

Master 2 Industrie de la Langue
2025 - 2026

Cahier des Charges

Interface de collation automatique pour les textes en moyen français du projet CreNum

UE : Projet Professionnel encadré par Monsieur Thomas Lebarbé
Yasaman AFSARI VELAYATI
Kemal Çelik

TABLE DE MATIERES

1. Présentation des parties	3
a. Maître d'œuvre	
b. Maître d'ouvrage	
2. Présentation du projet	4
a. Définition du projet	
b. État de l'art	
c. Étude du besoin	
d. Objectifs	
e. Public Visé	
f. Langue	
3. Solutions proposées	6
a. Utilisation de CollateX	
b. Développement d'une interface web	
c. Gestion et conservation	
4. Spécificités fonctionnelles	7
a. Import et structuration des données	
b. Collation et alignement automatique	
c. Visualisation parallèle des témoins	
d. Annotation et qualification des variantes	
e. Correction d'erreurs HTR et vérification	
f. Gestion des équivalences	
g. Sauvegarde	
5. Spécificités techniques	8
a. Technologies et environnement de développement	
b. Déploiement et accessibilité	
6. Planning	9
a. Calendrier académique	
b. Macro-phases	

1. Présentation des parties

a. Maître d'œuvre

- a) Maître d'œuvre
 - i. Coordonnées

Yasaman AFSARI VELAYATI

Yasaman.afsari-velayati@etu.univ-grenoble-alpes.fr

Kemal Çelik

Kemal.celik@etu.univ-grenoble-alpes.fr

- ii. Description

Notre équipe se forment de deux étudiants en Master 2 professionnel dans le domaine des Industries de la langue. Nous effectuons ce travail dans le contexte d'un projet professionnel, supervisé par Monsieur Thomas Lebarbé.

b. Maître d'ouvrage

- i. Coordonnées

Arnaud BEY

arnaud.bey@univ-grenoble-alpes.fr

Théo ROULET

theo.roulet@univ-grenoble-alpes.fr

2. Présentation du projet

a. Définition du projet

Ce projet de conception d'une interface de collation automatique de textes s'inscrit dans le cadre du projet CreNum : Édition critique numérique de la Chronique française de Guillaume Cretin. Porté par Ellen Delvallée (Chargée de recherche au sein du laboratoire Litt&Arts CNRS/UGA), il s'agit d'un projet de recherche en littérature et en histoire centré sur l'édition des cinq livres de la chronique en vers que le poète Guillaume Cretin écrivit au début du XVI^e siècle (1515-1525). L'un de ses objectifs est de recenser les variations que présente le texte à travers ses différents témoins manuscrits.

L'outil que nous avons pour mission de concevoir est une interface destinée à assister les chercheurs dans la comparaison de plusieurs versions manuscrites d'un même texte. Il s'appuie sur des transcriptions automatiques(HTR) des manuscrits réalisée par les ingénieurs travaillant au projet, et vise à faciliter l'identification, l'analyse et la qualification des variantes textuelles.

Contrairement à une collation entièrement manuelle, longue et limitée à quelques témoins à la fois, l'outil à réaliser cherche à mettre en place une approche semi-automatique permettant d'aligner simultanément plusieurs versions d'un texte, vers par vers et mot par mot. L'objectif n'est pas de remplacer l'expertise humaine, mais de faciliter sa mise en œuvre en proposant une visualisation synoptique, en réduisant le bruit et en permettant la qualification et le marquage rapide afin que les chercheurs puissent se concentrer sur les différences réellement significatives du point de vue philologique.

Cet outil repose sur l'utilisation et l'extension d'outils existants de collation (CollateX), combinés à une interface web interactive. Cette interface permet :

- la visualisation parallèle de plusieurs témoins,
- le repérage automatique des différences,
- la qualification manuelle des variantes (pertinente / non pertinente / à vérifier),
- et, le cas échéant, la correction d'erreurs de transcription.

