Édition comparative intermédiale de séries traductives : exploiter les homologies pour créer des visualisations modulables

Karolina Suchecka* — **Nathalie Gasiglia**** — **Karl Zieger*****

* *ALITHILA (EA 1061), Université de Lille, karolina.suchecka@univ-lille.fr*
** *STL (UMR 8163), Université de Lille, nathalie.gasiglia@univ-lille.fr*
*** *ALITHILA (EA 1061), Université de Lille, karl.zieger@univ-lille.fr*

RÉSUMÉ. *En examinant une série traductive du dixième livre des* Métamorphoses *d'Ovide (mythe d'Orphée et Eurydice), cet article interroge l'apport des logiciels PHŒBUS et MEDITE pour le traitement de textes proches du point de vue du contenu, mais éloignés au niveau de la structure, du lexique et de l'état de la langue. Notre objectif est de prouver que, après quelques enrichissements du corpus et en faisant coopérer ces deux outils, on peut concevoir une visualisation modulable et lisible par tous. Ceci présuppose que PHŒBUS et MEDITE traitent des corpus XML en prenant en compte des éléments ou leurs attributs. En suggérant des pistes d'évolution pour ces logiciels, nous essayons d'engager un dialogue étroit entre les informaticiens qui développent les outils et les chercheurs littéraires qui les exploitent.*

ABSTRACT. *The analysis of Ovid's* Metamorphoses *tenth book (myth of Orpheus and Eurydice) leads us to examine the treatment of texts that are closely related in terms of content, but not on the level of structure, vocabulary and state of language. Our objective is to prove that, after some enrichements to the corpus and by adjusting the tools to cooperate, it is possible to create a modular and readable visualization adapted to all readers. It assumes that these tools treat the XML structured corpus by allowing researches to focus on elements or their attributes. By suggesting evolutionary paths for both softwares, we try to engage a close dialogue between computer scientists who develop the tools and literary researchers who exploit them.*

MOTS-CLÉS : *PHŒBUS, MEDITE, XML:TEI, série traductive, homologies, alignement, visualisation spatialisée, graphe linéaire, visualisation modulable.*

KEYWORDS: *PHŒBUS, MEDITE, XML:TEI, chain of translation, homologies, alignment, spatialized visualization, linear graph, modular visualization.*

1. Introduction

Cette analyse s'inscrit dans le cadre de l'édition comparative intermédiale[1] des réécritures du mythe d'Orphée et Eurydice accompagnée d'une ressource pour le repérage intertextuel, qui a pour objectif de combiner un héritage culturel spécifique avec les technologies les plus récentes afin de permettre une présentation novatrice des œuvres et de leur structure. Dans le cadre de cet article, un échantillon du corpus a été exploité à l'aide des logiciels[2] PHŒBUS[3] et MEDITE[4] afin d'interroger leurs apports et leurs limites pour traiter des textes de contenus relativement proches mais éloignés au niveau de la structure, du lexique et de l'état de la langue. Ainsi, nous examinons une série traductive du dixième livre des *Métamorphoses* d'Ovide qui ouvre le cycle d'Orphée, en analysant un corpus de six traductions[5] : (1) en ancien français, en vers [AFv], cf. Anonyme (1315-1325), et en prose [AFp], cf. Walleys (1493) ; (2) juxtalinéaire, en vers [JXv] et en prose [JXp], cf. de Parnajon (1880) ; (3) littéraire, moderne ou contemporaine (ci-après dites "moderne"), en vers [Mv], cf. Cosnay (2006), et en prose [Mp], cf. Villenave (2003).

Nous procédons d'abord à une recherche d'homologies (de correspondances intertextuelles) avec PHŒBUS afin de visualiser, sous forme de graphes de relations[6], les proximités repérées entre les textes. Nous déterminons ainsi la nature des éléments les plus récurrents afin de détecter les correspondances entre les traductions et d'en proposer une visualisation (linéaire ou spatialisée, qui colore, place et relie les points en fonction des relations repérées). Les lecteurs peuvent alors accéder soit à l'édition d'une œuvre précise, en cliquant sur son nœud, soit à la comparaison des deux textes, en cliquant sur l'arc qui relie leurs nœuds.

Prouvant, avec PHŒBUS, que la série traductive présente bien des homologies, nous souhaitons apprécier plus textuellement les proximités et nous mobilisons pour cela MEDITE, un outil dédié à la comparaison de deux états d'un texte. Il signale les remplacements, les suppressions, les insertions et les déplacements. Son utilité pour la recherche génétique est prouvée[7], mais il nous intéresse de voir ce qu'il propose pour des textes aussi différents que ceux de notre corpus, afin d'apprécier s'il présente des problèmes de détection et de lisibilité qui mériteraient d'être surmontés.

1. Les œuvres intermédiales combinent plusieurs médias (textes, sons, images, etc.), comme le font les bandes dessinées ou les représentations de spectacles vivants.
2. L'équipe ACASA du LIP6, sous la direction de J.-G. Ganascia, les développe 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).
3. Cf. `http://obvil-dev.paris-sorbonne.fr/phoebus/`, et Boukhaled *et al.* (2015) et Ganascia *et al.* (2014).
4. Cf. `http://obvil.lip6.fr/medite/`, et Ganascia et Bourdaillet (2006), Fenoglio et Ganascia (2008) et Ganascia (2011).
5. Les citations reprennent l'orthographe des textes cités ou des traitements opérés par les outils.
6. Nous mobilisons pour cela l'outil Gephi (§ 4.2.2 à 4.3), cf. `https://gephi.org/`.
7. Cf. projet ANR Phœbus : eBalzac (`https://ebalzac.com/genetique`).

Nous voulons montrer que, avec un corpus enrichi et en faisant coopérer PHŒBUS et MEDITE[8], il est possible de concevoir une visualisation modulable et lisible adaptée à tous (des spécialistes, des élèves francophones ou non, ou des lecteurs passionnés). En effet, nous considérons que les éditions numériques comparatives doivent impérativement apporter des réponses pertinentes aux besoins et aux attentes des différents types de lecteurs et graduer la complexité de ce qui est présenté. Il s'agit non seulement d'exploiter les possibilités offertes par les outils testés, mais aussi de réfléchir aux modalités d'adaptation des contenus enrichis pour une lecture numérique, qui diffère de la lecture papier notamment par sa non-linéarité.

Notre contribution se propose ainsi de démontrer la plus-value qu'aurait l'intégration de traitements, dans PHŒBUS et MEDITE, de corpus balisés en XML:TEI P5 (le standard le plus employé dans les humanités numériques littéraires). Elle ambitionne aussi de montrer l'intérêt d'un dialogue étroit entre les informaticiens qui conçoivent et développent les outils et les chercheurs littéraires (non informaticiens) qui les utilisent. Nous essayons d'engager ce dialogue en observant certaines limites des traitements actuels et en suggérant des pistes d'évolution pour les deux logiciels.

2. Présentation du corpus

Les *Métamorphoses* d'Ovide ne sont pas la première œuvre où apparaît la figure d'Orphée. Nous le trouvons dans les *Argonautiques* d'Apollonios de Rhodes (III^e^ siècle av. J.-C.). L'histoire de son amour avec Eurydice apparaît au livre IV des *Géorgiques* de Virgile (37-30 av. J.-C.), elle est brièvement évoquée dans la narration de l'invention de l'apiculture par Aristée. Enfin, un récit détaillé du mythe figure dans les livres X et XI des *Métamorphoses* d'Ovide. Ce long poème, datant probablement du début du I^er^ siècle, regroupe quinze chants sur le thème des métamorphoses issus des mythologies grecque et romaine. L'histoire d'Orphée débute au moment de son mariage avec Eurydice et finit à la mort du poète. Dans notre corpus, nous nous arrêtons à la fin du livre X, ce qui correspond à un fragment de 85 vers[9].

Dans cet extrait, Orphée convoque Hyménée, le dieu des noces, en Thrace pour célébrer son mariage avec Eurydice, mais le flambeau devant symboliser leur union refuse de s'allumer. Ce mauvais présage pour les futurs mariés se réalise lors d'une promenade : Eurydice meurt d'une morsure de serpent. Inconsolable, Orphée descend aux Enfers pour obtenir que les dieux lui rendent sa femme. Émues par son chant,

8. Nous rejoignons, entre autres, les réflexions de Gallet *et al.* (2016) dans le cadre du projet HyperApollinaire : « Peut-être pourrons-nous 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 ».

