HTR sur le Ms Paris BnF latin 7558
Description du travail effectué

Alice Leflaëc

Mars 2023

1 Présentation du projet

L'objectif du travail était de s'entraîner à l'utilisation de l'HTR en transcrivant une vingtaine de folios du manuscrit Paris BnF lat. 7558 au moyen d'eScriptorium.

2 Description du jeu de données

Le manuscrit Paris BnF lat. 7558, en minuscule caroline à longues lignes, est originaire de Lyon et date de la première moitié du IX^{e} siècle. Il contient un ensemble de textes grammaticaux et poétiques en langue latine.

Nous avons fait le choix de transcrire les folios 104v-124r qui comportent six poèmes de la fin du IV^{e} siècle et du début du V^{e} siècle ainsi que deux poèmes du IX^{e} siècle. La paternité de certains de ces poèmes a fait l'objet de débats qui durent parfois encore. :

- f. 104v-111v : Ps.-Paulin de Nole, *Laus sancti Iohannis*. Autrefois attribué à Paulin de Nole, ce poème de la fin du IV^{e} siècle a été retiré du corpus des *Carmina* dans la récente édition de Franz Dolveck (Dolveck 2015, p. 25-30).
- f. 111v-114v : Anonyme, *Laudes Domini*. Ce poème anonyme, dont le ms Paris BNF lat. 7558 est l'unique témoin, a peut-être été écrit par un rhéteur d'Autun au début du IV^{e} siècle.
- f. 114v-118r : Paulin de Nole, *Ad Iouium* (uniquement la partie versifiée). Ce poème a parfois été attribué à Claudianus Mamertus, ce dont témoigne l'indication « *uulgo Claudiano Mamerto attributum* » ajoutée par une main moderne à côté du titre. Cependant, l'attribution à Paulin de Nole n'est plus remise en cause.
- f. 118r-121r : Ps.-Paulin de Nole, *De obitu Baebiani*. L'attribution, douteuse depuis longtemps, de ce poème à Paulin de Nole a été réfutée par Franz Dolveck dans son édition (Dolveck 2015, p. 26-27). Le ms Paris BNF lat. 7558 est l'unique témoin de ce poème.
- f. 121r-122r : Latinius Pacatus Drepanius, *Versus Drepani de cereo Paschali*. Anne-Marie Turcan-Verkerk attribue ce poème, longtemps intégré à l'oeuvre de Florus de Lyon, à Latinius Pacatus Drepanius, un orateur du IV^{e} siècle (Turcan-Verkerk 2003).
- f. 122r-124r : Florus de Lyon, *Carmina* 10 (f. 122r-123r) et 25 (f. 123r-124r). Ces deux poèmes du IX^{e} siècle ont été intégrés au corpus transcrit, car une main moderne ajoute « *ejusdem Drepanii* » à côté du premier poème. L'attribution à Florus de Lyon n'est, toutefois, pas remise en cause.

Pour plus de renseignements sur le manuscrit et sur les textes qu'il contient, voir la notice de la BNF.

3 Chaîne de traitement

Tout le travail d'océrisation a été effectué avec l'application eScriptorium et, plus précisément, son infrastructure genevoise FoNDUE.

1. **Récupération des images** : les images des folios ont été récupérées sur le site de la BnF.

   Comme nous ne souhaitions utiliser que quarante images sur les 356 pages du manuscrit, nous avons utilisé l'export par PDF et non en IIIF même si nous avons conscience que ce n'est pas la procédure idéale. Nous nous sommes rendu compte trop tard qu'il était possible de ne récupérer qu'une partie du manuscrit sur le site de la BnF tout en utilisant le IIIF.

2. **Segmentation** : Nous avons utilisé l'outil de segmentation proposé par eScriptorium puis avons corrigé, au besoin, les zones et les lignes manuellement. Dans un souci d'interopérabilité des données, nous avons utilisé le vocabulaire contrôlé proposé par SegmOnto (SegmOnto Documentation).

   Types utilisés pour les zones : *DamageZone* pour les zones abîmées dans le manuscrit (cf. le trou des éléments 28 et 29) ; *DropCapitalZone* pour les initiales ; *MainZone* pour le corps du texte ; *MarginTextZone* pour les annotations marginales et *NumberingZone* pour les numéros des folios et ceux des cahiers.

