Entraînement *from scratch* ou *fine-tuning* ? Modalités et enjeux d'un choix d'un projet HTR sur les registres de plaidoiries du Parlement de Paris (fin XIV^{e}-XV^{e} siècle)

À l'heure de la multiplication des projets de recherche cherchant à exploiter d'importants corpus de données archivistiques à l'aide de la transcription automatique d'écritures manuscrites, la question des investissements nécessaires pour ce faire apparaît d'autant plus sensible. Il ne s'agit évidemment pas seulement d'investissements financiers et d'institutions soutenant le projet, mais également du coût humain, en temps comme en compétences et formations, que computationnel, avec l'accessibilité à une plateforme de transcription et la possibilité d'entraîner des modèles sur des serveurs locaux. La présentation se concentrera en particulier sur les modalités de réutilisation de modèles de segmentation et de transcription de sources médiévales, disponibles sur deux plateformes, Transkribus et eScriptorium. Plusieurs modèles, issus de précédents travaux ayant publiés leurs jeux de données et métadonnées selon les principes FAIR, sont aujourd'hui disponibles en ligne sur ces plateformes. Cependant, le choix des modèles ainsi les tests d'entraînement afin de tester la solidité de ceux-ci ou le choix de développer des modèles *from scratch* demeurent des problématiques importantes à traiter dans le temps limité d'un projet de recherches. Dans le cas du choix de la réutilisation de modèles déjà disponibles, la création de la proximité des jeux de données avec les sources originales sera étudiée car elle pose différentes questions parmi lesquelles : selon quelles spécificités du corpus d'origine l'usage d'un méta-modèle de transcription, agrégeant les données de plusieurs projets scientifiques antérieurs, est-il pertinent pour affiner ce modèle sur un nouveau corpus de sources ? Est-ce plus rentable, en termes d'investissement en temps et en coût humain, de travailler sur un modèle entraîné sur un corpus de sources moins important mais peut-être plus proche du corpus de sources d'origine.

Ces questionnements sous-jacents accompagnent un projet de recherches mêlant histoire numérique et histoire médiévale, sur les registres de plaidoiries du Parlement de Paris à la fin du Moyen Âge. Il s'agit d'acquérir par l'HTR le texte non-corrigé des conflits judiciaires du début du XV^{e} siècle relevant du ressort de cette cour de justice afin d'y analyser la place des universitaires dans le milieu urbain. Si la mise en page des sources ne présente pas de difficultés irrémédiables, le nombre de pages à traiter, de l'ordre de plusieurs milliers, rend particulièrement sensible les questions liées à la flexibilité du modèle d'entraînement à obtenir.

Pauline Spychala

pspychala@dhi-paris.fr
https://orcid.org/0000-0002-0899-2046

Postdoctorante en histoire numérique à l'Institut historique allemand, docteure en histoire médiévale, membre associée au CRHEC (EA 4392, Créteil), qualifiée aux fonctions de maître de conférences CNU 21 (2022)

Parcours Professionnel

Depuis 2022 **Postdoctorante en histoire numérique** à l'Institut historique allemand (Paris)[1]
Depuis 2022 **Chargée d'enseignement** à l'Université Sorbonne Nouvelle
2023–2024 **Chargée d'enseignement** à l'Université Paris-Est Créteil
2021–2022 **ATER** à l'Université de Bretagne-Sud Lorient
2018–2020 **Chargée d'enseignement** à l'Université Paris-Est Créteil
2017–2020 **Doctorante contractuelle** à l'Université Paris-Est Sup
2015–2016 **Assistante ingénieur** à l'Institut de Recherche en Musicologie (UMR 8223, CNRS/BnF)

Formations et titres universitaires

2017–2021 **Doctorat en histoire médiévale, label « Doctorat Européen »**[2]
Sujet : « *Omnibus qui causa studiorum peregrinantur*. Mobilités sociales et géographiques des universitaires allemands, hongrois et slaves des universités françaises (1330-1500) »

2015–2017 **Master Études Médiévales : Littérature, textes et savoirs**, mention Très Bien
Co-habilité Universités Paris Sorbonne et Sorbonne Nouvelle, École nationale Supérieure Ulm, École nationale des Chartes
2e année en Erasmus+ à l'Université Ruprecht-Karl de Heidelberg (Allemagne)

2012–2015 **Licence Histoire, spécialité Histoire – Parcours Livres et documentation**, mention Bien
Double licence : **Histoire et Lettres Classiques** en première année
Université de Haute Bretagne Rennes II

Thèmes de recherche

- Transcription automatique d'écritures manuscrites (HTR)
- Histoire numérique
- Saint Empire et Europe centrale
- Universités médiévales
- Mobilités universitaires

Compétences informatiques

- C2I niveau 1 acquis
- Nettoyage de données : OpenRefine
- Langages de programmation : Html, Css, Tei, Python
- Bases de données et statistiques : Nodegoat, SPSS
- Visualisation géographique SIG : QGis, Magrit
- Reconnaissance automatique de textes : eScriptorium (Kraken), Transkribus

Dernière publication

[1] https://www.dhi-paris.fr/fr/recherche/histoire-numerique/processus-dintegration-des-universitaires-etrangers-a-paris.html.

[2] Les démarches d'obtention du diplôme de la Westfälische Wilhelms-Universität Münster sont en cours. Les établissements allemands délivrant encore des mentions, la thèse a reçu celle de *magna cum laude* (très honorable).

Roberto Berardinelli, Marie-Astrid Hugel, Ulrich Niggemann, Pauline Spychala (dir.), *Politisches Scheitern in der Vormoderne. Ein ambivalentes Phänomen in Europa (11. bis 18. Jahrhundert)*, Berlin, De Gruyter, 2023 ; DOI : https://doi.org/10.1515/9783111087122.