

Le roman Fierabas – rapport du travail

Carolina Macedo *alias* carolisteia,
Ekaterina Iakovleva *alias* yavlenieraket,
Gonzalo Freijedo *alias* gFreijedo

Mars 2023

1 Introduction

Ce travail, comme le récit de l’incunable sur lequel nous travaillons, se divise en trois parties : la première partie correspond à la phase technique et notre outil de travail est le logiciel *eScriptorium* ; la deuxième partie consiste dans un travail de recherche et de réflexion sur nos données et nos résultats ; finalement, pour la dernière partie nous avons déposé notre contenu sur la plateforme GitHub.

1.1 Plan du travail

Afin de transcrire nos documents, nous avons utilisé le logiciel *eScriptorium*. Chaque membre du groupe avait deux pages à transcrire, soit un total de 6 pages (*folios* 18 à 23). Après avoir téléchargé et sauvegardé les pages correspondantes à nos folios, nous avons créé un nouveau projet sur *eScriptorium*. Ensuite, nous avons emporté les images en format .jpg (nous n’avons pas réussi à télécharger les images dans le format iiif). Nous avons sélectionné les images à travailler, puis nous avons procédé à la segmentation du texte et création de régions de texte et des images. Dans l’étape suivante nous avons procédé à la transcription des images, ligne par ligne. Une fois réalisée la transcription, nous avons exporté nos données en tant que fichiers ALTO.

Finalement, nous avons déposé chacun de nos travaux sur GitHub. Nous avons travaillé initialement – soit localement, soit à distance – chacun sur une branche distincte (une branche par chaque membre de l’équipe) et ensuite avons *mergé* sur la branche principale tout ce contenu. Ce rapport a été réalisé sur l’éditeur collaboratif L^AT_EX Overleaf.

2 L’ HTR appliquée aux textes en français au XV^e siècle

Le principe de travail HTR (handwritten text recognition) peut être divisé en plusieurs étapes qui sont également liées à l’utilisation d’algorithmes d’apprentissage automatique formés sur de grands ensembles de données d’échantillons

d'écriture manuscrite : la première étape est la capture et la numérisation de l'image, la deuxième étape est l'analyse des images pour reconnaître d'abord l'individu les caractères et les mots (dans cette étape, HTR utilise un apprentissage automatique complexe, également pour les mots, il est utilisé le modèle de langue que l'utilisateur peut choisir, pour notre tâche, nous utilisons le modèle de langue latine), la troisième étape est l'utilisation du modèle de langue qui est connecté avec compréhension du sens des mots, puis montrer la sortie.

2.1 Les outils

La plateforme *eScriptorium*, utilisée dans ce travail, est un logiciel *open source* développé par des chercheurs de l'EPHE et de l'INRIA, qui permet l'analyse de documents images par la technologie de reconnaissance de l'écriture manuscrite (HTR), la reconnaissance optique de caractères (OCR - *Optical character recognition*), ainsi que des algorithmes d'apprentissage automatique¹. Grâce à ces technologies, les images qui sont données en *input* pour l'analyse seront décodées automatiquement. Comme rappelé par Peter Stokes (*apud* GAUTIER et al., 2022), *eScriptorium* permet d'effectuer plusieurs types de projets, tels que la transcription manuelle automatique, des transcriptions automatiques corrigées pour de l'édition, l'alignement des transcriptions existantes avec des photos de leur objet, des transcriptions imparfaites pour de l'analyse de données massives.

2.2 Les projets

Concernant des projets qui se consacrent à l'HTR dans le domaine des textes français au XV^e siècle, il faut mentionner le dépôt GitHub Cremma-medieval qui a été créé afin de mettre à disposition des *corpora* de transcription pour l'entraînement des modèles HTR pour les manuscrits médiévaux du XII^e au XV^e siècle. Le jeu de données CREMMA Medieval a été construit avec *eScriptorium* et Kraken.

Les recherches menées par Thibault Clérice, notamment (CLÉRICE, 2022), sont un exemple de ce qui se fait de plus récent dans le domaine. Dans cet article, l'auteur propose une nouvelle méthode basée sur le *deep learning* où il est possible de catégoriser chaque taux d'erreur de ligne en plages d'erreur pour résoudre les difficultés dans le contexte de la variation orthographique pour l'ancien français et le latin.

