https://doi.org/10.51305/ICL.CZ.9788076580336.11
Tackling the Toolkit, Plotting Poetry through Computational Literary Studies. ICL CAS: Prague 2021.

On Digital Comparative Editions and Textual Similarity Detection Tools: Towards a Hypertextual Cartography of a Rewritten Myth

Karolina Suchecka
ALITHILA, ULR 1061
Université de Lille, France
karolina.suchecka@univ-lille.fr
0000-0002-6485-7033

Nathalie Gasiglia
STL, UMR 8163
Université de Lille, France
nathalie.gasiglia@univ-lille.fr
0000-0001-7149-2274

Abstract

Our project aims to expose the intertextual relationships observable within a heterogeneous literary corpus. For this purpose, we examine the output of two text reuse detection tools, `Tracer` and `TextPAIR`. We suggest some solutions to overcome the specific limitations observed in those tools and to enhance data quality. We believe that automatic analysis of the rewriting process can make it more comprehensible if the analysis is combined with empirical research methods adapted to the corpus in question.

1 Introduction

Developments in computational technology and digital corpus accessibility are helping to improve the identification of similar passages, phrasing strategies, and references in different literary texts.[1]

In this article, we report on an investigation of rewritings and intertextual references within a corpus of French literary texts about the myth of Orpheus and Eurydice. Our aim is not only to identify text reuse but also to create a hypertextual and modular cartography to represent this phenomenon. At the same time, the heterogeneity of the corpus, which we discuss in Section 2, presents significant problems. Although the texts are closely related thematically, we observe numerous differences in their narrative structure and vocabulary. As works in the corpus were published at various points between the 15th and the 21st century, they include diverse forms of French, and thus, also very different constructions and lexical forms.

[1] Cf. for example Barzilay et al. (2001), Coffee et al. (2012), and Ganascia et al. (2014).

To detect similarities, we use two tools: `Tracer`[2] and `TextPAIR`.[3] Our study compares their results and capacities regarding literary texts. We first describe these tools briefly in Section 3. In Section 4, we then highlight some specific processing challenges related to the preparation of the corpus and the relevance and format of the results. Finally, in Section 5, we suggest a system that combines, reworks, and enriches the relationships detected. From morphosyntactic annotation to the word-level matching stage, we outline our method for adapting the general treatments offered by these tools to the specific requirements of our corpus. This method also focuses on several marked-out factors that can be interconnected.

2 Introducing the Corpus

Of the more than 30 representations of the myth of Orpheus and Eurydice in ancient Greek and Roman literature, *The Georgics* by Virgil (37–30 BC) and *The Metamorphosis* by Ovid (1st AD) are the most frequently reused in subsequent rewritings.[4]

Although these texts differ significantly, the main plot of the myth remains the same: inconsolable after the death of his wife, Eurydice, Orpheus descends into Hell and begs the deities to bring her back to life. Thanks to the beauty of his song, they grant him this favour, provided that he does not look back at Eurydice while she follows him out of the Underworld. Unfortunately, on reaching the surface, he turns around and loses her forever.

In addition to several French translation series,[5] our corpus contains more than 70 rewritings[6] that were created in different genres and during different periods.[7] Proximity to the canonical plot of the myth varies: while theatrical adaptations may be quite closely related, other works, especially poems and modern rewritings, only allude implicitly to the myth. The content and motifs used and their (re)interpretation also differ from one work to the next. Parodies, for example, often satirise Orpheus' music or his intense love for Eurydice. On

[2] Cf. https://www.etrap.eu/research/tracer/ (eTRAP Project, University of Göttingen) and Büchler (2013) and Büchler et al. (2012).

[3] Cf. https://artfl-project.uchicago.edu/text-pair (ARTFL Project, University of Chicago) and Allen et al. (2010) and Horton et al. (2010).

[4] This paper does not discuss the specific relationship between Ovid's and Virgil's versions of this myth, but it is important to note that these versions are not independent; Ovid deliberately alludes to and revises Virgil's text, cf. Kushner (1961) and Segal (1988).

[5] By "French translation series", we mean a subset of French translations of the same ancient source (either Ovid's *The Metamorphosis* or Virgil's *The Georgics*), which date, however, from different periods and reflect different translation practices (parallel, literary, adapted, etc.).

[6] Different theories provide different definitions of *intertextuality* and *rewriting*, especially in the mythological context (cf. Gignoux (2006) and Schnyder (2008)). This study does not discuss the complexity of these definitions and instead divides our corpus somewhat roughly into two subsets: (1) *translation series*, directly assigned to Ovid or Virgil and (2) *rewritings*, i.e. heterogeneous works that refer to the myth (adaptations, parodies, etc.).

[7] The texts in our corpus are available on https://github.com/karolinasuchecka/orphidys. We have also published our scripts there but please note that we are not yet at the final stage of our project and so the documentation is incomplete, particularly in English. We would be happy to receive any feedback and hope that developments in this study, which we will continue to update, will prove relevant and useful for other projects.

the other hand, modern rewritings introduce amalgams, new characters, and scenery that modify the meaning of the myth (Brunel 1997).

To counterbalance the diversity of these rewritings, it is helpful to begin by analysing the reuses detected within translations of Ovid and Virgil respectively and noting similarities and variations in these instances. Since they are very closely related, translation series of ancient sources have the potential to establish a common foundation and, at least indirectly, link different rewritings of the myth. These series, thus, seem to warrant special attention when developing a method to benchmark and improve detection results both for translations and rewritings. We hope to show that translation series play a crucial role in the adaptation of general treatments to the specific requirements of our corpus. For this purpose, we start by observing the initial detection results of `TextPAIR` and `Tracer`, which we introduce below.

3 Introducing the Tools

`TextPAIR` enables the detection of reuse based on a fairly adaptive parameterisation process. As such, the user can choose the minimum number of common words to be detected, submit a list of words to ignore, and determine whether the matching algorithm should take into account words, lemmas, or stemmas (word roots). The tool employs the sequence analysis techniques that are applied, for instance, to detect plagiarism. Initially, `TextPAIR` generates overlapping word sequences (*n-grams*)[8] for each text. It then compares these results with those from sequences in other texts.[9]

In contrast, `Tracer` requires the corpus to be submitted in text format and tokenised into syntactic units (sentences in the case of our corpus). Each unit obtains a unique ID that permits the regrouping of sentences belonging to the same text.[10] The parameters allowed include matching based on lemmas, synonyms, or word embedding. The tool also calculates a score that reflects the proximity between two segments.

