Mesures d'intertextualité entre Homère et les tragiques grecs

Marianne Reboul

Mesures d'intertextualité entre Homère et les tragiques grecs

Marianne Reboul
École Normale Supérieure de Lyon
marianne.reboul@ens-lyon.fr

Résumé

L'intertextualité en grec ancien, outre qu'elle est difficile à définir sur le plan théorique, est complexe à évaluer sur le plan technique, et ce pour au moins deux raisons. D'abord toutes les langues grecques ne se ressemblent pas, et ne sont donc pas comparables avec de simples calculs de distance d'édition, et leur traitement demande une normalisation de la graphie, du vocabulaire et, dans une certaine mesure, de la syntaxe, qui n'est possible que grâce aux techniques récentes apportées par les principaux modules de traitement automatique des langues anciennes. Ensuite, et bien que ce problème soit commun à tous les exercices de détection d'intertextualité, les références entre auteurs sont d'autant plus difficiles à repérer qu'elles peuvent ne pas être explicites. Les modèles contextuels récents sont efficaces pour ce type de détection sur des langues vivantes, mais le peu de données à disposition pour les langues anciennes rend le problème plus épineux encore, puisque les modèles de type BERT sont friands de données massives. Nous proposons d'exposer ici les différents modes de détection de l'intertextualité possibles, en les modulant pour un objet d'étude précis : nous souhaitons mesurer le degré de proximité sémantique, et potentiellement d'intertexte, entre les ouvrages d'Homère et ceux de deux des tragiques grecs, Sophocle et Eschyle. Ce travail est un travail encore en cours, et le code utilisé pour obtenir les résultats est susceptible d'amendements (Reboul, 2023).

1 Introduction

Dans cette étude, nous nous concentrons sur l'identification de l'intertextualité située à l'échelle de séquences de texte, ou *loci similes* (Manjavacas et al., 2020), et non des textes dans leur intégralité. L'intertextualité peut être au moins de deux types. Il peut d'abord y avoir un intertexte explicite : des pans d'expressions sont repris d'un auteur à l'autre, permettant de déterminer que l'auteur cible a eu connaissance de l'auteur source. Ce premier type d'intertexte est déjà problématique à définir avec certitude. Par exemple, dans le cas de deux types de langue différents, comme c'est le cas pour les variantes du grec ancien, il peut y avoir mention explicite d'un auteur plus ancien, mais pas forcément dans son dialecte d'origine (il peut, par exemple, s'agir d'un dialecte actualisé). Autre exemple, il peut y avoir mention explicite de textes antérieurs, sans pour autant que les extraits mentionnés soient significatifs sur le plan sémantique : une expression idiomatique, qui pourrait être composée de mots outils, ou de termes très usuels, devrait pouvoir être considérée comme un intertexte, même si celle-ci n'est pas propre au texte source. L'intertexte explicite peut aussi être difficile à détecter parce qu'il peut être parcellaire : une citation, volontairement ou non, inexacte peut être aussi considérée comme de l'intertexte. Pour la détection explicite de l'intertexte, il faut donc, semble-t-il, retenir au moins un critère : la spécificité de l'expression employée, qui ne peut appartenir ou être reconnue que comme un trait spécifique de l'auteur source (ou qui le serait devenu à force d'usage, même si l'auteur source peut lui-même en hériter).

Le second type d'intertexte, l'intertexte implicite, que l'on pourrait assimiler à une influence d'un auteur sur un autre (consciente ou inconsciente), est encore plus difficile à détecter avec certitude sur le plan technique et théorique. Nous partons du présupposé théorique que nous ne pouvons pas avec certitude affirmer qu'il y a intertexte implicite entre deux auteurs. Nous pensons qu'il y a des échos, plus ou moins importants, potentiellement détectables, mais qu'il ne nous est pas possible d'affirmer avec certitude que ces échos sont effectivement des intertextes. Sur le plan technique, la détection de l'intertexte est d'autant plus complexe qu'elle nécessite de faire appel à des mesures de similarités moins contrôlables que celles mises en place lors de la détection d'intertexte explicite. En effet, la détection d'intertexte non explicite implique d'avoir recours à des modèles sémantiques d'extraction de données, comme des modèles de plongement de mots ou des modèles contextuels,

dont le processus décisionnel est souvent nettement plus complexe à analyser, et dont les conclusions sont parfois ininterprétables pour les utilisateurs (pensons, par exemple, à la difficulté d'interprétation des *topics* dans une opération de *topic modelling*).

