Perceval Le Gallois
Reconnaissance automatique d'écriture à partir d'un imprimé d'ancien français de la légende du roi Arthur (XVIe siècle)

Donghan BIAN, José TORO BARBA, Gaëtan DROUET, Morgan HUART

30 mars 2024

1 Présentation générale du projet

1.1 Un projet réalisé via eScriptorium et Git/Github

Le projet HN-2023-Perceval_le_Gallois est un projet collectif pilote réalisé dans le cadre du master Humanités Numériques à l'Ecole nationale des Chartes, visant à tester le flux de travail de l'outil de Git par le biais d'un petit travail de Handwritten Text Recognition (HTR) et à créer des données potentiellement utilisables pour l'entraînement de modèle HTR dans le futur.

Dans ce projet, nous choisissons certaines pages extraites d'un manuscrit de Perceval le Gallois issu de l'année 1530 - une compilation arthurienne en prose - comme le texte de source pour la transcription. Dans le but d'assurer l'efficacité et la collaboration, nous avons employé différents outils durant ce projet selon leurs particularités : eScriptorium est une plateforme en open access en ligne permettant de réaliser la transcription à partir de certains manuscrits. Ainsi, le choix du modèle pour l'entraînement a été le modèle CatMus-Medieval 1.0.0, Pinche et al [1]., dont le corpus de référence est en ancien français. Puis, le logiciel Git permet l'annotation et la gestion des sources/fichiers locaux, alors que Github est un site web qui permet d'enregistrer et de contrôler les informations concernant les versions ainsi que les différentes mises à jour.

1. PINCHE A., et ali. "CATMus - Medieval : Consistent Approaches to Transcribing Manuscrits", Décembre 2023 [https://inria.hal.science/hal-04346939/document] - CLERICE Th., et ali., "CATMus Medieval, A multilingual large-scale cross-century datest in Latin script for handwritten text recognition and beyond [https://hal.parisnanterre.fr/hal-04453952v1]

Dans ce rapport, nous présenterons les détails de ce projet, y compris la présentation du texte de source et du modèle choisi, les choix en termes d'ontologie et de transcription, et les difficultés rencontrées, ainsi que les discussions concernant ce projet.

1.2 La présentation du manuscrit et du modèle

1.2.1 Manuscrit : Perceval Le Gallois

La source qui a été choisie pour ce projet est un imprimé intitulé Tresplaisante et recreative histoire du trespreulx et vaillant chevallier Perceval le Galloys, conservé à la bibliothèque de l'Arsenal en réserve, et accessible sur Gallica [2]. Il est écrit en ancien français, il est édité par Galliot Du Pré, Longis Jean, Saint-Denis et Aubry Bernard.Il a été imprimé par Jehan sainct Denys et Jehan Longis en 1530. Cet ouvrage contient les textes suivants : Conte du Graal de Chrétien de Troyes, ses deux prologues apocryphes Élucidation et Bliocadran, et deux continuations, la Première et la Deuxième, ainsi que celle de Manessier [3].

1.2.2 Modèle : CatMus-Medieval 1.0.0

Le premier modèle de transcription que nous avons utilisé a été le modèle HTR-United-Manu McFrench V3 qui propose un datasets composé de 18 155 images et de fichier XML ainsi que 41,5 millions de caractères couvrant 13 langues. Cependant ce modèle est utilisé dans la transcription de texte en français [4] moderne. De ce fait, le modèle le plus adéquat pour notre projet est le modèle : CatMus-Medieval 1.0.0, Pinche et al. L'entraînement de ce modèle est basé sur un corpus dont une partie importante est en ancien français. Dans l'évaluation faite par Pinche et al., ce modèle a obtenu une meilleure performance dans les tâches de français par rapport à d'autres modèles. Par conséquent, il est pertinent d'adopter ce modèle dans le but d'exécuter notre projet.