Our treatment attempts to reconcile the parameterisation approaches of the two tools. This entails generating tri-grams and using lemmatisation, flattened accents, and minimum three-word matching as well as ignoring word order.

4 Processing the Corpus

As our corpus has been assembled not only for processing with reuse detection tools, but also for presentation within a digital scholarly edition, it is structured according to the XML-TEI P5 standard. The genre of each piece is taken into ac-

[8] Cf. Jurafsky et al. (2009) and, for applications to intertextuality detection, Forstall, Coffee, et al. (2015).

[9] Cf. https://github.com/ARTFL-Project/text-pair.

[10] For text-format corpus preparation and required segment ID formatting, cf. https://tracer.gitbook.io/-manual/manual/corpus-preparation.

count; a common mark-up vocabulary is used to digitise all subgenres, whether they are plays, opera librettos, poetry, or novels.[11]

4.1 Preparing the Corpus

While TextPAIR accepts XML format without exploiting mark-ups, Tracer requires plain text. The latter leads to the loss of marked-up enhancements which might have improved the quality of the results.

Furthermore, Tracer requires the corpus to be split into syntactic units which are then assigned identifiers. Some programming skills are needed for automatic sentence splitting, and the results should be reviewed manually at least for a French literary corpus like our own. Plays and poems seem to be particularly hard to process since they are full of interjections and follow specific capitalisation rules.[12] Processing is, thus, limited to matches between sentences only. In contrast, TextPAIR does not impose this initial preparation requirement and can therefore also detect similarities across sentence boundaries.

4.2 Relevance of the Detected Pairs

To assess the relevance of reuse detected by these tools, a sample of the results is evaluated by a human reader. The initial processing is performed without linguistic enhancements except for lemmas extracted from the corpus annotation by the TreeTagger[13] tool. Five hundred pairs are, thus, randomly retrieved from the results for each of Tracer and TextPAIR.[14] To evaluate the relevance of each match, we follow the 5-point grading schema proposed by Coffee et al. (2012, p. 392).[15] Below we describe all of the potential types of results:

Type 1 A false match caused by bad parameterisation of the tool or failure to adapt the corpus during preparation.

Type 2 A false or irrelevant match based on stop-words (articles, auxiliaries, etc.), very common constructions, or different contexts.

Type 3 A match based on lexical words. This is affected by different contexts, difficulties in evaluating the relationship effectively, or references to different episodes of the myth.

[11] Cf. TEI Guidelines, https://tei-c.org/guidelines/p5/.

[12] Since optimal splitting improves the quality of results, we ultimately decided to enrich the corpus with specific mark-up (<milestone>).

[13] Cf. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

[14] We retain the same number of pairs for each tool rather than assigning it a percentage of the total number of detected relationships. This is because the fact that Tracer detects more reuse than TextPAIR (10 414 to 8851) does not imply that Tracer's cases are more relevant or should be more present within the evaluation sample.

[15] This schema is adapted to meet the specific requirements of our corpus. While Coffee et al. (2012) proposes general criteria related to formal similarity and context proximity, we also evaluate references to the same episodes in the myth, especially for matches of types 3–5.

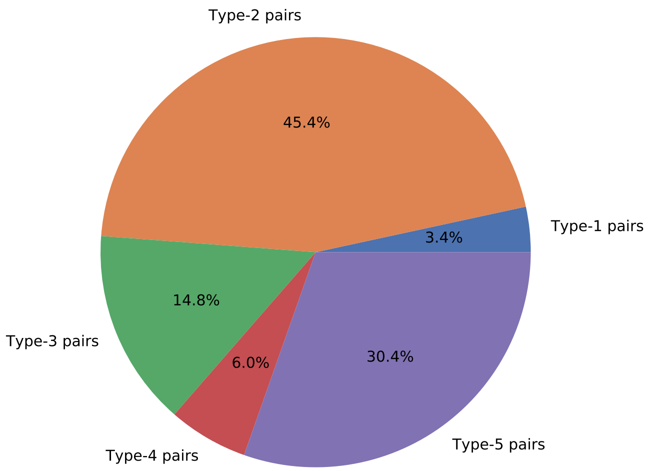


Figure 1: Human reader's evaluation of the results sample

Type 4 A relevant match where the two excerpts refer to the same episode. There are few identical words and expressions are different. The parallel is largely based on allusions.

Type 5 A relevant match that refers to the same episode. The majority of words and expressions are similar or identical.

As noted in Figure 1, irrelevant matches (types 1–2) represent 48.8 % of the evaluation sample, and type-2 pairs comprise the vast majority of these cases.[16] The analysis of these instances reveals several difficulties.

First, even though `TextPAIR` can detect long common segments, the results are most likely to be dominated by noisy data, especially for plays whose stage directions are filled with named entities (designations of speakers in the dialogue, etc.). The results are, thus, often false since the matching is only due to the presence of these named entities (Table 1).[17]

The majority of type-2 matches (259/454) arise within the `Tracer` results. Indeed, while the tool can filter certain words based primarily on their frequency[18] it does not allow us to ignore stop-words only. It is also not possible to exclude the most frequent words from our corpus (even if that would allow us to discard stop-words) since crucial elements for our research, particularly named entities, would then also be ignored.

[16] Type-1 pairs mainly reflect the issues described in Section 4.1. In the sections that follow, we therefore focus exclusively on pairs of types 2–5.

[17] All French excerpts are our own translations. We highlight common words in both the original quotes and their translations.

[18] The value of the minimum or maximum word frequency to be ignored is not customisable, however, cf. `Tracer` *Manual*, "Step 3: Selection. § Selection strategies" (https://tracer.gitbook.io/-manual/manual/configuration/step-3.-selection).

Offenbach et al. (1858)	Ovide (1702, trans. Duryer)
Eurydice : Une heure un quart ! Orphée : Au moins. Eurydice : Je n'écouterais pas !	Eurydice s'évanouit, et le malheureux Orphée n'embrassa que de l'air [...]. Cependant Eurydice qui mourut alors pour la seconde fois [...]
(Eurydice: An hour and a half! Orpheus: At least. Eurydice: I won't listen!)	(Eurydice vanishes and the unfortunate Orpheus embraces nothing but the air [...]. However, Eurydice dies a second time [...])