Le paramétrage du modèle de langue et de son utilisation sont aussi sujets de discussion et ont un fort impact sur les résultats obtenus. Enfin, les résultats eux-mêmes sont immanquablement objets d'interprétation : autant il est acceptable d'affirmer la présence d'un intertexte lorsque celui-ci est explicite et immédiatement reconnaissable, autant il est plus complexe d'en affirmer la présence lorsque l'intertexte est purement sémantique.

2 Propositions

Nous proposons donc une utilisation médiane de ces technologies, qui entre dans le sillage de ce qui a déjà été fait par Manjavacas et al. (2019). Nous précisons qu'il ne s'agit pas de technologies nouvelles, puisque les mesures et modèles que nous allons employer existent déjà. Nous proposons un cas d'application spécifique, et la réutilisation du code que nous soumettons implique une nécessaire adaptation à l'objet d'étude choisi. Nous paramétrons nos mesures de sorte qu'elles soient adéquates pour notre objet d'étude. Là encore nous rejoignons les conclusions de Manjavacas et al. (2019), qui tentent une approche mixte entre analyse lexicale et analyse sémantique pour un objet d'étude précis (et donc un paramétrage spécifique à l'objet d'étude).

Nous avons choisi de nous concentrer sur un sujet relativement consensuel, mais dont les conclusions n'ont pas fait jusqu'ici l'objet d'expérience concrète et mesurable sur le plan quantitatif, à savoir l'identification de similarités entre le corpus homérique et deux des tragiques grecs, Eschyle et Sophocle.

Nous avertissons le lecteur que les traductions qui sont produites dans cet article sont des traductions de l'auteur, susceptibles d'être amendées.

3 Corpus et techniques

Pour cette étude nous utilisons le corpus fourni par la *Perseus Digital Library*, et en particulier le corpus de *treebank-data* disponible sur GitHub (Library, 2004). Nous avons fait ce choix car il s'agit d'un corpus propre, et dont les lemmes ont été corrigés et vérifiés. Nous avons donc choisi de comparer l'*Iliade* et l'*Odyssée* d'Homère, selon les leçons retenues par la *Perseus Digital Library* (qui se fonde sur l'établissement de Monro et Allen (1902), et les tragédies d'Eschyle (*Agamemnon* (Aeschylus, 1999), les *Euménides* (Aeschylus, 1999), les *Choéphores* (Aeschylus, 1999), les *Perses* (Aeschylus, 1983), *Prométhée enchaîné* (Aeschylus, 1983), les *Sept contre Thèbes* (Aeschylus, 1983), les *Suppliantes* (Aeschylus, 1983)) et Sophocle (*Ajax* (Sophocles, 1913), *Antigone* (Sophocles, 1912), *Électre* (Sophocles, 1913), *Œdipe Roi* (Sophocles, 1912), les *Trachiniennes* (Sophocles, 1913)). Le corpus des auteurs n'est donc pas intégral. Ce choix nous pousse à exclure un auteur qui aurait eu sa place dans notre étude, à savoir Euripide. Les tragédies d'Euripide ne figurant pas dans le *treebank-data* de *Perseus*, il aurait fallu procéder à une lemmatisation automatique, qui aurait pu générer, suite aux erreurs produites par le lemmatiseur, un bruit conséquent dans les résultats.

Les outils que nous employons pour créer des scores de similarités entre des séquences textuelles sont les suivants (et sont tous implémentés dans le code fourni). Nous calculons les plongements de mots, que nous appellerons *embeddings* dorénavant, grâce au modèle transformer de type BERT `ancient-greek-bert` (Singh et al., 2021), et nous mesurons leur similarité grâce au module `faiss` (Johnson et al., 2019), qui permet une prise en charge de distance cosinus et euclidienne. Nous avons fait ce choix par souci d'optimisation, mais il est à noter que, du fait de l'indexation des données et de la *quantization* (qui regroupe les points similaires de l'espace vectoriel), ce choix peut donner lieu à des résultats plus approximatifs que l'utilisation plus longue et plus coûteuse de mesures de similarité classiques. Nous modulons les scores de similarité obtenus lors de la comparaison des *embeddings* grâce à deux autres mesures, une distance de Jaro-Winkler entre chaque lemme de chaque séquence source et chaque lemme de chaque séquence cible, et une mesure TF-IDF qui évalue la spécificité d'un terme dans son document par rapport à un corpus global, en tenant compte de la longueur des documents et en normalisant la fréquence des termes.

