Rapport de projet

3 avril 2024

Consigne : Création de set de données ou exercice en équipe sous la forme d'un dépôt GitHub.
Equipe : Arsène Géry, Francesca Hemery, Nana Maglakelidze, Francesco Savatteri, Albina Toumarkine
Source : *Le livre des faits de messire Bertrand du Guesclin*, édition de 1487, Bibliothèque nationale de France, Réserve des livres rares, RES-Y2-91

1 Introduction

Ce document est un compte-rendu du travail de groupe, consistant en une transcription réalisée sur e-Scriptorium d'une partie du Livre des Faits de messire Bertrand du Guesclin. Nous commencerons par une brève présentation du document, puis nous évoquerons les relations aux jeux de données existants d'autres projets de HTR du Moyen- ge tardif. Ensuite, nous exposerons nos choix de transcriptions et d'ontologie, illustrés par des extraits du manuscrit. Nous conclurons en dressant un bref retour sur expérience, et enfin nous mentionnerons notre bibliographie.

2 Présentation de la source

La manuscrit traité est une édition imprimée du Moyen- ge tardif, intitulé Livre des Faits de messire Bertrand du Guesclin par Guillaume Le Roy, publié en 1487 à Lyon. Cet ouvrage est rédigé en Moyen Français, et retrace la vie et les faits militaires de Bertrand du Guesclin, membre de la noblesse bretonne et personnage majeur de la guerre de Cent Ans. Le manuscrit est imprimé en caractères gothiques, qui sont caractéristiques des ouvrages imprimés de la fin du Moyen- ge. Nous avons transcrit les pages 14 à 23 du manuscrit, chaque membre ayant transcrit deux pages.

3 Le jeux existants de données

Traditionnellement, on distingue deux types de transcriptions : les transcriptions allographétique, qui visent à donner accès à toutes les formes de chaque lettre ou signe en transcrivant le plus fidèlement possible ce qui est inscrit sur le texte source, et les transcriptions graphématiques, qui préservent la suite des lettres et réduisent chaque forme à son sens dans un système alphabétique, en privilégiant la lisibilité de la transcription. Cependant, les projets de transcriptions de textes pour l'élaboration de corpus d'entraînement à l'HTR dépassent en partie cette dichotomie entre méthode allographétique et méthode graphématique. En effet, l'objectif est de produire des données suffisamment simples et unifiées pour être reproduites par des technologies de HTR, et en même temps de conserver les unités sémantiques essentielles. Ariane Pinche adopte des normes des deux types de transcriptions, et définit des "transcriptions graphématiques qui conservent les abréviations et la ponctuation originale nous ont semblé être les plus adaptées." (Ariane Pinche, "Guide de transcription pour les manuscrits du Xe au XVe siècle", 2022.)

Le projet CREMMALAB est un des projets fondateurs de transcription de manuscrits médiévaux pour la constitution de données d'entraînement à l'HTR sur des manuscrits d'ancien et de moyen Français. Les ressources mises en ligne dans le cadre du projet CREMMALAB ont été des documents essentiels pour l'élaboration des normes de transcription pour notre projet. Plus précisément, le clavier de caractères spéciaux présent sur le dépôt Github du projet, le guide de transcription d'Ariane Pinche et le compte-rendu de séminaire "L'allographie, entre besoins scientifiques et pragmatiques. Comment modéliser et optimiser les données d'entraînement pour l'HTR (I) ?" nous ont été particulièrement utiles, et reprennent des normes de transcription similaires, ou en tout cas régies par un même principe, celui de fournir des transcriptions fidèles à la source, mais ne tenant pas compte de certaines subtilités graphiques difficilement détectables par les technologies d'HTR.

Globalement, le guide d'Ariane Pinche et le compte-rendu du séminaire s'axent sur divers points communs :

- les signes spéciaux : qu'il soit question de signes fonctionnels (pied de mouche, crochet alinéaire) ou de signes diacritiques (tilde, cédille, macron), les deux guides préconisent de réunir sous un même symbole les signes graphiquement similaires (qui se ressemblent) ou sémantiquement identiques (qui veulent dire la même chose).

- la question délicate des espaces, et plus généralement l'enjeu de la seg-

mentation des mots : en effet, les textes médiévaux utilisent les espaces de façon arbitraire, laissant le lecteur déchiffrer le texte à l'aide de sa connaissance sémantique. Ariane Pinche préconise de séparer les mots sémantiques.