Table 1: False matching based on stage directions

Finally, while almost all type-5 matches are established within translation series (100 % for `Tracer` and 93 % for `TextPAIR`), rewritings mostly occur among the matches for types 3 and 4. For type-3 pairs, the contexts provided by both tools are not always sufficient to evaluate relevance. We therefore initially deal with relatively few pairs that contain rewritings (71) where we can be certain that the two parts are effectively related.

4.3 Output Formats

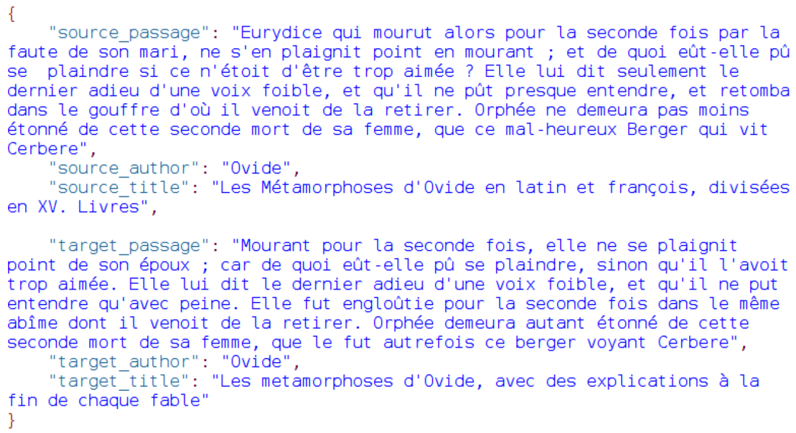
Another problem that we encounter is more formal in nature: working simultaneously with two differently designed tools creates the need for a treatment that can overcome these differences in order to arrive at comparable and compatible results. In addition, these tools generate an output consisting of either a list of properties associated with a value (Figure 2) or a selection of these values in a tabulated file[19]. These formats are machine-readable but difficult to interpret by a human.

Besides the data allowing the identification of each segment, no details are provided for the linguistic elements that enable matching. We would have liked to know, for example, which words are aligned and on what basis (common lemmas or stemmas or synonyms) as well as the number of common entities and the distance between them (i.e. the number of words that separate each common entity from the next one).

5 Post-Processing Results

To determine the optimal method for benchmarking results, we process the evaluation subset (Section 4.2) with a basic algorithm that detects unit-level

[19] Cf. *Tracer Manual*, “Results & computed files”, https://tracer.gitbook.io/-manual/manual/results-and-computed-files.



```
{
    "source_passage": "Eurydice qui mourut alors pour la seconde fois par la
faute de son mari, ne s'en plaignit point en mourant ; et de quoi eût-elle pû
se  plaindre si ce n'étoit d'être trop aimée ? Elle lui dit seulement le
dernier adieu d'une voix foible, et qu'il ne pût presque entendre, et retomba
dans le gouffre d'où il venoit de la retirer. Orphée ne demeura pas moins
étonné de cette seconde mort de sa femme, que ce mal-heureux Berger qui vit
Cerbere",
    "source_author": "Ovide",
    "source_title": "Les Métamorphoses d'Ovide en latin et françois, divisées
en XV. Livres",

    "target_passage": "Mourant pour la seconde fois, elle ne se plaignit
point de son époux ; car de quoi eût-elle pû se plaindre, sinon qu'il l'avoit
trop aimée. Elle lui dit le dernier adieu d'une voix foible, et qu'il ne put
entendre qu'avec peine. Elle fut engloûtie pour la seconde fois dans le même
abîme dont il venoit de la retirer. Orphée demeura autant étonné de cette
seconde mort de sa femme, que le fut autrefois ce berger voyant Cerbere",
    "target_author": "Ovide",
    "target_title": "Les metamorphoses d'Ovide, avec des explications à la
fin de chaque fable"
}
```

Figure 2: Extract of `TextPair` results format

lexical relations (common forms or lemmas). We proceed to perform automatic synonym detection.[20]

We first observe that in order to increase the number and relevance of detected pairs, there is a need for either a more complex NLP pipeline that exploits the taggers trained on each French language variety, or manual correction of the POS-tagging results. Incorrect lemmas produce many false alignments and hinder the effective detection of common linguistic elements, especially among texts in old and modern French.

Synonym detection also produces a lot of irrelevant matches, especially for polysemous verbs such as *retourner* (to turn around, to return). Nevertheless, synonymy appears to have real potential for improving results. Translations in verse are, for example, less connected with their subset since they privilege synonyms, polylexical constructions, and circumlocutions and have to respect rhyme and rhythm structures. But these characteristics are precisely what enable more subtle matches with the subset of rewritings: Ovide (1687, trans. Corneille) is a unique translation of Ovid that relates to a comic opera by Offenbach et al. (1858). A comparison of these texts shows that among the 18 common entities detected within 4 couples, 10 are pairs of synonyms.

`Tracer` can take synonymy into account to detect reuse.[21] It seems, however, that to fully exploit this functionality, general language synonyms should first be adapted to the specific requirements of the corpus.

Finally, among type-5 pairs, the same words, especially modifiers of sentence constituents, are not always correctly aligned ("*il ne porte ni visage serein ni présage heureux*" [he does not have a serene face nor reveal any happy omen] / "*il n'apporte ni parole rituelle ni visage heureux*" [he brings neither a ritual word

[20] Synonym lists were first extracted from a cumulative synonym dictionary for general language (the *Dictionnaire Électronique des Synonymes*, Crisco, Université de Caen, cf. https://crisco2.unicaen.fr/des/).

[21] Cf. *Tracer Manual*, "Pos-tagging, lemmatisation and WordNets", https://tracer.gitbook.io/-manual/manual/pos-tagging-lemmatisation-and-wordnets

nor a happy face]). In order to discard irrelevant matches and improve the detection of common units, it therefore seems necessary to delimit lexical groups. The integration of sentence constituents at the post-processing stage may also enhance the alignment of descriptive paraphrases (*“les âmes nouvelles”* [the new souls]/ *“les ombres arrivés récemment”* [the shadows recently appeared]).

Based on these observations, we develop a three-step post-processing chain that aims to discard irrelevant matches and detail the common linguistic elements of each sufficiently-related pair. First, we convert the outputs obtained with `TextPAIR` and `Tracer` into an enriched format (Section 5.1). We then perform a detailed analysis of each pair to detect the common entities and determine their number and degree of proximity. We calculate a new similarity score and then exclude insufficiently related pairs (Section 5.2). Finally, we integrate the results and enhance the annotations into an XML file (Section 5.3).

