Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Normaliser le moyen français : du graphématique au semi-diplomatique

Sonia Solfrini[1] Mylène Dejouy[1] Aurélia Marques Oliveira[1] Pierre-Olivier Beaulnes[1]
(1) Université de Genève, Boulevard des Philosophes 22, 1205 Genève, Suisse
sonia.solfrini@unige.ch

RÉSUMÉ

La pré-éditorialisation des documents anciens, comprise comme une automatisation partielle de la préparation éditoriale des données textuelles, est récemment devenue l'un des nouveaux fronts de la recherche en philologie computationnelle. Parmi les tâches qu'elle comprend, on trouve celle de la normalisation linguistique, qui rend le texte plus aisément lisible tant par les chercheurs que par les machines. Dans un premier temps, nous définissons cette tâche de TAL (Traitement Automatique du Langage) pour le moyen français et sa place dans une chaîne de traitement numérique qui permet la création de données *machine-actionable*, depuis les sorties de l'OCR (*Optical Character Recognition*). Ensuite, nous présentons et rendons disponible un ensemble de données d'environ 40 000 lignes, tirées d'un corpus d'imprimés du XVI[e] siècle, ainsi que les règles de normalisation semi-diplomatique qui ont guidé la préparation des données. Enfin, nous proposons un premier modèle de normalisation automatique, afin de confirmer la faisabilité de la tâche.

ABSTRACT

Normalising Middle French : From Graphematic to Semi-Diplomatic

The pre-editorialisation of ancient documents, understood as a partial automation of the editorial preparation of textual data, has recently emerged as a new frontier in computational philology research. Among the tasks involved in this process, linguistic normalisation enhances text readability for both researchers and machines. In this paper, first, we define this NLP (Natural Language Processing) task for Middle French and its role within a digital processing pipeline that transforms OCR (Optical Character Recognition) outputs into machine-actionable textual data. Next, we present a dataset of approximately 40 000 lines, extracted from 16th-century printed texts, along with our semi-diplomatic normalisation guidelines. Finally, we propose a first automatic normalisation model, to assess the feasibility of this task.

MOTS-CLÉS : Humanités Numériques, Normalisation automatique, Français du XVI[e] siècle, Moyen français, Règles de normalisation, Pré-éditorialisation des textes.

KEYWORDS: Digital Humanities, Automatic normalisation, 16th-Century French, Middle French, Normalisation guidelines, Pre-editorialisation of texts.

ARTICLE : **Accepté à** CORIA-TALN-RJC 2025.

Actes des 18e Rencontres Jeunes Chercheurs en RI (RJCRI) et 27ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL), pages 239–252.

1 Introduction

Contrairement à la traduction, qui fait passer un texte d'une langue à une autre, la normalisation linguistique vise à faire passer un texte d'un état de langue à un autre (par ex., d'une version graphématique du moyen français à une version semi-diplomatique), en s'alignant sur des normes préalablement définies. L'état de la langue cible correspond généralement à une convention philologique, comme les normes d'édition des textes anciens : ajout de l'apostrophe (*dune robbe* → *d'une robbe*), résolution des abréviations (*7* → *et*), etc. Notre objectif est double : rendre les textes plus aisément lisibles par les chercheurs et mieux exploitables par les machines (par ex., lemmatisation, reconnaissance d'entités nommées, etc.). S'agissant d'un processus particulièrement chronophage, la normalisation fait désormais l'objet de recherches visant à développer des outils qui permettent son automatisation. D'un point de vue technique, c'est une tâche qui s'inspire de la traduction automatique (TA).

À ce jour, peu de recherches ont été menées sur les outils permettant une normalisation automatique du moyen français [1] (voir § 4.1). Notre approche se distingue des précédentes par une normalisation semi-diplomatique du texte, intégrée à une chaîne de traitement numérique allant de l'acquisition du texte (analyse de mise en page et reconnaissance du texte) à l'annotation linguistique, en passant par l'encodage en XML-TEI (voir figure 1 et § 2).