9. L'extrait couvre les vers de « *Inde per immensum croceo uelatus amictu* » à « *Aetatis breue uer et primos carpere flores.* »

les âmes qui y subissent des peines éternelles oublient leurs châtiments et Orphée se voit accorder une dernière chance de récupérer sa bien-aimée : ils peuvent sortir des Enfers, à condition qu'Eurydice marche derrière lui et que celui-ci ne se retourne pas pour la regarder. Cependant, juste avant d'atteindre la surface, le poète regarde derrière lui pour s'assurer que sa femme le suit. La condition est ainsi brisée, et Eurydice meurt pour la deuxième fois et retourne aux Enfers. Stupéfait par cette perte, Orphée passe sept jours aux bords du Styx avant de retourner en Thrace. Très courtisé, il refuse ensuite de se lier à une autre femme, mais dirige ses passions vers de jeunes garçons.

Les six textes de notre corpus comportent des modifications d'importances variées. Dans les textes en ancien français, Eurydice meurt en fuyant le dieu champêtre Aristée[10] et l'évocation de deux amants changés en pierre est omise. L'amour homosexuel manque dans les deux traductions juxtalinéaires, qui se terminent au moment de la retraite d'Orphée sur le mont de Rhodope. Eurydice est accompagnée par les Naïades uniquement dans la traduction moderne en vers. En outre, beaucoup de reformulations, de substitutions synonymiques et de métaphorisations sont présentes et, parmi les trois traductions en vers, seule celle en ancien français est rimée.

3. Méthodologie de l'expérimentation

L'analyse simultanée de tous les textes du corpus est compliquée par les graphies des deux textes en ancien français, qui diffèrent des textes postérieurs, mais aussi entre elles : AFv est un peu plus ancien et porte des traces du dialecte normand, alors que la langue d'AFp est plus centralisée.

Au début de notre travail, nous faisons divers choix de transcription[11]. Nous gardons les graphies originales sans moderniser les textes, mais nous introduisons la ponctuation moderne et les majuscules. Les accents ne sont pas restitués, mais les abréviations sont développées. Un prétraitement des textes est effectué ensuite pour moderniser les graphies anciennes et pour lemmatiser chaque mot-occurrence.

Selon les usages actuels de l'édition numérique savante, nous avons initialement structuré notre corpus selon le standard XML:TEI P5, mais les logiciels utilisés imposent de ne pas conserver le balisage. Une première transformation XSL nous a donc permis de restituer les textes avec leurs graphies originales (nous parlons alors improprement de "textes bruts"), une deuxième nous a permis de remplacer les mots par les lemmes des graphies éventuellement modernisées. Enfin, comme la versification peut poser des problèmes d'alignement du fait des sauts de ligne et que nous considérons que sa préservation n'est pas indispensable pour comparer les contenus textuels, nous uniformisons la présentation de tous les textes en mettant une phrase par ligne.

10. Cet épisode est effectivement présent dans le mythe, mais chez Virgile. Pour venger Eurydice, les nymphes font perdre à Aristée toutes ses abeilles. Il ne les récupère que par des sacrifices expiatoires. Cf. Virgile, *Géorgiques*, l. IV, v. 315-558.

11. Nous remercions Matthieu Marchal pour son aide à la transcription des textes médiévaux. Pour les règles générales concernant l'édition des manuscrits, cf. Lepage (2001).

Nous traitons ensuite chaque couple de textes bruts puis lemmatisés. Par cette démarche, nous espérons démontrer la complémentarité du traitement des textes bruts et lemmatisés, le besoin de l'automatisation des traitements multitextes et, à terme, la possibilité de rendre PHŒBUS et MEDITE plus complémentaires et capables de traiter des corpus structurés en XML:TEI (textes et balisages inclus).

Dans la suite de cet article, nous présentons les résultats de notre expérimentation avec PHŒBUS et avec MEDITE, en détaillant leurs fonctionnalités. Notre choix de ces deux logiciels en particulier est motivé par le fait que ce sont des outils mis à disposition des chercheurs littéraires et exploitables par ces derniers, notamment grâce à l'interface graphique. Nous sommes pleinement conscients de l'existence d'outils et de techniques plus récents et, sans doute, plus performants, notamment en ce qui concerne l'alignement de traductions[12]. Toutefois, ils demandent souvent des compétences techniques et informatiques qui les rendent difficilement exploitables par les chercheurs littéraires non informaticiens[13].

4. Établissement des graphes de relations avec le logiciel PHŒBUS

Le logiciel PHŒBUS est conçu pour détecter des réutilisations textuelles (des plagiats aux reformulations), mais nous montrons ici que son exploitation s'avère très fructueuse aussi pour le traitement de séries traductives. Il peut en effet apporter un plus, par rapport à d'autres logiciels d'alignement, en permettant de déterminer la nature des éléments les plus récurrents, et donc de détecter des correspondances[14].

4.1. *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 (*stop-words*), comme les articles, les auxiliaires ou les prépositions, et procède à la racinisation (*stemming*) des mots conservés, à l'aide de l'algorithme Snowball (Porter, 2001 ; Tomlinson, 2004), afin de ne garder que les racines des mots retenus (cf. note 3) ce qui fait, par exemple,

12. L'exploitation d'outils d'alignement des traductions pour traiter des séries traductives a toutefois été critiquée par exemple par Barzilay et McKeown (2001) : « [...] *parallel corpus is far from the 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* ».

13. Nous renvoyons notamment aux travaux appuyés sur deux logiciels de détection des réutilisations textuelles, TextPAIR (ARTFL Project, Université de Chicago), cf. Horton *et al.* (2010) et Abdul-Rahman *et al.* (2016), et Tracer (Marco Büchler, Georg-August-Universität de Göttingen), cf. Büchler *et al.* (2012) et Franzini *et al.* (2014), et aux études de Ho (2011) et Reboul (2017).

14. Concernant l'intertextualité et les outils numériques, cf. par exemple Coffee *et al.* (2012), Forstall *et al.* (2014) et Ferrero et Simac-Lejeune (2015).

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. Nous utilisons les paramètres par défaut : le nombre maximal de "trous" autorisés entre deux occurrences et le nombre minimal de mots communs sont fixés à trois (sauf pour les textes en ancien français où nous admettons deux mots communs), et l'ordre des mots n'est pas pris en compte. Ainsi, les reprises et les réécritures moins directes ont les meilleures chances d'être repérées, puisque le logiciel détectera les correspondances même si les temps verbaux ou l'ordre des mots sont différents.