5.1 Results Compilation

To compile the results of each treatment, we take advantage of the division of the corpus into the sentences imposed by `Tracer` at the pre-processing stage. Each sentence is then enriched with a multi-level annotation, both for word groups and for words alone. Consider, for example, the case of Figure 3, a fragment of sentence 28 from `duryer1702`.[22]

Contained within the `<s>` element, the identifier of this sentence is provided with the `@xml:id` attribute while the IDs of matching sentences are included in `@corresp`. The nominal, verbal, and adjectival groups are marked as `<phr>` with the `@select` that supplies the lemma of the headword. *“La faute de son mari”* (the fault of her husband) and *“son mari”* (her husband) are marked as nominal groups.[23] The second group nested in the first one is also marked as a named entity *Orphée* using the `<persName>` element, which provides the ID of that entity with `@corresp`. As for word annotation (`<w>`), after manually correcting the POS-tagging results, we propose grammatical (`@pos`) and inflexional (`@msd`) codes, lemmas (`@lemma`), and a selection of synonyms (`@sameAs`).[24] The inflexional codes and synonyms are not provided for stop-words.

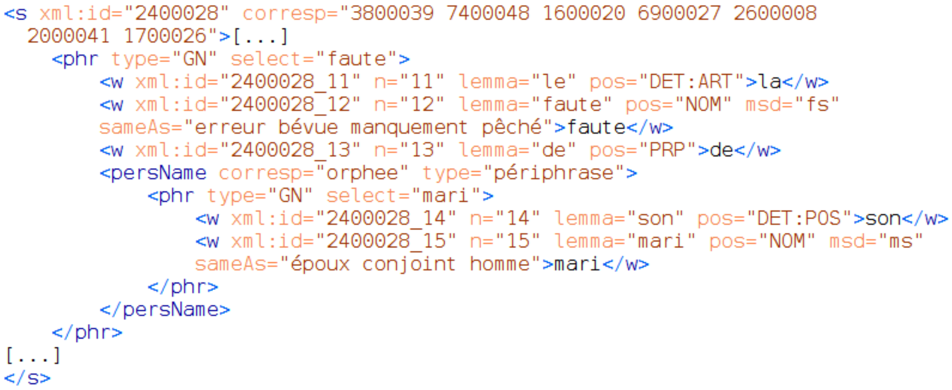
5.2 Searching for Lexical Relations

The algorithm that searches for lexical relations compares each pair of reused elements that was detected by at least one of the tools. It takes into account the enrichment provided through the attribute values and performs multi-level matching.

[22] “[...]*Eurydice qui mourut alors pour la seconde fois par la faute de son mari* [...]” [Eurydice, who dies for the second time through the fault of her husband] (Ovide 1702, trans. Duryer).

[23] In fact, “*la faute de son mari*” and “*son mari*” are part of prepositional groups introduced by *par* and *de* respectively.

[24] DES (footnote 20) classifies synonyms in order of their score, which is thought to represent proximity to the headword (cf. “§ L’ordre des synonymes”, *Présentation du DES*, http://crisco.unicaen.fr/dictionnaire-electronique-des-synonymes/presentation-du-des/). For our annotation, we initially choose the first 4 synonyms.



```
<s xml:id="2400028" corresp="3800039 7400048 1600020 6900027 2600008
  2000041 1700026">[...]
    <phr type="GN" select="faute">
        <w xml:id="2400028_11" n="11" lemma="le" pos="DET:ART">la</w>
        <w xml:id="2400028_12" n="12" lemma="faute" pos="NOM" msd="fs"
        sameAs="erreur bévue manquement pêché">faute</w>
        <w xml:id="2400028_13" n="13" lemma="de" pos="PRP">de</w>
        <persName corresp="orphee" type="périphrase">
            <phr type="GN" select="mari">
                <w xml:id="2400028_14" n="14" lemma="son" pos="DET:POS">son</w>
                <w xml:id="2400028_15" n="15" lemma="mari" pos="NOM" msd="ms"
                sameAs="époux conjoint homme">mari</w>
            </phr>
        </persName>
    </phr>
[...]
</s>
```

Figure 3: Mark-up of an extract of sentence 28 from duryer1702

First, only the lemmas provided as head of sentence constituents are taken into account and we calculate the ratio of exact lexical identity or synonym equivalence.[25] The processing of the evaluation sample shows (Figure 4) that even though some type-2 pairs obtain relatively high ratios (greater than 40 %), in the majority of cases (282/454) less than 30 % of headwords can be matched. For the type-5 pairs, the ratio of only 33 of the 303 pairs is less than 30 %.

The roughly equal distribution of ratios for the type-4 pairs implies that the processing of headwords alone is not sufficient. A more specific adaptation of linguistic resources (lists of synonyms adapted to the corpus, keywords, named entities, and their circumlocutions, etc.) may improve the relevance of headword matching and scoring. For now, and given that the initial results are mainly explored to enhance future treatments, we decide to discard all pairs of ratios below 30 % since irrelevant synonym matches may inflate the score.

When applying our method to sentence 28 of duryer1702, we observe that TextPAIR and Tracer detect very close relationships with 9 translations of Ovid. One reuse is also detected with a novel by Ballanche (1809) ([C], Figure 5).

For sentence A from bellegarde1701 (Ovide 1701, trans. Bellegarde), the correspondence is almost total with 5 identical headwords (marked in red in Figure 5) (*mourir* [to die], *fois* [time], *plaindre* [to complain], repeated twice, and *aimer* [to love]) and 1 group (marked in green) linked by synonymic equivalence (*mari* [husband]∼*époux* [spouse]). Only 2 groups remain unaligned. As for sentence B from corneille1687 (Ovide 1687, trans. Corneille), we find 5 identical headwords (*mourir*, *fois*, *plaindre*, *mari*, and *aimer*) and 1 derivative relationship (*plaindre*∼*plainte* [complaint]; marked in blue). Although no matching is observed for 8 groups from corneille1687, the threshold of 30 % similarity is exceeded for both correspondences and they accede to further processing. However, this does not extend to sentence C from ballanche1809 (Ballanche 1809) since only 2 out of the 9 groups (22 %) can be matched (*plaindre* is repeated twice).

[25] The number of matching headwords is expressed as a percentage of the total number of headwords within a sentence.