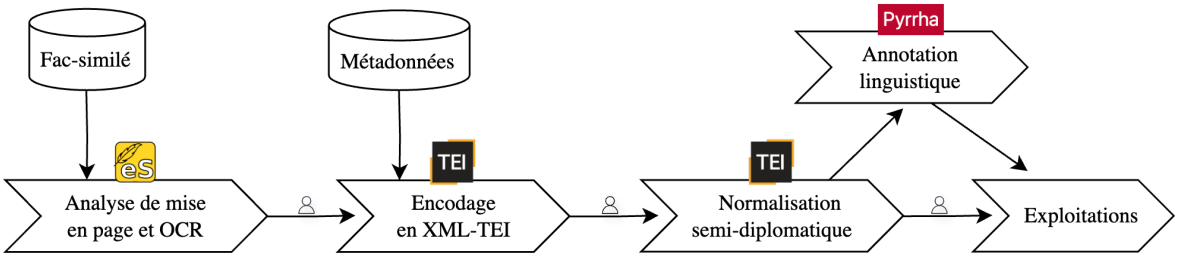

FIGURE 1 – Chaîne de traitement numérique d'un texte de notre corpus. L'icône représentant un personnage indique une intervention humaine ou une vérification manuelle. « eS » fait référence à *eScriptorium*, « OCR » à l'*Optical Character Recognition*, et « TEI » à la *Text Encoding Initiative*. La normalisation semi-diplomatique constitue un chaînon essentiel : elle permet non seulement divers modes d'exploitation du texte, mais offre également une base pour d'autres types de normalisation.

Notre corpus comprend des textes en français publiés dans les années 1530. La langue source correspond à une transcription graphématique (Stutzmann, 2011), produite par un modèle d'OCR et vérifiée manuellement. Ce type de transcription limite la variation graphique, c'est-à-dire toutes les variantes des caractères, en réduisant les différents formes possibles à une seule valeur dans l'alphabet latin

1. L'histoire de la langue française a longtemps été divisée en trois grandes périodes : l'ancien français, le moyen français et le français moderne. Toutefois, la définition des limites chronologiques du moyen français fait l'objet de débats, notamment quant à la question de sa délimitation aux XIV^e^ et XV^e^ siècles uniquement, ou à son éventuelle prolongation jusqu'au XVI^e^ siècle. Plus récemment, les questionnements sur la périodisation du français ont aussi été bousculés par la catégorie de français pré-classique, dont la délimitation temporelle précise, comprise entre les XVI^e^ et XVII^e^ siècles, reste sujette à discussion, un *terminus ad quem* étant fréquemment fixé aux alentours de 1630 (Combettes & Marchello-Nizia, 2010; Ayres-Bennett & Caron, 2016). Pour éviter ces problèmes de périodisation, des chercheurs ont opté pour une périodisation du français par siècle, notamment pour le XVI^e^ siècle (Gougenheim, 1973; Catach, 1968; Huchon, 1988; Fragonard & Kotler, 1994; Vachon, 2010). Étant donné que notre corpus comprend des textes publiés dans les années 1530, nous utiliserons indifféremment les termes de « français du XVI^e^ siècle » et de « moyen français », dont la borne finale est généralement située au milieu du XVI^e^ siècle (Combettes & Carlier, 2022).

contemporain (par ex., *s* long / *s* rond → ‹s›).

La langue cible correspond à une transcription de type semi-diplomatique. La notion de « semi-diplomatique » est assez complexe à définir, parce que les règles appliquées dépendent de plusieurs éléments : la période du texte lui-même (ancien français, moyen français, etc.), les traditions philologiques (italienne, française, etc., voir Duval 2006), ainsi que certains choix ecdotiques (par ex., le lecteur visé). Nos règles de normalisation (voir § 3) sont le fruit de réflexions menées à partir des normes d'édition francophones en usage chez les seiziémistes (Barbiche & Chatenet, 1993; Giraud, 1997) et les médiévistes (Meyer, 1910; Roques, 1926; Guyotjeannin & Vielliard, 2014), ainsi que des principes de la base *Epistemon* [2] (Demonet & l'équipe BVH, 2011) et de la notion de pré-éditorialisation (Pinche *et al.*, 2022; Vogeler, 2023; Gabay *et al.*, 2024).