Dans le produit du traitement d'AFp et de Mv (fig. 1), 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 AFp vs *peine* dans Mv), *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_end_index>) et de mots (de <text_1_word_start_index> à <text_2_word_end_index>). Enfin, les deux segments de textes mis en correspondance sont reproduits au sein des éléments <text_1_reuse> et <text_2_reuse>.

```
<phoebus file_1_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>1</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_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 rendit 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 1. *Extrait du produit (en XML) du traitement par PHŒBUS d'AFp et de Mv*

4.2. *Traitement du corpus brut*

Nous avons d'abord manipulé notre corpus sans le prétraiter (ni modernisation ni lemmatisation). Pour AFv, onze correspondances ont été trouvées avec AFp et une

avec un seul des textes postérieurs, JXv (« doloit pour sa double mort » / « stupéfait de la double mort »). Pour AFp, celles avec les textes juxtalinéaires et modernes sont un peu plus nombreuses, nous en recensons huit : une avec JXp, une avec JXv, deux avec Mv et quatre avec Mp. La comparaison des résultats montre la récurrence d'une correspondance, présentée dans le tableau 1, entre le texte de AFp[15], à gauche, et les autres textes, à droite. Chaque cooccurrence est délimitée par des crochets et suivie par l'identifiant du texte lié en indice. Les stemmes cooccurrents sont soulignés et suivis par les identifiants des textes liés en exposant.

[mains^{AFv} qui retenir la cuid^{AFv}a. Mais riens ne print fors^{AFv} vent^{AFv} et ainsi se part^{AFv}it Erudice de son amy et mourut de [[[second$^{AFv\ JXp\ Mp}$e mort^{AFv}. Mais de lui ne se peut plaindre$^{JXv\ JXp\ Mv\ Mp}$,]$_{Mp}$ fors de trop aim$^{JXv\ JXp\ Mv}$er. Le derni^{JXv}er$^{JXp\ Mv}$ salut^{AFv}] $_{JXp}$ luy rendi^{AFv}t que a peine$^{JXv\ Mv}$s l']$_{JXv\ Mv}$ entendi^{AFv}t Orpheus. Forment se plaignoit de la seconde mort de s'amie et voulut retourner pour trouver la morte mais la porte^{AFv} trouva^{AFv} ferm^{AFv}ee. Et le portier^{AFv} qui la gardoit^{AFv} lu]$_{AFv}$	**AFv :** « 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 entendi. Orpheüs forment se doloit pour sa double mort et voloit retorner pour querre la morte, mes il trouva fermé la porte et le portier qui le gardoit, » **JXv :** « 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 » **JXp :** « 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 ») **Mv :** « 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 »[16] **Mp :** « Eurydice meurt une seconde fois, mais sans se plaindre ; »[17]

Tableau 1. *Correspondance entre AFp et les autres textes*

La correspondance trouvée entre AFp et AFv 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 (JXp, Mp), mais c'est le seul, alors que nous en demandions deux pour qu'une correspondance soit établie.

15. « Orpheus tendit ses mains 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 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. »
16. L'autre occurrence 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 ».
17. Les trois autres occurrences 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 femenine » / « fuyait les femmes et l'amour ».

Au contraire, les autres correspondances repérées regroupent un nombre restreint de stemmes assez récurrents, qui peut être réduit à quatre lemmes au total : *plaindre, aimer, dernier* et *peine*.

Notons trois difficultés. 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*). Nous examinerons l'amélioration des résultats après la modernisation des textes et la lemmatisation (§ 4.3). Par ailleurs, les emplois de synonymes (*blasmer* dans AFv et *plaindre* dans les autres textes) bloquent les repérages de correspondances.

4.2.1. *Analyse des correspondances détectées*

Malgré ces difficultés, les correspondances croisées (où le même fragment d'un texte correspond, au moins partiellement, à d'autres entités du corpus) ont permis 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 (tab. 2) recense les extraits trouvés les plus complets, leurs @id et les nombres totaux de correspondances trouvées. Les nœuds étant identiques ou plus longs que les segments mis en correspondance dans le tableau 1, nous ne reproduisons pas les extraits qui y sont déjà cités.

@ID	Extraits	Nombres
AFv9	[tab. 1]	17
AFp10	[tab. 1]	30
JXp14	dissipe. Déjà elle [tab. 1] 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
JXv12	retirent. Et déjà mourant pour la seconde fois, elle se plaignit en quoi que ce soit de son époux : de [tab. 1] 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
Mv11	Mourant une deuxième fois, de son époux elle ne se [tab. 1] et elle roule au lieu où elle était avant. Devant la mort double de sa femme, Orphée resta immobile	24
Mp18	[tab. 1]	7
Mp19	crime de l'avoir trop aimée ! Adieu, lui dit-elle d'une voix faible qui fut à peine entendue	6

Tableau 2. *Table des nœuds*

Les nœuds du tableau 2 correspondent à la totalité des résultats trouvés pour l'épisode de la deuxième mort d'Eurydice dans tous les textes. Remarquons que le nœud AFv9, pour lequel la correspondance de quinze patterns a été trouvée avec AFp10, regroupe également « doloit pour sa double mort », la seule correspondance de AFv avec le texte en français moderne JXv (JXv12, deux patterns) : la taille de nœud est donc la

somme des patterns pour ces deux correspondances (dix-sept). Les correspondances de la plus petite taille ont été trouvées pour Mp avec deux extraits pour un total de vingt-six 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 se plaindre ;]$_{\text{Mp18}}$ 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]$_{\text{Mp19}}$; et elle rentre dans les abîmes.

La métaphorisation et la synonymisation, présentes dans cet extrait, empêchent le repérage de correspondances plus étendues. Les patterns repérés sont les mêmes que dans les autres textes (*mourir, seconde, plaindre, fois, mais* pour Mp18 et *adieu, aimée, peine* pour Mp19), 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 AFp et Mv (« 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, Mv11, pour lequel vingt-quatre patterns ont été trouvés au total. Ces relations spécifiques sont décrites dans la table des liens (tab. 3), où sont renseignés les nœuds sources et cibles, les lemmes des mots[18] qui ont permis d'identifier les correspondances et leurs nombres.

Nœuds sources	**Nœuds cibles**	**Lemmes des mots qui ont permis d'identifier les correspondances**	**Nombres**
AFv9	AFp10	mains, cuider, fors, vent, second, mort, plaindre, salut, rendre, entendre, porte, trouver, fermer, portier, garder	15
AFp10	Mp18	second, mais, plaindre	3
JXp14	Mp18	mourir, seconde, fois, plaindre	4
JXv12	Mv11	époux, quoi, plaindre, dernier, adieu, peine, lieu, Orphée, rester	9
JXv12	JXp14	déjà, seconde, époux, quoi, effet, plaindre, sinon, aimée, dernier, adieu, peine, oreille, replonger, nouveau, même, Orphée, mort, épouse	18
AFv9	JXv12	double, mort	2
AFp10	JXp14	seconde, plaindre, aimer, dernier	4
Mv11	Mp19	aimer, adieu, peine	3
AFp10	JXv12	plaindre, peine, dernier, aimer	4
AFp10	Mv11	plaindre, peine, dernier, aimer	4
JXp14	Mv11	époux, plaindre, peine, dernier, aimer, quoi, adieu, être	8
JXv12	Mp19	adieu, aimer, peine	3

Tableau 3. *Table des liens*

18. Nous reconstituons les stemmes après la racinisation et les lemmatisons afin d'améliorer la lisibilité des résultats.

Nous pouvons constater que quatre correspondances sont recensées pour le nœud Mv11 : avec JXv12 (neuf mots communs), Mp19 (trois), AFp10 (quatre) et JXp14 (huit). Plus généralement, pour les sept nœuds présentés dans le tableau 2, nous trouvons douze liens au total, dont le plus grand compte dix-huit stemmes (JXv12 / JXp14). Parmi les trente-huit lemmes recensés, nous relevons six occurrences de *plaindre* et d'*aimer*, cinq de *dernier*, *second* et *peine* (qui fait partie de la locution adverbiale *à peine*), quatre de *adieu* (contre une seulement de *salut*), trois de *mort*, etc. L'entité nommée *Orphée* n'a été détectée que pour deux comparaisons (JXv12 / Mv11 et JXv12 / JXp14). Simultanément, elle est absente de la correspondance JXp14 / Mv11, de taille assez importante (huit patterns), pour laquelle la correspondance a été établie sur *époux*. Nous nous pencherons sur la question des entités nommées pour l'alignement et la comparaison de notre corpus au § 6.

4.2.2. *Visualisation spatialisée*

