sarde, et Jean Loup Trassard », *loc. cit.*

³³ Jacqueline Bellas, « « Orphée » au XIX^e et au XX^e siècles », *loc. cit.*

³⁴ *Ibid.*

et des superbes machines du mariage d'Orphée et d'Eurydice (Buffequin1648 ; 21,74 %), la cantate *Orphée* de François de Lacroix (Lacroix1717 ; 21,74 %) et deux parodies, *Opuscule d'un célèbre auteur égyptien* (Col-lande1752 ; 21,95 %) et « Une histoire renouvelée des Grecs » (Eimann1854 ; 20,18 %).

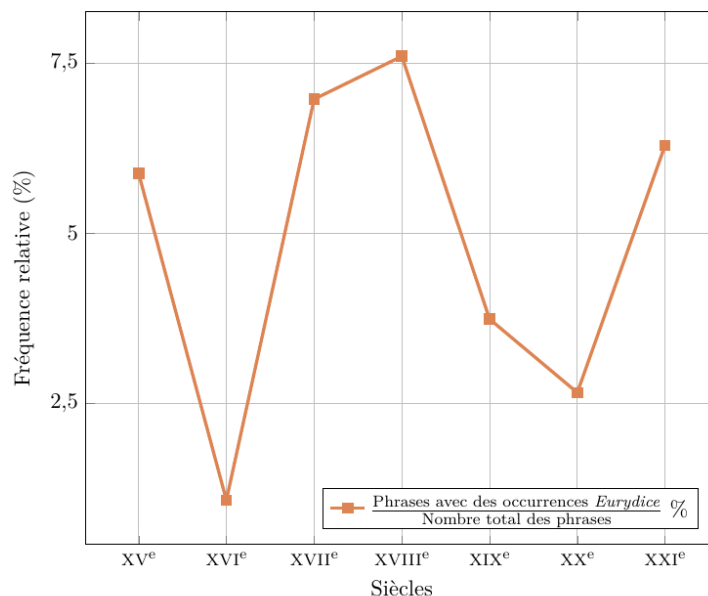


Figure 62. Répartition chronologique par siècle de publication

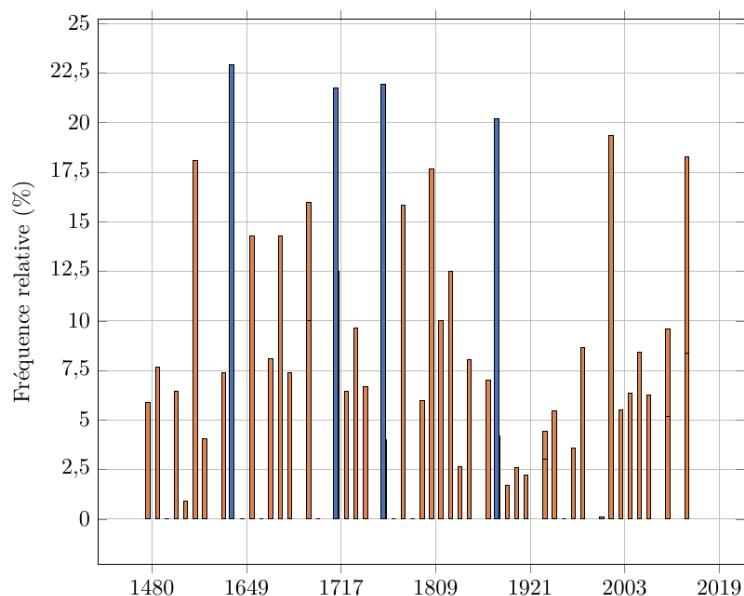


Figure 63. Répartition chronologique par année de publication

Au contraire, pour 9 œuvres, aucune occurrence n'est recensée et deux autres présentent moins d'1 % d'occurrences. Quatre sont des parodies : les pièces *Orphée ou Arlequin aux Enfers* (Fuzelier1711) et *La Descente d'Arlequin aux Enfers* (Regnard1823) sont des œuvres de la *commedia dell'arte*. Eurydice

est donc remplacée par le personnage canonique de Colombine. Dans la pièce de Dorvigny 1775, il s'agit de Javotte³⁵, alors qu'Orphée devient Roger-Bontemps³⁶. Dans le roman *Orphée Dafric* (Linking 1981) et dans le théâtre-rituel éponyme (Manjock 1981), qui s'inscrivent dans la continuité du mythe de l'Orphée noir, Eurydice devient Nyango³⁷.

Pour mieux observer et analyser tant les proximités que les différences dans la réinterprétation du récit original, nous revenons aux sources antiques du mythe pour constater qu'Eurydice y est mentionnée principalement par périphrases. Selon Jacques Heurgon :

[L]a femme d'Orphée [...] n'existe qu'en fonction de l'exploit ou de la faute de son mari ; on ne lui connaît ni parents ni patrie ; elle ne commence à vivre qu'à l'instant de sa mort ; elle n'a point de personnalité ; à peine a-t-elle un nom qui lui appartienne en propre ; en tout cas les poètes s'en souviennent rarement, et n'aiment pas l'employer ; et même ils ne savent pas s'il faut l'appeler, comme beaucoup de nymphes, Argiopè [...] ou, comme beaucoup de princesses, Eurydice.³⁸

Il serait donc erroné de se contenter d'observer uniquement les occurrences du nom d'Eurydice, sans au moins essayer d'identifier ses autres dénominations. En effet, en comptant ces mentions directes (nom ou périphrase), nous arrivons à des résultats comparables pour les deux séries traductives : 196 occurrences chez les traductions de Virgile, soit environ 12 mentions par texte, et 253 occurrences chez Ovide, soit 13 mentions en moyenne. Le nom *Eurydice* est mentionné un peu plus souvent chez Virgile : il s'agit de 62,75 % de la totalité des occurrences, contre 57 % chez Ovide. Dans les deux séries traductives, les périphrases concernent principalement le lien conjugal avec Orphée : *épouse* (20,4 % chez Virgile et 20,16 % chez Ovide) et *femme* (4,6 % et 7,5 %).

Néanmoins, les schématisations des dénominations d'Eurydice chez Virgile et chez Ovide montrent des différences significatives. Tout d'abord, les groupes lexicaux sont plus développés chez Virgile, et le lexique est nettement plus riche (Figure 64) : nous recensons une dizaine d'adjectifs et participes qualificatifs, antéposés ou postposés tant au nom d'Eurydice qu'à ses périphrases les plus fréquentes, des propositions subordonnées relatives (« qui aimait Orphée », etc.) et des adverbes comme *tant* ou *tendrement*. Le nom d'Eurydice est également introduit comme un complément du nom *ombre*.

³⁵ s. v. JAVOTTE² : « Pop. Femme bavarde et indiscrete », *TLLFi*, en ligne, <<https://www.cnrtl.fr/definition/javotte>>.

³⁶ s. v. ROGER-BONTEMPS : « Terme familier. Il se dit d'une Personne de belle humeur et qui vit sans aucune espèce de souci », *DAF*, 7^e éd., 1878, en ligne, <<https://www.dictionnaire-academie.fr/article/A7R1800>>.

³⁷ s. v. NYANGO : « [Cameroun] Dame. », *Base de données lexicographiques panfrancophone*, en ligne, <<https://www.bdlp.org/resultat?base=CA&query=801057,0?>>>.

³⁸ Jacques Heurgon, « Orphée et Eurydice avant Virgile », *Mélanges de l'école française de Rome*, vol. 49, n° 1, 1932, p. 15-16, en ligne, <https://www.persee.fr/doc/mefr_0223-4874_1932_num_49_1_7222>.

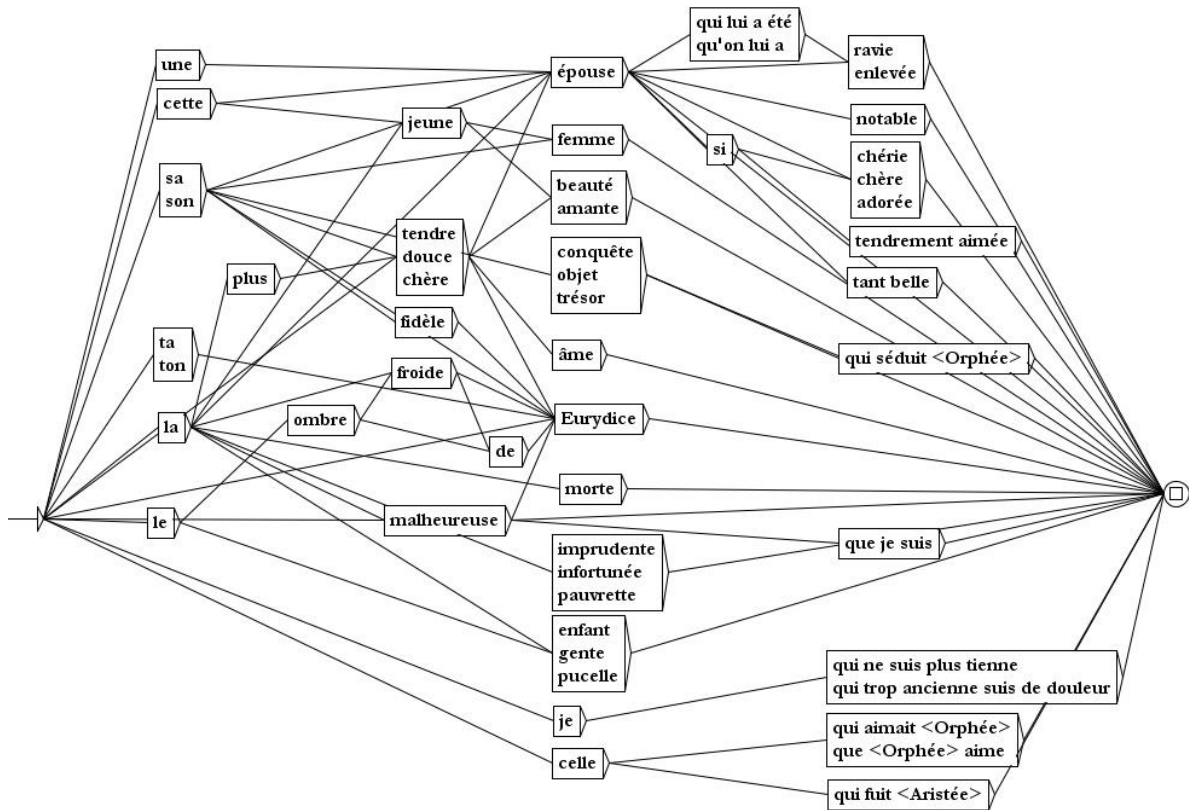


Figure 64. Schématisation des périphrases d'Eurydice dans les traductions des *Géorgiques* de Virgile

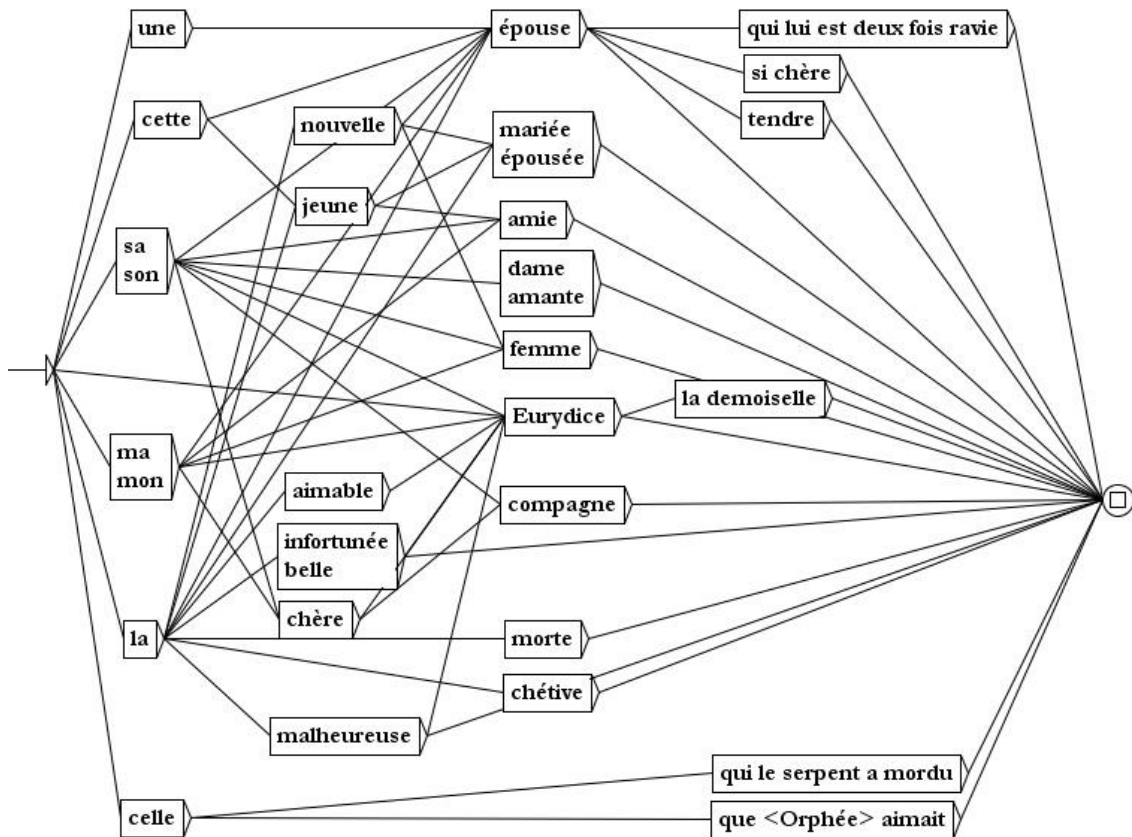


Figure 65. Schématisation des périphrases d'Eurydice dans les traductions des *Métamorphoses* d'Ovide

Chez Ovide (Figure 65), les périphrases sont un peu moins nombreuses, et c'est surtout la richesse des groupes nominaux qui diminue nettement. Si nous regardons plus en détail les champs lexicaux présents, chez Ovide Eurydice est majoritairement définie par rapport à Orphée. Six périphrases seulement (12 occ.) sont différentes et précisent sa physique (« la belle », 3 occ.), son état d'âme (« la malheureuse », « la chétive », « l'infortunée »), voire sa situation (« la morte », « celle qui le serpent a mordu »).

C'est aussi le cas chez Virgile, mais ces périphrases sont plus nombreuses et nous apprennent davantage sur le personnage d'Eurydice (« l'enfant », « la pucelle », « sa fidèle Eurydice »), qui semble par ailleurs davantage valorisé (« une épouse notable », « son trésor », « sa conquête »). Et, surtout, elle y paraît plus actrice de son destin (« l'imprudente », « l'objet qui le séduit »). Notre relevé porte également la trace de l'unique réplique d'Eurydice dans la version de Virgile, et qui disparaît chez Ovide : il s'agit des périphrases introduites par le pronom personnel *je* : « je qui ne suis plus tienne » et « je qui trop ancienne suis de douleur ».³⁹

Outre montrer les différences entre les deux sources antiques, cette schématisation permet également de délimiter de manière détaillée la figure “classique” d'Eurydice et de faciliter l'extraction de ses mentions dans le corpus des réécritures. En effet, nous pouvons présupposer que les transmodalisations qui suivent fidèlement les mythes traditionnels reproduiront les schémas dénommatifs des sources antiques (« sa chère épouse », « malheureuse Eurydice »). L'identification des occurrences du nom d'Eurydice qualifiées par un vocabulaire distinct de celui recensé dans les séries traductives (« cette femme inspirée », « courageuse Eurydice ») ou, au contraire, les épithètes fréquentes (*chère*, *malheureuse*) qui qualifient un nom commun non recensé (« chère Madame », « malheureuse femelle »), peuvent témoigner de la modification du rôle d'Eurydice dans la réécriture.

En effet, l'extraction et l'annotation dans les réécritures de toutes les phrases contenant le nom d'Eurydice montrent que, sur 124 épithètes, la plupart sont attestés dans les séries traductives : *chère* (45 occ., 36 %), *belle* (15 occ.), *adorable*, *adorée*, *aimable*, *blessée au talon*, *charmante*, *docile*, *fidèle*, *fragile*, *apeurée*, *éblouie par la lumière*, *infortunée*, *malheureuse*, *morte*, *seule*, *tendre*... Toutes les épithètes restantes sont recensées dans les modernisations idéologiques et dans les parodies :

- « Qui eût vu alors **Eurydice engagée en cet amoureux combat**, l'eut prise pour la Déesse Minerve » (Puget1624)

³⁹ La diminution du rôle d'Eurydice de la version de Virgile à celle d'Ovide, a déjà été relevée par la recherche traditionnelle, comme le travail de thèse d'Eva Kushner : « [Chez Ovide] Orphée ne représente ni l'Incantation, ni aucun autre idéal esthétique ou religieux, il n'est que lui-même, un homme en face de sa douleur. Pour la littérature latine, l'origine religieuse de la légende d'Orphée et de son échec était maintenant oubliée et la légende transformée en un conte d'amour » (*Le mythe d'Orphée dans la littérature française contemporaine, op. cit.*).

- « Je vais trouver enfin dans ton gouffre béant **l'éternelle Eurydice** ou l'éternel Néant ! » (Grandmougin1882)
- « **Eurydice vivante** m'eût donné l'ivresse du bonheur » (Schuré1889)
- « Tu n'es pas **l'autre Eurydice**, morte d'amour, morte là-haut » (Segalen1921)
- « **Madame Eurydice** reviendra des enfers » (Cocteau1927)
- « J'ai perdu **la divine et terrestre Eurydice** ! » (Loyson1948)

À l'inverse, l'observation des cooccurrences de l'adjectif *chère* montre que 26 occurrences sur 42 désignent clairement Eurydice (« chère Eurydice », « ma chère moitié », « une épouse qui t'est chère »), et plusieurs autres l'évoquent de manière plus ou moins implicite (« la plus chère récompense », « la chère image aux lignes pures, aux gestes longs » et « chère guirlande inoubliée »). Toutes les autres occurrences qualifient des noms abstraits ou inanimés (*patrie, contrée, mémoire, solitude, étreinte, douceur, folie, vie*), sauf deux : « ô toi, si chère à toutes nos compagnes » et « si votre sœur vous est chère » qui qualifient la Bacchante Phaëtuse dans la pièce *Orphée. Étude dramatique en deux parties* (Defontenay1854).

Phaëtuse symbolise chez Defontenay1854 le plaisir des sens. Subjuguée par la musique d'Orphée, qui apaise momentanément sa nature violente, elle le maudit en voyant ses avances repoussées :

Oh ! Malheur sur toi qui veux séduire sans pouvoir assouvir les passions que tu sais exciter ! Oh ! malheur ! car tu me laisses le cœur embrasé et tout à la fois plein de fureur. Malheur, te dis-je ! sur toi et les tiens, car si je ne puis te servir à genoux, je te tuerai..., toi et ton Eurydice !

Et c'est ce qu'elle fait, d'abord en conduisant Eurydice à la rencontre d'Aristée, puis en encourageant Cresphonte (qui symbolise la force brutale) à tuer Orphée.

Le corpus contient au total huit noms propres féminins extérieurs à la trame principale du mythe et aux mythologies grecque et romaine : Phaëtuse, Érigone, Moïna, Aglaonice, Nyango, Colombine, Javotte et Ismène. Quatre d'entre elles sont des Bacchantes, bien présentes dans les sources anciennes où aucune n'est nommée spécifiquement, mais qui sont périphrasées en tant que groupe et de manière très variée (« une troupe de Bacchantes », « Les Dames du pays, armées de peaux eu devant de l'estomac », « Les Ménades, couvertes de la dépouille des animaux et agitées des fureurs de Bacchus », « femmes des Cicones... »).

Le personnage de Bacchante joue un rôle particulier dans les œuvres qui se focalisent sur l'aspect initiatique du mythe orphique. La pièce de Cocteau1927 a bien fait connaître Aglaonice, introduite par Schuré1889, mais elle est loin d'être la seule Bacchante nommée spécifiquement. Un personnage semblable à celui de Phaëtuse, Moïna, est introduit par le roman de jeunesse *Orphée l'enchanté* (Jimenes2004) :

[Orphée] souriait, confiant : il avait pris une décision irrévocable et se trouvait en paix avec lui-même. Moïna se méprit sur le sens de son sourire. Elle crut qu'il lui était destiné. Aussi s'approcha-t-elle dans l'intention de se coller lascivement à Orphée. Il la repoussa d'une main ferme et elle fut dégrisée, comprenant à quoi s'adressait ce sourire : une ombre au-delà des vivants.

La Bacchante Érigone, quant à elle, apparaît dans Ballanche¹⁸²⁹. Comme pour Moïna et pour Phaëtuse, Orphée refuse l'amour d'Érigone. Cependant, ce personnage n'est pas aussi maléfique que les autres Bacchantes, car, pour devenir digne d'Orphée, Érigone essaie de s'élever jusqu'à lui et meurt ne pouvant supporter les accords immortels qu'elle a essayé de tirer de la lyre d'Orphée :

Jeune et belle ménade, dit Orphée, ton mal me cause une pitié profonde ; mais les chants qui peuvent fléchir le roi des morts peuvent-ils de même fléchir le tyran qui est en toi ? Érigone, élève ton âme vers les régions étoilées où réside l'éclatant Phanès, où habite la majesté des dieux. [...] Érigone, un effort qui fut au-dessus d'Eurydice, qui vainquit la fille de la vision, ne saurait être facile pour toi, pour toi faible mortelle, qui n'est point née d'un Titan.

Cette figure féminine renforce également le lien entre Orphée et Dionysos-Zagreus dont elle est l'amante dans la mythologie grecque. Elle se pend par désespoir en découvrant le corps de son père, Icarios, massacré par des bergers. Furieux, Dionysos lance une malédiction conduisant toutes les jeunes filles d'Athènes à se pendre à leur tour. Nous retrouvons implicitement dans cette vengeance l'épisode de la poursuite d'Eurydice par Aristée et la malédiction d'Orphée qui s'abat sur ce dernier.

Si, comme le souligne Bellas, deux trajectoires du mythe s'opposent à partir du XIX^e s., nous considérons que sa revalorisation n'est pas inhérente à l'abolition du personnage d'Eurydice, et celui-ci n'est pas régulièrement réduit à un objet passif de l'initiation du chantre. Le transfert des passions charnelles et violentes au personnage de la reine des Bacchantes permet de garder l'image d'Eurydice pure et initiée, *presque* égale à Orphée, ou encore, dans les réécritures qui rapprochent implicitement Eurydice et les Bacchantes, de supprimer l'aspect charnel de la cause de la mort d'Orphée. Les mythes traditionnels de ces deux personnages féminins sont modifiés dans une bonne partie des réécritures du XIX^e s. Par conséquent, dans les œuvres postérieures, les Bacchantes ne peuvent plus être considérées comme agresseurs exclusivement maléfiques, comme Eurydice n'est plus une douce victime qui n'existe qu'à travers son mari.

Cela est particulièrement visible dans l'opéra *Orphée-roi* (Segalen¹⁹²¹), qui reprend tous ces nouveaux personnages féminins : l'Eurydice de Segalen se rapproche plus d'Érigone, et meurt, elle aussi, en touchant la lyre d'Orphée :

EURYDICE

L'œuvre est beau...

Et je défaille... Orphée... sous ta voix. Je ne suis plus que l'écho de ta voix...

Je... ne... suis plus...

Elle s'incline et glisse doucement extasiée étendue aux pieds d'Orphée-roi qui, jusqu'au bout, mène l'hymne inexorable.

Brunel rapproche la Ménade déguisée en Eurydice, qu'Orphée fuit dans les Enfers, à Aglaonice⁴⁰. Mais elle nous semble se rapprocher tout autant de Phaëtuse, attirée par Orphée, contrairement à Aglaonice qui convoite Eurydice.

ORPHÉE
Eurydice, qu'est-tu devenue ! [...]
LA MÉNADE, *comme en folie*
Je t'ai dit « Viens » aussi. Je suis Reine ici. Je suis ivre ! Moi, prêtresse, je saisis mon
dieu !
Dans un transport elle enlace le corps redresse d'Orphée dont, sauvagement, elle entoure les hanches.

Cette *autre* Eurydice d'inscrit entièrement dans les caractéristiques des personnages introduits par les réécritures précédentes : voluptueuse, d'une féminité effrayante et sensuelle... Chez Segalen¹⁹²¹, les Reines de Bacchantes de Ballanche¹⁸²⁹, de Defontenay¹⁸⁵⁴ et de Schuré¹⁸⁸⁹ se confondent (à nouveau ?) en Eurydice et, finalement, la lyre avec laquelle elle doit désormais rivaliser n'est tout de même pas dépourvue de quelques traits traditionnels de cette femme, comme le montrent peut-être ces quatre uniques occurrences dans *Orphée-roi* du mot *lyre* écrit avec une majuscule :

- « tu est belle et indomptable, Lyre, amante enchantée » ;
- « Lyre, c'est à toi que vont les jeux aimants : tes hanches sont polies en nacrées ; la courbe de tes cornes est cambrée comme deux bras dansants : douloureuse » ;
- « Il a sa Lyre ressuscitée dans les bras » ;
- « seule, intacte, mortelle, à tous bienfaisante, irréal, harmonieuse, la Lyre s'élève peu à peu ».

Une autre manière de procéder à partir des données obtenues de l'annotation des entités nommées et de choisir d'observer la place qu'un texte ou un auteur en particulier prend au sein d'un ensemble. Pour ce type d'interrogation, Jean Cocteau est un des auteurs les plus intéressants de la partie moderne du corpus, mais aussi celui qui se prête le moins bien à ce type d'analyses. D'une part, alors que le mythe d'Orphée semble être un *topos* particulièrement récurrent chez le poète, nous n'incluons dans notre corpus que sa pièce de théâtre, *Orphée. Tragédie en un acte et un intervalle*, représentée pour la première fois en 1926, mais dont la première édition imprimée date de 1927 (Cocteau¹⁹²⁷). Au sein de nos analyses quantitatives, cette œuvre est très peu visible (Figure 66).

⁴⁰ « Du plus profond de l'antre on voit naître une Forme voilée. Ce n'est qu'une fausse Eurydice, une Ménade déguisée qui, après avoir été jalouse de la fille du Citharède, veut se venger. L'idée vient des *Grands Initiés*, de l'antre où Aglaonice, la prêtresse d'Hécate et l'amante d'Eurydice, éprouve l'hierophante objet de sa haine. Cet antre plein de vapeurs méphitiques rappelle moins le traditionnel antre de Trophonios que la caverne de Zarathoustra, celle d'où il doit sortir, à la fin du livre, pour aller vers la force et le soleil. » Pierre Brunel, « *Orphée-roi* de Victor Segalen, ou le miracle de la lyre », *op. cit.*, paragr. 18.

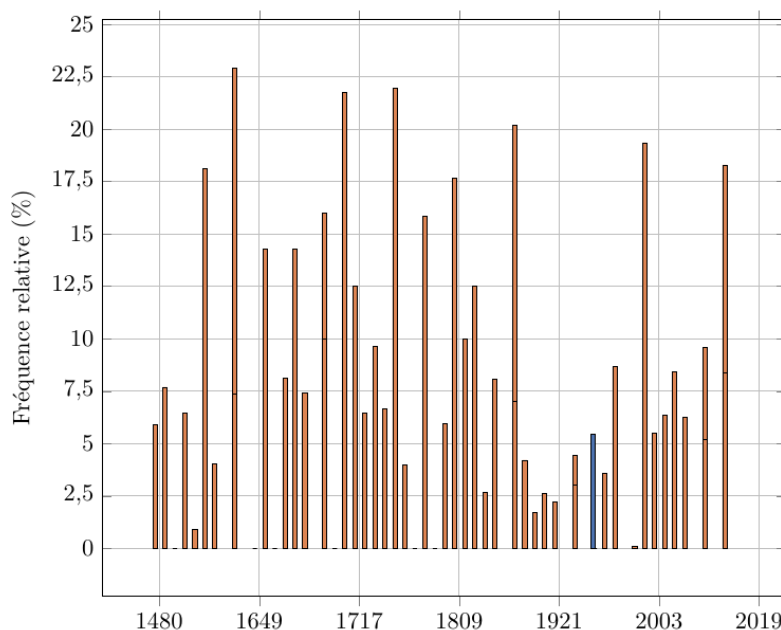


Figure 66. Fréquence des mentions d'Eurydice chez Cocteau1927 (en bleu) par rapport aux autres œuvres du corpus

Seulement 58 sur 1 066 phrases contiennent au moins une mention du nom d'Eurydice (soit environ 5,5 %). C'est une fréquence légèrement inférieure à la moyenne (9 %), mais très proche de la médiane (6,5 %), alors que, généralement, on a tendance à s'intéresser plutôt aux résultats limitrophes : les fréquences particulièrement élevées ou particulièrement basses.

De même, le recensement des mentions et périphrases d'Eurydice dans la pièce résulte en un graphe très simple (Figure 67), bien plus simple que celui obtenu avec les séries traductives des sources anciennes (Figure 64–Figure 65).

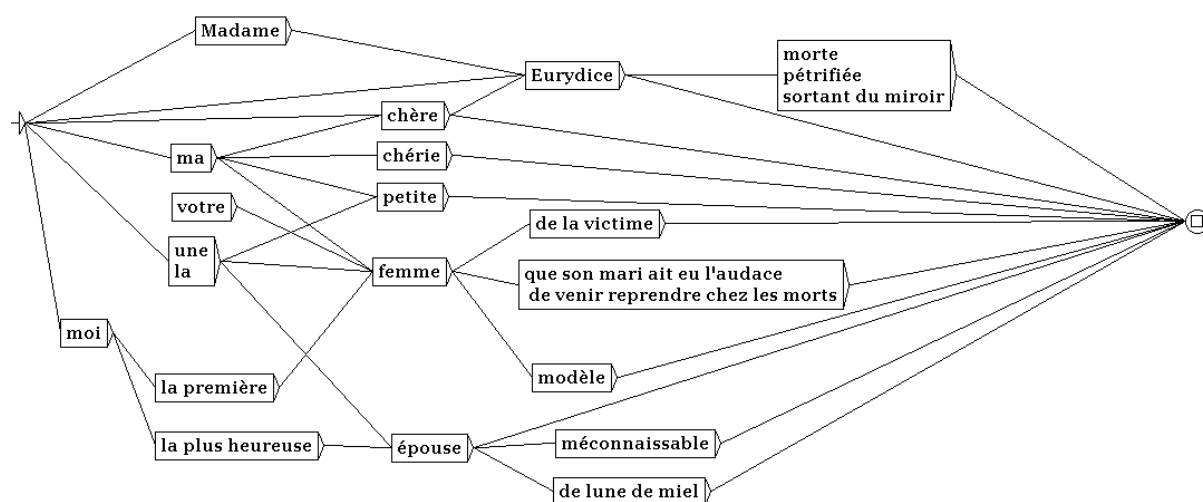


Figure 67. Schématisation des périphrases d'Eurydice chez Cocteau 1927

Sur 69 occurrences du nom *Eurydice*, 11 seulement sont complétés par un attribut ou une épithète : « Madame Eurydice » (6 occ.), « Eurydice est morte » (3 occ.), « chère Eurydice » (prononcé par Heurtebise), « Eurydice, pétrifiée » et « Eurydice sortant du miroir », dans les didascalies.

Pour comparer, le nom apparaît 105 fois dans la pièce *Eurydice* d'Anouilh (Anouilh 1941), ce qui n'est pas beaucoup non plus, mais ses épithètes et attributs sont extrêmement riches, et surtout très parlants : « petite Eurydice », « charmante Eurydice », « mademoiselle Eurydice », « Eurydice est ma maîtresse », « Eurydice est une drôle de fille », « Eurydice est un être rare », « votre petite Eurydice fidèle et ordonnée était un rêve », « J'aurais préféré Eurydice défigurée, souffrante, vieille »...

Ensuite, sur 25 occurrences du mot « femme » chez Cocteau 1927, 11 désignent Eurydice, dont 6 sont prononcées par Heurtebise (« votre femme », 5 occ. et « une femme modèle », 3 par Orphée (« ma femme ») et une par le commissaire (« la femme de la victime »). Enfin, dans une seule occurrence le sens du mot « femme » n'est pas synonyme d'épouse, et elle est prononcée par Eurydice : « moi la première femme que mon mari ait eu l'audace de venir reprendre chez les morts ». Même en se définissant elle-même, Eurydice n'existe qu'à travers son mari. D'ailleurs, dans cette même phrase, elle se décrit comme « la plus heureuse des épouses ».

Il faut creuser bien davantage pour trouver les désignations qui échappent à la représentation très classique d'Eurydice. L'observation des occurrences du pronom « elle » révèle de manière assez parlante le fait qu'Eurydice n'est pas la seule héroïne chez Cocteau 1927. En effet, 33 occurrences la concernent (19 dans les répliques), mais 35 autres désignent soit la Mort soit la Bacchante Aglaonice. Et, étrangement, c'est l'image d'Aglaonice qui en ressort la plus flatteuse : « Elle a du talent », « Elle a de belles images », « Elle est capable de choses très bien »... C'est très différent pour La Mort (« elle est avare », « elle aime mieux prendre que donner », « elle ne rend jamais ce qu'on lui laisse prendre » et pour Eurydice (« Vous ne savez pas de quoi elle est capable », « Elle boude », « Elle me le payera », « elle m'aurait vite envoyé promener », « Elle est absente », « Elle avait même abandonné le domicile conjugal », « Elle est morte, morte, remorte »), sauf une occurrence : « elle a tué le diable sous la forme d'un cheval ». Peut-être que, chez Cocteau 1927, Eurydice est en train de s'émanciper et les figures féminines prennent petit à petit leur pouvoir pour revenir en force 24 ans plus tard, dans le film *Orphée* ?

Grâce à la collaboration avec Fanny Van Exaerde⁴¹, nous pouvons affirmer que les premiers traitements quantitatifs sur le scénario du film (Tableau 91) confirment l'augmentation de l'importance du personnage de La Mort, qui prononce moins de 5 % de répliques dans la pièce, mais qui, en tant que La Princesse, devient le troisième personnage le plus présent dans le film avec 17 % de

⁴¹ Fanny Van Exaerde, « « Une poésie cinématographique. » Étude historique, génétique et intermédiaire des scénarios et dialogues de Jean Cocteau », Thèse de doctorat, Université de Lille, 2023.

répliques, une quarantaine de plus qu'Eurydice. Nous observons par ailleurs l'apparition du personnage d'Aglaonice, seulement mentionnée dans la pièce, dans la liste des personnages (même si elle ne prononce que 13 répliques). Cette importance est davantage valorisée par la lecture rapprochée et l'étude génétique des scénarios, et notamment au sein d'une séquence inédite du Club des Bacchantes qui dévoile la grossesse d'Eurydice et l'aide d'Aglaonice pour l'interruption volontaire de celle-ci par l'ingestion de l'absinthe⁴².

Tableau 91. Répartition du nombre des répliques dans la pièce de théâtre (1926) et dans le film (1950)

Pièce de théâtre			Scénario du film		
Personnage	Nb.	%	Personnage	Nb	%
Heurtebise ¹	165	29	Orphée ²	213	26
Orphée ²	155	27	Heurtebise ¹	174	21
Eurydice ³	123	21	La princesse ⁵	143	17
Le commissaire ⁴	48	8	Eurydice ³	109	13
La mort ⁵	28	5	Les trois juges	38	4,5
Le greffier ⁶	28	5	Cégeste ⁷	28	3
Azraël ⁷	15	2,5	Le commissaire ⁴	21	2,5
Raphaël	15	2,5	Le monsieur	19	2
Voix du facteur	6	1	Aglaonice	13	1,5
			Des policiers	12	1,45
			Le journaliste	11	1,33
			Des écrivains	10	1,21
			La radio	7	0,85
			Des jeunes filles	6	0,72
			Des motocyclistes	6	0,72
			Voix de l'auteur	6	0,72
			Le greffier ⁶	3	0,36
			Cris	2	0,24
			Une grosse femme	2	0,24

⁴² Voir Fanny Van Exaerde, *Absinthe, avortement et Bacchantes dans le scénario d'Orphée (1949-1950) de Jean Cocteau*, Ritualités sociales et littéraires. Boire et écrire, Université de Lille, 7 février 2025.

Conclusion du chapitre

L'annotation automatique des noms propres et leurs périphrases que nous proposons combine donc l'analyse empirique avec des manipulations automatiques et semi-automatiques pour se donner les moyens de procéder à des explorations fructueuses en composant avec la taille restreinte du corpus et en exploitant l'hétérogénéité au moyen d'annotations qui valorisent les points communs et de lexiques qui compilent ce qui a été observé. La validation et l'analyse humaine du corpus améliorent la qualité des données obtenues, même si cela doit parfois se faire au détriment de la quantité.

Il est certain que les traitements quantitatifs ne permettent pas d'observer et de représenter quoi que ce soit s'ils ne sont pas conjugués à une analyse détaillée, à une connaissance approfondie du corpus et à une méfiance critique concernant les résultats obtenus. Mais la conscience du biais permet d'aboutir à une exploitation équilibrée du corpus numérisé qui montre quelques résultats probants. Il semblerait effectivement qu'on puisse identifier automatiquement les mentions d'Eurydice, après avoir effectué un relevé manuel et exhaustif de ses dénominations traditionnelles et que, dans certains cas, il est possible de déterminer automatiquement s'il s'agit d'une dépoétisation ou d'une revalorisation. De même, l'annotation des entités nommées permet de dévoiler les liens entre Eurydice et les Bacchantes, non seulement dans les réécritures symboliques, mais également celles parodiques ou destinées pour la jeunesse. Les outils informatiques ne restent que des outils dans les mains des chercheurs. Ils offrent certes de nouvelles possibilités d'interrogation, mais demandent à être maniés prudemment et, surtout, ils ne sont pas conçus pour produire des conclusions ou pour dépasser l'érudition humaine. Comme le disait Franco Moretti dans *Graphs, Maps and Trees*, « la quantification pose le problème, la forme apporte la solution »⁴³.

En présentant notre méthode de l'annotation automatique, nous ambitionnions aussi d'illustrer d'un certain point de vue la place des traitements linguistiques au sein de notre projet, tout en insistant sur l'importance du dialogue et de l'échange (des connaissances, des pratiques, mais aussi des données). Malgré un rattachement principal en littérature comparée, le projet que nous portons inclut des questions linguistiques soulevées par nos traitements, qu'ils soient informatiques ou empiriques. Nous constatons ainsi que des traitements linguistiques fins, pour suppléer aux limites des logiciels Tracer et TextPAIR, ou réalisés pour eux-mêmes, induisent une réelle amélioration des détections des correspondances textuelles. En définitive, si les Humanités numériques peuvent continuer à se développer, c'est en approfondissant leur dimension collaborative. Cette coopération n'implique pas seulement des échanges entre disciplines, mais aussi la combinaison de différentes

⁴³ Franco Moretti, Étienne Dobenesque et Laurent Jeanpierre, *Graphes, cartes et arbres: modèles abstraits pour une autre histoire de la littérature*, Paris, les Prairies ordinaires, coll. « Penser-croiser », 2008, p. 58.

approches au sein d'un même domaine (Linguistique et TAL, recherches littéraires traditionnelles et informatisées, *close* et *distant reading*...)

Dans le champ particulièrement vaste et hétérogène des Humanités numériques, le rôle de la langue peut différer d'un projet à l'autre, mais nous sommes certaines que l'intégration de méthodes linguistiques de qualité, notamment pour la préparation et l'analyse préliminaire du corpus, ne peut être que bénéfique. L'interopérabilité et le partage des données et des usages, l'organisation des rencontres interdisciplinaires, et la conception de projet au sein d'équipes diversifiées et polyvalentes rendent possibles de stimulantes découvertes, qui ne sont pas toujours réalisables dans le cadre restreint de son domaine de spécialité. Ainsi, au sein des Humanités numériques, des projets très différents, focalisés sur la terminologie médicale, la conception, d'un moteur de recherche thématique, ou encore l'annotation des chaînes de co-référence, implémentent des méthodes qui s'avèrent exploitables pour le repérage des périphrases et l'identification des entités qu'elles nomment au sein d'un corpus littéraire de réécritures du mythe d'Orphée et Eurydice. Nous espérons, de notre côté, parvenir à produire des données mieux exploitables par des linguistes, en restant attentives aux besoins spécifiques de ce domaine, et à valoriser d'autant plus l'histoire commune des recherches littéraires et linguistiques⁴⁴ dans le contexte numérique.

Cependant, pour pouvoir valoriser les données produites et faciliter l'exploitation de ce que nous concevons, il est important de mettre à disposition de la communauté une interface graphique qui présente les résultats de manière structurée, lisible et interactive. C'est ce que nous souhaitons proposer avec notre prototype d'une édition comparative numérique que nous présentons au chapitre suivant.

⁴⁴ Concernant les relations entre la littérature et la linguistique, voir par exemple Jean-Michel Adam et Ute Heidmann, *Le texte littéraire*, *op. cit.*

3. Présentation de prototype d'une édition comparative numérique¹

En 2015, Elena Pierazzo consacre un chapitre de son essai *Digital Scholarly Editing*² à la figure du lecteur et tente de répondre à quelques questions essentielles, à savoir : qui sont les lecteurs des éditions numériques savantes ? quels sont leurs besoins ? comment lisent-ils ? À ces questions nous en ajoutons d'autres : quelles sont les nouvelles possibilités de visualisation ? comment concevoir des interfaces innovantes et intuitives ? et, finalement, qu'est-ce que les lecteurs, habitués à l'imprimé, même si les générations nativement numériques se joignent progressivement au public potentiel, accepteront et adapteront à leurs usages ?