   Types utilisés pour les lignes : *DefaultLine* pour les vers ; *DropCapitalLine* pour l'initiale dans la zone *DropCapitalZone* ; *HeadingLine* pour les titres, les *inplicit* et les *explicit* des poèmes et *InterlinearLine* pour toutes les annotations et tous les termes entre les lignes principales (*DefaultLine*).

   Remarques sur les annotations marginales et interlinéaires et sur les lettres suscrites :
   — Pour toutes les annotations marginales, le type *MarginTextZone* a été utilisé.
   — Les corrections interlinéaires d'une seconde main [1], sont nombreuses et présentent fréquemment un intérêt pour un éventuel travail d'édition. Il s'agit, le plus souvent, de simples lettres suscrites. Il a été fait le choix de traiter ces annotations comme des lignes interlinéaires.
   — Quand une lettre, premièrement oubliée, a été ajoutée par le copiste lui-même, celle-ci a été transcrite, lorsque la combinaison des lettres était possible, comme une lettre suscrite.

3. **Transcription** : Comme il a été dit précédemment, le texte transcrit est un manuscrit de la première moitié du IX^{e} siècle en minuscule caroline à longue ligne. Trois modèles, utilisant Kraken et récupérés sur HTR United, semblaient intéressants :
   — *cremma-medieval_best*
   — *HTR_medieval_documentary_best*
   — *cortado*

   Le modèle *cremma-medieval_best* s'est très rapidement avéré le plus pertinent et c'est celui avec lequel nous avons travaillé. Nous avons transcrit le texte grâce à eScriptorium puis nous avons corrigé manuellement la transcription.

1. Cette seconde main a été identifiée par P. Lejay (Lejay 1890, p. 172) et P.F. Hovingh (Hovingh 1960, p. 198) comme celle de Guillaume Morel, éditeur de certains textes du manuscrit en 1560.

4. **Exportation des données transcrites** : Les données ont été exportées hors du logiciel FoNDUE avec le format ALTO qui permet de donner une information géométrique pour chaque élément sur l'image.

5. **Dépôt GitHub** : Les données transcrites ont, pour finir, été déposées sur GitHub dans une organisation intitulée « Ms Paris BnF Lat. 7558 ».

4 Principes de transcription appliqués

Pour la transcription, nous nous sommes fondée sur les préconisations énoncées par Ariane Pinche (Pinche 2022) et sur les principes appliqués dans le projet CREMMA Medii Aevi : Literary manuscript text recognition in Latin (Clérice, Vlachou-Efstathiou et Chagué 2023).

Voici les principaux principes appliqués à notre jeu de données :

— **Lettres « u », « v », « i » et « j »**
Pas de distinction entre le « u » et le « v » ou le « i » et le « j ». Comme le préconise Pinche 2022 (p. 5), les « u » et les « v » ont été écrits avec la lettre « u » et les « i » et les « j » avec la lettre « i ». Clérice, Vlachou-Efstathiou, Chagué semblent utiliser, pour le « u » et le « v », le « u » pour la minuscule et le « V » pour la majuscule (Clérice, Vlachou-Efstathiou et Chagué 2023, Appendix, p. 16). Nous avons préféré mettre un « u » dans les deux cas.

— **Ponctuation**
Les principes de ponctuation proposés par Clérice, Vlachou-Efstathiou et Chagué 2023 ont été appliqués, à savoir la virgule pour les virgules (« , »), le point pour tous les points en bas (« . ») et le double point pour tous les autres signes de ponctuation (« : »).

— **Modifications dans une *DefaultLine***
Quand le copiste ou une autre main a remplacé une lettre par une autre lettre en écrivant directement sur la lettre, lorsqu'une lettre, une cédille ou un signe de ponctuation ont été ajoutés à la ligne principale (*DefaultLine*), nous avons transcrit la seconde strate du texte sans indiquer la première.
Le texte effacé, mais encore lisible, ou raturé a été transcrit en l'encadrant par de doubles crochets droits ⟦ ⟧ (U+27E6 et U+ 27E7) (voir, par exemple, l'image 36). Si le texte n'était pas lisible, nous avons utilisé ces mêmes crochets sans texte à l'intérieur (cf. Pinche 2022, p. 18-19).
Les points ajoutés sous certaines lettres par une seconde main n'ont pas été transcrits, de même que les carets, parfois utilisés pour insérer une correction interlinéaire.