Les tableaux 2 et 3, fournis à un logiciel de visualisation de graphes comme Gephi, permettent de générer différents schémas. Nous proposons d'abord une représentation spatialisée, particulièrement utile pour des analyses détaillées de correspondances croisées. Elle montre notamment les groupes des nœuds les plus larges, que nous plaçons aux périphéries du graphe (fig. 2) afin d'optimiser la visualisation et de faciliter l'analyse détaillée de chaque groupe. Une couleur spécifique est attribuée à chaque texte afin de permettre un aperçu général de la structure des correspondances détectées. Beaucoup de relations ont été repérées uniquement entre deux textes : il s'agit majoritairement des correspondances AFv (bleu) / AFp (vert), mais nous en recensons également deux Mv (rose) / JXp (orange), une JXv (noir) / Mp (violet) et une AFv / JXv. L'épaisseur des liens est, quant à elle, dépendante du nombre de patterns impliqués dans la mise en correspondance. Six graphes ont une taille supérieure à quatre nœuds, mais celui qui concerne l'épisode de la deuxième mort d'Eurydice (fig. 3) est le seul qui regroupe tous les textes du corpus. Chaque nœud établit entre deux (Mp18, Mp19, AFv9) et cinq (AFp10) liens. Curieusement, pour ceux de Mp, les deux nœuds présents ne correspondent pas aux mêmes textes : Mp18 établit des liens avec AFp et JXp, et Mp19 avec JXv et Mv. AFv9 est lié uniquement à JXv et AFp, mais ce dernier lien est d'une taille très importante (quinze patterns communs).

Le graphe le plus large compte onze nœuds : il s'agit du chant d'Orphée implorant les dieux des Enfers de lui rendre Eurydice (fig. 4). Parmi ces nœuds, quatre appartiennent à Mp, trois à JXp et deux à JXv et Mv, aucune correspondance n'a donc permis d'alignement avec les textes en ancien français. Quatorze liens au total sont établis entre les nœuds : moins que pour le graphe de l'épilogue (fig. 5), qui en compte quinze (mais seulement neuf nœuds) et qui intègre également deux extraits de AFp. Le troisième graphe le plus large correspond à la réaction des âmes résidant aux Enfers au chant d'Orphée (fig. 6), il compte neuf nœuds et quatorze liens.

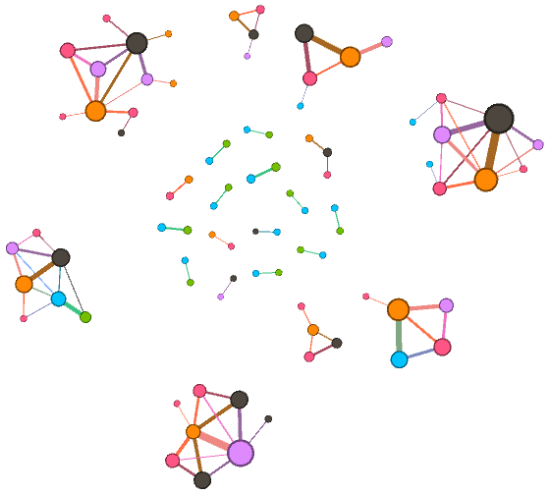

Figure 2. *Visualisation spatialisée de la totalité des correspondances (AFv : vert ; AFp : bleu ; JXv : noir ; JXp : orange ; Mv : violet ; Mp : rose)*

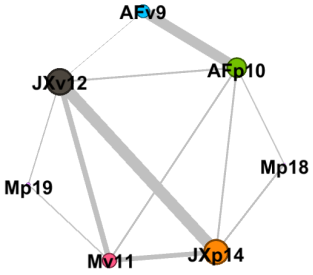

Figure 3. *Graphe de l'épisode de la deuxième mort d'Eurydice*

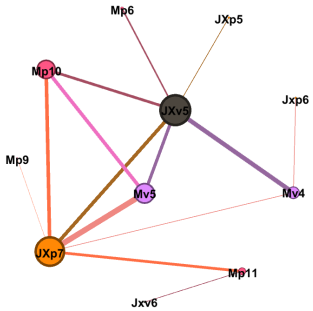

Figure 4. *Graphe du chant d'Orphée*

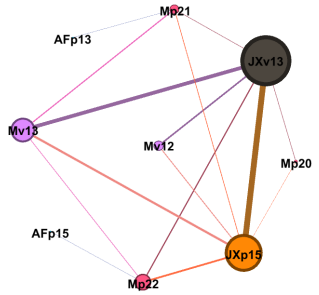

Figure 5. *Graphe de l'épilogue*

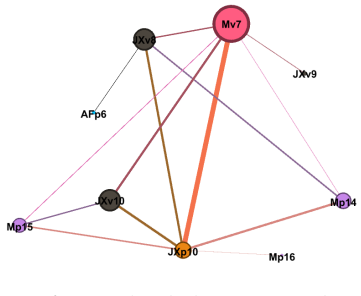

Figure 6. *Graphe de la réaction des âmes des Enfers au chant d'Orphée*

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. Une édition numérique proposant ce type de visualisation permet de restreindre les recherches aux lemmes spécifiques ou à la taille des graphes et de se focaliser sur un texte précis. En pointant le curseur sur un nœud, le texte correspondant s'affiche dans une infobulle et, en pointant sur le lien, on accède à la liste des correspondances. Un clic sur un nœud ou sur un lien permet de visualiser la comparaison détaillée entre deux correspondances, ce que nous présentons au § 7.

4.2.3. *Visualisation linéaire*

Une autre visualisation, linéaire, nous semble plus adaptée à l'analyse généralisée de l'évolution des correspondances au fil du texte. Sa lisibilité (fig. 7) n'est pas aussi optimale que celle de la visualisation spatialisée, notamment là où les relations deviennent complexes. Cependant, elle offre une vue globale des relations entre les textes : le lecteur peut constater plus aisément que très peu de relations ont été détectées entre les textes en ancien français et les textes modernes ou que l'alignement de toute une partie du chant d'Orphée n'a pas été fait entre AFv et AFp (repère `[B]`).

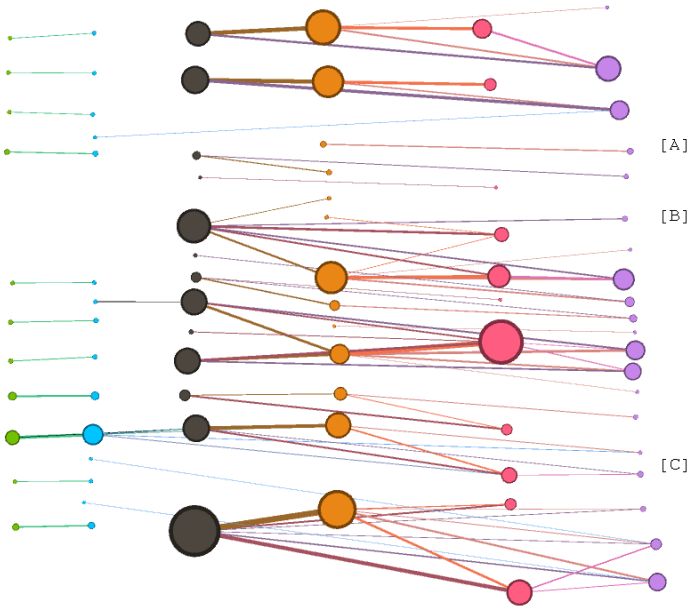

Figure 7. *Graphe linéaire*

Très peu de correspondances ont également été détectées entre le mariage d'Orphée et Eurydice et le début du chant (`[A]`). Si elles sont là, elles concernent plutôt des couples de textes, une relation plus complexe n'est pas observable (comme elle l'est

par exemple au niveau du [C]). Notons également qu'il y a extrêmement peu de problèmes d'alignement : les résultats proposés par PHŒBUS renvoient bien aux mêmes épisodes.

Nous reviendrons sur l'analyse de ces visualisations après avoir vu ce que fournit le logiciel MEDITE (§ 5), mais avant, confrontons les résultats de PHŒBUS pour le corpus brut avec ceux obtenus en traitant le corpus modernisé, pour les textes en ancien français, et lemmatisé.

4.3. *Traitement du corpus modernisé et lemmatisé*

Sans nous attarder sur la modernisation, notons que la lemmatisation des mots[19] a été effectuée avec TreeTagger[20] et revue manuellement. PHŒBUS a ensuite opéré avec les mêmes paramètres que précédemment (fig. 1). Comme attendu, les repérages des correspondances se sont nettement améliorés pour les textes en ancien français, tant en ce qui concerne le nombre des nœuds, que leur taille et le nombre des liens avec les textes plus modernes. Sur quatorze nœuds détectés pour AFv, seulement cinq liens l'ont été uniquement avec AFp, sans aucune possibilité d'alignement avec les textes postérieurs. Pour AFp (dix nœuds), une relation binaire a été détectée en plus avec Mp (« moult de lui se plaindre. Orphée être celui qui premier apprendre celui » / « soupirer ; tout se plaindre de son refus. mais ce être lui qui, par son exemple, apprendre au » [lem.]). Cette seule occurrence est d'ailleurs très significative, puisqu'elle ne concerne pas uniquement les questions de réécriture, mais également celles de l'adaptation du texte : il s'agit de la mention finale de l'amour homosexuel, absente des traductions juxtalinéaires. Mv l'aborde de manière très euphémique (« Chez les peuples thraces, il fut l'auteur de ceci : transférer l'amour sur les tendres garçons et cueillir l'avant de la jeunesse, le printemps bref, les premières fleurs… »). Dans AFv, nous observons une formulation plus descriptive (« Ce fu cil qui premierement aprist ceulz de Trace a retraire d'amour femeline et a faire des joennes malles lor deduit, dont or sont cil de Trace tuit. »).

En appliquant exactement la même démarche que dans le traitement du texte brut pour la constitution des nœuds généraux de la totalité des correspondances trouvées, nous arrivons à un nombre de nœuds très limité, mais de taille extrêmement importante. Quarante-trois nœuds au total ont été constitués (quatre-vingt-cinq pour le corpus brut) : quatorze pour AFv, dix pour AFp, trois pour JXv et JXp, deux seulement pour Mv et dix pour Mp. Quasiment la totalité des textes est recensée et le nœud le plus grand (JXv2) compte 318 patterns communs (249 pour JXp3 et 245 pour Mv2). La visualisation spatialisée, contrairement à celle du texte brut, permet de constater que

19. « ils appellent Eurydice. Elle se tenait parmi les ombres nouvellement arrivées » → « il appeler Eurydice. elle se tenir parmi le ombre nouvellement arriver » [lem.]. Par « [lem.] », nous signalons que le texte cité a été lemmatisé.

20. Cf. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

nous obtenons uniquement une grande galaxie de relations extrêmement complexes (fig. 8).

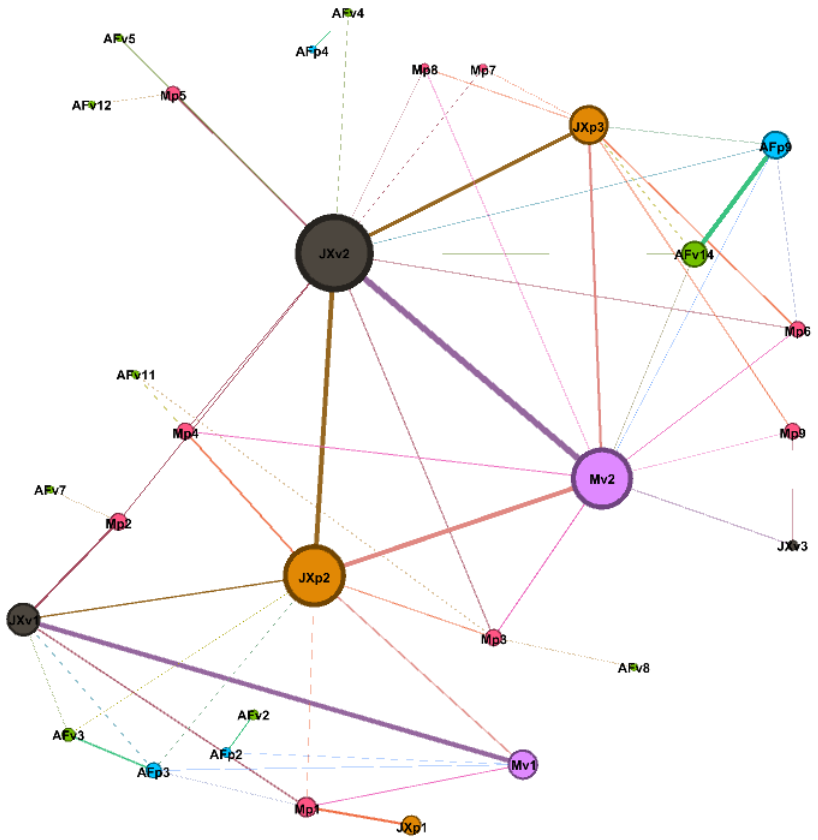

Figure 8. *Visualisation spatialisée du corpus modernisé et lemmatisé*

Cela prouve que le traitement des textes lemmatisés améliore la détection des correspondances et pas seulement pour les textes en ancien français. Toutefois, avec des nœuds devenus parfois très larges, le traitement détaillé des correspondances est moins lisible : nous voyons que JXv2 est lié avec JXp2, ce qui est très pertinent, mais il l'est aussi avec JXp3, ce qui l'est moins. Le moment exact où la correspondance a été trouvée et le nombre d'extraits détectés ne sont pas visibles, puisque la taille du nœud correspond à une autre relation très importante qui a été détectée entre JXv2 et Mv2 et qui regroupe plus des deux tiers du texte (du commencement du chant d'Orphée jusqu'à la seconde mort d'Eurydice). L'analyse semble être plus aisée pour les textes plus fragmentés, comme AFv, AFp et Mp. Simultanément, le nombre de résultats qui ne correspondent pas aux mêmes fragments du récit, qui sont donc de fausses correspon-

dances, augmente de manière significative, surtout pour les petites correspondances de trois patterns. Ainsi, pour la comparaison AFv / JXp, seulement trois résultats sur six ont été pertinents pour notre étude[21].

Au contraire, la visualisation linéaire (fig. 9) semble gagner en lisibilité et fournir davantage d'informations que lors du traitement du texte brut.

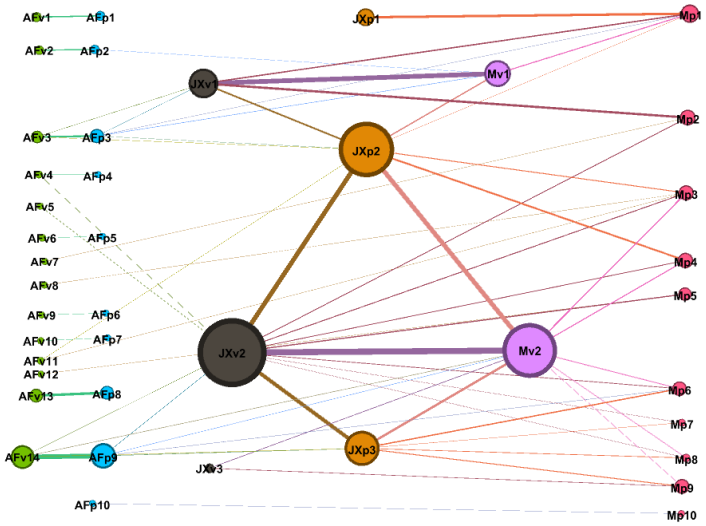

Figure 9. *Visualisation linéaire du corpus modernisé et lemmatisé*

Si des relations binaires restent présentes entre AFv et AFp, elles deviennent beaucoup plus significatives. Ainsi, nous constatons que le texte correspondant à AFv1 et AFp1 est absent des autres traductions, puisqu'il correspond au résumé de la métamorphose précédente (« dessus avoir oïr le fable comment Yphis fille fils devenir et prendre femme. » / « dessus avoir oïr comment Yphis devenir un beau jouvenceau et comment il épouser Yenta son ami. » [lem.]). AFv n'est plus presque uniquement en relation avec sa version en prose, mais il établit également des liens assez fréquents avec les textes modernes, notamment Mp, qui permet ensuite son alignement avec les textes plus éloignés (principalement Mv, pour lequel deux correspondances ont été trouvées, et dont une seulement a été retenue). Mv est le texte qui compte le moins de nœuds : alors que nous y trouvons la quasi-totalité du texte, la taille de Mv2 (245) et le nombre des liens qu'il établit avec d'autres nœuds (onze) sont inférieurs notamment à ceux de JXv2, qui regroupe 318 patterns repartis entre douze liens.

