Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

Reconnaissance automatique des appellations d'œuvres visuelles antiques

Aurore Lessieux[1] Iris Eshkol-Taravella[1] Anne-Violaine Szabados[2] Marlene Nazarian[2]
(1) Modeles, Dynamiques, Corpus (MoDyCo) UMR7114, Université Paris Nanterre, France
(2) Archeologies et Sciences de l'Antiquité (ArScAn) UMR7041, LIMC, France
aurorelessieux@parisnanterre.fr, ieshkoll@parisnanterre.fr,
anne-violaine.szabados@cnrs.fr, marlene.nazarian@cnrs.fr

RÉSUMÉ

Le projet pluridisciplinaire MonumenTAL a pour objectif de repérer et répertorier les appellations d'œuvres d'art visuel de l'Antiquité classique dans des textes en français publiés du XVIIIe au XXIe siècle en utilisant les méthodes du TAL. Il repose sur une collaboration étroite entre historiens de l'art (LIMC), linguistes-TAListes (MoDyCo) et bibliothécaires (BnF). Le traitement proposé implique plusieurs étapes : sélection du corpus d'étude, élaboration d'une typologie des appellations, constitution d'un corpus annoté par les experts du domaine et développement d'un outil de reconnaissance automatique des appellations fondé sur des méthodes symboliques.

ABSTRACT

Recognition of classical visual works appellations.

The multidisciplinary project MonumenTAL aims to identify and list the names of works of classical visual works of art in French texts published from the 18th to the 21st century using NLP methods. It is based on a close collaboration between art historians (LIMC), NLP researchers (MoDyCo) and librarians (BnF). The process includes several stages : selection of the study corpus, elaboration of a typology of appellations, constitution of a corpus annotated by experts in the field and development of a tool for automatic recognition of appellations based on symbolic methods.

MOTS-CLÉS : appellations d'œuvre, humanités numériques, histoire de l'art, TAL, REN, méthodes symboliques.

KEYWORDS: artwork title, Digital humanities, Art history, NLP, NER, symbolic method.

1 Introduction

Les appellations des œuvres visuelles de l'Antiquité classique sont au cœur du projet MonumenTAL – Monuments antiques et Traitement Automatique de la Langue – qui vise leur repérage automatique et leur étude par les historiens d'art et les archéologues. Compte tenu du grand nombre d'œuvres concerné et de la multiplicité de leurs dénominations, des méthodes et des outils relevant du TAL ont été privilégiés. Il s'agit d'une recherche pluridisciplinaire, qui repose sur une collaboration étroite entre des linguistes-TAListes (MoDyCo), des historiens d'art et des philologues (LIMC) et des conservateurs et bibliothécaires de la BnF.

Atelier TAL et Humanités Numériques (TAL-HN), pages 36–44.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

Cette présentation porte sur le processus du développement d'un module de détection automatique des appellations d'œuvres d'art antiques dans les textes en français produits du XVIIIe au XXIe siècle. Le traitement proposé implique plusieurs étapes : constitution du corpus d'étude, mise en place d'une typologie des appellations, développement d'un corpus annoté manuellement par les experts du domaine, la création d'un outil de reconnaissance automatique des appellations d'œuvre fondé sur des méthodes symboliques et ses résultats.

2 Appellations d'œuvres

Les appellations d'œuvres d'art varient. Cette variation est dûe à différents facteurs. Les œuvres visuelles de l'Antiquité classique, telles que les statues ou les scènes figurées sur des artefacts et des décors architecturaux, sont souvent connues sous plusieurs titres attribués au fil des siècles, certains dès la période antique (Prioux, 2011), selon des principes de nommage qui témoignent de l'érudition de l'auteur et de divers contextes de transmission textuelle de la connaissance. À titre d'exemple, l'œuvre antique connue sous l'appellation de l'*Apollon au diadème* dans les écrits de Pline devient l'*Apollon du Vatican* ou l'*Apollon du Belvédère* à l'époque moderne.