La conception d'une édition numérique devient ainsi un projet pleinement interdisciplinaire. Aux tâches classiques, comme l'établissement du texte et la conception de l'apparat critique, s'ajoutent celles du Web design, de la visualisation de l'information et même du génie logiciel. Simultanément, le désir de profiter pleinement de nouvelles possibilités offertes par le numérique se heurte aux difficultés de cerner les besoins et les appétences du public cible. Aux exigences traditionnelles de l'édition savante, qui conduisent souvent à des éditions numériques imitant l'objet livre originel (design « skeuomorphe »), s'ajoutent celles du Web, formulées entre autres par Ben Shneiderman³, qui posent quant à elles d'autres problèmes, relatifs à la compatibilité et l'interopérabilité des outils et des données. Plus encore, selon Tim McLoughlin⁴ par exemple, face au numérique, nous ne lisons plus de manière attentive et continue, nous naviguons sans but précis ou nous cherchons des informations dont nous avons besoin.

Pourtant, au sein de notre édition comparative des réécritures du mythe d'Orphée et Eurydice⁵ nous souhaitons explorer certaines des possibilités offertes par le numérique : dégagée des limites du papier, l'édition numérique peut présenter simultanément une quantité importante d'œuvres, mais aussi donner à voir, de manière interactive et attrayante, les relations qu'elles entretiennent

¹ Ce chapitre reprend partiellement notre article (co-signé avec N. Gasiglia) « Édition comparative numérique des réécritures du mythe d'Orphée et Eurydice. Cartographie hypertextuelle des correspondances », dans Pascal Mougin (dir.), *Littérature et design : visualités et visualisations du texte en régime numérique*, Dijon, les presses du réel, coll. « Figures », 2023, p. 243-266.

² Elena Pierazzo, *Digital Scholarly Editing: Theories, Models and Methods*, Ashgate, Farnham, Surrey, 2015, p. 147-164.

³ « Mantra de la recherche visuelle d'informations : vue d'ensemble d'abord, zoom et filtrage, puis détails sur demande » (« The visual information-seeking mantra is: overview first, zoom and filter, then details on demand », nous traduisons). Voir Ben Shneiderman, « The eyes have it: a task by data type taxonomy for information visualizations », *Proceedings 1996 IEEE Symposium on Visual Languages*, 1996, p. 337, en ligne, <DOI : 10.1109/VL.1996.545307>.

⁴ Tim McLoughlin, « Bridging the Gap », *Jahrbuch für Computerphilologie*, vol. 10, 2008, p. 37-54, en ligne, <<http://computerphilologie.digital-humanities.de/jg08/mclough.pdf>>.

⁵ Cette édition peut être consultée localement à partir des annexes de ce chapitre ([IV 3 prototype](#)), dans le dossier [HTML](#). Il peut être nécessaire de lancer un serveur local pour pouvoir visualiser les graphes. Cela peut être fait en ligne de commande « `python -m http.server 8000` », à condition d'avoir installé Python sur sa machine. La page d'accueil sera ensuite disponible dans le navigateur à partir d'adresse <http://localhost:8000>.

entre elles. Les vues analytiques des réécritures peuvent être diachroniques, synchroniques, cycliques⁶, mais aussi littéraires, linguistiques, iconographiques, ou encore intertextuelles, intermédiaires, intergénériques... Quels repères pouvons-nous donner au lectorat pour l'accompagner au mieux dans notre interface ? Devons-nous nous restreindre à ce qu'il connaît, comme, pour les éditions savantes, des annotations et l'alignement des variantes ou, pour les usages du Web, des chaînes de caractères ou des zones cliquables ou rendues actives en les survolant, des fenêtres pop-up ou des menus déroulants ? Des approches issues de la visualisation de l'information, comme les graphes de relations, peuvent-elles présenter l'intertextualité et être interprétables pour le public cible ? Il est difficile de répondre à ces questions, mais une visualisation plus complexe que l'annotation classique est certainement nécessaire pour valoriser le sens et la structure de nos données.

Dans la suite de ce chapitre, nous exposons l'état actuel de nos propositions de visualisations, en nous référant à d'autres usages observés dans des cadres similaires. Notre souhait est de soumettre ce prototype d'édition numérique à l'appréciation des collègues littéraires autant qu'aux spécialistes du design et de la visualisation des données textuelles.

a) Première visualisation des relations intertextuelles

Comment montrer les correspondances entre plusieurs textes ? Dans certaines éditions numériques génétiques, le choix se porte sur l'alignement horizontal, avec des fonctionnalités variables, comme le surlignement des variantes ou des blocs communs, la possibilité d'activer ou de désactiver des fenêtres pour comparer différentes versions ou se concentrer seulement sur un état précis, d'afficher le fac-similé, etc. C'est ce que proposent par exemple des outils comme Versioning Machine⁷ ou Edition Visualization Technology⁸. Comme l'a montré Ben Fry dans son projet *On the Origin of Species. The Preservation of Favoured Traces*⁹, ce mode d'affichage, de plus en plus répandu¹⁰,

⁶ Dans ce contexte, un « cycle » est un cadre temporel. Voir par exemple la définition qu'en donne Franco Moretti : « [...] les cycles constituent des *structures temporaires à l'intérieur du flux de l'histoire*. Il s'agit, après tout, de la logique sous-jacente à la tripartition de Braudel : le temps bref est du pur flux sans structure, la longue durée est la pure structure sans flux, et les cycles sont la zone frontalière – instable – entre les deux. ». Voir Franco Moretti, Étienne Dobenesque et Laurent Jeanpierre, *Graphes, cartes et arbres*, op. cit.

⁷ Voir <http://v-machine.org/>.

⁸ Voir <http://evt.labcd.unipi.it/>.

⁹ Voir <https://benfry.com/traces/>.

¹⁰ Parmi les éditions électroniques qui proposent ce type d'affichage, citons par exemple : celle des manuscrits de *Madame Bovary* de Flaubert (université de Rouen, <https://flaubert.univ-rouen.fr/%C5%93uvres/%C5%93uvres-publi%C3%A9es/madame-bovary/>) ; celle de *La Comédie humaine* de Balzac (ANR Phœbus, <https://ebalzac.com/>) ; ou celle des séries traductives de *l'Odyssée* par Marianne Reboul (<https://odysseuspolymetis.github.io/paralogos/>). Sur les outils de collation automatisée dans les éditions génétiques, voir par exemple : Jean-Gabriel Ganascia, *MEDITE - a unilingual text aligner for humanities. Application to textual genetics and to the edition of text variants*, Copenhagen ; Elena Pierazzo, « What future for digital scholarly editions? From Haute Couture to Prêt-à-Porter », *International Journal for Digital Humanities*, vol. 1, Springer, mai 2019, p. 1-12, en ligne, <<https://hal.archives-ouvertes.fr/hal-02117714>> ; Gregor Middell, « Review of “Juxta Web Service, LERA, and Variance Viewer. Web based collation tools for TEP” », loc. cit. Nous renvoyons également le lecteur aux précédentes parties de ce travail, notamment II.3 « Prolégomènes des approches computationnelles de l'intertextualité ».

donc potentiellement connu des lecteurs, peut donner lieu à des visualisations graphiques innovantes et permettant une lecture discontinue et partielle.

Toutefois, ces approches s'adaptent mal à notre corpus, trop hétérogène pour qu'on puisse en offrir un alignement lisible et attrayant. Souhaitant tout donner à voir, nous choisissons de ne pas le faire simultanément. Nous nous inspirons de Mark Heyer¹¹, qui propose de différencier trois types de lecture en référence à la métaphore de la collecte de nourriture par nos ancêtres :

- le *broutage*, à savoir une lecture continue et intégrale (pour s'imprégner de tout) ;
- le *foretage*, où l'on parcourt le texte sans chercher quelque chose de précis (comme la lecture d'un journal) ;
- la *chasse*, lors de la recherche d'une information spécifique (comme la consultation d'un dictionnaire).

Nous présumons que nos lecteurs seront dans une situation de *foretage* ou de *chasse*. Cela n'empêche pas que, si les outils sont adaptés, la lecture puisse se transformer en *broutage*, voire en lecture d'étude (*close reading*), modalité absente de la typologie d'Heyer, mais qu'il est nécessaire d'ajouter dans le cadre des éditions savantes, majoritairement conçues pour cela¹². Nous suivons aussi la première règle de Shneiderman : « la vue d'ensemble d'abord »¹³. Selon nous, l'affichage qui s'y prête le mieux puise dans la théorie des graphes et s'inscrit dans la continuité de la visualisation d'informations (*infovis*) décrite, entre autres, par Lev Manovich¹⁴. Les graphes sont potentiellement connus, puisqu'ils sont employés, par exemple, pour présenter des réseaux des relations artistiques ou sociales¹⁵.

Ainsi, nous proposons d'entrer dans l'édition par une page d'accueil qui affiche un graphe de toutes les relations détectées entre les textes du corpus (Figure 68). Les codes-couleurs, les formes des nœuds et les étiquettes donnent de premières informations concernant le contenu. La légende, affichée en haut à droite, précise les codes employés¹⁶. En plus de guider les utilisateurs vers d'autres vues, ce graphe informe déjà les lecteurs, qu'ils entrent dans l'édition sans but précis ou avec un

¹¹ Voir Mark Heyer, « The Creative Challenge of CD-ROM », dans Steve Lambert et Suzanne Ropiequet (dir.), *CD-ROM. The New Papyrus. The Current and Future State of the Art*, Redmond, WA, Microsoft Press, 1986, p. 347-357. Nous citons d'après Christian Vandendorpe, *Du papyrus à l'hypertexte. Essai sur les mutations du texte et de la lecture*, La Découverte, 1999, p. 206.

¹² Pour cette typologie enrichie, voir Elena Pierazzo, *Digital Scholarly Editing*, op. cit., p. 148. « However comprehensive, the list of types of readings given above is incomplete. In addition we could include the so-called "close reading", that is the type of attentive reading we reserve for texts we wish to memorize and learn from. »

¹³ Ben Shneiderman, « The eyes have it », loc. cit., p. 337.

¹⁴ Lev Manovich, *What is Visualization?*, 2010, en ligne, <<http://manovich.net/index.php/projects/what-is-visualization>>.

¹⁵ Voir par exemple <https://www.moma.org/interactives/exhibitions/2012/inventingabstraction/?page=connections>.

¹⁶ La classification du corpus, bien plus complexe (voir I.3 « Représentation de la variété des réécritures du mythe ») est ici simplifiée pour ne pas cumuler les codes.

objectif de recherche : la taille et la localisation des nœuds sont intentionnelles et chargées de sens, car plus un nœud est grand, plus le pourcentage de phrases avec correspondances en son sein est élevé. Huit textes sont isolés au centre (quatre traductions d'Ovide et quatre réécritures) alors que les réécritures sont plutôt en périphérie. Plus précisément, quatre cercles concentriques peuvent être distingués au sein du graphe. Le deuxième en partant du centre, par exemple, est constitué de 24 traductions et neuf réécritures. Ces dernières s'avèrent être étroitement liées avec les sources anciennes (contrairement aux réécritures du centre qui le sont autant avec des traductions qu'avec des réécritures).

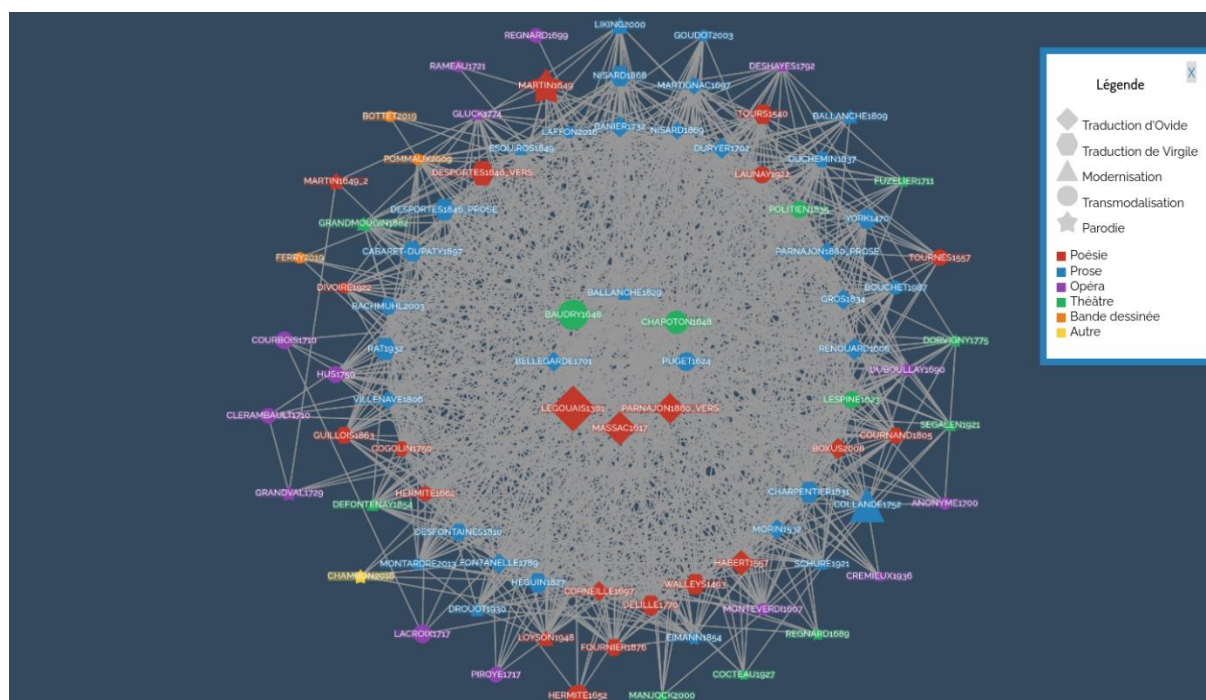


Figure 68. Page d'accueil de l'édition : visualisation générale du corpus

C'est le cas du texte d'Esquiro1849, une transposition en prose du mythe à l'ancienne Égypte¹⁷, qui est en relation avec 24 traductions et neuf réécritures (majoritairement celles symboliques, focalisées sur l'Orphée-initiateur). Un clic sur le nœud représentant l'œuvre permet de mieux observer ces relations, d'afficher la notice bibliographique de l'œuvre et d'accéder à son édition spécifique (Figure 69).

Certes, cette solution limite le focus à un seul texte à la fois. Mais elle offre déjà une riche documentation aux lecteurs attentifs. Pour chaque texte, deux types de visualisation de correspondances sont proposés, afin de répondre aux différentes attentes d'exploration et de proposer des solutions aussi lisibles pour les œuvres courtes que pour les œuvres longues et qu'elles aient plus

¹⁷ Voir III.4c) « Problèmes d'exploration et de visualisation ».

ou moins de relations intertextuelles : une édition dite « hyperannotée » et une autre dite « hypertextuelle ».

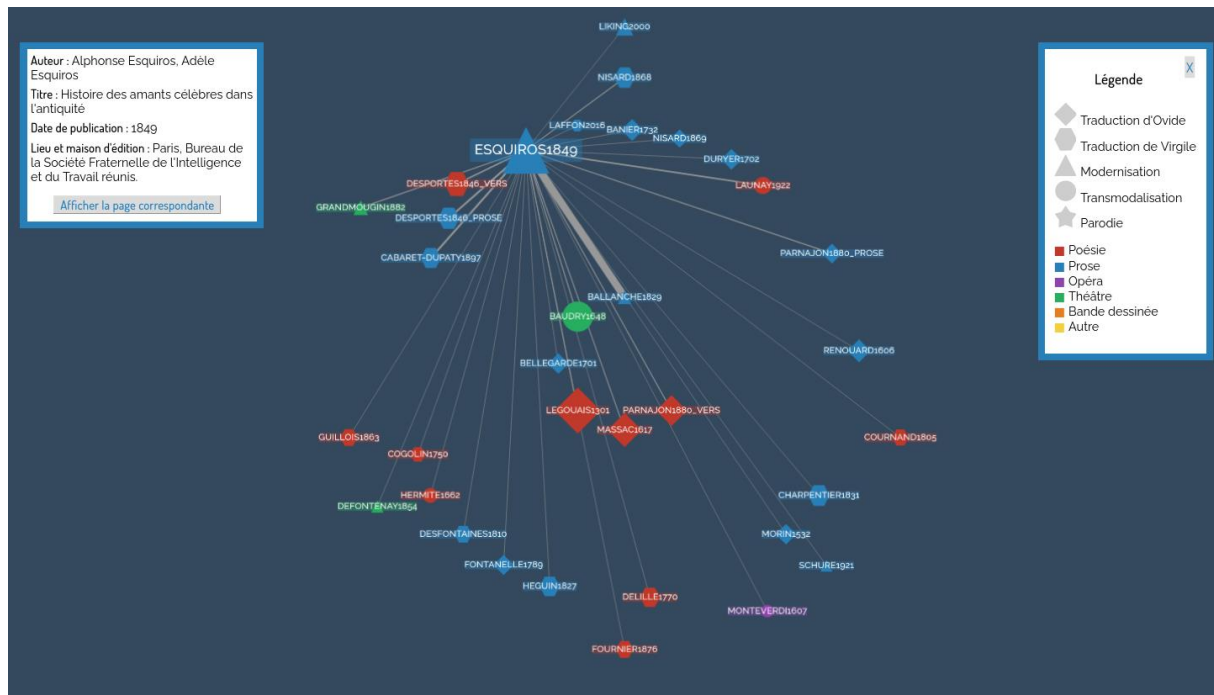


Figure 69. Visualisation générale : clic sur le nœud Esquiros1849.

b) Édition hyperannotée

Les lecteurs entrent en contact avec l'œuvre par le biais de la visualisation la plus intuitive parmi celles proposées : l'édition hyperannotée (Figure 70). Elle est principalement destinée à la lecture et à la prise de connaissance du texte. Pour cela, elle respecte les usages de l'édition et de la mise en forme de textes littéraires (structuration spécifique au genre, accès au fac-similé, etc.)¹⁸.

Cette visualisation utilise des objets connus, comme les hyperliens et les notes hypertextuelles, qui affichent leur contenu en infobulle et auxquelles les lecteurs numériques sont déjà habitués, ou encore des actions déclenchées sur une zone par le survol du curseur ou un clic¹⁹.

¹⁸ Ioana Galleron et al., *Les publications numériques de corpus d'auteurs. Guide de travail, grille d'analyse et recommandations*, op. cit. Voir aussi le chapitre 0.4 « Numériser un corpus dans l'esprit des humanités numériques ».

¹⁹ Nous nous référons ici à la théorie des « affordances » (James Gibson, « The Theory of Affordances », *The Ecological Approach to Visual Perception*, Boston, Houghton Mifflin, 1979, p. 127-137.) dont l'application à la visualisation d'informations est décrite par exemple dans Stéfán Sinclair, Stan Ruecker et Milena Radzikowska, « Information Visualization for Humanities Scholars », *Literary Studies in the Digital Age: An Evolving Anthology*, MLA Commons, 2013, en ligne, <<https://dlsanthology.mla.hcommons.org/information-visualization-for-humanities-scholars/>>.

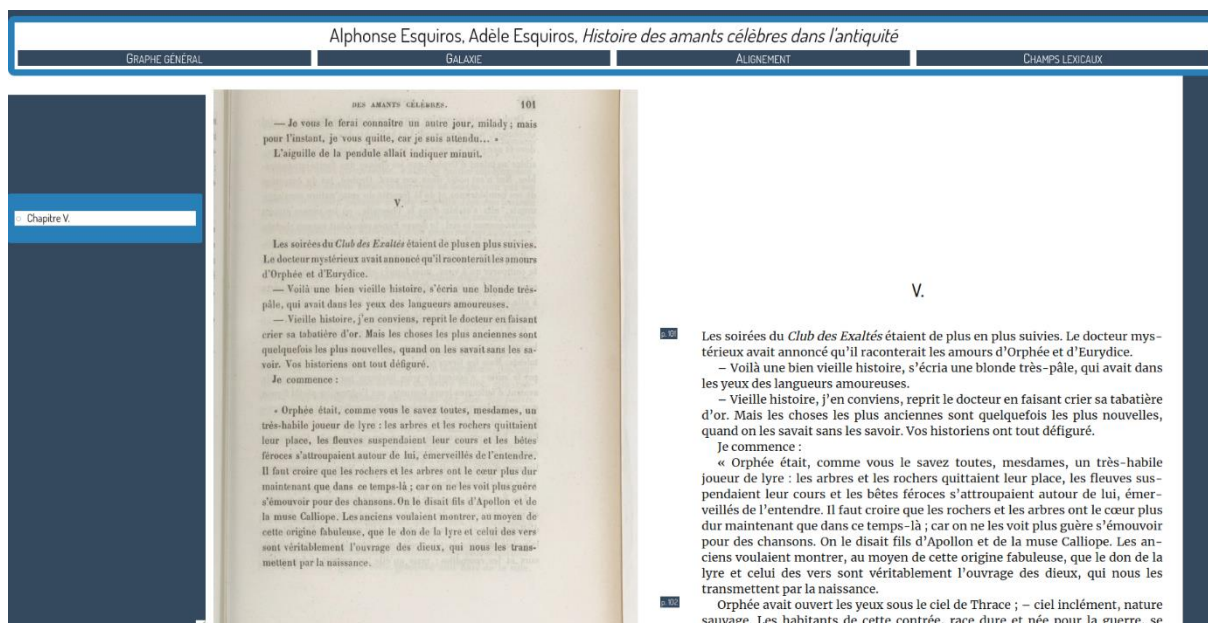


Figure 70. Exemple de page d'édition avec fac-similé

Nous nous conformons à ces usages, qui rendent l'interface intuitive, mais en introduisant un dispositif un peu plus habituel (que les lecteurs peuvent esquiver s'ils souhaitent ignorer complètement l'apparat critique proposé). Il s'agit des notes intertextuelles qui sont signalées avec des appels spécifiques (étoiles bleues), afin de les différencier des celles de l'éditeur (chiffres arabes) ou du traducteur (chiffres romains). Le survol du curseur sur l'appel de note intertextuelle met en surbrillance l'extrait du texte concerné et un clic fait apparaître une fenêtre de comparaison modulable (Figure 71).

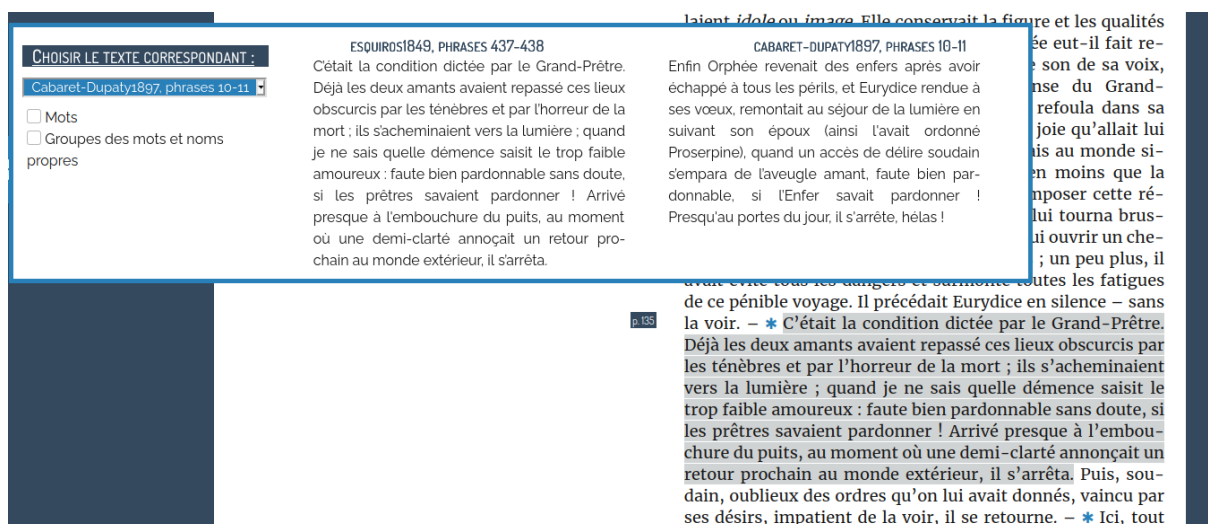


Figure 71. Contenu d'une note intertextuelle

Les couples d'extraits sont alignés afin de faciliter la comparaison. Un formulaire, situé en haut à gauche, permet de choisir le texte lié à afficher : dans l'exemple des phrases 437-438 d'Esquiros1849 (Figure 71-Figure 74), plusieurs correspondances ont été trouvées avec des traductions en prose de Virgile, par exemple pour les phrases 10-11 de Cabaret-Dupaty1897. Par défaut, le texte des correspondances s'affiche seul. Des annotations peuvent ensuite être ajoutées en cochant les cases « Mots » et « Groupes de mots et noms propres ». Les entités appariées sont alors mises en valeur et une légende des codes-couleur employés s'affiche sous la case cochée.

L'annotation des « Mots » (Figure 72) signale, en rouge, les reprises mot à mot, en vert, les paraphrases lexicales (synonymie) et, en orange, les relations morpho-syntaxiques (différences de flexion, de temps verbal, etc.) ou morpho-sémantiques (au sein d'une famille dérivationnelle) entre les unités lexicales pleines.

CHOISIR LE TEXTE CORRESPONDANT :

Cabaret-Dupaty1897, ph. 10-11

☒ Mots

Mot à mot

Synonymie

Relations lexicales flexionnelles et dérivationnelles

☐ Groupes de mots et noms propres

ESQUIROS1849, PHRASES 437-438

C'était la condition dictée par le Grand-Prêtre. Déjà les deux amants avaient repassé ces lieux obscurcis par les ténèbres et par l'horreur de la mort ; ils s'acheminaient vers la lumière ; quand je ne sais quelle démence saisit le trop faible amoureux : faute bien pardonnable sans doute, si les prêtres savent pardonner ! Arrivé presque à l'embouchure du puits, au moment où une demi-clarté annonçait un retour prochain au monde extérieur, il s'arrêta.

CABARET-DUPATY1897, PHRASES 10-11

Enfin Orphée revenait des enfers après avoir échappé à tous les périls, et Eurydice rendue à ses vœux, remontait au séjour de la lumière en suivant son époux (ainsi l'avait ordonné Proserpine), quand un accès de délire soudain s'empara de l'aveugle amant, faute bien pardonnable, si l'Enfer savait pardonner ! Presqu'au portes du jour, il s'arrête, hélas !

Figure 72. Annotation des mots

L'annotation en « Groupes de mots en noms propres », elle, va au-delà des mots et code les relations de proximité au niveau des combinaisons d'unités lexicales (Figure 73).

CHOISIR LE TEXTE CORRESPONDANT :

Cabaret-Dupaty1897, ph. 10-11

☐ Mots

☒ Groupes de mots et noms propres

Reprise mot à mot

Reformulation

Metaphore

Changement ou périphrase d'un nom propre

ESQUIROS1849, PHRASES 437-438

C'était la condition dictée par le Grand-Prêtre. Déjà les deux amants avaient repassé ces lieux obscurcis par les ténèbres et par l'horreur de la mort ; ils s'acheminaient vers la lumière ; quand je ne sais quelle démence saisit le trop faible amoureux : faute bien pardonnable sans doute, si les prêtres savent pardonner ! Arrivé presque à l'embouchure du puits, au moment où une demi-clarté annonçait un retour prochain au monde extérieur, il s'arrêta.

CABARET-DUPATY1897, PHRASES 10-11

Enfin Orphée revenait des enfers après avoir échappé à tous les périls, et Eurydice rendue à ses vœux, remontait au séjour de la lumière en suivant son époux (ainsi l'avait ordonné Proserpine), quand un accès de délire soudain s'empara de l'aveugle amant, faute bien pardonnable, si l'Enfer savait pardonner ! Presqu'au portes du jour, il s'arrête, hélas !

Figure 73. Annotation des groupes de mots et de noms propres

Pour la plupart, il s'agit de regrouper des reprises mot à mot (« faute bien pardonnable »), des combinaisons de plusieurs unités lexicales synonymes (« s'acheminaient vers la lumière ») et « remon-

taient au séjour de la lumière ») ou présentant des différences morpho-syntaxiques (« savaient pardonner » et « savait pardonner »). Une annotation plus complexe peut également être introduite²⁰, comme celle des métaphorisations (« repassé ces lieux obscurcis par les ténèbres et par l'horreur de la mort » et « revenait des enfers après avoir échappé à tous les périls ») et des références aux entités nommées. Ces dernières sont souvent analytiquement parlantes.

Par exemple la transposition du mythe dans Esquiros1849 au-delà de la mythologie est observable dans la correspondance analysée : « l'Enfer » de la traduction Cabaret-Dupaty1879 devient « les prêtres » et sa reine, « le Grand-Prêtre ». Ensuite, deux périphrases d'Orphée sont à noter : « le trop faible amoureux » et « l'aveugle amant ». Si nous arrêtons notre annotation au niveau des unités lexicales, *faible* et *aveugle* ne seraient pas synonymes. Il faut alors passer par une entité plus large, ici le syntagme nominal, pour observer que les deux adjectifs qualifient la même entité dans la même situation narrative. Enfin, la périphrase collective « les deux amants », soulignée en gris (Figure 73²¹), peut être doublement mise en correspondance avec deux des trois évocations des entités nommées individuelles (« Orphée » et « Eurydice », ou « son époux » et « Eurydice »).

Les deux couches d'annotation peuvent enfin être superposées. La visualisation devient alors plus complexe, le surlignage des groupes correspondants se transformant en délimitation entre crochets (Figure 74). Mais les lecteurs qui l'exploitent sont probablement en situation d'étude, ils connaissent bien les phrases reliées et l'annotation progressivement introduite. Ils peuvent donc s'y repérer et approfondir leur analyse.

Figure 74. Cumul de deux couches d'annotation

²⁰ Dans cette présentation du prototype, nous montrons l'annotation souhaitée, ici issue de l'enrichissement manuel de l'annotation automatique, que nous avons présenté au chapitre précédent (IV.1 « Post-traitement des résultats de la détection automatique des correspondances »). Alors que nous y proposons une méthode d'annotation (semi)-automatique des périphrases (IV.2 « Annotation des entités nommées et leurs périphrases »), les métaphorisations ne sont pas, pour le moment, prises en compte.

²¹ Le survol du curseur sur une zone en surbrillance met en évidence la ou les entités correspondantes en les soulignant en gris.

c) Édition hypertextuelle

Alors que l'édition hyperannotée permet d'explorer des relations intertextuelles, une vision de la totalité des correspondances détectées est nécessaire pour les percevoir et les analyser globalement. Pour ce faire, nous recourons à une visualisation qui rappelle celle proposée pour entrer dans l'édition hyperannotée (Figure 68-Figure 69), à savoir des graphes de correspondances.

Différents types de graphes ont déjà été exploités pour présenter des relations intertextuelles²². L'interface d'exploration TRAViz²³, intégré dans le logiciel Tracer, propose un alignement des phrases à l'aide d'un graphe orienté acyclique²⁴. Cela permet d'examiner les variantes entre des phrases similaires et de comparer efficacement des textes dont les modifications se limitent à celles flexionnelles ou lexicales. Mais, comme nous l'avons vu pour la correspondance entre les phrases 427-428 de la réécriture Esquiros1849 et 10-11 de la traduction Cabaret-Dupaty1897, beaucoup de relations sont synonymiques (9/16). Or, TRAViz ne met en correspondance que les mots identiques.

Un autre dispositif de visualisation, Galaxies²⁵, propose de regrouper les couples de correspondances détectées par TextPAIR en analysant et en enrichissant les résultats, afin de construire des graphes de relations. Plusieurs affichages sont proposés, permettant par exemple de focaliser l'attention sur les mots communs trouvés entre les couples. Les couleurs codent alors les quatre mots les plus fréquemment liés entre un nœud et la totalité de ses correspondances. Comme une couleur est assignée à chaque mot commun, la présence des palettes distinctes à des endroits différents d'un graphe donne une représentation visuelle des glissements thématiques au sein des couples détectés et mis en correspondance (directement ou indirectement). Mais il s'agit plutôt d'une interface d'exploration pour des utilisateurs qui connaissent bien leur corpus et qui sont à l'aise avec ce type de visualisation. Le logiciel est également plus adapté au traitement de textes sans parenté, entre lesquels on observe généralement peu de relations longues et complexes²⁶. Pour notre corpus, en revanche, certains graphes sont très complexes, allant jusqu'à 516 nœuds, tandis que la différence de longueur des textes, prise en compte dans le calcul de la taille de nœuds, pose d'importants problèmes de visualisation. Le graphe ci-dessous (Figure 75), établi à partir des différentes traductions d'Ovide, témoigne de ces difficultés.

²² Voir par exemple Matteo Romanello, « Exploring Citation Networks to Study Intertextuality in Classics », *Digital Humanities Quarterly*, vol. 10, n° 2, juin 2016.

²³ Voir, en ligne, <https://traviz.vizcovery.org/source.html>, ainsi que Stefan Jänicke et al., « Visualizations for Text Re-use ». Nous avons commenté cette interface au III.3 « Présentation des logiciels retenus, TextPAIR et Tracer ».

²⁴ Voir III.3c), n. 60.

²⁵ Ce logiciel en cours de développement au sein de l'équipe ACASA (UMR LIP6), dirigée par Jean-Gabriel Ganascia à Sorbonne Université, n'est pas encore disponible au public. Pour sa présentation, voir Jean-Gabriel Ganascia, « Graphes et intertextualité », *Humanités numériques*, Centre Universitaire Méditerranéen, Nice, 2019; Andrea Del Lungo et Karolina Suchecka, « Projet eBalzac : construire une bibliothèque hypertextuelle des sources intellectuelles », *op. cit.*

²⁶ Une relation est complexe quand un extrait est lié à plusieurs autres, liés à leur tour à d'autres extraits, etc.

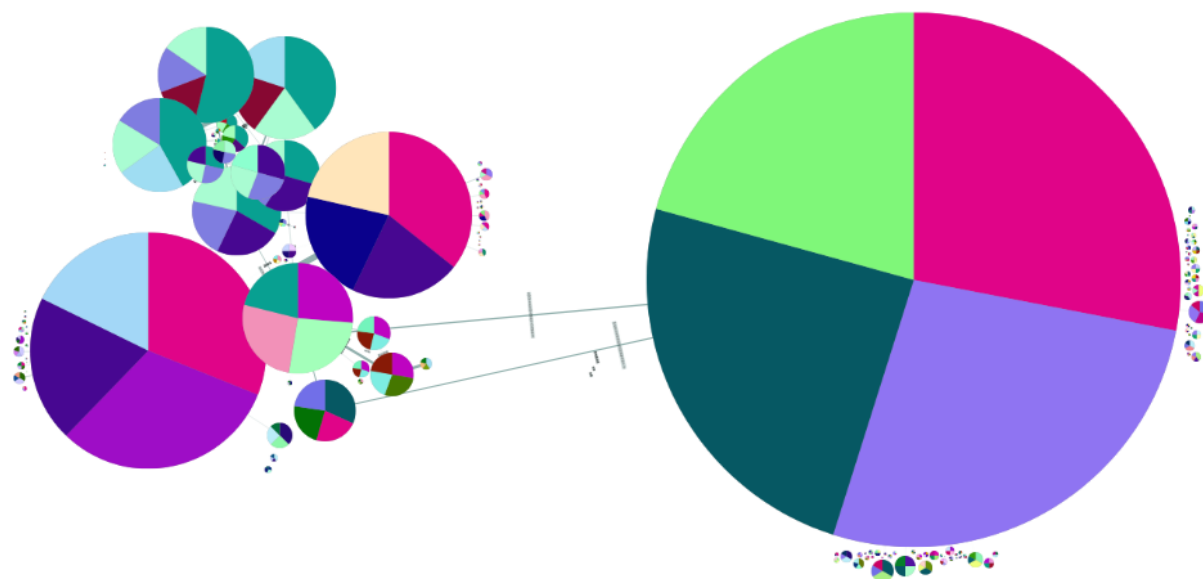


Figure 75. Galaxies : 45 correspondances entre des traductions d'Ovide

Pour donner une vision plus large des correspondances au sein des œuvres traitées, nous nous aidons de graphes, mais nous tâchons d'en rendre la compréhension et l'exploitation intuitives. Nous décidons de proposer un graphe multidimensionnel²⁷ qui permet aux lecteurs d'accéder progressivement à des informations de plus en plus complexes, comme nous le faisons déjà pour les couches d'annotation de la visualisation plus classique. Le premier niveau présente des relations entre des œuvres (Figure 76).

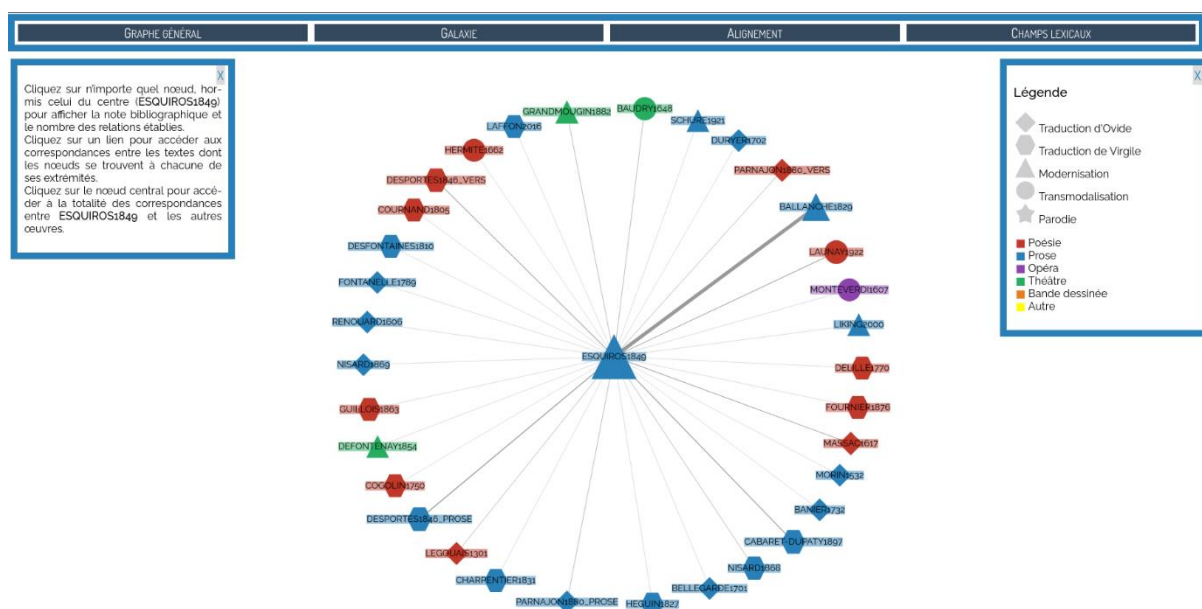


Figure 76. Relations entre Esquiros1849 et d'autres œuvres.

²⁷ Martin Grandjean, « Les réseaux de la coopération intellectuelle. La Société des Nations comme actrice des échanges scientifiques et culturels dans l'entre-deux-guerres », thèse de doctorat, François Vallotton et Bertrand Müller (dir.), Université de Lausanne, 2018.

C'est une visualisation simple où le texte traité apparaît au centre, les textes liés étant répartis en cercle tout autour. Les formes et les couleurs des nœuds sont ici les mêmes que sur la page d'accueil. Elles constituent des repères stables pour les lecteurs. Toutefois, comme les possibilités d'action sont plus complexes, des instructions figurent en haut à gauche de l'écran. Elles indiquent en l'occurrence trois types de clics.

Clic sur un nœud extérieur

Un clic sur un nœud représentant un texte lié affiche sa note bibliographique et le nombre de phrases en correspondance avec le texte traité (Figure 77).

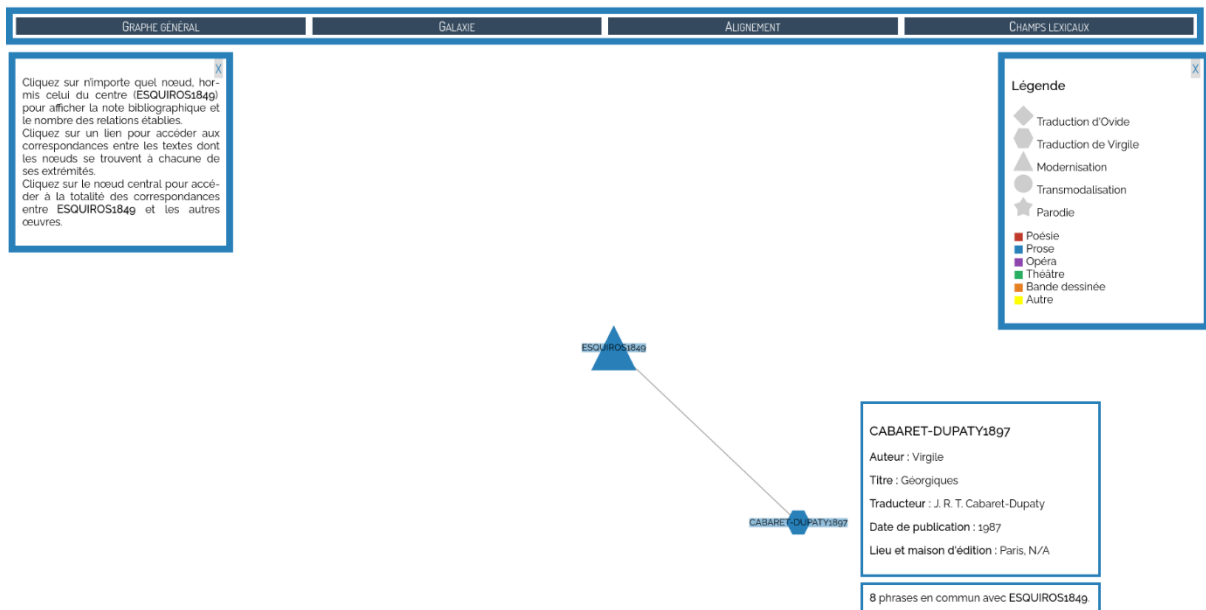


Figure 77. Précisions sur Cabaret-Dupaty1897 et sur sa relation avec Esquiros1849.