Les règles de normalisation semi-diplomatique que nous proposons sont moins interventionnistes que celles traditionnellement en usage dans les éditions critiques. Ces dernières peuvent néanmoins être appliquées dans un second temps, à partir de notre normalisation. Cette approche minimaliste facilite également l'élaboration de règles communes en diachronie longue, ouvrant la voie à la création d'outils partagés à travers les siècles et les différents états du français. La question des transcriptions plus interventionnistes des éditions « interprétatives », ou même des « translations » (Garnier, 2019), pour lesquelles le lexique et la syntaxe peuvent être retouchées, sont écartées de cette étude.

Transcription graphématique	**Normalisation semi-diplomatique**
¶ Les noms *7* accoustremẽs des personnaiges de ceste pre¬ sente moralite. Foy uestue dune belle robbe blanche/ Esperance/ uestue dune robbe de uiolet/ Charite/ uestue descarlate/	¶ Les noms et accoustremens des personnaiges de ceste pre¬ sente moralite. Foy vestue d'une belle robbe blanche/ Esperance/ vestue d'une robbe de violet/ Charite/ vestue d'escarlate/

TABLE 1 – La transcription graphématique d'un extrait de notre corpus et sa normalisation semi-diplomatique réalisée selon nos règles. Il s'agit de l'extrait qui apparaît en figure 2.

Le passage d'une transcription graphématique à une normalisation semi-diplomatique, comme présenté dans le tableau 1, s'avère fondamental dans divers contextes, tels que la constitution de corpus annotés, l'exploration des données, ou encore l'utilisation d'outils de TAL dans les meilleures conditions. Cette étape contribue également à la mise en place de chaînes de traitement numérique, allant de l'acquisition du texte à son annotation linguistique. La normalisation semi-diplomatique s'affirme ainsi comme un chaînon essentiel, en ce qu'elle ouvre la voie non seulement à divers modes d'exploitation du texte (voir figure 1), mais aussi à d'éventuels autres types de normalisation linguistique, plus interventionnistes, comme celles qui alignent intégralement le système graphique du texte avec l'orthographe du français contemporain.

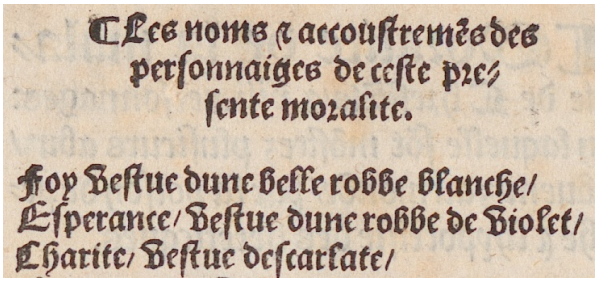

FIGURE 2 – Un extrait de notre corpus (Malingre, 1533). Il s'agit de l'extrait transcrit dans le table 1.

Dans cet article, nous présentons un ensemble de données d'environ 40 000 lignes (voir § 2), ainsi

2. La base *Epistemon* : https://www.bvh.univ-tours.fr/Epistemon/index.asp.

que les règles de normalisation que nous avons suivies (voir § 3). Nous proposons ensuite un premier modèle permettant d'automatiser ce processus pour le moyen français (voir § 4). Enfin, nous concluons en évoquant les perspectives et les prochaines étapes de ce travail (voir § 5).

2 Description du corpus

Identifiant	Titre bref	Auteur	Date
CRRPV01	*Premiere partie de l'union de plusieurs passaiges de l'escripture saincte*	Herman Bodius	1533
CRRPV03	*Instruction des enfans*	[Pierre Robert Olivétan]	1533
CRRPV04	*Grans pardons et indulgences, le tresgrand Jubile de plainiere remission*	[Anonyme]	[1533]
CRRPV08	*Livre des marchans*	[Antoine Marcourt]	1533
CRRPV09	*Maniere et fasson qu'on tient en baillant le sainct baptesme*	[Guillaume Farel]	1533
CRRPV11	*Moralité de la maladie de Chrestienté, a .xiij. personnages*	[Matthieu Malingre]	1533
CRRPV12	*Plusieurs belles et bonnes chansons*	[Matthieu Malingre]	1533
CRRPV13	*Noelz nouveaulx*	[Matthieu Malingre]	[1533]
CRRPV16	*Chansons nouvelles demonstrantz plusieurs erreurs et faulsetez*	[Matthieu Malingre]	[1534]
CRRPV19	*Faictz de Jesus Christ et du Pape*	[Anonyme]	[1534]
CRRPV20	*Articles veritables sur les abuz de la Messe papale*	[Antoine Marcourt]	[1534]
CRRPV21	*Petit traicté tres utile et salutaire de la saincte eucharistie*	[Antoine Marcourt]	1534
CRRPV22	*Declaration de la messe*	[Antoine Marcourt]	[1534]
CRRPV23	*Summaire, et briefve declaration d'aucuns lieux fort necessaires*	[Guillaume Farel]	1534
CRRPV25	*Letres certaines d'aucuns grandz troubles advenuz à Geneve l'an 1534*	[Guillaume Farel]	[1534]