3.1 Constitution et brassage du corpus

Le corpus est intégralement tiré des données de *treebank-data* issues du dépôt GitHub de *Perseus*. Ainsi nous évitons autant que possible les erreurs d'étiquetage et de lemmatisation. Nous conservons

en mémoire la structure des données, à savoir le découpage en phrases qui est proposé par *Perseus*, grâce à l'élément `<sentence>`. Nous retenons ensuite, pour chaque sous-élément `<word>`, les attributs `@form` et `@lemma`. Cependant, les attributs de formes ne servent qu'à la visualisation des résultats : l'ensemble des calculs est effectué sur les lemmes, afin d'éviter une trop grande disparité orthographique. Les textes sont ensuite mis en basse casse et les diacritiques sont enlevés. Les textes sont divisés en deux groupes, les textes sources (ceux d'Homère) et les textes cibles (des autres auteurs). Nous proposons ensuite de mesurer la similarité des séquences textuelles à l'échelle de la phrase, en la pondérant avec des calculs de similarité sur des éléments-mots locaux.

3.2 Modèle BERT pour le grec ancien

Pour mesurer le degré de similarité sémantique entre différentes séquences de texte, nous utilisons un modèle de type `transformer` (Devlin et al., 2019) qui permet d'obtenir une représentation vectorielle des textes étudiés. Le modèle prend en compte non seulement la représentation individuelle des mots dans l'espace vectoriel, mais aussi le contexte dans lequel les mots sont utilisés. Nous obtenons les lemmes de chaque vers des fichiers XML du corpus *treebank-data*, et nous tirons une représentation vectorielle de chaque vers grâce au modèle BERT. Le modèle a été entraîné sur une grande quantité de textes (notamment pour la phase qui ne tient pas à l'étiquetage syntaxique, qui n'arrive qu'en phase d'affinage), et est particulièrement efficace pour traiter les variations morphologiques entre les différents dialectes grecs (Singh et al., 2021, p.135). Il nous a donc semblé adéquat pour notre usage, puisqu'il permet une mise en perspective plus large de la langue que des générateurs d'*embeddings* de mots fixes (contrairement aux *embeddings* contextuels de BERT), qui eux seraient entraînés sur un corpus plus restreint. Puisque nous souhaitons percevoir l'intertexte au niveau du vers et au-delà, les vecteurs contextuels, entraînés sur un corpus massif et varié, semblent plus adaptés. Il est à noter que le modèle a été entraîné sur des données qui ont subi un pré-traitement minimal, que nous reproduisons sur nos propres données, à savoir une mise en basse casse et une suppression des diacritiques.

Nous avons ajouté une possibilité de visualisation de l'attention dans le code. Sans entrer dans les détails de ce qu'est l'« attention » pour un modèle de type `transformer`, rappelons que dans un phrase, tous les mots n'ont pas un poids équivalent pour permettre de comprendre l'ensemble de la phrase, et le mécanisme d'attention crée des connexions entre des mots qui ne sont pas forcément directement liés dans la phrase. Ce code ne peut en revanche être exécuté sans une RAM de GPU conséquente et un disque dur d'au moins 80 gigaoctets (les fichiers de stockage mémoire des variables permettant de faire moins grand usage de la RAM).