Clic sur un lien

Un clic sur un lien affiche les correspondances détectées entre le texte traité et le texte lié (Figure 78). Afin d'augmenter la lisibilité des étiquettes des extraits en correspondance, les couleurs codant le genre textuel ne sont plus présentes qu'en bordure des nœuds.

Pour la plupart des textes, ce deuxième niveau est composé de plusieurs petits graphes, permettant de localiser facilement les passages plus étendus, susceptibles de révéler l'inspiration d'un texte par l'autre ou, *a minima*, relatant le même épisode du mythe. Dans l'exemple ci-dessus, nous constatons que la relation entre les phrases 437-438 de la réécriture Esquiros1849 et les phrases 10-11 de la traduction Cabaret-Dupaty1897 (voir Figure 71-Figure 74) est suivie d'une série des liens entre des phrases voisines. La correspondance proche entre les deux textes peut donc être étendue aux

phrases 437-447 de la réécriture et 10-19 de la traduction, malgré les manques dans la détection automatique qui empêchent la réunion de la totalité du segment au sein d'un même nœud.

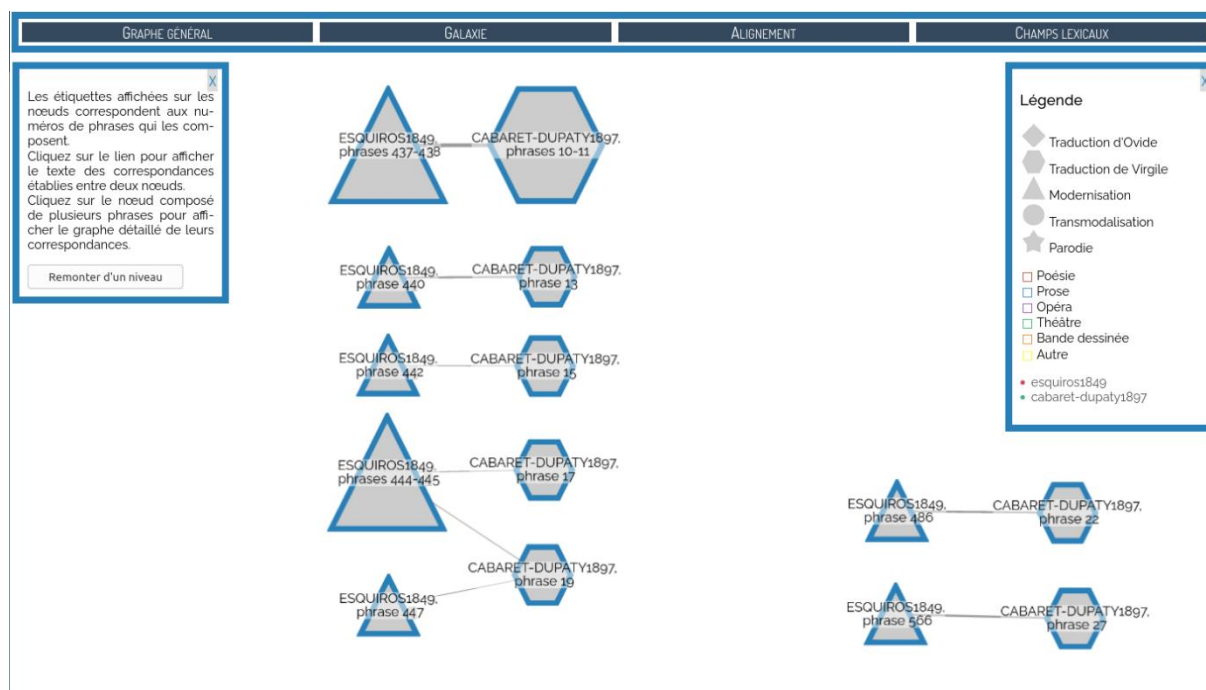


Figure 78. Correspondances regroupées entre Esquiros1849 et Cabaret-Dupaty1897

Un clic sur un lien de cette nouvelle vue affiche une fenêtre de comparaison textuelle. Il est possible d'ouvrir ainsi plusieurs fenêtres, ce qui permet notamment d'étudier le déploiement de la reformulation. L'exemple ci-dessous (Figure 79) montre que la lamentation d'Eurydice précédant sa seconde mort est évoquée dans deux textes assez proches, avec plusieurs reprises mot à mot. Toutefois, la reformulation n'est pas linéaire et les propos qui sont regroupés dans une phrase de Cabaret-Dupaty1897 (phrase 19) apparaissent dans deux phrases d'Esquiros1849 (phrases 445 et 447, séparées par l'exclamation « Adieu ! » en 436, mais sans correspondance).

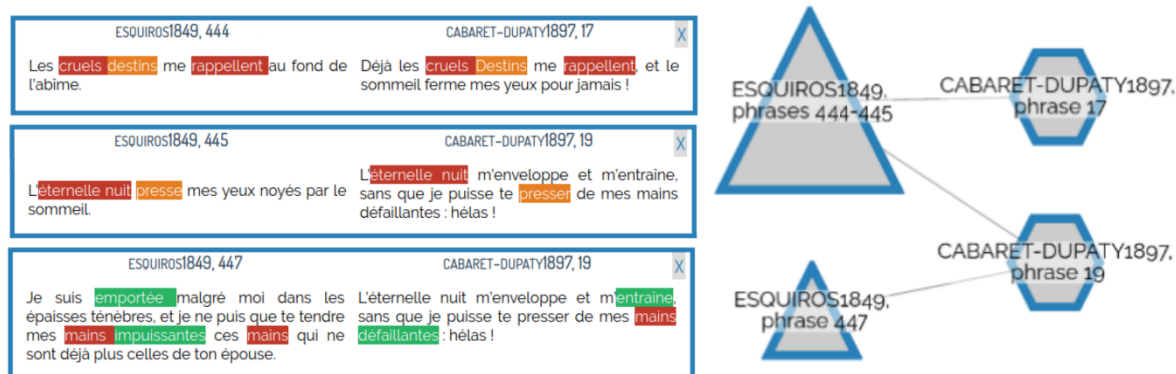


Figure 79. Trois fenêtres de comparaison d'Esquiros1849 et de Cabaret-Dupaty1897.

Clic sur le nœud central

Enfin, un clic sur le nœud central de la vue d'ensemble permet de visualiser la totalité des correspondances détectées entre le texte traité et les autres œuvres (Figure 80). Cette visualisation, bien que plus complexe, est importante du point de vue analytique, même si elle n'intéressera que les lecteurs en situation d'étude ou de recherche ciblée. Pour rendre la représentation aussi lisible que possible, nous lissons les tailles maximale et minimale des nœuds (pour éviter les disproportions observables par exemple dans la visualisation par Galaxies à la Figure 75) et nous optons pour une spatialisation automatiquement optimisée du graphe (qui évite autant que possible la superposition de nœuds, de liens et d'étiquettes). Consciente du risque de chevauchements résiduels, nous laissons toutefois la possibilité d'agrandir les graphes et de déplacer librement les nœuds.

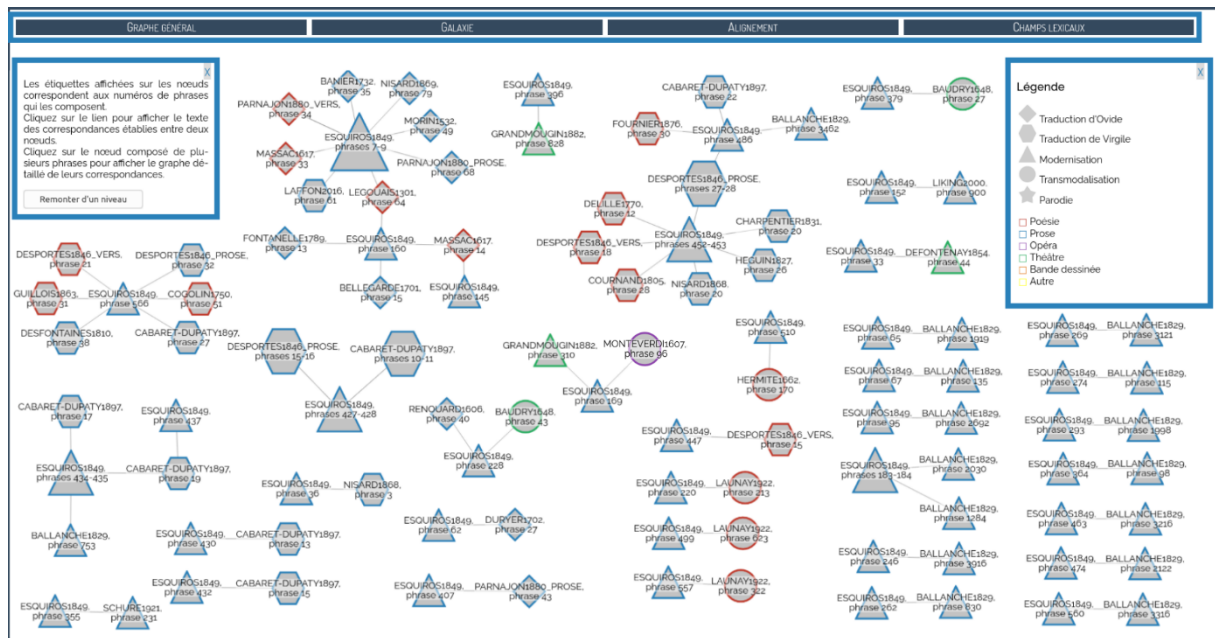


Figure 80. Correspondances regroupées entre Esquiros1849 et 33 autres œuvres

Parmi la totalité des correspondances entre Esquiros1849 et les autres textes, quatre graphes sont composés de 5 à 14 nœuds. Pour la phrase 566²⁸, par exemple, six correspondances sont trouvées avec différentes traductions de Virgile. À la relation avec la traduction Cabaret-Dupaty1897 (Figure 79) s'ajoute celle avec la phrase 753 de Ballanche1829. Les correspondances entre Esquiros1849 et Ballanche1828 sont nombreuses, mais assez disparates : les 12 relations phrase à phrase ne forment pas d'extraits de réutilisation plus étendus, ce qui implique soit qu'il s'agit d'allusions ponctuelles et très explicites, soit que la proximité du contenu ou de la formulation est accidentelle. Dans tous les cas, cette visualisation révèle la nécessité d'une comparaison plus détaillée de ces deux œuvres.

²⁸ « Orphée errait seul dans les glaces de la Scythie, le long des bords du Tanais, et dans les campagnes voisines des monts Riphées éternellement couverts de neige. »

Si une présentation relativement aisée des correspondances générales peut être conservée pour la réécriture Esquiros1849, ce n'est pas toujours le cas. Pour certaines traductions, nous observons de nombreuses correspondances croisées qui composent un seul graphe complexe. Un nœud peut alors regrouper jusqu'à un tiers du texte, et être lié à un grand nombre d'extraits correspondants, liés à leur tour à d'autres nœuds... C'est le cas par exemple de la traduction Cabaret-Dupaty1879 pour laquelle un nœud de 13 phrases initiales est lié à 58 nœuds composés d'entre 1 et 11 phrases et provenant de 31 textes différents (14 autres traductions de Virgile, 7 traductions d'Ovide et 10 réécritures, Figure 81)

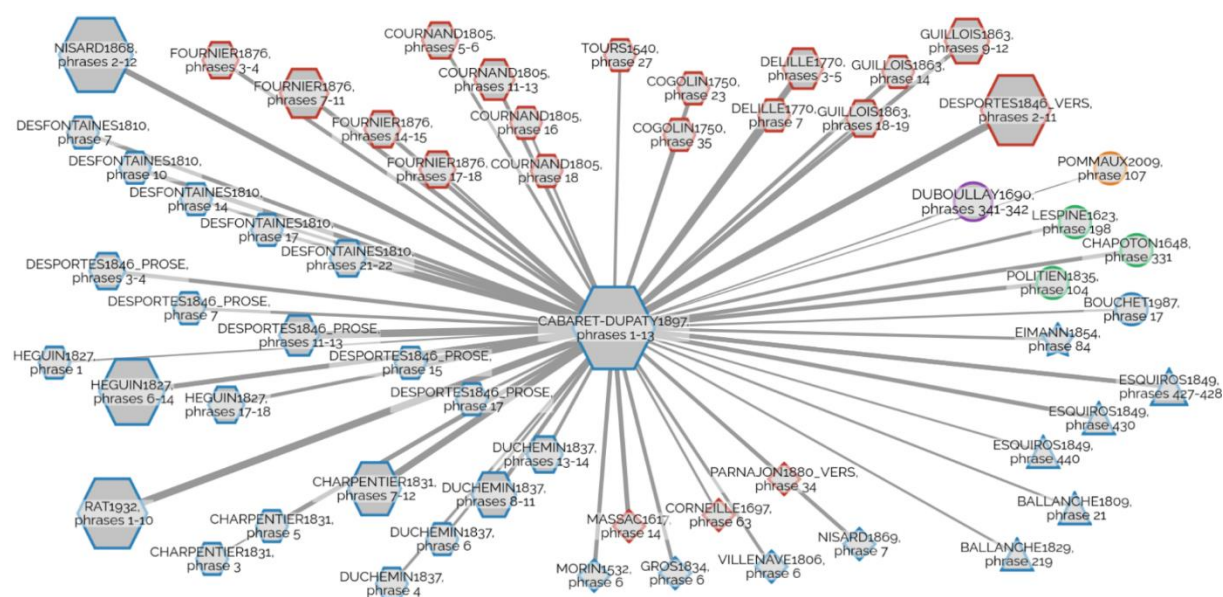


Figure 81. Correspondances regroupées entre Cabaret-Dupaty1879 (1-13) et 31 autres textes

Alors que ce long extrait regroupe plusieurs épisodes du mythe, les textes correspondants peuvent référer uniquement au décès initial d'Eurydice, à l'enchantement des Enfers par Orphée, ou à son regard fatal et à la seconde mort de sa femme. La possibilité d'observer de près cette répartition est analytiquement importante. Ainsi, en cliquant sur un nœud multiphrastique de la vue précédente, nous pouvons accéder au dernier niveau de la visualisation, qui limite la taille de chaque nœud à une phrase et qui propose deux modalités d'affichage. L'une (Figure 82), comparable à la visualisation cumulative, permet d'identifier l'auteur et l'emplacement de la phrase représentée par un nœud (dont la taille est, cette fois, conditionnée par le nombre de relations l'impliquant). L'autre (Figure 83) présente les unités linguistiques communes détectées.

42 relations ont été trouvées pour les phrases 10-12 de Cabaret-Dupaty1897, reliées entre elles grâce aux trois relations communes des phrases 10 et 11, et deux relations des phrases 11 et 12. Les lemmes communs les plus fréquents sont représentés à l'aide de codes-couleur. Parmi eux, des noms d'entités (*Eurydice*, 17 occ. ; *Averne*, 10 occ. ; etc.), des mots à consonance négative (*délire*,

13 occ. ; *peine*, 11 occ. ; etc.) et des verbes-clés de l'épisode de la seconde mort d'Eurydice (*remonter*, 16 occ. ; *revenir*, 15 occ., etc.). L'observation des lemmes communs permet de percevoir les thèmes principaux du graphe. Pour l'analyser, l'utilisateur peut avoir recours tant à ses connaissances linguistiques et littéraires (en accédant aux fenêtres de comparaison textuelle) qu'à ses compétences d'analyse de graphes (*clustering*, densité, degré, etc.).

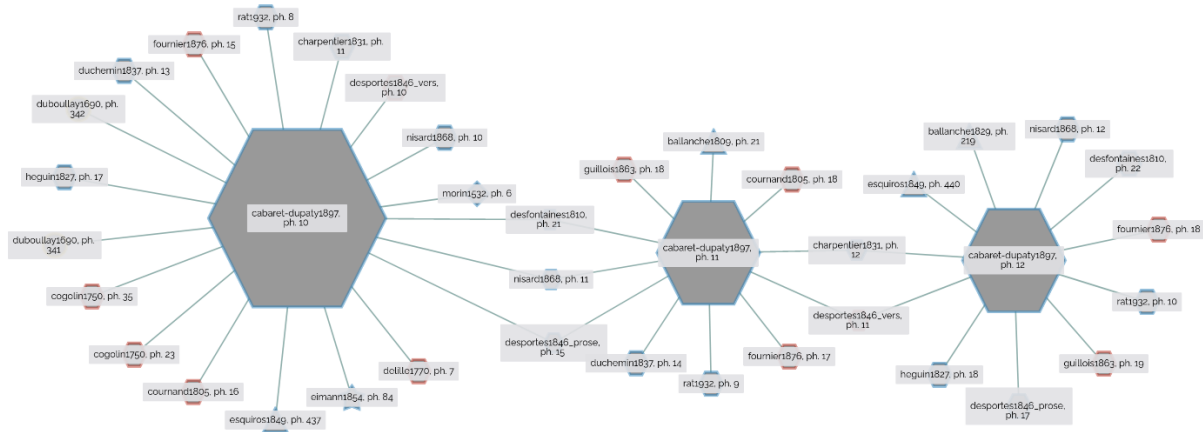


Figure 82. Correspondances détaillées : 42 relations trouvées pour les phrases 10-12 de Cabaret-Dupaty1897

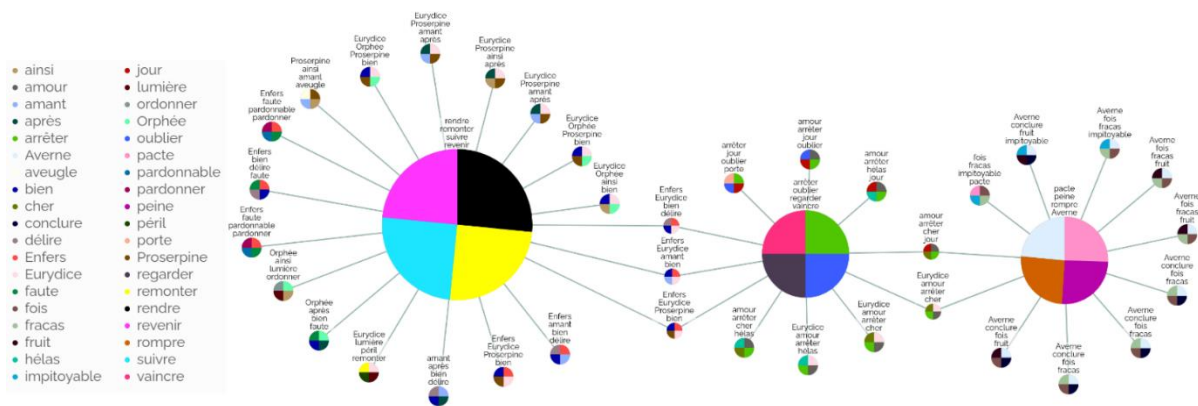


Figure 83. Correspondances détaillées : focus sur les lemmes communs

Alors que, dans les affichages précédents, la visualisation visait à faciliter la navigation en demeurant compréhensible pour tout lecteur, cet objectif passe désormais au second plan : nous considérons que le lecteur qui accède à ce dernier niveau de visualisation possède la capacité de compréhension nécessaire.

Conclusion du chapitre

La cartographie hypertextuelle des réécritures du mythe d'Orphée et Eurydice regroupe ainsi les résultats d'une comparaison intertextuelle d'un ensemble hétérogène d'œuvres en proposant un affichage classique, mais richement annoté, ainsi qu'un autre, spatialisé, construit avec des graphes de correspondances. Évolutive et possiblement exploitable pour différents types de corpus, cette visualisation modulable ambitionne apporter des réponses pertinentes aux besoins et aux attentes des utilisateurs, en permettant de graduer la complexité de ce qui est présenté. Le tableau ci-dessous résume les quatre niveaux d'exploitation proposés pour chaque affichage, le degré de complexité et la nature des informations fournies à chaque niveau correspondant aux trois types de lecture repérés par Heyer, enrichis de la lecture d'étude (*close reading*) ajoutée par Pierazzo²⁹

Tableau 92. Récapitulatif des différentes interfaces et fonctionnalités de l'édition

	Page d'accueil : graphe général de correspondances entre les textes : Figure 68- Figure 69	
	Édition classique hyperannotée	Visualisation en graphes
Niveau 1 (<i>broutage</i>)	Lecture du texte : Figure 70	Relations entre l'œuvre au centre et les textes correspondants (nœuds = textes) : Figure 76- Figure 77
Niveau 2 (<i>foretage</i>)	Exploration des notes intertextuelles : Figure 71	Relations entre l'œuvre centrale et un autre texte (nœuds = extraits textuels) : Figure 78
Niveau 3 (<i>chasse</i>)	Affichage d'annotation sur des mots ou des groupes plus complexes : Figure 72-Figure 73	Relations entre l'œuvre centrale et les textes correspondants (nœuds = extraits textuels) : Figure 80
Niveau 4 (<i>close-reading</i>)	Superposition de deux annotations : Figure 74	Relations à l'intérieur d'un large nœud textuel (nœuds = phrases) : Figure 82-Figure 83

Notre interface s'éloigne des habitudes éditoriales, alors que nous observons plutôt une tendance à la standardisation des interfaces avec des outils comme Omeka³⁰, TEI Publisher³¹, etc. Elle peut désorienter, voire rebuter des lecteurs habitués à l'imprimé ou aux éditions skeuomorphes. Mais, simultanément, nous ne nous attendons pas à la même homogénéité dans l'architecture du Web, où chaque site est différent. Maintenant que l'imitation de l'imprimé a permis l'acceptation du numérique, il est temps de se montrer « disruptif » et d'introduire de nouvelles fonctionnalités.

L'édition comparative numérique que nous proposons est novatrice, mais elle adopte certains principes en usage dans l'édition numérique. Par exemple, afin de permettre d'appliquer ce développement à d'autres projets, l'interface peut être générée automatiquement à partir d'un corpus

²⁹ Voir n. 11 et 12.

³⁰ Voir <https://omeka.org/>. Nous rappelons que notre corpus peut également être exploré dans une interface plus "classique" conçue avec Omeka, voir 0.4 « Numériser un corpus dans l'esprit des humanités numériques ».

³¹ Voir <https://teipublisher.com/index.html>.

structuré en XML-TEI et des résultats des traitements opérés par TextPAIR et Tracer. Par ailleurs, les données produites sont disponibles en accès libre sur les plateformes dédiées.

Même si notre approche n'est pas applicable telle quelle à tout type de corpus, même si elle nécessite quelques ajustements, souvent manuels, pour améliorer la qualité des données, et même si l'interface générée s'expose au risque d'obsolescence (du fait de mises à jour des langages informatiques, des navigateurs, et des systèmes d'exploitation), nous défendons la pertinence éditoriale et analytique de ce projet et sa capacité à être transposable à d'autres.

4. Quelques perspectives pour le futur du projet

Dans le cadre de ce projet doctoral, un vaste corpus des réécritures et des traductions du mythe d'Orphée et Eurydice a été collecté et numérisé. Ses composantes sont décrites et structurées à l'aide des formats de catalogage universels pour faciliter leur réutilisation. Nous avons également déposé la totalité des données dans Nakala où elles peuvent être collectées, référencées et préservées. Plusieurs logiciels de détection automatique des correspondances textuelles ont été testés, avant d'en choisir deux, qui se sont révélés les plus performants et complémentaires. Pour améliorer leurs performances, nous avons conçu des lexiques complexes (des lemmes, des entités nommées, etc.). Même si ces lexiques ont été en grande partie générés automatiquement, nous les avons corrigés manuellement. Une série d'algorithmes a ensuite été développée pour un traitement enrichi d'un corpus littéraire avec deux logiciels de détection automatique des correspondances textuelles, ainsi qu'un prototype d'édition hypertextuelle incluant une visualisation au moyen de graphes des connaissances. Enfin, nous avons évalué les résultats ainsi obtenus et décrit notre démarche.

Cette démarche expérimentale est extrêmement chronophage. Relativement peu de développements existants ont pu être exploités et certains aspects de notre travail sont inédits. Par ailleurs, nous sommes limitée par nos propres compétences, qu'elles soient éditoriales, informatiques ou analytiques. Le développement rapide des nouvelles technologies extrêmement performantes de l'intelligence artificielle et l'apparition des grands modèles de langages est survenu au cours de notre projet. Même si, pour l'instant, nous n'avons pas trouvé de solutions performantes de l'IA pour améliorer ce que nous proposons, cette question mérite de s'y pencher de manière approfondie.

En somme, nous ne sommes pas en mesure de présenter, dans le temps restreint du doctorat et en travaillant seule, un projet entièrement abouti. Nous nous arrêtons donc à une étape satisfaisante et mettons tout en place pour que notre développement puisse être enrichi, amélioré et réutilisé pour d'autres corpus et espérons nous-mêmes pouvoir continuer ce travail dans le futur.

Dans ce chapitre, nous ouvrons trois perspectives pour le futur du projet. Pour les deux premières, la prise en compte des séries traductives et l'amélioration du traitement des synonymes, nous proposons quelques observations concrètes sans aboutir à les intégrer au développement existant. Pour la dernière perspective, nous ouvrons quelques pistes pour l'analyse approfondie des résultats en montrant un échantillon de ce que nous obtenons pour trois sous-ensembles du corpus : les opéras, les parodies et la littérature de jeunesse. Ce faisant, nous espérons démontrer l'utilité de notre développement tant pour l'analyse littéraire, que la traductologie, la linguistique du corpus et la lexicographie et encourager de futures recherches et collaborations impliquant notre projet de l'édition numérique intertextuelle des réécritures du mythe d'Orphée et Eurydice.

a) Prise en compte des séries traductives¹

Nous avons vu *supra* que le taux d'erreurs des mises en relations d'extraits textuels reste élevé et que le nombre d'extraits impliqués est trop important pour examiner chaque couple. Le système de post-traitement automatique des résultats permet d'écarter une partie des détections fautives et d'examiner la pertinence des couples restants à l'aide d'une annotation des entités communes (mots, lemmes, racines ou synonymes)², mais certaines difficultés persistent.

Nous avons également constaté l'importance des traductions des sources antiques au sein des résultats. En effet, pour la majorité des réécritures de notre corpus, le mythe d'Orphée et Eurydice est le seul point commun identifié, et c'est le retour vers ses sources antiques qui permet d'observer et d'analyser tant les proximités que les différences dans la réinterprétation du récit original. Prenons l'exemple de la nouvelle d'Eimann1854 qui cite explicitement la traduction de Virgile (celle de Delille1770³). Nous aurions pu supposer que c'est également sur cette traduction que l'auteur s'appuie quand il reformule le mythe :

Là, il chanta ; ses chants émurent l'Enfer, attendrirent Proserpine, et enfin il revenait
sur la terre, précédant Eurydice, – qu'on avait bien voulu lui rendre, à la condition qu'il
ne la regarderait qu'à la lumière, – lorsque

Presqu'aux portes du jour, troublé, hors de lui-même,
Il s'arrête, il se tourne... il revoit ce qu'il aime !

Mais, si les extraits cités sont correctement reliés par nos outils à la traduction de Delille1770, nous observons que pour la phrase qui introduit la citation, quatre correspondances ont été trouvées avec d'autres textes, dont deux traductions en prose de Virgile et deux traductions d'Ovide, une en vers et une en prose (Figure 84).

Ces dernières relations, dont deux se font avec des textes postérieurs à la nouvelle, montrent que le texte d'Eimann1854 diffère de la version poétisée de Delille1770 et incitent à chercher d'autres influences. Plus encore, si nous nous focalisons sur les quatre extraits des traductions liées à la phrase 84 d'Eimann1854, nous observons qu'elles entretiennent également des relations avec d'autres réécritures. Cabaret-Dupay1897, par exemple, est mis en correspondance avec Esquiros1849, Banier1732 est en relation avec Montardre2013 et Laffon2016. Même si ces réécritures ne sont pas liées directement, elles le sont indirectement, grâce à leurs relations avec les traductions.

¹ Cette section reprend partiellement notre article (co-signé avec N. Gasiglia) « TAL et Littérature comparée. Détection automatique des correspondances textuelles entre les réécritures d'un mythe », *Actes de la 29^e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier TAL et Humanités Numériques (TAL-HN)*, Avignon, France, ATALA, 2022, p. 88-98, en ligne, <<https://aclanthology.org/2022.jeptalnrecital-humanum.10>>. Les données et visualisations présentées se trouvent dans les annexes de ce chapitre, [IV 4 perspectives](#).

² Voir IV.1 « Post-traitement des résultats de la détection automatique des correspondances ».

³ Voir II.1 « Intertextualité dans la théorie littéraire ».

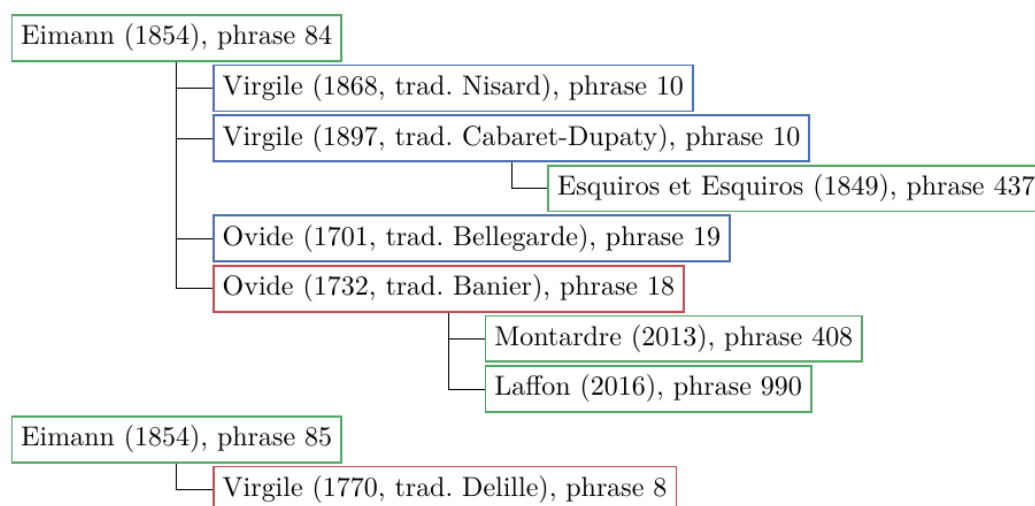


Figure 84. Diagramme simplifié des relations détectées avec les phrases 84-85 d'Eimann 1854

Notre développement permet déjà d'explorer ce type de correspondances, notamment grâce à la visualisation en graphes. Mais nous pensons que les séries traductives peuvent aider davantage à l'amélioration et l'enrichissement des résultats initiaux obtenus avec TextPAIR et Tracer.

En effet, en soumettant uniquement le corpus des traductions à ces logiciels, nous constatons que la précision des résultats est bien plus élevée que pour le corpus des réécritures (confronté à lui-même et aux traductions). Sur 860 couples détectés par Tracer, seulement 12 sont invalides. Nous post-traitons les 848 couples pertinents pour faciliter l'exploitation des résultats enrichis et visualisons les résultats à l'aide de graphes. Cela nous permet de délimiter de manière précise différents segments du récit mythologie et d'extraire des mots-clés les plus saillants de chaque épisode. Ces informations nous paraissent précieuses pour l'amélioration des futures détections des correspondances au sein du corpus des réécritures.

Graphes pour délimiter des segments du mythe

La totalité des relations entre les traductions de Virgile constitue un seul graphe très complexe, au sein duquel certains textes sont regroupés en leur quasi-totalité au sein d'un même nœud (Figure 85). Cela prouve une forte proximité entre ces textes et une précision plutôt satisfaisante de la détection des correspondances pour ce sous-ensemble. Les relations entre les traductions d'Ovide se répartissent en sept graphes complexes (c'est-à-dire constitués de plus de 10 nœuds) et une grande quantité de graphes très simples, de deux à quatre nœuds. Cette répartition montre certes qu'un travail d'amélioration de la détection est encore nécessaire non seulement pour le sous-corpus des réécritures, mais aussi pour celui des traductions. Cependant, elle nous permet de nous focaliser sur les caractéristiques de chaque graphe afin de dégager les proximités thématiques de leurs nœuds.

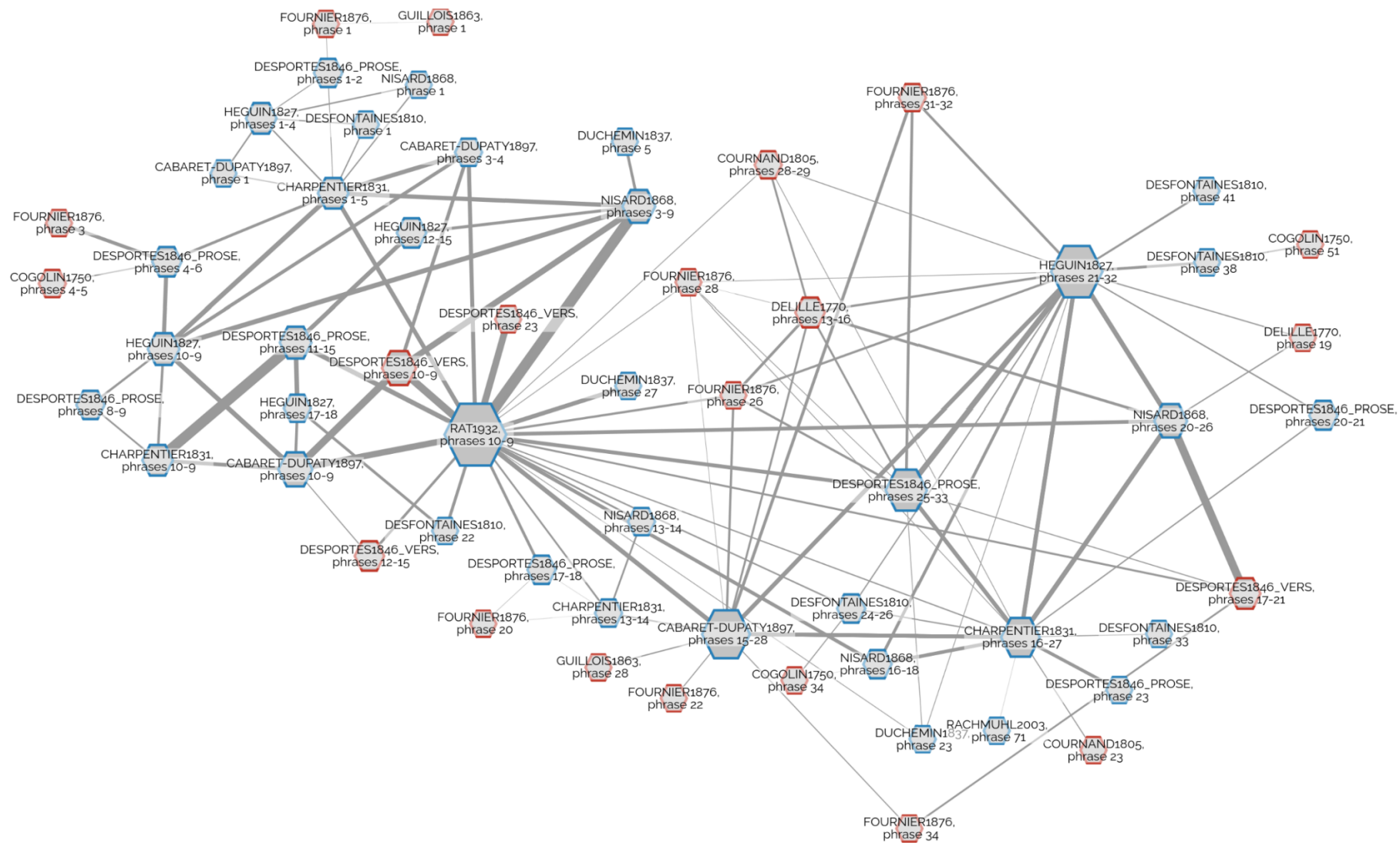


Figure 85. Graphe regroupant la totalité des relations détectées entre les traductions de Virgile

Le graphe à la Figure 86, par exemple, renvoie dans sa totalité à l'épisode de la seconde mort d'Eurydice. Des relations d'une à trois phrases y sont établies parmi 14 traductions publiées à partir du XVI^e siècle. Six traductions, dont les quatre en ancien français, n'apparaissent pas dans ce graphe.

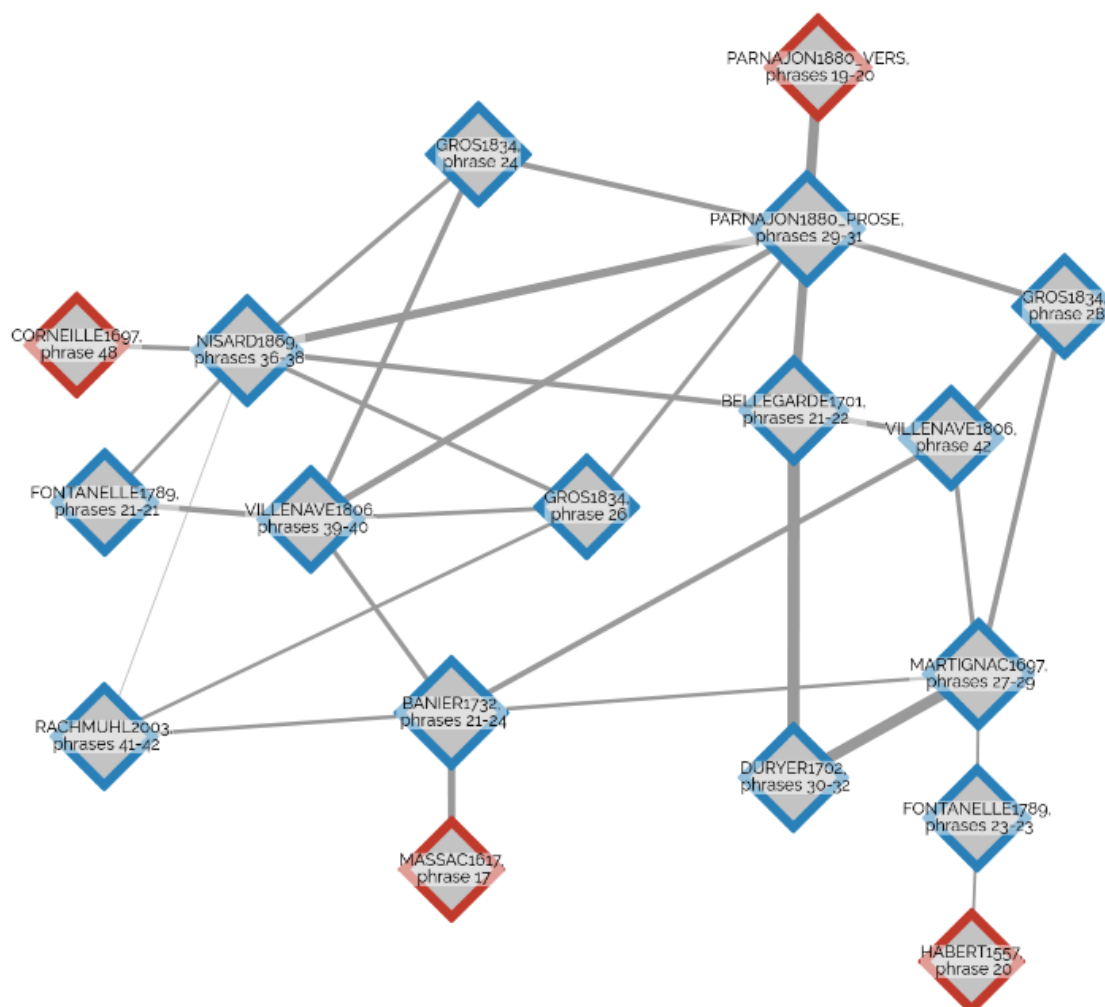


Figure 86. Graphe de 18 nœuds évoquant la seconde mort d'Eurydice dans les traductions d'Ovide

Pour déterminer s'il y a lieu de prendre en compte d'autres relations avec des extraits d'une des traductions manquantes relatifs à cet épisode, nous couplons cette visualisation cumulative à une autre, qui limite la taille de chaque nœud à une phrase, mais qui propose deux modalités d'affichage, une qui permet d'identifier l'auteur et l'emplacement de la phrase représentée par un nœud, dont la taille est, cette fois, conditionnée par le nombre de relations l'impliquant, et une autre, qui présente les unités communes détectées.

Une partie des phrases de notre graphe cumulatif de 18 nœuds, représentant des extraits, se retrouve dans un réseau de 11 nœuds, dont les mots-clés les plus fréquents sont, notamment, *aimer*, *Eurydice*, *mourir* et *plaindre* (Figure 87). Nous le retrouvons aussi dans un graphe de 9 nœuds. Celui-ci regroupe les quatre manuscrits anciens, qui restent isolés dans nos résultats puisqu'ils n'ont pas

de relations directes avec les traductions modernes. La présence des mots-clés identiques, ou seulement proches, nous permet donc de regrouper les deux ensembles de textes et de considérer qu'ils évoquent tous la seconde mort d'Eurydice.