21. Aucune correspondance n'a été trouvée en comparant les textes bruts.

4.4. *Conclusions relatives à l'emploi de PHŒBUS*

Malgré la croissance de la fréquence des détections erronées, les différences de résultats du traitement de notre corpus, soit brut racinisé soit lemmatisé, prouvent que la possibilité de moduler le type de données prises en compte (tokens racinisés ou lemmes de formes éventuellement antérieurement modernisées) s'avère utile pour les utilisateurs, tant pour la visualisation que pour l'analyse des correspondances, puisque ce qui est révélé est complémentaire. Un traitement prenant en compte les différents types de données pourrait aboutir à la création de graphes multidimensionnels qui permettraient, par exemple, de basculer d'une vue générale (traitement du texte lemmatisé) à une vue plus détaillée (traitement du texte brut). Notons aussi que ceci aurait une validité comparable pour la détection de liens non avérés entre des textes sans relation évidente, ce qui est la destination première de PHŒBUS.

Cette observation nous conduit à considérer que, si nous posons qu'il faudrait pouvoir traiter plusieurs états des textes alternatifs, il nous faut les faire coexister dans le corpus manipulé plutôt que de décliner le corpus, comme nous l'avons fait ici, en une version brute et une version modernisée et lemmatisée. Cette coexistence est usuelle dans les corpus XML:TEI au sein desquels les mots sont balisés (<w>) et porteurs d'attributs pour les graphies modernisées (@ana) et les lemmes (@lemma).

Ce traitement alternatif pourrait également permettre de lier la détection des homologies avec la comparaison des versions en rendant les logiciels PHŒBUS et MEDITE complémentaires et en ouvrant de nouvelles possibilités d'affichage pour l'édition numérique savante. Mais, pour ce faire, voyons ce que nous permet le logiciel MEDITE.

5. Comparaison modulable avec le logiciel MEDITE

MEDITE est un outil de comparaison des versions d'une œuvre qui puise entre autres dans l'algorithme d'alignement par fragments grâce à la détection des homologies, une méthode de détection des séquences (utilisée initialement pour l'alignement de macromolécules – ADN ou protéines) afin de faire ressortir leurs régions homologues (cf. note 4). Actuellement, il propose une comparaison de deux textes bruts (cf. *supra*) : 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. Les possibilités de modulation des paramètres sont un peu plus larges que dans PHŒBUS : 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 cinquante pour cent (« abcd » est 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 cinquante pour cent. Ici, nous modifions uniquement le ratio des rempla-

cements à un pour cent, 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.

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[22].

5.1. *Traitement des textes dits bruts*

Un exemple des résultats de la comparaison pour deux textes détectés par PHŒBUS comme étant très proches, JXv et Mv, permet de constater que, actuellement, la complémentarité des résultats de PHŒBUS et de MEDITE (fig. 10) existe, mais qu'elle n'est pas pleinement perceptible :

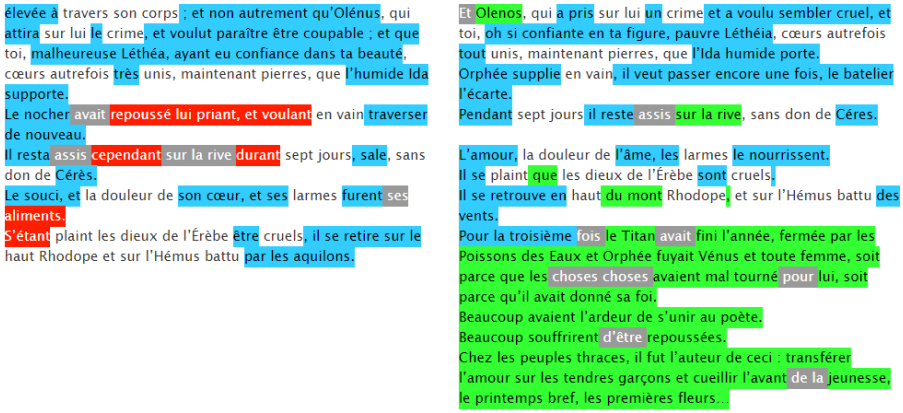

Figure 10. *Comparaison en texte brut de JXv et Mv*

La plupart des variantes détectées ont été analysées comme étant des remplacements. Parmi les quatre suppressions, une seulement (« S'étant ») n'est pas la suite d'un remplacement interrompu par la détection d'un déplacement. Les insertions sont plus nombreuses pour Mv, notamment à cause de l'absence de l'épisode de l'amour

22. Outre les projets de génétique littéraire, comme l'ANR Phœbus : eBalzac ou HyperApollinaire, cf. notes 7 et 8, MEDITE est 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), Lafont-Terranova *et al.* (2017). Guerry (2018), lui, diversifie ainsi les attendus : à partir des résultats de 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 ».

homosexuel dans les traductions juxtalinéaires. Toutefois, certains déplacements détectés ailleurs dans JXv viennent interrompre cette insertion, ce qui produit même une erreur de reconstitution du texte en redoublant le mot « choses » (« que les choses choses avaient mal tourné »). Nous n'allons pas nous attarder sur la question des déplacements, dont la détection peut être modulable.

Pour l'extrait présenté à la figure 10, l'alignement des blocs communs ne pose pas de problèmes majeurs (même si l'on peut noter le bloc commun « en vain », qui ne correspond pas au même contexte, « voulant en vain traverser de nouveau » / « supplie en vain »). Il peut y avoir plus de bruits (fausses détections) ou de silences (absences de détections), notamment pour des extraits très différents (fig. 11) :

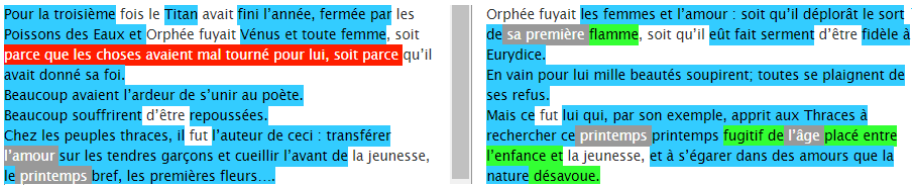

Figure 11. *Comparaison en texte brut Mv / Mp*

Dans la comparaison de Mv et Mp, le bloc commun « d'être » ne correspond pas au même contexte (« il avait donné sa foi » / « il eût fait serment d'être fidèle à Eurydice » et « Beaucoup souffrirent d'être repoussées » / « toutes se plaignent de ses refus »). La répétition de la conjonction « soit » n'a pas été repérée, alors qu'elle permettrait de faire correspondre « parce que les choses avaient mal tourné pour lui » et « qu'il déplorât le sort de sa première flamme ». Enfin, l'accumulation de signalements de remplacements rend les résultats de la comparaison trop généraux.

5.2. *Traitement du corpus modernisé et lemmatisé*

Si MEDITE était capable de prendre en compte des formes lemmatisées, cela permettrait, au choix, de ne pas repérer les différences de flexion ou de conjugaison, ou de n'en sélectionner que certaines (fig. 12) :

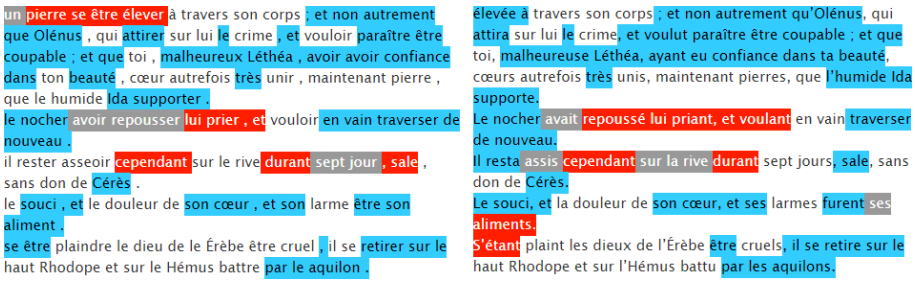

Figure 12. *Résultats pour JXv en format lemmatisé (à gauche) et brut (à droite)*

Pour cet extrait, cinq occurrences de ce type de variantes sont observables : deux occurrences de « vouloir » (« voulut » / « a voulu » et « voulant » / « veut »), une de « rester » (« resta » / « reste »), une de « repousser » (« repoussé » / « repoussées ») et une de « être » (« être » / « sont »). Parmi elles, la première, la troisième et la quatrième pourraient être détectées sans lemmatisation puisque MEDITE peut faire une recherche sur des sous-chaînes de caractères incluses dans les mots plutôt que sur les mots entiers. La deuxième permet d'éviter le seul problème d'alignement pour cet extrait (« en vain »), en séparant les deux parties de la phrase (« le nocher avait repoussé lui priant, et voulant » et « traverser de nouveau »). La cinquième permet d'augmenter la taille des blocs communs repérés (« plaindre le dieu de le Érèbe être cruel » [lem.]). En revanche, « repousser » engendre l'apparition d'un déplacement qui n'est pas avéré (« Le nocher avoir repousser » / « beaucoup souffrir de être repousser » [lem.]).

Pour certaines recherches, les formes actualisées restent significatives et méritent d'être conservées, tandis que, pour d'autres, effectuer la comparaison sur les lemmes serait préférable. La prise en compte des valeurs d'attributs @lemma dans la structure XML permettrait de moduler aisément cet affichage, en traitant soit tous les lemmes soit une partie d'entre eux. En effet, la prise en compte des codes catégoriels et flexionnels (@pos et @msd) rendrait possible une spécification encore plus poussée (comme « remplacement flexionnel passé simple / passé composé »). Enfin, même si la comparaison était effectuée sur les lemmes, leur enregistrement en tant qu'attributs permettrait d'afficher le texte original afin d'améliorer la lisibilité des résultats.

6. Extension des traitements aux entités nommées et autres objets textuels balisables

Pour des textes comparables, mais très différents en surface, comme ceux de notre corpus, il importe de traiter le contenu mais aussi d'autres indices qui peuvent être fournis par un balisage XML idoine et semi-automatisable grâce à des outils de TAL. Pris en compte par PHŒBUS et MEDITE pour leurs traitements, les indices inclus dans le XML permettraient d'intégrer dans la comparaison les entités nommées et leurs périphrases, mais aussi les interprétations du récit source qui affectent tant le contenu que les protagonistes invoqués (comme l'absence de la mention de l'amour homosexuel), la synonymie (« elle fut replongée de nouveau au même lieu » / « elle roule au lieu où elle était avant »), l'hyperonymie (« il fut présent à la vérité » / « il est là »), la métaphorisation (« elle ne put s'animer bien qu'il l'agite » / « le dieu qui l'agite ne peut ranimer ses mourantes clartés »), etc.

Faute de pouvoir tout exposer, observons concrètement un cas d'entité nommée (<persName>) avec deux désignations d'Hyménée (« Hymen » et « diex de noçoiement ») pour lesquelles nous proposons (fig. 13) un balisage XML:TEI qui fournit, en tant qu'attributs, l'identifiant de l'entité nommée (@corresp) et, pour les mots (<w>), les codes grammaticaux (@pos) et flexionnels (@msd), les formes lemmatisées (@lemma), les modernisations (@ana), les champs lexicaux (@corresp) et les synonymes (@sameAs) :

