# Le projet Intelligence artificielle littéraire (IAL) : définir formellement le concept de variation au sein de l'*Anthologie grecque* ?

Yann Audin[1], Mathilde Verstraete[2], Dominic Forest[3], Marcello Vitali-Rosati[4]

[1]Université de Montréal – yann.audin@umontreal.ca

[2]Université de Montréal – mathilde.verstraete@umontreal.ca

[3]Université de Montréal – dominic.forest@umontreal.ca

[4]Université de Montréal – marcello.vitali.rosati@umontreal.ca

## Abstract

The Literary Artificial Intelligence (IAL) project investigates the possibility of formalizing literary concepts using computational and algorithmic principles. Our corpus of study is the *Greek Anthology*, and the concept studied the of variation, abundantly present in this corpus. This contribution summarizes the exploratory experiments conducted on a first sample, namely the French translation of the book VI. We provide an in-depth analysis of the preliminary results, outline the methodology employed, and lay the foundations for the next phases of this pilot project.

**Keywords**: Theories of literature, Greek literature, Greek anthology, Variation, Vector space model, Intertextuality, Digital humanities, Modeling.

## Résumé

Le projet Intelligence Artificielle Littéraire (IAL) interroge la possibilité d'une formalisation de concepts littéraires à travers des principes computationnels et algorithmiques. Notre corpus d'étude est l'*Anthologie grecque*, le concept étudié est celui de la variation, abondamment présent dans ce corpus. Cette contribution résume les expérimentations exploratoires conduites sur un premier échantillon, à savoir la traduction française du livre VI. Nous procédons à une analyse approfondie des résultats préliminaires, exposons la méthodologie employée, et établissons les fondements pour les prochaines phases de ce projet pilote.

**Mots clés** : Théories de la littérature, Littérature grecque, Anthologie grecque, Variation, Modèle vectoriel, Intertextualité, Humanités numériques, Modélisation.

## 1. Introduction

En 2012, Stéphane Marche se positionnait clairement en faveur du paradigme des deux cultures, où sciences humaines et sociales s'opposent aux sciences naturelles (Snow, 1959). Son article, « Literature is not Data », se ralliait à une longue tradition théorique, celle d'une différence irréconciliable entre le sens et la syntaxe (voir Searle, 1980) et d'une incalculabilité du littéraire (Meunier, 2017 ; Da, 2019). Comme d'autres projets des dernières décennies, le présent article se positionne à contre-courant de cette idée : nous cherchons ici à mettre à l'épreuve cette supposée incompatibilité.

Le projet IAL (Intelligence Artificielle Littéraire) dérive du projet d'édition numérique collaborative de l'Anthologie grecque, mené à la Chaire de recherche du Canada sur les

Écritures numériques (CRCEN) depuis 2014[1]. Nous étudions la possibilité de formuler une définition formelle (Piper, 2017 ; Rhody, 2012 ; Meunier, 2017) – ou plus précisément computationnelle ou algorithmique – d'un concept littéraire. Une telle définition ne reposerait ni sur l'exemplification, ni sur le langage naturel, enclin aux ambiguïtés et à la subjectivité. Ainsi, notre objectif est de penser, concevoir et implémenter des algorithmes de fouille de données et de traitement automatique du langage permettant de définir formellement un concept littéraire (ici, la variation) au sein d'un corpus donné (soit, l'*Anthologie grecque*). Ce corpus dispose d'une diversité de formes intertextuelles, dont la variation, phénomène particulièrement abondant dans le genre épigrammatique (Laurens, 2012 ; Tarán, 1979). De nombreux projets se sont déjà intéressés à la recherche d'intertextualités dans les textes anciens, par le biais d'approches aux visées heuristiques (notamment Schubert, 2020 ; Pöckelmann et al., 2020 ; Coffee et al., 2012). La spécificité de IAL se trouve dans son dessein : plutôt que de rechercher des occurrences de variations, nous tentons ici d'en produire une définition formelle ; si l'algorithme est capable de retrouver les variations précédemment relevées, c'est qu'il incarne la définition dudit concept. La formalisation du concept de variation pourrait être élaborée en combinant des approches algorithmiques d'analyse textuelle et de fouille de données, intégrant également les implications théoriques découlant de ces méthodes (Roberts, 2000 ; Rockwell, 2016).

