Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

# Reconnaissance automatique de thématiques dans une correspondance historique

**Équipe ou projet dans le laboratoire :** ADA (Algorithmique Discrète et Applications : https://ligm.u-pem.fr/equipe/ada/)

**Nom et adresse e-mail du tuteur :** Philippe Gambette, philippe.gambette@univ-eiffel.fr

**Filières visées :** Informatique, Datascience et intelligence artificielle

## Présentation générale du sujet

Ce projet vise à concevoir une approche de reconnaissance automatique d'extraits liés aux thématiques de la famille ou du pouvoir dans les lettres écrites par Catherine de Médicis. L'approche proposée vise d'une part à utiliser un modèle de langue pour transformer en vecteurs les mots des lettres, d'autre part à tester l'intérêt de la prise en compte des métadonnées relatives aux lettres pour améliorer la qualité de l'étiquetage automatique.

Ce travail sera encadré par Philippe Gambette, enseignant-chercheur en informatique au LIGM, et sera mené en collaboration avec Beatrice Mundo, étudiante en deuxième année de master LSCN (Littérature, savoirs et culture numérique) à l'université Gustave Eiffel, qui a déjà procédé à la création du corpus numérique de lettres et à son étiquetage, et dont le mémoire de master est encadré par Caroline Trotot, enseignante-chercheuse en littérature au LISAA.

## Objectif du projet

La problématique proposée correspond à une tâche d'apprentissage supervisé pour un étiquetage sémantique, qui consiste à déterminer, pour chaque phrase des lettres, si elle est classée, ou non, dans la catégorie « famille », ou dans la catégorie « pouvoir ». La solution développée s'appuiera sur un ensemble de 90 lettres déjà étiquetées pour ces deux thématiques, qui sera découpé en corpus d'entrainement, de validation et de test.

Un enjeu particulier concerne le fait que les lettres sont écrites en moyen français, c'est-à-dire avec les graphies et la syntaxe du français du XVIe siècle. Ainsi, deux approches pourront être comparées : procéder à une modernisation automatique préalable à l'utilisation d'un modèle de langue entrainé sur des textes en français contemporain, ou utiliser directement un modèle de langue entrainé sur du moyen français. Par ailleurs, des métadonnées supplémentaires sur les lettres, notamment le type de destinataire, pourront être ajoutées en vue de tester si ceci améliore la qualité des résultats. Un autre enjeu sera l'explicabilité de la méthode développée : en particulier, fournir un indice sur les mots de la phrase ou les métadonnées qui ont conduit à sa catégorisation dans les thématiques « famille » ou « pouvoir », sera particulièrement utile pour la vérification de la pertinence de tels passages trouvés dans des lettres pas encore étiquetées.

Les solutions implémentées seront programmées en Python, sous licence libre, en ayant recours à divers algorithmes existants d'apprentissage automatique. Des modèles de langue spécifiques au français, comme CamemBERT, FlauBERT, ou DAlemBERT, pour le moyen français, pourront aussi être utilisés.

Une extension de ce travail pourra concerner l'adaptation de cette approche à des textes obtenus par reconnaissance optique de caractères (OCR) sur une édition du XIXe siècle de ces lettres, et donc présentant éventuellement des erreurs d'OCR.

## Bibliographie


- Malte Ostendorff, Peter Bourgonje, Maria Berger, Julian Moreno-Schneider, Georg Rehm, Bela Gipp (2019), "Enriching BERT with Knowledge Graph Embeddings for Document Classification", *KONVENS 2019* (*Proceedings of the 15th Conference on Natural Language Processing*)
- Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, Prithviraj Sen, "A Survey of the State of Explainable AI for Natural Language Processing", *AACL 2020* (*Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*), p. 447-459
- Chuanming Dong, Philippe Gambette, Catherine Dominguès (2021), "Extracting event-related information from a corpus regarding soil industrial pollution", *KDIR 2021* (*Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*), volume 1, p. 217-224
- Rachel Bawden, Jonathan Poinhos, Eleni Kogkitsidou, Philippe Gambette, Benoît Sagot, Simon Gabay (2022), "Automatic Normalisation of Early Modern French", *LREC 2022* (*Proceedings of the 13th Language Resources and Evaluation Conference*), p. 3354-3366
- Simon Gabay, Pedro Ortiz Suarez, Alexandre Bartz, Alix Chagué, Rachel Bawden, Philippe Gambette, Benoît Sagot (2022), "From FreEM to D'AlemBERT: a Large Corpus and a Language Model for Early Modern French", *LREC 2022* (*Proceedings of the 13th Language Resources and Evaluation Conference*), p. 3367-3374