Des algorithmes pour définir des concepts littéraires ?

Mathilde Verstraete, Yann Audin, Dominic Forest, Marcello Vitali-Rosati

▶ **To cite this version:**

Mathilde Verstraete, Yann Audin, Dominic Forest, Marcello Vitali-Rosati. Des algorithmes pour définir des concepts littéraires ? : IAL, la variation, et l'Anthologie grecque. Humanistica 2024, Association francophone des humanités numériques, May 2024, Meknès, Maroc. hal-04563545

HAL Id: hal-04563545
https://hal.science/hal-04563545

Submitted on 29 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Des algorithmes pour définir des concepts littéraires ? IAL, la variation, et l'*Anthologie grecque*

Mathilde Verstraete, Yann Audin, Dominic Forest, Marcello Vitali-Rosati
Université de Montréal
{mathilde.verstraete,yann.audin,dominic.forest,marcello.vitali.rosati}@umontreal.ca

Résumé

Le projet *Intelligence Artificielle littéraire* (IAL) propose de formaliser des concepts littéraires selon des principes computationnels et algorithmiques. Notre corpus est constitué de l'*Anthologie grecque* et le concept étudié, particulièrement abondant dans ce recueil, est celui de la variation. Cette contribution fait la synthèse des expérimentations exploratoires menées sur un premier échantillon (traduction française du livre VI). Nous y analysons également les résultats préliminaires de cette recherche, faisons état de la méthodologie employée, et posons les bases pour la suite de ce projet pilote.

1 Introduction

Est-il possible de donner une définition formelle, non ambiguë d'un concept littéraire ? Est-il possible d'implémenter ladite définition dans un algorithme qui réussisse à la saisir ? [1] En effet, si l'on peut donner une définition formelle d'un concept littéraire et l'implémenter dans un algorithme, cet algorithme sera ensuite capable de *saisir* le concept et donc, d'une certaine manière, de le comprendre. Dans une démarche de ce type, il ne s'agit pas d'utiliser des algorithmes à des fins heuristiques, mais de les construire pour vérifier si notre définition formelle d'un concept donné est assez bien formulée de sorte que l'algorithme parvienne à l'identifier.

Le projet *Intelligence artificielle littéraire* (IAL) a l'ambition de vérifier la faisabilité d'une telle approche en la testant avec un concept littéraire et un corpus spécifiques : celui de « variation » au sein de l'*Anthologie grecque*. Parmi les épigrammes de ce corpus, plusieurs sont des variations d'autres épigrammes : reprises, réécritures, reformulations, résultat d'un « processus d'engendrement d'une épigramme par une autre » (Laurens, 2012, p. 120). Après avoir identifié ces variations, nous tentons d'une part, de donner une définition formelle de ce concept, d'autre part, de l'implémenter dans des algorithmes. Si nos algorithmes sont capables de repérer toutes les variations précédemment identifiées comme telles – sans faux positifs –, alors notre algorithme incarnera notre définition du concept de variation.

La recherche d'intertextualité dans les corpus anciens a déjà fait couler beaucoup d'encre (cf. notamment Coffee et al., 2012; Schubert, 2020; Pöckelmann et al., 2020). Cependant, à la différence de ces travaux, notre objectif initial n'est pas de repérer de nouveaux exemples d'intertextualité [2], mais plutôt de formaliser *une* définition d'un type spécifique d'intertextualité (la variation) pour lequel nous en avons déjà répertorié les occurrences. Seront donc privilégiées des approches computationnelles transparentes et simples et dont il est possible de comprendre le modèle épistémologique. Des approches opaques, même performantes – comme celles des LLMs –, ne nous permettraient pas d'approcher le but de ce projet, soit de formuler une définition formelle du concept de variation.

2 Le corpus

IAL est issu des acquis d'un précédent projet mené par la Chaire de recherche du Canada sur les Écritures numériques : l'édition numérique collaborative de l'*Anthologie grecque* (*AG*) [3].