Par ailleurs, certaines appellations désignent à la fois l'œuvre originale et ses copies ou dérivés plus tardifs, ainsi que, en conséquence, son type iconographique que l'on pourrait considérer comme un bien culturel conceptuel. C'est le cas par exemple de l'*Aphrodite anadyomène*, ou l'*Anadyomène* (Wissowa, 1894), qui peut désigner l'original – une peinture perdue d'Apelle –, une réplique ou un type iconographique et en l'occurrence un thème iconographique (*Aphrodite sortie des eaux*). À cela s'ajoute la fréquente substitution, selon les époques, d'un nom de personnage grec par celui de son équivalent latin, par exemple Vénus pour Aphrodite dans *Vénus de Milo* (une statue grecque d'Aphrodite) ou dans le titre de l'article Wikipedia « Vénus anadyomène ».

3 Corpus

Afin de prendre en compte la diversité des appellations et leurs évolutions, le corpus textuel sélectionné, principalement sur Gallica (bibliothèque numérique de la BnF), mais aussi sur OpenEdition, Persée et Wikipedia pour garantir un accès libre et durable aux ressources, couvre une période allant du XVIIIe au XXIe siècle et inclut des traductions de textes antiques grecs et latins. Il réunit un panel varié représentatif des écrits des spécialistes (catalogues de musée structurés ou rédigés, essais, articles de revue, ...) et des amateurs (guides de musée ou de voyage, manuels, articles d'encyclopédie ou de revue, ...). Depuis le début du projet, plus d'un millier de ressources textuelles en français ont été sélectionnées pour constituer un corpus global. Pour le développement d'un outil de reconnaissance automatique des appellations d'œuvre, un sous corpus de trente publications a été utilisé. Pour atteindre la représentativité des données, c'est-à-dire, pour rendre compte des différentes pratiques d'appellation, les textes de ce sous corpus sont de natures variées (quatorze écrits de spécialistes en histoire de l'art et seize d'amateurs) et de diverses époques pour un total de 591 711 tokens.

Une importante partie du corpus n'étant pas nativement numérique, les publications ont été soumis à un processus d'océrisation propre aux différents sites proposant les textes. L'OCR n'est pas un procédé totalement fiable. Suivant l'état de conservation des ouvrages (défauts d'impression, vieillissement du papier, lisibilité, etc), la qualité de l'image numérisée joue sur l'aptitude des systèmes OCR à

en extraire le texte. On retrouve diverses erreurs d'océrisation pouvant aller de problèmes dans les diacritiques et les lettres d'un mot, à des suites de lettres ou de symboles qui n'ont pas sens. Ainsi, les textes récupérés après l'océrisation ont exigé la mise en place d'un script de pré-traitement pour en permettre un traitement numérique et facilement applicable par l'ensemble des membres du projet (talistes et historiens de l'art). Il a donc été nécessaire d'éliminer les erreurs d'océrisation et de supprimer plusieurs éléments indésirables : les erreurs de transcription communes ("VA" pour "L'A" dans L'Apollon, L'Amour, L'Athéna), la désorganisation de la mise en forme (passages à la ligne, etc.), les numéros en chiffres arabes (pagination, inventaires de collection, etc.).

4 Typologie des appellations d'oeuvres visuelles antiques

Grâce à une première étape d'analyse manuelle du corpus, associée à des réunions assurant le partage de compétences et de connaissances entre historiens d'art et linguistes, une typologie des appellations d'œuvres visuelles - principalement des statues, des peintures et des pierres gravés - a pu être établie et servir de modèle pour l'annotation manuelle.

Trois catégories d'appellations sont distinguées :