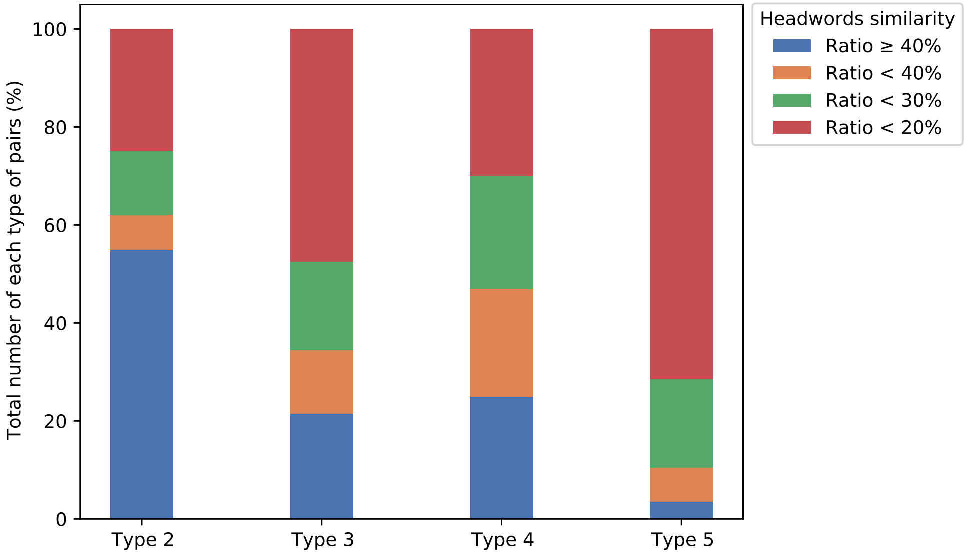


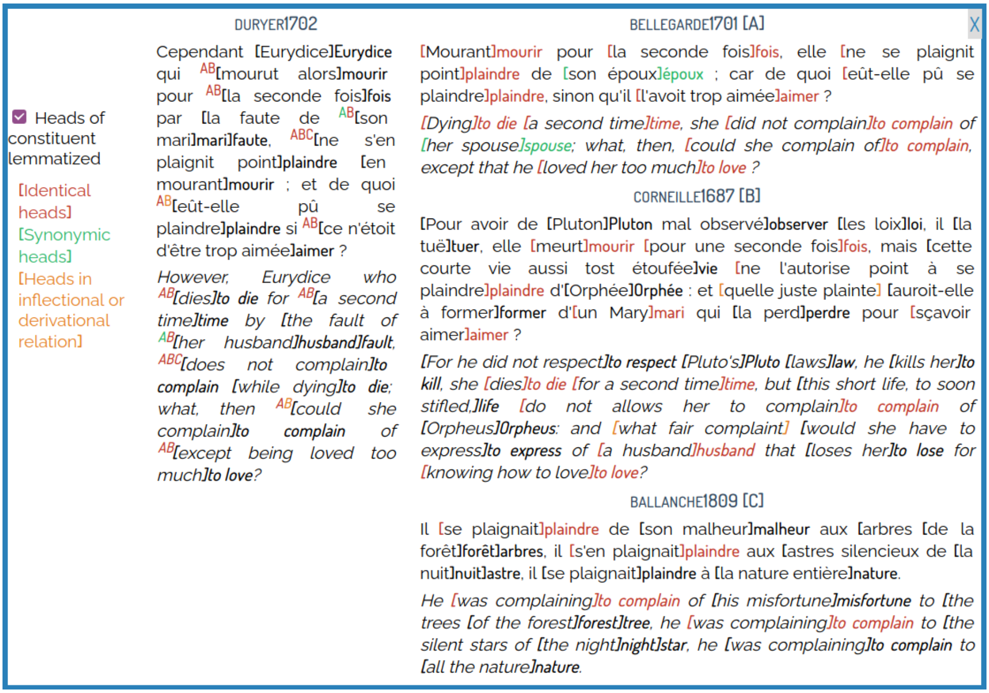
Figure 4: Headwords similarity ratios within evaluation sample

Excluding irrelevant pairs by benchmarking the results is indeed important. But our aim is also to focus on linguistic elements that enable matching and eventually to distinguish the degree of similarity between each pair (quotation, paraphrase, allusion, etc.). To do so, we proceed to word-level matching that excludes stop-words. Currently we consider 3 types of equivalence: (1) unit-level (form∼form, lemma∼lemma, stemma∼stemma), (2) synonym-level, and (3) mixed (form∼synonym, lemma∼synonym).

Several methods can evaluate the lexical proximity between text reuses.[26] Some of these approaches show the relevance of word frequency. However, as all the texts in our corpus share the same mythological theme, the relevant relationships can be based on words of highly variable frequencies. For example, *époux* (spouse) and *mari* (husband) are recurrent periphrases for Orpheus. Their respective frequencies in the corpus are relatively low (146 and 94 out of 432 078), but they appear regularly in the detection results. Indeed, within the evaluation sample, 55 pairs contain the word *époux*, among which 31 are of type 5. However, we also regularly find other keywords from the myth within the type 4–5 pairs derived from the first 100 most frequent words (*amour* [love], *dieu* [god], *mort* [death], *voix* [voice], and *femme* [women/wife]). Therefore, in addition to the headword proximity ratio, we propose a simple lexical proximity assessment method that focuses mainly on the type of equivalence.

First, if a unit-level match is found, 5 points are added to the sentence score and the word pair is no longer taken into account in further processing. For synonym-level and mixed relations, the processing continues in order to achieve optimal matching. Each synonym-level equivalence is worth 0.5 points and each

[26] Cf. for example Büchler (2013, pp. 116–119) and Forstall and Scheirer (2019).



☑ Heads of constituent lemmatized

[Identical heads]
[Synonymic heads]
[Heads in inflectional or derivational relation]

DURYER1702

Cependant [Eurydice]**Eurydice** qui AB[mourut alors]**mourir** pour AB[la seconde fois]**fois** par [la faute de AB[son mari]**mari]faute**, ABC[ne s'en plaignit point]**plaindre** [en mourant]**mourir** ; et de quoi AB[eût-elle pû se plaindre]**plaindre** si AB[ce n'étoit d'être trop aimée]**aimer** ?