Le projet *AG* rassemble un large éventail d'informations sur 4 134 épigrammes [4] (écrites entre le VI^e^ siècle avant J.-C. et le X^e^ siècle après par plus de 300 auteurs) qui constituent l'*Anthologie*

1. Ces questions peuvent être interprétées comme une expression particulière de la question plus générale : « une machine peut-elle penser ? », formulée dans le célèbre article de Turing (1950).

2. Les résultats fournis par les algorithmes nous ont toutefois fait découvrir des variations qui avaient échappé à notre vigilance toute humaine.

3. La plateforme du projet est disponible à l'adresse anthologiagraeca.org.

4. Le nombre d'épigrammes peut varier selon les éditions ; notre API (anthologiagraeca.org/api) en compte 4 134.

grecque. Ce travail éditorial, toujours en cours, est réalisé par de multiples éditeurs et a été documenté en plusieurs endroits (voir notamment Mellet et Verstraete, 2024, Vitali-Rosati et al., 2021, 2020, Mellet, 2020, Dumouchel, 2018). De ce projet découle un corpus complet et hautement structuré : un terrain de jeu idéal pour initier de nouvelles expériences littéraires computationnelles.

3 La variation dans l'*Anthologie grecque*

L'*Anthologie grecque* constitue un corpus diversifié de formes intertextuelles parmi lesquelles on retrouve celle de la variation (Tarán, 1979), une forme largement encouragée par les pratiques rhétoriques de l'époque. Celle-ci se manifeste par la reprise et l'adaptation d'un texte provenant d'un prédécesseur ou d'un contemporain. Le genre de l'épigramme se prête particulièrement bien à cette pratique poétique, puisque la simplicité de sa forme permet aux auteurs de briller en quelques vers seulement.

Pierre Laurens, auteur d'un volume conséquent dédié au genre épigrammatique, identifie trois types de variation (Laurens, 2012, p. 117-130). La variation *stylistique* concerne d'abord et avant tout les mots et leur agencement, introduisant des innovations par la modification du lexique et du style. La variation *rhétorique*, quant à elle, porte sur la forme générale des épigrammes, modifiant l'agencement des phrases ; « l'impression est celle d'une multiplication à l'infini des possibilités d'expression d'une même idée » (Laurens, 2012, p. 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, p. 130).

4 Objectifs et méthodologie

Comme explicité dans l'introduction, nous avons pour objectif de produire un modèle formel du concept de variation puis de l'implémenter dans un ou plusieurs algorithmes avec l'hypothèse que si cet ensemble d'algorithmes est capable d'identifier les variations précédemment annotées comme telles, cet algorithme représentera une définition formelle de la variation [5]. Nous avons dès lors commencé par annoter les variations dans un sous-ensemble du corpus, à savoir la traduction française du livre VI (Waltz, 1931), contenant 359 épigrammes avec plus de 300 variations répertoriées par des annotateur·ice·s [6]. Le choix de ce sous-corpus nous permet de sonder le terrain sur le plan technique, d'étudier un volume de textes raisonnable pour un lecteur humain et de travailler dans un premier temps sur une langue plus facile à traiter avec des librairies et les méthodes standardisées.

Nous définissons la variation comme une relation entre deux épigrammes, et cherchons donc une méthode ou une combinaison de méthodes algorithmiques qui permette d'identifier une variation (et son type) d'une non-variation. Au présent stade du projet, IAL se concentre sur les représentations textuelles simples comme la vectorisation de type sac de mots (pondéré avec tf-idf (Salton et McGill, 1983) ou binaire), les ensembles de n-grammes ($n = 2$ termes), et les listes ordonnées. Les paires d'épigrammes sont comparées à l'aide de mesures de distance [7] et de similarité, soit la similarité cosinus, l'indice de Jaccard, et la distance de Damerau-Levenshtein (Damerau, 1964; Levenshtein, 1966). Chaque combinaison de représentation textuelle et de mesure de similarité explorée est testée sur 48 combinaisons de nettoyage des données textuelles :

- avec ou sans normalisation de la casse ;
- avec ou sans suppression des marques de ponctuation ;
- avec ou sans l'application d'un anti-dictionnaire ;
- avec ou sans suppression des accents ;
- racinisation, lemmatisation, ou aucun traitement supplémentaire.