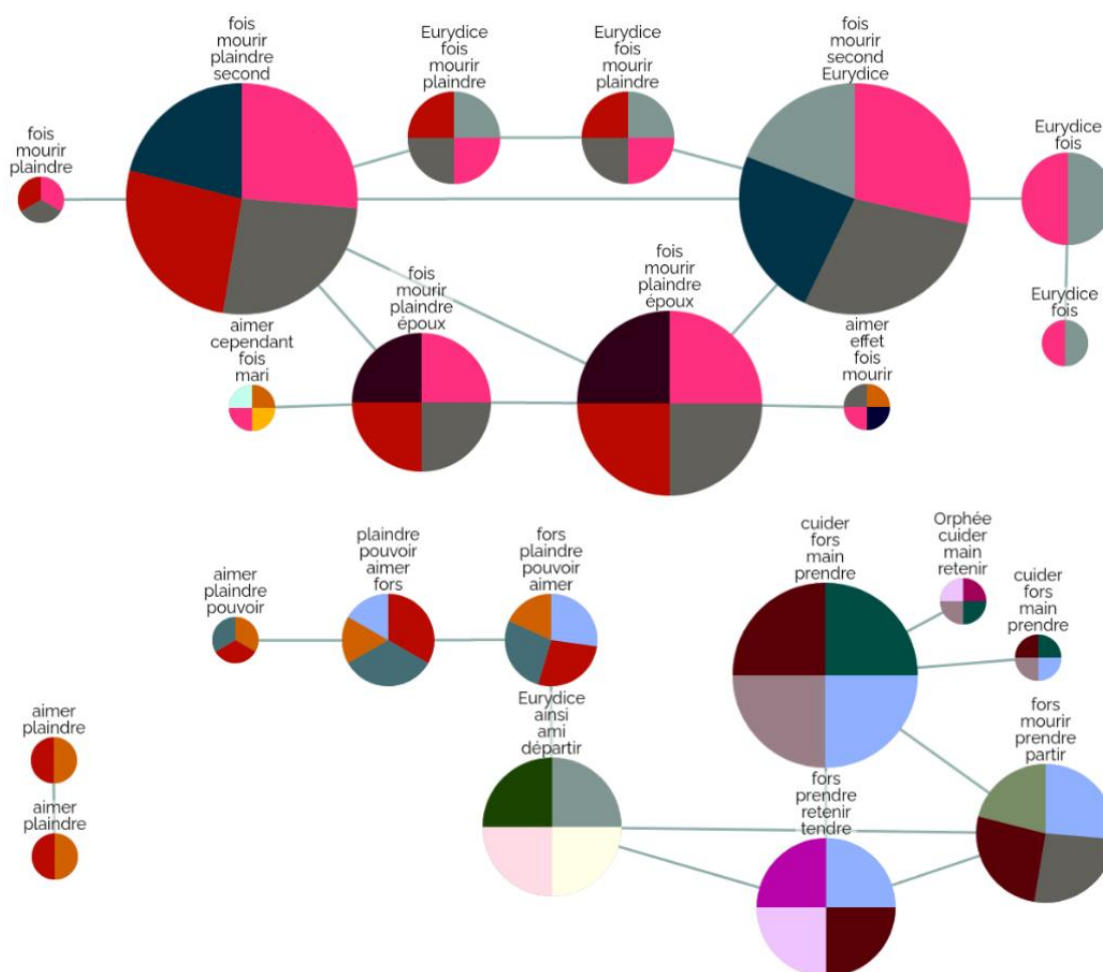


Figure 87. Rapprochement de trois graphes dont les correspondances partagent les mêmes mots communs

Exploitation des mots-clés extraits des relations entre traductions

Ce regroupement permet également une analyse plus subtile des entités communes détectées. 445 relations lexicales ont été trouvées pour 57 phrases de 18 textes. 130 lemmes communs sont recensés. Pour 30 d'entre eux, nous comptons au moins 20 occurrences. *Second*, *mourir* et *plaire* sont les plus fréquents (34-32 occ.), suivis des noms et de périphrases des deux amants (*Orphée*, *Eurydice*, *ami*, *époux*, 14-13 occ.) et de verbes (*tendre*, 21 occ. ; *entendre*, 20 occ., *retourner*, 14 occ., *aimer*, 12 occ. ; *embrasser*, 12 occ. et *regarder*, 12 occ.).

Ces lemmes communs peuvent notamment être exploités dans l'analyse des réécritures qui développent certains épisodes de manière importante. C'est surtout le cas des pièces de théâtre et des opéras qui se focalisent principalement sur l'amour d'Orphée et Eurydice. Chez Gluck (Moline1774) la sortie des Enfers et la seconde mort d'Eurydice, par exemple, sont racontées dans la première

scène du troisième acte, ce qui correspond à 38 répliques découpées en 110 phrases. Dans les séries traductives, le traitement de l'épisode ne dépasse pas 15 phrases chez Virgile et 8 chez Ovide. Cette différence d'échelle de développement handicape Tracer et TextPAIR.

En revanche, nous constatons que, parmi 385 mots forts présents dans les 110 phrases de Moline1774, 175 se trouvent dans les mots-clés détectés dans les traductions de Virgile et d'Ovide pour cet épisode (71 d'Ovide, 80 de Virgile et 24 occ. des noms propres de deux amants). L'analyse de leur répartition permet d'observer les extraits de Moline1774 qui se rapprochent le plus des traductions, mais il est également intéressant de se pencher sur les unités linguistiques qui ne font pas partie des mots-clés, car elles rendent bien compte des modifications majeures introduites dans la réécriture analysée.

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 avant d'obtenir une explication. En effet, nous recensons, d'un côté, une quantité importante de mots à consonance négative, exprimant l'incompréhension et le refus d'Eurydice (*indifférence, barbare, ingrat, jaloux, injustice, affliger, abhorrer, soupçon, outrager, barbarie, ennemi, fuir, abandonner*) et, de l'autre, beaucoup de mots illustrant le désarroi d'Orphée (*contrainte, effroi, refuser, implorer, secret, trouble, désespérer, désoler, martyre*).

Toutefois, alors que cette méthode s'avère exploitable pour l'analyse des réinterprétations plus descriptives, suffisamment longues pour ce type de comparaisons linguistiques massives, elle est moins probante pour les réécritures plus courtes, mais fortement reformulées. C'est par exemple le cas du poème *La Lyre d'Orphée* de Tristan l'Hermite (Hermite1662). Alors que plusieurs relations ont été détectées entre ce poème et différentes traductions d'Ovide, la réécriture de la seconde mort d'Eurydice reste sans correspondance. L'auteur effectue, pour cet extrait, un mélange particulier des versions :

De ton affection ma disgrâce est éclos,
Et si j'en hais l'effet, j'en dois aimer la cause.
Encore que tes yeux me donnent le trépas,
Cette atteinte me tue et ne me blesse pas :
Ta foi, charmant époux, n'en peut être blâmée ;
Tu n'aurais point failli si j'étais moins aimée

Chez Virgile, Eurydice se plaint de la faute d'Orphée, avant de disparaître pour toujours. Chez Ovide, elle ne formule aucune plainte, parce que la faute d'Orphée n'est qu'une preuve de la puissance de son amour. L'Hermite donne la parole à Eurydice, mais pour louer la force du sentiment de son époux. À cela s'ajoute un langage fortement poétisé (« Console-toi de même et ne plaint point ma cendre/ Dans les torrents de pleurs que tu pourrais épandre ») et très métaphorique (« Ta curiosité trop peu considérée/ Me remet dans les fers dont tu m'avais tirée ») et l'introduction de

nouveaux personnages qui remplacent ceux habituellement présents dans cet épisode, comme Hadès ou Perséphone (« Pourquoi du vieux Minos n’as-tu gardé les lois ? »).

Par conséquent, seulement 59 des mots-clés extraits des traductions ont été trouvés sur 193 mots forts, dont une occurrence de chacun de deux noms propres. Contrairement à ce qui se produisait pour le texte de Moline¹⁷⁷⁴, l’analyse des 134 mots restants n’apporte pas d’informations quant à la réinterprétation du mythe, car ce sont majoritairement des synonymes, périphrases ou métaphorisations des mots-clés déjà repérés pour cet épisode. L’application de notre traitement des unités lexicales aux mots restants permet d’apparier automatiquement 81 mots forts de cet extrait d’Hermite¹⁶⁶² aux mots-clés du mythe (6 correspondances au niveau des racines et 75 relations synonymiques), ce qui restreint le nombre des mots non appariés à 43. Même si, pour certains verbes, comme *obtenir* ou *garder*, qui sont très polysémiques, plusieurs équivalents sont extraits du *DES*, dont certains ne correspondent pas aux sens attestés dans le poème, un synonyme au moins était pertinent par rapport au mot fort du texte traité.

Nous pensons que l’intégration de ce que nous observons au sein du corpus des traductions permettrait d’améliorer davantage les résultats de la détection automatique, l’alignement des entités communes, ainsi que l’observation et l’analyse de réinterprétations plus subtiles et plus implicites : poétisations, métaphorisations, mélanges de deux versions, ou encore modifications de la trame originelle du mythe. Par exemple, la restriction de la recherche des synonymes parmi les mots attestés et mis en relation dans les séries traductives permet potentiellement d’améliorer la pertinence des appariements synonymiques. Par exemple, pour l’expression « sur son visage il détournât ses yeux », nous avons détecté tout d’abord la présence du mot « œil » parmi les mots-clés de l’épisode, puis la relation synonymique entre *détourner* et *se retourner*. Mais, dans le futur, nous aimerions pouvoir 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 cooccurents 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, comme le *DES*, leur caractère cumulatif (une liste de synonymes, pas une par sens) pose problème. Nous nous y penchons en détail dans la prochaine section de ce chapitre.

b) Amélioration du traitement des synonymes¹

La lexicographie numérique élargit les raisons qui poussent à consulter des dictionnaires² et en facilite de nouvelles utilisations, notamment à l'aide d'outils informatiques d'extractions ciblées d'informations. Dans cette section, nous proposons d'expérimenter les possibilités offertes par le recours aux dictionnaires de synonymie numériques au sein de notre projet et proposons des pistes pour exploiter pleinement leur potentiel informationnel.

Comme nous l'avons déjà souligné, les ressources lexicales talistes exploitables pour la détection de la synonymie (comme *WOLF*³) et les méthodes statistiques (comme le « plongement lexical »⁴) s'avèrent peu opératoires pour notre petit corpus. Afin d'atteindre une annotation aussi fine que possible, avec des synonymes pertinents pour chaque mot-occurrence qui peut en accepter, nous heurtons parfois à des items très polysémiques pour lesquels les choix de synonymes à retenir peuvent manquer d'évidence, même en nous tournant vers des dictionnaires spécialisés traitant spécifiquement ces relations lexicales. En effet, notre objectif final est d'automatiser l'appariement d'une sélection pertinente de synonymes à chaque mot-occurrence lors du prétraitement du corpus. D'une part, cela implique de disposer d'un dictionnaire interrogeable par un programme et proposant des articles subdivisés en sous-listes de synonymes en fonction des emplois pour lesquels les substituts sont valides, comme ce que proposent en principe les dictionnaires de synonymie distinctive⁵. D'autre part, il faut aussi savoir exploiter informatiquement les sélecteurs sémantiques ou contextuels qui introduisent souvent les sous-listes pour les confronter aux indices repérables dans le contexte de chaque mot-occurrence à annoter.

Dans la suite de cette section, nous évaluons la pertinence de la méthode d'appariement implémentée actuellement dans notre chaîne de post-traitement appuyée sur le *Dictionnaire électronique des synonymes*⁶, qui est un dictionnaire de synonymie cumulative de la langue générale. Cet outil n'est pas optimal pour notre objectif, nous le montrerons, mais il est aisé à exploiter dans une chaîne de

¹ Cette section reprend l'article co-écrit avec N. Gasiglia « Exploration fonctionnelle de dictionnaires de synonymes pour un projet d'humanités numériques », *Linx. Revue des linguistes de l'université Paris X Nanterre*, n° 86, Département de Sciences du langage, Université Paris Ouest, septembre 2023, en ligne, <<https://journals.openedition.org/linx/10203>>.

² Schématiquement : aide à la compréhension, à l'expression ou à l'amélioration de la compétence métalinguistique apportées à des utilisateurs humains.

³ *WordNet Libre du Français*, Inria, AND EDyLex, voir <https://pauillac.inria.fr/%7Eesagot/index.html#wolf> et Benoît Sagot, « Représentation de l'information sémantique lexicale : le modèle wordnet et son application au français », *Revue Française de Linguistique Appliquée*, vol. XXII, 2017, en ligne, <<https://inria.hal.science/hal-01583995>>.

⁴ Par exemple avec l'algorithme *Word2vec* développé par le groupe de recherche de Google (voir Tomas Mikolov et al., *Efficient Estimation of Word Representations in Vector Space*, *op. cit.*)

⁵ C'est souvent une liste unique de synonymes qui est fournie dans les dictionnaires en ligne, on parle alors de la synonymie cumulative. Exploiter une liste unique de synonymes impliquerait de savoir repérer en son sein les substituts pertinents pour chaque mot-occurrence du corpus susceptible de recevoir une annotation synonymique pertinente.

⁶ Voir IV.1b) « Détection des noyaux des syntagmes et les unités lexicales liés, puis calcul du score de ressemblance ».

traitements numériques et il nous fournit un point de comparaison pour la suite des expérimentations. Nous nous tournons ensuite vers trois dictionnaires de synonymie distinctive, proposés par les principaux éditeurs du marché français et dont il existe une version numérique au moins, le *Dictionnaire des synonymes et nuances* du Robert (2020⁷), le *Grand Dictionnaire Synonymes & Contraires* de Larousse (2004⁸) et le *Dictionnaire des synonymes* d'Hachette Éducation (1996⁹), qui proposent chacun une manière différente d'introduire des sous-ensembles de synonymes. Nous comparons leur potentiel pour nous aider à révéler les relations synonymiques valides dans les contextes d'emploi d'une petite sélection de mots-occurrences de notre corpus. Nous proposons enfin d'expérimenter une méthode d'appariement automatique qui exploite les sous-listes d'un de ces dictionnaires de synonymie distinctive en s'appuyant sur le repérage de cooccurents extraits des contextes et confrontés à ceux relevés dans un dictionnaire de cooccurents, qui ne sera pas un produit d'éditeur du secteur éditorial de référence, mais un dictionnaire offert à la consultation au sein de logiciels d'aide à la correction, *Antidote*¹⁰ et le *Robert correcteur*¹¹.

Méthode d'appariement appuyée sur la synonymie cumulative

Commençons par observer concrètement un couple de correspondances détecté entre la traduction de Virgile par Héguin1827 et celle de Nisard1868¹² (Figure 88).

	HÉGUIN1827, PHRASE 9	NISARD1868, PHRASE 6
forme~forme	Il fit plus : il osa ¹ franchir ² les	Il osa ¹ même descendre ² dans les
lemme~lemme	gouffres ² du Ténare ³ , ses forêts ⁵	gouffres ² du Ténare ³ ; il vit le
radical~radical	noires d'épouvante ⁷ , et le seuil du	profond royaume de Pluton ⁹ , et ses
synonyme~synonyme ou mixte	palais de Pluton ⁹ ; aborder les	bois ⁵ que remplissent l'horreur ⁷ et
	Mânes ¹⁰ et leur formidable ¹¹ roi ¹² , et	les ténèbres , et les mânes ¹⁰ , et le
	tenter de fléchir ces cœurs ¹³	terrible ¹¹ roi ¹² des enfers, et ces
	impitoyables que des prières ¹⁴	cœurs ¹³ que les prières ¹⁴ des
	humaines ¹⁵ n'ont jamais ¹⁶ fléchis ¹⁷ .	humains ¹⁵ n'ont jamais ¹⁶ attendris ¹⁷ .

Figure 88. Relations lexicales entre un couple de deux traductions de Virgile

Pour ces deux extraits, notre algorithme a détecté 8 appariements de formes identiques (*osa*, *gouffres*, *Ténare*, *Pluton*, *roi*, *cœur*, *prières* et *jamais*) et un au niveau de lemmes (*Mânes*~*mânes*). Nous avons également pu aligner l'adjectif *humaines* avec le nom *humains*. Enfin, quatre relations synonymiques ont

⁷ Dominique Le Fur (dir.), *Dictionnaire des synonymes et nuances*, Paris, Le Robert, coll. « Collection les usuels », 2005.

⁸ Chantal Lambrechts (dir.), *Grand Dictionnaire Synonymes & contraires*, Paris, Larousse, 2005.

⁹ Marc Baratin et Marianne Baratin-Lorenzi, *Dictionnaire des synonymes*, Paris, Hachette Éducation, 1996.

¹⁰ Voir <https://www.antidote.info/fr> et <https://www.antidote.info/fr/documentation/guide-utilisation/avant-propos?guide=a11>.

¹¹ Voir <https://robert-correcteur.lerobert.com/aide-redaction.php>.

¹² Nous incluons l'alignement des séries traductives présenté à la section précédente, car il contient beaucoup de couples proches, facilitant l'exposition des difficultés que nous rencontrons avec l'alignement des synonymes.

été trouvées : *forêts~bois*, *épouvante~horreur*, *formidable~terrible* et *fléchis~attendris*. Idéalement, elles auraient pu être complétées par trois autres relations (en violet sur la Figure 88) :

- les forêts peuvent être « noires d’épouvante » ou remplies par « l’horreur et les ténèbres ». Alors que *épouvante* et *horreur* sont bien mis en correspondance, un lien apparaît aussi entre *noires* et, conjointement, *ténèbres* et *remplissent*, alors que, si les ténèbres sont noires, la relation synonymique entre *noir* et *remplir* ne serait valide que dans un contexte où un espace serait saturé d’un contenu (un lieu serait noir de monde ou rempli d’une foule dense)¹³ ;
- les Enfers sont le *seuil du palais* ou le *profond royaume* de Pluton ;
- *franchir* et *descendre* sont équivalents dans ce contexte, mais partagent peu de traits sémantiques communs, notamment dans la langue générale.

Les logiciels ont du mal à détecter ces reformulations, comme les métaphores et les périphrases.

Le verbe *descendre*, qui nous intéressera dans la suite de cette section, est un des mots-clés les plus saillants du mythe. Très polysémique, il pose d’importants problèmes pour l’appariement des unités lexicales synonymes. Différentes traductions d’Ovide l’emploient majoritairement dans quatre contextes (Tableau 93) pour lesquels nous présentons les traductions qui utilisent *descendre*, combien cela représente de traductions sur les 21 du corpus et quels sont les autres équivalents traductionnels rencontrés.

Tableau 93. Contextes d’emploi du verbe *descendre* dans les 21 traductions d’Ovide et alternatives traductionnelles

Texte d’Ovide	Exemple de traductions avec <i>descendre</i> et part de ce choix d’équivalent traductionnel	Autres équivalents traductionnels présentés au sein d’un exemple et part de chacun
« ad Styga Taenaria est ausus descendere porta »	« Il ose descendre vers le Styx, par la porte du Ténare » 20/21	« il osa franchir » 1/21
« o positi sub terra numina mundi, in quem reccidimus »	« Ô divinités du monde souterrain, où nous descendons » 8/21	« dans lequel nous retombons », 4 occ. « où retombe tout ce qui naît pour mourir » « où tous les mortels sont précipitez » « où les mortels enfin vont étouffer leurs jours » « qui voyez tout le genre humain arriver dans ces lieux » « où l’inflexible mort tour à tour nous attire » « a ce submise est toute humanité », 2 occ. « auxquels nous appartiendrons tous un jour » « où tout ce qui respire doit se rendre » 13/21
« Eurydicenque uocant ; umbras	« Ils appellent Eurydice : elle était parmi les ombres récemment descendues »	« ombres (venues/ arrivées) (nouvellement/ récemment) », 8 occ. « ombres qui venaient d’arriver en ce lieu »

¹³ Dans le texte latin, nous lisons « Et caligantem nigra formidine lucum ingressus », le bois (*lucum*) serait donc sombre (*caligantem*) et la crainte (*formidine*) noire (*nigra*).

erat illa recentes inter » ¹⁴	5/21	« les ombres nouvelles », 4 occ. « celes qui nouvelement sont venues à dampement » « mainte ame nouvelle » « son ombre encor recente errait parmi les ombres » 16/21
« Umbra subit terras, et quae loca uiderat ante. »	« L'ombre d'Orphée descend sous terre, et reconnaît les lieux par lui déjà visités » 13/21	« va sous les terres », 2 occ. « aux enfers lors s'en va » « si-tost que la mort y fait errer son ombre » « l'ombre d'Orphée se glissa sous terre » « l'ombre d'Orphée devala dans les Enfers » (2 occ.) 7/21 Non traduit (« Orphée ombre là-bas ») 1/21

La descente évoquée est toujours celle aux Enfers, symbolisant la mort. Dans la plupart des traductions qui n'emploient pas le verbe *descendre*, celui-ci est remplacé par un autre verbe de mouvement. Deux seulement associent dans leurs sens les notions de mouvement et d'orientation de ce dernier vers le bas, *dévaler* et *retomber* qui ont deux sèmes communs avec *descendre*, contrairement à (*s'en*) *aller*, *arriver*, (*se*) *glisser*, *se précipiter*, *se rendre* et *venir* qui n'ont qu'un sème commun (le mouvement). La notion de la mort est introduite directement uniquement dans 3 traductions : deux fois pour décrire la fonction des Enfers (« où l'inflexible mort nous attire tour à tour » et « où les mortels enfin vont étouffer leurs jours ») et une fois concernant la mort d'Orphée (« si-tost que la mort y fait errer son ombre »). Mais, dans les deux derniers extraits, la notion de mouvement est tout de même présente avec les verbes *aller* et *errer* (mais c'est la mort qui se déplace, pas un humain).

Le *DES* du Crisco que nous utilisons dans notre appariement des synonymes est riche (plus de 50 000 entrées et 209 000 relations synonymiques réciproques¹⁵) et les corrélats les plus saillants sont classés par nombre décroissant de sens élémentaires communs¹⁶ avec la vedette (Figure 89).

¹⁴ « Umbras recentes » (les ombres nouvelles) ne sont qualifiées par aucun adjectif verbal dans le texte latin. Dans les textes français correspondants, nous valorisons la totalité de la construction utilisée pour traduire l'adjectif *recentes*.

¹⁵ La notion de *réciprocité* vient d'un post-traitement qui a consisté à rendre symétriques les relations synonymiques extraites des dictionnaires dépouillés : « La base de départ a été constitué à partir de sept dictionnaires classiques. Un premier travail [...] a permis d'en extraire les relations synonymiques. [On] a ensuite concaténé, homogénéisé et symétrisé les données. [...] Le DES est symétrique : cela signifie que si un mot est synonyme d'un autre, le réciproque est vrai. En d'autres termes, les liens ne sont pas orientés. Cette propriété découle de la définition de la synonymie en tant que substituabilité en contexte », voir <https://crisco.unicaen.fr/dictionnaire-electronique-des-synonymes-des/presentation-du-dictionnaire-electronique-des-synonymes/>.

¹⁶ La notion de *clique* intervient ici. « Une clique – terme emprunté à la théorie des graphes – est un ensemble maximal de mots tous synonymes entre eux. [...] Elles sont plus proches des concepts que les mots, car elles neutralisent partiellement la polysémie de ceux-ci, ne retenant idéalement qu'un sens donné, commun à tous les mots de la clique », voir <https://crisco.unicaen.fr/dictionnaire-electronique-des-synonymes-des/presentation-du-dictionnaire-electronique-des-synonymes/>.

72 synonymes

abaisser, abattre, aborder, affaler, amener, avaler, baisser, bousiller, cartonner, condescendre, conduire, couler, critiquer, débarquer, débouler, décaler, déchoir, décliner, décrocher, décroître, **dégringoler**, demeurer, déposer, déroger, dessouder, dévaler, diminuer, émaner, envahir, éreinter, exécuter, faire irruption, incliner, jeter, loger, naître, partir, passer, pencher, pendre, perquisitionner, plonger, procéder, provenir, rabaisser, résider, rétrograder, s'abaisser, s'affaïsser, s'arrêter, s'avilir, s'effondrer, s'enfoncer, s'incliner, sauter, se coucher, se dégrader, se déposer, se loger, se ravalier, **se retirer**, se ruer, séjourner, sonder, sortir, suivre, surbaisser, tirer, tomber, transporter, tuer, **venir**

22 antonymes

ascensionner, augmenter, culminer, dresser, élever, embarquer, escalader, exhausser, gravir, grimper, guinder, hausser, hisser, jucher, lever, monter, rehausser, relever, remonter, s'élever, se dresser, se hausser

Classement des premiers synonymes

baisser	■
déchoir	■
s'abaisser	■
dégringoler	■
tomber	■
décliner	■
provenir	■
pencher	■
sauter	■
venir	■

Figure 89. Extrait de l'article descendre du DES¹⁷

Nous testons notre méthode en interrogeant les 847 couples de phrases issues des traductions de Virgile et Ovide et détectées comme correspondantes par TextPAIR et Tracer, puis validées avec notre algorithme de comparaison des mots forts¹⁸. Parmi elles, 20 % des relations lexicales sont obtenues grâce aux relations synonymiques inventoriées dans le *DES*. Toutefois, leur analyse révèle plusieurs problèmes.

Tout d'abord, sur 30 couples de phrases mises en correspondance, dont l'une des deux contient le verbe *descendre*, 51 appariements sont proposés, dont seulement 12 sont corrects. Nous y recensons 7 verbes (*dévaler*, *entrer*, *franchir*, *passer*, *pénétrer*, *retomber* et *venir*), dont les constructions sont comparables ($N_{[+animé]} V N_{[+lieu]}$) et donc aucun ne figure parmi les synonymes les plus fréquents de *descendre* dans le *DES*. Ces verbes ont des traits sémantiques partiellement différents (Tableau 94), même si, dans les occurrences du corpus, ils expriment tous le déplacement d'un humain. *Dévaler* et *retomber* sont les seuls à indiquer, comme *descendre*, l'orientation du mouvement vers le bas, mais le second sans valeur volontaire de l'action¹⁹. *Franchir* et *passer* impliquent en plus un franchissement de quelque chose, et *entrer* et *pénétrer* une pénétration dans quelque chose. Selon le nombre de sèmes communs, il est plus ou moins difficile d'identifier une relation synonymique entre les verbes considérés seuls, dans la mesure où les sujets et compléments jouent un rôle important dans l'interprétation des énoncés.

¹⁷ Cet article est consultable à l'URL suivante : <https://crisco4.unicaen.fr/des/synonymes/descendre>.

¹⁸ Voir la section précédente (4a) « Prise en compte des séries traductives »).

¹⁹ Si une construction factitive (*faire V*) induit une interprétation en tant qu'action involontaire, le trait [+/- volontaire] est moins facilement décidable pour chaque occurrence verbale que le trait [+/- vers le bas] par exemple.

Tableau 94. Table sémique de descendre et des 7 verbes qui lui sont appariés automatiquement dans l'échantillon

	[vers le bas]	[volontaire]	[rapide]	[franchissant]	[pénétrant]
<i>descendre</i>	+	+	-	-	-
<i>dévaler</i>	+	+	+	-	-
<i>entrer</i>	-	+	-	-	+
<i>franchir</i>	-	+	-	+	-
<i>passer</i>	-	+	-	+	-
<i>pénétrer</i>	-	+	-	-	+
<i>retomber</i>	+	-	-	-	-
<i>venir</i>	-	+	-	-	-

Pire, pour certains couples d'énoncés, plusieurs appariements sont proposés pour un même mot fort. Dans les traductions des *Métamorphoses*, l'apostrophe adressée à Pluton et Proserpine par Orphée contient *descendre* dans la version de Villenave1806 : « Divinités du monde souterrain où *descendent* successivement tous les mortels ». Lors de notre traitement, celle-ci est appariée simultanément aux trois verbes (*retomber*, *naître* et *mourir*) de Nisard1869 : « Ô divinités de ce monde souterrain où *retombe* tout ce qui *naît* pour *mourir* ». 11 synonymes communs sont proposés dans le *DES* pour *descendre* et *mourir*. *Naître* se trouve parmi les synonymes de *descendre* et a 6 synonymes communs avec ce verbe. *Retomber*, qui a le moins de liens avec *descendre* (3), occupe pourtant la même position dans la phrase et traduit le même mot latin (*recidimus*).

Enfin, certains appariements sont pertinents pour certaines occurrences, et pas pour d'autres. Pour *passer*, par exemple, l'appariement avec *descendre* fonctionne entre « passer les noirs torrents du Styx » et « descendre vers le Styx », qui expriment l'entrée aux Enfers (avec l'idée de franchissement de quelque chose pour *passer* et de déplacement vers le bas et en direction de quelque chose pour *descendre vers*), mais pas pour « passer au milieu d'un léger peuple de fantômes », qui exprime la traversée des Enfers, avec l'idée de pénétration dans une foule pour la franchir. La différence n'est donc pas uniquement liée aux sens possibles des verbes ni à leurs constructions (transitive directe pour « passer les noirs torrents du Styx » et avec une localisation introduite par *vers* ou *à* dans les deux autres exemples), mais elle l'est plutôt par l'expression des lieux eux-mêmes. En effet, la présence du cooccurent *Styx* dans les deux premiers contextes invite à les rapprocher et à interpréter *passer* comme <franchir ce cours d'eau> et *descendre vers* comme <se diriger dans sa direction>.

Le fonctionnement de notre algorithme n'est donc pas suffisamment discriminant. En effet, au moins pour des verbes très polysémiques comme *descendre* ou *passer*, l'appariement synonymique doit prendre en compte les cooccurents et les constructions syntaxiques qui contribuent à la construction du sens de chaque occurrence verbale. Si la richesse et la facilité d'exploitation du *DES* est appréciable dans le cadre de traitements automatiques, son caractère cumulatif pose problème. Nous cherchons plutôt une ressource qui propose des sous-listes de synonymes, accompagnées de

sélecteurs sémantiques, cooccurrence ou syntaxiques qui faciliteraient l'identification des synonymes valides pour l'emploi particulier de chacun des mots-occurrences et nous tournons donc vers des dictionnaires de synonymie distinctive.

Comparaison du potentiel informationnel de trois dictionnaires de synonymie distinctive

Les principaux éditeurs français, Le Robert, Larousse et Hachette, offrent trois dictionnaires de synonymes, qui introduisent différemment les sous-ensembles de synonymes des items en adresse d'articles.

Le *Dictionnaire des synonymes et nuances* du Robert (2020), dont les synonymes sont repris dans l'onglet « Synonymes » du *Dico en ligne*²⁰ (ci-après Dsn-Dico-Robert), distingue les sous-ensembles par la position initiale et la mise en gras du synonyme le plus représentatif du sens partagé par les items de chaque sous-ensemble. Selon l'éditeur, c'est le moyen le plus efficace de « rappeler un mot ou une expression que l'on connaît, mais qui fait défaut au moment où l'on rédige »²¹. Seuls les mots de sens équivalent ou très voisin sont fournis, mais les éditeurs invitent à « circuler dans le dictionnaire et [à] élargir ainsi son choix »²². Dans le volume imprimé, 200 000 synonymes sont proposés pour 25 000 mots et expressions et 1 400 mots proches sont comparés. *Dico en ligne* indique seulement « 200 000 synonymes organisés par sens ».

Le *Grand dictionnaire Synonymes & Contraires* de Larousse (2004), consultable en ligne sur le site *Fr.TheFreeDictionary*²³ (GdSC-FreeDic-Larousse), est présenté dans le volume imprimé comme « [u]n dictionnaire pratique, dans sa consultation et dans la circulation entre les mots »²⁴. Il propose, lui aussi, 200 000 synonymes (nombre d'entrées non précisé). Ce dictionnaire conduit « le lecteur au mot qu'il recherche par des définitions concises, systématiques et précises »²⁵, permettant de savoir pour quelle acception de l'entrée un ensemble de synonymes est pertinent.

Enfin, le *Dictionnaire des synonymes* d'Hachette Éducation (1996), indisponible en ligne, mais édité en 2001 sur le CD-ROM du *Dictionnaires français Hachette* (Ds-DfH-Hachette), présente chaque sous-ensemble introduit par une contextualisation de l'item en adresse qui en discrimine l'emploi syntaxico-sémantique. Ce dictionnaire est un peu moins fourni que les deux précédents (120 000 synonymes pour 18 000 entrées, selon la 4^e de couverture du volume imprimé de 1996). Son « Avant-propos » (non signé) indique qu'il vise à fournir « un terme qui puisse permuter avec celui qu'on a en tête, dans la même tournure et, bien sûr, sans changement de sens »²⁶.

²⁰ L'article **descendre** est disponible à l'URL suivant : <https://dictionnaire.lerobert.com/definition/descendre>.

²¹ Voir « Préface », p. V [non numéroté], § « Un nouveau concept de dictionnaire de synonymes », paragraphe 2).

²² *Idem*, p. VI, § « Les synonymes », paragraphe 3).

²³ L'article **descendre** est disponible à l'URL suivant : <https://fr.thefreedictionary.com/descendre>.

²⁴ Voir « Avant-propos », p. II [non numéroté], § « Quels ont été nos choix ? », paragraphe 2.

²⁵ *Idem*, paragraphe 5.

²⁶ Voir « Avant-propos », p. VII, paragraphe 7.

Les traitements du verbe *descendre* que ces trois dictionnaires proposent diffèrent fortement (Figure 90-Figure 92).

SYNONYMES

Synonymes de **descendre**

verbe intransitif

- **baisser**, décroître, diminuer, faiblir
- [astre] **se coucher**, [nuit] **tomber**, [mer] **se retirer**
- **rétrograder**, baisser, être relégué
- **loger**, résider, séjourner

Synonymes de **descendre de**

- **venir de**, émaner de, être issu de, provenir de

Synonymes de **descendre jusqu'à**

- **condescendre à**, consentir à
- **s'abaisser à**, s'avilir au point de, se ravalier au point de

verbe transitif

- **dégringoler**, dévaler
- **suivre**, emprunter, longer
- **décharger**, débarquer
- **décrocher**, déposer
- **abaisser**, baisser, [une voile] **affaler**, amener
- **critiquer**, éreinter, faire le procès de, **arranger** (familier), **assassiner** (familier), **casser du sucre sur le dos de** (familier), **débiner** (familier), **dézinguer** (familier), **esquinter** (familier), **étriller** (familier), **habiller pour l'hiver** (familier), **sabrer** (familier), **tailler un costard à** (familier), **taper sur** (familier)
- [familier] **tuer**, **buter** (familier), **flinguer** (familier), **dessouder** (argot)
- [familier] **boire**, **absorber**, **avalier**, **ingurgiter**, **téter**, **carburer à** (familier), **écluser** (familier), **s'enfiler** (familier), **s'envoyer** (familier), **se jeter derrière la cravate** (familier), **pomper** (familier), **siffler** (familier), **se taper** (familier)

Figure 90. Dsn-Dico-Robert

Synonymes et Contraires

descendre

verbe intransitif **descendre**

1. Aller de haut en bas.
≡ **décliner**, **rétrograder** -FAMILIER: **dégringoler**.
≠ **grimper**, **monter**, **se hausser**, **s'élever**.
2. Loger quelque part.
≡ **résider**, **s'arrêter**, **séjourner**.
3. Perdre de la valeur.
≡ **baisser**, **chuter**, **décliner**, **diminuer**, **s'effondrer** -FAMILIER: **dégringoler**.
4. S'abaisser à faire.
≡ **déchoir** -FAMILIER: **s'enfoncer**.
≠ **s'élever**.
5. Être issu de.
≡ **naître**, **provenir**.

descendre

verbe transitif **descendre**

1. Déposer à un endroit.
≡ **amener**, **conduire**.
≠ **embarquer**.
2. Parcourir de haut en bas.
≡ **dévaler** -FAMILIER: **débouler**, **dégringoler**.
≠ **gravir**.
3. Déplacer vers le bas.
≡ **abaisser**, **baisser**, **incliner**, **pencher** -LITTÉRAIRE: **rabaisser**.
≠ **élever**, **hausser**, **hisser**, **lever**, **monter**, **rehausser**, **relever**.
4. POPULAIRE. Tuer quelqu'un.
≡ **abattre**, **achever**, **assassiner**, **égorger**, **éliminer**, **envoyer ad patres**, **exécuter**, **fusiller**, **massacrer**, **tuer** -FAMILIER: **faire passer le goût du pain**, **le goût du goût du pain**, **trucider** -LITTÉRAIRE: **occire** -POPULAIRE: **buter**, **faire la peau**, **flinguer**, **ratatiner**, **rectifier**, **refroidir**, **zigouiller** -VIEUX: **suriner**.

"CITE" Le Grand Dictionnaire des Synonymes et Contraires © Larousse 2004

Figure 91. GdSC-FreeDic-Larousse

descendre v.

1. (Descendre une voile ou un pavillon) ▷ AFFALER, AMENER, BAISSER.
2. (Le niveau de l'eau descend) ▷ BAISSER, DÉCROÎTRE, DIMINUER.
3. (Descendre dans le meilleur hôtel de la ville) ▷ ALLER, S'ARRÊTER.
4. Fig. (Descendre d'une famille de magistrats) ▷ ÊTRE ISSU DE, VENIR DE.
5. Fig. (Descendre jusqu'aux attaques personnelles les plus basses) ▷ S'ABAISSER, S'AVILIR.
6. Fam. (Descendre un verre) ▷ AVALER, BOIRE.
7. Fam. (Descendre qqn) ▷ ABATTRE, ASSASSINER, LIQUIDER (fam.), OCCIRE (litt.), RECTIFIER (fam.), REFROIDIR (fam.), SUPPRIMER, TRUCIDER (fam.), TUER, ZIGOUILLER (fam.).
8. Fam. (La critique a descendu son dernier film) ▷ DÉCHIRER, DÉMOLIR (fam.), ÉREINTER, ESQUINTER (fam.).

Figure 92. Ds-DfH-Hachette

Pour comparer les capacités des trois dictionnaires à révéler les relations synonymiques valides pour les mots-occurrences de notre corpus, observons concrètement le couple de traductions de Virgile déjà présenté (Figure 88) et voyons si les synonymes saillants, les définitions ou les contextualisations sont exploitables (et, de surcroît, peuvent l'être informatiquement) pour valider le repérage d'un synonyme dans l'ensemble qu'ils introduisent. Pour ce faire, nous considérons plusieurs segments, en consultant les articles des items attestés dans la traduction Héguin¹⁸²⁷ (donnée en premier ci-après) afin d'y chercher les items synonymes attestés dans la réécriture de Nisard¹⁸⁶⁸ (donnée en second).

Pour les segments « ses forêts noires d'épouvante » *vs* « ses bois que remplissent l'horreur et les ténèbres », l'entrée *forêt* ne figure pas dans le GdSC-FreeDic-Larousse, mais dans le Dsn-Dico-Robert *bois* est son synonyme principal et, dans Ds-DfH-Hachette, c'est l'unique équivalent fourni :

A. *s.n.* **forêt**

- a. Dsn-Dico-Robert : 1^{er} • « **bois**, futaie, plantation »
- b. GdSC-FreeDic-Larousse : Ø
- c. Ds-DfH-Hachette : « 1. (Se promener dans une forêt) ▷ BOIS. »

Les trois dictionnaires permettent d'apparier sans obstacle *épouvante* avec *horreur* puisque *s.n.* **épouvante** le nom *horreur* est effectivement attesté dans une liste qui s'avère introduite par un sélecteur de sens que, en tant que lecteur, nous pouvons considérer comme valide²⁷.

²⁷ Un test de validité informatisé, quant à lui, constitue une question complexe que nous n'avons pas réussi à aborder dans le cadre de notre recherche – elle reste une perspective stimulante pour les travaux futurs.

B. *s.v.* **épouvante**²⁸

- a. Dsn-Dico-Robert : « terreur, affolement, effroi, frayeur, horreur, panique, épouvantement (*viens*) »
- b. GdSC-FreeDic-Larousse : « Sentiment de peur intense $\equiv$ affolement, angoisse, crainte, effroi, frayeur, horreur, panique, terreur. »
- c. Ds-DfH-Hachette : (Des attentats qui sèment l'épouvante) $\triangleright$ AFFOLEMENT, EFFROI, HORREUR, PANIQUE, PEUR, TERREUR. »

Pour les segments « les Mânes et leur formidable roi » *vs* « les mânes et le terrible roi des enfers », l'appariement de *formidable* et de *terrible* est également possible en exploitant les trois dictionnaires, mais le contrôle de pertinence échoue puisque le sens traité n'est pas le même dans le Dsn-Dico-Robert (où il signifie « effrayant », comme dans le corpus), que dans le GdSC-FreeDic-Larousse (où il est défini par « qui sort de l'ordinaire ») et le Ds-DfH-Hachette qui contextualise le même emploi (« Une soirée formidable »). Par ailleurs, les dictionnaires indiquent que *terrible*, avec ce sens, relève d'un emploi familier, alors que nous n'observons pas de changement de registre au sein des phrases traitées.