3.3 TF-IDF

Une fois les *embeddings* contextuels constitués, nous calculons, grâce à une mesure TF-IDF (Sparck Jones, 1972), la pertinence de chaque terme des textes sources par rapport à l'ensemble du corpus fourni par le dépôt *treebank-data* de *Perseus*, afin d'identifier les termes les plus spécifiques du vocabulaire homérique. Cette partie de l'étude bénéficierait grandement d'un élargissement du corpus de comparaison. Cependant, nous excluons d'autres textes pour la même raison que nous avons exclu de notre étude les tragédies d'Euripide : nous faisons le choix de ne traiter que des textes déjà annotés et corrigés, pour éviter le bruit que pourrait générer un lemmatiseur externe. Les termes spécifiques à Homère que nous rencontrons et qui se trouvent aussi dans les tragédies étudiées sont nombreux. À titre d'exemple (les exemples retenus ont été sélectionnés aléatoirement), citons « δύσμορος » (« malheureux »), présent à la fois chez Homère et chez Sophocle, qui a un score de spécificité de 0.8 (le score allant de 0 à 1) ; « γοάω » (« gémir »), présent à la fois chez Homère et Eschyle, avec un score de spécificité de 0.8 ; ou encore « βλώσκω » (« s'en aller »), présent chez les trois auteurs, avec une spécificité de 0.75. Nous pouvons ensuite, pour chaque phrase, calculer le score maximum obtenu pour chaque mot et moduler le score de proximité des *embeddings* : si une phrase source a à la fois un score de proximité vectorielle et un score de TF-IDF élevés, le score global sera lui aussi plus élevé. Si le score de proximité vectorielle est haut, mais que les termes employés ne sont pas spécifiques à Homère, le score est pondéré. Cela nous permet, théoriquement, de discerner les phrases proches sémantiquement les unes les autres, tout en augmentant le score si deux termes (ou plus) spécifiques à Homère se trouvent à la fois dans la

source et dans la cible.

3.4 Distance de Jaro-Winkler

Nous avons ajouté une mesure supplémentaire aux deux mesures précédentes pour pouvoir prendre en compte, *a minima*, le style « inconscient » (Cafiero et Camps, 2022) des auteurs, qui peut résider, entre autres, dans l'usage des mots-outils [1]. Les mots-outils obtiennent presque toujours un score proche de 0 lors d'une analyse de type TF-IDF parce qu'ils ne sont spécifiques d'aucun auteur. En revanche, il est intéressant de les conserver, tant dans le calcul des *embeddings* (lorsqu'il s'agit d'*embeddings* contextuels) que dans une étude qui tient davantage de la stylométrie (Kestemont, 2014). En effet, les mots-outils permettent de conserver l'empreinte d'un style. Il existe de nombreuses autres techniques que celle que nous avons employée pour les conserver et en tirer profit (n-grammes, Delta de Burrows, de Elder, de Hoover, etc.). Nous avons choisi la distance de Jaro-Winkler (1999), méthode permettant une granularité fine (à l'échelle de la phrase ou du mot notamment, notre but n'étant pas de faire de l'identification d'auteur) et une mise en place technique plus aisée et plus modulaire. D'autre part, nous avons pensé que ce type de mesure serait plus facilement adaptable à d'autres objets d'étude que le nôtre.

Dans notre cas, la distance de Jaro-Winkler est utilisée pour mesurer la similarité entre un mot source et un mot cible. À la différence de la distance d'édition classique (Levenshtein, 1966), la distance de Jaro-Winkler nous permet de donner plus de poids à l'ordre des caractères dans un mots, et plus particulièrement au préfixe. Ainsi, elle est adaptée pour le grec qui, même lemmatisé (et donc sans flexions), est une langue agglutinante : il est donc fréquent de retrouver des termes mêlés, et la distance de Jaro-Winkler permet de mieux appréhender leur ressemblance avec d'autres termes. D'autre part, la distance de Jaro-Winkler pondère davantage les variations en début de chaîne que celles en fin de chaîne : même si les flexions ont été ôtés, pour le cas de mots agglutinés, elle est donc plus efficace qu'un calcul de distance plus traditionnel. Pour plus de détails sur la comparaison des différents calculs de distance dans les chaînes de caractères, voir Navarro (2001). Pour chaque séquence de caractères source, nous prenons chaque séquence de caractères cible, nous la tokénisons en mots, et nous mesurons la distance de Jaro-Winkler entre ces mots. Cela n'est cependant pas effectué sur tous les mots : les mots doivent être relativement longs (par défaut nous avons proposé une longueur de cinq caractères minimum). Le score maximum obtenu sur l'ensemble de la séquence est celui qui viendra infléchir le score global de similarité.