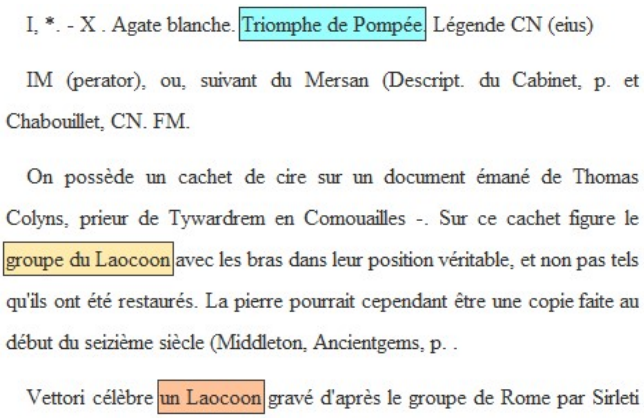
- **Appellations Précises** : elles sont considérées comme les plus courantes, suffisent pour reconnaître l'image, l'artefact ou le type iconographique et sont structurées de manière régulière et récurrente. Elles se focalisent soit sur le personnage principal figuré (personnage antique divin, mythique ou historique) : *Vénus Victrix*, *Hercule Farnèse*, *Aphrodite Cnidienne*, *Zeus olympien de Phidias*, *Gladiateur mourant du Capitole*, *statue de la Vénus de Milo* ; soit associent le type d'objet avec le nom de l'artiste (*peinture d'Apelle*, *Dinos de Sophilos*), de la collection (*Tête Kaufmann*, *Vase Médicis*) ou d'un lieu de provenance ou de conservation (*Aphrodite de Cnide*).
- **Thèmes** : ils renvoient vers des thèmes iconographiques qui sont exprimés par une brève description de la scène figurée ou le nom d'une légende, d'un mythe (*Hercule combattant le lion de Némée*). On y distingue principalement des événements liés aux personnages antiques, notamment les événements de la vie (naissance, mariage ou mort), les combats, les sacrifices, les processions, les banquets (*Naissance d'Athéna, Mariage de Pélée et Thétis, Bataille contre les Perses, Enlèvement de Perséphone, Sacrifice d'Iphigénie, Jugement de Pâris*), des descriptions du personnage principal et de sa place par rapport à un objet (*Bacchus dans un char, Amour avec la dépouille du lion de Némée*) ou à un animal (*Hercule sur le sanglier d'Erymanthe*), son attitude (*Hercule au repos*), sa condition (*le Gaulois blessé*), ou son costume (*Antigone en armure*) ; et des descriptions de personnages accomplissant ou subissant une action (*Combat de gladiateurs, Dionysos portant Bacchus enfant, Hercule dompté par l'Amour, Mercure changé en chèvre*).
- **Appellations Courtes** : elles sont limitées au nom du personnage précédé d'un déterminant (*le Laocoon, cette Vénus*). Ce type d'expression était notamment utilisé par les auteurs du XVIIIe et du début du XIXe siècle pour les statues célèbres car peu de sculptures antiques étaient connues à l'époque : *l'Apollon* suffisait à identifier *l'Apollon du Belvédère*, tout comme *le Sauroctone* pour *Apollon sauroctone* ou *la Cnidienne* pour *Aphrodite de Cnide*.

5 Annotation manuelle

Le corpus (30 textes) a été annoté manuellement selon cette typologie par trois historiens de l'art en utilisant Glozz (Widlöcher & Mathet, 2012). Le projet Monumental réunit les chercheurs de deux disciplines : histoire d'art et TAL. Pour pouvoir travailler ensemble, il fallait se familiariser avec les méthodes et pratiques propres à chaque discipline. Ainsi, chaque étape de travail a nécessité des aller-retours importants entre les participants : ateliers de formation pratique à l'annotation manuelle, ateliers de prise en main de l'outil d'annotation.

Le corpus annoté comprend 1701 Appellations Précises, 1277 Thèmes et 640 Appellations Courtes. L'évaluation de l'annotation manuelle a été faite sur le corpus annoté par deux annotateurs en utilisant le kappa de Cohen (1960). On obtient un kappa de 0,93 pour les Appellations Précises, 0,84 pour les Thèmes et 0,88 pour les Appellations Courtes. L'Appellation Précise a le meilleur score inter-annotateur car il s'agit de la catégorie pour laquelle les constructions sont les plus identifiables : elles sont le plus souvent formées à partir du personnage représenté (personnage divin, mythique ou historique ancien) ou du type d'objet. La catégorie Thème, qui comprend les descriptions iconographiques, est plus difficile à annoter car elle se confond facilement avec une partie de description complète et détaillée d'une scène d'œuvre d'art figurée (*Héraclès combattant le lion de Némée*). La différenciation nécessite une solide connaissance des expressions traditionnellement employées, et donc du domaine artistique, qui peut varier d'un annotateur à l'autre.

La Figure 1, montre un exemple d'annotation manuelle effectuée sous le logiciel Glozz sur un extrait du catalogue raisonnés de collections, "Pierres gravées des collections Marlborough et d'Orléans", de S. Reinach (1895). L'appellation "*groupe du Laocoon*" (jaune) est une Appellation Précise, l'appellation "*Triomphe de Pompée*" (bleu) est un Thème, et l'appellation "*un Laocoon*" (orange) est une Appellation Courte.


