Transcrire un manuscrit en grec ancien

Maxime Guénette, Mathilde Verstraete, Marcello Vitali-Rosati, Alix Chagué

Transcrire un manuscrit en grec ancien : un modèle de reconnaissance automatique pour le *codex Palatinus graecus* 23

Maxime Guénette, Mathilde Verstraete, Marcello Vitali-Rosati
Université de Montréal
{maxime.guenette, mathilde.verstraete, marcello.vitali.rosati}@umontreal.ca

Alix Chagué
ALMAnaCH, Inria Paris ; Université de Montréal ; École Pratique des Hautes Études, Paris
alix.chague@inria.fr

Résumé

Cette contribution a pour but de présenter les résultats de nos expérimentations d'entraînement d'un modèle de transcription automatique (HTR) pour le grec ancien à partir d'un corpus d'entraînement élaboré sur le *Heidelbergensis Palatinus graecus* 23 et avec l'environnement logiciel eScriptorium/Kraken. Ce manuscrit byzantin datant de la fin du Xe siècle est un témoin capital pour l'épigrammatique grecque, en ce qu'il est la source principale nous livrant l'*Anthologie palatine*. Sa structure claire et son écriture soignée en font un candidat idéal pour l'entraînement d'un modèle pour le grec ancien.

1 Introduction

Rares sont les documents de la littérature grecque antique à être parvenus jusqu'à nous. Selon les calculs d'Irigoin (2002), 97.5% de la production dramatique de l'Athènes du Ve siècle a été perdue. Malgré cette perte considérable, Alphonse Dain (1961) estimait que plus de cinquante mille manuscrits grecs étaient encore conservés.

L'utilisation et l'implémentation de la reconnaissance automatique d'écritures (ou *Handwritten Text Recognition*, HTR dans la suite de l'article) propose un possible élément de réponse pour accéder aux textes de ces manuscrits. L'HTR est basé sur l'apprentissage automatique et permet d'établir des modèles de transcription potentiellement capables de reconnaître l'écriture de textes en grec ancien.

Comme le souligne Chahan Vidal-Gorène (2023), il existe peu de modèles de transcription pour le grec ancien et autres graphies non-latines permettant de rendre disponibles et interrogeables les manuscrits dans ces langues. L'absence de ces modèles de transcription est essentiellement due au manque de données annotées (vérité de terrain) publiques et libres pour l'entraînement de modèles de transcription (Tsochatzidis et al., 2021) [1].

Nous présentons ici une expérimentation qui s'inscrit dans le contexte du projet d'édition numérique collaborative de l'*Anthologie grecque* (désigné par la suite comme projet *AG*) que mène la Chaire de recherche du Canada sur les Écritures numériques (CRCEN) depuis 2014 [2]. Cette expérimentation vise à développer un modèle HTR pour le *Palatinus graecus* 23, principal témoin de l'*Anthologie palatine*, et manuscrit étudié dans le cadre du projet *AG*.

2 Le manuscrit

Le *Heidelbergensis Palatinus graecus* 23 (*P.*) est conservé à la Ruprecht-Karls-Universität (Heidelberg) qui en a réalisé une numérisation de bonne qualité [3]. Un partenariat entre le projet *AG* et la bibliothèque a permis d'identifier et d'isoler les épigrammes dans la numérisation du manuscrit pour les lier à notre plateforme, le tout en s'appuyant sur le protocole IIIF (Fernández Riva, 2020).

Notre premier échantillon se compose des pages 143-195 du manuscrit (épigrammes VI.13-285) [4]. *P* est le fruit de plusieurs copistes, lemmatistes,

1. En janvier 2024, le catalogue HTR-United compte deux jeux de données d'entraînement pour le grec ancien, contre 37 pour le français et 14 pour l'anglais, manuscrits et imprimés confondus. Cependant, il existe des transcriptions n'ayant pas été répertoriées dans HTR-United (*e. g.* Tsochatzidis et al., 2021). Sur HTR-United, voir Chagué et al. (2021) ou Chagué and Clérice (2023).