TABLE 2 – Liste des textes normalisés par les auteurs de cet article et utilisés pour les expériences, avec une sélection de métadonnées.

Le corpus primaire de notre projet [3] comprend les publications des imprimeurs Pierre de Vingle et Jean Michel, lorsqu'ils étaient actifs à Neuchâtel et/ou Genève, entre 1532 et 1535 pour le premier, et entre 1538 et 1545 pour le second (Berthoud, 1980; Kemp, 2004; Solfrini *et al.*, 2023b). Ce corpus compte une cinquantaine d'ouvrages liés à la Réforme protestante, couvrant une grande diversité de genres : moralités, chansons spirituelles, placards, traités polémiques. Au niveau de l'histoire de la langue, cette période du XVI^{e} siècle est marquée par l'instauration progressive des signes auxiliaires (accents, tréma, cédille, etc.). Un manuel imprimé par Pierre de Vingle illustre bien cette évolution du français et ses défis : *L'Instruction des enfants, contenant la manière de prononcer et d'écrire en françois* (1533) de Pierre-Robert Olivétan. Dans son « Épître au lecteur », l'auteur met en avant l'introduction de l'apostrophe ainsi que des accents aigu et grave (Carbonnier-Burkard, 2014). Ces innovations ne sont cependant pas employées dans les autres publications du même imprimeur. Par ailleurs, ces ouvrages se distinguent également par l'usage fréquent d'abréviations et par l'emploi d'une typographie gothique (voir figure 2).

Océrisation. Chaque texte de notre corpus est transcrit à l'aide d'un modèle d'OCR – d'abord *Cortado* 2.0.0 (Pinche & Clérice, 2022), puis une version « fine-tunée » (Solfrini & Gabay, 2024) de *CATMuS* 1.0.0 (Pinche *et al.*, 2023) –, avant d'être vérifié manuellement (Solfrini *et al.*, 2024). Les données sont préparées avec *eScriptorium* (Kiessling *et al.*, 2019) et publiés avec les outils d'intégration continue du projet HTR-United (Chagué *et al.*, 2021; Chagué & Clérice, 2023). L'analyse de mise en page (ou zonage) des documents respecte les normes du vocabulaire *SegmOnto* (Gabay *et al.*, 2024), tandis que la transcription respecte un guide de transcription graphématique pour les imprimés en caractères gothiques (Solfrini *et al.*, 2023a). À ce jour, près d'une trentaine d'ouvrages ont été océrisés.

3. Le projet SETAF : https://www.unige.ch/setaf.

Encodage et normalisation. Chaque document est ensuite encodé en XML-TEI, de façon semi-automatique, selon un schéma personnalisé qui est appliqué à l'ensemble du corpus. Cet encodage s'inspire du schéma du projet Gallic(orpor)a [4] (Pinche *et al.*, 2022). Il permet de préserver et de manipuler les différentes strates d'un même texte, notamment sa transcription graphématique et sa version normalisée, dérivée de cette transcription initiale. À ce jour, quinze textes ont été normalisés manuellement conformément à nos règles de normalisation semi-diplomatique (voir table 2 et § 3).