I, *. - X . Agate blanche. Triomphe de Pompée. Légende CN (eius)

IM (perator), ou, suivant du Mersan (Descript. du Cabinet, p. et Chabouillet, CN. FM.

On possède un cachet de cire sur un document émané de Thomas Colyns, prieur de Tywardrem en Comouailles -. Sur ce cachet figure le groupe du Laocoon avec les bras dans leur position véritable, et non pas tels qu'ils ont été restaurés. La pierre pourrait cependant être une copie faite au début du seizième siècle (Middleton, Ancientgems, p. .

Vettori célèbre un Laocoon gravé d'après le groupe de Rome par Sirleti

FIGURE 1 – Annotation sur Glozz d'un extrait du catalogues raisonnés de collections, "Pierres gravées des collections Marlborough et d'Orléans", S. Reinach, 1895.

Ce corpus annoté a servi de corpus de référence pour le développement et l'évaluation d'un outil de reconnaissance automatique des appellations d'œuvres antiques fondé sur des méthodes symboliques.

6 Reconnaissance automatique des appellations

6.1 La méthode choisie

L'annotation des appellations d'œuvres antiques se rapproche de l'annotation des entités nommées dans le sens où les appellations d'œuvres antiques et les entités nommées ont des critères définitoires similaires (Nouvel *et al.*, 2016) : toutes deux renvoient à une entité référentielle unique et elles sont autonomes du point de vue référentiel – une entité nommée au même titre que l'appellation d'une œuvre antique permet à elle seule l'identification du référent. La tâche de reconnaissance des entités nommées exploite des méthodes et techniques variées : les méthodes symboliques (Díez Platas *et al.*, 2020; Collin & Guerraz, 2015), l'apprentissage supervisé de surface (Brandsen *et al.*, 2020) et profond (Brandsen *et al.*, 2021).

Pour ce projet, le choix s'est porté sur l'utilisation des méthodes symboliques pour plusieurs raisons. Une étude de E. Prioux (Prioux, à paraître) a démontré une récurrence dans les composants et la structure interne des appellations d'œuvres et les thèmes iconographiques de l'art classique. De plus, dans une visée pluridisciplinaire, l'emploi d'une méthode opaque comme l'apprentissage profond aurait été un frein dans la réutilisation de l'outil par un public non TAListe. La capacité de l'outil sélectionné à proposer une représentation claire et visuelle du processus de reconnaissance automatique, son accessibilité et sa prise en main simple et rapide ont été des critères décisifs dans le choix de la méthode et de l'outil dans MonumenTAL. D'autre part, en considérant la taille du corpus annoté qui nous occupe, le choix d'une méthode symbolique nous a semblé être la plus pertinente. Le développement de l'outil de reconnaissance automatique des appellations est fondé donc sur des méthodes symboliques.

6.2 Outil d'annotation automatique

Les méthodes symboliques s'appuient sur la création de patrons syntaxiques utilisant des dictionnaires électroniques certains déjà existants et créés spécifiquement pour le projet, et formalisés sous Unitex (Paumier *et al.*, 2009). Les patrons Unitex ont été développés sur 40% des annotations du corpus annoté manuellement, puis testés sur 30% et enfin évalués sur les 30% restants.

33 patrons et 12 dictionnaires de noms communs et de noms propres, relatifs aux différents composants des appellations ont été créés. L'élaboration des dictionnaires s'est appuyée sur le thésaurus TheA (Thésaurus-Antiquité) du laboratoire LIMC et son enrichissement par des termes récupérés dans les appellations annotées manuellement. Le terme est sélectionné s'il est utilisé au moins une fois dans une appellation. Parmi ces dictionnaires, on retrouve par exemple : le dictionnaire PERSONNAGE qui regroupe des noms d'êtres et de personnages divins, mythiques, allégoriques, historiques, de créatures hybrides, de fonctions (prêtre, empereur), présents en tant que sujet représenté d'une œuvre visuelle antique ; le dictionnaire SUPPORT qui rassemble la liste des types d'objets antiques portant une représentation visuelle (statue, buste, reliefs, camée, peinture) ; des dictionnaires d'artistes antiques (ARTISTE), de collections et collectionneurs (COLLECTION), de lieux antiques et modernes (LIEU).