2. Voir https://anthologiagraeca.org.

3. La numérisation est accessible à l'adresse suivante : https://doi.org/10.11588/diglit.3449#0515. À noter que le manuscrit a été séparé en deux entre les livres XIII et XIV (à la p. 614) ; la seconde partie (pp. 615-709), le *Parisinus Supplementum graecum* 384 se trouve à la Bibliothèque Nationale de France (https://gallica.bnf.fr/ark:/12148/btv1b8470199g).

4. Selon la dernière pagination et celle adoptée par la Bibliothèque de Heidelberg, les pages 177, 188 et 189 ne figurent pas dans le manuscrit. Notre échantillon comprend donc 50 pages.

correcteurs, mais notre échantillon ne comporte que l'écriture du copiste A (et quelques annotations d'un correcteur et d'un lemmatiste) (Waltz, 1931). Il s'agit d'une écriture en minuscule du Xe siècle, aux caractères larges et plutôt épais (Agati, 1984, 52).

3 Méthodologie

3.1 Choix des outils

Notre choix pour la transcription et l'entraînement d'un modèle s'est porté sur eScriptorium (Kiessling et al., 2019). eScriptorium se distingue d'autres logiciels de transcription automatique par sa gratuité, son statut *open source*, mais aussi son ouverture : adossé au moteur d'HTR Kraken (Kiessling et al., 2017), il permet très aisément d'exporter les données de transcription vers des formats standards (XML ALTO, XML PAGE, TEI, texte brut), mais aussi d'exporter les modèles de transcription et de segmentation générés par l'intermédiaire de la plateforme [5]. Nous avons bénéficié d'un accès au serveur CREMMA sur lequel est déployée l'application eScriptorium afin de profiter des fonctionnalités de travail collaboratif [6].

eScriptorium permet l'import d'images hébergées sur un serveur IIIF à partir d'un lien vers le manifeste du manuscrit. Nous avons ainsi collecté les fichiers JPEG des pages 143-195 de *P*. Environ 100 heures ont été consacrées à la transcription manuelle qui s'est appuyée sur l'édition de Stadtmüller (1894). Au total, environ 78,000 caractères et 2,400 lignes ont été encodés.

3.2 Détection et classification des zones de texte

Pour la classification des zones de texte, nous avons suivi les préconisations de l'ontologie SegmOnto (Gabay et al., 2021) : `MainZone` pour les épigrammes, `MarginTextZone` pour les scholies [7].

Nous avons aussi opté pour une catégorisation des lignes simplifiée – permise par la structure du manuscrit. La majorité des lignes (tant les épigrammes que leurs scholies) ont été annotées sous la catégorie des *DefaultLine*, hormis quelques lignes interlinéaires – *InterlinearLine* – pour les ajouts ou corrections (**Figure 1**).

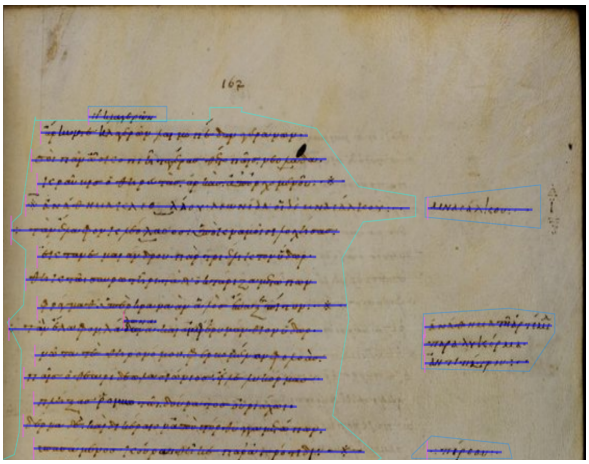

FIGURE 1 – Segmentation de *P*, p. 167 (VI.109 (*passim*)-111).