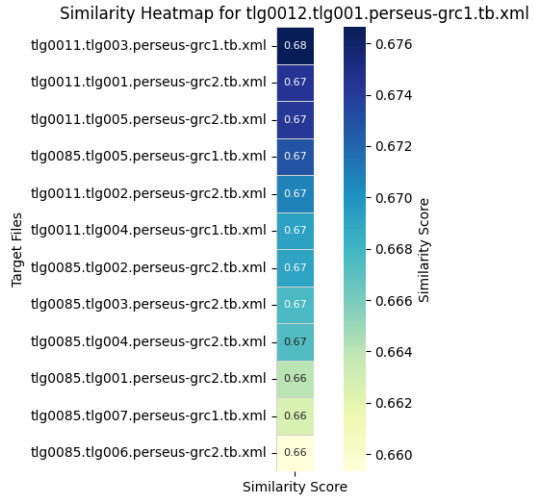
Les scores sont donc pondérés de la façon suivante : si la phrase source est très courte, et le score de spécificité maximum du TF-IDF supérieur à un certain seuil (par défaut 0.6), nous calculons le score global de similarité en faisant la moyenne du score de proximité des *embeddings* avec le score maximum de TF-IDF des lemmes communs entre les deux séquences. Ainsi, nous n'augmentons pas le score global si la phrase est trop courte, et donc potentiellement avec un seul mot véritablement spécifique à la source. Si la phrase est plus longue, et que le score maximum de TF-IDF des lemmes communs dépasse le seuil par défaut, nous ajoutons au score global une distance maximum de Jaro-Winkler. Ainsi, pour des séquences plus longues, plus les lemmes communs sont spécifiques et se ressemblent, plus ils ont de poids. Enfin, si aucune de ces conditions n'est remplie, seul le score de proximité des *embeddings* est pris en compte. Les séquences sont ensuite triées par score, et nous pouvons obtenir un tableau de données exploitable graphiquement.

4 Résultats

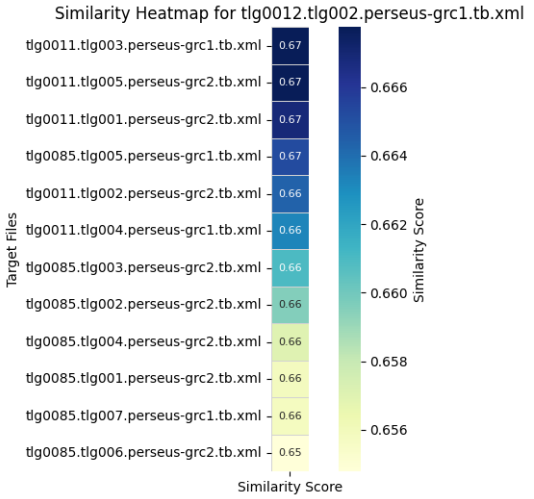
4.1 Classification globale

Qu'il s'agisse de l'*Iliade* ou de l'*Odyssée*, le résultat est sans appel, puisque c'est le fichier `tlg0011.003`, c'est-à-dire *Ajax* de Sophocle, qui est le plus proche, tous scores confondus, des deux textes sources. Cependant, il est à noter que c'est moins l'influence du score de Jaro-Winkler que celui induit par le modèle BERT qui a le plus de poids dans le score global, et ainsi, nous retrouvons dans les séquences sources et dans les séquences cibles de chaque texte non seulement quelques lemmes communs, mais surtout des accents, des sens communs. Si l'intertexte n'est pas explicite, il est possible que l'influence d'un texte sur un autre soit pourtant bien réelle. Et c'est ce qui conditionne l'apparition de forts scores de similarité avec *Ajax* : non seulement le verbe en soi, mais aussi l'idée, par exemple avec la ré-exploitation de

1. Pour la liste des mots outils que nous utilisons, voir Berra (2020).



(a) Score de similarité globale entre les textes de Sophocle et Eschyle et l'*Iliade* d'Homère.



(b) Score de similarité globale entre les textes de Sophocle et Eschyle et l'*Odyssée* d'Homère.

FIGURE 1 – Similarité entre les textes homériques et les textes d'Eschyle et Sophocle.

mythes ou de figures communes, font que l'inspiration homérique est tangible.

Autre exemple chez Eschyle cette fois, où la pièce la plus proche de l'*Iliade* est *Agamemnon*. Globalement, Eschyle emploie beaucoup plus de termes spécifiques à Homère que Sophocle. Pourtant, la similarité sémantique globale est moindre. Là encore, il s'agit d'un travail en cours, mais notre hypothèse est que l'héritage homérique est bien plus conscient ou explicitement affirmé chez Eschyle que chez Sophocle, mais que Sophocle reprend pourtant, avec une trame plus constante et plus dissimulée, les accents des textes homériques. Globalement, à part avec *Agamemnon*, qui est le texte le plus « homérique » selon nos résultats, Sophocle est nettement plus proche d'Homère, alors qu'Eschyle a plus de vocabulaire spécifique commun avec l'auteur des épopées.