C. *s.v.* **formidable**

- a. Dsn-Dico-Robert : 4^e • « **effrayant**, épouvantable, redoutable, terrible (*familier*) »
- b. GdSC-FreeDic-Larousse : « 1. Qui sort de l'ordinaire $\equiv$ effrayant, étonnant, étourdissant, extraordinaire, fantastique, gigantesque, imposant, inouï, magnifique, prodigieux, stupéfiant -FAMILIER : énorme, époustouflant, faramineux, renversant, terrible. »
- c. Ds-DfH-Hachette : « 2. Fam. (Une soirée formidable) $\triangleright$ DÉMENT (fam.), DU TONNERRE (fam.), ÉPATANT (fam.), EXTRAORDINAIRE, FABULEUX (fam.), FANTASTIQUE, GÉANT (fam.), GÉNIAL (fam.), MERVEILLEUX, PRODIGIEUX, SENSATIONNEL, SUPER (fam.), TERRIBLE (fam.). »

Pour les segments « ces cœurs impitoyables que des prières humaines n'ont jamais fléchis » *vs* « ces cœurs que les prières des humains n'ont jamais attendris », et spécifiquement le couple *fléchir* et *attendrir*, le Dsn-Dico-Robert et le GdSC-FreeDic-Larousse proposent de différencier les emplois transitif et intransitif de *fléchir*, ce qui est pertinent pour ce verbe afin de limiter le nombre des synonymes à prendre en compte. Le Ds-DfH-Hachette ne caractérise pas la construction de la contextualisation, mais elle est transitive.

D. *s.v.* **fléchir**

- a. Dsn-Dico-Robert : *verbe transitif*, 1^{er} • « **émouvoir**, apitoyer, adoucir, attendrir, désarmer, ébranler, faire céder, gagner, toucher, se concilier, gagner à sa cause »
- b. GdSC-FreeDic-Larousse : *verbe transitif* « 2. Faire céder quelqu'un $\equiv$ adoucir, apitoyer, attendrir, désarmer, ébranler, toucher. » »
- c. Ds-DfH-Hachette : « 3. Fig (Fléchir qqn à force de prières) $\triangleright$ APITOYER, ATTENDRIER, DÉSARMER, ÉBRANLER, ÉMOUVOIR, FAIRE CÉDER, TOUCHER. »

La contextualisation de Ds-DfH-Hachette contient en outre le cooccurrent *prière*, attesté dans les deux phrases du corpus, ce qui nous permet de confirmer la mise en relation des synonymes. Le Dsn-Dico-Robert ne propose pas de cooccurrents des synonymes listés, sauf au sein de l'expression « gagner à sa cause » qui est lexicalisée, et le GdSC-FreeDic-Larousse propose une définition (« Faire céder

²⁸ Les articles **épouvante** comportent chacun une seule liste, donc (i) dans le Dsn-Dico-Robert, le premier synonyme n'étant pas distinctif, il n'est pas mis en gras, et (ii) dans le Ds-DfH-Hachette, la liste est non numérotée.

quelqu'un ») qui n'est pas informatiquement exploitable dans ce contexte puisqu'elle ne contient pas d'indice tel que la présence du cooccurrent *prière*.

Alors que la transitivité peut être un des indices formels facilitant le choix des synonymes, comme pour *fléchir* dans le contexte précédent, les 67 % d'occurrences de *descendre* (160/239) qui expriment, pour un sujet animé, le fait de <se déplacer vers le bas> peuvent se combiner avec des compléments de lieu ou des COD exprimant un lieu : « Il osa même descendre dans les gouffres du Ténare », « Orphée descend vers le Styx », « Il descend de la montagne », etc. Les emplois sont donc intransitifs ou transitifs, mais souvent avec une localisation mentionnée : $N_{[+animé]} V_{[+mouvement + vers \text{ le bas}]} Prép? N_{[+lieu]}$.

Les entrées de *descendre* dans les trois dictionnaires (Figure 90-Figure 92) permettent de voir qu'ils ne prennent pas en compte la rection de manière identique : elle est explicite dans les dictionnaires du Robert et de Larousse, mais celui de Larousse s'en tient à l'opposition intransitif/ transitif alors que celui du Robert isole en plus, à la fin du traitement des emplois intransitifs, ceux des constructions *descendre de* et *descendre jusqu'à*.

Bizarrement, le Ds-DfH-Hachette ne traite pas le verbe *descendre* en tant que verbe de mouvement vers le bas réalisé par son sujet. Dans le Dsn-Dico-Robert et GdSC-FreeDic-Larousse, nous trouvons comme synonyme du verbe transitif *descendre*, le verbe *dévaler*, que nous avons vu être le plus souvent apparié à *descendre* lors de notre expérimentation d'appariement synonymique avec le DES :

E. *s.v. descendre*

- a. Dsn-Dico-Robert : *verbe transitif*, 1^{er} • « **dégringoler**, dévaler »
- b. GdSC-FreeDic-Larousse : *verbe transitif* « 2. Parcourir du haut en bas ≡ dévaler -FAMILIER : débouler, dégringoler » ; *verbe intransitif* « 1. Aller de haut en bas ≡ décliner, rétrograder -FAMILIER : dégringoler. »
- c. Ds-DfH-Hachette : Ø

Dans le corpus, *dévaler* apparaît 14 fois, majoritairement dans des emplois associés à des mentions de lieu et avec un sujet animé : « il me faut devaller par sanglantes alarmes », « il s'élance à travers les fourrés et les roches pour dévaler l'autre versant », « l'ombre d'Orphée devalla dans les Enfers », « sans long intervalle au pays Thracien il devalle », « cet Amant fidele osa [...] devaler en des lieux ou n'arriva jamais la lumière des Cieux », « en enfer est sans demorance l'ame dou devin devalee », « Suis-je si malheureux de t'avoir espousée pour devaller si tost dans la plaine Élysée ? », « je veux pour te revoir devaller aux Enfers », « vouloir devaller dans la pasle demeure », « pour finir mes travaux je veux y devaller », « chez Pluton devaller pour vous mettre en repos », « Je ne suis devalé dedans ce creux Tartare ».

Un autre synonyme, *dégringoler*, également listé dans le DES et fourni par le Dsn-Dico-Robert, *s.v. descendre verbe transitif* 1., figure dans le GdSC-FreeDic-Larousse pour les deux rections. Dans le corpus, il correspond à la même idée de mouvement vers le bas et un N[+lieu] de destination est

toujours mentionné, que le sujet soit animé ou non : « Notre amant se met en chemin / Et dans l'Empire souterrain / Il dégringole en moins de rien » (Grandval1729), « [Q]uelque chose gronda. Cela évoqua pour lui des rochers d'abord vacillant pour dégringolant au fond d'une gorge » (Jimenes2004).

Les autres synonymes du sous-ensemble intransitif du GdSC-FreeDic-Larousse, *décliner* et *rétrograder*, n'expriment pas le déplacement physique d'une entité animée. Le corpus le confirme :

- 5 occurrences sur 6 de *décliner* sont intransitives et employées avec un sujet non animé *soleil* ou *jour* : « Le jour décline », « Le soleil, moins ardent et plus large, décline », « Le soleil décline », « le soleil pâli décline et se consume », « le soleil déclinait, allongeant les ombres ». L'occurrence restante est transitive, elle exprime un refus formulé par un sujet animé : « Orphée déclina son offre ».
- Parmi les 4 occurrences de *rétrograder*, aucune ne mentionne de lieu physique, puisqu'elles expriment le retour à un état antérieur : « sa pensée [...] rétrogradait vers les temps où les rois étaient des dieux », « Parviendrons-nous à [...] rétrograder lorsque l'épreuve aura été au-dessus des forces de l'adepte ? », « Le Grec qui n'avait pas cru aux mystères d'Éleusis était condamné à rétrograder », « Ceux qui sont nés dans le christianisme [...] ne seront-ils pas condamnés à rétrograder vers l'état antérieur ? »

La définition « Aller de haut en bas » du GdSC-FreeDic-Larousse laisse donc ouvert le choix du trait plus ou moins animé de l'entité qui se déplace et du lieu concret ou abstrait de déplacement.

Dans le Dsn-Dico-Robert, *décliner* n'apparaît dans aucune liste et *rétrograder* introduit un sous-ensemble de synonymes du verbe intransitif (3^e • « **rétrograder**, baisser, être relégué ») qui n'expriment pas de déplacement physique vers le bas. Le sous-ensemble le plus proche du sens de *décliner* (2^e • « [astre] **se coucher**, [nuit] tomber, [mer] se retirer »), *s.v.* **descendre** *verbe intransitif* toujours, propose des synonymes dont les sujets sont sélectionnés par des cooccurents typiques *astre*, *nuit* et *mer* afin de réunir des verbes de sens légèrement différent. Un de ces cooccurents, *nuit*, est effectivement attesté dans notre corpus comme sujet de *tomber* (Tableau 95). Même si, sur les 35 occurrences de la construction *nuit*_[sujet] V, 13 seulement comportent des verbes exprimant un déplacement vers le bas (10 occ. de *tomber* et 3 de *descendre*), toutes expriment la transition du jour à la nuit. La présence du sujet *nuit* conditionne donc très fortement, dans notre corpus, les interprétations verbales, sans qu'il soit utile de prendre en compte des indications de lieu éventuellement présentes.

Tableau 95. Résultats d'interrogation du corpus pour nuit_[sujet] V

Construction	Verbes et expressions	Exemple d'occurrence
V _[+ mouvement +vers le bas]	<i>tomber</i> (10 occ.)	« la nuit tombe »
	<i>descendre</i> (3 occ.)	« la nuit descendit sur la terre »
V _[+/- mouvement -vers le bas]	<i>venir</i> (10 occ.)	« [la] nuit bénie qui vient verser sur nous sa douceur infinie »
	<i>s'étendre</i> (2 occ.)	« la nuit anticipée s'étendit sur les eaux »
	<i>s'avancer</i> (2 occ.)	« la nuit transparente s'avancait en silence »
	<i>se faire</i> (2 occ.)	« la nuit se fait »
	<i>approcher</i> (1 occ.)	« la nuit approche »
	<i>arriver</i> (1 occ.)	« la nuit arrive subitement »
	<i>régner</i> (1 occ.)	« la nuit règne »
Expressions verbales _[-mouvement - vers le bas]	<i>épaissir l'air</i> (1 occ.)	« Quand la nuit comme une grande fumée épaissit l'air »
	<i>envelopper quelqu'un</i> (1 occ.)	« L'éternelle nuit m'enveloppe »
	<i>couvrir de son manteau</i> (1 occ.)	« La nuit couvre de son manteau la surface de la terre »

Alors que nous devons rapidement renoncer à exploiter le Ds-DfH-Hachette, qui sélectionne les emplois de manière trop particularisante avec des contextualisations et dont le traitement de la polysémie semble être moins exhaustif, les contenus du Dsn-Dico-Robert et du GdSC-FreeDic-Larousse sont assez complémentaires. Toutefois, en ce qui concerne la structure des articles, le Dsn-Dico-Robert est susceptible d'être plus facilement exploitable informatiquement. Il propose :

- des cooccurents entre crochets qui sélectionnent la valeur sémantique de l'item traité : *s.v. descendre, verbe intransitif*, 2^e • « [nuit] tomber » ;
- les propositions introductrices des compléments : *s.v. descendre, verbe intransitif*, « **descendre de** venir de, émaner de, être issu de, provenir de » ;
- des constructions lemmatisées : *s.v. passer, verbe intransitif*, « **passer de... à • se couler**, se glisser [...] » ;
- des expressions : *s.v. descente* « **descente aux enfers** chute, dégringolade, disgrâce » ;
- des marquages (diachroniques, diatopiques, etc.) assignés spécifiquement à chaque synonyme auquel la marque est postposée ou à l'item en adresse quand la marque précède la liste : « rejeton (*vieux ou plaisant*) », *s.v. descendant*, ou, pour une valeur de *descendre* non attestée dans notre corpus « [familier] tuer, buter (*familier*), flingueur (*familier*), des-souder (*argot*) », *s.v. descendre verbe transitif*, 7^e •. La première marque (entre crochets) concerne *descendre* et les suivantes (entre parenthèses) le synonyme qui précède chaque marque.

Par ailleurs, les différentes constructions des verbes sont traitées au sein de subdivisions du même article, alors que le GdSC-FreeDic-Larousse les dégroupé.

Enfin, même si l'introduction des sous-ensembles par le synonyme le plus saillant n'apporte pas d'indices supplémentaires pour déterminer si le sens de l'occurrence en corpus est proche de celui pour lequel les synonymes sont valides, cet item est plus facilement exploitable que les définitions proposées par le GdSC-FreeDic-Larousse. En effet, celles-ci sont trop hétérogènes dans leurs formulations pour pouvoir y localiser et en extraire automatiquement des informations-clés et elles sont parfois volontairement implicites afin de permettre de regrouper des synonymes d'occurrences de l'item en adresse présentant des sens voisins, mais distinguables, comme « Aller de haut en bas » qui peut correspondre à une glose d'occurrences verbales acceptant des sujets +/- animés et des déplacements entre des lieux concrets ou des éléments abstraits (dont des états).

Vers une nouvelle méthode d'appariement appuyée sur de la synonymie distinctive ?

Le Dsn-Dico-Robert paraît être le mieux adapté à nos besoins. Mais pour exploiter pleinement son potentiel, il est nécessaire d'adapter le traitement élaboré pour le *DES*, notamment afin d'améliorer la prise en compte du contexte de chaque occurrence et l'identification du sens d'un mot-occurrence polysémique. En effet, nous vérifions sans surprise que la présence de certains cooccurents sujets ou au sein des mentions de localisations joue un rôle important dans la discrimination des emplois du verbe *descendre* du corpus.

Pour affiner la sélection des synonymes pertinents, nous aimerions donc mieux exploiter ces cooccurents. De nouveau, nous nous tournons vers des ressources dictionnaires afin d'évaluer ce qu'elles peuvent apporter à notre traitement. Les logiciels de correction orthographique, grammaticale et typographique, *Antidote* et le *Robert correcteur*, proposent des dictionnaires de cooccurrences²⁹, qui semblent répondre à notre besoin et dont nous cherchons aussi à apprécier s'ils peuvent constituer un type de ressources complémentaires des dictionnaires de synonymes. Les inventaires de cooccurents proposés dans ces deux dictionnaires sont classés par sens de l'item traité et indiquent les relations syntaxiques entre l'item et son cooccurent.

Dans la phrase de la traduction d'Ovide de Nisard 1869, « [...Orphée] osa descendre vers le Styx par la porte du Ténare », deux éléments du contexte du verbe *descendre* peuvent être directement trouvés dans le dictionnaire de cooccurrences d'*Antidote* (mais pas dans l'onglet « Combinaisons » du *Robert correcteur*) : « par la porte » et « oser ». Le troisième peut l'être indirectement, puisque si

²⁹ Ils proposent aussi des dictionnaires de synonymes. Celui d'*Antidote* est de facture classique, alors que celui fourni par les Éditions Diagonal au sein du *Robert correcteur* sélectionne les emplois au moyen de traits distinctifs : *s.n.* **descendre**, « **descendre** [Sans complmt][Aller de haut en bas] » vs « **descendre** (qqch) [... le niveau ou la hauteur de qqch », etc. Même si dans cette investigation nous ne prenons pas en compte ces dictionnaires, les traits sélecteurs d'emploi du dictionnaire de synonymes du *Robert correcteur* retiennent notre attention.

« vers le Styx » est trop spécifique pour figurer parmi les cooccurents, nous y trouvons une expression de même construction avec un hyperonyme (« vers le fleuve »³⁰).

Ces trois cooccurents ne sont listés que pour le sens « Aller vers le bas », *s.v.* **descendre**, qu'ils permettent donc de retenir automatiquement comme interprétation valide dans ce contexte. Ce constat fait, il reste toujours le problème de l'articulation avec le dictionnaire des synonymes. Or, alors qu'un lien peut être fait avec la définition « Aller de haut en bas » de GdSC-FreeDic-Larousse (2 mots commun, *aller* et *bas*), ce n'est pas le cas pour le Dsn-Dico-Robert dans lequel les sous-ensembles ne sont pas introduits par des définitions.

Plus encore, nous constatons que l'articulation n'est pas non plus possible avec le dictionnaire des synonymes d'*Antidote* lui-même, qui regroupe des structures assez hétérogènes (Figure 93), en fonction des articles : *s.v.* **descendre**, les sous-ensembles sont introduits, comme dans le Dsn-Dico-Robert, par le synonyme le plus représentatif, mais *s.v.* **dégringoler**, par exemple, nous trouvons des définitions, comme dans le GdSC-FreeDic-Larousse.

4 Synonymes de **descendre**, verbe

- ▲ Diminuer — décroître, baisser, décliner, diminuer, s'amoindrir.
 - Redescendre à marée basse — refluer (mer), baisser, rebaisser, se retirer.
- ▲ Demeurer — séjourner, rester, loger, s'arrêter, se relaisser.
- ▲ Déplacer — baisser, abaisser.
- ▲ (EMPLOI FAMILIER) Ingérer — lamper, avaler d'un coup, avaler d'un trait, boire d'un coup, boire d'un trait.
 - FAMILIER – siffler.
 - BELGIQUE, FAMILIER – squetter, afonner.
 - QUÉBEC, FAMILIER – caler.
 - VIEUX – fouetter (du vin), sabler.

4 Synonymes de **dégringoler**, verbe

- ▲ Descendre rapidement — débouler, dévaler.
- ▲ Baisser, en parlant d'un prix — chuter, s'effondrer, baisser, diminuer, tomber.

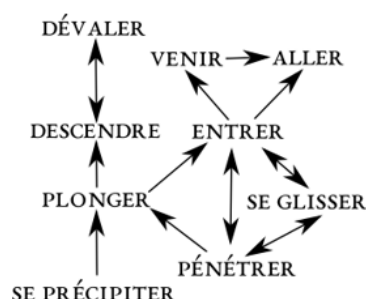
Figure 93. Articles **descendre** et **dégringoler** de l'onglet « Synonymes » d'*Antidote*

³⁰ Dans le *Robert correcteur*, nous trouvons seulement « descendre le fleuve » *s.v.* **descendre** « Sens 1 : aller, parcourir de haut en bas », § « Avec complément direct ».

Nous butons sur les procédures qui permettraient d'exploiter les combinaisons de mots des dictionnaires spécialisés (d'*Antidote*, du *Robert correcteur* ou du Dsn-Dico-Robert) confrontés aux cooccurrents de nos mots occurrences afin d'en déterminer le sens, mais nous pouvons au moins nous inspirer de leur principe de regroupement des cooccurrents fournis par valeurs sémantiques de l'item traité. Celui du *Robert correcteur* est particulièrement intéressant, car les cooccurrents de chaque emploi d'un item y sont présentés au moyen de graphes³¹. Cette approche est transposable à notre recherche, comme nous le constatons en explorant un sous-corpus constitué des seules traductions de Virgile et d'Ovide en repérant les contextes de chacune dans lesquels une occurrence de *descendre* est associée à un sujet N[+animé] et au nom de lieu *Enfers* ainsi que les contextes correspondants dans les autres traductions, dont les auteurs ont préféré un autre verbe ou une expression verbale à *descendre*. Nous constatons qu'en fonction du sujet, les équivalents de *descendre* dans les traductions qui choisissent un autre verbe ne sont pas les mêmes. Pour le sujet *Orphée*, nous trouvons des verbes comme *se précipiter*, *pénétrer* ou *venir* dont le sens implique un mouvement volontaire. Pour le sujet *Eurydice*, ou les noms communs *morts* et *mortels* (au pluriel pour nommer l'ensemble de ceux qui sont morts ou mortels), l'action n'est pas volontaire (*tomber*, *chuter*, *être précipité*, etc.).

Les premiers résultats d'exploitation des cooccurrents pour sélectionner les ensembles de synonymes sont encourageants. L'ensemble des verbes du corpus qui prennent *Orphée* comme sujet et mentionnent *Enfers* dans un complément nous permet déjà, avec une exploitation hypertextuelle du Dsn-Dico-Robert, d'aboutir à un graphe des relations synonymiques des verbes du corpus pour cette sélection de contextes (Figure 94). En son sein, tous les verbes attestés dans le corpus sont liés en fonction de leur mention dans les articles du dictionnaire (comme item en adresse ou comme synonyme d'une des listes), tantôt directement et réciproquement, comme *descendre* et *dévaler*, présents chacun dans l'article de l'autre (b. et c.), tantôt indirectement, comme *descendre* et *glisser* (e.) pour lequel plusieurs cheminements dans le graphe sont possibles, mais l'un d'eux consiste à observer que *descendre dans* est un synonyme de *plonger* (g.), qui est lui-même synonyme de *pénétrer* (f.), lui-même synonyme de *se glisser* (e.).

³¹ Voir également Laurette Chardon, « L'espace sémantique du Dictionnaire électronique des synonymes (DES) et les méthodes de regroupement de sens : l'exemple de sec », *Syntaxe & Sémantique*, vol. 21, n° 1, Presses universitaires de Caen, 2020, p. 87-126, en ligne, <<https://shs.cairn.info/revue-syntaxe-et-semantique-2020-1-page-87>>. Chardon expose les différentes méthodes de calcul et de visualisation de l'espace sémantique des mots à l'aide de graphes et leur application pour la détermination des cliques et le regroupement des sens lexicographiques. Cette approche, implémentée dans le DES, est accessible sous l'onglet « espace sémantique » de chaque article (pour *descendre*, voir <https://crisco4.unicaen.fr/espsem/descendre>).



- a. s.v. **aller** Ø
- b. s.v. **descendre**, v. tr. 1^{er} • « **dégringoler**, **dévaler** »
- c. s.v. **dévaler**, v. tr. « **descendre**, **dégringoler**, **débouler** (familier) »
- d. s.v. **entrer**, v. tr. 1^{er} • « **aller**, **s'introduire**, **se couler**, **s'engager**, **s'engouffrer**, **s'enfoncer**, **faire irruption**, **se faufiler**, **se glisser**, **s'infiltrer**, **s'insinuer**, **pénétrer**, **rentrer**, **se plonger**, **venir** »
- e. s.v. **glisser** • **se glisser**, v. pron. « **se couler**, **entrer**, **se faufiler**, **s'infiltrer**, **s'insinuer**, **s'introduire**, **pénétrer** »
- f. s.v. **pénétrer**, v. intr. 1^{er} • « **s'introduire**, **s'avancer**, **s'aventurer**, **se couler**, **s'enfoncer**, **s'engager**, **entrer**, **fendre**, **se glisser**, **s'infiltrer**, **s'insinuer**, **plonger**, **se fourrer** (familier) »
- g. s.v. **plonger**, v. tr., (se) **plonger dans**, 1^{er} • « **entrer dans**, **descendre dans**, **s'enfoncer dans** »
- h. s.v. **précipiter** • **se précipiter**, v. pron. 1^{er} • « **sauter**, **se jeter**, **plonger**, **tomber** »
- i. s.v. **venir**, v. intr. l.1^{er} • « **aller**, **se déplacer**, **passer**, **se rendre**, **radiner** (familier), **rappliquer** (familier), **s'amener** (familier), **se pointer** (familier), **viendre** (barbarisme, plaisant) »

Figure 94. Visualisation en graphe et en listes de synonymes pertinents extraits de l'article *descendre* et de ceux de ses correspondants dans le corpus (items représentés dans le graphe et soulignés dans les listes extraites des articles du Dsn-Dico-Robert)

Les relations au sein de ce graphe sont assez complexes : tous les composants sont liés au moins à deux autres nœuds, sauf *dévaler* et *se précipiter*, qui étaient synonymes dans les GdSC-FreeDic-Larousse, mais qui ne sont pas inventoriés dans leurs listes respectives (c. et h.) au sein du Dsn-Dico-Robert, et qui ne sont pas dans plus d'une liste de synonymes ou en adresse des articles mobilisés du Dsn-Dico-Robert (s.v. **descendre** et dans son article propre pour *dévaler* (b. et c.), et dans son article pour *se précipiter* (h.)). Notons aussi qu'*entrer* occupe une position importante, car il est lié à 5 nœuds différents : son article (d.) liste *aller* et *venir*, et il figure dans les listes des articles **pénétrer** (f.), **plonger** (g.) et **(se) glisser** (e.).

Cette méthode permet d'exploiter pleinement le caractère distinctif du Dsn-Dico-Robert et d'automatiser l'appariement d'une sélection pertinente de synonymes pour chaque mot-clé attesté dans le sous-corpus des traductions. Plus encore, la proximité des appariements ainsi établis peut être graduée, grâce, entre autres, à des calculs issus de la théorie des réseaux, comme la proximité³² ou la centralité³³. Enfin, la liste des équivalents attestés dans le sous-corpus peut être étoffée par les autres synonymes des sous-ensembles sélectionnés, ce qui nous arme mieux pour annoter avec

³² La proximité se calcule à partir du nombre d'arcs qu'il faut parcourir pour arriver d'un nœud à l'autre par le chemin le plus court : dans notre graphe (Figure 94), *se précipiter* est plus proche de *descendre* (2 arcs à parcourir) que de *venir* (3 arcs).

³³ *Entrer* est considéré comme étant au centre du graphe non seulement parce qu'il est lié à plusieurs nœuds, mais aussi parce qu'il est nécessaire de traverser son nœud pour arriver de *se précipiter* et de *plonger* vers *venir*, *aller*, *se glisser* et *pénétrer*, ou encore pour arriver à *venir* et *aller* depuis *se glisser* et *pénétrer*.

leurs synonymes les mots-occurrences de l'ensemble du corpus qui peuvent en recevoir, en espérant que ceci permette, en fin de traitement, de trouver des énoncés non encore appariés que les synonymes de leurs annotations permettraient de lier.

Après avoir constaté quelques difficultés d'exploitation d'une ressource de synonymie cumulative dans le cadre de notre projet, nous avons comparé trois dictionnaires de synonymie distinctive, le Dsn-Dico-Robert, le GdSC-FreeDic-Larousse et le Ds-DfH-Hachette, pour juger de leur capacité à révéler les relations synonymiques présentes dans notre corpus. C'est finalement le Dsn-Dico-Robert qui apparaît comme le mieux exploitable pour nos besoins, à condition de concevoir une méthode d'appariement (semi-)automatique capable de prendre en compte les cooccurents les mieux repérés de chaque mot à appairier et d'exploiter les équivalents déjà attestés dans le corpus.

Alors que l'expérimentation que nous avons menée sur le verbe *descendre*, polysémique et très fréquent dans notre corpus, montre que la méthode que nous proposons est probante, son déploiement à la totalité du corpus reste difficile. Les ressources dont nous avons besoin, comme les dictionnaires de synonymie distinctive et de cooccurrences, appartiennent à des acteurs commerciaux et ne sont donc pas exploitables pour des traitements automatiques (sauf accord à obtenir auprès des éditeurs, alors qu'ils n'ont *a priori* pas d'enjeux à participer à des projets de recherche non financés). Les ressources linguistiques disponibles librement en format numérique sont des ressources de synonymie cumulative. Les représentations distributionnelles, comme les plongements lexicaux, même si elles gagnent en popularité, s'appliquent, pour le moment, difficilement aux petits corpus comme le nôtre. Par exemple, parmi les verbes proches pour *descendre* proposés par Word2Vec seulement *venir* est attesté dans le corpus et la majorité des verbes ne sont pas synonymes des occurrences dans notre corpus : *traverser*, *présenter*, et même *divertir*.

Selon nous, les ressources lexicographiques existantes ont un réel potentiel d'exploitation dans les projets d'humanités numériques, comme le nôtre, comme pour certaines tâches de traitement automatique des langues naturelles. En exploitant les traitements mis en œuvre dans une sélection choisie de dictionnaires imprimés et numériques de facture traditionnelle – dictionnaires de synonymie distinctive et de cooccurrences –, nous constatons l'amélioration significative de la pertinence de l'appariement (semi-)automatique des synonymes présents dans notre corpus. Même si elle posait la question des droits de consultation par un automate et celle de la couverture des nomenclatures et des inventaires synonymiques offerts à la consultation, la possibilité d'exploiter effectivement ces ressources à grande échelle ouvrirait sans doute de nouvelles perspectives probantes. Elle serait en outre susceptible de profiter en retour aux lexicographes, dans le cadre d'accords de recherche et développement, puisque, par exemple, ils pourraient ainsi proposer, en marge

des articles de leurs dictionnaires de synonymes, des attestations, au sein de corpus d'œuvres littéraires, d'emplois de certaines paires d'équivalents (un peu comme dans notre Tableau 93), chaque paire de synonymes en contextes comparables ayant été validée dans le cadre d'une annotation de corpus intégrant des synonymes par des analystes littéraires engagées dans des projets éditoriaux numériques.

Conclusion du chapitre

Les manipulations exposées dans ce chapitre sont motivées par des objectifs littéraires, mais sollicitent des compétences linguistiques et informatiques, dont certaines sont à acquérir ou à trouver chez des partenaires. Les difficultés que nous rencontrons à les intégrer dans notre développement montrent la complexité des questions auxquelles peuvent faire face les chercheurs des humanités numériques qui exploitent des outils informatiques.

Nous sommes convaincue 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.

Conclusion de la partie

Dans cette dernière partie, nous avons présenté les développements conçus dans le cadre de notre projet : le post-traitement des résultats initiaux obtenus avec TextPAIR et Tracer, l'annotation automatique des entités nommées et leurs périphrases et le prototype d'une édition comparative hyperannotée et hypertextuelle des réécritures du mythe d'Orphée et Eurydice. Nous avons également ouvert deux pistes d'amélioration de ces développements pour le futur du projet que nous envisageons avec optimisme.

Au fil des chapitres, nous avons essayé de démontrer que les humanités numériques ne sont pas antagonistes aux pratiques traditionnelles de la recherche littéraire et linguistique, mais, au contraire, permettent de repenser certaines méthodes et proposer des façons innovantes d'accès aux textes. Notre projet illustre bien, à notre sens, les bénéfices d'une approche interdisciplinaire, sans pour autant ignorer les exigences qu'elle implique.

En effet, si nous avons montré que 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 prouvé 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.

Notre traitement spécifique des résultats de TextPAIR et Tracer combine l'analyse empirique avec des manipulations automatiques et semi-automatiques pour se donner les moyens de procéder à des explorations fructueuses en composant avec la taille du corpus et la qualité de ce que nous obtenons initialement. Nous arrivons à exploiter l'hétérogénéité des résultats au moyen d'annotations qui valorisent les points communs et des lexiques qui compilent ce qui a été observé. La validation et l'analyse humaine du corpus améliorent la qualité des données obtenues, même si cela doit parfois se faire au détriment de la qualité.

Si certaines limites persistent – qu'elles soient techniques, méthodologiques ou éditoriales –, elles nous incitent à regarder au-delà de notre projet de thèse et d'envisager sa poursuite : l'amélioration de la chaîne du traitement grâce aux échanges avec d'autres chercheurs et développeurs constitue une perspective enthousiasmante au terme de notre projet. C'est dans cette optique ouverte, curieuse et collaborative que nous achevons notre travail, convaincue que les Humanités numériques ne trouvent leur pleine pertinence que lorsqu'elles favorisent la mise en commun des savoirs, des outils et des expériences.

Conclusion

L'ambition de notre projet consiste en l'exploration des réécritures françaises du mythe d'Orphée et Eurydice à travers une approche interdisciplinaire, au croisement de la critique littéraire, la linguistique et l'informatique, et dans l'esprit des humanités numériques. Nous en tirons une conviction principale : la complexité des phénomènes intertextuels, et en particulier dans le cadre particulier d'une tradition mythologique aussi vivante que mouvante, est extrêmement difficile à saisir, mais gagne à être considérée en alliant la rigueur méthodologique, la sensibilité herméneutique et les outils computationnels.

Nous avons commencé notre recherche par la constitution d'un corpus représentatif qui prend en compte la tension entre l'objectivité documentaire et la subjectivité interprétative. Un protocole rigoureux, mais ouvert aux particularités des corpus littéraires et laissant place à des décisions motivées, a ainsi permis le recueillement de 53 œuvres publiées entre le XVI^e et le XXI^e siècle. Ce corpus regroupe tant les grands classiques littéraires et artistiques que des textes en marge du canon, voire complètement inconnus. L'édition en XML-TEI, l'archivage pérenne et l'ouverture des données sur la plateforme Nakala ont garanti à ce que l'ensemble de nos travaux soit interopérable et réutilisable, conformément aux usages des autres projets (qu'ils soient issus des humanités numériques ou non) soucieux de respecter la politique nationale de la science ouverte. Cette première étape a constitué à la fois une fondation solide pour notre étude et a produit une ressource partagée pour les chercheurs qui souhaiteraient explorer ces textes à leur tour.

La deuxième étape de notre travail consistait à cerner les contours théoriques de l'intertextualité, à identifier les procédés linguistiques susceptibles d'en révéler les traces et à expérimenter quelques méthodes d'exploration computationnelle du corpus. En constatant que les relations observées résistent au cloisonnement dans des conceptualisations trop rigides, nous avons prôné une approche rhizomatique de l'intertextualité, que nous considérons la plus à même de reconnaître la pluralité, la transversalité et parfois l'opacité des liens entre les textes. Ce point de vue plus ouvert a guidé notre réflexion sur les phénomènes de paraphrase et de reformulation tout en nous permettant de souligner les limites de leur objectivation linguistique, en particulier dans un corpus diachronique et hétérogène. Ainsi, l'expérimentation que nous avons menée sur le sous-corpus des traductions françaises d'Ovide a démontré un certain intérêt de notre projet pour les études traductologiques, la linguistique du corpus et la lexicographie.

Nous avons ensuite confronté nos observations empiriques avec ce que nous constatons en ayant recours aux approches computationnelles. Nous avons évalué de manière systématique plusieurs outils de détection automatique des réutilisations textuelles. Ce travail de comparaison, réalisé à partir d'un corpus-test et en suivant une grille d'analyse précise, nous a permis d'identifier deux logiciels, TextPAIR et Tracer, qui répondent au mieux aux besoins spécifiques de notre projet.

Nous avons optimisé les paramétrages de ces outils à l'aide d'un corpus-test, puis nous avons évalué les résultats obtenus pour l'ensemble du corpus et constaté quelques difficultés. Cette démarche, inédite à ce jour, est destinée à servir de feuille de route à d'autres projets intéressés par la détection automatique des réutilisations textuelles, tant pour le choix des outils, que pour leur paramétrage initial et l'évaluation des résultats.

Enfin, la dernière étape de notre travail a été consacrée au filtrage, à l'enrichissement et à la valorisation des résultats obtenus, par la conception d'outils d'annotation et de visualisation qui facilitent une lecture critique et exploratoire des relations détectées. Le développement des méthodes de post-traitement adaptées a nécessité des ajustements techniques, une réflexion approfondie sur les paramètres les plus propices à la détection des relations pertinentes et un important travail d'annotation manuelle, ce qui constitue, nous l'espérons, un gage de précision et de qualité. Le prototype de l'édition comparative hyperannotée illustre, quant à lui, les possibilités offertes par la collaboration entre analyse littéraire et techniques informatiques.

Ainsi, notre projet a soulevé des défis particuliers,

- tant du point de vue littéraire : s'éloigner de l'analyse rapprochée de textes pour constituer un corpus bien plus large que la taille usuelle de ceux de la recherche traditionnelle,
- que linguistique : explorer les phénomènes présents au sein d'un corpus hétérogène pour interroger la pertinence des outils et lexiques existants,
- et informatique : adapter des outils généralistes aux besoins particuliers des recherches en SHS.

Dans son ensemble, il nous a amenée à formuler plusieurs constats. D'abord, il apparaît clairement que les méthodes computationnelles, aussi puissantes soient-elles, ne se substituent pas à l'intelligence interprétative, mais la complètent, prolongent, voire l'ouvrent à de nouvelles formes d'exploration. Ensuite, le choix des outils, des méthodes et des paramètres doit toujours être guidé par les spécificités du corpus et les objectifs de la recherche, qui ne peuvent être déterminés que par l'analyse empirique d'un expert humain. Enfin, l'interdisciplinarité s'est imposée dans notre travail tant comme une véritable nécessité intellectuelle que comme un défi constant, exigeant une vigilance théorique, une rigueur méthodologique et un effort de faire coopérer des vocabulaires, pratiques et attentes parfois dissonants.

Pour arriver à ces conclusions, nous avons engagé un travail important, mais chronophage et expérimental, dont l'aboutissement n'est pas atteignable dans le temps imparti du doctorat. Par

ailleurs, nous étions parfois limitée par nos propres compétences, qu'elles soient éditoriales, informatiques ou analytiques. Par conséquent, l'engagement d'une véritable équipe interdisciplinaire et le financement à long terme sont des perspectives dont nous rêvons pour la suite du projet. En outre, le développement rapide des nouvelles technologies extrêmement performantes de l'intelligence artificielle et l'apparition des grands modèles de langage est survenu au cours de notre projet. Même si, pour l'instant, nous n'avons pas trouvé de solutions performantes de l'IA pour améliorer ce que nous proposons, cette question mérite de s'y pencher de manière approfondie.

Ainsi, nous ne prétendons pas avoir clôturé ce travail. Plusieurs chantiers restent inachevés, tantôt parce que, après avoir identifié le problème, nous ne parvenons pas à concevoir la solution, tantôt, tout simplement, faute de temps et de ressources (financières, humaines, technologiques) nécessaires. Nous espérons avoir tout mis en œuvre pour assurer la continuité et la reproductivité de notre recherche, que ce soit au cours de cette rédaction, que lors de la préparation et le partage des données produites. Nous la concluons donc non pas en tant qu'aboutissement, mais comme une étape et une démonstration des possibilités offertes par le croisement des regards interdisciplinaires et des outils informatiques.

Les perspectives de prolongement qui s'offrent à nous sont multiples, et certaines ont déjà été engagées.

Tout d'abord, même si nous avons proposé quelques analyses littéraires des relations révélées par notre traitement (notamment pour le sous-ensemble des réécritures symboliques), il ne s'agit que d'un petit pourcentage de tout ce qui peut être observé. Nous espérons que les études des cas que nous avons présentées tout au long de cette rédaction permettent déjà d'apprécier la pertinence du traitement que nous avons conçu. À présent, il paraît nécessaire de soumettre notre édition comparative aux collègues littéraires et linguistes : à des experts du mythe d'Orphée et Eurydice ou d'un des auteurs du corpus, à des spécialistes du processus de paraphrase et de reformulation, etc. Ainsi, nous pourrions évaluer ensemble tant la capacité de notre développement à fournir de nouveaux matériaux interprétatifs que la praticité de notre édition pour l'exploration et l'analyse du réseau des réécritures.

Parmi de nombreuses pistes qui restent à explorer par la critique littéraire, nous pourrions par exemple nous focaliser sur plusieurs sous-ensembles du corpus susceptibles de dévoiler des résultats intéressants, comme les opéras et les cantates, la littérature de jeunesse ou encore les parodies.

Pour ces dernières, nous constatons par exemple que les unités linguistiques communes sont pour la plupart identiques à celles appariées dans le sous-corpus des traductions, mais que leurs fréquences sont différentes. C'est particulièrement saillant pour le nom d'Eurydice, qui est le plus

souvent apparié dans le sous-corpus des parodies (403 relations), presque quatre fois plus que le nom d'Orphée (122 relations, en deuxième position). Dans le sous-corpus des traductions, c'est le nom d'Orphée qui apparaît le premier, avant celui d'Eurydice, mais les fréquences sont plus proches (130 *vs* 103 appariements).

Nous voyons également que beaucoup de relations entre parodies sont établies grâce aux mots communs relatifs au théâtre (« théâtre », « scène », « rideau », « sortir »), voire plus précisément à la *commedia dell'arte* (« Colombine », « Arlequin »). Nous y trouvons également des liens relatifs à la comédie (« ah ! » dans « ah ! ah ! ah ! », « rire ») ou à la dépréciation du sujet mythique (« diable », « violon », « appât », « vin », « vielleur », « tromper », « vieux », « payer », « veuf », « faux », « aveugle », « squelette », « sabot », « malice »...), qui sont absents des unités communes détectées au sein du sous-corpus des traductions.

Les relations détectées automatiquement entre les parodies prouvent la capacité de nos outils à accompagner le chercheur dans l'observation efficace des procédés qui produisent le décalage humoristique. Pour la cantate de Grandval1729, que notre traitement aligne sans difficulté avec son texte-source sérieux (la cantate de Clérembault1710), quatre relations sont établies avec d'autres parodies, et elles partagent toutes un vocabulaire particulier, généralement éloigné des mots-clés du mythe (Tableau 96)

Tableau 96. *Quatre relations détectées entre Grandval1729 et quatre autres parodies*

Grandval1729, phrase 40 : Je viens expres au Sabat , En robbe longue, en rabat, Chercher mon Euridice Qui manque à mon grabat .	Martin1649, phrase 1 : Un Vilon yure à la Feste La nuict m'a tant rompu la teste, M'a tant lassé dans mon grabat Par sa musique de sabat [...]
Grandval1729, phrase 24 : [...] Un vilain maudit Serpent remply de malice La guettant sur le gazon L'a piquée en trahison	Martin1649-2, phrase 10 : [...] Maudit soit-il, la male peste Du serpent couvert d'un gazon Qui t'a mordu en trahison
Grandval1729, phrase 20 : Par cette Épouse si chere Lachez l' objet de mes vœux	Deshayes1792, phrase 198 : Euridice, objet de mes vœux Est inutile dans ces lieux !
Grandval1729, phrase 33 : Allez Orphée, allez au Diable, Servez d'exemple à l' univers : Pour rejoindre un objet aimable Il est beau d'aller aux Enfers :	Deshayes1792, phrase 255 : Ah ! Si l' objet de ma constante flamme Bravait pour moi la mort et les enfers Heureuse encor de régner sur son âme Je le suivrais au bout de l' Univers .