*However, Eurydice who AB[dies]**to die** for AB[a second time]**time** by [the fault of AB[her husband]**husband]fault**, ABC[does not complain]**to complain** [while dying]**to die**; what, then AB[could she complain]**to complain** of AB[except being loved too much]**to love**?*

BELLEGARDE1701 [A]

X

[Mourant]mourir pour [la seconde fois]fois, elle [ne se plaignit point]plaindre de [son époux]époux ; car de quoi [eût-elle pû se plaindre]plaindre, sinon qu'il [l'avoit trop aimée]aimer ?

[Dying]to die [a second time]time, she [did not complain]to complain of [her spouse]spouse; what, then, [could she complain of]to complain, except that he [loved her too much]to love ?

CORNEILLE1687 [B]

[Pour avoir de [Pluton]**Pluton** mal observé]**observer** [les loix]**loi**, il [la tuë]**tuer**, elle [meurt]mourir [pour une seconde fois]fois, mais [cette courte vie aussi tost étoufée]**vie** [ne l'autorise point à se plaindre]plaindre d'[Orphée]**Orphée** : et [quelle juste plainte] [auroit-elle à former]**former** d'[un Mary]mari qui [la perd]**perdre** pour [sçavoir aimer]aimer ?

*[For he did not respect]**to respect** [Pluto's]**Pluto** [laws]**law**, he [kills her]**to kill**, she [dies]to die [for a second time]time, but [this short life, to soon stifled,]**life** [do not allows her to complain]to complain of [Orpheus]**Orpheus**: and [what fair complaint] [would she have to express]**to express** of [a husband]husband that [loses her]**to lose** for [knowing how to love]to love?*

BALLANCHE1809 [C]

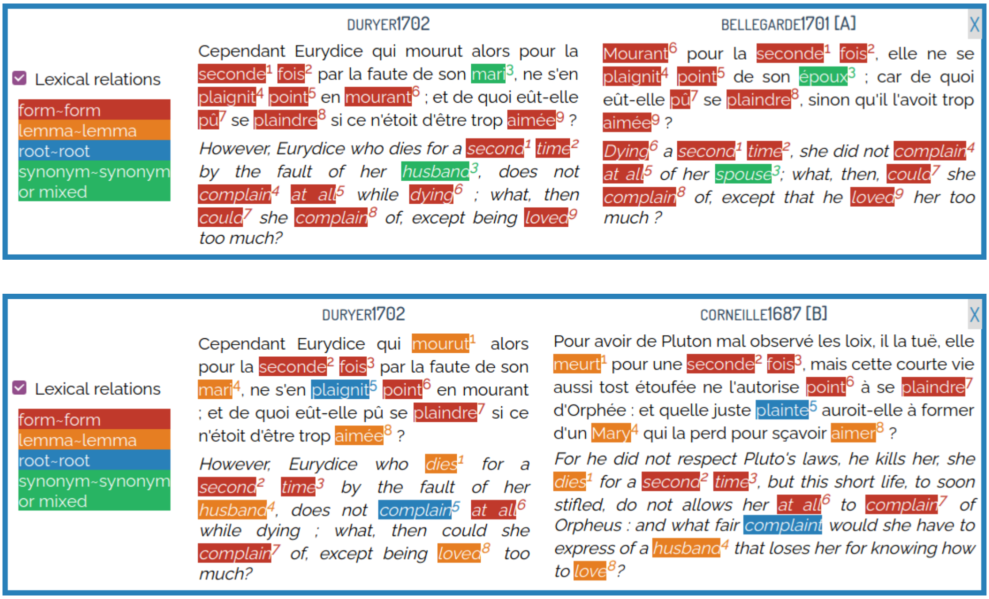
Il [se plaignait]plaindre de [son malheur]**malheur** aux [arbres [de la forêt]**forêt]arbres**, il [s'en plaignait]plaindre aux [astres silencieux de [la nuit]**nuit]astre**, il [se plaignait]**plaindre** à [la nature entière]**nature**.

*He [was complaining]to complain of [his misfortune]**misfortune** to [the trees [of the forest]**forest]tree**, he [was complaining]to complain to [the silent stars of [the night]**night]star**, he [was complaining]**to complain** to [all the nature]**nature**.*

Figure 5: Example of matching headwords

mixed relationship counts for 1.5 points. Points may accumulate if multiple synonym or mixed matches are found. As 4 synonyms are provided for each word, a pair can obtain 3 points maximum (1 mixed and 3 synonym-level matches). The equivalence between *mari* and *époux*, for example, gets 2 points: this reflects 1 mixed match between the lemma *mari* found among the *époux* synonyms, and 1 synonym-level match for *conjoint* (≈marriage partner), which is common to both words. This scoring of synonym matches enables us to distinguish different degrees of proximity. The equivalence of *mari*∼*homme*, for example, accumulates 1.5 points (lemma∼synonym) and *époux*∼*ami* (friend) only 0.5 (synonym *compagnon* [companion]). Currently a synonym match counts for less than a unit-level one since the error rate for the former remains quite high as long as the lexical resource used is intended for general language.

As Figure 6 illustrates, the correspondence with `bellegarde1701`, thus, receives the score of 42 points (8 equivalences on a word-to-word basis and 1 between synonyms). In contrast, the match with `corneille1697` accumulates 40 points (4 same words, 3 same lemmas, and 1 equivalence between stemmas). A moderate difference between the scores for two texts may suggest a comparable degree of proximity. Nevertheless, while both `duryer1702` and `corneille1679` suggest Orpheus' responsibility for the second death of Eurydice ("*par sa faute*" [through his own fault]/ "*il la tuë*" [he kills her]), this relationship seems too complex to be detected automatically. As for "*Cette courte vie aussi tost étoufée*" (this short life, too soon stifled) and "*Pour avoir*



☑ Lexical relations

🟥 form~form
🟧 lemma~lemma
🟦 root~root
🟩 synonym~synonym or mixed

DURYER1702

Cependant Eurydice qui mourut alors pour la seconde^{1} fois^{2} par la faute de son mari^{3}, ne s'en plaignit^{4} point^{5} en mourant^{6} ; et de quoi eût-elle pû^{7} se plaindre^{8} si ce n'étoit d'être trop aimée^{9} ?

However, Eurydice who dies for a second^{1} time^{2} by the fault of her husband^{3}, does not complain^{4} at all^{5} while dying^{6} ; what, then could^{7} she complain^{8} of, except being loved^{9} too much?

BELLEGARDE1701 [A]

Mourant^{6} pour la seconde^{1} fois^{2}, elle ne se plaignit^{4} point^{5} de son époux^{3} ; car de quoi eût-elle pû^{7} se plaindre^{8}, sinon qu'il l'avoit trop aimée^{9} ?