3.3 Transcription du texte

La transcription du manuscrit a soulevé des interrogations en plusieurs endroits :

— L'utilisation indifférenciée de majuscules et minuscules, caractéristique fréquente de cette écriture et époque (**Figure 2**) ;
— L'emploi d'abréviations et de caractères dits flottants : une ou plusieurs lettres suscrites complètent un mot (**Figure 2** et **3**) ;
— Un recours fréquent aux ligatures, notamment « ει, αγ, εγ, σσ, ελ, ου » (**Figure 4** et **2**).

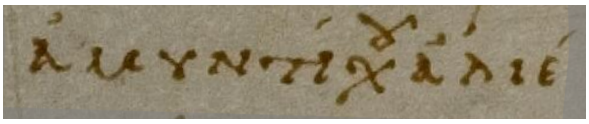

FIGURE 2 – *P*, p. 147 ; transcription : « ἀμυντίχου ἀλιέ ».

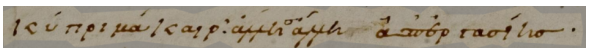

FIGURE 3 – *P*, p. 143 ; transcription : « κύπρι μάκαιρ' · ἄλλησ ἄλλη ἀπ' ἐργασίησ · ».

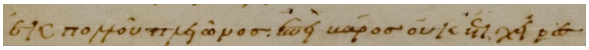

FIGURE 4 – *P*, p. 158 ; transcription : « ἐκ πολλοῦ πλειῶνοσ· ἐπεὶ βάροσ οὐκέτι χεῖρεσ ».

5. Ces fonctionnalités d'export permettent la publication des modèles et de la vérité de terrain pour une réutilisation par le grand public.
6. https://cremmacall.sciencescall.org/.
7. Nous avons délibérement gardé une segmentation simple : d'autres zones pourraient toutefois être utilisées (`NumberingZone` pour les zones de pagination ou `DamageZone` pour les parties endommagées du manuscrit, etc.).

Données d'entraînement	Modèles de base	Mode de normalisation Unicode	Nombre d'époques	Taux de précision (%)
From scratch	Aucun	NFC	47	87,36
From scratch	Aucun	NFD	71	89,96
Fine-tuning	CREMMA Medieval	NFD	47	89,16

TABLEAU 1 – Tableau 1. Résultat des premiers modèles avec la vérité terrain de *P*, pp. 143-195 (soit 50 pages).

Données d'entraînement	Modèles de base	Mode de normalisation Unicode	Nombre d'époques	Taux de précision (%)
From scratch	Aucun	NFC	48	91,00
From scratch	Aucun	NFD	40	90,85
Fine-tuning	CREMMA Medieval	NFD	48	91,05

TABLEAU 2 – Tableau 2. Résultat des seconds modèles avec la vérité terrain de *P*, pp. 143-215 (soit 70 pages).

Afin de standardiser la façon dont nous transcrivions le texte du manuscrit, nous avons :

— Normalisé tous les sigma avec le caractère « σ » (et non « ς ») [8] ;
— Transcrit uniquement les lettres visibles des caractères flottants et abréviations ;
— Assigné un symbole Unicode aux signes significatifs grâce à un tableau de correspondance [9] ;
— Normalisé et simplifié la ponctuation avec le point médian (·) afin de faciliter la tâche de reconnaissance de nos modèle [10].

3.4 Entraînement

À partir de la transcription manuelle de ces 50 premières pages, nous avons produit un modèle de segmentation capable de reconnaître les diverses zones et types de lignes de *P* précédemment établies avec SegmOnto [11]. Il n'existe pas de métrique qui permette d'évaluer de manière fiable la réussite d'un modèle pour la détection des lignes, la détection des zones et leur typage tout à la fois, mais dans notre cas, le modèle donnait visuellement des résultats concluants.

eScriptorium permet aussi l'entraînement de modèles de transcription nécessaires pour la reconnaissance du texte de *P* [12]. Trois premiers modèles de transcription ont été produits : (1) un modèle entraîné *from scratch*, c'est-à-dire sans pré-apprentissage, sur nos transcriptions en utilisant la normalisation Unicode NFC (*Normalization Form Canonical Composition*) ; (2) un modèle sans pré-apprentissage basé sur des données normalisées en NFD (*Normalization Form Canonical Decomposition*) [13] ; (3) un modèle avec *fine-tuning* à partir du modèle CREMMA Medieval (Pinche, 2022), initialement entraîné pour les manuscrits médiévaux latins [14].