1.2.3 Ontologie de segmentation des Zones : SegmOnto [5]

Pour la segmentation des zones nous nous sommes référés aux présentations de Camps J-B., et Pinche A., qui propose une SegmOnto par zones et par lignes. En effet, la mise en page de l'ouvrage est structuré de la même manière, les composants que nous

2. Notice de l'ensemble de l'ouvrage : http://ark.bnf.fr/ark:/12148/cb30925590q
3. Notice de titre conventionnel : http://ark.bnf.fr/ark:/12148/cb171503036
4. CHAGUÉ A., et ali., "Manu McFrench, from zero to hero : impact of using a generic handwriting recognition model for smaller datasets", 2022, p.1-7. [https://inria.hal.science/hal-04094241/document]
5. CAMPS J-B., SegmOnto, SegmOnto, 10 déc. 2021 [https://hal.science/hal-03481089/file/SegmOnto.pdf] - PINCHE A.,"Des images au texte : comment apprendre à des ordinateurs à lire des manuscrits médiévaux ?" [https://hal.science/hal-03585216/file/CommentApprendreAuxOrdi-2.pdf]

constatons sur les pages sont comme suivant : les textes principaux, les drop capitals, les running titles, et les quire marks. Il est donc pertinent d'employer cette ontologie pour faire la segmentation.

2 Méthodologie

2.1 Organisation fonctionnelle et aspects techniques

2.1.1 eScriptorium

Afin de faciliter le travail, nous avons choisi les pages qui doivent être traitées dans notre projet sur IIIF et les avons extraites sous forme de PDF depuis IIIF. Ce document PDF a été ensuite importé sur la plateforme eScriptorium, où une collaboration de rédaction et transcription a été réalisée. La plateforme eSscriptorium offre des fonctionnalités puissantes pour segmenter les zones différentes sur une certaine page selon leur nature, délimiter les espaces de textes, et ordonner les lignes de textes. Après la segmentation automatique, les ajustements manuels ont été appliqués pour préciser les frontières de zones. Cette opération permet d'améliorer la précision de détection de modèle. Sur cette base, les examens manuels ont été effectués afin de corriger les erreurs produites pendant la transcription automatique et d'appliquer les critères spécifiques de transcription établies dans notre projet. La saisie de certains caractères spécifiques dans l'écriture de l'ancien français a été réalisée par le biais d'un clavier personnalisé importé dans eScriptorium. Après avoir terminé la transcription, les fichiers ALTO ont été exportés depuis la plateforme.

2.1.2 Git et Github

Git et Github servent ensemble comme une suite d'outils puissants de collaboration et de gestion des fichiers qui a été très utile dans notre projet. Il permet aux membres de collaborer et de travailler de manière asynchrone, ces derniers peuvent enregistrer les informations à chaque étape et opération de rédaction. Chaque collaborateur de ce projet a traité respectivement deux pages de manuscrit. Dans un premier temps les fichiers ont été enregistrés dans le répertoire local de Git, puis ils ont été soumis au répertoire collectif de Github à distance par le biais de branches indépendantes. Cette opération a assuré la sécurité de répertoire total et la traçabilité de chaque opération. De ce fait, les autres fichiers, tels que les sources brut, la transcription sous format txt, ont été soumis dans les branches indépendantes. Après que tous les fichiers ont été collectés dans le répertoire à distance, les branches ont été fusionnées pour créer un répertoire complet.

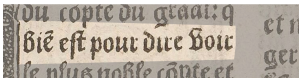

FIGURE 1 – Folio 1 ligne 8

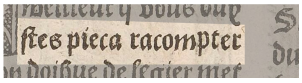

FIGURE 2 – Folio 1 ligne 11

2.2 Réalisation de transcription

2.2.1 Etablissement des normes de transcription

Nous avons choisi de ne pas corriger les fautes d'orthographe, ni de la standardiser ; nous n'avons pas corrigé les oublis de caractère (par exemple, point manquant entre deux phrases). Nous avons gardé les abréviations et les différents caractères présents dans le texte, afin d'être fidèles à l'écriture originale en restant dans les limites du pratique. Ainsi, nous avons conservé les s longs et courts, la différentiation entre les u et les v, nous avons transcrit les virgule par des /, et conservé tous les diacritiques présents dans le texte. De la même façon, nous n'avons pas tenté de "corriger" les i lorsqu'ils deviendrait des j en orthographe moderne.

2.2.2 Les problèmes rencontrés s/s long et u/v

Lors de la transcription des textes, nous avons eu quelques difficultés dans la transcription graphique des caractères u/v ainsi que les caractères s et le s long.

Hétérogénéité du corpus CatMus-Medieval 1.0.0, Pinche et al. : Dans le cadre de notre projet, le modèle choisi n'a pas permis de présenter une égalité suffisante dans le traitement : u/v ainsi que s/s long sont parfois distingués par valeur phonétique et mais il existe dans certains cas une distinction graphique.