Soulignons également l'article sur la transcription de manuscrits médiévaux pour l'apprentissage automatique par Estelle Guéville (GUÉVILLE et WRISLEY, 2023). C'est une observation méthodologique très importante de la transcription des manuscrits médiévaux par traitement automatique. Les auteurs soutiennent que pour poursuivre de nouveaux types d'installations d'analyse par la recherche informatique avancée, nous devons théoriser les modes de transcription, non seulement pour la lecture humaine, mais aussi pour le traitement par machine

1. *eScriptorium* s'en sert du moteur d'HTR *open source*, Kraken, développé par Benjamin Kiessling dans le cadre du projet *Scripta* (PSL).

et cela peut être une nouvelle étape de l’histoire du processus de transcription des manuscrits médiévaux. en utilisant HTR.

De plus en plus émergent des articles sur la technique de la reconnaissance optique de caractères (OCR), notamment du moteur Kraken. SCHOEN et SARETTO, 2022 en est un exemple. Les auteurs sont parvenus à entraîner le modèle sur l’écriture manuscrite anglaise d’un scribe du XV^e siècle avec un taux de précision de 97 %.

3 Présentation et description des données sources

Le document qui sert de base à notre travail fut le premier roman médiéval imprimé (LAMBERT, 2017) : il s’agit de *Le roman Fierabras*, ou simplement *Fierabras*, du vaudois Jean Bagnyon. Il a été mis en prose d’après la chanson de geste *Li rommans de Fierabras d’Alixandre*, écrite au XII^e siècle, dont plusieurs manuscrits en diverses langues subsistent aujourd’hui². Pour ce travail, nous nous occuperons de l’*editio princeps*, publiée en 1478 par Adam Steinschaber, premier imprimeur à Genève.

Il existe actuellement six copies de cet incunable distribuées par plusieurs bibliothèques³, y compris deux digitalisées⁴.

En France, ce roman a été imprimé pour la première fois en 1485, à Lyon, par Guillaume le Roy et ce n’est qu’à partir de 1520 qu’il sera imprimée à Paris, par Michel Le Noir, sous le titre *La conquête du grant roy Charlemaigne des Espaignes. Et les vaillances des douze pers de France. Et aussi celles du vaillant Fierabras*.

Le roman est écrit en français moyen, marqué originellement par une *scripta* régionale romande, puis standardisée au fil des éditions (LAMBERT, 2017). Cette première édition dans un format *in-folio* possède 115 *folios* de 30/31 lignes chacun, rédigés dans un caractère gothique (G120 de Steinschaber⁵) et avec l’ajout des initiales à la main, comme l’on peut lire sur la notice bibliographique de la BNF⁶.

Écrite entre 1465 et 1470 à la demande du chanoine de Lausanne, Henri Bolomier, l’oeuvre se divise en trois parties, dont les récits de l’histoire des rois de France jusqu’à Clovis et, puis, du règne de Charlemagne, puisent dans le *Speculum historiale* de Vincent de Beauvais, d’après Brun (BRUN, 2020).

2. Voir une descriptions de certains disponibles en <https://www.arlima.net/eh/fierabras.html>

3. Ces copies sont distribuées en Bruxelles, à la Bibliothèque Royale (B 1363 Rés); en Suisse, à la Bibliothèque de Genève (Hf 350 Rés); quatre copies en France : Bibliothèque et archives du château de Chantilly (IV-G-037), Bibliothèque de l’Arsenal (RES FOL-B-932) et Bibliothèque Nationale de France (RES Y2- 76 et RES Y2- 20).

4. Une copie disponible à la Bibliothèque Nationale de France, et sur laquelle nous nous appuyons pour rédiger ce travail (consultable sur la page <https://gallica.bnf.fr/ark:/12148/btv1b8600180x/f9.item>) et l’autre à la Bibliothèque de Genève (https://www.e-rara.ch/gep_g/doi/10.3931/e-rara-33196).

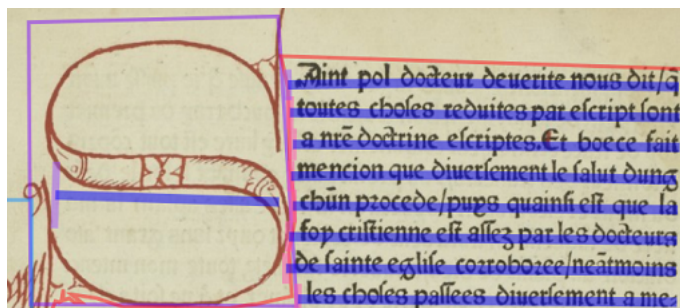
5. <https://tw.staatsbibliothek-berlin.de/ma05151>