3.5 Résultats préliminaires et *fine-tuning*

Les résultats pour nos premiers modèles ayant été entraînés de zéro sont encourageants (**Tableau 1**). Selon Hodel et al. (2021), un taux de précision de 90% pour un modèle de transcription est acceptable pour une analyse computationnelle [15]. Il faut cependant noter que le mode NFD complique parfois la tâche de saisie manuelle et de correction et qu'un pré-traitement de la prédiction pour rebasculer en mode NFC avant la correction manuelle serait utile [16].

Nous avons ré-entraîné les modèles après avoir augmenté notre vérité de terrain de 20 pages (196-215). Ce *fine-tuning* permet d'améliorer notre taux de précision en fournissant de nouvelles données aux modèles et ainsi les rendre plus performants (**Tableau 2**). L'augmentation de la quantité de données a permis un léger gain de précision pour les modèles entraînés de zéro (+0,89 points en moitié

8. Cette décision reflète la pratique du scribe qui utilise exclusivement « σ ».
9. eScriptorium permet de personnaliser un clavier virtuel, avec des raccourcis vers tout caractère UTF-8.
10. *P* présente une ponctuation sous la forme d'un point en hauteur, au milieu ou en bas de la ligne.
11. `MainZone` ou `MarginTextZone` pour les zones de texte, `DefaultLine` ou `InterlinearLine` pour les types de ligne
12. L'entraînement de modèles dans l'application eScriptorium du serveur CREMMA ayant été rendu impossible en août 2023 pour des raisons techniques, nous sommes passés directement par le logiciel Kraken.
13. Pour comprendre la distinction entre NFC et NFD et l'intérêt de jouer sur ce paramètre pour l'entraînement de modèles pour le grec ancien, il faut considérer que le graphème « ὰ » en NFC est encodé comme un seul caractère, tandis qu'en NFD, il est encodé comme la combinaison de deux caractères, l'alpha et l'accent. En général, le mode NFD peut s'avérer utile pour les textes comportant beaucoup de combinaisons de diacritiques.
14. Dans le cadre de ce dernier entraînement, le modèle pour le grec ancien était entraîné de manière à ne conserver en mémoire que les caractères présents dans la vérité terrain issues de *P*. Cela correspond à l'option *resize* avec la valeur *new* dans Kraken 4.3.
15. Les auteurs considèrent aussi un taux de précision supérieur à 95% comme étant très bon, et un taux supérieur à 97,5% comme excellent.
16. Le changement d'encodage d'un texte peut facilement être changé par des transcodeurs (voir notamment Cayless, 2011).

moins d'époques) ou à partir du modèle CREMMA Medieval (+1,9 points).

3.6 Modèles et données

Les trois modèles produits sont disponibles sur Zenodo (Guénette et al., 2024a,b,c). Les *datasets* sont également accessibles à l'adresse suivante `https://gitlab.huma-num.fr/ecrinum/anthologia/htr_cpgr23`.

4 Conclusion

En définitive, nous observons (1) que l'augmentation de la quantité de données d'entraînement permet d'obtenir des modèles de transcription légèrement plus performants et (2) qu'il est possible de s'appuyer sur les mécanismes de *transfer learning* pour un modèle de transcription avec un *fine-tuning* à partir d'un modèle entraîné sur des manuscrits de la même période mais rédigés dans un alphabet tout à fait différent, comme c'est le cas du modèle CREMMA Medieval [17].