```
<persName corresp="#Hyménée">
  <w pos="NAM" msd="sing" lemma="Hymen" ana="Hymenée" corresp="#déité">Hymen</w>
</persName>
<persName corresp="#Hyménée">
  <w pos="NOM" msd="sing−CS" lemma="dieu" ana="dieu" sameAs="déité divinité
  providence">diex</w>
  <w pos="PRP" lemma="de" ana="de">de</w>
  <w pos="NOM" msd="sing" lemma="noçoiement" ana="noce" corresp="#mariage" sameAs=
  "épousailles mariage">noçoiement</w>
</persName>
```

Figure 13. *Balisage de deux <persName> désignant Hyménée : « Hymen » et « diex de noçoiement »*

Dans l'extrait traité au § 5.2, seuls « Dieux de l'Érèbe », « Rhodope », « Hémus » et quelques pronoms personnels désignant Orphée ont pu être détectés comme blocs communs. Ces correspondances sont encore plus faibles pour les textes en ancien français.

La possibilité offerte par MEDITE d'ignorer les signes diacritiques est suffisante pour traiter certaines variations des désignations d'entités, comme pour le cas de figure « Cérès » / « Céres ». Mais cette fonctionnalité s'avère risquée pour le traitement des textes littéraires, en ignorant, par exemple, les différences entre présent et participe passé ou les lettres logogrammiques, souvent importantes pour l'analyse. Et PHŒBUS ne propose pas de tels traitements. La structuration XML que nous proposons (fig. 14 pour certaines entités de JXv et de Mv) est certes riche, mais en grande partie automatisable en modifiant les sorties de TreeTagger, par exemple. Elle offre l'intérêt de multiplier les indices et donc de permettre la détection des différentes dénominations d'entités.

Ainsi, pour la variante « humide Ida » et « Ida humide », qui présente un simple déplacement de l'adjectif, la proximité des blocs communs et la taille de l'entité font que ce déplacement n'est pas reconnu par MEDITE. Même si, dans ce cas précis, il s'agit effectivement d'un déplacement, généraliser ce traitement produirait des erreurs indésirables, comme pour cette variante entre Mp et Mv : « il ne porte ni visage serein, ni présage heureux » / « il n'apporte ni parole rituelle ni visage heureux » où l'on voudrait plutôt voir un remplacement synonymique (« visage serein / visage heureux »), une variante dénominative (« présage heureux / parole rituelle ») et le déplacement général de tout le syntagme nominal. Alors que ce dernier exemple est très difficilement automatisable, pour « humide Ida », sa présence au sein de l'élément <persName> offrirait un indice supplémentaire autant sur l'intégralité du nom et de son épithète que de la correspondance des formes « humide Ida » et « Ida humide ».