L'observation comparative des relations établies entre toutes les parodies permettrait d'approfondir, voire nuancer, les analyses empiriques (qui ont par ailleurs été menées uniquement sur deux œuvres de ce sous-corpus composé d'œuvres majoritairement non canoniques : l'opéra-bouffe de Crémieux1858 et la pièce de Cocteau1921).

Il conviendrait également de vérifier si notre traitement permet d'analyser la manière dont les reformulations sont construites au sein de ce sous-corpus particulier. Les périphrases, métaphorisations ou changements de registre sont ici de précieux indices de la manière dont la langue véhicule l'humour et le rire. Ils permettent également de mieux déterminer ce qui caractérise les dépréciations de la trame principale du mythe.

Par ailleurs, le recensement des mots-clés caractéristiques des œuvres parodiques peut faciliter l'identification des passages humoristiques dans des textes qui ne sont pas, *a priori*, repérés comme en contenant. Dans ce contexte, l'amélioration des traitements (semi-)automatiques que nous proposons actuellement permettrait potentiellement l'identification des particularités linguistiques des textes parodiques, notamment en détectant les facteurs communs et les unités linguistiquement pertinentes, introduisant souvent une dépréciation ou une ridiculisation, comme certaines périphrases des entités (« Pluton » / « Prince enfumé »), certaines substitutions synonymiques (« se précipiter » / « dégringoler ») ou hyperonymiques (« femme » / « femelle »), certaines métaphorisations (« descendre aux Enfers » / « aller au Diable »), etc.

Plus encore, la compilation de ces particularités, ainsi que celles des autres sous-ensembles du corpus des réécritures, au sein d'une base de données lexicale (le *Trésor d'Orphée et Eurydice pour le TAL*) pourrait servir de ressource à un outil de détection automatique des allusions (brèves ou plus appuyées) au mythe d'Orphée et Eurydice dans n'importe quel corpus textuel, et de caractérisation des reprises (parodie, modernisation ou simple transposition) grâce au repérage du vocabulaire spécifique attesté. Cette méthode pourrait ensuite être enrichie par d'autres chercheurs afin de permettre la détection des allusions à d'autres œuvres et d'autres *topoi* plus ou moins récurrents.

Une autre perspective pour la suite du projet concerne l'application du développement à d'autres corpus et questions de recherche.

Nous l'avons déjà expérimenté au sein du projet eBalzac, qui s'intéresse, entre autres, aux relations intertextuelles entre les œuvres de Balzac et un corpus plus vaste des textes littéraires et scientifiques de la première moitié du XIX^e s. Nous avons mené deux expérimentations distinctes, qui nous ont permis de faire différentes observations¹.

La première confrontait la totalité des textes de *La Comédie humaine* au corpus romanesque des auteurs français contemporains ou antérieurs à Balzac à l'aide du logiciel TextPAIR. Les résultats initiaux ont été très abondants et incluaient par exemple un réseau complexe de 3 856 relations,

¹ Andrea Del Lungo et Karolina Suchecka, « Projet eBalzac : construire une bibliothèque hypertextuelle des sources intellectuelles », *op. cit.*; Andrea Del Lungo et Karolina Suchecka, « Le projet eBalzac : un hypertexte des sources scientifiques », *Balzac et les disciplines du savoir*, Classiques Garnier, à paraître.

mais dont les mots les plus courants étaient « vingt », « sept », « an », « trois » et « bien ». La pertinence des couples composant ce réseau était difficile à évaluer et le nombre des relations peu pertinentes (« La femme entre », « Il est dix-huit heures, etc.) restait très important. Cependant, en écartant les mille mots les plus fréquents dans la totalité du corpus et en allouant un poids plus important aux mots rares chez Balzac, nous sommes arrivés à n'obtenir que des relations pertinentes et qui ont permis de nombreuses découvertes, comme la documentation des aspects historiques des œuvres comme *Sur Catherine de Médicis* en s'appuyant sur les travaux de Chateaubriand ou encore des réécritures à quatre mains de Balzac (*Une fille d'Ève*, *Madame de Firmiani*) et Théophile Gautier (*Mademoiselle de Maupin*).

La visualisation à l'aide de graphes a également permis de rendre compte d'une forte proximité sémantique entre Balzac et Eugène Sue, assez inattendue au vu de l'approche critique que Balzac manifeste envers l'écriture de ce dernier. Les résultats ont également un certain intérêt pour les linguistes et les historiens de la langue : par exemple, plusieurs relations sont établies sur des expressions figées qui ne sont plus utilisées de nos jours, mais dont les occurrences sont nombreuses dans plusieurs œuvres du corpus (« méchant comme un âne rouge », « vouloir manger » ou « faire manger de la vache enragée », « conduite est un chef-d'œuvre politique », « une tête de saumon vaut mieux que dix mille grenouilles », etc.).

Cependant, la méthode de filtrage des résultats en écartant les mots très fréquents chez Balzac n'a pas été probante pour la confrontation du corpus de *La Comédie humaine* avec les ouvrages scientifiques dans les domaines des sciences naturelles, de la médecine et de la physiologie. En effet, nous n'aboutissions alors qu'à une dizaine des couples extrêmement proches, déjà révélées par les recherches empiriques. Pour obtenir des résultats plus riches, mais toujours pertinents, pour ce corpus scientifique très spécifique, nous avons eu recours à la détection des topiques caractéristiques en son sein. Nous sommes alors arrivés à délimiter environ 90 lemmes qui peuvent être considérés comme mots-clés des textes scientifiques (les parties du visage, le vocabulaire lié au cerveau et au système nerveux, la terminologie médicale, etc.). La restriction des résultats aux couples liés au moins par un de ces mots-clés a permis de retenir 34 % des couples initialement détectés par Tracer et de les classer en fonction du nombre des mots-clés communs. Le triage des résultats à base de leur pertinence a ensuite facilité de manière importante leur analyse qualitative, plus laborieuse que dans le cas de l'expérimentation avec le corpus romanesque, mais tout de même probante. Nous avons notamment pu affirmer que l'approche de Balzac face aux théories phrénologiques est particulière : il les adapte, voire les déforme, pour créer de nouveaux paradigmes. Les traitements automatiques ne sont pas en mesure de mettre en avant les réécritures aussi complexes,

mais peuvent au moins présenter aux experts des données filtrées, classées et modélisées pour faciliter leur interprétation.

Ainsi, grâce aux expérimentations menées sur d'autres corpus et posant d'autres questions de recherche, nous avons pu constater la flexibilité de notre développement et sa capacité à traiter les corpus très hétérogènes. Simultanément, nous avons pu enrichir notre méthodologie et réaliser l'importance de faire connaissance avec le corpus afin d'optimiser la détection et le filtrage des résultats. En ce qui concerne la détection de l'intertextualité quantitative, il semblerait en effet qu'il ne soit pas tant question de proposer des solutions prêt-à-porter, mais de concevoir sur mesure des chaînes de traitements qui conjuguent les méthodes quantitatives et qualitatives, les techniques informatiques et les analyses empiriques.

La dernière perspective enfin, concerne la méthodologie d'une recherche doctorale qui se veut dans l'esprit des humanités numériques. En effet, la particularité de ce champ est qu'il met l'accent sur la collaboration interdisciplinaire, ainsi que le partage des pratiques et des productions. Cette ouverture facilite la construction d'un réseau et d'une communauté aux intérêts communs, mais dont les objets et méthodes de recherche sont souvent très différents. De telles expériences interdisciplinaires sont extrêmement enrichissantes non seulement pour les jeunes chercheurs. Nous étions heureuse de pouvoir profiter de ces échanges. Ils nous ont permis de rencontrer des historiens programmeurs qui nous ont appris leurs méthodes de visualisation des réseaux complexes. Nous avons eu la chance de collaborer avec des ingénieurs et des chercheurs en science d'information et de communication afin de trouver la meilleure manière de partager nos données avec la communauté des chercheurs. Lors des colloques rejoignant les communautés aux horizons très différents, nous avons pu mieux comprendre la manière dont les linguistes conceptualisent le corpus, ou encore échanger avec des chercheurs en informatique qui travaillent sur le corpus littéraire, par exemple en exploitant le théâtre burlesque pour rendre les agents conversationnels moins misogynes. Nous espérons continuer à faire partie de cette communauté, mais aussi d'y amener davantage notre discipline de spécialité, la littérature comparée. En effet, nous nous attachons à démontrer non seulement l'utilité des approches issues des humanités numériques pour l'analyse comparative des œuvres littéraires, mais aussi la nécessité d'inscrire notre discipline dans la politique de la science ouverte. Nous sommes encore relativement peu nombreux à exploiter et exposer ces méthodes lors des rencontres comparatistes. Or, beaucoup de collègues manifestent de l'intérêt face aux outils, techniques et ressources que nous présentons.

En même temps, les enjeux éthiques du numérique et de l'informatique prennent davantage d'ampleur. D'autres types de corpus émergent, intermodaux, nativement numériques, etc. L'exploration des archives et bibliothèques numériques peut procurer la sensation de disposer d'une bibliothèque d'Alexandrie où tous les ouvrages sont présents de manière égalitaire et représentative. Cela risque, comme nous avons tâché de le démontrer au cours de ce travail, de provoquer des biais et de nuire davantage à la réputation des approches quantitatives dans la construction des corpus littéraires. Enfin, l'émergence des agents conversationnels hautement efficaces, au moins à première vue, et des grands modèles des langues offre sans doute de nouvelles possibilités d'analyse, mais pose en même temps des problèmes méthodologiques, éthiques et écologiques. Toutes ces questions sont étroitement liées avec notre recherche doctorale et elles doivent être posées dans l'immédiat. Nous avons pu entamer une réflexion sur certaines d'entre elles, que ce soit au sein de cette rédaction qu'à l'occasion des autres travaux et collaborations. D'autres, survenues au cours de notre thèse, n'ont pas encore pu être abordées, mais constituent un prolongement naturel, presque inévitable, de notre parcours académique.

Bibliographie

Corpus principal¹

- Aneau, Barthelemy et al., *La métamorphose d'Ovide figurée*, Lyon, Jan de Tournes, 1557, f° 61r-62r, 68v-69r. [Tournes1557]
- Anonyme, *L'Orphée grotesque, avec le bal rustique. En vers burlesques. Suivi de Suite de l'Orphée avec les Bacchantes ou Les Rudes Joyeuses*, Paris, Sebastien Martin, 1649. [Martin1649-1]
- , *Le Ballet des machines représentant le V^e acte de la Mort d'Orphée et d'Euridice*, Paris, 1700. [Martin1649-2]
- Anouilh, Jean, *Eurydice, suivi de Roméo et Jeannette*, Paris, Gallimard, coll. « Folio », 1971 [1942]. [Anouilh1942]
- Ballanche, Pierre-Simon, « Sixième Fragment », *Œuvres de M. Ballanche*, Paris, J. Barbezat, 1809, p. 488-492. [Ballanche1809]
- , *Essais de palingénésie sociale. Orphée*, Paris, Jules Didot, 1827-1829. [Ballanche1829]
- Berlioz, Hector, « Orphée. Cantate », *Séance publique de l'Académie Royale des Beaux-Arts, du 6 octobre 1827*, Didot, 1827, p. 11. [Berlioz1827]
- , *La mort d'Orphée. Monologue et Bacchanale à grands chœurs et à grand orchestre*, 1828. [Berlioz1828]
- Bottet, Béatrice et Aurore Pinet, « Les Métamorphoses d'Ovide », *La Mythologie en BD*, vol. 7, Paris, Casterman, 2016. [Bottet2016]
- Bruneau, Clotilde, Luc Ferry et Diego Oddi, *Orphée et Eurydice. Précédé de Déméter et Perséphone*, Paris, Glénat, 2019. [Ferry2019]
- Buffequin, Denis, *Dessein du poème et des superbes machines du mariage d'Orphée et d'Eurydice, qui se représentera sur le théâtre du Marais par les comédiens entretenus par Leurs Majestés*, Paris, R. Baudry, 1648. [Buffequin1648]
- Chapoton, François, *La grande Journée des machines, ou Le mariage d'Orphée et d'Euridice*, Paris, Chez Toussaint Quintet, 1648. [Chapoton1648]
- Clérembault, Nicolas, *Orphée, cantate française, à voix seule avec symphonie*, Paris, Bureau d'édition de la « Schola cantorum », 1710. [Clérembault1710]
- Cocteau, Jean, *Orphée. Tragédie en un acte et un intervalle*, Paris, Stock, 2005 [1927]. [Cocteau1927]
- Collande, Alexandre-Jacques-Pierre Le Gendre, *Opuscule d'un célèbre auteur égyptien. Contenant l'histoire d'Orphée, par laquelle on pourroit soupçonner qu'il est peu de femmes fideles*, Paris, Sébastien Jorry, 1752. [Collande1752]
- Courbois, Philippe, « Orphée. Troisième cantate à voix seule et violon », *Cantates françaises à I et II voix sans symphonie et avec symphonie*, Paris, Foucalt, 1710, p. 32-47. [Courbois1710]
- Defontenay, Charlemagne Ischir, « III. Orphée », *Études dramatiques. I. Barkokébas. II. Le Vieux de la*

¹ Voir aussi : les annexes, [I Présentation-corpus](#), « [I 3 Corpus-détaille.xlsx](#) ».

- montagne. III. Orphée. IV. Prométhée*, Paris, Ledoyen, 1854. [Defontenay1854]
- Deshayes, Prosper-Didier et Jean Rouhier-Deschamps, *Le Petit Orphée : Opéra comique*, Paris, Tourbon, 1792. [Deshayes1792]
- Divoire, Fernand, *Orphée*, Paris, Les Poètes de la Renaissance du Livre, 1922. [Divoire1922]
- Dorvigny et Pierre-Louis Moline, *Roger-Bontems et Javotte, parodie d'Orphée et Eurydice*, Paris, Veuve Duchesne, 1775. [Dorvigny1775]
- Eimann, D.-L., « Une Histoire renouvelée des Grecs », dans Alexandre Dumas (dir.), *Histoires cavalières*, Paris, Charpentier, 1854, p. 23-35. [Eimann1854]
- Esquiros, Alphonse et Adèle Esquiros, *Histoire des amants célèbres de l'antiquité*, Paris, Bureau de la Société Fraternelle de l'Intelligence et du Travail réunis, 1849, p. 101-146. [Esquiros1849]
- Fuzelier, Louis, « Orphée ou Arlequin aux Enfers, divertissement », dans Paul Fièvre (dir.), *Scaramouche pedant, divertissement*, Riom, G. Valleyre, 1711, p. 15-20. [Fuzelier1711]
- Goudot, Marie, *Les mystères d'Orphée*, Paris, L'école des loisirs, 2003. [Goudot2003]
- Grandmougin, Charles, *Orphée : drame en 4 actes 16 tableaux*, Paris, Calmann Lévy, 1882. [Grandmougin1882]
- Grandval, Nicolas Racot de, *Orphée. Cantate avec symphonie*, Paris, Boivin, Le Clerc, 1729. [Grandval1729]
- Guillion, Albert, *Orphée. Cantate composée à l'occasion de la fête de M. Henry Monton Berton, membre de l'Institut et chevalier de la Légion d'honneur*, 1770. [Guillion1770]
- Hus, Jean-Baptiste, *La Mort d'Orphée, ou les Fêtes de Bacchus, ballet héroïque de la composition de Mr Hus, représenté sur le théâtre de la Comédie-française*, Paris, Chez la V^{re} Delormel, et fils, 1759. [Hus1759]
- Hyacinthe-Loyson, Jean, *Orphée, poème*, Paris, Nouvelle Librairie de France, 1948. [Loyson1948]
- Jimenes, Guy, *Orphée l'enchanteur*, Paris, Nathan, 2004. [Jimenes2004]
- Lacroix, François de, *Orphée. Cantate avec symphonie*, Toulouse, 1714. [Lacroix1714]
- Laffon, Martine, *Orphée, l'ensorceleur*, Paris, Flammarion Jeunesse, 2016. [Laffon2016]
- Launay, Louis de, *Orphée : poème*, Paris, Jouve et Cie éditeurs, 1922. [Launay1922]
- Lemaire, Louis, *Orphée, cantatille nouvelle pour un dessus avec accompagnement de violons et flûtes*, 1743. [Lemaire1743]
- Lespine, Charles de, *Le mariage d'Orphée, sa descente aux enfers, sa mort par les bacchantes. Tragédie*, Paris, H. Sara, 1623. [Lespine1623]
- L'Hermite, Tristan, « Orphée aux Enfers parlant à Pluton », *Recueil de diverses poésies héroïques et burlesques. Contenant la Belle recluse, la Vieille layde, l'Amour honneste, le Doute amoureux, la Nuit amoureuse, l'Inquiétude amoureuse, les Soupirs de Silvie, Caprice burlesques, Orphée aux enfers, l'Aurore du bois de Vincennes et autres pièces curieuses, recueillies par le sieur T. L'Hermite*, Paris, la V^{re} G.

- Loyson et J.-B. Loyson, 1652, p. 37-39. [Hermite1652]
- , « La Lyre d'Orphée », *Les Amours de feu Mr Tristan et autres pièces très-curieuses [: La Lyre d'Orphée ; les Baisers de Dorinde]*, Paris, Chez Gabriel Quinet, 1662 [1641], p. 195-216. [Hermite1641]
- Liking, Werewere, *Orphée Dafric*, Paris, L'Harmattan, coll. « Encres noires », 2000 [1981]. [Liking1981]
- Lully, Louis de et Michel Du Boullay, *Orphée, tragédie en musique représentée par l'Académie royale de musique*, Paris, aux dépens de l'Académie, 1690. [Duboullay1690]
- Ma Njock, Manuna, *Orphée d'Afrique, théâtre-rituel*, Paris, L'Harmattan, coll. « Encres noires », 2000 [1981]. [Manjock1981]
- Moline, Pierre-Louis et Christophe Willibald von Glück, *Orphée et Eurydice, drame héroïque*, Paris, Académie Royale, 1774. [Moline1774]
- Montardre, Hélène, *Orphée aux Enfers*, Paris, Nathan, 2013. [Montardre2013]
- Offenbach, Jacques et Hector Crémieux, *Orphée aux Enfers*, Paris, Calmann-Lévy, 1858. [Crémieux1858]
- Piroye, Charles, *Le retour d'Eurydice aux enfers, suite de la cantate d'Orphée*, Paris, J-B-Christophe Ballard, 1717. [Piroye1717]
- Pommaux, Yvan, *Orphée et la morsure du serpent*, Paris, l'école des loisirs, 2009. [Pommaux2009]
- Puget de La Serre, Jean, « Les Amours d'Orphée », *Les amours des dieux, de Cupidon et Psyché, du Soleil et Clytie, de Jupiter et Danaé, de Jupiter et Io, de Jupiter et Calisto, de Neptune et Anphitrite, avec celles d'Orphée & sa descente aux enfers*, Paris, la Vve G. Loyson et J.-B. Loyson, 1624, p. 791-948. [Puget1624]
- Rachmuhl, Françoise, « Vie et mort d'Orphée », *16 métamorphoses d'Ovide*, Paris, Castor Poche Flammarion, 2003, p. 139-147. [Rachmuhl2003]
- Rameau, Jean-Philippe, *Orphée. Cantate*, Lille, 1721. [Rameau1721]
- Regnard, Jean-François, « La Descente d'Arlequin aux Enfers », *Œuvres complètes de J. F. Regnard*, vol. 5, Paris, Brière, 1823 [1689], p. 117-152. [Regnard1689]
- , « Orphée aux enfers, opéra », *Œuvres complètes de J. F. Regnard*, vol. 4, Paris, Brière, 1823 [1699], p. 343-351. [Regnard1699]
- Ronsard, Pierre de, « Orphée », *Les œuvres de Pierre de Ronsard*, Paris, N. Buon, 1609 [1563], p. 711-714. [Ronsard1563]
- Schuré, Édouard, « Livre V. Orphée (Les mystères de Dionysos) », *Les grands initiés : esquisse de l'histoire secrète des religions*, Paris, Perrin, 1889, p. 219-264. [Schuré1889]
- Segalen, Victor, *Orphée-roi*, Paris, Georges Crès, 1921. [Segalen1921]

Corpus secondaire²

- Legouais, Chrétien, *Ovide moralisé*, Rouen, Manuscrits de la BM de Rouen., 1301, f° 246v-249r, 250r-251r, 261v-261v, 271r-272r. [Legouais1301]
- Ovide, *Le grand Olympe des histoires poétiques du prince de poésie Ovide Naso en sa Métamorphose, œuvre authentique et de hault artifice, pleine de bonnesté récréation, traduyct de latin en françoys et imprimé nouvellement*, Lyon, R. Morin, 1532, 81r-83r, 2r-3r p. [Morin1532]
- , *Les 15 livres de la métamorphose d'Ovide, interprétés en rimes françoises*, trad. François Habert, Rouen, 1557, p. 593-601, 653-658. [Habert1557]
- , *Les Métamorphoses d'Ovide, de nouveau traduites en françois, avec XV discours contenant l'explication morale des fables*, trad. Nicolas Renouard, Paris, 1606, p. 445-451, 493-496 [Renouard1606]
- , *Les métamorphoses d'Ovide, mises en vers françois*, trad. Raimond de Massac et Charles de Massac, Paris, F. Pomeray et P. Rocolet, 1617, p. 441-446, 483-486. [Massac1617]
- , *Les metamorphoses, mises en vers françois*, trad. Thomas Corneille, vol. 2-3, Paris, Berthelemy Girin, Michel Brunet, 1687, p. 395-408, 1-7. [Corneille1687]
- , « Les Métamorphoses », *Les œuvres d'Ovide, traduction nouvelle, avec des remarques*, trad. Étienne Algay de Martignac, vol. 5-6, Lyon, Horace Molin, 1697, p. 437-449, 3-11. [Martignac1697]
- , *Les métamorphoses, avec des explications à la fin de chaque fable*, trad. Jean-Baptiste Morvan de Bellegarde, Paris, Michel David, 1701, p. 1-12, 66-70. [Bellegarde1701]
- , *Les Métamorphoses en latin et françois*, trad. Pierre du Ryer, Amsterdam, P. & J. Blaev, Janssons à Waesberge, Boome, & Goethals., 1702, p. 307-312, 343-345. [Duryer1702]
- , *Les Métamorphoses d'Ovide en latin, traduites en françois avec des remarques et des explications historiques*, trad. M. l'abbé Banier, Amsterdam, R. et J. Wetstein et G. Smith, 1732, p. 325-331, 360-363. [Banier1732]
- , *Les Métamorphoses d'Ovide, traduction nouvelle, avec le texte latin, suivie d'une analyse, de l'explication des fables, de notes géographiques, historiques, mythologiques et critiques*, trad. Mathieu-Guillaume-Thérèse Villenave, vol. 3, Paris, F. Gay, Ch. Guestard, 1806, p. 296-306, 380-386. [Villenave1806]
- , « Les Métamorphoses », *Œuvres complètes d'Ovide*, trad. Étienne Gros, vol. 5-6, Paris, 1834, p. 290-299, 4-10. [Gros1834]
- , « Métamorphoses », dans Paul Deproost (dir.), *Bibliotheca Classica Selecta*, trad. A.-M. Boxus et J. Poucet, Bruxelles, Université Catholique de Louvain (UCL), 2018, en ligne, <<http://bcs.fltr.ucl.ac.be/METAM/Met10/M10-1-142.htm>>. [Boxus2018]
- , *Métamorphoses d'Ovide. Traduction nouvelle conforme au texte latin du P. Jouvenci*, trad. Joseph-Gaspard Dubois-Fontanelle, vol. 2, Rouen (Paris), de l'imprimerie privilégiée (chez Nyon le jeune), 1789, p. 35-41, 59-63. [Fontanelle1789]
- , « Choix des Métamorphoses », dans Société de professeurs et de latinistes (dir.), *Les auteurs latins*

² Voir aussi : les annexes, [II Intertextualité](#), « [II 2 Corpus-secondaire.xlsx](#) ».

- expliqués d'après une méthode nouvelle par deux traductions françaises*, trad. Félix de Parnajon, Paris, Hachette, 1880, p. 418-429, 440-447. [Parnajon1880-1 et Parnajon1880-2]
- , « Métamorphoses », dans Désiré Nisard (dir.), *Œuvres complètes : avec la traduction en français*, trad. Louis Puget et Théodore Guiard, Paris, Firmin Didot Frères, 1869, p. 418-420. [Nisard1869]
- , « Livre X : Orphée et Eurydice, Livre XI : La mort d'Orphée », *Les visages d'Orphée*, trad. Annick Béague, Jacques Boulogne, Alain Deremetz et Françoise Toulze, Villeneuve d'Ascq, Presses universitaires de Septentrion, 1998, p. 36-38, 39-40. [Beague1998]
- Virgile, « Les Georgiques », *Les œuvres de Virgile. Translatees de latin en [vers] francoyses*, trad. Michel Guillaume (dit de Tours), Paris, Jean Petit, Pierre Vidoue, 1540 [1519], p. 19-29. [Tours1519]
- , *Aristée, épisode du quatrième livre des Géorgiques*, trad. Joseph Cogolin, (Paris), 1750, p.19-29. [Cogolin1750]
- , *Les Géorgiques de Virgile, traduction nouvelle en vers françois*, trad. Jacques Delille, Paris, C. Bleuet, 1770, p. 233-239. [Delille1770]
- , *Les Géorgiques de Virgile traduites en vers français, avec le texte latin à côté, accompagnées de notes relatives à l'agriculture, à l'astronomie, à la géographie, à l'histoire, à la mythologie et à la poésie*, trad. Antoine de Cournand, Paris, Bernard, 1805, p. 322-331. [Courmand1805]
- , « Géorgiques », *Œuvres de Virgile traduites en françois, le texte vis-à-vis la traduction, avec des remarques*, trad. Pierre-François Desfontaines, vol. 1, Lyon, J.H. Ayné, 1810, p. 211-215. [Desfontaines1810]
- , *Les bucoliques et les géorgiques de Virgile, traductions nouvelles*, trad. Charles Héguin de Guerle, Paris, Auguste Delalain, 1827, p. 281-289. [Héguin1827]
- , « Géorgiques », *Œuvres complètes de Virgile, traduction nouvelle*, trad. Jean-Pierre Charpentier, vol. 1, Paris, C. L. F. Panckoucke, 1831, p. 259-265. [Charpentier1831]
- , *Œuvres de Virgile traduites en vers français : Bucoliques, Géorgiques*, trad. Louis Duchemin, Paris, Hachette, 1837, p. 295-301. [Duchemin1837]
- , *Les Géorgiques, traduction vraie*, trad. Un naturaliste, Paris, Veuve Guillois, 1863, p. 65-68. [Guillois1863]
- , *Les Géorgiques : essais de traductions lyriques*, trad. François Fournier, Firmin et Cabirou, Montpellier, 1876, p. 122-126. [Fournier1876]
- , « Géorgiques », *Œuvres complètes de Virgile*, trad. J.-R.-T. Cabaret-Dupaty, Paris, Hachette, 1897, p. 101-104. [Cabaret-Dupaty1897]
- , *Les Bucoliques et les Géorgiques*, trad. Maurice Rat, Paris, Classiques Garnier, 1932. [Rat1932]
- , « Géorgiques », dans Désiré Nisard (dir.), *Lucrèce, Virgile, Valérius Flaccus : œuvres complètes, avec la traduction en français*, trad. Désiré Nisard, Paris, Firmin Didot Frères, 1868, p. 225-227. [Nisard1868]
- , *Géorgiques, quatrième livre*, trad. Auguste Desportes, Paris, Hachette, 1863. [Desportes1863]

—, « Géorgique, IV, 450-557 », *Les visages d'Orphée*, trad. Annick Béague, Jacques Boulogne, Alain Deremetz et Françoise Toulze, Villeneuve d'Ascq, Presses universitaires de Septentrion, 1998, p. 32-33. [Béague1998]

Walleys, Thomas, *La Bible des poètes, Métamorphose [d'Ovide moralisée]*, trad. Colard Mansion, Paris, Antoine Vérard, 1493, f° 107v-109v, 119r-199v. [Walleys1493]

York, Maître de Marguerite d', *Ovide moralisé en prose II*, Bruges, 1470, f° 132v-135r, 147r-148r. [York1470]

Études consacrées au mythe d'Orphée et ses réécritures

Archimandritis, Georgios, « Le mythe d'Orphée dans le théâtre et le cinéma du XX^e siècle », Presses universitaires du Septentrion, 2002.

Béague, Annick et al., *Les visages d'Orphée*, Villeneuve d'Ascq, Presses universitaires du Septentrion, coll. « Savoirs Mieux », 2019.

Bellas, Jacqueline, « “Orphée” au XIX^e et au XX^e siècles : interférences littéraires et musicales », *Cahiers de l'AIEF*, vol. 22, n° 1, 1970, p. 229-246, en ligne, <https://www.persee.fr/doc/caief_0571-5865_1970_num_22_1_962>.

Bleschet, Émilie, *Les représentations du mythe d'Orphée du XVI^e au XIX^e siècle*, mémoire de maîtrise, Philippe Martin (dir.), Lyon, Enssib, Université Lumière Lyon 2, 2016, en ligne, <<https://www.enssib.fr/bibliotheque-numerique/documents/67143-les-representations-du-mythe-d-orphee-du-xvie-au-xixe-siecle.pdf>>.

Boom, Romijn Helene, « Pierre-Simon Ballanche et l'orphisme : une tentative d'herméneutique », Thèse de doctorat, University of Cape Town, 1988.

Bouloumié, Arlette, « La Résurgence du mythe d'Eurydice et ses métamorphoses dans l'œuvre d'Anouilh, de Pascal Quignard, d'Henri Bosco, de Marguerite Yourcenar, de Michèle Sarde, et Jean Loup Trassard », *Loxias*, Université de Nice, Nice, 2004.

Brunel, Pierre, « *Orphée-roi* de Victor Segalen, ou le miracle de la lyre », *Mythocritique : théorie et parcours*, Paris, Presses Universitaires de France, coll. « Écriture », 1992, p. 190-202.

—, « Orphée moderne », *Apollinaire entre deux mondes. Le contrepoint mythique dans Alcools. Mythocritique II*, Paris, Presses Universitaires de France, coll. « Écriture », 1997, p. 63-82.

Busca, Maurizio, « La mise en recueil des *Métamorphoses* d'Ovide aux XVI^e et XVII^e siècles en France », *Pratiques et formes littéraires. Cahiers du Gadges*, n° 17, Institut d'Histoire des Représentations et des Idées dans les Modernités, décembre 2020, en ligne, <<http://publications-prairial.fr/pratiques-et-formes-litteraires/index.php?id=199>>.

Cellier, Léon, « Le Romantisme et le mythe d'Orphée », *Cahiers de l'AIEF*, vol. 10, n° 1, 1958, p. 138-157, en ligne, <https://www.persee.fr/doc/caief_0571-5865_1958_num_10_1_2128>.

Cerrito, Stefania, *Une relecture renaissance de l'Ovide moralisé en prose Le Grand Olympe des Histoires poétiques du prince de la poésie Ovide (Lyon, 1530-1532)*, Classiques Garnier, coll. « Ovidiana », 2021, en ligne, <DOI : 10.48611/isbn.978-2-406-11517-5.p.0173>.

- Décaudin, Michel, « Deux aspects du mythe orphique au XX^e siècle : Apollinaire, Cocteau », *Cahiers de l'AIEF*, vol. 22, n° 1, 1970, p. 215-227, en ligne, <https://www.persee.fr/doc/caief_0571-5865_1970_num_22_1_961>.
- Dekens, Julie, « Rester aux Enfers : le bonheur paradoxal d'Eurydice », *TRANS-. Revue de littérature générale et comparée*, n° 17, Presses Sorbonne Nouvelle, février 2014, en ligne, <<http://journals.openedition.org.ressources-electroniques.univ-lille.fr/trans/910>>.
- Durand, Gilbert, « Les nostalgies d'Orphée. Petite leçon de mythanalyse », *Réligiologiques*, n° 15, 1997, en ligne, <<https://www.religiologiques.uqam.ca/no15/durand.html>>.
- Fauquet, Joël-Marie, « Berlioz and Gluck », dans Peter Bloom (dir.), *The Cambridge Companion to Berlioz*, Cambridge, Cambridge University Press, 2000, p. 199-210.
- Friedman, John Block, *Orphée au Moyen Âge*, trad. Jean-Michel Roessli, Fribourg, Cerf, coll. « Vestigia », 1999.
- Gyssels, Kathleen, « Sartre postcolonial ? Relire Orphée noir plus d'un demi-siècle après », *Cahiers d'études africaines*, vol. 45, n° 179-180, Éditions de l'École des hautes études en sciences sociales, décembre 2005, p. 631-650, en ligne, <<http://journals.openedition.org/etudesafri-caines/14952>>.
- Hauptmann, Barunka, « “Son corps déchiqueté” : les genres à l'œuvre dans le mythe d'Orphée au XIX^e siècle », Thèse de doctorat, Université de Paris IV, 2013.
- Heurgon, Jacques, « Orphée et Eurydice avant Virgile », *Mélanges de l'école française de Rome*, vol. 49, n° 1, 1932, p. 6-60, en ligne, <https://www.persee.fr/doc/mefr_0223-4874_1932_num_49_1_7222>.
- Joukovsky, Françoise, *Orphée et ses disciples dans la poésie française et néolatine du XVI^e siècle*, Droz, 1970.
- Juden, Brian, « Particularités du mythe d'Orphée chez Ballanche », *Cahiers de l'AIEF*, vol. 22, n° 1, 1970, p. 137-152, en ligne, <DOI : 10.3406/caief.1970.956>.
- Kushner, Eva, *Le mythe d'Orphée dans la littérature française contemporaine*, Paris, A.G. Nizet, 1961.
- , « Le Personnage d'Orphée chez Ronsard », *Lumières de la Pléiade*, Paris, Vrin, coll. « Neuvième stage international d'études humanistes, Tours 1965 », 1966, p. 271-302.
- , « Orphée et l'orphisme chez Victor Segalen », *Cahiers de l'AIEF*, vol. 22, n° 1, 1970, p. 197-214, en ligne, <https://www.persee.fr/doc/caief_0571-5865_1970_num_22_1_960>.
- Lavaud, Florence, *Le personnage d'Orphée : du mythe latin au roman médiéval*, mémoire de maîtrise, Limoges, Faculté des Lettres et Sciences Humaines, 1995.
- Paulson, William R., « Fragment et autobiographie dans l'œuvre de Ballanche : étude et textes inédits », *Nineteenth-Century French Studies*, vol. 15, n° 1/2, 1986, p. 14-32.
- Porra, Véronique, « Sur quelques Orphée noirs. Reproduction, adaptation et hybridation du mythe d'Orphée en contexte post(-)colonial », *Revue de littérature comparée*, vol. 344, n° 4, Klincksieck, 2012, p. 441-455, en ligne, <<https://www.cairn.info/revue-de-litterature-comparee-2012-4-page-441.htm>>.

Riesz, János, « « Orphée Noir » – « Schwarzer Orpheus » – « Black Orpheus » : Quand les marges se mettent en marche vers un centre commun », *Revue de littérature comparée*, vol. 314, n° 2, Klincksieck, 2005, p. 161-177.

Sartre, Jean-Paul, « Orphée noir », dans Léopold Sédar Senghor (dir.), *Anthologie de la nouvelle poésie nègre et malgache de langue française*, Paris, Presses Universitaires de France, 1948.

Turmel, Émilie, « Orphée et l'orphisme dans *À la recherche du temps perdu* de Marcel Proust », *Études françaises*, vol. 51, n° 1, Les Presses de l'Université de Montréal, 2015, p. 141-161, en ligne, <<https://www.erudit.org/fr/revues/etudfr/2015-v51-n1-etudfr01694/1028526ar/>>.

Verrier, Jean, « Segalen lecteur de Segalen », *Poétique*, n° 27, 1976, p. 338-350.

Théorie littéraire et littérature comparée

Adam, Jean-Michel, *Souvent textes varient : génétique, intertextualité, édition et traduction*, Paris, Classiques Garnier, coll. « Investigations stylistiques », 2018.

Adam, Jean-Michel et Ute Heidmann, *Le texte littéraire : pour une approche interdisciplinaire*, Louvain-la-Neuve, Bruylant-Academia, coll. « Au cœur des textes », 2009.

Apter, Emily S., *Against world literature: on the politics of untranslatability*, London ; New York, Verso, 2013.

Arrivé, Michel, *Les Langages de Jarry : essai de sémiotique littéraire*, Paris, Klincksieck, 1972.

—, « Intertextualité et intratextualité chez Ferdinand de Saussure », *Le Plaisir de l'intertexte*, Berne, Peter Lang, 1986, p. 11-36.

Bakhtine, Mikhaïl, *La Poétique de Dostoïevski*, trad. Isabelle Kolitcheff, Paris, Seuil, 1970 [1929].

—, *L'œuvre de François Rabelais et la culture populaire au Moyen Âge et sous la Renaissance*, trad. Andrée Robel, Paris, Gallimard, coll. « Tel », 1970 [1965].

—, *Problèmes de la poétique de Dostoïevski*, trad. Guy Verret, Lausanne, L'Âge d'Homme, 1970 [1929].

Barthes, Roland, « Texte (Théorie du) », dans *Encyclopædia Universalis*, vol. 15, Paris, 1973, p. 1013-1017.

—, *Le Plaisir du texte*, Paris, Seuil, 1975.

Begum-Lees, Rebecca, « Que(e)r(y)ing Iphis' Transformation in Ovid's *Metamorphoses* », dans Alison Surtees et Jennifer Dyer (dir.), *Exploring Gender Diversity in the Ancient World*, Edinburgh University Press, 2020, p. 106-117, en ligne, <<https://www.cambridge.org/core/books/exploring-gender-diversity-in-the-ancient-world/queerying-iphis-transformation-in-ovids-metamorphoses/6DEDAFEB8F7236AF2691C2FD25C3CF38>>.

Béhar, Henri, « Le collage, ou le pagure de la modernité », *Cahiers du XX^e siècle*, n° 5, 1975, p. 43-68.

—, « La réécriture comme poétique – ou le même et l'autre », *Romantic Review*, n° 72/1, janvier 1981.

- Biaisi, Pierre-Marc de, « Intertextualité (Théorie de l') », dans *Encyclopædia Universalis*, vol. Corpus, Paris, 1989, p. 514-516.
- Brunel, Pierre, *Mythocritique : théorie et parcours*, Paris, Presses universitaires de France, coll. « Écriture », 1992.
- Chevrel, Yves et Camille Dumoulié (dir.), « Introduction », *Le mythe en littérature. Essais en hommage à Pierre Brunel*, 1. éd, Paris, Presses universitaires de France, coll. « Écriture », 2000.
- Chevrel, Yves, « Les études de réception », dans Pierre Brunel (dir.), *Précis de littérature comparée*, 1^{re} éd., Paris, Presses Universitaires de France, 1989, p. 177-214.
- Compagnon, Antoine, *La Seconde main, ou le travail de la citation*, Paris, Seuil, 1979.
- , *La notion de genre*, Fabula, 2001, en ligne, <<https://www.fabula.org/compagnon/genre.php>>.
- Domino, Maurice, « La réécriture du texte littéraire Mythe et Réécriture », *Semen. Revue de sémiotique linguistique des textes et discours*, n° 3, Presses universitaires de Franche-Comté, 1987, en ligne, <<http://journals.openedition.org/semen/5383>>.
- Dumézil, Georges, *Mythe et épopée*, Paris, Gallimard, coll. « Bibliothèque des sciences humaines », 1968.
- Durand, Gilbert, *Figures mythiques et visages de l'œuvre : de la mythocritique à la mythanalyse*, Paris, Berg international, coll. « L'Ile verte », 1979.
- , *Introduction à la mythodologie : mythes et sociétés*, Paris, Albin Michel, coll. « La pensée et le sacré », 1996.
- Eco, Umberto, *Interprétation et surinterprétation*, trad. Jean-Pierre Cometti, Paris, Presses universitaires de France, coll. « Formes sémiotiques », 1996.
- Eigeldinger, Marc, « Introduction », *Mythologie et intertextualité*, Genève, Slatkine, 1987, p. 9-20.
- Engélibert, Jean-Paul et Yen-Mai Tran-Gervat (dir.), *La littérature dépliée : reprise, répétition, réécriture*, Rennes, Presses Universitaires de Rennes, coll. « Interférences », 2008.
- Ferrer, Daniel, « Quelques remarques sur le couple intertextualité-genèse », *La création en acte. Devenir de la critique génétique*, vol. 289, Brill, 2007, p. 203-216, en ligne, <<https://brill.com/view/book/9789401203531/B9789401203531-s015.xml>>.
- Fitch, Brian, « L'Intra-intertextualité interlinguistique de Beckett, la problématique de la traduction en soi », *Texte*, n° 2, 1983, p. 85-100.
- Franco, Bernard, *La littérature comparée. Histoire, domaines, méthodes*, Paris, Armand Colin, 2016.
- Garcia, Anne, « La réécriture dans l'œuvre de José Emilio Pacheco. Une poétique du déjà-lu », Thèse de doctorat, Université Paris-Est, 2016.
- Gefen, Alexandre, *L'idée de littérature : de l'art pour l'art aux écritures d'intervention*, Paris, Éditions Corti, coll. « Les essais », 2021.
- Genette, Gérard, *Introduction à l'architexte*, Paris, Seuil, coll. « Poétique », 1979.