## 2. Corpus de l'*Anthologie grecque*

L'*Anthologie grecque* est un recueil regroupant la poésie épigrammatique grecque antique de la période classique à la période byzantine. Elle résulte de compilations successives, modifiées, additionnées et réarrangées par des compilateurs divers, de Méléagre (I^er^ s. av. J.-C.) aux scribes de l'*Anthologie palatine* en passant – entre autres – par Agathias (VI^e^ s. apr. J.-C.) ou Céphalas (X^e^ s. apr. J.-C.) (Gutzwiller, 1998). L'expression « Anthologie grecque » désigne un ensemble constitué par deux parties. D'une part, l'*Anthologie palatine*, un manuscrit datant du X^e^ siècle (le *codex Heidelbergensis Palatinus graecus* 23) retrouvé en 1606 par Claude Saumaise à la Bibliothèque palatine de Heidelberg[2]. D'autre part, l'*Appendix Planudea*, soit les épigrammes absentes du manuscrit palatin mais présentes dans l'*Anthologie de Planude*, une compilation datant du début du XIV^e^ siècle, rédigée par Maxime Planude dans le *Marcianus gr*. 481 (Aubreton, 1968). Le genre de l'épigramme évolue de manière significative dans l'histoire de la littérature grecque, mais également au sein même de l'*Anthologie*. Dès l'époque hellénistique, l'épigramme – initialement composée pour être gravée – se détache de son support et se diversifie dans son propos, de sorte à devenir un véritable genre littéraire, caractérisé par sa brièveté et son trait railleur.

L'édition de l'*Anthologie grecque* menée à la CRCEN et la structure de données qu'elle a permis de produire constituent le corpus de départ de ce projet, donnant accès au texte original (image et transcription), à diverses traductions multilingues, aux commentaires marginaux du manuscrit (scholies, gloses, [inter]titres, etc.), mais aussi à des commentaires contemporains. La plateforme et le modèle de données sur lequel elle se construit favorisent la mise en évidence des relations intertextuelles présentes à l'intérieur du manuscrit, notamment par l'usage de marqueurs codés (auteurs, thèmes, mots-clés, etc.).

[1] Le manuscrit a été séparé en deux parties à la page 614 (entre les livres XIII et XIV). La seconde partie (pp. 615-709) se trouve à la Bibliothèque Nationale de France sous le nom *Parisinus Suppl. Gr.* 384.

[2] Nous avons choisi de commencer par les approches algorithmiques les plus simples et dont il est facile de comprendre les implications de modélisation. Pour cette raison nous avons, pour le moment, évité d'utiliser des approches comme celles décrites par Schubert (2020) ou Pöckelmann et al. (2020), plus opaques.

## 3. La variation dans l'*Anthologie grecque*

L'*Anthologie grecque* présente différentes formes intertextuelles, parmi lesquelles le phénomène de « variation » (Tarán, 1979). Concrètement, la variation – largement encouragée par les pratiques rhétoriques antiques – se manifeste par la reprise et l'adaptation d'un texte provenant d'un prédécesseur ou contemporain. Le genre de l'épigramme se prête particulièrement bien à cette pratique poétique, offrant un terrain propice à des variations quasiment infinies. La simplicité de sa forme permet aux auteurs de briller en quelques vers seulement, tout en rendant impossible l'épuisement complet de la richesse d'un sujet en un seul poème. En variant un même sujet, les auteurs explorent toutes ses facettes, créant ainsi un défi artistique où le triomphe est d'autant plus grand que le thème a été traité de manière plus exhaustive. Pierre Laurens, auteur d'un volume considérable dédié au genre épigrammatique (2012), identifie trois niveaux où s'opère la variation. La variation *stylistique* concerne les mots et leur agencement, introduisant de « multiples mais infimes modifications » (Laurens, 2012, 128) par quelques éléments déplacés ou par des substitutions d'ordre lexical ou stylistique (Laurens, 2012, 124). La variation *rhétorique*, quant à elle, porte sur la forme générale des épigrammes ; « l'impression est celle d'une multiplication à l'infini des possibilités d'expression d'une même idée » (Laurens, 2012, 127). Enfin, la variation *paradigmatique* conserve la structure de l'épigramme, mais en fait varier le sujet même, lequel est considéré comme une variable parmi d'autres, résultat de « la répétition d'une structure simple, combinée avec la variation (...) du sujet » (Laurens, 2012, 130). Ces trois types de variation sont considérés comme des concepts littéraires distincts dans notre projet. L'un des exemples les plus fameux illustrant la variation stylistique est le groupe de variations sur les trois frères (VI.11-16 et 179-187). Les poèmes rapportent la dédicace de trois frères (un chasseur, un pêcheur et un oiseleur) au dieu Pan. Chacun lui offre un filet adapté à son style de chasse. L'épigramme (l'originale serait la VI.13) a été reprise 14 fois, plusieurs poètes l'ont même variée à plusieurs reprises. À ce groupe s'ajoutent les épigrammes VI.174 ou la VI.17 (parodique), illustrant la variation paradigmatique[3].