Dying^{6} a second^{1} time^{2}, she did not complain^{4} at all^{5} of her spouse^{3}; what, then, could^{7} she complain^{8} of, except that he loved^{9} her too much ?

☑ Lexical relations

🟥 form~form
🟧 lemma~lemma
🟦 root~root
🟩 synonym~synonym or mixed

DURYER1702

Cependant Eurydice qui mourut^{1} alors pour la seconde^{2} fois^{3} par la faute de son mari^{4}, ne s'en plaignit^{5} point^{6} en mourant ; et de quoi eût-elle pû se plaindre^{7} si ce n'étoit d'être trop aimée^{8} ?

However, Eurydice who dies^{1} for a second^{2} time^{3} by the fault of her husband^{4}, does not complain^{5} at all^{6} while dying ; what, then could she complain^{7} of, except being loved^{8} too much?

CORNEILLE1687 [B]

Pour avoir de Pluton mal observé les loix, il la tuë, elle meurt^{1} pour une seconde^{2} fois^{3}, mais cette courte vie aussi tost étouffée ne l'autorise point^{6} à se plaindre^{7} d'Orphée : et quelle juste plainte^{5} auroit-elle à former d'un Mary^{4} qui la perd pour sçavoir aimer^{8} ?

For he did not respect Pluto's laws, he kills her, she dies^{1} for a second^{2} time^{3}, but this short life, to soon stifled, do not allows her at all^{6} to complain^{7} of Orpheus : and what fair complaint would she have to express of a husband^{4} that loses her for knowing how to love^{8}?

Figure 6: Example of word-level matching

de Pluton mal observé les loix” (for he did not respect Pluto’s laws), they are stylistic additions to `corneille1687`, a verse translation.

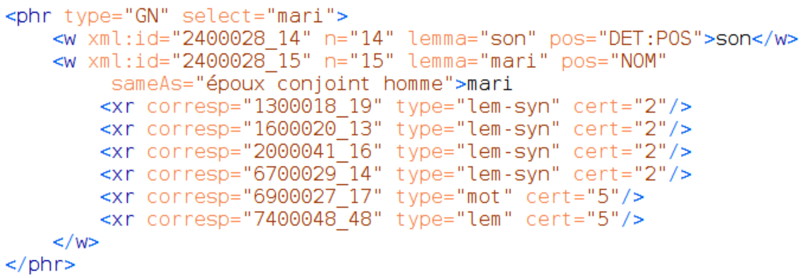
To compare sentences of different lengths and complexities, we therefore calculate a lexical similarity ratio. This is expressed as a percentage of the maximum score a pair would accumulate if all the lexical words of the shortest sentence were matched at unit-level. The maximum score for `bellegarde1701` is 45 points, and so the relationship with `duryer1702` receives 93.3 % lexical ratio. As for `corneille1687`, the maximum score for the sentence is 110 points. It is, however, the shorter sentence in `duryer1702` that we take into account when calculating the lexical ratio. As 8 words out of 13 match, the lexical similarity of the pair is 65 % (in contrast with only 36 % when we calculate the ratio in `corneille1687`).

Applying this method to our evaluation sample suggests that the threshold for the lexical ratio is lower than the scoring based on sentence constituents.[27] If we assume that only pairs with a lexical similarity above 20 % (and a headwords similarity above 30 %) are relevant, then without losing type-4 and 5 pairs, we can discard 76 additional type-2 pairs. Meanwhile 96/454 remain.

To enable this kind of global analysis, it is important to preserve the enriched results in the adapted format. Such an approach will allow the results to be exploited for information retrieval that can lead, often through traditional analysis, to a significant improvement in the initial results.[28]

[27] For 58 % of type 2 pairs, the ratio is less than 20 and for 15 %, it is between 20 and 30. Only 4 % of type 5 pairs do not exceed the threshold of 20, and 16 % are between 20 and 30.

[28] This also allows for the visualisation of the results, a process that we do not detail here. The examples provided in Figure 5 and Figure 6 are retrieved from an operating interface developed



```
<phr type="GN" select="mari">
    <w xml:id="2400028_14" n="14" lemma="son" pos="DET:POS">son</w>
    <w xml:id="2400028_15" n="15" lemma="mari" pos="NOM"
         sameAs="époux conjoint homme">mari
        <xr corresp="1300018_19" type="lem-syn" cert="2"/>
        <xr corresp="1600020_13" type="lem-syn" cert="2"/>
        <xr corresp="2000041_16" type="lem-syn" cert="2"/>
        <xr corresp="6700029_14" type="lem-syn" cert="2"/>
        <xr corresp="6900027_17" type="mot" cert="5"/>
        <xr corresp="7400048_48" type="lem" cert="5"/>
    </w>
</phr>
```

Figure 7: Enriched mark-up of an extract from sentence 28 in `duryer1702`

5.3 Enriched Results Exploitation

The results of the word matching are included in the initial annotation of the sentence using an `<xr>` element, child of `<w>`, which specifies the ID of the matching word (`@corresp`), the type of relationship detected (`@type`), and its score (`@cert`). In this way, the extract from sentence 28 in `duryer1702` presented in Figure 3 is enriched by 6 elements `<xr>` (Figure 7). Indeed, 6 out of 10 correspondences contain an equivalence with the word *mari*, in most cases based on synonymy with *époux* (4 occurrences).

The enriched results are then included into an XML database that compiles all the matches detected beforehand by `TextPAIR` and `Tracer`. This compilation allows for a wide range of manipulation and analyses. Some of these are specific and focused on a chosen text, while others are general and include all the correspondences.

Sentence 28 of `duryer1702`, for example, can be included in a larger passage that consists of 6 sentences and relates the entire episode of Eurydice's second death, from Orpheus' fatal look to his dismay over his wife's vanishing. Based on the 156 lexical relations found in this excerpt, we can establish at least partial connections with 16 sentences from 11 different texts. Almost half of the matches (71) are found in sentence 28. The most frequently connected words are the proper names of the two lovers, their periphrases, and 3 verbs: *plaindre*, *mourir*, and *aimer*. These could be considered the most salient keywords for this episode if the same trend is confirmed through an analysis of the overall results obtained from other translations of Ovid. For each episode and each variant of the myth, we aim to determine the keywords and their synonyms based on the evidence within the translation series. Using these findings, we hope to improve the results for rewritings. To take one example, the keywords observed for the episode of Eurydice's second death could be exploited to establish more relevant matching between the translation by `duryer1702` and the novel by `ballanche1809`:

to facilitate exploring specific or complex reformulations. By the end of our project, this interface should be adapted and converted into an open access digital comparative edition.

