Éditer numériquement *Li histoire de Kanor et de ses freres* ms. BnF fr. 1446

Florian-Pierre Zanardi

5 mars 2025

1 Introduction

Le Roman de Kanor est le dernier roman du *Cycle des Sept Sages de Rome*. Parmi les six témoins qui nous sont parvenus, le ms. BnF fr. 1446 (C) donne à lire une version non-cyclique, qui contient une «re-chapitulation» et une version remaniée des romans de *Pelyarmenus* et de *Kanor*[1]. Ce témoin pose des problèmes de datation, de critique textuelle, quant l'antériorité du manuscrit[2], et littéraire, quant à sa composition.

2 État de l'art

Le projet FNS *Canoniser les Sept Sages*[3] souhaite à terme produire une base de données qui inclut tous les témoins du cycle. Une équipe a déjà OCRisé le *Pelyarmenus* et a fourni une édition XML-TEI[4].

Le témoin C du *Roman de Kanor* n'a jamais été édité et n'a encore fait l'objet d'aucun traitement numérique, alors que le ms. 1446 renferme plusieurs œuvres qui intéressent la critique[5].

3 Le corpus du projet

Le corpus se compose de 70 folios issus du ms. BnF fr. 1446[6] qui forment l'intégralité du roman. Les images ont été récupérées par le point d'accès IIIF grâce au script Pyllica[7]. Deux familles de mains sont repérables, notamment de type gothique *littera hybrida*. Le trait est rapide et la préservation des traces écrites médiocre.

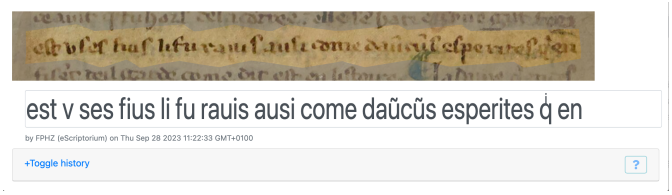

Figure 1 – *Exemple de la* `main1`

4 Modèles utilisés

Nous avons utilisé un modèle de segmentation et un modèle OCR. Seul le modèle OCR a été entraîné.

4.1 Préparation des données

Les données du corpus ont été préparées sur une instance *eScriptorium*[8] des serveurs de l'INRIA en 2023. Elle utilise le moteur OCR Kraken[9]. Ces données ont été soumises au catalogue HTR-United[10].

4.2 Modèle de segmentation

Modèle Le modèle «global model rg best» (références précises à trouver) a servi à la segmentation automatique du texte. Une reprise manuelle a été nécessaire ponctuellement.

Règles de segmentation Pour la mise en page, les données suivent le vocabulaire contrôlé SegmOnto afin de décrire les différentes zones d'un folio. Les zones retenues sont réduites au minimum : `Title`, `MainZone`, `DropCapitalZone`, `NumeringZone` et `StampZone`[11].