Nos modèles sont cependant entraînés à reconnaître presque uniquement la main du copiste A dans *P*. Pour une transcription complète du manuscrit, il sera nécessaire d'entraîner de nouveaux modèles avec des échantillons variés incluant les mains d'autres copistes, correcteurs, et lemmatistes. À nos modèles actuels pourraient alors être appliqué un *fine-tuning* avec ces nouveaux échantillons pour être représentatifs de l'entièreté de *P*.

Bibliographie

Maria Luisa Agati. 1984. Note paleografiche all'"Antologia Palatina". *Bollettino Classici Lincei*, 5 : 43–59.

Hugh A. Cayless. 2011. Transcoder.

Alix Chagué and Thibault Clérice. 2023. "I'm here to fight for ground truth": HTR-United, a solution towards a common for HTR training data. In *Digital Humanities 2023 : Collaboration as Opportunity*.

Alix Chagué, Thibault Clérice, and Laurent Romary. 2021. HTR-United : Mutualisons la vérité de terrain ! In *DHNord2021 - Publier, partager, réutiliser les données de la recherche : les data papers et leurs enjeux*, Lille, France. MESHS.

Alphonse Dain. 1961. Paléographie grecque. In *L'Histoire et ses méthodes*, pages 532–552. Gallimard, Paris.

17. Le dépôt hébergeant les données d'entraînement ainsi que les modèles se trouve à l'adresse suivante : https://gitlab.huma-num.fr/ecrinum/anthologia/htr_cpgr23.

Gustavo Fernández Riva. 2020. Guidelines for the Annotation of the Digital Facsimile of the *Anthologia Palatina* (Cod. Pal. graec. 23).

Simon Gabay, Jean-Baptiste Camps, Ariane Pinche, and Nicola Carboni. 2021. SegmOnto, A Controlled Vocabulary to Describe the Layout of Pages.

Maxime Guénette, Mathilde Verstraete, Alix Chagué, and Marcello Vitali-Rosati. 2024a. HTR Model Palatinus graecus 23 (Meleagre-NFC). Publisher : Zenodo.

Maxime Guénette, Mathilde Verstraete, Alix Chagué, and Marcello Vitali-Rosati. 2024b. HTR Model Palatinus graecus 23 (Meleagre-NFD). Publisher : Zenodo.

Maxime Guénette, Mathilde Verstraete, Alix Chagué, and Marcello Vitali-Rosati. 2024c. HTR Model Palatinus graecus 23 (Meleagre-NFD-finetuned). Publisher : Zenodo.

Tobias Hodel, David Schoch, Christa Schneider, and Jake Purcell. 2021. General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example. *Journal of Open Humanities Data*, 7(13) : 1–10.

Jean Irigoin. 2002. La transmission des textes et son histoire. *Publications de l'Académie des Inscriptions et Belles-Lettres*, 13(1) :1–20.

Benjamin Kiessling, Matthew Thomas Miller, Romanov Maxim G, and Sarah Bowen Savant. 2017. Important New Developments in Arabographic Optical Character Recognition (OCR). *Al-ʿUṣūr al-Wusṭā*, 25 : 1–13.

Benjamin Kiessling, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. 2019. eScriptorium: An Open Source Platform for Historical Document Analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, pages 19–24.

Ariane Pinche. 2022. HTR model Cremma Medieval.

Hugo Stadtmüller. 1894. *Anthologia graeca epigrammatum Palatina cum Planudea*, volume 1. B. G. Teubneri, Leipzig.

Lazaros Tsochatzidis, Symeon Symeonidis, Alexandros Papazoglou, and Ioannis Pratikakis. 2021. HTR for Greek Historical Handwritten Documents. *Journal of Imaging*, 7(12) : 260.

Chahan Vidal-Gorène. 2023. La reconnaissance automatique d'écriture à l'épreuve des langues peu dotées. *Programming Historian en français*.

Pierre Waltz. 1931. *Anthologie grecque*, volume III. Les Belles Lettres, Paris.