| | |
|---|---|
| *Les trois frères t'ont consacré, chasseur Pan, ces filets, pris par chacun à son genre de chasse : Pigrès, pour les oiseaux ; Damis, pour les quadrupèdes ; Cléitor, pour le peuple de la mer. Envoie-leur en échange une bonne chasse à l'un par les airs, au second par les bois, à l'autre par les grèves.* | *À Pan, trois frères ont consacré ces instruments de leur profession : Damis un panneau pour les bêtes des montagnes, Cleitor ces filets à poissons, Pigrès cet infrangible collet à prendre les oiseaux. Car jamais de leur chasse l'un dans les bois, l'autre dans les airs, l'autre sur les eaux, leur logis ne les a vus revenir les rets vides.* |
| *AP*, VI.13 (Léonidas de Tarente) | *AP*, VI.14 (Antipater de Sidon) |

## 4. Objectifs

Notre objectif est de créer un modèle formel du concept de variation, capable de classifier une paire d'épigrammes comme étant ou non une variation, et si oui de quel type. Notre hypothèse est que si ce modèle (constitué d'un ensemble d'algorithmes) est en mesure d'identifier les variations dans l'*Anthologie grecque*, il en représentera la définition formelle. Pour cette raison,

[3] Ce nombre représente chaque paire $[E_i,E_j]$ pour lesquelles $i > j$, de sorte qu'aucune paire ne contienne deux fois le même poème et en assumant que $m(E_i, E_j) = m(E_j, E_i)$ où $m(E_i, E_j)$ est une mesure de similarité entre $E_i$ et $E_j$.

nous privilégions les méthodes algorithmiques non opaques : l'interprétation de ce modèle est centrale à notre démarche qui est de nature herméneutique et non pas heuristique[4]. En effet, le travail d'annotation du corpus (l'identification des variations) est du ressort des philologues ; dans ce projet, les méthodes algorithmiques n'ont pas comme fonction d'en découvrir de nouvelles. Une définition formelle de la variation pourrait prendre la forme d'une (combinaison de) mesure(s) de similarité entre deux épigrammes et d'une valeur seuil au-delà de laquelle la paire de poèmes est considérée comme une variation. Le processus de définition formelle de la variation dans l'*Anthologie grecque* est itératif (Ramsay, 2011 ; Rockwell, 2016) : le modèle est affiné par des retours aux textes et à la lecture rapprochée. Nous nous concentrons dans cette phase sur la variation stylistique puisqu'elle est définie par les cooccurrences de termes et les modifications lexicales.

## 5. Méthodologie

Puisque nous considérons la variation comme une relation entre deux épigrammes, nous sommes confrontés à des limites computationnelles importantes : les 4 134 épigrammes présentes sur notre plateforme représentent 8 542 911 paires à analyser[5]. Nous avons donc décidé de commencer par nous concentrer sur l'un des 16 livres de l'*Anthologie* pour réduire le temps de computation nécessaire. Le livre VI est l'un des mieux annotés sur la plateforme et les épigrammes votives qu'il contient sont particulièrement propices aux variations (Laurens, 2012, 104-108). Il contient 358 épigrammes, soit 63 903 paires puisque nous assumons que le concept de variation est commutatif (si A est une variation de B, B est aussi une variation de A). Plus de 300 variations ont été répertoriées dans ce sous-ensemble du corpus. En nous limitant au livre VI, nous pouvons sonder le terrain sur le plan technique en minimisant les possibles erreurs d'annotation. La preuve de concept présentée dans les résultats préliminaires du présent document utilise les données textuelles des traductions françaises de l'*Anthologie grecque* en suivant l'édition des Belles Lettres (Waltz, 1931)[6].

La méthode employée pour la classification algorithmique des paires constituant des variations est modélisée en quatre étapes : 1) le nettoyage des données textuelles, 2) la représentation des données, 3) l'implémentation de mesures de similarité ou de distance entre les épigrammes et 4) l'analyse des données à partir d'algorithmes de fouille. Cet assemblage computationnel représente la transformation des épigrammes en objets mathématiques, leurs comparaisons et la recherche de règles formelles permettant d'identifier quelles combinaisons de valeurs de similarité représentent ou non une variation.

### *5.1. Traitement des données textuelles*

Les deux premières étapes du projet consistent à nettoyer les données textuelles et à les transformer en objets mathématiques (soit, des listes séquentielles et des vecteurs, selon les méthodes). Dans la phase exploratoire, toutes les combinaisons possibles de nettoyage des données sont considérées et évaluées individuellement. Pour ce faire, chaque expérience est effectuée avec les options suivantes : 1) avec ou sans application d'un anti-dictionnaire ; 2)

[4] Damerau-Levenshtein est une distance d'édition d'arbre que nous transformons en similarité avec la formule suivante : similarité = 1 − distance/longueurMaximale.