4.2 Rapprochements de séquences

Par rapport aux autres textes du corpus, *Ajax* fait partie de ceux qui ont le moins de lemmes explicites en commun avec les deux épopées. Mais, lorsque c'est le cas, ces lemmes font l'objet d'une attention forte du modèle, et ces lemmes ont eux aussi un score TF-IDF élevé. Donc, même si le nombre de lemmes communs est faible, les images communes peuvent faire penser à de l'intertexte. Par exemple, une des séquences qui comporte un fort score de similarité globale est :

> ἀλλάἰσχυνόμενοι φάτιν ἀνδρῶν ἠδὲ γυναικῶν, μή ποτέ τις εἴπῃσι κακώτερος ἄλλος Ἀχαιῶν ἦ πολὺ χείρονες ἄνδρες ἀμύμονος ἀνδρὸς ἄκοιτιν μνῶνται, οὐ δέ τι τόξον ἐΰξοον ἐντανύουσιν· [2]

Odyssée, chant XXI, vers 323 à 326

Et son correspondant dans *Ajax* est :

> εἰ δ'ὑποβαλλόμενοι κλέπτουσι μύθους οἱ μεγάλοι βασιλῆς ἢ τᾶς ἀσώτου Σισυφιδᾶν γενεᾶς, μὴ μή, ἄναξ, ἔθ'ὧδ ἐφάλοις κλισίαις ὄμμ'ἔχων κακὰν φάτιν ἄρῃ. [3]

Ajax, vers 188 à 192

Les termes communs entre les deux séquences sont peu nombreux et relativement communs (« κακός », « φάτις »), mais il s'agit pourtant de séquences qui ont beaucoup en commun. D'abord, ce sont des exhortations d'un corps étranger, celui des prétendants d'une part, celui du chœur de l'autre. Les valeurs antithétiques des différents personnages sont aussi mises en évidence : dans les deux cas, un être vil, presque anonyme, vient salir la réputation d'un ou d'une victime. Enfin, un autre sème présent dans l'une et l'autre séquence est celui de l'inaction, ou de l'incapacité de réaliser l'action qui permettrait de distinguer le héros, celui qui bande l'arc d'une part, celui qui sort de sa folie de l'autre.

2. Que l'on pourrait traduire par « Mais honteux de ce que diraient hommes et femmes, de peur que le plus vil des Achéens ne dise que des hommes bien inférieurs convoitent l'épouse d'un homme irréprochable, sans pouvoir tendre son bel arc. »

3. Que l'on pourrait traduire par « Mais si c'est une calomnie tramée dans l'ombre par les Atrides, ou par l'infâme rejeton de la race de Sisyphe, je t'en conjure, ne reste pas le regard fixe près des tentes, en excitant une rumeur fâcheuse. »

Autre exemple, que nous prenons chez Eschyle cette fois. Bien souvent, la proximité chez Eschyle est importante grâce à la reprise exacte de termes spécifiques à Homère. Dans cet exemple, les termes communs entre les deux séquences, plus fréquents que chez Sophocle, ont un indice moyen de TF-IDF, mais le sens des séquences se rejoint seulement par touches. Dans les vers suivants, c'est Nestor qui hésite, comme la mer, balancé par les mouvements de sa conscience :

> ὡς δ ὅτε πορφύρῃ πέλαγος μέγα κύματι κωφῷ ὀσσόμενον λιγέων ἀνέμων λαιψηρὰ κέλευθα αὔτως, οὐ δ ἄρα τε προκυλίνδεται οὐδετέρωσε, πρίν τινα κεκριμένον καταβήμεναι ἐκ Διὸς οὖρον, ὣς ὃ γέρων ὥρμαινε δαϊζόμενος κατὰ θυμὸν διχθάδι, ἢ μεθ ὅμιλον ἴοι Δαναῶν ταχυπώλων, ἦε μετ Ἀτρεΐδην Ἀγαμέμνονα ποιμένα λαῶν. [4]
>
> *Iliade*, chant XIV, vers 16 à 22

Chez Eschyle, c'est au Chœur des vieillards qu'il revient de présenter l'intrigue, sur une tonalité manifestement épique :

