# Intelligence Artificielle Littéraire

Yann Audin, Mathilde Verstraete, Marcello Vitali-Rosati

# *Intelligence Artificielle Littéraire*: dialogues entre philologues et algorithmes

**Yann Audin, Mathilde Verstraete, Marcello Vitali-Rosati**
Université de Montréal
Chaire de recherche du Canada sur les écritures numériques
{yann.audin,mathilde.verstraete,marcello.vitali.rosati}@umontreal.ca

## Résumé

Cette contribution interroge la possibilité d'une définition formelle – ou, plus précisément, computationnelle et algorithmique – d'un concept littéraire. La Chaire de recherche du Canada sur les écritures numériques (CRCEN) mène depuis 2014 et 2021 les projets d'édition numérique collaborative de l'*Anthologie grecque* et d'*Intelligence artificielle littéraire* (IAL), dérivé du premier. Ces projets ont en commun qu'ils déplacent la création de connaissance en dehors des chercheur·e·s uniques en faisant de la collaboration le lieu de production de sens. Cette contribution fait état de la méthodologie du projet et des premiers résultats, tout en explorant les opportunités nées de telles collaborations pour la théorie littéraire et l'édition.

Peut-on définir de manière formelle – ou, plus précisément, de manière computationnelle et algorithmique – un concept littéraire ? En 2012, dans son article *Literature is not Data*, Stéphane Marche se positionnait clairement en faveur des *deux cultures* (les sciences humaines/sociales s'opposant aux sciences naturelles, voir Snow 1959). Il 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). Cette division des domaines d'études, née d'une réaction romantique à la montée des sciences et du scientisme (Binder, 2020) mena à une présupposition qui marqua plus d'un siècle de recherche : celle de l'incompatibilité entre les langages naturels et formels.

Comme d'autres initiatives, le projet *Intelligence Artificielle Littéraire* (IAL) se positionne à contre-courant de cette orthodoxie et cherche à tester cette supposée incompatibilité et approche la littérature comme un ensemble de données complexes. Pour rendre compte de cette complexité, il est nécessaire de développer des modèles formels riches. IAL s'appuie sur le projet d'édition (numérique et collaborative) de l'*Anthologie grecque* [1] développé à la Chaire de Recherche du Canada sur les Écritures Numériques (CRCEN) depuis 2014 et l'utilise comme corpus. L'objectif de IAL n'est pas d'utiliser les développements algorithmiques à des fins heuristiques ou au profit de l'analyse littéraire, mais plutôt de mettre la computation au service de la théorie littéraire.

## 1 Corpus étudié – l'*Anthologie grecque*

Dès les années 1970 fleurissent des projets de numérisation, d'édition et de valorisation de textes classiques. La CRCEN s'inscrit dans cette tradition par le développement de la plateforme. Monument de la littérature classique, l'*Anthologie grecque* recueille la majorité de la poésie épigrammatique grecque, des périodes classique à byzantine (Cameron, 1993; Gutzwiller, 1997, 1998). L'*Anthologie grecque* se distingue de l'*Anthologie palatine* – nom donné au *codex palatinus graecus 23* (Waltz, 1929) – par son absence de frontière : cette appellation correspond au regroupement du manuscrit palatin et de l'*Anthologie de Planude* (une autre compilation, datant du XIII$^{e}$ siècle). En ce sens, elle est vouée à être amplifiée et refuse toute clôture. Pour ces raisons, le projet se positionne en continuation de ce processus d'anthologisation de l'épigrammatique grecque.

Grâce au travail de nombreux·ses contributeur·ice·s, la plateforme d'édition dédiée au projet présente – pour chaque épigramme – son emplacement dans le manuscrit palatin (récupéré à partir de l'API de l'outil d'annotation du manuscrit de la Bibliothèque de Heidelberg grâce au protocole IIIF), plusieurs traductions en diverses langues, des informations concernant son auteur et ses thématiques (via des mots-clés suivants les standards du *Web* des données ouvertes et liées), des commentaires, des références internes (variations) et

1. https://anthologiagraeca.org.