- , *Palimpsestes. La littérature au second degré*, Paris, Seuil, coll. « Poétique », 1982.
- Gignoux, Anne-Claire, *Initiation à l'intertextualité*, Paris, Ellipses, coll. « Initiation à », 2005.
- , « De l'intertextualité à la réécriture », *Cahiers de Narratologie. Analyse et théorie narratives*, n° 13, REVEL, 2006, en ligne, <<http://journals.openedition.org.ressources-electroniques.univ-lille.fr/narratologie/329>>.
- Goethe, Johann Wolfgang von, « Traductions », *Divan occidental-oriental*, trad. Henri Lichtenberger, Paris, Aubier, 1940 [1819], p. 430-433.
- Goyet, Francis, « “Imitatio” ou intertextualité ? », *Poétique*, n° 71, septembre 1987, p. 313-320.
- Heidmann, Ute, « Différencier au lieu d'universaliser. Comparer les façons de (r)écrire des mythes », *Interférences littéraires/Littéraire interférenties*, n° 17, 2015, p. 15-34, en ligne, <<http://interferenceslitteraires.be/index.php/illi/article/view/117>>.
- Jenny, Laurent, « La Stratégie de la forme », *Poétique*, n° 27, 1976, p. 257-281.
- , « Les genres littéraires », *Méthodes et problèmes*, Genève, Département de français moderne, 2003, en ligne, <<https://www.unige.ch/lettres/framo/enseignements/methodes/genres/index.html>>.
- , « Méthodes sans questions, questions sans méthodes », *Versants. Revue suisse des littératures romanes*, vol. 1, n° 63, « À quoi bon l'enseignement de la littérature ? », octobre 2017, p. 75-82, en ligne, <<https://bop.unibe.ch/versants/article/view/3574>>.
- Kristeva, Julia, « Pour une sémiologie des paragrammes », *Tel Quel*, n° 29, 1967, p. 53-75.
- , « Problèmes de la structuration du texte », *La Nouvelle Critique*, n° spécial, « Linguistique et littérature », 1968, p. 55-64.
- , *Séméiotikè. Recherches pour une sémanalyse*, Paris, Seuil, coll. « Tel quel », 1969.
- , *La révolution du langage poétique : l'avant-garde à la fin du xix^e siècle, Lautréamont et Mallarmé*, Paris, Seuil, coll. « Points », 1974.
- Lavocat, Françoise, « Le comparatisme comme herméneutique de la défamiliarisation », *Vox-Poetica*, avril 2012, en ligne, <<http://www.vox-poetica.org/t/articles/lavocat2012.html>>.
- , *Fait et fiction : pour une frontière*, Paris, Seuil, coll. « Poétique », 2016.
- Limat-Letellier, Nathalie, « Historique du concept d'intertextualité », dans Marie Miguet-Ollagnier (dir.), *L'intertextualité*, Besançon, Presses universitaires de Franche-Comté, coll. « Annales littéraires », 1998, p. 17-64, en ligne, <<http://books.openedition.org/pufc/4507>>.
- Lombez, Christine, « Réécriture et traduction », dans Jean-Paul Engélibert et Yen-Mai Tran-Gervat (dir.), *La littérature dépliée : Reprise, répétition, réécriture*, Rennes, Presses universitaires de Rennes, coll. « Interférences », 2008, p. 71-80.
- Marghescou, Mircea, *Le concept de littérarité : critique de la métalittérature*, Paris, Kimé, coll. « Détours littéraires », 2009 [1974].

- McCawley, James, « D. 1968. The Role of Semantics in a Grammar », dans Emmon Bach et R. Harms (dir.), *Universals in Linguistic Theory*, Holt, Rinehart, and Winston, 1968, p. 125-170.
- Mendenhall, T. C., « The Characteristic Curves of Composition », *Science*, vol. 9, n° 214, 1887, p. 237-249, en ligne, <<https://www.jstor.org/stable/pdf/1764604>>.
- Mombert, Sarah et Pascal Durand (dir.), « Ce que causer veut dire », *Entre presse et littérature : Le Mousquetaire, journal de M. Alexandre Dumas (1853-1857)*, Liège, Presses universitaires de Liège, coll. « Bibliothèque de la faculté de philosophie et lettres de l'université de Liège », 2019, p. 45-63, en ligne, <<http://books.openedition.org/pulg/5385>>.
- Moretti, Franco, « The Slaughterhouse of Literature », *MLQ: Modern Language Quarterly*, vol. 61, n° 1, Duke University Press, 2000, p. 207-227, en ligne, <<https://muse.jhu.edu/article/22852>>.
- , « Conjectures on World Literature », *New Left Review*, n° 1, février 2000, p. 54-68.
- , « Hypothèses sur la littérature mondiale », *Études de lettres*, n° 258 (2), trad. Raphaël Micheli, 2001, p. 9-24, en ligne, <<https://enseignement.typepad.fr/files/moretti-hypotheses.pdf>>.
- , « Graphes (Modèles abstraits pour une autre histoire de la littérature) », *Graphes, cartes et arbres. Modèles abstraits pour une autre histoire de la littérature*, trad. Étienne Dobenesque, Paris, Les Prairies ordinaires, 2008, en ligne, <<http://www.vox-poetica.org/t/articles/moretti.html>>.
- , *Distant reading*, London ; New York, Verso, 2013.
- , « Les études littéraires entre herméneutique et quantification », *Revue italienne d'études françaises. Littérature, langue, culture*, n° 10, trad. Anne-Frédérique Schläpfer, Revue italienne d'études françaises, novembre 2020, en ligne, <<http://journals.openedition.org/rief/6508>>.
- Pavel, Thomas, *Univers de la fiction*, trad. Thomas Pavel, Paris, Seuil, coll. « Poétique », 1988.
- Pouey-Mounou, Anne-Pascale, « Du Bellay, Ronsard, Tristan : autour de quelques sonnets des *Amours* », *Cahiers Tristan L'Hermite*, n° XLIII, Paris, Classiques Garnier, « Tristan et la poésie du XVI^e siècle », 2021, p. 35-58, en ligne, <<https://classiques-garnier.com/cahiers-tristan-l-hermite-2021-n-xliii-tristan-et-la-poesie-du-xvie-siecle-du-bellay-ronsard-tristan-autour-de-quelques-sonnets-des-amours.html>>.
- Ricardou, Jean, *Pour une théorie du nouveau roman*, Paris, Seuil, coll. « Tel quel », 1971.
- , « Pour une théorie de la réécriture », *Poétique*, n° 77, 1989, p. 3-15.
- Riffaterre, Michael, « La trace de l'intertexte », *La Pensée*, n° 215, octobre 1980, p. 4-18.
- Rolet, Serge, « L'affleurement du contre-texte chez Volochinov et Bakhtine », dans Alain Trouvé (dir.), *La lecture littéraire*, n° 12, Reims, « Le contre-texte », 2014, p. 17-30.
- Rougemont, Denis de, *Lumières du mythe*, Paris, Presses Universitaires de France, 1983.
- Rouget, François, « *Ordo neglectus* : remarques sur la disposition des recueils poétiques de Tristan L'Hermite », *Cahiers Tristan L'Hermite*, n° XLIII, Paris, Classiques Garnier, « Tristan et la

poésie du XVI^e siècle », 2021, p. 151-181, en ligne, <https://classiques-garnier.com/doi/garnier?article=CatMS42_151>.

Ruprecht, Hans-George, « Intertextualité », *Texte*, n° 2, 1983, p. 13-22.

Saussy, Haun, « La lecture, pratique dissidente », *Le Comparatisme comme approche critique*, vol. 3. Objets, méthodes et pratiques comparatistes, Paris, Classiques Garnier, coll. « Rencontres », 2017, p. 59-73, en ligne, <<https://classiques-garnier.com/le-comparatisme-comme-approche-critique-comparative-literature-as-a-critical-approach-tome-3-objets-methodes-et-pratiques-comparatistes-objects-methods-practices-la-lecture-pratique-dissidente-en.html>>.

Schaeffer, Jean-Marie, *Qu'est-ce qu'un genre littéraire ?*, Paris, Seuil, coll. « Poétique », 1989.

Schnyder, Peter (dir.), *Métamorphoses du mythe : réécritures anciennes et modernes des mythes antiques*, Paris, Orizons, coll. « Universités – domaine littéraire », 2008.

Seris, Émilie, « Ronsard, imitateur de Politien ? », dans Perrine Galand (dir.), *Camenae*, n° 22, « La réception d'Ange Politien en France au XVI^e siècle », décembre 2018, en ligne, <<https://www.saprat.fr/toutes-les-revues-en-ligne-camenae/camenae-n-22-decembre-2018-la-reception-drange-politien-en-france-au-xvie-siecle-699.htm>>.

Siess, Jürgen, « “Parallèle”, un concept opératoire en Littérature comparée ? », *Revue de Littérature comparée*, n° 298, LXXV, avril-juin 2001, p. 225-230.

Todorov, Tzvetan, *Symbolisme et interprétation*, Paris, Seuil, coll. « Poétique », 1978.

Linguistique et traitement automatique du langage

Adam, Jean-Michel, *La linguistique textuelle*, 3^e éd., Paris, A. Colin, coll. « Cursus », 2011 [2005].

Almogi, Orna et al., « A Hackathon for Classical Tibetan », *Journal of Data Mining & Digital Humanities*, vol. Special Issue on..., n° Towards a Digital Ecosystem:..., janvier 2019, p. 2047, en ligne, <<https://jdmhd.episciences.org/2047>>.

Altschul, S. F. et al., « Basic local alignment search tool », *Journal of Molecular Biology*, vol. 215, n° 3, octobre 1990, p. 403-410, en ligne, <DOI : 10.1016/S0022-2836(05)80360-2>.

Arbach, Najib et Saandia Ali, « Aspects théoriques et méthodologiques de la représentativité des corpus », *Corela. Cognition, représentation, langage*, n° HS-13, Université de Poitiers, décembre 2013, en ligne, <<https://journals.openedition.org/corela/3029>>.

Barzilay, Regina et Michael Elhadad, « Using Lexical Chains for Text Summarization », *Intelligent Scalable Text Summarization*, 1997.

Barzilay, Regina et Noemie Elhadad, « Sentence Alignment for Monolingual Comparable Corpora », *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 2003, p. 25-32.

Barzilay, Regina et Kathleen R. McKeown, « Extracting Paraphrases from a Parallel Corpus », *Proceedings of the 39th Annual Meeting of the ACL*, Toulouse, Association for Computational Linguistics, 2001, p. 50-57.

- Bhagat, Rahul et Eduard Hovy, « What Is a Paraphrase? », *Computational Linguistics*, vol. 39, n° 3, septembre 2013, p. 463-472, en ligne, <https://doi.org/10.1162/COLI_a_00166>.
- Biber, Douglas, « Representativeness in Corpus Design », *Literary and Linguistic Computing*, vol. 8, n° 4, janvier 1993, p. 243-257, en ligne, <<https://doi.org/10.1093/lc/8.4.243>>.
- Biber, Douglas, Edward Finegan et Dwight Atkinson, « ARCHER and its challenges: Compiling and Exploring a Representative Corpus of Historical English Registers », dans Udo Fries, Peter Schneider et Gunnel Tottie (dir.), *Creating and using English language corpora. Papers from the 14th International Conference on English Language Research on Computerized Corpora, Zurich 1993*, Amsterdam, Rodopi, 1994, p. 1-13.
- Calabrese-Steimberg, Laura, « Nom propre et dénomination événementielle : quelles différences en langue et en discours ? », *Corela. Cognition, représentation, langage*, n° 7-1, Université de Poitiers, juin 2009, en ligne, <<http://journals.openedition.org/corela/173>>.
- Cardon, Rémi et Natalia Grabar, « Construction d'un corpus parallèle à partir de corpus comparables pour la simplification de textes médicaux en français », *Traitement Automatique des Langues*, vol. 61, n° 2, 2020, p. 15-39.
- Chomsky, Noam, *Structures syntaxiques*, trad. Michel Braudeau, Paris, Seuil, 1969 [1957].
- Consortium, BNC, *British National Corpus, XML edition*, Oxford Text Archive, 2007.
- Coseriu, Eugenio, *Textlinguistik : Eine Einführung*, Francke, Tübingen-Basel, 1994 [1980].
- , « Le double problème des unités “DIA-S” », *Les cahiers DIA : Études sur la diachronie et la variation linguistique*, 1998, p. 9-16.
- Ehrmann, Maud, « Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation », Thèse de doctorat en Informatique et Traitement Automatique des Langues, Paris 7, 2008.
- Ehrmann, Maud, Matteo Romanello, Simon Clematide, Phillip Ströbel et Raphaël Barman, *Language Resources for Historical Newspapers: the Impreso Collection*, Marseille, European Language Resources Association (ELRA), 2020, p. 958-968.
- Eshkol-Taravella, Iris et Natalia Grabar, « Repérage et analyse de la reformulation paraphrastique dans les corpus oraux », *Actes de la 21ème conférence sur le Traitement Automatique des Langues Naturelles*, 2014, p. 304-315.
- Ferrero, Jérémy et Alain Simac-Lejeune, « Détection automatique de reformulations – Correspondance de concepts appliquée à la détection du plagiat », *Revue des Nouvelles Technologies de l'Information*, vol. Extraction et Gestion des Connaissances, RNTI-E-28, 2015, p. 287-298.
- Francis, W. Nelson, « A Standard Corpus of Edited Present-Day American English », *College English*, vol. 26, n° 4, National Council of Teachers of English, 1965, p. 267-273, en ligne, <<https://www.jstor.org/stable/373638>>.
- , « Language corpora B.C. », dans Jan Svartvik (dir.), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991*, Berlin, New York, De Gruyter Mouton,

- 2011, p. 17-32, en ligne, <<https://www.degruyter.com/document/doi/10.1515/9783110867275.17/html>>.
- Fuchs, Catherine, « La paraphrase entre la langue et le discours », *Langue française*, vol. 53, n° 1, 1982, p. 22-33, en ligne, <https://www.persee.fr/doc/lfr_0023-8368_1982_num_53_1_5113>.
- Gibbon, Dafydd, Roger Moore et Richard Winski (dir.), *Spoken language system and corpus design*, Berlin, New York, De Gruyter, coll. « Handbook of standards and resources for spoken language systems », 1998.
- Ghamnia, Adel et al., « Extraction de relations : combiner les techniques pour s'adapter à la diversité du texte », *28^{es} Journées Francophones d'Ingénierie des Connaissances (IC 2017)*, Caen, France, 2017.
- Grabar, Natalia et Pierre Zweigenbaum, « Lexically-Based Terminology Structuring: Some Inherent Limits », *Terminology*, vol. 10, 2004, p. 23-54.
- , *Utilisation de corpus de spécialité pour le filtrage de synonymes de la langue générale*, Dourdan, 2005, p. 81-90.
- Grice, H. Paul, « Logic and conversation », dans John P. Kimball, Jerry L. Morgan et Peter Cole (dir.), *Syntax and semantics*, New York San Francisco London, Academic press, Harcourt Brace Jovanovich, 1975, p. 41-58.
- Hajlaoui, Najeh et al., « DCEP -Digital Corpus of the European Parliament », *Proc. LREC 2014*, Reykjavik, Iceland, 2014, p. 3164-3171.
- Harris, Zellig S., « Co-Occurrence and Transformation in Linguistic Structure », *Language*, vol. 33, n° 3, Linguistic Society of America, 1957, p. 283-340, en ligne, <<https://www.jstor.org/stable/411155>>.
- , « Introduction to Transformations », dans Zellig S. Harris (dir.), *Papers in Structural and Transformational Linguistics*, Dordrecht, Springer Netherlands, coll. « Formal Linguistics Series », 1970 [1956], p. 383-389, en ligne, <https://doi.org/10.1007/978-94-017-6059-1_22>.
- Hatmi, Mohamed, « Reconnaissance des entités nommées dans des documents multimodaux », Thèse de doctorat en Informatique et Traitement Automatique des Langues, Université de Nantes, 2014.
- Hunston, Susan, « Collection strategies and design decisions », dans Anke Lüdeling et Merja Kytö (dir.), *Corpus Linguistics. An International Handbook*, Berlin, De Gruyter Mouton, 2008, p. 154-168, en ligne, <<https://www.degruyter.com/document/doi/10.1515/9783110211429.1.154/html>>.
- Ibrahim, Ali, Boris Katz et Jimmy Lin, « Extracting Structural Paraphrases from Aligned Monolingual Corpora », *Proceedings of the Second International Workshop on Paraphrasing*, Sapporo, Japon, Association for Computational Linguistics, 2003, p. 57-64.
- Iyer, Parvati et Abhispita Singh, *Document Similarity Analysis for a Plagiarism Detection System*, Pune, India, 2005.
- Jakobson, Roman, *Essais de linguistique générale*, trad. Nicolas Ruwet, Paris, Éditions de Minuit, 1963.

- Jolles, André, *Formes simples*, trad. Antoine Marie Buguet, Paris, Seuil, 1972.
- Kytö, Merja (dir.), *Manual to the diachronic part of the Helsinki corpus of English texts: coding conventions and lists of source texts*, 3. ed, Helsinki, Dep. of Engl. Univ. of Helsinki, 1996.
- Landragin, Frédéric, « Expressions référentielles et chaînes de référence en français : le projet Democrat et son exploration des rapports entre linguistique textuelle et linguistique de corpus », *Écho des études romanes*, vol. 18, n° 1, Écho des études romanes, novembre 2022, p. 49-65, en ligne, <<https://doi.org/10.32725/eer.2022.004>>.
- Laurent, Nicolas, « Le nom propre dans les Lettres de l'année 1671 de Madame de Sévigné », *Séminaire Madame de Sévigné*, n° 1, « Séminaires de recherche », 2013, en ligne, <<http://publisshs.univ-rouen.fr/ceredi/index.php?id=519>>.
- Leech, Geoffrey, « New resources, or just better old ones? The Holy Grail of representativeness », *Corpus Linguistics and the Web*, vol. 59, Brill, coll. « Language and Computers », 2007, p. 133-149, en ligne, <<https://brill.com/view/book/edcoll/9789401203791/B9789401203791-s009.xml>>.
- Le Meur, Céline, Sylvain Galliano et Edouard Geoffrois, *Conventions d'annotations en Entités Nommées*, ESTER, coll. « Rapport technique de la campagne Ester », 2004.
- Longo, Laurence et Amalia Todirascu, « Une étude de corpus pour la détection automatique de thèmes », *Texte et corpus*, vol. 4, janvier 2013, p. 143-155.
- Martin, Louis, Benjamin Muller, Perdo J. O. Suárez, Yoann Dupont et Laurent Romary, *Camembert: a Tasty French Language Model*, 2020.
- Melčuk, Igor, « Paraphrase et lexique : la théorie Sens-Texte et le Dictionnaire explicatif et combinatoire », *Lexique*, n° 6, 1992, p. 9-58, en ligne, <http://olst.ling.umontreal.ca/pdf/Mel%27%C4%8Duk_1992_Paraphrase_DEC-3.pdf>.
- , *Un modèle linguistique fonctionnel : le modèle Sens-Texte*, conférence organisée par S. Krylosova, Paris, Inalco, 22 octobre 2015, en ligne, <https://www.inalco.fr/sites/default/files/asset/document/inalco_modele_linguistique_fonctionnel.pdf>.
- Murat, Michel, « La périphrase : remarques autour d'une figure », *L'information grammaticale*, vol. 13, n° 1, 1982, p. 38-40, en ligne, <https://www.persee.fr/doc/igram_0222-9838_1982_num_13_1_2379>.
- Qi, Peng et al., *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages*, Association for Computational Linguistics (ACL) System Demonstrations, 2020.
- Ruder, Sebastian, « Dependency parsing », dans *NLP-progress*, en ligne, <http://nlpprogress.com/english/dependency_parsing.html>.
- Sagot, Benoît, « Représentation de l'information sémantique lexicale : le modèle wordnet et son application au français », *Revue Française de Linguistique Appliquée*, vol. XXII, 2017, en ligne, <<https://inria.hal.science/hal-01583995>>.

- Sagot, Benoît et Darja Fišer, *WordNet Libre du Français (WOLF)*, Institut national de recherche en sciences et technologies du numérique (Inria), 2012.
- Savan, David, « La sémiotique de Charles S. Peirce », *Langages*, vol. 14, n° 58, 1980, p. 9-23, en ligne, <https://www.persee.fr/doc/lgge_0458-726x_1980_num_14_58_1844>.
- Sampson, Geoffrey, *English for the computer: the SUSANNE corpus and analytic scheme*, Oxford : New York, Clarendon Press ; Oxford University Press, 1995.
- Sinclair, John, « Reference corpora », *EAGLES. Preliminary recommendations on Corpus Typology*, 1996, en ligne, <<http://www.ilc.cnr.it/EAGLES96/corpusyp/node18.html>>.
- Spärck Jones, Karen, « Computational Linguistics: What About the Linguistics? », *Computational Linguistics*, vol. 33, n° 3, 2007, p. 437-441, en ligne, <<https://www.aclweb.org/anthology/J07-3008>>.
- Teubert, Wolfgang, « La linguistique de corpus : une alternative », *Semen. Revue de sémio-linguistique des textes et discours*, n° 27, trad. Aurélie Lebaud, Presses universitaires de Franche-Comté, juin 2009, en ligne, <<https://journals.openedition.org/semen/8923#tocto1n10>>.
- The London–Lund corpus of spoken English: Description and research*, vol. 82, Lund University Press, coll. « Lund Studies in English », 1990.
- Varga, Dániel et al., « Parallel corpora for medium density languages », dans Nicolas Nicolov et al. (dir.), *Current Issues in Linguistic Theory*, vol. 292, Amsterdam, John Benjamins Publishing Company, 2007, p. 247-258, en ligne, <<https://benjamins.com/catalog/cilt.292.32var>>.
- Vila, Marta, M. Antònia Martí et Horacio Rodríguez, « Paraphrase Concept and Typology. A Linguistically Based and Computationally Oriented Approach », *Procesamiento de Lenguaje Natural*, n° 46, 2011, p. 83-90.
- , « Is This a Paraphrase? What Kind? Paraphrase Boundaries and Typology », *Open Journal of Modern Linguistics*, vol. 4, n° 1, Scientific Research Publishing, 2014, p. 205-218, en ligne, <<http://www.scirp.org/journal/PaperInformation.aspx?PaperID=43894>>.
- Vila Rigat, Marta, « Paraphrase Scope and Typology. A Data-Driven Approach from Computational Linguistics / Abast i tipologia de la paràfrasi. Una aproximació empírica des de la lingüística computacional », Thèse de doctorat, Universitat de Barcelona, 2013.
- Wahl, François, « Texte [375]. Le texte comme productivité », dans Oswald Ducrot et Cvetan Todorov (dir.), *Dictionnaire encyclopédique des sciences du langage*, Paris, Seuil, coll. « Points Série essais », 1993, p. 443-448.

Études consacrées à la détection automatique de l'intertextualité

- Barber, Mathew, « Adventures in Alignments: Training an Algorithm to Recognise Text Reuse », dans *KITAB*, 7 août 2020, en ligne, <<https://kitab-project.org/Adventures-in-Alignments-Training-an-Algorithm-to-Recognise-Text-Reuse/>>.
- Bergroth, Lasse, Harri Hakonen et Timo Raita, « A Survey of Longest Common Subsequence Algorithms », *Proceedings of the Seventh International Symposium on String Processing Information Retrieval*, Corogne, 2000, p. 39-48.

- Bernstein, Neil, Kyle Gervais et Wei Lin, « Comparative rates of text reuse in classical Latin hexameter poetry », *Digital Humanities Quarterly*, vol. 9, n° 3, décembre 2015.
- Boukhaled, Mohamed-Amine, Zied Sellami et Jean-Gabriel Ganascia, « Phoebus : un Logiciel d'Extraction de Réutilisations dans des Textes Littéraires », *22ème Conférence sur le Traitement Automatique des Langues Naturelles*, Caen, France, 2015.
- Büchler, Marco et al., « Detection of citations and text reuse on ancient greek texts and its applications in the classical studies: eAQUA project », *Proceedings of the Digital Humanities*, 2010.
- , « Informationstechnische Aspekte des Historical Text Re-use », Université de Leipzig, 2013.
- , « Mining and analysing one billion requests to linguistic services », *2016 IEEE International Conference on Big Data (Big Data)*, 2016, p. 3230-3239.
- Burns, Patrick J., « Measuring and Mapping Intergeneric Allusion in Latin Poetry using Tesseract », *Journal of Data Mining & Digital Humanities*, vol. Special Issue on..., n° Towards a Digital Ecosystem:..., août 2017, p. 3821, en ligne, <<https://jdmhdh.episciences.org/3821>>.
- , « Profiling of Intertextuality in Latin Literature Using Word Embeddings », *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online, Association for Computational Linguistics, 2021, p. 4900-4907.
- Coffee, Neil et al., « Intertextuality in the Digital Age », *Transactions of the American Philological Association*, vol. 142, n° 2, 2012, p. 383-422, en ligne, <DOI : 10.1353/apa.2012.0010>.
- , « The Tesseract Project: intertextual analysis of Latin poetry », *Literary and Linguistic Computing*, vol. 28, n° 2, 2012, p. 221-228, en ligne, <https://www.academia.edu/3830928/The_Tesseract_Project_Intertextual_Analysis_of_Latin_Poetry>.
- , « The Tesseract Project: intertextual analysis of Latin poetry », *Literary and Linguistic Computing*, vol. 28, n° 2, 2012, p. 221-228, en ligne, <https://www.academia.edu/3830928/The_Tesseract_Project_Intertextual_Analysis_of_Latin_Poetry>.
- , « Intertextuality as Viral Phrases: Roses and Lilies », De Gruyter Saur, 2019.
- Coffee, Neil et Christopher W. Forstall, « Claudian's engagement with Lucan in his historical and mythological hexameters », dans Valéry Berlincourt, Lavinia Galli Milić et Damien Nelis (dir.), *Lucan and Claudian: context and intertext*, Winter Verlag, Heidelberg, 2016, p. 255-284.
- Cooney, Charles et Clovis Gladstone, *Tracing Swerves of Influence : text reuse and the reception of Lucretius in 18th-century England*, DH2017, Montréal, 10 août 2017.
- , « Opening New Paths for Scholarship: Algorithms to Track Text Reuse in ECCO », dans Simon Burrows et Glenn Roe (dir.), *Digitizing Enlightenment: Digital Humanities and the Transformation of Eighteenth-Century Studies*, Oxford University Studies in the Enlightenment, Voltaire Foundation in association with Liverpool University Press, 2020.
- Del Lungo, Andrea et Karolina Suchecka, « Le projet eBalzac : un hypertexte des sources scientifiques », *Balzac et les disciplines du savoir*, Classiques Garnier, à paraître.

- , « Projet eBalzac : construire une bibliothèque hypertextuelle des sources intellectuelles », dans Clarisse Bardiot, Émilien Ruiz et Esther Dehoux (dir.), *La Fabrique numérique des corpus en sciences humaines et sociales*, Presses universitaires du Septentrion, 2022.
- Düring, Marten et al., *Impresso Text Reuse at Scale*, 2023.
- Fedchenko, Valentina, Dario Maria Nicolosi et Glenn Roe, « À la recherche des réseaux intertextuels : défis de la recherche littéraire à grande échelle », *Humanités numériques*, n° 9, Humanistica, juin 2024, en ligne, <<https://journals.openedition.org/revuehn/3940#bodyftn2>>.
- Forstall, Christopher et Walter Scheirer, *Quantitative Intertextuality: Analyzing the Markers of Information Reuse*, Cham, Springer International Publishing, 2019, p. 23-52, en ligne, <<https://doi.org/10.1007/978-3-030-23415-7>>.
- Forstall, Christopher, Neil Coffee, Thomas Buck, Katherine Roache et Sarah Jacobson, « Modeling the scholars: Detecting intertextuality through enhanced word-level n-gram matching », *Digital Scholarship in the Humanities*, vol. 30, n° 4, 2015, p. 503-515, en ligne, <DOI : 10.1093/llc/fqu014>.
- Forstall, Christopher, Sarah Jacobson et Walter Scheirer, « Evidence of Intertextuality: Investigating Paul the Deacon's *Angustae Vitae* », *Literary & Linguistic Computing (LLC)*, 2011, p. 285-296.
- Franzini, Greta, « English translations of Pan Tadeusz: a comparison with TRACER », *Corpus-based Research in the Humanities workshop (CRH)*, Varsovie, Pologne, 2016.
- , « Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm », *Frontiers in Digital Humanities*, vol. 5, Frontiers, avril 2018, en ligne, <<https://www.frontiersin.org/journals/digital-humanities/articles/10.3389/fdigh.2018.00004/full>>.
- , « Towards a Historical Text Re-use Detection », dans C. Biemann et A. Mehler (dir.), *Text Mining. From Ontology Learning to Automated Text Precession Applications*, Suisse, Springer International Publishing, 2014, p. 221-238, en ligne, <DOI : 10.1007/978-3-319-12655-5_11>.
- , « Using and evaluating TRACER for an Index fontium computatus of the Summa contra Gentiles of Thomas Aquinas », dans Elena Cabrio, Alessandro Mazzei et Fabio Tamburini (dir.), *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018 : 10-12 December 2018, Torino*, Torino, Accademia University Press, coll. « Collana dell'Associazione Italiana di Linguistica Computazionale », 2018, p. 199-205, en ligne, <<https://books.openedition.org/aaccademia/3354>>.
- Funk, Kellen et Lincoln A. Mullen, « A Servile Copy: Text Reuse and Medium Data in American Civil Procedure », *Rechtsgeschichte – Legal History*, n° 24, Max-Planck-Institut fuer europäische Rechtsgeschichte, Klostermann Verlag, 2016, en ligne, <<http://dx.doi.org/10.12946/rg24/341-343>>.
- Ganascia, Jean-Gabriel, « Détection automatique de phénomènes intertextuels », *Genesis. Manuscripts – Recherche – Invention*, n° 51, Sigales, décembre 2020, p. 63-77, en ligne, <<https://journals.openedition.org/genesis/5671>>.

- Ganascia, Jean-Gabriel, Pierre Glaudes et Andrea Del Lungo, « Automatic Detection of Reuses and Citations in Literary Texts », *Digital Scholarship in the Humanities*, vol. 29, n° 3, 2014, p. 412-421.
- Gawley, James O et A Caitlin Diddams, « Comparing the intertextuality of multiple authors using Tesseract: A new technique for normalization », *Digital Scholarship in the Humanities*, vol. 32, n° suppl_2, décembre 2017, p. ii53-ii59, en ligne, <<https://doi.org/10.1093/llc/fqx038>>.
- Horton, Russell, Mark Olsen et Glenn Roe, « Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections », *Digital Studies / Le Champ numérique*, vol. 2, n° 1, 2010, en ligne, <DOI : 10.16995/DSCN.258>.
- Jänicke, Stefan et al., « Visualizations for Text Re-use », *Proceedings of the 5th International Conference on Information Visualization Theory and Applications*, 2014, p. 59-70.
- Kokkinakis, Dimitrios et Mats Malm, « Detecting Reuse of Biblical Quotes in Swedish 19th Century Fiction using Sequence Alignment », *Corpus-based Research in the Humanities workshop (CRH)*, Varsovie, Pologne, 2016, p. 79-86.
- Korf, Ian, Mark Yandell et Joseph Bedell, *BLAST. An Essential Guide to the Basic Local Alignment Search Tool*, 1^{re} éd., Sebastopol, CA, O'Reilly & Associates, 2003.
- Manjavacas, Enrique, Brian Long et Mike Kestemont, « On the Feasibility of Automated Detection of Allusive Text Reuse », *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, Minneapolis, USA, Association for Computational Linguistics, 2019, p. 104-114.
- Mullen, Lincoln A., *Textreuse: Detect Text Reuse and Document Similarity*, R, 2020.
- Oberreuter, Gabriel et Juan D. Velásquez, « Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style », *Expert Systems with Applications*, vol. 40, n° 9, juillet 2013, p. 3756-3763, en ligne, <<https://www.sciencedirect.com/science/article/pii/S0957417412013231>>.
- O'Neill, Helen et al., « Text mining Mill: Computationally detecting influence in the writings of John Stuart Mill from library records », *Digital Scholarship in the Humanities*, vol. 36, n° 4, décembre 2021, p. 1013-1029, en ligne, <<https://doi.org/10.1093/llc/fqab010>>.
- Romanello, Matteo, « Exploring Citation Networks to Study Intertextuality in Classics », *Digital Humanities Quarterly*, vol. 10, n° 2, juin 2016.
- Romanello, Matteo, Aurélien Berra et Alexandra Trachsel, « Rethinking Text Reuse as Digital Classicists », *Digital Humanities conference*, Lausanne, 2014.
- Romanello, Matteo et Simon Hengchen, « Detecting Text Reuse with Passim », *Programming Historian*, 2021, en ligne, <<https://programminghistorian.org/en/lessons/detecting-text-reuse-with-passim>>.
- Smith, David A. et al., « Detecting and modeling local text reuse », *IEEE/ACM Joint Conference on Digital Libraries*, IEEE, 2014, p. 183-192.

- Smith, David A., Ryan Cordell et Elizabeth Maddock Dillon, « Infectious texts: Modeling text reuse in nineteenth-century newspapers », *2013 IEEE International Conference on Big Data*, IEEE, 2013, p. 86-94.
- Smith, Temple F. et Michael S. Waterman, « Identification of common molecular subsequences », *Journal of Molecular Biology*, vol. 147, n° 1, 1981, p. 195-197, en ligne, <<https://www.sciencedirect.com/science/article/pii/0022283681900875>>.
- Suchecka, Karolina et Nathalie Gasiglia, « Réécritures d'un mythe et outils de détection des réutilisations. De l'Orphée de Virgile à celui de Ballanche », *Humanités numériques*, n° 4, Humanistica, 2021, en ligne, <<https://journals.openedition.org/revuehn/2467>>.
- , « TAL et Littérature comparée. Détection automatique des correspondances textuelles entre les réécritures d'un mythe (NLP and Comparative Literature) », *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier TAL et Humanités Numériques (TAL-HN)*, Avignon, France, ATALA, 2022, p. 88-98.
- , « Focus sur les traitements linguistiques d'un projet d'humanités numériques. Édition et cartographie hypertextuelle de réécritures », *Essais*, vol. 21, « Linguistique et interdisciplinarité », 2024, en ligne, <<https://journals.openedition.org/essais/13158>>.
- , « On Digital Comparative Editions and Textual Similarity Detection Tools: Towards a Hypertextual Cartography of a Rewritten Myth », dans Petr Plecháč et al. (dir.), *Tackling the Toolkit: Plotting Poetry through Computational Literary Studies*, Institute of Czech Literature of the Czech Academy of Sciences, 2022, p. 167-182, en ligne, <https://www.plotting-poetry.org/books/tackling_toolkit/chapter?chap=11>.
- Suchecka, Karolina, Nathalie Gasiglia et Karl Zieger, « Édition comparative intermédiaire de séries traductives : exploiter les homologues pour créer des visualisations modulables », *Traitement Automatique des Langues*, vol. 60, n° 3, « TAL et humanités numériques », décembre 2019, p. 37-62, en ligne, <https://www.atala.org/sites/default/files/TAL-60-3_SucheckaEtAl_EditionComparativeIntermedialeDeSeriesTraductives-final.pdf>.
- Tharsen, Jeffrey et Clovis Gladstone, « Using Philologic For Digital Textual and Intertextual Analyses of the Twenty-Four Chinese Histories 二十四史 », *Journal of Chinese History 中國歷史學刊*, vol. 4, n° 2, juillet 2020, p. 558-563, en ligne, <<https://www.cambridge.org/core/journals/journal-of-chinese-history/article/abs/using-philologic-for-digital-textual-and-intertextual-analyses-of-the-twentyfour-chinese-histories/E390EDB0BEFC2F96F1B16265BF8A4395>>.
- Vesanto, Aleksu et al., « A System for Identifying and Exploring Text Repetition in Large Historical Document Corpora », *Proceedings of the 21st Nordic Conference on Computational Linguistics*, Gothenburg, Sweden, Association for Computational Linguistics, 2017, p. 330-333.
- , « Applying BLAST to Text Reuse Detection in Finnish Newspapers and Journals, 1771-1910 », *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, Gothenburg, Linköping University Electronic Press, 2017, p. 54-58.
- , « Detecting and Analyzing Text Reuse with BLAST », mémoire de Master, University of Turku, 2019.

Vierthaler, Paul et Mees Gelein, « A BLAST-based, Language-agnostic Text Reuse Algorithm with a MARKUS Implementation and Sequence Alignment Optimized for Large Chinese Corpora », *Journal of Cultural Analytics*, vol. 4, n° 2, 2019.

Édition numérique et numérisation

Ariel, Mira, « Accessibility theory: An overview », dans Ted Sanders, Joost Schilperoord et Wilbert Spooren (dir.), *Human Cognitive Processing*, vol. 8, Amsterdam, John Benjamins Publishing Company, 2001, p. 29, en ligne, <<https://benjamins.com/catalog/hcp.8.04ari>>.

Casenave, Joana, *L'édition critique numérique : une nouvelle approche du patrimoine littéraire et documentaire*, Paris, Honoré Champion, coll. « Lettres numériques », 2023.

Chartier, Roger, « Qu'est-ce qu'un livre ? Grandeurs et misères de la numérisation », dans Gérard Fussman (dir.), *La mondialisation de la recherche : Compétition, coopérations, restructurations*, Paris, Collège de France, coll. « Conférences », 2013, en ligne, <<http://books.openedition.org/cdf/1579>>.

Del Lungo, Andrea, « L'édition numérique eBalzac. Une nouvelle lecture de *La Comédie humaine* », *The Balzac Review / Revue Balzac*, n° 4, « L'édition / Publishing », 2021, p. 119-140, en ligne, <https://classiques-garnier.com/doi/garnier?article=RbzMS04_119>.

Del Lungo, Andrea et Giovanni Pietro Vitali, « La littérature du XIX^e siècle au prisme des humanités numériques », *Romantisme*, vol. 191, n° 1, 2021, p. 64-75, en ligne, <<http://www.cairn.info/revue-romantisme-2021-1-page-64.htm>>.

Del Lungo, Andrea et Karolina Suchecka, « Projet eBalzac : construire une bibliothèque hypertextuelle des sources intellectuelles », dans Clarisse Bardirot, Émilien Ruiz et Esther Dehoux (dir.), *La Fabrique numérique des corpus en sciences humaines et sociales*, Presses universitaires du Septentrion, 2022.

Epron, Benoît et Marcello Vitali-Rosati, *L'édition à l'ère numérique*, Paris, la Découverte, coll. « Repères », 2018.

Gabay, Simon et Thibault Clérice, « CATMuS-Print [Large] », *Reconnaissance des écritures dans les imprimés. CATMuS print: un modèle générique, multilingue et diachronique*, Zenodo, 2024, en ligne, <<https://zenodo.org/records/10592716>>.

Galleron, Ioana, Marie-Luce Demonet, Cécile Meynard, Fatiha Idmhand, Elena Pierazzo, Geoffrey Williams, Julia Roger et Pierre-Yves Buard, *Les publications numériques de corpus d'auteurs. Guide de travail, grille d'analyse et recommandations*, Huma-Num CAHIER, 2018.

Gallet, Olivier et al., « Apollinaire numérique », *Revue d'histoire littéraire de la France*, vol. 116, n° 3, 2016, p. 533-546.

Gibson, James, « The Theory of Affordances », *The Ecological Approach to Visual Perception*, Boston, Houghton Mifflin, 1979, p. 127-137.

Glorieux, Frédéric, « Le traitement de textes (odt) pour produire des documents structurés (XML/TEI) — Odette », dans *J'attends des résultats*, 2015, en ligne, <<https://resultats.hypotheses.org/267>>.

- Guillot, Céline et al., *La « philologie numérique » : tentative de définition d'un nouvel objet éditorial*, ATILF/SLR, 2017.
- Idmhand, Fatiha et Ioana Galleron, *Guide pour la FAIRisation des données des corpus d'auteurs*, Rapport de recherche, version 2, Huma-Num, 6 septembre 2020.
- Jeanneney, Jean Noël, *Quand Google défie l'Europe : plaidoyer pour un sursaut*, 3^e éd., Paris, Fayard/Mille et une nuits, coll. « Essai », 2010.
- Manovich, Lev, *What is Visualization?*, 2010, en ligne, <<http://manovich.net/index.php/projects/what-is-visualization>>.
- McGann, Jerome, « Electronic Archives and Critical Editing », *Literature Compass*, vol. 7, n° 2, 2010, p. 37-42, en ligne, <<https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1741-4113.2009.00674.x>>.
- McLoughlin, Tim, « Bridging the Gap », *Jahrbuch für Computerphilologie*, vol. 10, 2008, en ligne, <<http://computerphilologie.digital-humanities.de/jg08/mclough.pdf>>.
- Mori, S., C.Y. Suen et K. Yamamoto, « Historical review of OCR research and development », *Proceedings of the IEEE*, vol. 80, n° 7, 1992, p. 1029-1058, en ligne, <<https://ieeexplore.ieee.org/document/156468>>.
- Nunberg, Geoffrey, « Google's Book Search: A Disaster for Scholars », *The Chronicle of Higher Education*, section The Review, 2009.
- Pierazzo, Elena, *Digital Scholarly Editing: Theories, Models and Methods*, Ashgate, Farnham, Surrey, 2015.
- , « What future for digital scholarly editions? From *Haute Couture* to *Prêt-à-Porter* », *International Journal for Digital Humanities*, vol. 1, Springer, 2019, p. 1-12, en ligne, <<https://hal.archives-ouvertes.fr/hal-02117714>>.
- Robinson, Peter, « Editing Without Walls », *Literature Compass*, vol. 7, n° 2, 2010, p. 57-61, en ligne, <<https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1741-4113.2009.00676.x>>.
- , « Toward a Theory of Digital Editions », *Variants*, vol. 10, 2012, p. 105-131.
- Sampling Criteria for ELTeC – Proposal*, COST Action Distant Reading, WG 1 expert meeting, Prague, 13 février 2018, en ligne, <https://distantreading.github.io/sampling_slides.pdf>.
- Schöch, Christof, « Théâtre Classique – review », *RIDE*, n° 8, « Text Collections », 2018, en ligne, <<https://ride.i-d-e.de/issues/issue-8/theatre-classique/>>.
- Stokes, Peter, « eScriptorium recognised in prestigious DataX grant from Princeton University », dans *eScripta*, [text/html], eScripta, 28 mars 2022, en ligne, <<https://escripta.hypotheses.org/486>>.
- Shneiderman, Ben, « The eyes have it: a task by data type taxonomy for information visualizations », *Proceedings 1996 IEEE Symposium on Visual Languages*, 1996, en ligne, <DOI: 10.1109/VL.1996.545307>.

