Analyse des registres manuscrits de la comédie-Italienne au XVIIIe siècle : de l'image au langage

Granet Adeline[1], Roman-Jimenez Geoffrey[1], Rubellin Françoise[2], Quiniou Solen[1], Morin Emmanuel[1], Mouchère Harold[1], Viard-Gaudin Christian[1]
E-mail: adeline.granet@univ-nantes.fr

1. UMR 6004, LS2N, Université de Nantes, 44000, Nantes, France.
2. EA 4276, L'AMO, Université de Nantes, 44000, Nantes, France.

Résumé

La Comédie-Italienne et les théâtres de la Foire sont deux formations théâtrales du XVIIIe siècle encore largement sous-étudiées dans le domaines des sciences humaines et sociales. Le projet CIRESFI vise à analyser l'acculturation des italiens dans le contexte politique de l'époque.
Les registres comptables de la Comédie-Italienne sont aujourd'hui disponibles au format numérique et l'extraction automatique de ces informations, encore inexploitées, permettra de faciliter leur analyse.
Nos travaux se concentrent sur la transcription des titres des pièces. Celle-ci repose sur deux axes consécutifs : la segmentation automatique des titres et leur transcription par reconnaissance semi-supervisée d'écriture.
Les zones de titres ont été extraites au travers de l'utilisation de la méthode DMOS pour la reconnaissance de structure des documents. La séparation entre les lignes de textes est effectuée à l'aide de l'algorithme *Seam Carving*.
La transcription des lignes de titre est réalisée à partir d'un réseau de neurones récurrent profond et d'un étiquetage par classification temporelle (BLSTM-CTC). Pour une transcription non-supervisée des titres, ce modèle est entraîné sur différentes ressources annotées du XVIIIe siècle.
Ces premiers travaux basés sur les titres de pièces permettront d'enrichir notre modèle en vue de transcrire la totalité des informations disponibles au sein de ces documents.