[5] La variation rhétorique n'apparaît pas dans ce groupe. Nous renvoyons le lecteur aux épigrammes VI. 69 et 70 par exemple, où s'opère notamment un changement dans l'énonciation du poème.

[6] Les résultats présentés dans cet article s'attachent donc au corpus traduit. Une phase ultérieure du projet portera sur le grec ancien, mais son traitement (comme la lemmatisation et la racinisation) présente des défis importants, particulièrement sur un corpus aux abondantes fluctuations linguistiques diachroniques et stylistiques.

avec ou sans suppression des marques de ponctuation ; 3) avec ou sans normalisation de la casse ; 4) avec ou sans suppression des accents et 5) racinisation ou lemmatisation ou aucun traitement supplémentaire.

La performance de chaque combinaison peut ainsi être déterminée à partir de la capacité des mesures de similarité produites et des listes de variations répertoriées par les annotateurs. Nous considérons quatre représentations textuelles différentes, soit : 1) sac de mots (binaire) ; 2) sac de mots (pondéré par tf-idf) ; 3) ensembles de bigrammes (n-grammes de termes, n=2) et 4) liste de formes (par ordre d'apparition, avec répétitions des formes récurrentes). Ces représentations isolent des caractéristiques spécifiques aux épigrammes, tout en permettant l'utilisation de mesures de similarité différentes.

### *5.2. Mesures de similarité et fouille de données*

Nous considérons plusieurs mesures de similarité, soit (1) la similarité cosinus, (2) le coefficient de Jaccard et (3) la similarité de Damerau-Levenshtein[7]. Ces mesures ne peuvent pas être utilisées avec chaque représentation textuelle ; par exemple, la similarité de Damerau-Levenshtein ne fonctionne qu'avec une liste ordonnée. Les combinaisons de représentations et de mesures de similarité considérées sont les suivantes : 1) sac de mots (binaire) et similarité cosinus ; 2) sac de mots (binaire) et coefficient de Jaccard ; 3) sac de mots (pondéré tf-idf) et similarité cosinus ; 4) ensemble de bigrammes et coefficient de Jaccard et 5) liste des formes (par ordre d'apparition) et similarité de Damerau-Levenshtein.

Chaque mesure de similarité est testée pour chacune des 48 combinaisons de nettoyage des données textuelles pour déterminer la plus performante avec le coefficient de corrélation de Pearson. Cette étape de l'expérience permet de confirmer ou d'infirmer notre hypothèse selon laquelle les variations sont généralement associées à de plus hautes valeurs de similarité. Nous utilisons comme première variable la mesure de similarité pour chaque paire d'épigrammes et comme seconde variable une valeur [0, 1] où 0 représente une non-variation et 1 représente une variation. Une valeur positive du coefficient de corrélation de Pearson démontrerait alors une tendance entre une haute similarité de la mesure testée et une variation. Si la corrélation statistique permet choisir la permutation de nettoyage des données la plus efficace, elle ne peut pas créer un modèle formel de la variation ; elle ne fait que juger grossièrement de l'efficacité d'une mesure de similarité. Nous considérons trois approches algorithmiques permettant le partitionnement binaire des données, soit (1) le perceptron, (2) la régression logistique et (3) l'arbre de décision.

## 6. Résultats préliminaires

Pour chaque mesure de similarité, les 63 903 paires d'épigrammes du sous-corpus ont été évaluées en fonction des 48 cas de figure de nettoyage des données textuelles avec le coefficient de corrélation de Pearson pour déterminer les combinaisons les plus performantes (numérotées 1 à 5 dans le Tableau 1) pour chaque mesure de similarité.

[7] Le projet – déjà documenté par ailleurs (voir notamment Vitali-Rosati et al., 2020 ; 2021 ; Mellet, 2020) – a donné lieu à la plateforme anthologiagraeca.org(/api), qui rassemble un éventail d'informations sur les 4 134 épigrammes (datant des VI^e^ siècle av. J.-C. au X^e^ siècle apr. J.-C. par plus de 300 auteurs) qui composent l'*Anthologie grecque*. Le nombre de poèmes peut varier selon les éditions : notre API en comptabilise 4 134.