> δέκατον μὲν ἔτος τόδ᾽ἐπεὶ Πριάμου μέγας ἀντίδικος, Μενέλαος ἄναξ ἠδ᾽Ἀγαμέμνων, διθρόνου Διόθεν καὶ δισκήπτρου τιμῆς ὀχυρὸν ζεῦγος Ἀτρειδᾶν στόλον Ἀργείων χιλιοναύτην, τῆσδ᾽ἀπὸ χώρας ἦραν, στρατιῶτιν ἀρωγὰν, μέγαν ἐκ θυμοῦ κλάζοντες Ἄρη τρόπον αἰγυπιῶν, οἵτ᾽ἐκπατίοις ἄλγεσι παίδων ὕπατοι λεχέων στροφοδινοῦνται πτερύγων ἐρετμοῖσιν ἐρεσσόμενοι, δεμνιοτήρη πόνον ὀρταλίχων ὀλέσαντες· [5]
>
> *Agamemnon*, vers 40 à 53

Remarquons plusieurs traits particuliers à ces deux passages : d'abord, la présence d'une comparaison filée avec la mer, où l'esprit des personnages se fond dans le mouvement marin (qui s'accompagne du vocabulaire maritime traditionnel, plus spécifique à Homère, chez les deux auteurs). Notons aussi la récurrence de la thématique du duo, avec Nestor balancé entre deux partis à prendre d'un côté, et le couple des Atrides de l'autre. Enfin les deux séquences ont effectivement des lemmes communs, à savoir « θυμός » (le « cœur » ou le « désir »), « μέγας » (« grand »), « Ἀγαμέμνων », « Ζεύς » et « Ἀτρείδης » (« Atride »), qui viennent augmenter le score de similarité. Notons aussi qu'il faudrait, pour mieux mettre en lumière les correspondances entre les deux séquences, et notamment lorsque celles-ci sont liminaires, comparer la scansion de l'une et l'autre séquence, pour voir s'il est possible, dans le rythme, de percevoir un rapprochement. Cela sera fait dans une étude à venir. Autre cas spécifique à Eschyle, moins présent chez Sophocle, les séquences courtes : des termes explicitement homériques, mais réinvestis, avec une coloration légèrement différente, par transposition. Par exemple, dans les *Sept contre Thèbes*, pour la séquence « παντοδαπὸς δὲ καρπὸς χαμάδις πεσὼν ἀλγύνει κυρήσας· » [6], l'inspiration homérique est tangible dans la séquence correspondante « ὃ δ ἐξ ἵππων χαμάδις πέσε , λύντο δὲ γυῖα. » [7], qui raconte la chute et la mort d'Iphinoos. L'image des fleurs ainsi tombant ne peut qu'évoquer à l'auditoire la chute des héros de l'*Iliade* et ainsi laisser présager la chute des prétendants au trône de Thèbes.

Notons enfin que le notebook que nous proposons permet d'interroger les sources homériques, en entrant une chaîne de caractères, pour visualiser les proximités les plus fortes dans l'ensemble du corpus. D'autres visualisations sont aussi proposées, entre autres un *clustering* des fichiers sources (qui présuppose que les deux fichiers sources ne peuvent pas appartenir au même cluster).

5 Perspectives

Notre perspective à court terme est d'élargir le corpus, afin d'augmenter la masse de données brassées par le programme (et donc rendre les résultats plus significatifs), et afin, sur le plan littéraire, d'in-

4. Que l'on pourrait traduire par « Comme lorsque dans une onde silencieuse la haute mer rougit, ouvrant les voies rapides des vents sifflants, elle ne se roule ni d'un côté ni de l'autre, sans qu'un vent plus fort venu de Jupiter lui soit envoyé, le vieillard roulait en lui-même partagé dans un entre-deux, sans savoir s'il irait vers la foule des grecs aux coursiers rapides, ou vers Agamemnon fils d'Atrée, pasteur des peuples. »

5. Que l'on pourrait traduire par « Voici la dixième année depuis que le grand ennemi de Priam, le roi Ménélas, et Agamemnon, doués par Zeus d'un double trône et d'un double sceptre, les deux puissants fils d'Atrée, ont entraîné loin de cette terre les mille nefs de la flotte Argienne, force guerrière, et ont poussé l'immense clameur d'Arès, comme crient les aigles, qui, avec les douleurs du deuil de leurs enfants, s'élevant au-dessus de leur nid, volent en cercle avec les rames de leurs ailes. »

6. Que l'on pourrait traduire par « Des fruits de toute espèce tombent au sol, triste spectacle que cette chute » (*Ajax*, vers 357 à 358).