```
<!--Ida humide-->
<persName corresp="#Ida">
  <w pos="NAM" msd="sing" lemma="Ida">
  Ida</w>
  <w pos="ADJ" msd="sing" lemma="humide"
  sameAs="embué mouillé moite">
  humide</w>
</persName>
```

```
<!--humide Ida-->
<persName corresp="#Ida">
  <w pos="ADJ" msd="sing" lemma="humide"
  sameAs="embué mouillé moite">
  humide</w>
  <w pos="NAM" msd="sing" lemma="Ida">
  Ida</w>
</persName>
```

```
<!--Dieux de l'Érèbe-->
<persName corresp="#Charon #Hadès #Perséphone">
  <w pos="NOM" msd="pl" lemma="Dieu" sameAs="déité divinité providence">Dieux</w>
  <w pos="PRP" lemma="de">de</w>
  <w pos="DET:ART" msd="sing" lemma="le">l'</w>
  <persName corresp="#Érèbe">
    <w pos="NAM" msd="sing" lemma="Érèbe">Érèbe</w>
  </persName>
</persName>
```

```
<!--nocher-->
<persName corresp="#Charon">
  <w pos="DET:ART" msd="sing" lemma=
  "le">le</w>
  <w pos="NOM" msd="sing" lemma="nocher"
   sameAs="navigateur nautonier pilote">
  nocher</w>
</persName>
```

```
<!--batelier-->
<persName corresp="#Charon">
  <w pos="DET:ART" msd="sing"  lemma=
  "le">le</w>
  <w pos="NOM" msd="sing" lemma=
  "batelier" sameAs="nautonier passeur
  pilote">batelier</w>
</persName>
```

```
<!--malheureuse Léthéa-->
<persName corresp="#Léthéa">
  <w pos="ADJ" msd="sing fem" lemma=
  "malheureux" sameAs="pauvre funeste
  déplorable">malheureuse</w>
  <w pos="NAM" msd="sing" lemma=
  "Léthéa">Léthéa</w>
</persName>
```