Transcription et paléographie : En ce qui concerne la question de la paléographie de cet imprimé, nous avons décidé de transcrire le texte au plus proche de la réalité afin de ne pas dénaturer le texte et de conserver toutes les informations paléographiques qu'il pourrait nous fournir. L'article de D. Stutzmann prend le parti de restituer le texte selon sa valeur phonétique et non par rapport à sa graphie.

2.2.3 Utilisation de la plateforme eScriptorium

La prise en main de la plateforme eScriptorium est dans un premier temps assez difficile et déroutante.

Segmentation : L'outil pour dessiner la ligne de texte est compliqué à prendre en

main, car le moindre clic crée une nouvelle ligne. De plus, la segmentation automatique est difficile et on a du mal à séparer les deux colonnes de textes, de ce fait une segmentation manuelle présente de meilleurs avantages.

Transcription : Les transcriptions obtenues sont pour la plupart pertinentes et conformes au texte initial. Néanmoins, il existe des confusions entre certaines lettres v/u ou encore s/s longs. Nous avons pris le parti de distinguer les lettres pour respecter la graphie du texte imprimé.

3 Discussion

Dans ce projet, nous réalisons une tâche HTR à travers d'un flux de travail combinant eScriptorium et Git/Github. Et voici quelques réflexions que nous avons pu tirer depuis nos opérations en tant que discussions et conclusions :

Le choix de modèle pour la transcription automatique : le modèle initial que nous avons choisi, comme indiqué au-dessus, n'est pas adapté au manuscrit d'objet dans notre projet, ce qui souligne l'importance du rôle de modèle au sein des projets HTR. La convenance entre le modèle et le projet doit être validée par le biais d'essais de plusieurs modèles.

Les inconvénients techniques d'eScriptorium : comme indiqué au-dessus, certains outils offert par défaut par eScriptorium ne sont pas aussi facile à opérer, ce qui empêche dans un certain sens l'efficacité d'opération dans l'exécution de projet. De plus, les bogues inattendues sont aussi présentes dans les phases différentes de l'opération sur eScriptorium. Ce serait plus efficace si eScriptorium dispose des fonctions perfectionnées et de la prise en charge des plugins.

Les avantages de Git/Github : Git et Github sont souvent considérés comme des outils destinés aux projets de programmation. Leurs avantages sur le contrôle de versions et l'annotation des informations sont aussi applicables dans les tâches HTR. La potentialité de Git/Github en termes de gestion de fichiers est énormément explorable dans ce genre de tâches.

En conclusion, malgré quelques inconvénients rencontrés dans le flux de travail, la combinaison de eScriptorium et Git/Github a bien complété les besoins de tâches HTR et peut bien servir comme des supports nécessaires pour ce genre de travail.

Références

[1] CHAGUÉ, A. et al. "Manu McFrench, from zero to hero : impact of using a generic handwriting recognition model for smaller datasets", 2022, p.1-7. https://inria.

hal.science/hal-04094241/document.

[2] CLERICE, Th. et al. “CATMus Medieval, A multilingual large-scale cross-century dataset in Latin script for handwritten text recognition and beyond”. https://hal.parisnanterre.fr/hal-04453952v1.

[3] CAMPS, J-B. SegmOnto, SegmOnto, 10 déc. 2021. URL : https://hal.science/hal-03481089/file/SegmOnto.pdf.

[4] PINCHE, A. et al. “CATMus - Medieval : Consistent Approaches to Transcribing Manuscrits”, Décembre 2023. https://inria.hal.science/hal-04346939/document.

[5] PINCHE, A. “Des images au texte : comment apprendre à des ordinateurs à lire des manuscrits médiévaux ?”. https://hal.science/hal-03585216/file/CommentApprendreAuxOrdi-2.pdf.

[6] Notice bibliographique. (s. d.). Retrieved from http://ark.bnf.fr/ark:/12148/cb30925590q (Accessed on 2024-03-21).

[7] Notice de titre conventionnel. (s. d.). Retrieved from http://ark.bnf.fr/ark:/12148/cb171503036 (Accessed on 2024-03-21).

Acronymes

CatMus-Medieval Consistent Approaches to Transcribing ManuScripts. 1, 2, 4

HTR Handwritten Text Recognition. 1, 5