Les mesures de similarité employées sont comparées avec les variations recensées par les annotateur·ice·s dans une base de données dans laquelle une variation entre deux épigrammes est représentée par une valeur de 1, et une non-variation par une valeur de 0. La corrélation statistique permet d'identifier quel nettoyage de données convient le mieux à chaque méthode, et des algorithmes de découverte de connaissance dans les bases de données

5. Les fondements théoriques et conceptuels du projet ont été exposés plus en détail lors d'une communication précédente (Audin et al., 2023).

6. L'annotation a été réalisée par des annotateur·ice·s ayant une connaissance solide du corpus – il leur a été demandé de parcourir l'édition de Waltz (1931) et d'y repérer les variations. Notons que Waltz indique souvent des renvois à d'autres épigrammes, sans qu'il ne s'agisse forcément de variations (il parle, entre autres, d'imitation, de reprise, de variation, de doublet).

7. Dans chaque cas où une mesure de distance est utilisée, celle-ci est convertie en mesure de similarité pour tenter de corréler une haute similarité avec une variation.

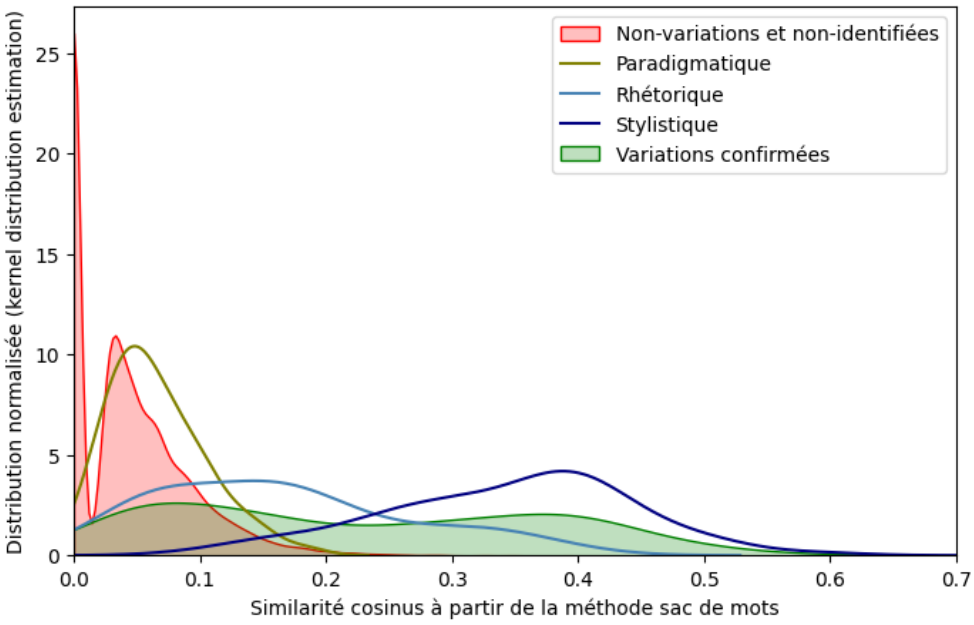

FIGURE 1 – Distribution des mesures de similarité cosinus pour la représentation sac de mots binaire entre paires d'épigrammes. Les courbes de distribution normalisées (estimation par noyau) permettent d'estimer la densité de probabilité qu'une paire d'épigrammes d'un type précis (variation, non-variation, variation stylistique, paradigmatique et rhétorique) ait une similarité donnée par l'axe x, et permettent de comparer des populations (de paires d'épigrammes) de tailles différentes sur une même échelle.

permettent l'analyse multivariable des mesures de similarité. Nous nous limitons dans les choix de ces algorithmes aux méthodes qu'il nous est possible d'analyser afin de contribuer à l'élaboration d'une définition formelle de la variation. Ainsi, les méthodes connexionnistes et les grands modèles de langue sont à éviter : ces derniers sont opaques, et nous ne pouvons dériver des définitions de ces modèles qui sont souvent considérés comme des boîtes noires (Carabantes, 2020).