externes. L'*Anthologie grecque* constitue un corpus précieux et diversifié de formes intertextuelles, dont celle de la variation. Plusieurs des auteurs de l'*Anthologie* s'inspirent en effet les uns des autres (Waltz, 1929). La *variatio*, forme très spécifique d'intertextualité commune et prisée dans la littérature grecque, consiste à reprendre une pièce d'un autre auteur et de la réécrire avec des variations stylistiques, rhétoriques ou paradigmatiques (Laurens, 2012). Le procédé était particulièrement apprécié des épigrammatistes : la simplicité de la forme permettait aux auteurs de s'illustrer en l'espace de quelques vers. Ainsi, l'épigramme est souvent décrite comme un art de la variation en tant que tel (Tarán, 1979). La nature même du corpus anthologique (la réunion de poèmes hétéroclites néanmoins reliés par des *topoï* communs) en fait une source d'intertextualité inépuisable. Certains thèmes reviennent particulièrement fréquemment, et les épigrammes se répondent, parfois avec plusieurs siècles de décalage (Gutzwiller, 1998). Cette forme, encouragée par les pratiques rhétoriques de l'époque, commande la production de la littérature grecque et son évolution (Laurens, 2012).

Si ces variations sont nombreuses et semblent évidentes au lecteur ou à la lectrice humain·e, il apparaît plus complexe de s'arrêter sur une définition claire et précise. À partir de la plateforme d'édition, du modèle de données, et des données produites, nous avons initié, en 2021, le projet (IAL), interrogeant la possibilité de créer des algorithmes qui nous conforteraient dans l'établissement de définitions formelles de concepts littéraires. Une première expérience se construit donc autour du corpus de l'*Anthologie grecque*, et du concept de la *variatio*.

## 2 IAL - Technique et pensée

La problématique au cœur du projet IAL adapte une question fondamentale de la recherche informatique à la littérature : celle du rapport entre l'intelligence humaine et l'intelligence artificielle. L'aspect de cette question qui nous intéresse ici est la définition de concepts littéraires ; une tâche qui, dans le paradigme des *deux cultures* (Snow, 1959; Binder, 2020), tombe clairement dans le domaine de l'humain et du langage naturel. En cherchant à produire une définition mathématique, algorithmique, du concept de *variatio* au sein de l'*Anthologie grecque*, IAL est un pont entre le langage naturel et le langage formel. Pour tester la possibilité de la formalisation d'un concept littéraire, IAL cherche à découvrir une équivalence entre un modèle algorithmique et une définition. Le concept de *variatio* dans le corpus de l'*Anthologie grecque* a l'avantage de présenter un défi définitionnel, tout en étant situé dans un cadre fixe et limité : la modélisation est donc possible à partir de données complètes.

Du point de vue de la méthode, IAL s'articule autour de deux expériences, une première de modélisation textuelle (Piper, 2017) et une autre d'extraction de connaissances dans les bases de données (Fayyad et al., 1996). Puisque le concept de variation mobilise toujours au moins deux épigrammes, nous pouvons appliquer les notions de distance et de similarité textuelle, induisant la question des caractéristiques ou modèles utilisés pour générer ces distances. Plusieurs méthodes ont ainsi été testées sur les épigrammes du corpus, en commençant par des méthodes simples comme celle du sac de mot (*bags of words*) et le TF-IDF (*term frequency-inverse document frequency*). Après plusieurs expériences et adaptations des paramètres, ces deux techniques ont fini par fournir des résultats relativement convenables, mais loin d'être satisfaisants. Des méthodes de *Word Mover's Distance* ont également été testées (Kusner et al., 2015; Pöckelmann et al., 2020; Schubert, 2020), mais les résultats s'avérèrent plus mitigés, notamment à cause des difficultés intrinsèques à notre corpus (Bamman et Crane, 2011; Pantella, 2009) : faible volume de textes (et de données d'entraînement) ; vocabulaire riche et particulièrement mouvant (selon le style, l'époque et le lieu) ; questions liées à la lemmatisation ; etc. Des méthodes plus simples, comme les distances de Levenshtein (Levenshtein, 1966) et de Damerau-Levenshtein (Damerau, 1964) produisirent quant à elles des résultats prometteurs.

Plusieurs questions sont centrales dans l'application de ces algorithmes, les expériences avec la distance de Damereau-Levenshtein montrent par exemple de bien meilleurs résultats lorsque les mots vides sont conservés. Similairement, *Word Mover's Distance* et les modèles linguistiques en grec ancien sont plus efficaces lorsque le corpus n'est ni lemmatisé, ni racinisé. Ces enjeux sont importants pour la modélisation des variations – le modèle devant être capable d'identifier correctement ces dernières –, mais aussi pour la production d'une définition (ou au moins d'une définition efficace) de la *variatio*. En effet, la conceptualisation d'une variation passera par l'analyse de chaque

composante de la modélisation, et les différentes formes de pré-traitement font partie de ce processus. Les caractéristiques du grec ancien, de même que les limites formelles de la théorie littéraire, forcent des interactions entre les programmeurs et spécialistes des méthodes d'analyse numérique d'un côté, et les philologues, littéraires et linguistes de l'autre.