| # | Représentation | Mesure de similarité | Application d'un anti-dict. | Normalisation de la casse | Suppression ponctuation | Suppression accentuation | Traitement supplémentaire |
|---|---|---|---|---|---|---|---|
| 1 | Sac de mots (binaire) | Similarité cosinus | OUI | OUI | NON | OUI | Racinisation |
| 2 | Sac de mots (binaire) | Coefficient de Jaccard | NON | OUI | OUI | NON | Racinisation |
| 3 | Sac de mots (pondéré tf-idf) | Similarité cosinus | OUI | NON | NON | NON | Racinisation |
| 4 | Ensemble de bigrammes | Coefficient de Jaccard | NON | OUI | NON | NON | Racinisation |
| 5 | Liste de formes (par ordre d'apparition) | Similarité de Damerau-Levenshtein | OUI | OUI | OUI | NON | NON |

*Tableau 1. Nettoyages des données textuelles retenus pour chaque mesure de similarité et représentation après évaluation avec la corrélation statistique en fonction des annotations.*

Bien que ces combinaisons de nettoyage des données textuelles soient les plus performantes pour chaque mesure de similarité, ces dernières ne sont pas toutes aussi performantes entre elles. Nous pouvons visualiser l'efficacité des méthodes individuelles en comparant les profils de distribution des similarités entre les variations et les non-variations. La Figure 1 compare ces distributions en fonction du type de variation pour la mesure de similarité #1.

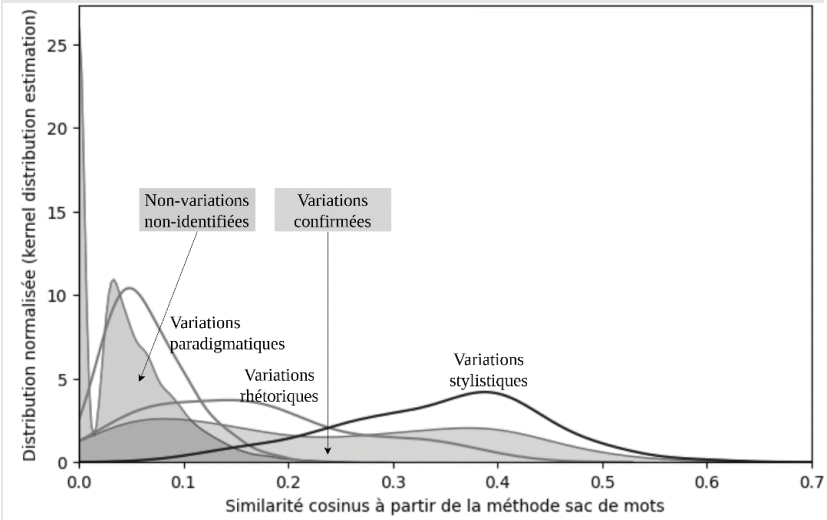


*Figure 1. Distribution des paires d'épigrammes en fonction de la mesure de similarité #1 (cf. Tableau 1), visualisation avec distribution par noyau.*

Dans le cas présenté, ainsi que dans toutes les autres combinaisons testées et retenues, la distribution des variations stylistiques est distincte de la distribution des non-variations (coefficient de Pearson positif). Également, dans tous les cas étudiés, la distribution de similarité des variations paradigmatiques est presque indistinguable de celle des non-variations, et les variations rhétoriques se situent entre les deux. La méthodologie et les approches algorithmiques utilisées jusqu'à présent semblent mieux adaptées à un modèle de la variation stylistique. La Figure 2 montre les coefficients de Pearson pour chaque caractéristique et la cible « variation stylistique » pour laquelle la valeur est 0 pour les non-variations ou 1 pour les variations de ce type. Les corrélations confirment une relation linéaire positive entre la présence d'une variation et une haute similarité.

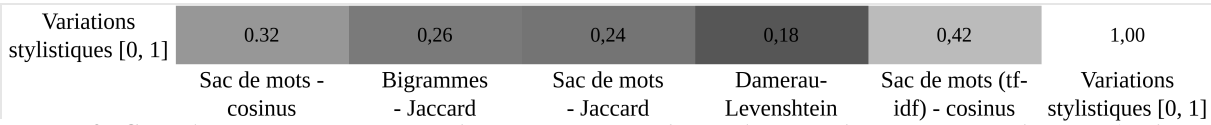


*Figure 2. Corrélation statistique pour les cinq mesures de similarité et la variation stylistique (indiqué par une valeur de 1 pour une variation, et 0 pour une non-variation).*