5 Résultats préliminaires

La première étape, liée à la paramétrisation du nettoyage des données textuelles à l'aide de la corrélation statistique, a démontré que l'étude des termes était prometteuse pour l'identification de la variation stylistique, mais peu concluante pour l'identification des variations rhétoriques et paradigmatiques. En effet, les mesures de similarité proposées, combinées aux représentations des données textuelles basées sur les fréquences et occurrences des termes, révèlent des similarités en moyenne plus hautes pour les variations stylistiques que pour les autres types de variations, et, de manière plus importante, que pour les non-variations. La **Figure 1** illustre cette première étape pour l'un des cas de figure, et met en évidence le potentiel de la méthode dans l'identification des variations stylistiques par rapport aux deux autres types.

La distribution des variations stylistiques se distingue de la distribution des non-variations pour chaque combinaison de nettoyage des données, représentation des données textuelles, et mesure de distance. Ainsi, à ce stade, nous mettons de côté les deux autres types de variation qui ne sont pas bien prédites par le modèle de calcul de distance et similarité. Après évaluation de chaque mesure de distance (avec la corrélation stylistique), nous avons privilégié cinq métriques, basées sur la combinaison des modes de nettoyage des données textuelles, des modes de représentations et des métriques de similarité :

— Similarité cosinus, représentation sac de mots, application d'un anti-dictionnaire, normalisation de la casse, suppression des accents et racinisation ;
— Coefficient de Jaccard, ensemble de 2-grammes (termes consécutifs), normalisation de la casse et racinisation ;
— Coefficient de Jaccard, représentation sac de mots binaire, normalisation de la casse, suppression des marques de ponctuation et des accents et racinisation ;

— Distance de Damerau-Levenshtein [8], liste ordonnée des termes, application d'un anti-dictionnaire, suppression de la ponctuation et normalisation de la casse ;
— Similarité cosinus, représentation sac de mots pondérée avec tf-idf, application d'un anti-dictionnaire et racinisation.

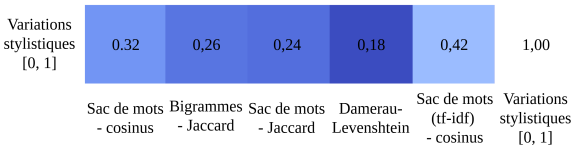

FIGURE 2 – Corrélation statistique entre les variations stylistiques attestées et les différentes mesures de similarité calculées entre les paires d'épigrammes

Ces approches ont été appliquées à chaque paire d'épigrammes du corpus, et évaluées à l'aide de la corrélation statistique. Les résultats de similarité générés sont enregistrés dans une base de données qu'il nous est possible d'analyser avec des méthodes transparentes.

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

FIGURE 3 – Poids et importances des métriques de similarité pour le perceptron et la régression logistique

Les valeurs de corrélation positives dans la **figure 2** illustrent que la similarité des représentations textuelles utilisées est corrélée avec la présence d'une variation stylistique. Nous utilisons trois algorithmes pour réaliser une tâche de classifications binaires, soit la régression logistique, le perceptron et l'arbre de décision. La **figure 3** contient les poids et importance des métriques de similarité, et permet d'identifier trois d'entre elles comme étant plus performantes dans la prédiction des variations stylistiques. Ces dernières sont la similarité cosinus avec représentation sac de mots binaire, la similarité cosinus avec représentation sac de mots pondéré, et le coefficient de Jaccard avec représentation sac de mots binaire.