Enfin, le projet vise à conserver la trace des décisions prises par les chercheurs, telles que les annotations, les corrections ou les équivalences orthographiques afin de capitaliser ce travail pour de futures collations et d'autres parties de l'œuvre. Il s'inscrit ainsi dans une démarche d'humanités numériques, à l'interface entre informatique, linguistique et sciences historiques.

b. État de l'art

Actuellement, il existe plusieurs outils et méthodes pour la collation de textes, utilisés en domaine des humanités numériques. Ces outils ont pour objectif de comparer différentes versions d'un même texte afin d'identifier les variantes.

Parmi les outils les plus connus, on trouve des logiciels de collation automatique comme CollateX, Juxta Classical Text Editor. Ces outils permettent d'aligner plusieurs témoins textuels et

de repérer automatiquement les différences entre eux. CollateX, en particulier, est un outil libre et multi-plateforme qui s'utilise dans la recherche et qui propose des algorithmes d'alignement mot à mot ou caractère par caractère.

Cependant, ces outils présentent plusieurs limites. D'une part, ils sont souvent conçus pour des textes modernes ou normalisés, et s'adaptent difficilement à des textes anciens, comme ceux du XVI^e siècle, où l'orthographe n'est pas fixée. Les variations graphiques peuvent souvent générer un grand nombre de différences non pertinentes, ce qui complique l'analyse. D'autre part, ces outils produisent généralement des résultats bruts, sous forme de fichiers texte, JSON ou XML, qui ne sont pas toujours facilement exploitables pour des chercheurs non spécialistes en informatique.

Il existe ainsi des plateformes proposant des interfaces de visualisation des variantes, mais elles sont souvent soit complexes à prendre en main, soit limitées en termes d'interaction. En particulier, peu d'outils permettent aux chercheurs de qualifier manuellement les différences (pertinente, non pertinente, erreur de transcription) ou de corriger directement les erreurs issues de HTR.

Enfin, la plupart des solutions ne proposent pas de mécanisme simple pour conserver et réutiliser les décisions prises par les chercheurs lors d'une collation précédente. Chaque nouvelle comparaison nécessite donc souvent de recommencer une grande partie du travail.

C'est dans ce contexte que s'inscrit ce projet, qui vise à s'appuyer sur des outils existants tout en proposant une interface plus adaptée aux besoins spécifiques des chercheurs travaillant sur des manuscrits anciens, en particulier en tenant compte des erreurs d'HTR et des variations orthographiques propres aux textes du XVI^e siècle.

c. Étude du besoin

Par ailleurs, les transcriptions obtenues par HTR sont toujours susceptibles de contenir des erreurs. Le travail sur des textes en moyen français complique leur détection automatique, en raison des variations orthographiques possibles. Les outils existants ne permettent pas facilement de distinguer ces erreurs des variantes réellement significatives.

Le besoin exprimé par les chercheurs est donc de pouvoir filtrer les différences et ne conserver que des variantes réellement significatives. Celles-ci peuvent être de plusieurs natures : absence d'un vers, d'un groupe de mots ou d'un mot, changement d'orthographe ayant un effet sur la versification (syllabe), ou changement de la nature ou de la fonction d'un mot modifiant le sens du texte. Face à cela, ils préfèrent le recours à leur expertise plutôt qu'à des solutions entièrement automatisées.

Cependant, l'utilisation d'un dictionnaire du français moyen pour identifier certaines erreurs d'HTR a été évoquée.

d. Objectifs

L’objectif de ce travail est de concevoir un outil de collation semi-automatique permettant aux chercheurs de comparer efficacement plusieurs versions manuscrites d’un même texte du XVI^e siècle à partir de transcriptions HTR. À partir des résultats de collation fournis par CollateX, le projet vise à proposer une interface web simple et fonctionnelle qui permet de visualiser les témoins en parallèle, d’identifier les différences, et surtout d’aider les chercheurs à distinguer les variantes textuelles réellement pertinentes des erreurs de transcription ou des variations orthographiques non significatives. L’outil doit également permettre de distinguer des variantes à conserver ou à exclure grâce à l’annotation manuelle des chercheurs et dans certains cas, de corriger ces différences, tout en conservant les décisions prises afin de réutiliser ce travail lors de futures collations ou pour d’autres parties de l’œuvre.