- Sinclair, Stéfano, Stan Ruecker et Milena Radzikowska, « Information Visualization for Humanities Scholars », *Literary Studies in the Digital Age: An Evolving Anthology*, MLA Commons, 2013.
- Suchecka, Karolina et Nathalie Gasiglia, « Édition comparative numérique des réécritures du mythe d'Orphée et Eurydice. Cartographie hypertextuelle des correspondances », dans Pascal Mougin (dir.), *Littérature et design : visualités et visualisations du texte en régime numérique*, Dijon, les presses du réel, coll. « Figures », 2023, p. 243-266.
- Suchecka, Karolina, Victoria Le Fournier et Andrea Del Lungo, « Phœbus e-Balzac : édition numérique exhaustive d'un monument littéraire », dans Clarisse Bardiot, Esther Dehoux et Émilien Ruiz (dir.), *La fabrique numérique des corpus en sciences humaines et sociales*, Villeneuve d'Ascq, Presses universitaires du Septentrion, coll. « Humanités numériques et science ouverte », 2022, p. 278-296, en ligne, <http://www.septentrion.com/html/WYSIWYGfiles/files/CONTENEUR2108_Bardiot_Dehoux_Ruiz.pdf>.
- Vitali-Rosati, Marcello et Michael E. Sinatra, « Pratiques de l'édition numérique », Montréal, Presses de l'Université de Montréal, coll. « Parcours numérique », 2014.
- Walsh, John A., « Comic Book Markup Language: An Introduction and Rationale », *Digital Humanities Quarterly*, vol. 6, n° 1, 2012, p. 43.
- Wieviorka, Michel, *L'impératif numérique ou La nouvelle ère des sciences humaines et sociales*, Paris, CNRS, 2013.

Humanités numériques et informatique

- Azoulay, Benjamin et Benoît de Courson, *Gallicagram : un outil de lexicométrie pour la recherche*, SocArXiv, 2021.
- Blei, David M., « Probabilistic topic models », *Communications of the ACM*, vol. 55, n° 4, avril 2012, p. 77-84, en ligne, <<https://dl.acm.org/doi/10.1145/2133806.2133826>>.
- Baillargeon, Sophie, *Utilisation de packages R*, Université de Laval, 2021.
- Brett, Megan R., « Topic Modeling: A Basic Introduction », *Journal of Digital Humanities*, vol. 2, n° 1, 2012, en ligne, <<http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/>>.
- Burdick, Anne et al., *Digital Humanities*, Cambridge, Mass, MIT Press, 2012.
- Burnard, Lou, « Du *literary and linguistic computing* aux *digital humanities* : retour sur 40 ans de relations entre sciences humaines et informatique », dans Pierre Mounier (dir.), *Read/Write Book 2 : Une introduction aux humanités numériques*, Marseille, OpenEdition Press, coll. « Read/Write Book », 2012, p. 45-58, en ligne, <<http://books.openedition.org/oep/242>>.
- Burrows, John, « 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship », *Literary and Linguistic Computing*, vol. 17, n° 3, septembre 2002, p. 267-287, en ligne, <<https://doi.org/10.1093/lc/17.3.267>>.
- Cabral, Maria de Jesus et al. (dir.), *Lire de près, de loin. Close vs distant reading*, Paris, Classiques Garnier, 2014.

- Cafiero, Florian et Jean-Baptiste Camps, « Why Molière most likely did write his plays », *Science Advances*, vol. 5, n° 11, American Association for the Advancement of Science, novembre 2019, en ligne, <<https://www.science.org/doi/10.1126/sciadv.aax5489>>.
- Cohen, Dan, « Docker : qu'est-ce que c'est et comment l'utiliser ? », dans *DataScientest*, 23 décembre 2020, en ligne, <<https://datascientest.com/docker-guide-complet>>.
- Craig, D. H. et Arthur F. Kinney, *Shakespeare, computers, and the mystery of authorship*, New York, Cambridge University Press, 2009.
- Doueihi, Milad, *Pour un humanisme numérique*, Paris, Seuil, 2011.
- Eder, Maciej, Jan Rybicki et Mike Kestemont, « Stylometry with R: A Package for Computational Text Analysis », *The R Journal*, vol. 8, n° 1, 2016, p. 107-121, en ligne, <<https://journal.r-project.org/archive/2016/RJ-2016-007/index.html>>.
- Evert, Stefan et al., « Understanding and explaining Delta measures for authorship attribution », *Digital Scholarship in the Humanities*, vol. 32, 2017, p. 4-16, en ligne, <DOI : 10.1093/llc/fqx023>.
- Fenoglio, Irène et Jean-Gabriel Ganascia, « Le logiciel MEDITE: approche comparative de documents de genèse », dans Aurèle Crasson (dir.), *L'édition du manuscrit - De l'archive de création au scriptorium électronique*, vol. 10, 2008, p. 209-228.
- Ganascia, Jean-Gabriel, « MEDITE – A Unilingual Text Aligner for Humanities. Application to Textual Genetics and to the Edition of Text Variants », *Supporting Digital Humanities (SDH 2011)*, Copenhagen, 2011.
- Ganascia, Jean-Gabriel et Jean Bourdaillet, « Alignements unilingues avec MEDITE », *Actes des Huitièmes Journées Internationales d'Analyse Statistique des Données Textuelles*, Paris, France, 2006, p. 427-437.
- Gefen, Alexandre, « Les enjeux épistémologiques des humanités numériques », *Socio. La nouvelle revue des sciences sociales*, n° 4, Éditions de la maison des sciences de l'homme, avril 2015, p. 61-74, en ligne, <<https://journals.openedition.org/socio/1296>>.
- George, Shini et Srividhya Vasudevan, « Comparison of LDA and NMF Topic Modeling Techniques for Restaurant Reviews », *Indian Journal of Natural Sciences*, vol. 10, n° 62, octobre 2020.
- Goulet, Vincent, *Introduction à la programmation en R*, 5^e éd., Canada, Bibliothèque et Archives nationales du Québec, 2016.
- Guerri, François-Xavier, « Góngora et ses premiers biographes : une analyse comparative moyennant des outils numériques », *e-Spania. Revue interdisciplinaire d'études hispaniques médiévales et modernes*, n° 29, CLEA (Civilisations et Littératures d'Espagne et d'Amérique du Moyen Âge aux Lumières), EA 4083, février 2018, en ligne, <<https://journals.openedition.org/e-spania/27502>>.
- Heyer, Mark, « The Creative Challenge of CD-ROM », dans Steve Lambert et Suzanne Ropiequet (dir.), *CD-ROM. The New Papyrus. The Current and Future State of the Art*, Redmond, WA, Microsoft Press, 1986, p. 347-357.

- Juola, Patrick, « How a Computer Program Helped Show J.K. Rowling write A Cuckoo's Calling », *Scientific American*, 20 août 2013.
- Kirschenbaum, Matthew, « What Is “Digital Humanities”, and Why Are They Saying Such Terrible Things about It? », *differences*, vol. 25, n° 1, janvier 2014, p. 46-63, en ligne, <<https://mkirschenbaum.files.wordpress.com/2014/04/dhterriblethingskirschenbaum.pdf>>.
- Labbé, Cyril et Dominique Labbé, « La distance intertextuelle et l'attribution d'auteur. Corneille et Molière », *Journal of Quantitative Linguistics*, vol. 8, n° 3, 2000, p. 213-231, en ligne, <<https://halshs.archives-ouvertes.fr/halshs-00137675>>.
- Lafont-Terranova, Jacqueline et al., « Modéliser le processus d'écriture d'un scripteur de haut niveau : intérêt et limites du repérage automatique des opérations de réécriture à l'aide du logiciel MEDITE », *L'écriture, ses supports, ses archives*, 2017.
- Michel, Jean-Baptiste et al., « Quantitative Analysis of Culture Using Millions of Digitized Books », *Science*, vol. 331, n° 6014, American Association for the Advancement of Science, janvier 2011, p. 176-182, en ligne, <<https://www.science.org/doi/10.1126/science.1199644>>.
- Middell, Gregor, « Review of “Juxta Web Service, LERA, and Variance Viewer. Web based collation tools for TEI” », *RIDE*, vol. 11, 2020, en ligne, <<https://ride.i-d-e.de/issues/issue-11/web-based-collation-tools/>>.
- Mikolov, Tomas et al., *Efficient Estimation of Word Representations in Vector Space*, arXiv:1301.3781, 2013.
- Miyagawa, So et al., « Building Linguistically and Intertextually Tagged Coptic Corpora with Open Source Tools », *Leveraging Open Data*, Tokyo, 2018, p. 139-141.
- Mohr, John W. et Petko Bogdanov, « Topic models: What they are and why they matter », *Poetics*, vol. 41, n° 6, « Topic Models and the Cultural Sciences », décembre 2013, p. 545-569, en ligne, <<https://www.sciencedirect.com/science/article/pii/S0304422X13000685>>.
- Mounier, Pierre, *Les humanités numériques Une histoire critique*, Paris, Éditions de la Maison des sciences de l'homme, 2018.
- Paatero, Pentti et Unto Tapper, « Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values », *Environmetrics*, vol. 5, n° 2, 1994, p. 111-126, en ligne, <<https://onlinelibrary.wiley.com/doi/abs/10.1002/env.3170050203>>.
- Porter, Martin F., « Snowball: A language for stemming algorithms », *Retrieved March*, vol. 1, 2001.
- Potter, Rosanne G., « Literary Criticism and Literary Computing: The Difficulties of a Synthesis », *Computers and the Humanities*, vol. 22, n° 2, Springer, 1988, p. 91-97, en ligne, <<https://www.jstor.org/stable/30200105>>.
- Puren, Marie et Simon Gabay, « Affaire Grégory : la stylométrie permettra-t-elle enfin d'identifier le corbeau ? », *The Conversation*, 2021, en ligne, <<https://theconversation.com/affaire-gregory-la-stylometrie-permettra-t-elle-enfin-didentifier-le-corbeau-153899>>.

- Reboul, Marianne, « Comparaison semi-automatique des traductions en langue française de l’Odyssée d’Homère (1547-1955) », Thèse de doctorat, 2017.
- , *OdysseusPolymetis/odysseus_interface: open-source code for publication (PhD)*, Zenodo, 2021.
- , *Comparaison semi-automatique des traductions françaises de l’Odyssée d’Homère (1547-1955)*, Paris, Classiques Garnier, coll. « Cultures et pratiques savantes du numérique », 2022.
- Rybicki, Jan, « Vive la différence : Tracing the (authorial) gender signal by multivariate analysis of word frequencies », *Digital Scholarship in the Humanities*, vol. 31, n° 4, 2016, p. 746-761.
- Schöch, Christof, « Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik », dans Lars Schneider et Christof Schöch (dir.), *Literaturwissenschaft im digitalen Medienwandel*, PhiN, coll. « Beihefte zu Philologie im Netz », 2014, p. 130-157, en ligne, <<https://hal.archives-ouvertes.fr/hal-00957091>>.
- Schreibman, Susan, Ray Siemens et John Unsworth (dir.), *A Companion to Digital Humanities*, Oxford, Blackwell, 2004.
- , (dir.), *A New Companion to Digital Humanities*, 1^{re} éd., Wiley, 2016.
- Shapira, Dana et James A. Storer, « Edit Distance with Move Operations », *Proceedings of the 13th Annual Symposium on Combinatorial Pattern Matching*, London, UK, UK, Springer-Verlag, coll. « CPM ’02 », 2002, p. 85-98.
- Sinatra, Michaël E. et Marcello Vitali-Rosati, « Chapitre 3. Histoire des humanités numériques », dans Michael E. Sinatra (dir.), *Pratiques de l’édition numérique*, Montréal, Presses de l’Université de Montréal, coll. « Parcours numérique », 2014, p. 49-60, en ligne, <<http://books.openedition.org/pum/317>>.
- Spiro, Lisa, « “This Is Why We Fight”: Defining the Values of the Digital Humanities », dans Matthew K. Gold (dir.), *Debates in the Digital Humanities*, University of Minnesota Press, 2012, p. 16-35, en ligne, <<https://www.jstor.org/stable/10.5749/j.ctttv8hq.6>>.
- Tan, Pang-Ning, Michael Steinbach et Vipin Kumar, *Introduction to data mining*, 2^e éd., Boston, Pearson Addison Wesley, 2018.
- Tomlinson, Stephen, *Lexical and Algorithmic Stemming Compared for 9 European Languages with Hummingbird SearchServer at CLEF 2003*, dans Peters C et al. (dir.), vol. 3237, s. 1., 2004, p. 286-300.
- Underwood, Ted, « Topic modeling made just simple enough. », dans *The Stone and the Shell*, 2012, en ligne, <<https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/>>.
- Vinck, Dominique, *Humanités numériques : la culture face aux nouvelles technologies*, 2^e éd. revue et augmentée, Paris, le Cavalier bleu éditions, coll. « Idées reçues », 2020.
- Wise, Michael J, « Neweyes: A System for Comparing Biological Sequences Using the Running Karp-Rabin Greedy String-Tiling Algorithm », *Unpublished Basser Department of Computer Science*, 1993.

Lexicographie ; dictionnaires et encyclopédies

- Brunel, Pierre, *Dictionnaire des mythes littéraires*, édition revue et augmentée, Monaco, Editions du Rocher, 1994.
- Baratin, Marc et Marianne Baratin-Lorenzi, *Dictionnaire des synonymes*, Paris, Hachette Éducation, 1996.
- Chardon, Laurette, « L'espace sémantique du Dictionnaire électronique des synonymes (DES) et les méthodes de regroupement de sens : l'exemple de sec », *Syntaxe & Sémantique*, vol. 21, n° 1, Presses universitaires de Caen, 2020, p. 87-126, en ligne, <<https://shs.cairn.info/revue-syntaxe-et-semantique-2020-1-page-87>>.
- Chardon, Laurette et Jacques François, « Les vedettes du Dictionnaire Électronique des Synonymes et les relations d'adjacence entre leurs synonymes », *Lexique*, n° 27, 2020, p. 21, en ligne, <<https://halshs.archives-ouvertes.fr/halshs-03192815>>.
- Darembert, Charles Victor et Edmond Saglio, « Dis pater », dans *Dictionnaire des antiquités grecques et romaines*, vol. 2, Paris, Hachette, 1892, p. 284.
- Eluerd, Roland, *Grammaire descriptive de la langue française*, Paris, Armand Collin, 2008 [2002].
- Féraud, Jean-François, *Dictionnaire critique de la langue française*, Marseille, Chez Jean Mossy Père et Fils, 1787.
- Ferrara-Léturgie, Alice, « La transformation des dictionnaires de synonymes au XX^e siècle : de la synonymie distinctive à la synonymie cumulative », *La révolution du dictionnaire*, Hermann, 2014, p. 157-174, en ligne, <<https://www.cairn.info/la-revolution-du-dictionnaire-2014--9782705686796-page-157.htm?ref=doi>>.
- Girard, Gabriel, *La justesse de la langue françoise ou Les différentes significations des mots qui passent pour synonymes*, Paris, L. d'Houry, 1718.
- , *Dictionnaire universel des synonymes de la langue française : contenant les synonymes de Girard et ceux de Beauzée, Roubaud, d'Alembert, Diderot et autres écrivains célèbres*, Paris, France, E. Penaud, 1846.
- Grevisse, Maurice, *Le Bon Usage*, Paris, Duculot-Hatier, 1980.
- Grimal, Pierre, *Dictionnaire de la mythologie grecque et romaine*, Paris, Presses Universitaires de France, coll. « Grands dictionnaires », 1999.
- Jacquin, Gérard, « L'Œuvre de Robert Estienne et les dictionnaires de la Renaissance », *Revue 303 arts recherches créations. Patrimoine linguistique*, n° 120, Nantes, Région des Pays de la Loire, Association 303, 2012, p. 28-33.
- Lambrechts, Chantal (dir.), *Grand Dictionnaire Synonymes & contraires*, Paris, Larousse, 2005.
- La Porte, Maurice de, *Les Épithètes*, Paris, Gabriel Buon, 1571.
- Le Fur, Dominique (dir.), *Dictionnaire des synonymes et nuances*, Paris, Le Robert, coll. « Collection les usuels », 2005.

Pastoureau, Michel, *Jaune : histoire d'une couleur*, Paris, Seuil, 2019.

Rich, Anthony, *Dictionnaire des Antiquités romaines et grecques*, trad. Adolphe Chéruef, Paris, Firmin Didot Frères, 1861.

Suchecka, Karolina et Nathalie Gasiglia, « Exploration fonctionnelle de dictionnaires de synonymes pour un projet d'humanités numériques », *Linx. Revue des linguistes de l'université Paris X Nanterre*, n° 86, Université Paris Ouest, 2023, en ligne, <<https://journals.openedition.org/linx/10203>>.

Tran, Mickaël et Denis Maurel, « Prolexbase. Un dictionnaire relationnel multilingue de noms propres », *TAL*, vol. 47, n° 3, 2006, p. 115-139.

Autres

Ballanche, Pierre-Simon, *Du sentiment considéré dans ses rapports avec la littérature et les arts*, Lyon, Paris, Ballanche et Barret, Calixte Volland, 1801.

—, *Antigone*, Paris, Jules Didot, 1819.

Batteux, Charles, *Les Beaux-arts réduits à un même principe*, Paris, Durand, 1746.

Blaze, François Henri Joseph, *Théâtres lyriques de Paris : l'académie impériale de musique, histoire littéraire, musicale, choréographique, pittoresque, morale, critique, facétieuse, politique et galante de ce théâtre*, Paris, Castil-Blaze, 1855.

Commelin, Pierre, *Mythologie grecque et romaine*, Paris, Garnier frères, coll. « Classiques Garnier », 1960.

Deleuze, Gilles et Félix Guattari, *Rhizome : introduction*, Paris, Éditions de Minuit, 1976.

Grandjean, Martin, « Les réseaux de la coopération intellectuelle. La Société des Nations comme actrice des échanges scientifiques et culturels dans l'entre-deux-guerres », thèse de doctorat préparée sous la direction de François Vallotton et Bertrand Müller, Université de Lausanne, 2018.

Huffer, Damien et Shawn Graham, « The Insta-Dead: The rhetoric of the human remains trade on Instagram », *Internet Archaeology*, n° 45, 2017, en ligne, <<http://intarch.ac.uk/journal/issue45/5/index.html>>.

Inqualifiables fureurs. Poétique des invocations inspirées aux XVI^e et XVII^e siècles, Paris, Classiques Garnier, coll. « Études et essais sur la Renaissance », 2019.

Jarrassé, Dominique, « Trois gouttes d'art nègre. Gobinisme et métissage en histoire de l'art », *Les actes de colloques du musée du quai Branly Jacques Chirac*, n° 1, juillet 2009, en ligne, <<http://journals.openedition.org/actesbrantly/96>>.

Louvat, Bénédicte, « Influence italienne et satire médicale dans la *Pastorale du berger Célidor et de Florimonde sa bergère* (Béziers, 1629) », *Quand Minerve passe les monts : modalités littéraires de la circulation des savoirs, Italie-France, Renaissance-XVII^e siècle*, International, Pessac, Presses Universitaires de Bordeaux, 2021, p. 45-55.

- Ovide, *Les métamorphoses*, trad. Jean-Pierre Néraudau, Paris, Gallimard, coll. « Collection Folio Classique », 2006.
- , *Les métamorphoses*, trad. Marie Cosnay, Paris, Éditions de l'Ogre, 2017.
- Van Exaerde, Fanny, « « Une poésie cinématographique. » Étude historique, génétique et intermédiaire des scénarios et dialogues de Jean Cocteau », Thèse de doctorat, Université de Lille (2022-....), 2023.
- , *Absinthe, avortement et Bacchantes dans le scénario d'Orphée (1949-1950) de Jean Cocteau*, Ritualités sociales et littéraires. Boire et écrire, Université de Lille, 7 février 2025.
- Villenave, Mathieu-Guillaume-Thérèse, « Avertissement », *Les Métamorphoses d'Ovide, traduction nouvelle, avec le texte latin, suivie d'une analyse, de l'explication des fables, de notes géographiques, historiques, mythologiques et critiques*, vol. 1, Paris, F. Gay, Ch. Guestard, 1806, p. V-XI.
- Zaugg, Roberto et Andrea Graf, « Guerres napoléoniennes, savoirs médicaux, anthropologie raciale », *Histoire, médecine et santé*, n° 10, 2016, p. 17-43, en ligne, <<https://journals.openedition.org/hms/1014>>.

Sitographie

Ressources du projet de thèse

Code source, GitHub, v2, 2025-08, <<https://github.com/karolinasucheka/orphidys/>>.

Code source, GitLab, v2, 2025-08, <<https://gitlab.univ-lille.fr/karolina.sucheka/>>.

Dépôt pérenne du code source, Software Heritage, 2025-08, <<https://archive.softwareheritage.org/swh:1:dir:398a23260f8eb1ae6aed57f46360d16a7aa15e06;origin=https://github.com/karolinasucheka/orphidys;visit=swh:1:snp:f74d267bde2e982b97a7262e7693609fe499a32b;anchor=swh:1:rev:0b1b52cb23f8044fedac662db20b5e87a208666e>>.

Corpus, Nakala, 2024-12-04, <<https://nakala.fr/collection/10.34847/nkl.ba7766ru>>.

Lexiques, Ortolang, 2025-07-23, <<https://www.ortolang.fr/market/lexicons/orphidys>>.

Vitrine pour les données, Nakala, 2025-07-21, <<https://orphidys.nakala.fr/>>.

Logiciels, scripts et interfaces

Akbik Alan, *et al.*, *Flair NLP*, Humboldt University of Berlin, v0.15.1, 2025-02-05, <<https://github.com/flairNLP/flair>>.

BLAST+, NCBI, v2.17, 2025-07-01, <<https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/>>.

Büchler, Marco, *et al.*, *TRACER*, eTRAP, 2018, <<https://doi.org/21.11101/0000-0007-C9CA-3>>.

Clérice, Thibault, *et al.*, *Pyrrha*, v4.0.1, 20-03-2024, <<https://github.com/hipster-philology/pyrrha/tree/4.0.1>>.

CollateX, Interedition Development Group, v1.7.1, 2015-12-17, <<https://collatex.net/>>

Compilatio Magister, Compilatio, 2025, <<https://www.compilatio.net/>>.

Cytoscape JS, Donnelly Centre (University of Toronto), v 3.32.0, 2025-04-30, <<https://js.cytoscape.org/>>.

Eder, Maciej, Joanna Byszek et Jan Rybicki, *Stylo*, Computational Stylistics, v0.7.5, 2024-04-04, <https://computationalstylistics.github.io/stylo_nutshell/#>.

Edition Visualization Technology, Roberto Rosselli Del Turco (dir.), Université de Pise, v. 1.3.2, 2024-10-08, <<https://github.com/evt-project/evt-builder/>>.

eScriptorium, INRIA, v0.13, 2025-04-02, <<https://escriptorium.inria.fr/>>.

Ganascia, Jean-Gabriel, *Médite*, LIP6-OBVIL, Sorbonne Université, v.3.8, 2014-05-31, <https://webia.lip6.fr/~ganascia/Medite_Project>.

—, *Phæbus* [archivé], LIP6-OBVIL, Sorbonne Université, 2015-01-22, <https://web.archive.org/web/20220121001231/>.

Gilman, Ian, *OpenSeadragon*, v5.0.1, 2024-12-09, <<https://openseadragon.github.io/>>.

- Git*, Linux Foundation, v 2.50.1, 2025-06-16, <<https://git-scm.com/>>.
- GitHub – Plateforme de développement collaboratif*, GitHub, Inc., <<https://github.com/>>.
- GitLab – Plateforme DevOps intégrée*, GitLab Inc., <<https://about.gitlab.com/>>.
- Gladstone, Clovis, *et al.*, *TextPAIR (Pairwise Alignment for Intertextual Relations)*, University of Chicago / ARTFL Project, v2.2.0.1, 2024-06-09, <<https://github.com/ARTFL-Project/text-pair>>.
- , *PhiloLogic*, University of Chicago / ARTFL Project, v4.7.5.5, 2024-05-28, <<https://github.com/ARTFL-Project/PhiloLogic4>>.
- Glorieux, Frédéric, *Debook*, LIP6-OBVIL, Sorbonne Université, 2015, <<https://obtic.humanum.fr/obvil-web/developpements/debook>>
- , *Teinte*, LIP6-OBVIL, Sorbonne Université, 2015, <<https://obvil.github.io/Teinte/teinte.html>>.
- , *Odette*, LIP6-OBVIL, Sorbonne Université, 2015, <<https://github.com/oeuvres/Odette>>.
- , *Bookmeka*, LIP6-OBVIL, 2015, <<https://github.com/oeuvres/Bookmeka>>.
- Hathout, Nabil, *Morphonette*, Université Toulouse II, v.0.1, 2010, <<http://redac.univ-tlse2.fr/lexiques/morphonette.html>>.
- Jänicke, Stefan, *TRAVIZ*, Vizcovery / eTRACES, <<http://www.traviz.vizcovery.org/index.html>>.
- Java*, Oracle, v24, 2024-09, <<https://www.oracle.com/java/technologies/downloads/>>.
- Kestemont, Mike, *Harry Potter intertextuality notebook*, <https://github.com/mikekestemont/potter/blob/master/harry/intertextuality/intertextuality.ipynb>.
- Kraken*, École Pratique des Hautes Études, Université PSL, v 5.2, 2024-04-21, <<https://kraken.re/main/index.html>>.
- Li, Yaoliang et Lincoln Mullen, *textreuse : Detect Text Reuse and Document Similarity*, rOpenSci, <<https://github.com/ropensci/textreuse>>.
- Losada, José Luis, *Tracer-scripts*, v1, 2020, <<https://github.com/editio/tracer-scripts>>.
- Namer, Fiammetta, *DeriF*, CNRTL, v3.1, 2005, <<https://www.cnrtl.fr/outils/DeriF/>>.
- NLTK Sentence Segmentation*, Text Analysis Online, 2016, <<https://textanalysisonline.com/nltk-sentence-segmentation>>.
- Omeka S*, Corporation for Digital Scholarship / Omeka Team, v 4.1.1, 2024-06-24, <<https://omeka.org/>>.
- Pandore-Toolbox*, ObTIC, Sorbonne Université, <<https://github.com/obtic-sorbonne/Toolbox-site>>.

- Qi, Peng, Yuhao Zhang, Yuhui Zhang, *et al.*, *Stanza*, Stanford NLP Group, v1.10.1, 2024-12-29, <<https://stanfordnlp.github.io/stanza/>>.
- Schmid, Helmut, *TreeTagger*, Institut für Informatik, LMU München, v3.2.5, 1995, <<https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>>.
- Schreibman, Susan, *The Versioning Machine*, v5.0, 2016-01-21, <<http://v-machine.org/>>.
- Smith, David, *Passim*, Andrew W. Mellon Foundation and the National Endowment for the Humanities, v2, 2024-02-07, <<https://github.com/dasmiq/passim>>.
- spaCy*, Explosion AI, v 3.8.7, 2025-05-23, <<https://spacy.io/>>.
- TEI Publisher*, e-editiones, v 9.1.0, 2024-10-16, <<https://teipublisher.com/index.html>>.
- Tesseract OCR*, v 5.5.1, 2025-05-25, <<https://tesseract-ocr.github.io/tessdoc/>>.
- Tesserae*, University of Notre-Dame, v5, 2018, <<https://github.com/tesserae/tesserae>>.
- Transkribus*, READ-COOP, v4, 2024-04-03, <<https://www.transkribus.org/fr>>.
- The Wayback Machine*, Internet Archive, <<https://web.archive.org/>>.
- Varga, Daniel *et al.*, *Hualign*, Hunglish Corpus, 2014, <<https://github.com/danielvarga/hu-align>>.
- Vesanto, Alexi, *Text Reuse Detection with BLAST*, 2018, <<https://github.com/avjves/textreuse-blast>>.

Corpus d’auteurs, projets de recherche et éditions numériques

- Auvray-Assayas, Clara, *Édition numérique de La Nature des dieux de Cicéron*. URL : <https://www.uni-caen.fr/puc/sources/ciceron/>.
- Coffee, Neil, *et al.*, *Tesserae Interface* [archivé], University at Buffalo, v5, 2018, [sauvegarde du 2025-05-25], <<https://web.archive.org/web/20250525083114/https://tesserae.caset.buffalo.edu/>>.
- Cordell, Ryan et David Smith, *The Viral Texts Project. Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines*, National Endowment for the Humanities, American Council of Learned Societies, Mellon Foundation, 2012-2024, <<https://viraltexts.org/>>.
- Corpus Thomisticum*, Fundación Tomás de Aquino, 2013, <<http://www.corpusthomisticum.org/wintrofr.html>>.
- eBalzac : édition génétique hypertexte*, Andrea Del Lungo, Jean-Gabriel Ganascia et Pierre Glaudes (dir.), ANR, laboratoires CELLF et LIP6, Sorbonne Université, 2015-2019, <<https://ebalzac.com/>>.
- European Literary Text Collection (ELTEC)*, Carolin Odebrecht, Lou Burnard et Christoph Schöch (dir.), COST Action *Distant Reading* (CA16204), v1.1, 2021-04, <doi.org/10.5281/zenodo.4662444>.

Fièvre, Paul, *Théâtre classique*, 2007-2015, <<http://www.theatre-classique.fr/index.html>>.

Flaubert, Yvan Leclerc et François Vanoosthuyse (dir.), laboratoires IRIHS et CÉRÉDI, Université de Rouen, 2001, <<https://flaubert.univ-rouen.fr>>.

Fry, Ben, *Ont the Origin of Species: The Preservation of Favoured Traces*, 2009, <<https://ben-fry.com/traces/>>.

Heppler, Jason, *What is Digital Humanities?*, 2015-01, <<https://whatisdigitalhumanities.com/>>.

Huguenots Online, Andrew Pettegree (dir.), Leiden et Boston : Brill, 2008, <<https://primarysources.brillonline.com/browse/huguenots>>.

Impresso – Media Monitoring of the Past II. Beyond Borders: Connecting Historical Newspapers and Radio, Maud Ehrmann, Simon Clematide, Raphaëlle Ruppen Coutaz, Marten Düring (dir.), Swiss National Science Foundation, Luxembourg National Research Found, 2023, <<https://impresso-project.ch/>>.

Inventing Abstraction 1910-1925, Museum of Modern Art, New York, 2012, <<https://www.moma.org/interactives/exhibitions/2012/inventingabstraction/?page=connections>>.

Knowledge, Information Technology, & the Arabic Book (KITAB), Sarah Bowen Savant (dir.), Aga Khan University International, The Mellon Foundation, Horizon, 2020, <<https://kitab-project.org/>>.

Morrissey, Robert, Clovis Gladstone, Mark Olsen, Charles Cooney et Jordan Pruett, *Intertextual Hub*, University of Chicago / ARTFL Project, National Endowment for the Humanities, 2025, <<https://intertextual-hub.uchicago.edu/>>.

Nouvelle édition numérique de fac-similés de référence (Nénufar), Hervé Bohbot (dir.), DGLFLF, CORLI, Praxiling, Université de Montpellier Paul-Valéry, <<https://nenufar.huma-num.fr/presentation/>>.

Reboul, Marianne, *Odysseus Polymetis*, 2017, <https://odysseuspolymetis.github.io/odysseus_interface/>.

Rosetti Archive, Jerome McGann (éd.), Institute for Advanced Technology in the Humanities, University of Virginia 1992, <<https://www.rossettiarchive.org/>>.

The Canterbury Tales Project, Peter Robinson et Barbara Bordalejo (dir.), ITSEE, University of Birmingham, 1991, <<http://canterburytalesproject.com/index.html>>.

Corpus linguistiques, bibliothèques numériques et leurs interfaces d'exploration

Azoulay, Benjamin et Benoît de Courson, *Gallicagram*, Université Paris-Saclay, 2021, <<https://shiny.ens-paris-saclay.fr/app/gallicagram>>.

N-grams, Google Books, <<https://books.google.com/ngrams>>.

Google Books, Google, 2011, <<https://books.google.fr/>>.

Frantext, Alex Boulton (dir.), CNRS, Atilf, Université de Lorraine, v25.1, 2025-01, <<https://www.frantext.fr/>>.

Gallica, BnF, 1997, <<https://gallica.bnf.fr/>>.

Europeana, Common European Data Space for Cultural Heritage, 2008, <<https://www.europeana.eu>>.

Internet Archive, Brewster Kahle, Rick Prelinger, Kathleen Burch et David Rumsey (dir.), 2005, <<https://archive.org/about/>>.

Wikisource, Wikimedia Foundation, 2012, <<https://fr.wikisource.org/wiki/Wikisource:Accueil>>.

A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown), compilé par W. N. Francis et H. Kučera, Brown University, 1964, 1971, 1979, <<https://www.sketchengine.eu/brown-corpus/>>.

British National Corpus online, BNC Consortium, Oxford University Computing Services, 200-2001, 2007, <<http://www.natcorp.ox.ac.uk/>>.

Dictionnaires, lexiques et WordNets

Antidote, Druide Inc., v12, 2024, <<https://www.antidote.info/fr>>.

Base de données lexicographiques panfrancophone, Claudre Poirier et Michèle Francard (dir.), AUF, TLFQ, 2001-2014, <<https://www.bdlp.org/>>.

Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT), ANR, laboratoires Lattice, LILPa, ICAR et IHRIM, corpus déposé sur Ortolang, v1.1, 2019, <<https://www.ortolang.fr/market/corpora/democrat/v1.1>>.

Dico en ligne, Le Robert, <<https://dictionnaire.lerobert.com/>>.

Dictionnaire de l'Académie Française, Académie française, <<https://www.dictionnaire-academie.fr/>>.

Dictionnaires d'autrefois, ARTFL Project, University of Chicago, Atilf – CNRS, Université de Lorraine <<https://portail.atilf.fr/dictionnaires/>>.

Dictionnaire électronique des Synonymes (DES), Laurette Chardon (dir.), CRISCO, Université de Caen, v4, 2024, <<https://crisco4.unicaen.fr/des/>>.

Le Robert correcteur, Le Robert, <<https://robert-correcteur.lerobert.com/aide-redaction.php>>.

Prolexbase, Mockaël Tran (dir.), Laboratoire d'Informatique Fondamentale et Appliquée, Université de Tours, v3.1, 2025-06-25, <<https://www.ortolang.fr/market/lexicons/prolex/v3.1>>.

Sagot, Benoît et Darja Fišer, *WordNet Libre du Français (WOLF)*, v1.0b4, 2008, <<http://pauillac.inria.fr/~sagot/index.html#wolf>>.

The Ancient Greek WordNet, William Michael Short, Alexander Forte, James Tauber et Silvia Luraghi (dir.), Center for Hellenic Studies, University of Exeter et University of Pavia, <<https://greekwordnet.chs.harvard.edu/>>.

The FreeDictionary, Farlex, <<https://fr.thefreedictionary.com/>>.

Trésor de la Langue Française informatisé, Atilf – CNRS, Université de Lorraine, <<http://atilf.atilf.fr/tlfi.htm>>.

Standards, entrepôts et outils de gestion des données

DCMI Metadata Terms, Dublin Core Metadata Initiative, 2008-01-14, <<https://www.dublincore.org/specifications/dublin-core/dcmi-terms/2008-01-14/>>

DMP OPIDoR, Inist-CNRS, v4.3.1, 2016-2025, <<https://dmp.opidor.fr/>>.

Lagoze, Carl, Herbert Van de Sompel, Michael Nelson et Simeon Warner, *The Open Archives Initiative Protocol for Metadata Harvesting*, v2.0, 2002-06-14, <<https://www.openarchives.org/OAI/openarchivesprotocol.html>>.

Nakala, IR* Huma-Num, <<https://nakala.fr/about>>.

Recherche Data Gouv, Isabelle Blanc (dir.), Ministère de l'Enseignement Supérieur et de la Recherche, 2022, <<https://entrepot.recherche.data.gouv.fr/dataverse/root/?q=>>.

Text Encoding Initiative, TEI Consortium, <<https://tei-c.org/>>.

Structures et institutions

Alliance of Digital Humanities Organizations, <<https://adho.org/?lang=fr>>.

CenterNet. An International Network of Digital Humanities Center, <<https://dhcenternet.org/>>.

Centre de recherche interuniversitaire sur les humanités numériques (CRIHN), Michael Sinatra (dir.), Fonds de recherche Société et culture Québec, <<https://www.crihn.org/?redirwww>>.

Consortium-HN Ariane (Analyses, Recherches, Intelligence Artificielle et Nouvelles Éditions numériques), Fatiha Idmhand, Ioana Galleron et Sabine Loudcher (coordinatrices), 2023-2027, <<https://cst-ariane.huma-num.fr>>.

Consortium-HN CORLI (CORpus, Langues et Interactions), Christophe Parisse et Céline Poudat (coordonateurs), <<https://corli.huma-num.fr/fr/accueil/>>.

Consortium Corpus d'auteurs pour les humanités : informatisation, édition, recherche (Cahier), 2011-2021, <<https://cahier.hypotheses.org/>>.

Digital Object Identifier, DOI Foundation, 2022, <<https://www.doi.org/>>.

Données de la recherche apprentissage numérique (DoRANum), Inist-CNRS, <<https://dorum.fr/plan-gestion-donnees-dmp/>>.

Humanistica, Association francophone des humanités numériques/ digitales, Marie Puren (dir.), <<https://www.humanisti.ca/>>.

Identifiants et Référentiels pour l'enseignement supérieur et la recherche (IdRef), Nicolas Morin (dir.), Abes, <<https://www.idref.fr/autorites.jsp>>.

IR* *Huma-Num*, Olivier Baude (dir.), <<https://www.huma-num.fr/quest-ce-que-l-ir-huma-num/>>.

Licence CC BY-NC-ND 4.0, Creative Commons Corporation, 2021-05-19, <<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>>.

Mellon Foundation, <<https://www.mellon.org/>>.

National Endowment for the Humanities, <<https://www.neh.gov/>>.

Ouvrir la science !, Fonds national pour la science ouverte (FNSO), Ministère de l'Enseignement supérieur et de la Recherche, 2019, <<https://www.ouvirlascience.fr/accueil/>>.

Notices et autres ressources

[Bnf] « Charte documentaire de la numérisation des collections de la Bibliothèque nationale de France », BnF 2017, <https://www.bnf.fr/sites/default/files/2019-03/charte_documentaire_numerisation_collections.pdf>.

[Bnf] « Gallica en chiffres [archivé] », *Gallica*, BnF, capture de 2022-12-28, <<https://web.archive.org/web/20221228212743/https://gallica.bnf.fr/GallicaEnChiffres>>.

[Bnf] « Numérisation des livres indisponibles [archivé] », *Gallica – BnF*, capture de 2023-03-27, <<https://web.archive.org/web/20230327143905/https://gallica.bnf.fr/edit/und/numeration-des-livres-indisponibles>>

[Bnf] *Registre des livres indisponibles en réédition électronique (ReLIRE)*, projet porté par la BnF, 2012-2025, <<https://relire.bnf.fr/accueil>>.

[Bnf] « Techniques et formats de conversion en mode texte », BnF, <<https://www.bnf.fr/fr/techniques-et-formats-de-conversion-en-mode-texte>>.

DEscription et MODélisation des Chaînes de Référence : outils pour l'Annotation de corpus (en diachronie et en langues comparées) et le Traitement automatique (DEMOCRAT), ANR, laboratoires Lattice, LilPa, ICAR, IHRIM, 2016-2020, <<https://www.lattice.cnrs.fr/democrat/motivations.html>>.

Documentation sur Prolex 3.1, <<https://repository.ortolang.fr/api/content/prolex/2/Prolexbase.pdf>>.

Franzini, Greta, Emily Franzini, Kirill Bulert, Marco Büchler et Maria Moritz, *TRACER Text Reuse Detection Machine : The user manual*, 2016, <<https://tracer.gitbook.io/-manual/>>.

« Output », *eTRAP*, <<https://www.etrapp.eu/academic-output/>>.

« Principes FAIR », *Centre pour la Communication Scientifique Directe*, <<https://www.ccsd.cnrs.fr/principes-fair/>>.

« TextPAIR (Pairwise Alignment for Intertextual Relations », <*The ARTFL Project*>, <<https://artfl-project.uchicago.edu/text-pair>>.