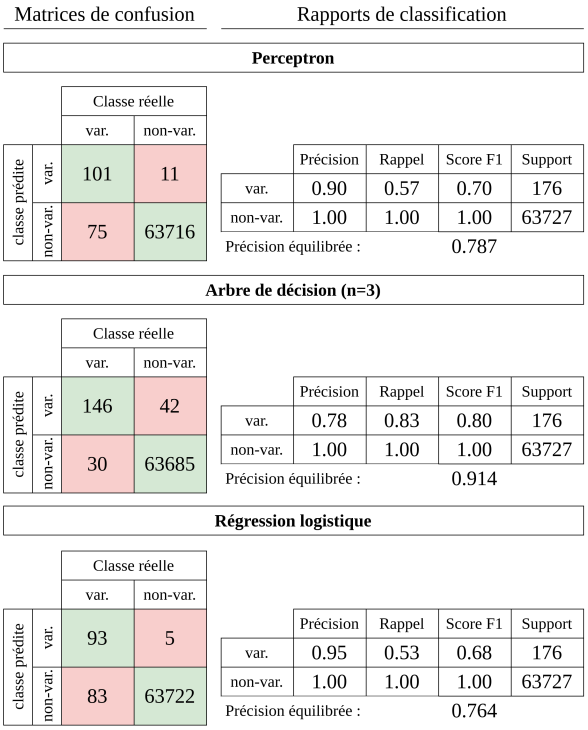

FIGURE 5 – Matrice de confusion, rapports de classification et précision équilibrée

Quant à l'arbre de décision, dont nous présentons les trois premiers niveaux dans la **Figure 4**, il permet d'obtenir des résultats parfaits (classification correcte de tous les cas de figure) lorsqu'il est composé de 14 niveaux. Cela semble mettre en lumière une limite de cette approche : les règles générées par l'arbre de décision semblent difficilement généralisables et donc moins performantes pour classifier des épigrammes qui n'ont pas été employées dans l'ensemble d'apprentissage. De plus, un arbre de décision de 14 niveaux est malaisément interprétable et donc impossible à traduire en définition formelle. Les trois plus hauts niveaux de l'arbre confirment toutefois que les deux métriques basées sur le cosinus sont plus discriminantes que les trois autres. La **Figure 5** montre la matrice de confusion, ainsi que le rapport de classification et la précision équilibrée pour les trois méthodes de classification utilisée, soit le perceptron, l'arbre de décision (pour les trois niveaux de la **Figure 4**) et la régression logistique.

Pour les méthodes utilisées, 20 à 50% des variations répertoriées échappent aux méthodes, et

8. Convertie en similarité en utilisant

$$1 - \frac{distanceDamerauLevenshtein}{longueurMaximale}.$$

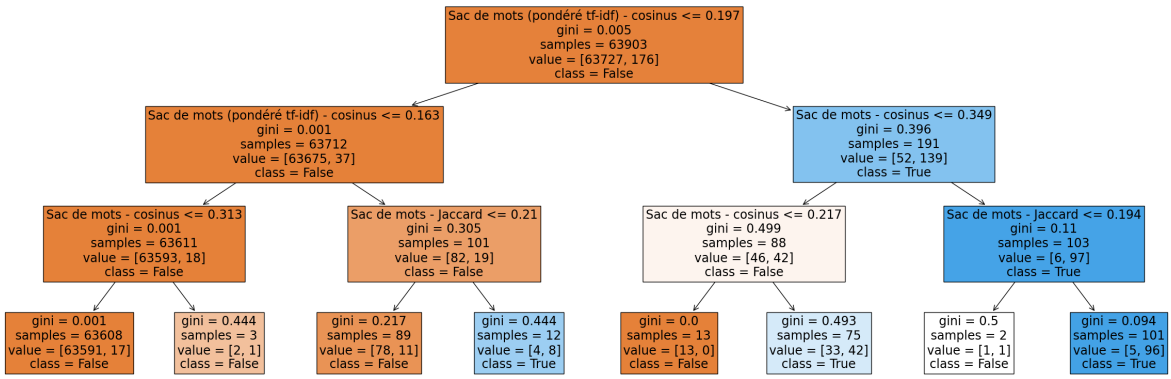

FIGURE 4 – Arbre de décision (profondeur de 3 niveaux pour en permettre la lisibilité et en faciliter l'analyse), 14 niveaux sont nécessaires pour obtenir des résultats optimaux

plusieurs non-variations sont mal classifiées [9].