7. Que l'on pourrait traduire par « Et celui-ci tomba de ses chevaux à terre, et ses membres se relâchèrent. » (*Iliade*, chant VII, vers 16).

clure des auteurs qui n'auraient pas dû être exclus du corpus sur le plan strictement théorique. Cela implique donc par conséquent de procéder à un traitement approfondi des corpus, en les lemmatisant, en les étiquetant et en les vérifiant.

Nous rejoignons les conclusions de Manjavacas et al. (2020) lorsque les auteurs de l'article expliquent que l'un des enjeux majeurs de l'utilisation de techniques d'analyse sémantique globale (comme le *topic modelling* dans leur cas) est de rendre les résultats interprétables pour l'utilisateur, afin de pouvoir faire évoluer au mieux l'approche nécessairement mixte de la détection d'intertexte. Nous comptons donc poursuivre nos investigations sur les retours graphiques qui peuvent être fournis (par les modèles contextuels notamment) pour expliciter les résultats obtenus.

Nous avons paramétré notre approche afin qu'elle permette au mieux la détection explicite et implicite de l'intertexte homérique, à l'échelle de la phrase ou du vers, pour le grec uniquement. Pour une compréhension plus globale de l'intertextualité, il faudrait aussi pouvoir inclure d'autres langues. Par exemple, il faudrait pouvoir appliquer ce type d'expériences à des textes latins, et, pour Homère, commencer par Virgile.

Bibliographie

Aeschylus. 1983. *Aeschylus. 1 : Suppliant Maidens, Persians, Prometheus, Seven Against Thebes*. Loeb classical library. Harvard University Press - W. Heinemann, Cambridge (MA) - Londres.

Aeschylus. 1999. *Aeschylus. 2 : Agamemnon, Libation-bearers, Eumenides*, repr édition. The Loeb classical library. Harvard University Press - W. Heinemann, Londres.

Aurélien Berra. 2020. aurelberra/stopwords v2.3.0.

Florian Cafiero et Jean-Baptiste Camps. 2022. *Affaires de style : du cas Molière à l'affaire Grégory, la stylométrie mène l'enquête*. Le Robert, Paris.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, et Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, et Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3) :535–547.

Mike Kestemont. 2014. Function Words in Authorship Attribution. From Black Magic to Theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 59–66, Gothenburg, Sweden. Association for Computational Linguistics.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8) :707–710.

Perseus Digital Library. 2004. GitHub - PerseusDL/treebank_data: Perseus Treebank Data.

Enrique Manjavacas, Folgert B. Karsdorp, et Mike Kestemont. 2020. *A Statistical Foray into Contextual Aspects of Intertextuality*, volume 2723, pages 77–96. CEUR Workshop Proceedings.

Enrique Manjavacas, Brian Long, et Mike Kestemont. 2019. On the feasibility of automated detection of allusive text reuse. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 104–114, Minneapolis, USA. Association for Computational Linguistics.

David Binning Monro et Thomas William Allen. 1902. *Homeri Opera, recognovit brevique adnotatione critica instruxit*, scriptorum classicorum bibliotheca oxoniensis édition, volume 1. Clarendon, Oxford.

Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1) :31–88.

Marianne Reboul. 2023. OdysseusPolymetis/humanistica2023: v.0.1.

Pranaydeep Singh, Gorik Rutten, et Els Lefever. 2021. A Pilot Study for BERT Language Modelling and Morphological Analysis for Ancient and Medieval Greek. In *The 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2021)*, pages 128–137.

Sophocles. 1912. *Sophocles. 1 : Oedipus Tyrannus, Oedipus at Colonus, Antigone*. The Loeb classical library. Harvard University Press - W. Heinemann, Cambridge (MA).

Sophocles. 1913. *Sophocles. 2 : Ajax, Electra, Trachiniae, Philoctetes*. The Loeb classical library. Harvard University Press - W. Heinemann, Cambridge (MA).

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1) :11–21.

William E. Winkler. 1999. The state of record linkage and current research problems. *Statistical Research Division, US Bureau of the Census, Wachington, DC*.