La Figure 2 présente un graphe Unitex visant à reconnaître des Appellations Précises. Ce graphe utilise le lexique sous forme de dictionnaires construits préalablement. Il reconnaît les constructions suivantes :

— PERSONNAGE + COLLECTION : *Hercule Farnèse*

— PERSONNAGE + ARTISTE : *Vénus de Praxitèle*
— PERSONNAGE + LIEU : *Aphrodite de Cnide*
— SUPPORT + PERSONNAGE : *statue d'Apollon*
— SUPPORT + COLLECTION : *tête Leconfield*
— SUPPORT + ARTISTE : *statue de Stéphanos*
— SUPPORT + PERSONNAGE + COLLECTION : *statue de la Vénus de Médicis*
— SUPPORT + PERSONNAGE + ARTISTE : *statue de Vénus de Scopas*
— SUPPORT + PERSONNAGE + LIEU : *statue de la Diane de Versailles*

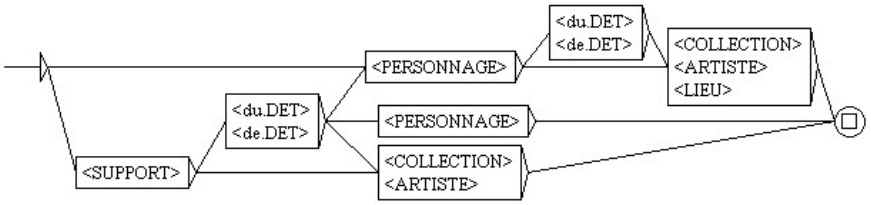


FIGURE 2 – Graphe Unitex pour la reconnaissance d'Appellations Précises.

Pour améliorer la précision de certains patrons syntaxiques, une sélection de termes spécifiques est nécessaire. À titre d'exemple, le dictionnaire PERSONNAGE contient aussi bien des noms de divinités, que des fonctions et des gentilés. Son utilisation en l'état pour une reconnaissance d'Appellation Courte (déterminant + PERSONNAGE) produit bien trop de bruit - là où *l'Apollon* est assuré d'avoir pour référent une œuvre, *l'empereur* ou *les Romains* font plus probablement référence à différents individus. Afin d'améliorer la précision de la reconnaissance des AC, il est ainsi nécessaire d'utiliser une sous-partie du dictionnaire qui contient uniquement des personnages bien individualisés.

6.3 Évaluation de la reconnaissance automatique des appellations

Les résultats de la reconnaissance automatique par catégorie sont présentés dans la table 1.

Type d'appellation	Précision	Rappel	F-mesure
Appellation Précise	0,72	0,84	0,78
Thème	0,85	0,72	0,78
Appellation Courte	0,87	0,84	0,89

TABLE 1 – Evaluation de l'outil de reconnaissance des appellations d'œuvres.

L'Appellation Courte est la catégorie la mieux reconnue avec une F-mesure de 0,89. La précision de l'Appellation Courte (0,87) dépend principalement de la pertinence des termes contenus dans le dictionnaire PERSONNAGE. En effet, cette appellation est construite à l'aide d'un déterminant et d'un terme du dictionnaire PERSONNAGE. La présence du déterminant permet de distinguer

le référent : par exemple, dans le contexte des œuvres visuelles, *la Vénus* fait référence à un objet matériel, tandis que *Vénus*, sans déterminant, peut aussi renvoyer à la divinité romaine. Cependant, certains noms de personnages sont génériques, comme le satyre, ou correspondent à des groupes, comme les Muses, et le déterminant ne suffit plus à distinguer ce à quoi renvoie *le satyre* ou *la Muse* : dans de nombreux cas, *Le Satyre* renvoie à un satyre parmi d'autres dans la description d'une œuvre ou dans un commentaire. Le Rappel (0,84) s'explique par le fait qu'il existait déjà une sélection dans le dictionnaire PERSONNAGE. Le choix a été fait de se concentrer sur la Précision de l'annotation et de laisser de côté les noms de certains personnages historiques comme les empereurs : dans les textes, *Auguste* fait plus souvent référence à l'empereur Auguste qu'à l'une de ses statues (par exemple *Auguste de Prima Porta*). Ainsi, pour améliorer la Précision de la catégorie Appellation Courte, le dictionnaire PERSONNAGE a évolué pour distinguer les individus des personnages génériques, groupes et fonctions.