### *6.1. Modélisation de la variation*

Les trois modèles entraînés avec les données produites par les mesures de similarité sont l'arbre de décision, le perceptron et la régression logistique. L'arbre de décision est la seule méthode permettant d'obtenir une classification parfaite des 63 903 paires d'épigrammes du livre VI. Toutefois, un tel arbre de décision demande 14 niveaux de décisions successifs ce qui constitue deux problèmes majeurs. Le premier est la difficulté à analyser un tel arbre qui complexifie la tâche herméneutique de création d'une définition formelle de la variation. Le second est le fait qu'une telle complexité représente potentiellement un phénomène de surentraînement : les premiers niveaux de l'arbre identifient des tendances générales, tandis que les niveaux inférieurs se rapportent à des cas spécifiques de l'échantillon de données : ces niveaux ne sont pas généralisables. Un tel modèle risque de se montrer inadéquat pour des situations sortant de son ensemble d'entraînement. La Figure 3 montre les trois premiers niveaux de l'arbre de décision qui offre une performance équilibrée supérieure à celle des deux autres méthodes considérées. L'arbre de décision privilégie les variables les plus discriminantes dans ses premiers niveaux, nous identifions ainsi les métriques basées sur la représentation sac de mots comme étant de meilleurs indicateurs de la présence d'une variation stylistique.

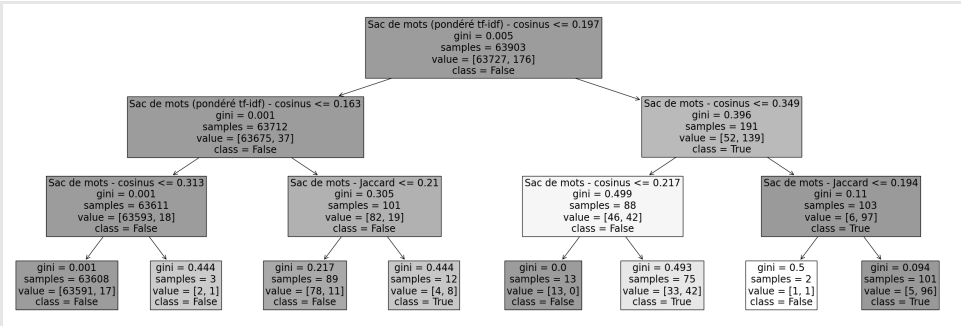


*Figure 3. Les trois premiers niveaux de décision de l'arbre de décision.*

Le Tableau 2 présente les biais, poids et importances pour chaque mesure de similarité du perceptron et de la régression logistique.

| **Métrique de similarité** | | **Perceptron** (poids) | **Régression logistique** (importance) |
|---|---|---|---|
| **Métrique** | **Représentation** | | |
| Cosinus | Sac de mots, pondérée tf-idf | 4.76 | 10.315 |
| Cosinus | Sac de mots, binaire | 2.85 | 12.167 |
| Jaccard | Sac de mots, binaire | 4.22 | 6.298 |
| Jaccard | bigrammes | 1.32 | 2.052 |
| Damerau-Levenshtein | séquentielle | −2.12 | 1.809 |
| Biais | | −3.00 | −8.737 |

*Tableau 2. Poids des mesures de similarité pour le perceptron et la régression logistique.*

Les poids et importances des mesures de similarité 4 et 5 (coefficient de Jaccard avec représentation par bigrammes et la similarité de Damerau-Levenshtein) révèlent que ces métriques sont moins discriminantes que les trois premières. Les mesures de similarité cosinus et le coefficient de Jaccard avec représentation par sac de mots (binaire ou pondéré par tf-idf) sont les plus significatives, ce qui est également le cas dans les niveaux les plus hauts (et donc les plus discriminants) de l'arbre de décision. Ces deux méthodes (avec l'arbre de décision tronqué au troisième niveau, présenté dans la Figure 3) offrent des précisions balancées entre 0.764 et 0.914. Cela démontre le potentiel des méthodes pour arriver à une classification efficace des variations stylistiques.

**Perceptron**

Matrices de confusion

| | | Classe réelle: var. | Classe réelle: non-var. |
|---|---|---|---|
| classe prédite | var. | 101 | 11 |
| classe prédite | non-var. | 75 | 63716 |

Rapports de classification

| | Précision | Rappel | Score F1 | Support |
|---|---|---|---|---|
| var. | 0.90 | 0.57 | 0.70 | 176 |
| non-var. | 1.00 | 1.00 | 1.00 | 63727 |

Précision équilibrée : 0.787

**Arbre de décision (n=3)**

Matrices de confusion

| | | Classe réelle: var. | Classe réelle: non-var. |
|---|---|---|---|
| classe prédite | var. | 146 | 42 |
| classe prédite | non-var. | 30 | 63685 |