4.3 Modèle de transcription

Une première transcription a été lancée à partir du modèle standard

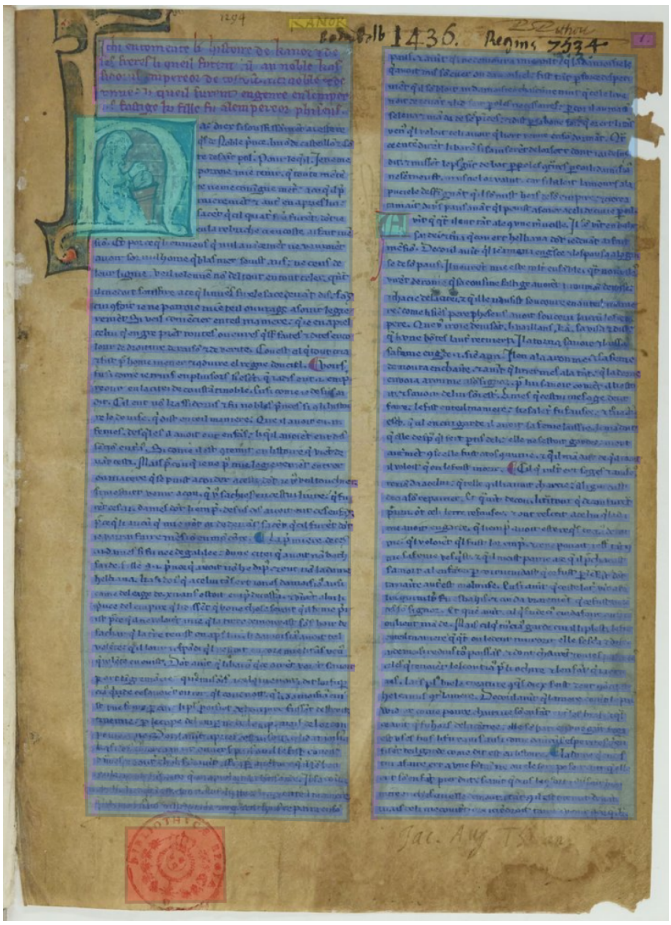

Figure 2 – *Zones retenues f. 1*

`cremma-medieval-lat-AF-best`[12]. Ce modèle est entraîné à partir d'un jeu de données centré sur des manuscrits des XIIIe et XIVe siècles en ancien français et en écriture gothique *textualis*.

Règles de transcription Pour corriger les données, nous avons suivi la synthèse d'Ariane Pinche pour les manuscrits du Xe au XVe siècle en français[13]. Ce document offre une procédure d'édition qui permet de normaliser la création des vérités de terrain. Nous avons opté pour une transcription pseudo-graphématique qui conserve ponctuation originale, abréviations et graphies, mais non les tirets de fin de ligne ni les espaces entre morphèmes élidés (ex. `quil` et non `qu il`). La nature de notre travail a impliqué un contrôle manuel systématique de tous les folios sur l'interface eScriptorium.

Entraînement Ariane Pinche a mené deux entraînements à partir des données contrôlées (*golden dataset*) créées. Un premier sur la base des 10 premiers folios, puis des 20 premiers.

Résultats Les fichiers ALTO de la transcription sont disponibles sur le GitHub du projet, ainsi que le modèle `Kanor` du ms. 1446.

5 Normalisation du corpus

Un document «Principes de transcription» recense les différentes règles appliquées pour obtenir l'édition normalisée, qui est notre support de travail pour l'étude littéraire. Ces principes suivent les conseils actuels en termes d'édition de texte, notamment ceux de Pascale Bourgain et Françoise Vielliard[14].

6 Une édition XML-TEI

Le pan numérique de cette édition entre dans un projet doctoral plus vaste dont il n'est qu'une fraction. Nous avons borné nos prétentions à une édition littéraire qui comprend plusieurs aspects.

6.1 Objectifs

La création de cette édition numérique répond à un besoin de compréhension littéraire et littéral de ce témoin complexe, ainsi que de son intégration spécifique dans un cycle romanesque important au Moyen Âge. Quelques objectifs : unifier les différents témoins du Cycle grâce à un alignement des unités textuelles ; unifier les références internes en réduisant les difficultés dues aux variations graphiques ;

6.2 Moyens

Nous avons respecté les guides d'encodage de la TEI [1] et utilisé le logiciel Oxygen XML Editor.

La rédaction d'un modèle ODD commun avec l'équipe *Pelyarmenus* permet une opérabilité de nos éditions. Un guide ODD propre au *Kanor* est rédigé pour rendre explicites les problèmes, nombreux, et les choix retenus[15]. Ce guide présente et justifie ces choix.

1. `https://tei-c.org/`

Un exemple efficace de cette collaboration réside dans la constitution d'une base commune de noms propres (personnages/lieux) avec les balises `<person xml:id="xxx">` et `<place xml:id="xxx">`.

6.3 Résultats

À ce jour, cette édition a permis de rendre lisible au public un texte complexe tant dans sa conception que dans sa matérialité, notamment grâce à un site internet [2]. Il intègre certaines possibilités offertes par le balisage XML-TEI. Par ailleurs, ce travail a permis d'intégrer le témoin C dans l'alignement synoptique mené par Camille Carnaille.