Les Thèmes obtiennent une F-mesure de 0,78 due à leur ambiguïté avec les descriptions détaillées de scène, une forme d'expression spécifique des textes sur l'art (comme les descriptions d'œuvres d'art dans un catalogue). Ainsi, les patrons pour les Thèmes sont aussi stricts que possible : les thèmes iconographiques avec des personnages en action peuvent être exprimés au participe présent ou passé (*Dionysos portant Bacchus enfant*), mais aussi au présent (*Dionysus porte Bacchus enfant*). Cependant, comme le présent est le temps du discours dans les textes, rechercher ce type de construction au présent impliquerait de récupérer un grand nombre de séquences incluses dans des descriptions d'œuvres d'art, ce qui créerait un bruit considérable dans les résultats.

La Précision des Appellations Précises (0,72) s'explique principalement par deux facteurs : (1) l'ambiguïté de certaines appellations qui peuvent avoir plusieurs référents : une œuvre d'art matérielle représentant une divinité, une divinité (*Zeus Olympien*), un temple ou un sanctuaire particulier (*Artémis d'Ephèse*) ; (2) notre volonté d'outrepasser les limites restrictives des dictionnaires en autorisant certains patrons à reconnaitre des appellations du type "SUPPORT ou PERSONNAGE + X", avec X un nom propre n'étant pas présent dans les dictionnaires COLLECTION, ARTISTE et LIEU. Par exemple, cette approche s'est avérée efficace pour repérer les œuvres dont l'appellation associe le nom du personnage et le lieu de trouvaille (fouille archéologique) et est mentionnée dans le titre de l'article de journal consacré à sa découverte, suivant une habitude des archéologues du XIXe siècle, comme *Le Mars de Coligny* dans J. Buche," *Le Mars de Coligny (Musée de Lyon)*", Monuments et mémoires de la Fondation Eugène Piot n°10.1, 1903). Ce procédé fonctionne également bien pour les stèles funéraires portant les noms inscrits des défunts car leur désignation associe traditionnellement ce type d'objet au nom du défunt (ex. : *Stèle (funéraire) de Dexileos*). Ce type de construction permet d'identifier de nouvelles appellations, de compléter les dictionnaires et d'enrichir le thésaurus TheA du LIMC. Le Rappel de 0,84 peut s'expliquer par des problèmes dus à la qualité variable du traitement OCR des textes qui ne peuvent être corrigés automatiquement : *Apliro-dite* pour *Aphrodite*, *Gladiateur iYJgafias* pour *Gladiateur d'Agasias*.

Notons que les résultats de l'annotation automatique des désignations des œuvres d'art peuvent varier d'un texte à un autre en raison de la qualité du traitement OCR des textes, du type de publication ou de l'époque.

On peut mettre en relation nos résultats (F-mesure comprise entre 0,78 et 0,89) avec des travaux sur des données comparables. Même si des données traitées ne sont pas identiques, les chercheurs utilisent également des méthodologies symboliques afin de reconnaître des entités nommées dans les textes médiévaux espagnols (Díez Platas *et al.*, 2020) où ils obtiennent la F-mesure entre 0,74 et 0,87, ou dans les titres de films (Collin & Guerraz, 2015) où ils obtiennent la F-mesure de 0,63.

En comparaison avec ces travaux nous obtenons de meilleurs résultats. Si l'on choisit le domaine des humanités numériques, Brandsen et al. utilisent d'autres techniques comme l'apprentissage automatique de surface (F-mesure de 0,70) (Brandsen *et al.*, 2020) et l'apprentissage profond (F-mesure moyenne de 0,735) (Brandsen *et al.*, 2021). Ces résultats sont comparables à ceux que nous avons obtenus en utilisant les méthodes symboliques.

7 Perspectives

Au travers de cette étude pluridisciplinaire nous avons su mettre en commun et tirer parti de connaissances et de pratiques de différents domaines : les chercheurs en TAL ont dû s'adapter à des textes qui ne proviennent pas de leur discipline, et les historiens de l'art ont dû se confronter à des méthodes qui leur étaient étrangères comme l'annotation manuelle. Celle-ci a nécessité de passer par plusieurs étapes afin d'établir une typologie des appellations d'œuvres d'art antique efficace sur un corpus français. Cette collaboration a permis de mettre en place un processus d'annotation automatique avec des méthodes symboliques obtenant, pour les désignations d'œuvres d'art classiques, une F-mesure variant entre 0,78 et 0,89.