— **Ligatures et abréviations**
Sur le modèle de ce que proposent Pinche 2022 et Clérice, Vlachou-Efstathiou et Chagué 2023, nous avons fait le choix d'une transcription graphématique conservant les abréviations. Nous avons suivi leurs recommandations pour le traitement des ligatures et des abréviations. Ainsi, les ligatures ne sont pas notées et les signes abréviatifs sont des signes UTF-8 rattachés au projet MUFI (*Medieval Unicode Font Initiative*).

Les principales abréviations utilisées sont les suivantes [2] :

Type	**Nom du signe**	**Unicode**	**Usage**
Tilde horizontal	Combining tilde	U+0303	Abréviation par contraction de la lettre *-m*
	p + combining tilde	p+U+0303	Abréviation de *prae-*
Signes spéciaux	Latin small letter b with stroke	U+0180	Utilisé pour l'abréviation de *nobis*
	Latin small letter d with stroke	U+0111	Utilisé pour l'abréviation de *quod*
	Latin small letter l with stroke	U+0142	Abréviation de certains termes comme uł pour *uel* ; sacła pour *saecula* et ppłm pour *populum*
	Latin small letter p with stroke	U+A751	Abréviation de *per-*
	Latin small letter p with flourish	U+A753	Abréviation de *pro-*
	Combining us above	U+A770	Usage du signe tironien pour la désinence *-us*
	Combining ur above	U+1DD1	Abréviation de la désinence *-ur*
	Combining cedilla	U+0327	Abréviation de la diphtongue *-ae* Cette cédille combinatoire a été rencontrée avec un *e*, mais aussi un *c* et &.

5 Suite envisagée pour le projet

À l'heure actuelle, les données transcrites avec le modèle *cremma-medieval_best* et corrigées manuellement ont été déposées sur GitHub. L'objectif est de les utiliser pour l'entraînement d'un modèle en recourant aux outils développés par HTR United. Une première étape sera de contrôler l'homogénéité du jeu de caractères utilisé grâce à l'outil de contrôle et d'encodage du texte *Choco-Mufin* [3]. Il faudra également répartir les données en trois sets pour l'entraînement, l'évaluation et le test en mélangeant les pages pour une bonne répartition des données.

2. Les noms employés dans la deuxième colonne sont empruntés au guide la MUFI.
3. Thibault Clérice et Ariane Pinche, *Choco-Mufin, a tool for controlling characters used in OCR and HTR projects*, 2021.

Bibliographie

Clérice, Thibault, Malamatenia Vlachou-Efstathiou et Alix Chagué (2023). *CREMMA Medii Aevi : Literary manuscript text recognition in Latin*. URL : https://hal-enc.archives-ouvertes.fr/hal-03828353v4.

Dolveck, Franz (2015). *Paulini Nolani Carmina*. Corpus Christianorum Series Latina 21. Turnhout : Brepols.

Gabay, Simon et al. (2021). *SegmOnto, A Controlled Vocabulary to Describe the Layout of Pages, version 0.9*. URL : https://github.com/SegmOnto.

Hovingh, Pieter Frans (1960). *Claudii Marii Victorii Alethia*. T. 128. Corpus Christianorum Series Latina. Turnhout : Brepols.

Lejay, Paul (1890). "Marius Victor, l'éditeur Morel et le ms. latin 7558 de Paris". In : *Revue de Philologie, de littérature et d'histoire ancienne* 14.1, p. 71-781.

Pinche, Ariane (2022). *Guide de transcription pour les manuscrits du Xe au XVe siècle*. URL : https://hal.science/hal-03697382.

Turcan-Verkerk, Anne-Marie (2003). *Un poète latin chrétien redécouvert : Latinius Pacatus Drepanius, panégyriste de Théodose*. Bruxelles : Latomus.