e. Public Visé

L’outil est développé principalement pour deux chercheurs en littérature et en histoire, impliqués directement dans le projet et dans l’analyse des manuscrits. Il n’est pas destiné à un très large public ni à des utilisateurs sans accompagnement technique. L’outil reste toutefois réutilisable par d’autres utilisateurs disposant de compétences techniques suffisantes, capables de préparer leurs données en amont ou de modifier eux-mêmes le paramétrage de CollateX dans le code, ce qui est accessible à un ingénieur ou à un utilisateur avancé.

f. Langue

La langue utilisée dans le cadre de ce projet est principalement le français, à la fois pour l’interface de l’outil et pour la rédaction de la documentation. Les textes étudiés sont également en moyen français (XVI^e siècle), ce qui implique une prise en compte des variations orthographiques et linguistiques propres à cette période.

3. Solutions proposées

a. Utilisation de CollateX

Dans ce projet, la collation des textes repose sur l’utilisation de CollateX qui permet de comparer plusieurs versions d’un même texte en les alignant automatiquement, mot par mot. À partir des textes fournis en entrée, l’outil détecte les différences entre les témoins et génère un alignement mettant en évidence les variantes.

L’intérêt principal de CollateX est qu’il automatise une grande partie du travail de comparaison, tout en laissant au chercheur la responsabilité de l’interprétation des résultats. Il permet de comparer simultanément un nombre quelconque de témoins, avec un alignement n à n. Dans le cadre de ce projet, CollateX est utilisé comme base pour identifier les divergences entre les transcriptions HTR, avant leur visualisation et leur analyse dans une interface web dédiée.

b. Développement d'une interface web

Le projet inclut la conception d'une interface web dédiée à la consultation des résultats de la collation. Les différentes versions d'un même texte y sont affichées côte à côté, sous forme de colonnes, chaque ligne devant correspondre à un vers. CollateX comparant les témoins mot à mot, il faut donc ensuite reconstituer les vers pour l'affichage. Le but est de proposer une interface de lecture en parallèle qui s'adapte au besoin et aux habitudes de lecture du chercheur, sans perdre le contexte du vers, et même d'un ensemble de vers. Les différences identifiées par l'outil de collation apparaissent de manière visible, afin d'orienter rapidement la lecture du chercheur.

Cette interface offre également la possibilité d'interagir avec les variantes. Chaque différence peut être examinée, qualifiée selon son importance et, si nécessaire, corrigée lorsqu'il s'agit d'une erreur de transcription. L'ensemble a été pensé pour rester simple et lisible, tout en répondant aux besoins spécifiques des chercheurs travaillant sur des textes anciens.

c. Gestion et conservation

Les décisions prises par les chercheurs, comme les annotations, les corrections d'erreurs HTR et les équivalences orthographiques, sont conservées de manière persistante afin quaucun travail ne soit perdu.

4. Spécificités fonctionnelles

a. Import et structuration des données

L'outil prend en entrée plusieurs versions d'un même texte (issues de l'HTR), fournies sous forme de fichiers (par exemple TXT ou JSON). Les données sont ensuite structurées pour la comparaison : découpage par chapitres et par vers (ou lignes), suppression éventuelle de la ponctuation si nécessaire, et préparation d'un format cohérent entre témoins afin de faciliter l'alignement.

b. Collation et alignement automatique

Une fois les textes préparés, l'outil lance une collation via CollateX pour aligner automatiquement les témoins et chaque ligne correspond à un vers. L'alignement se fait mot par mot, et la sortie contient les zones identiques et les zones divergentes. Cette étape constitue la base de la comparaison, avant l'analyse humaine.

c. Visualisation parallèle des témoins

Les résultats sont affichés dans une interface web sous forme de colonnes, chaque colonne correspondant à un témoin. L'utilisateur lit ainsi un même vers en parallèle et repère plus facilement les différences, qui sont mises en évidence (par exemple par couleur) à la fois à l'échelle du vers, mais également à l'échelle des mots contenus par ce vers. La navigation doit rester simple pour parcourir rapidement un grand volume de vers.