Remerciements

Ariane Pinche pour le temps passé à entraîner un modèle sur le ms. 1446 ainsi que l'INRIA pour la mise à disposition gratuite d'une instance eScriptorium. Camille Carnaille pour les nombreux échanges concernant l'apprentissage du XML-TEI et les choix de balisage communs aux deux projets.

Références

[1] Florian-Pierre Zanardi. "La «rechapitulation» du Cycle des Sept Sages et le *Roman de Kanor* dans le manuscrit Paris, BnF fr. 1446 : éditer une rédaction alternative". In : *XVIIe Congrès de la Société Internationale de Littérature Courtoise*. Université de la Colombie-Britannique, Vancouver, Canada, 24-28 juill. 2023.

[2] Albert Henry. "Burir et sa famille en ancien français". In : *Festschrift Kurt Baldinger zum 60. Geburtstag*. Tübingen, 1979, p. 511-522.

[3] Yasmina Foehr-Janssens et Simone Ventura. *Canoniser les Sept Sages : Livre, langues, écriture sérielle (XIIIe-XVe siècles)*. `https://www.unige.ch/c7s`. Projet financé par le FNS et le FNRS.

[4] Camille Carnaille, Prunelle Deleville et Sophie Lecomte. *Canoniser les Sept Sages de Rome (C7S)*. `https://github.com/CycleDes7Sages/CycleDes7Sages.github.io`. Projet financé par le FNS (Suisse) et le FNRS (Wallonie-Bruxelles).

[5] Nicolas Chardonnens et Béatrice Wahlen. "Heurs et malheurs d'un brouillon. Des contes desrimez de Baudouin Butor à Perceforest". In : *Lieux de mémoire antiques et médiévaux : Texte, image, histoire : la question des sources*. Sous la dir. de Bernard Andenmatten et al. BSN Press, 2012, p. 257-291. doi : `10.3917/bsn.ndenm.2012.01.0257`.

[6] Bibliothèque nationale de France. *Manuscrit numérisé BnF fr. 1446: ark:/12148/btv1b10023851v*. `https://gallica.bnf.fr/ark:/12148/btv1b10023851v`. Manuscrit numérisé disponible sur Gallica, la bibliothèque numérique de la BnF.

[7] Pierre-Carl Langlais et Julien Schuh. *Pyllica : Un outil de récupération automatisée de données sur Gallica*. `https://github.com/Dorialexander/Pyllica`. Version mise à jour disponible sur GitHub. 2025.

[8] Benjamin Kiessling et al. "eScriptorium: An Open Source Platform for Historical Document Analysis". In : *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. T. 2. 2019.

[9] Benjamin Kiessling. "Kraken - An Universal Text Recognizer for the Humanities". In : *Digital Humanities Conference 2019 - DH2019*. Utrecht, The Netherlands : ADHO, 2019.

[10] HTR-United Team. *HTR-United : Catalogue de jeux de données pour la transcription et la segmentation automatiques*. `https://htr-united.github.io/`.

[11] Simon Gabay et al. *SegmOnto, A Controlled Vocabulary to Describe the Layout of Pages*. `https://segmonto.github.io/`. Version 0.9, accessed March 5, 2025. 2023.

2. `kanor-c7s.github.io/`

[12] Ariane Pinche. *CREMMALab Project: Handwritten Text Recognition for medieval manuscripts*. Digital Humanities. Poster. Juill. 2022. URL : https://hal.science/hal-03724041.

[13] Ariane Pinche. "Guide de transcription pour les manuscrits du Xe au XVe siècle". working paper or preprint. Juin 2022. URL : https://hal.science/hal-03697382.

[14] Pascale Bourgain et Françoise Vielliard. *Conseils pour l'édition des textes médiévaux. III : Textes littéraires*. T. 4. Orientations et méthodes. Paris : École Nationale des Chartes, 2002, p. 253.

[15] Florian-Pierre Zanardi. *Guide d'encodage de la rédaction non cyclique de l'Histoire de Kanor et de ses freres d'après le manuscrit BnF fr. 1446 (C)*. https://github.com/kanor-c7s/kanor_xml_tei/blob/main/ODD/ODD-c7s-fro7.pdf. 2024.