Rapports de classification

| | Précision | Rappel | Score F1 | Support |
|---|---|---|---|---|
| var. | 0.78 | 0.83 | 0.80 | 176 |
| non-var. | 1.00 | 1.00 | 1.00 | 63727 |

Précision équilibrée : 0.914

**Régression logistique**

Matrices de confusion

| | | Classe réelle: var. | Classe réelle: non-var. |
|---|---|---|---|
| classe prédite | var. | 93 | 5 |
| classe prédite | non-var. | 83 | 63722 |

Rapports de classification

| | Précision | Rappel | Score F1 | Support |
|---|---|---|---|---|
| var. | 0.96 | 0.53 | 0.68 | 176 |
| non-var. | 1.00 | 1.00 | 1.00 | 63727 |

Précision équilibrée : 0.764

*Tableau 3. Matrices de confusion, rapports de classification et précision équilibrée pour le perceptron, l'arbre de décision (trois premiers niveaux) et la régression logistique.*

### *6.2. Retour aux textes*

Les méthodes utilisées se concentrent essentiellement sur la similarité du vocabulaire utilisé au sein des paires d'épigrammes. Les résultats préliminaires s'avèrent efficaces pour repérer la variation stylistique. Nous notions plus haut que ce type de variation use notamment d'une recherche de synonymes. Nos *vrais positifs* semblent pourtant invalider cette définition *a priori*, qui invoquait une innovation par le lexique : l'apport des méthodes algorithmiques nous pousse plutôt à caractériser la variation stylistique notamment par une reprise élevée de mots entre deux textes. Ensuite, chaque méthode produit quelques dizaines de *faux positifs* qui demandent une attention spéciale. Il est possible d'y trouver des variations qui n'ont pas été répertoriées par les philologues, ou de déterminer d'autres facteurs discriminants des non-variations. Les algorithmes ont également mis en lumière des paires d'épigrammes dont le statut est plus ambigu. L'exemple ci-dessous, qui n'avait pas été repéré initialement par les annotateurs, a un haut score de similarité et provoqua des désaccords quant à son classement – s'agit-il d'une variation stylistique, rhétorique, d'un simple topos ?

*Cette ceinture aux belles franges et, en même temps, ce vêtement, c'est Atthis qui, en relevant de ses couches, les a suspendus, fille de Lêto, au-dessus des portes de ton temple virginal, parce que tu l'as délivrée du fardeau de sa grossesse et que sans douleur elle a mis au monde un enfant vivant.*

*AP*, VI.202 (Léonidas de Tarente)

*Cette ceinture, fille de Lêto, ce vêtement brodé de fleurs et ce soutien-gorge qui enveloppait étroitement ses seins, c'est Timaessa qui te les a consacrés, quand elle eut, après neuf mois, échappé au fardeau pénible d'un douloureux enfantement.*

*AP*, VI.272 (Persès de Thèbes)

Les *faux négatifs* sont plus nombreux : certaines variations stylistiques évidentes pour des lecteurs humains ne sont pas détectées par les algorithmes, notamment à cause d'un usage trop important de synonymes. Revenons au groupe des trois frères, dont il a été question plus haut ; l'épigramme VI.13 apparaît comme un faux négatif (avec les trois algorithmes sélectionnés) en face du l'épigramme VI.11 :

*Le chasseur Damis a consacré ce long panneau, Pigrès ce filet aux fines mailles pour attraper les oiseaux et le rameur de nuit Cleitor ce tramail à mettre les rougets : c'est à toi, Pan, que tous trois ont dédié ces instruments de leur travail ; sois propice à ces frères si pieux, accorde-leur leur provende de volatiles, de venaison et d'habitants des eaux.*

*AP*, VI.11 (Satrius)

Enfin, entre 17 et 43% (cf. Tableau 3) des variations répertoriées échappent encore aux modèles, ce qui indique que d'autres représentations des données textuelles et mesures de similarité doivent être envisagées dans les prochaines étapes du projet pour discriminer ces paires d'épigrammes des non-variations. Comme indiqué plus haut, nous nous sommes principalement intéressés aux variations stylistiques. Dans le cadre des variations rhétoriques et paradigmatiques, la différence de vocabulaire peut s'avérer plus radicale ; pour les premières, les structures narratives priment sur le langage choisi, pour les deuxièmes, le changement du sujet-même de l'épigramme nécessite de recourir à un nouveau champ lexical. Ces résultats portent à croire que nos modèles doivent encore être sophistiqués, et ce, non seulement du point de vue computationnel, mais aussi du point de vue de nos définitions en langage naturel – comme l'illustrent les exemples précédents.