d. Annotation et qualification des variantes

À l'échelle d'un vers ou d'un groupe de mots, le chercheur peut indiquer si une différence est pertinente, non pertinente ou à vérifier. L'idée est de partir du principe que la majorité des différences ne sont pas significatives, et de ne conserver que celles qui présentent un intérêt du point de vue philologique. Ces annotations doivent être associées à une position précise dans le texte (numéro ou identifiant du chapitre et du vers, et numéro du mot).

e. Correction d'erreurs HTR et vérification

Certaines différences viennent d'erreurs de transcription automatique plutôt que de véritables variantes. L'outil doit donc permettre, au moins dans les cas nécessaires, de corriger une forme fautive (par exemple une lettre ou un mot mal reconnu). Lorsque c'est possible, une vérification par retour à l'image numérisée du manuscrit peut être envisagée (via un lien vers la page).

f. Gestion des équivalences

Pour réduire le bruit, le système doit gérer des équivalences orthographiques fréquentes (ex. Y/I, S/Z, doubles consonnes, mots collés/séparés, variantes régulières comme différentes graphies d'un même verbe). Le chercheur peut déclarer qu'une différence doit être "ignorée toujours", comme dans un correcteur orthographique, afin qu'elle ne soit plus signalée comme variante lors des collations lancées sur le chapitre suivant.

g. Sauvegarde

Toutes les décisions prises (annotations, corrections, équivalences) doivent être sauvegardées de manière persistante afin de capitaliser le travail et de le réutiliser sur d'autres chapitres ou d'autres livres.

5. Spécificités techniques

a. Technologies et environnement de développement

Le projet s'appuie principalement sur Python et la collation est réalisée avec la bibliothèque CollateX (version Python), et l'application peut être construite avec un framework web léger comme Flask, afin d'éviter une solution trop lourde. Une partie front-end simple (HTML/CSS/JavaScript) est considérée pour l'affichage et l'interaction.

b. Déploiement et accessibilité

La solution doit rester simple à utiliser pour des utilisateurs. Elle se définit en une application web accessible via un navigateur est privilégiée. L'outil doit être facile à lancer et privilégier la fonctionnalité plutôt que le graphisme, afin de permettre des itérations rapides avec les utilisateurs.

6. Planning

a. Calendrier académique

Tâches	Date
Présentation des projets	05/11/2025
Choix des projets par les étudiants et notification aux commanditaires	07/11/2025
Soumission des cahiers des charges	19/12/2025
Acceptation des cahiers des charges par les commanditaires	26/12/2025
Rendu des projets, documentations et manuels	20/02/2026
Soutenance publique	27/02/2026

b. Macro-planning

Tâches	Date de début	Date de fin	Responsables
Préparation & set up	26/12/2025	29/12/2025	Les deux
Réception des transcriptions HTR et préparation des fichiers	29/12/2025	03/01/2026	Kemal
Mise en place de la collation automatique avec CollateX	03/01/2026	08/01/2026	Les deux
Développement du premier affichage web en colonnes	08/01/2026	15/01/2026	Yasmine
Navigation dans le corpus par chapitre et par vers	15/01/2026	19/01/2026	Yasmine
Ajout des fonctions de qualification des variantes	19/01/2026	26/01/2026	Yasmine
Ajout de la fonction ignorer toujours et gestion des équivalences	26/01/2026	01/02/2026	Kemal
Ajout des fonctions de correction des erreurs HTR	01/02/2026	06/02/2026	Kemal
Mise en place de la sauvegarde des décisions des utilisateurs	06/02/2026	10/02/2026	Kemal
Mise en place de l'export des résultats et des décisions	10/02/2026	14/02/2026	Yasmine
Tests avec les chercheurs et ajustements	14/02/2026	18/02/2026	Les deux
Rédaction de la documentation et du manuel utilisateur	18/02/2026	20/02/2026	Yasmine
Préparation de la soutenance et démonstration	20/02/2026	27/02/2026	Les deux