Grâce à ces travaux, la méthodologie présentée ici est actuellement réutilisée pour annoter des appellations d'œuvres d'art visuel dans des corpus textuels en latin et en grec ancien, ainsi que pour un ensemble de textes en français traitant de domaines artistiques autres que la statuaire et les pierres gravées. L'application du procédé à des publications sur la poterie peinte antique et les décors architecturaux (reliefs, mosaïques, peintures), qui présentent des scènes à plusieurs personnages et donc des sujets thématiques complexes, devrait d'abord permettre d'améliorer l'outil de reconnaissance d'appellation d'œuvres antiques, puis d'étendre son utilisation à d'autres mouvements et périodes artistiques.

La gestion des vocabulaires sous forme de thésaurus, induite par l'enrichissement des dictionnaires, permet aujourd'hui d'envisager des appellations normalisées pour les schémas iconographiques afin de différencier les œuvres originales et leurs copies (*Aphrodite anadyomène d'Apelle* / *type de l'Aphrodite anadyomène*), en tenant compte de statistiques sur les occurrences des termes. En perspective, il est envisagé d'exploiter cette distinction, déjà présente dans les vocabulaires produits, entre une appellation normalisée correspondant à une appellation préférée (prefLabel dans le standard SKOS pour les thésaurus) ou fréquente et ses variantes (une œuvre peut avoir plusieurs appellations au fil du temps), par exemple en indexant les occurrences dans les textes avec la forme normalisée.

Références

BRANDSEN A., VERBERNE S., WANSLEEBEN M. & LAMBERS K. (2020). Creating a dataset for named entity recognition in the archaeology domain. In *Proceedings of the 12th Language Resources and Evaluation Conference*, p. 4573–4577, Marseille, France : European Language Resources Association.

BRANDSEN A., VERBERNE S., WANSLEEBEN M. & LAMBERS K. (2021). Can BERT dig it ? - named entity recognition for information retrieval in the archaeology domain. *CoRR*. DOI : 10.48550/arXiv2106.07742.

COLLIN O. & GUERRAZ A. (2015). Classification d'entités nommées de type « film ». In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles*, p. 364–370, Caen, France : Association pour le Traitement Automatique des Langues.

DÍEZ PLATAS M. L., ROS MUÑOZ S., GONZALEZ-BLANCO E., RUIZ P. & ÁLVAREZ MELLADO E. (2020). Medieval Spanish (12th–15th centuries) named entity recognition and attribute annotation system based on contextual information. *Journal of the Association for Information Science and Technology*. DOI : 10.1002/asi.24399, HAL : hal-02970312.

NOUVEL D., EHRMANN M. & ROSSET S. (2016). *Named Entities for Computational Linguistics*. ISTE Ltd and John Wiley and Sons Inc. DOI : 10.1002/9781119268567, HAL : hal-01359440.

PAUMIER S., NAKAMURA T. & VOYATZI S. (2009). UNITEX, a Corpus Processing System with Multi-Lingual Linguistic Resources. In *eLexicography in the 21st century : new challenges, new applications (eLEX'09)*, volume 1, p. 173–175, France. HAL : hal-00621564.

PRIOUX E. (2011). Images de la statuaire archaïque dans les aitia de callimaque / archaic sculpture in callimachus. *Aitia*, **1**. DOI : 10.4000/aitia.74.

PRIOUX E. (à paraître). Titres et désignations antiques des œuvres d'art célèbres. *Duarte, P. and Le Bars-Tosi, F. (dir.), Vocabulaire des collectionneurs de l'Antiquité à la fin du XIXe siècle. Héritage méditerranéen*.

SZABADOS A.-V. (2014). From the limc vocabulary to lod. current and expected uses of the multilingual thesaurus thea. *Orlandi, S. et alii (éd.), Information Technologies for Epigraphy and Cultural Heritage. Proceedings of the EAGLE 2014 International Conference*, p. 59–75.

WIDLÖCHER A. & MATHET Y. (2012). The Glozz platform : a corpus annotation and mining tool. In CONCOLATO, CYRIL, SCHMITZ & PATRICK, Éds., *Proceedings of the ACM Symposium on Document Engineering (DocEng'12)*, p. 171–180, Paris, France. HAL : hal-01023774.