## 7. Conclusion

La première phase du projet IAL offre des résultats encourageants pour la formalisation du concept de variation stylistique dans le corpus de l'*Anthologie grecque*. Nous sommes en mesure de corréler la présence d'une variation stylistique avec la cooccurrence des termes, et plus précisément des fréquences similaires de formes autrement plus rares dans le reste du corpus (modélisé par la pondération tf-idf). En effet, la mesure de similarité cosinus appliquée aux représentations de type sac de mots (pondéré ou non) peut servir d'indicateur potentiel de variations stylistiques. Les variations paradigmatiques et rhétoriques, en revanche, demanderont un travail de modélisation plus important : il sera nécessaire d'élaborer de nouvelles approches algorithmiques pour capturer formellement ces concepts. Concernant la variation stylistique, la prochaine phase du projet consistera à étudier les faux positifs et les faux négatifs pour affiner les mesures de similarité, ou supplémenter ces dernières avec de nouvelles métriques et représentations. À cet égard, nous envisageons implémenter des approches telles que le *Word Mover's Distance* et l'analyse des structures de mots vides. Ces développements visent à étendre nos expérimentations à l'ensemble de l'*Anthologie grecque*, tout en adaptant nos méthodes computationnelles au grec ancien, langue qui pose d'importantes difficultés de traitement. L'intégration de modèles computationnels à l'étude littéraire et philologique de l'*Anthologie grecque* représente une nouvelle perspective sur ce corpus, et un exemple concret de collaboration herméneutique humain-machine.

## Bibliographie

Aubreton R. (1968). La tradition manuscrite des épigrammes de l'Anthologie palatine. *Revue des Études Anciennes*, 70 (1-2), 32-82.

Da N.Z. (2019). The computational case against computational literary studies. *Critical Inquiry*, 45 (3), 601-639.

Coffee N., Koenig J.-P., Poornima S., Forstall C.W., Ossewaarde R. et Jacobson S.L. (2012). The Tesserae Project: intertextual analysis of Latin poetry. *Literary and Linguistic Computing*, 28 (2), 221–228.

Gutzwiller K.J. (1998). *Poetic Garlands: Hellenistic Epigrams in Context*. Los Angeles-Londres : University of California Press.

Laurens P. (2012). *L'abeille dans l'ambre : Célébration de l'épigramme de l'époque alexandrine à la fin de la Renaissance* (2e édition). Les Belles Lettres.

Marche S. (2012). Literature is not data: Against digital humanities. *LA Review of Books*, 28.

Mellet M. (2020). Penser le palimpseste numérique. Le projet d'édition numérique collaborative de l'Anthologie palatine. *Captures*, 5 (1).

Meunier J.-G. (2017). Humanités numériques et modélisation scientifique. *Questions de communication,* 31, 19-48.

Pöckelmann M., Dähne J., Ritter J. et Molitor P. (2020). Fast Paraphrase Extraction in Ancient Greek Literature. *It - Information Technology*, 62 (2), 75-89.

Piper A. (2017). Think small: on literary modeling. *PMLA*, 132 (3), 651-658.

Pöckelmann M., Dähne J., Ritter J. et Molitor P. (2020). Fast paraphrase extraction in ancient greek literature. *It - Information Technology*, 62 (2), 75–89.

Ramsay S. (2011). *Reading Machines: Toward and Algorithmic Criticism*. University of Illinois Press.

Rhody L. M. (2012). Topic Model Data for Topic Modeling and Figurative Language. *Journal of Digital Humanities*, 2 (1).

Roberts C. W. (2000). A conceptual framework for quantitative text analysis. *Quality and Quantity*, 34, 259-274.

Rockwell G. et Sinclair S. (2016). *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. MIT.

Schubert Ch. (2020). Intertextuality and Digital Humanities. *It - Information Technology*, 62 (2), 53-59.

Searle J.R. (1980). Minds, brains, and programs. *The Behavioral and Brain Science*, 3, 417-457.

Snow C.P. (1959). *The two cultures*. Oxford University Press.

Tarán S.L. (1979). *The Art of Variation in the Hellenistic Epigram*, vol. 9. Brill.

Vitali-Rosati M. Monjour S., Casenave J., Bouchard E. et Mellet M. (2020). Editorializing the Greek Anthology: The palatin manuscript as a collective imaginary. *Digital Humanities Quarterly*, 14 (1).

Vitali-Rosati M., Mellet M., Monjour S., Fauchié A., Guicherd T., Larlet D. et Agostini-Marchese E. (2021). L'épopée numérique de l'Anthologie grecque : entre questions épistémologiques, modèles techniques et dynamiques collaboratives. *Sens public*.

Waltz P. (1931). *Anthologie grecque (libre VI),* t. 3. Les Belles Lettres.