> Mais, ô faiblesse d'un cœur qui aime ! [...] Vaincu par cette puissance contre laquelle l'homme lutte en vain, Orphée se retourne. [...] Eurydice s'évanouit [...] et sa parole plaintive, inarticulée, meurt dans le vague des airs [...].[29]

Working progressively and starting from the most explicit adaptations of the myth, we, thus, plan to extend our method to increasingly allusive and symbolic connections, parodies, reinterpretations, and modernisations. The objective is to improve the understanding of the role of the linguistic processes observed within these texts.

6 Conclusion

The automatic detection of intertextual relations is an exciting prospect for literary and linguistic researchers as well as for the creators of digital scholarly editions. False detection and recognition problems are integral to this process and confirm the importance of combining distant and close reading. An in-depth understanding of the corpus is essential not only for the analysis of the results, but also for their improvement and enrichment so that new treatments can be applied to the same texts. These new approaches may prove more fruitful and capable of revealing increasingly subtle and surprising relationships.

As part of our project, we are endeavouring to establish some shared practices for preparing digital editions in the humanities. To this end, we employ interoperability, open access data, and usage sharing. In addition, we propose an automated processing method that generates enriched files marked up to the XML-TEI standard. Nevertheless, we cannot claim that our treatment can be generalised easily. It may be adapted with varying success to different corpora with manual adjustments needed to improve the data quality. It will also take significant work and time to obtain the first meaningful results. For some texts, no relevant relationships will ever be found. Still all of this seems to us to be inherent to literary research, whether it is traditional and empirical or supported by computational techniques.

References

Allen, Timothy and Charles Cooney (2010). "Plundering Philosophers: Identifying Sources of the Encyclopédie". In: *Journal of the Association for History and Computing*.

Ballanche, Pierre-Simon (1809). "Sixième Fragment". In: *Œuvres de M. Ballanche*. Paris: J. Barbezat, pp. 488–492.

Barzilay, Regina and Kathleen R. McKeown (2001). "Extracting Paraphrases from a Parallel Corpus". In: *Proceedings of the 39th Annual Meeting of the ACL*. Toulouse: Association for Computational Linguistics, pp. 50–57. DOI: 10.3115/1073012.1073020.

[29] [But, O weakness of the loving heart ! [...] Defeated by this power against which man struggles in vain, Orpheus turns around. [...] Eurydice vanishes and her unarticulated complaints die in the air], Ballanche (1809, p. 491).

Brunel, Pierre (1997). “Orphée Moderne”. In: *Apollinaire Entre Deux Mondes. Le Contrepoint Mythique Dans Alcools. Mythocritique II*. Écriture. Paris: Presses Universitaires de France, pp. 63–82.

Büchler, Marco (2013). “Informationstechnische Aspekte Des Historical Text Re-Use”. PhD thesis. Leipzig: Université de Leipzig.

Büchler, Marco, Gregory Crane, Maria Moritz, and Alison Babeu (2012). “Increasing Recall for Text Re-Use in Historical Documents to Support Research in Humanities”. In: *Theory and Practice of Digital Libraries*. Ed. by P. Zaphiris, G. Buchanan, E. Rasmussen, and F. Loizides. Berlin: Springer Berlin Heidelberg, pp. 95–100.

Coffee, Neil, Jean-Pierre Koenig, Poornima Shakti, Roelant Ossewaarde, Christopher Forstall, and Sarah Jacobson (2012). “Intertextuality in the Digital Age”. In: *Transactions of the American Philological Association* 142.2, pp. 383–422. DOI: 10.1353/apa.2012.0010.

Forstall, Christopher, Neil Coffee, Thomas Buck, Katherine Roache, and Sarah Jacobson (2015). “Modeling the Scholars: Detecting Intertextuality through Enhanced Word-Level n-Gram Matching”. In: *Digital Scholarship in the Humanities* 30.4, pp. 503–515. DOI: 10.1093/llc/fqu014.

Forstall, Christopher and Walter Scheirer (2019). “Lexical Matching: Text Reuse as Intertextuality”. In: *Quantitative Intertextuality: Analyzing the Markers of Information Reuse*. Cham: Springer International Publishing, pp. 55–78. DOI: 10.1007/978-3-030-23415-7_3.

Ganascia, Jean-Gabriel, Pierre Glaudes, and Andrea Del Lungo (2014). “Automatic Detection of Reuses and Citations in Literary Texts”. In: *Digital Scholarship in the Humanities* 29.3, pp. 412–421.

Gignoux, Anne-Claire (2006). “De l’intertextualité à La Récriture”. In: *Cahiers de Narratologie. Analyse et théorie narratives* 13. DOI: 10.4000/narratologie.329.

Horton, Russell, Mark Olsen, and Glenn Roe (2010). “Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections”. In: *Digital Studies / Le Champ numérique* 2.1. DOI: 10.16995/DSCN.258.

Jurafsky, Daniel and James H. Martin (2009). “N-Gram Language Models”. In: *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. New York: Prentice-Hall, Inc., pp. 189–232.

Kushner, Eva (1961). *Le Mythe d’Orphée Dans La Littérature Française Contemporaine*. Paris: A.G. Nizet.

Offenbach, Jacques and Hector Crémieux (1858). *Orphée Aux Enfers*. Paris: Calmann-Lévy.

Ovide (1687). *Les Metamorphoses, Mises En Vers François*. Trans. by Thomas Corneille. Vol. 2-3. Paris: Berthelemy Girin, Michel Brunet.

Ovide (1701). *Les Métamorphoses, Avec Des Explications à La Fin de Chaque Fable*. Trans. by Jean-Baptiste Morvan de Bellegarde. Paris: Michel David.

Ovide (1702). *Les Métamorphoses En Latin et François*. Trans. by Pierre du Ryer. Amsterdam: P. & J. Blaev, Janssons à Waesberge, Boome, & Goethals.

Schnyder, Peter, ed. (2008). *Métamorphoses Du Mythe: Réécritures Anciennes et Modernes Des Mythes Antiques*. Universités - Domaine Littéraire. Paris: Orizons.

Segal, Charles (1988). *Orpheus: The Myth of the Poet*. Baltimore: John Hopkins University Press.