

eBalzac

Analyse des relations hypertextuelles entre presse et roman dans l'œuvre de Balzac

1 Information

- Encadrant :
 - Andrea Del Lungo, Professeur de Littérature française, UMR CELLF, Sorbonne Université
- Lieu :
 - Initiative Humanités numériques/Obtic, Sorbonne Université, Faculté des Sciences, Jussieu
 - Maison de la recherche, Sorbonne Université - 28 rue Serpente, 75006 Paris
- Durée : 3 ans (date de début souhaitée : automne 2026)
- Financement : Contrat doctoral de l'Initiative Humanités Numériques, Sorbonne Université.
- Salaire brut mensuel : 2300 €
- Mots clés : Humanités numériques littéraires ; Reconnaissance et correction automatiques de caractères ; Édition numérique et hypertextuelle, OCR/HTR (presse XIXe) ; post-correction automatique ; détection de réemplois ; alignement de textes ; XML-TEI.

2 Contexte

Le projet eBalzac consiste à mettre en résonance *La Comédie humaine* de Balzac avec un vaste corpus d'écrits contemporains qui ont pu la nourrir : œuvres romanesques d'autres auteurs de l'époque ; recueils collectifs de littérature panoramique (auxquels Balzac apporta des contributions) ; ouvrages scientifiques susceptibles d'avoir influencé l'écriture balzacienne, notamment dans les domaines de la médecine, de la physiologie et des sciences naturelles. Son objectif est de permettre des recherches et des comparaisons intertextuelles élaborées, à l'intérieur de l'œuvre de Balzac, et dans le corpus plus vaste de textes littéraires et scientifiques de l'époque, entre 1800 et 1850, afin de faire émerger des correspondances, de repérer des emprunts, des citations, des reprises, des plagiat éventuels, et de constituer ainsi une cartographie de l'univers intellectuel de Balzac à partir des traces que d'autres textes ont laissées dans l'œuvre.

Piloté par Andrea Del Lungo, le programme a bénéficié d'un financement ANR sur une période de 4 ans, de 2015 à 2019. Il est actuellement rattaché à l'Obtic (Observatoire des Textes, des Idées et des Corpus) et à l'Initiative humanités numériques de Sorbonne Université.

Le site internet de référence, ebalzac.com, ouvert en 2017, est articulé autour de trois axes :

- [L'édition électronique](#), à commencer par *La Comédie humaine*, dans une version inédite en ligne et philologiquement exacte, qui intègre les corrections apportées par Balzac sur son exemplaire personnel.
- [L'étude génétique](#) des différents textes balzaciens, comparés à l'aide du logiciel Médite, qui permet de visualiser les principales opérations génétiques et d'étudier le processus de création chez Balzac.
- [L'étude de l'hypertexte](#) qui constitue un objet scientifique expérimental et nouveau, rappelé supra, en voie de construction.

3 Sujet de thèse

Le site ebalzac présente l'édition de *La Comédie humaine* et du théâtre de Balzac. Dans la perspective de pouvoir offrir à l'utilisateur l'accès aux œuvres complètes de Balzac, actuellement indisponibles même en version papier, une partie importante de sa production mérite d'être valorisée : il s'agit des textes journalistiques, que Balzac publia abondamment dans la presse de l'époque, dans des revues ou des quotidiens.

Une partie de ces textes ont été publiés dans les deux volumes des *Œuvres diverses* de la collection de la Pléiade, mais une autre grande partie reste à éditer : elle correspond aux publications dans la presse après 1834, et jusqu'à la mort de l'auteur.

Le travail demandé dans le cadre de ce contrat doctoral serait d'ordre éditorial : il consiste à numériser les textes à partir des originaux, via OCR ou des logiciels d'IA, afin d'obtenir une version parfaitement corrigée du point de vue philologique, à encoder en XML-TEI pour qu'elle soit versée dans le site. Sur le plan technique, le travail mobilisera une chaîne OCR/HTR adaptée à la presse du XIXe siècle et des méthodes de post-correction automatique assistée, avec validation philologique. La chaîne de traitement (OCR/HTR → correction → TEI) sera versionnée et documentée afin d'assurer la reproductibilité et la traçabilité des transformations. Une version diplomatique (philologique) et une version normalisée (pour analyses TALN) seront maintenues en parallèle, avec alignement entre les deux.

À partir de ce travail éditorial, une perspective d'ordre génétique (analyse du parcours de création de l'auteur) sera développée dans le sens de l'édition hypertextuelle décrite supra, cette fois, non dans la recherche de sources extérieures à l'œuvre, mais plutôt vers l'analyse de son hypertextualité interne. On sait que Balzac réutilise dans les romans des fragments textuels initialement parus dans la presse, et que son immense œuvre est le fruit d'un travail de « montage ». La détection des réemplois et des passages d'une partie à l'autre de l'œuvre permettrait d'ouvrir de nouvelles perspectives de recherche, et de construire à l'intérieur du site ebalzac un parcours intertextuel sous la forme d'un réseau. Cette détection pourra s'appuyer sur des approches de similarité textuelle et d'alignement de passages, afin d'identifier des reprises exactes ou réécrites. Il pourrait aussi, dans ce sens, fournir un modèle de représentation graphique applicable à d'autres corpus. Un protocole d'évaluation et de traçabilité (échantillon de référence, suivi qualité, validation) sera mis en place tout au long de la chaîne OCR → correction → TEI.

Ce corpus journalistique pourrait ainsi fournir dans le cadre de la thèse un matériau inédit pour le développement d'autres types de recherche, par exemple l'extraction d'entités nommées,

l'analyse des sentiments, ou sur la détection des opinions politiques. Cette partie du développement pourra être laissée à la libre appréciation de la candidate ou du candidat.

4 Profil du/de la candidat·e

Les candidat·e·s doivent répondre aux critères suivants :

- Être titulaire d'un Master en littérature, humanités numériques ou sciences du langage avec une forte ~~une~~ sensibilité aux outils numériques appliqués aux corpus. Avoir une forte appétence pour les sciences humaines numériques et la littérature.
- Une expérience en Python pour le traitement de textes est souhaitée ; un accompagnement pourra être assuré pour monter en compétences selon les besoins du projet.
- Notions en TALN : similarité textuelle et alignement de textes (détection de réemplois).
- Une familiarité avec le traitement de textes (scripts, expressions régulières...) et avec l'écosystème XML-TEI/XSLT est appréciée ; ces compétences pourront être consolidées au cours de la thèse. Capacité à travailler en équipe pluridisciplinaire
- Avoir une excellente maîtrise du français écrit et oral (indispensable)

5 Modalités de candidature

Les personnes intéressées sont invitées à envoyer un dossier complet par e-mail à l'encadrant :

- Andrea Del Lungo – adellungo@free.fr

Le dossier doit comprendre :

1. Une **lettre de motivation (1 page max.)** détaillant l'intérêt pour le sujet et les compétences pertinentes pour la thématique proposée.
2. Un **CV complet** (formations, expériences, publications).
3. Un **relevé de notes** (Master).
4. Deux **lettres de recommandation** (*optionnel*).