6. <https://catalogue.bnf.fr/ark:/12148/cb33392022z>

4 *Fierabras* sur *eScriptorium*

4.1 Segmentation

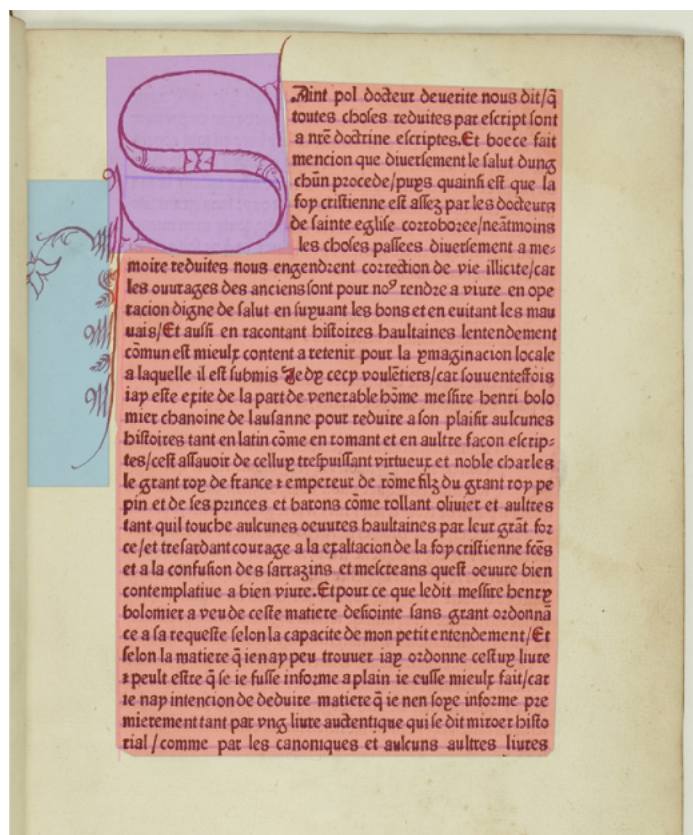
Le premier stade de traitement de notre image consiste en effectuer sa segmentation. Nous avons utilisé soit la segmentation manuelle soit la fonctionnalité de segmentation automatique proposée par le logiciel. Le taux d'exactitude pour les pages de texte simple est excellent. Pour les pages avec des ornements, nous avons effectué la segmentation manuellement.



En ce qui concerne l'ontologie, nous avons utilisé l'ontologie par défaut du logiciel, en ajoutant seulement deux balises pour définir les régions « Lettrine » et « Décoration » pour les zones que nous illustrons ci-dessous :

	Lettrine
	Décoration

Le document est ainsi divisé en trois régions : en violet, les lettrines ; en bleu, les décorations ; et en rouge, tout le texte :



4.2 Transcription

D'un point de vue paléographique notre texte est très facilement lisible et nous n'avons pas eu des difficultés majeures lors de l'exercice de transcription.


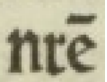
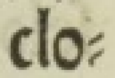
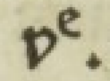
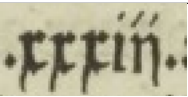
Étant donné que, d'un côté, « transcrire, c'est décrire un texte et sa mise en forme graphique » (STUTZMANN, 2011) et que, d'autre côté, l'objectif de la plateforme *eScriptorium* (et des travaux qui y sont menés) est l'entraînement de modèles HTR/OCR, nous avons décidé de faire une transcription aussi proche que possible du document original. Pour ce faire, nous n'avons pas procédé à la normalisation de l'orthographe ni à la résolution des abréviations⁷, en respectant, dans la mesure du possible, les caractères spéciaux. Cependant, dans certains cas, nous avons dû faire des choix pour des caractères qui ne sont pas exactement les mêmes que ceux représentés dans l'image à transcrire, et ce malgré l'incorporation du clavier Cremmalab⁸. Nous avons utilisé le modèle GalliCorpora+ (French Early Modern Print) en choisissant la *scripta* Latin et

7. « La résolution des abréviations étant un acte interprétatif lié à la spécificité de chacun des documents, elle relève d'un autre geste que celui de la prédiction textuelle » (PINCHE, 2022).

8. <https://github.com/HTR-United/cremma-medieval/blob/main/CremmaLab.json>

nous avons obtenu un taux d'exactitude de 96.6%. La plupart de problèmes de reconnaissance sont presque tous de caractère paléographique, dans le sens qu'ils avaient lieu lors que deux lettres se ressemblaient graphiquement, comme le *f* et le *f*(s long), le *r* et le *t*, ou encore le *e* et le *c*, par exemple.