- la ponctuation : la variété des signes de ponctuation dans les manuscrits médiévaux est simplifiée et rendue par deux ou trois signes : la virgule, le point-virgule et la virgule.
- les majuscules et les lettrines : la distinction entre ces deux éléments est simplifiée au profit de la représentation unique d'une majuscule (Ariane Pinche).

4 Les choix en termes d'ontologie et de transcription

Globalement, nous avons fait le choix de suivre le guide d'Ariane Pinche. Cependant, nous avons souhaité distinguer le tilde et le macron (tilde droit) dans l'ensemble de nos transcriptions. Concernant les espaces, un sujet débattu dans la littérature nous avons choisi de suivre les indications d'Ariane Pinche (Pinche 2022).

5 Retours d'expérience

Au terme de ce travail, nous pouvons constater que ce projet nous a offert l'opportunité d'acquérir de nouvelles expériences et compétences en Humanités numériques. L'expérience pratique sur un projet de sources numériques nous a incités à collaborer efficacement autour d'objectifs et d'enjeux définis collectivement.
Tout au long du projet, notre utilisation de eScriptorium a été grandement facilitée par l'interface accessible de la plateforme. Les performances des modèles de segmentation et de transcription se sont révélées assez élevées pour l'incunable que nous avons étudié. Par ailleurs, nous avons remarqué la flexibilité d'utilisation offerte par la plateforme : la possibilité de travailler en groupe, d'importer aisément un clavier personnalisable, ainsi que la variété des choix concernant les sources et les formats des documents, sont des caractéristiques qui font de eScriptorium une plateforme adaptée à une grande diversité de cadres de recherche.

Ce projet nous a surtout permis de collaborer avec Git et Github, et pour certains d'entre nous, de découvrir comment Git peut contribuer au bon déroulement d'un projet. En cela, la transcription de sources s'est avérée être une expérience très pertinente, puisqu'elle nécessite une gestion continue de différentes versions d'un même document, d'autant plus dans le cadre d'un projet de groupe.
Au-delà des aspects pratiques évidents, nous avons également remarqué en quoi la collaboration via Git et Github pouvait faciliter l'organisation d'un travail en groupe, en incitant chacun à suivre certaines bonnes pratiques. Cela inclut le suivi et la vérification des différentes versions, la réponse aux objectifs clairement identifiés à travers les issues, ainsi que la communication efficace des contributions aux autres membres du projet.

Références

Camps (Jean-Baptiste), Vidal-Gorène (Chahan) et Vernet (Marguerite), "Handling Heavily Abbreviated Manuscripts : HTR engines vs text normalisation approaches", dans *International Conference on Document Analysis and Recognition 2021*, 2021, p. 306, doi : 10.1007/978-3-030-86159-9_21.

Clérice (Thibault), "Ground-truth Free Evaluation of HTR on Old French and Latin Medieval Literary Manuscripts", dans *Workshop on Computational Humanities Research*, 2022, url : https://www.semanticscholar.org/paper/Ground-truth-Free-Evaluation-of-HTR-on-Old-French-Cl%C3%A9rice/8d34d21cd69a2a49b0be94795d384e79972fec19 (visité le 28/03/2024).

Compte-rendu de la séance n° 3, url : https://cremmalab.hypotheses.org/compte-rendu-seance-3 (visité le 28/03/2024).

cremma-medieval/htr-united.yml at main · HTR-United/cremma-medieval, url : https://github.com/HTR-United/cremma-medieval/blob/main/htr-united.yml (visité le 28/03/2024).

Guéville (Estelle) et Wrisley (David Joseph), *Transcribing Medieval Manuscripts for Machine Learning*, oct. 2023, doi : 10.46298/jdmdh.9805, arXiv : 2207.07726 [cs].

Pinche (Ariane), *Guide de transcription pour les manuscrits du Xe au XVe siècle*, juin 2022, url : https://hal.science/hal-03697382 (visité le 28/03/2024).

Schoen (Jenna) et Saretto (Gianmarco E.), "Optical Character Recognition (OCR) and Medieval Manuscripts : Reconsidering Transcriptions in the Digital Age", *Digital Philology : A Journal of Medieval Cultures*, 11–1 (2022), p. 174-206, url : https://muse.jhu.edu/pub/1/article/853521 (visité le 28/03/2024).