6 Conclusion et travaux futurs

Cette contribution présente les objectifs du projet et fait état de nos expériences exploratoires et de leurs premiers résultats. Cela nous indique que la similarité cosinus des représentations de type sac de mots (pondéré ou binaire) est un possible indicateur de variation stylistique, et que ce type de représentation performe également avec le coefficient de Jaccard. L'étude des termes et de leurs cooccurrences est une avenue prometteuse pour la modélisation du concept de variation stylistique. La précision équilibrée des trois méthodes (entre 0.76 et 0.91) nous encourage à tester de nouvelles mesures de similarité et représentations : la définition formelle de la variation stylistique nécessite des itérations successives. La suite de nos travaux consistera à implémenter de nouvelles méthodes que nous appliquerons pour l'identification de variations tant stylistiques que rhétoriques et paradigmatiques. Enfin, nous étendrons nos résultats à l'ensemble de l'*Anthologie grecque* et adapterons nos méthodes computationnelles au grec ancien, tout en tenant compte des diverses spécificités du corpus.

Bibliographie

Yann Audin, Mathilde Verstraete, et Marcello Vitali-Rosati. 2023. Intelligence Artificielle Littéraire. In *Humanistica 2023*, Épistémologie, Genève, Suisse. Association francophone des humanités numériques.

Manuel Carabantes. 2020. Black-box artificial intelligence: an epistemological and critical analysis. *AI & SOCIETY*, 35(2) :309–317.

Neil Coffee, Jean-Pierre Koenig, Shakthi Poornima, Christopher W. Forstall, Roelant Ossewaarde, et Sarah L. Jacobson. 2012. The Tesserae Project: intertextual analysis of Latin poetry. *Literary and Linguistic Computing*, 28(2) :221–228.

Frederick J Damerau. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3) :171–176.

Suzanne Dumouchel. 2018. Séance 2 EPHN : Marcello Vitali Rosati, Université de Montréal : « Une API pour l'Anthologie grecque: repenser le codex Palatinus à l'époque du numérique ».

Pierre Laurens. 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, Paris.

Vladimir Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10(8) :707–710.

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

Margot Mellet et Mathilde Verstraete. 2024. Passés et présents anthologiques. Le projet d'édition numérique collaborative de l'Anthologie grecque. In *Communautés et pratiques d'écritures des patrimoines et des mémoires*, presses universitaire de paris nanterre édition, Intelligences numériques. Paris. Publication.

Marcus Pöckelmann, Janis Dähne, Jörg Ritter, et Paul Molitor. 2020. Fast paraphrase extraction in ancient greek literature. *it-Information Technology*, 62(2) :75–89.

Gerard Salton et Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill.

Charlotte Schubert. 2020. Intertextuality and digital humanities. *it-Information Technology*, 62(2) :53–59.

9. Il n'est pas à exclure que les faux positifs puissent contenir des variations non relevées par les philologues. La deuxième phase du projet prévoit étendre l'annotation de notre corpus, afin de limiter les erreurs et la subjectivité des annotateurs.

Sonya lidia Tarán. 1979. *The Art of Variation in the Hellenistic Epigram*, volume IX of *Columbia studies in the classical tradition*. E. J. Brill, Leyde.

Alan M. Turing. 1950. Computing Machinery and Intelligence. *Mind*, LIX(236) :433–460.

Marcello Vitali-Rosati, Margot Mellet, Servanne Monjour, Antoine Fauchié, Timothée Guicherd, David Larlet, et Enrico Agostini-Marchese. 2021. L'épopée numérique de l'Anthologie grecque : entre questions épistémologiques, modèles techniques et dynamiques collaboratives. *Sens public*.

Marcello Vitali-Rosati, Servanne Monjour, Joana Casenave, Elsa Bouchard, et Margot Mellet. 2020. Editorializing the Greek Anthology: The palatin manuscript as a collective imaginary. *Digital Humanities Quarterly*, 014(1).

Pierre Waltz. 1931. *Anthologie grecque. Anthologie Palatine*, volume III (livre VI) of *Collection des Universités de France*. Les Belles Lettres, Paris.