Voici quelques exemples de décisions de transcription :

	7
	nrē
	clo=
	v ^e .
	.xxxiiñ.

5 En guise de conclusion : difficultés

Un certain nombre de difficultés que nous avons rencontrées étaient en partie liées au fonctionnement et la manipulation de l'interface d'*eScriptorium* elle-même. Par exemple, il y avait un problème d'ajout manuel de lignes dans un document. Les lignes s'ajoutaient de manière spontanée, puis il était difficile de les supprimer. La délimitation et transcription des lettrines aussi n'a pas été facile. De l'exportation de notre travail en fichiers ALTO - XML ont résulté quelques anomalies liées au fait que la transcription était modifiée lors de la conversion.

Concernant la partie Git, certains d'entre nous ont travaillé depuis un dépôt local et parfois il y a eu des conflits entre les versions des différentes branches locales et en distance ; d'autres petits problèmes par rapport aux contenus entre local et gitHub, n'ont pas été toujours faciles à résoudre.

Bibliographie

- BRUN, L. (2020). Jean Baignon. Récupérée 29 mars 2023, à partir de https://www.arlima.net/il/jean_bagnyon.html
- CLÉRICE, T. (2022). Ground-truth Free Evaluation of HTR on Old French and Latin Medieval Literary Manuscripts. Récupérée 5 avril 2023, à partir de <https://hal-enc.archives-ouvertes.fr/hal-03828529>
- GAUTIER, D., HUGUET, A., MASSOT, M.-L., TRICOCHÉ, A., CARLIN, M., MOREUX, J.-P., & AURÉLIA, R. (2022). Compte-rendu de la journée d'étude "Point HTR 2022" Transkribus / eScriptorium : Transcrire, annoter et éditer numériquement des documents d'archives. Récupérée 5 avril 2023, à partir de <https://hal.science/hal-03692413/document>
- GUÉVILLE, E., & WRISLEY, D. J. (2023). Transcribing Medieval Manuscripts for Machine Learning [arXiv :2207.07726 [cs]]. <https://doi.org/10.48550/arXiv.2207.07726>
- LAMBERT, A. (2017). Notice de Jean Bagnyon, Fierabras, Lyon, Guillaume Le Roy, 16 novembre [1484-1487], in-2° (Liège, Bibliothèque Alpha, XV.B119) [Publisher : Université de Liège. Transitions (Département de recherches sur le Moyen Âge tardif & la première Modernité), Liège, Belgium]. *Arm@rium Universitatis Leodiensis : la Bibliothèque Numérique du Moyen Âge et de la Première Modernité de l'Université de Liège*. Récupérée 29 mars 2023, à partir de <https://orbi.uliege.be/handle/2268/220013>
- LÖKKÖS, A. (1978). *Catalogue des incunables imprimés à Genève 1478-1500* [OCLC : 5282442]. Bibliothèque publique et universitaire. Récupérée 29 mars 2023, à partir de <https://bac-lac.on.worldcat.org/oclc/1032845626>
- MANDACH, A. d. (2014). La geste de 'Fierabras' ou le jeu du réel et de l'invraisemblable [Code : Au carrefour des routes d'Europe : la chanson de geste. Tome II]. In *Au carrefour des routes d'Europe : la chanson de geste. Tome II* (p. 843-857). Presses universitaires de Provence. <https://doi.org/10.4000/books.pup.2358>
- PINCHE, A. (2022). Guide de transcription pour les manuscrits du Xe au XVe siècle. Récupérée 29 mars 2023, à partir de <https://hal.science/hal-03697382>
- SCHOEN, J., & SARETTO, G. E. (2022). Optical Character Recognition (OCR) and Medieval Manuscripts : Reconsidering Transcriptions in the Digital Age [Publisher : Johns Hopkins University Press]. *Digital Philology : A Journal of Medieval Cultures*, 11(1), 174-206. <https://doi.org/10.1353/dph.2022.0010>
- STUTZMANN, D. (2011). Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin ? [Publisher : BoD], 247. Récupérée 29 mars 2023, à partir de <https://shs.hal.science/halshs-00596970>
- VIELLIARD, F., & GUYOYJEANNIN, O. (2014). *Conseils pour l'édition des textes médiévaux, fascicule I. Conseils pour l'édition des textes médiévaux :*

conseils généraux ([Nouvelle éd.] revue et mise à jour). Ecole nationale des Chartes.