```
<!--pauvre Léthéia-->
<persName corresp="#Léthéa">
  <w pos="ADJ" msd="sing" lemma="pauvre"
  sameAs="misérable piteux
  malheureux">pauvre</w>
  <w pos="NAM" msd="sing" lemma=
  "Léthéia">Léthéia</w>
</persName>
```

Figure 14. *Table d'entités nommées, de JXv à gauche et de Mv à droite, et leur structuration en XML:TEI*

La syntaxe TEI permet de couvrir tous les cas de figure, comme la possibilité d'imbriquer des entités (noms de lieux dans des noms de personnes, par exemple) et de cumuler des identifiants renseignés au sein de l'attribut @corresp : cette nécessité est visible dans l'occurrence « Dieux de l'Érèbe », qui désigne plusieurs entités (Perséphone, Hadès et Charon), tout en renvoyant à l'entité Érèbe. Le balisage des entités nommées permet également de repérer les périphrases et de les lier à l'entité désignée, comme c'est le cas, par exemple, de « rive », qui, tant dans JXv que dans Mv, désigne le Styx, ou « nocher » et « batelier » désignant Charon. Cela permet de lier ces occurrences à celles plus directes présentes dans notre corpus (JXp : « Orphée essaie de fléchir Charon ; vainement il veut traverser de nouveau le Styx »), mais également entre elles, et de spécifier leur nature grâce aux synonymes renseignés (@sameAs). Même

si la relation synonymique n'est pas aussi directe que pour « malheureuse Léthéa » et « pauvre Léthéia », nous retrouvons la correspondance au niveau des synonymes communs (« nautonier » et « pilote »). Simultanément, l'indice supplémentaire fourni par l'encadrement du texte par des éléments <persName> ou <placeName> et l'identification uniformisée grâce à @corresp permettent un affichage modulable, par exemple concentré uniquement sur les entités nommées. En signalant exclusivement les variantes en leur sein, nous pouvons nous permettre une analyse et un étiquetage plus détaillés et effectuer un alignement partiel en faisant abstraction des autres éléments du texte.

Si nous insistions ici sur la plus-value à attendre pour le traitement par MEDITE, il serait tout à fait équivalent (pour des raisons comparables) pour celui par PHŒBUS.

7. Coopération de PHŒBUS et de MEDITE

Bien évidemment, l'intégration de toutes les modalités de traitement évoquées *supra* et considérées comme étant pertinentes pour les recherches avec PHŒBUS et MEDITE serait coûteuse. Toutefois, nous considérons que le traitement de corpus XML, en permettant aux chercheurs de focaliser leur attention sur des sous-arbres et de demander que des calculs soient opérés en leur sein et intégrés aux résultats d'analyse affichés, marquerait un réel progrès.

Une modulation plus poussée du traitement nous semble pouvoir aboutir à des résultats très probants et susceptibles de faire émerger des besoins plus spécifiques, comme la recherche focalisée sur un champ lexical (amour, souffrance, etc.[23]) et la création de graphes multidimensionnels qui combinent plusieurs types de traitements (fig. 15).

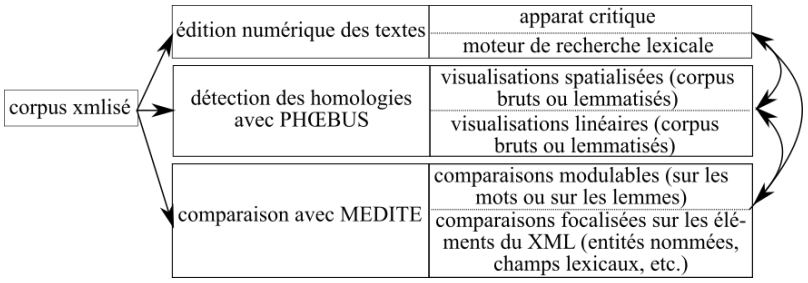

Figure 15. *Schéma éditorial qui prend en compte la coopération entre PHŒBUS et de MEDITE : possibilité d'obtenir plusieurs sorties à partir d'un seul corpus xmlisé.*

Ainsi, la visualisation linéaire des homologies détectées avec PHŒBUS permettrait de cibler les recherches dans les comparaisons textuelles proposées par MEDITE.

23. Cf. fig. 13, où l'appartenance du mot « noçoiement » au champ lexical du mariage est signalée au sein de l'@corresp.

Simultanément, la possibilité de s'appuyer sur une visualisation linéaire pour le travail de comparaison permettrait de localiser plus facilement les extraits qui intéressent le plus le lecteur et de naviguer aisément entre les couples de comparaisons. Par exemple, un clic sur un nœud précis le sélectionnerait comme premier texte et ouvrirait un menu permettant de choisir le texte à lui comparer, puis transporterait le lecteur vers chaque partie intéressante de la comparaison (tout en conservant la possibilité de naviguer dans la totalité des textes). Les visualisations spatialisées (fig. 2 à 6) permettraient, quant à elles, de focaliser l'attention sur un fragment précis pour lequel des homologies ont été détectées entre plusieurs textes. Elles seraient également plus adaptées pour le traitement des corpus sans relation évidente. Dans ce cas de figure, un clic sur l'arête qui lie deux nœuds textuels afficherait leur comparaison avec MEDITE (fig. 16).

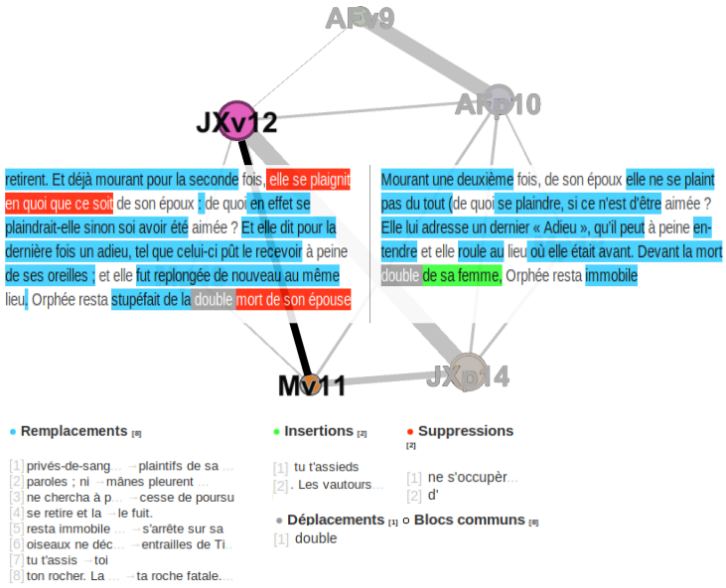

Figure 16. *Exemple d'une coopération possible entre PHŒBUS et MEDITE : le clic sur l'arête qui lie deux nœuds textuels affiche leur comparaison avec MEDITE.*

Une préparation du corpus au préalable par une structuration XML:TEI adaptée à ces besoins spécifiques permettrait d'aboutir à cette liberté de traitement en fonction des contenus de chaque corpus et des objectifs des chercheurs. Elle donnerait la possibilité de lier le traitement de PHŒBUS avec celui de MEDITE et de focaliser la comparaison sur les facteurs communs balisés dans le XML (§ 4 à 6) et des sous-arbres d'éléments (certains syntagmes, les entités nommées, etc.). Pour lier les différents extraits dans les deux types de traitements, chaque mot (<w>) se verrait attribuer un identifiant unique (@id).

8. Conclusion

Notre contribution a cherché à montrer la plus-value de l'évolution proposée des modalités de traitement par PHŒBUS et MEDITE, que ce soit pour la série traductive intermédiale des réécritures du mythe d'Orphée et Eurydice qui nous occupe ou, plus largement, pour l'épanouissement d'autres projets littéraires numériques diversifiés et innovants. Pendant la phase préparatoire de cet article, la participation d'un des coauteurs (durant quelques mois) aux travaux de l'équipe ACASA, a permis d'observer la prise en compte de certaines des observations que nous avons formulées[24]. En remarquant quelques limites des traitements actuels et en suggérant des pistes d'évolution pour les deux logiciels, nous espérons poursuivre ce dialogue étroit entre les chercheurs et les ingénieurs en informatique qui développent les outils et les chercheurs littéraires qui les exploitent.

9. Bibliographie

Abdul-Rahman A., Roe G., Olsen M., Gladstone C., Morrissey R., Cronk N., Chen M., « Constructive visual analytics for text similarity detection », *Computer graphics forum*, vol. 36, n° 1, p. 237-248, février, 2016.

Anonyme, *Ovide moralisé*, Rouen, 1315-1325. Transcrit à partir de Ms. O.4 (fol. 246v-248r). [Attribué de manière incertaine à de Vitry P. et à Legouais Ch.].

Barzilay R., McKeown K. R., « Extracting paraphrases from a parallel corpus », *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Toulouse, France, p. 50-57, 2001.

Büchler M., Crane G., Moritz M., Babeu A., *Increasing recall for text re-use in historical documents to support research in humanities*, Springer Berlin Heidelberg, Berlin, p. 95-100, 2012.

Boukhaled M.-A., Sellami Z., Ganascia J.-G., « Phoebus : un logiciel d'extraction de réutilisations dans des textes littéraires », *22ème conférence sur le Traitement Automatique des Langues Naturelles*, Caen, France, p. 391-396, 2015.

Coffee N., Koenig J.-P., Shakti P., Ossewaarde R., Forstall C., Jacobson S., « Intertextuality in the digital age », *Transactions of the American Philological Association*, vol. 142, n° 2, p. 383-419, septembre, 2012.

Cosnay M., « Ovide, Les Métamorphoses, X », *Musagora*, 2006.

de Parnajon F., « Ovide, choix de Métamorphoses », *Les auteurs latins expliqués d'après une méthode nouvelle par deux traductions françaises*, Hatier, Paris, p. 418-427, 1880.

Del Lungo A., Suchecka K., « Projet eBalzac : construire une bibliothèque hypertextuelle des sources intertextuelles », *DHNord 2019 "Corpus et archives numériques"*, MESH, Lille, octobre, 2019.

Fenoglio I., Ganascia J.-G., « Le logiciel MEDITE : approche comparative de documents de genèse », *L'édition du manuscrit – De l'archive de création au scriptorium électronique*, vol. 10, p. 209-228, 2008.

24. Cf. Ganascia (2019), Del Lungo et Suchecka (2019).

Ferrero J., Simac-Lejeune A., « Détection automatique de reformulations – Correspondance de concepts appliquée à la détection du plagiat », *Actes de la 15ème conférence internationale sur l'extraction et la gestion des connaissances*, Luxembourg, p. 287-298, 2015.

Forstall C., Coffee N., Buck T., Roache K., Jacobson S., « Modeling the scholars : Detecting intertextuality through enhanced word-level n-gram matching », *Literary and linguistic computing*, vol. 30, n° 4, p. 503-515, mai, 2014.

Franzini G., Franzini E., Büchler M., Mueller M., Burns P., *Towards a historical text re-use detection*, Springer International Publishing, Suisse, p. 221-238, décembre, 2014.

Gallet O., Michel L., Murat M., Pradeau C., « Apollinaire numérique », *Revue d'histoire littéraire de la France*, vol. 116, n° 3, p. 533-546, 2016.

Ganascia J.-G., « 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.

Ganascia J.-G., « Graphes et intertextualité », *Humanités numériques*, Centre Universitaire Méditerrannéen, Nice, septembre, 2019.

Ganascia J.-G., Bourdaillet J., « Alignements unilingues avec MEDITE », *Actes des huitièmes journées internationales d'analyse statistique des données textuelles*, Paris, France, p. 427-437, 2006.

Ganascia J.-G., Glaudes P., Del Lungo A., « Automatic detection of reuses and citations in literary texts », *Digital scholarship in the Humanities*, vol. 29, n° 3, p. 412-421, juin, 2014.

Guerry F.-X., « Góngora et ses premiers biographes : une analyse comparative moyennant des outils numériques », *e-Spania*, 2018.

Ho Y., *Corpus stylistics in principles and practice : A stylistic exploration of John Fowles' The Magus*, Advances in stylistics, Bloomsbury Publishing, 2011.

Horton R., Olsen M., Roe G., « Something borrowed : Sequence alignment and the identification of similar passages in large text collections », *Digital Studies / Le Champ numérique*, 2010.

Lafont-Terranova J., Badin F., Niwese M., Comte E., Chevrot G., Colin D., « 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 », MSH Val de la Loire, 2017.

Lepage Y., *Guide de l'édition de textes en ancien français*, Champion, Paris, 2001.

Porter M. F., « Snowball : A language for stemming algorithms », *Retrieved March*, 2001.

Reboul M., *Comparaison semi-automatique des traductions en langue française de l'Odyssée d'Homère (1547-1955)*, 2017. Thèse de doctorat en Littérature comparée, Masson, J.-Y. (dir.), Université Paris IV.

Tomlinson S., « Lexical and algorithmic stemming compared for 9 european languages with Hummingbird SearchServer at CLEF 2003 », *in* P. C., G. J., B. M, K. M. (eds), *Comparative evaluation of multilingual information access systems*, p. 286-300, 2004.

Villenave G. T., « Ovide, Les Métamorphoses, X », *Bibliotheca Classica Selecta*, 2003.

Walleys T., *La Bible des poëtes. Métamorphose d'Ovide moralisée par Thomas Walleys et traduite par Colard Mansion*, Paris, 1493. Transcrit à partir de A. Vérard (fol. 107v-109v).