La modélisation du concept littéraire de variation dans l'*Anthologie grecque* passe par la comparaison entre les métriques de similarité utilisées et les variations découvertes par une équipe de philologues et d'étudiant·e·s en langues classiques. Ainsi, des algorithmes de découverte de connaissance dans les bases de données permettront de créer un modèle qui sera ensuite analysé pour formuler une définition formelle de la variation. Pour cette raison, des algorithmes transparents (pleinement explicables) sont d'abord considérés, ce qui met le projet en opposition à la *doxa* prédominante d'une utilisation uniquement applicative et heuristique des réseaux de neurones et des modèles opaques. IAL fait le pari que certains concepts littéraires simples peuvent être modélisé, mais ces démarches ouvrent deux autres questions de nature épistémologique. D'abord, l'équivalence entre une définition en langage naturel et un modèle est-elle valide ou accidentelle ? Ensuite, quelles sont les limites de la formalisation des définitions littéraires ?

## 3 Conclusion

Le projet n'étant qu'à ses prémisses, seules quelques méthodes algorithmiques ont été testées. Les perspectives qui nous semblent les plus prometteuses résulteraient d'une combinaison de différentes méthodes algorithmiques alliée à des expérimentations en essais-erreurs selon des paramétrages précis. Ce processus itératif de raffinement du modèle coïncide avec la formalisation plus précise d'une définition du concept de *variatio* dans notre corpus. La traduction de résultats algorithmiques littéraires en langage naturel pourrait mener à un important bond conceptuel quant à notre compréhension du concept de variation, de l'*Anthologie grecque*, et, à plus grande échelle, d'autres concepts littéraires au sein d'autres corpus.

## Bibliographie


David Bamman et Gregory Crane. 2011. Structured knowledge for low-resource languages: The latin and ancient greek dependency treebanks. In Caroline Sporleder, Antal van den Bosch, et Kalliopi Zervanou, éditeurs, *Language Technology for Cultural Heritage*, Foundations of Human Language Processing and Technology, chapitre 10. Springer.

Jeffrey M. Binder. 2020. Romantic disciplinarity and the rise of the algorithm. *Critical Inquiry*, 46(4).

Alan Cameron. 1993. *The Greek Anthology. From Meleager to Planudes*. Oxford University Press, Oxford. Publisher : Clarendon Press.

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

Usama Fayyad, Gregory Piatetsky-Shapiro, et Padhraic Smyth. 1996. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3) :37.

Kathryn Gutzwiller. 1997. The Poetics of Editing in Meleager's Garland. *Transactions of the American Philological Association (1974-)*, 127 :169–200. Publisher : [Johns Hopkins University Press, American Philological Association].

Kathryn J Gutzwiller. 1998. *Poetic Garlands : Hellenistic epigrams in context*. University of California Press, Los Angeles/Londres.

Matt Kusner, Yu Sun, Nicholas Kolkin, et Kilian Weinberger. 2015. From Word Embeddings To Document Distances. *Proceedings of Machine Learning Research*, 37 :957–966.

Pierre Laurens. 2012. *L'abeille dans l'ambre : Célébration de l'épigramme de l'époque alexandrine à la fin de la Renaissance*. Les Belles Lettres.

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

Stephen Marche. 2012. Literature is not Data: Against Digital Humanities. Section : essay.

Jean-Guy Meunier. 2017. Humanités numériques et modélisation scientifique. *Digital Humanities and Scientific Modelling*, pages 19–48.

Maria Pantella. 2009. Thesaurus linguae graecae. Accessed December, 2022.

Andrew Piper. 2017. Think small: On literary modeling. *PMLA*, 132(3) :651–658.

Marcus Pöckelmann, Janis Dähne, Jörg Ritter, et Paul Molitor. 2020. Fast paraphrase extraction in Ancient Greek literature. *it - Information Technology*, 62(2) :75–89. Publisher : De Gruyter Oldenbourg.

Charlotte Schubert. 2020. Intertextuality and Digital Humanities. *Information Technology*, 62(2) :53–59.

John R. Searle. 1980. Minds, brains, and programs. *The Behavioral and Brain Science*, 3 :417–457.

C. P. Snow. 1959. *The Two Cultures and the Scientific Revolution*. The Syndics of the Cambridge University Press.

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.

Pierre Waltz. 1929. *Anthologie grecque. t. I*, volume I of *Budé*. Les Belles Lettres, Paris.