Annotation linguistique. Une annotation linguistique est produite à partir des lignes normalisées. Elle comprend la lemmatisation des tokens, ainsi que l'étiquetage syntaxique et morphologique des lemmes. Le référentiel que nous avons choisi pour lemmatiser les textes est *LGeRM* [5] (Souvay & Pierrel, 2009), développé dans le cadre du *DMF* [6]. Concernant l'étiquetage syntaxique et morphologique, nous suivons le jeu d'étiquettes *Cattex-max* (Prévost *et al.*, 2013), dont les principes sont indiqués dans les *Principes d'annotation Cattex09* (Guillot *et al.*, 2013) et développés dans le *Manuel d'annotation linguistique pour le français moderne (XVIe - XVIIIe siècles)* (Gabay *et al.*, 2022a). Pour réaliser ces annotations, nous utilisons l'application *Pyrrha* (Clérice *et al.*, 2024), qui permet d'utiliser le modèle *FreEM* (*French Early Modern*) (Gabay *et al.*, 2022b), basé sur *LGeRM* et *Cattex-max*. Cet outil permet également la correction manuelle des annotations, ce qui garantit une meilleure qualité des données. À ce jour, environ 10 000 tokens issus de nos documents ont été vérifiés manuellement.

3 Règles de normalisation

3.1 Nos règles

Dans une perspective de pré-éditorialisation du texte, notre objectif principal est de préparer un corpus textuel en vue de diverses tâches courantes en philologie numérique. Par exemple, une phase d'enrichissement du texte peut inclure des tâches d'annotation linguistique, telles que la lemmatisation et l'étiquetage morpho-syntaxique. C'est dans ce cadre que nous avons établi nos règles de normalisation semi-diplomatique, afin de garantir un passage optimal d'une tâche à l'autre de notre chaîne de traitement numérique (voir § 1 et figure 1), et notamment de simplifier la tâche de tokenisation qui prépare celle de lemmatisation.

Tout d'abord, nos règles ne s'alignent pas sur l'orthographe du français contemporain : par ex., *personnaige* ne devient pas *personnage* ; *vestue* ne devient pas *vêtue* ; et *robbe* ne devient pas *robe*. Concernant l'alphabet, la distinction entre ‹u›/‹v› et ‹i›/‹j› (lettres dites « ramistes ») est introduite : par ex., *vniuers* devient *univers*. Nos règles impliquent également de corriger les fautes d'impression manifestes (par ex., des fautes de composition du prote, des caractères illisibles), qui sont distinguées des particularités du moyen français.

Tous les types d'abréviations sont développés, à l'exception des références bibliques. Les signes auxiliaires (accents, trémas, cédilles, traits d'union) ne sont ni ajoutés ni corrigés. La segmentation de la chaîne graphique est normalisée par l'ajout de l'apostrophe (par ex., *larbre* → *l'arbre*) et par la segmentation des séquences agglutinées qui seraient disjointes en français contemporain (par ex., *tresriche* → *tres riche*). Enfin, vu qu'ils ont un impact relativement mineur sur les tâches ultérieures de notre chaîne de traitement et que leur normalisation serait trop chronophage, la ponctuation et la

4. Le projet Gallic(orpor)a : https://gallicorpora.github.io.
5. *LGeRM* (Lemmes Graphies et Règles Morphologiques) : http://stella.atilf.fr/LGeRM/.
6. *DMF* (Dictionnaire du Moyen Français) : http://zeus.atilf.fr/dmf/.

grammaire originelles sont préservées. La syntaxe et le lexique originels, qui ne sont jamais retouchées dans le cadre d'une normalisation semi-diplomatique, sont également préservés.

Concernant le signe ‹-›, celui-ci est ambigu car il peut signaler un trait d'union (par ex., *lui-même*, *Pierre-Olivier*), aussi bien qu'un trait d'hypénisation indiquant la coupure des mots en fin de ligne (par ex., *man-|ger*). Ces deux usages sont désambiguïsés en utilisant deux signes différents, ‹-› (*Hyphen-Minus*, U+002D) et ‹¬› (*Not Sign*, U+00AC, voir Gabay *et al.* 2023). Cela permet de savoir quand il est nécessaire de reconstituer automatiquement un mot tronqué (par ex., *man¬|ger* → *manger*) et de simplifier ainsi les traitements ultérieurs, notamment la lemmatisation.

3.2 Compatibilité avec les règles d'autres projets

Catégorie	**Éléments à normaliser selon nos règles**	**Éléments à normaliser selon RCnum**	**Éléments à normaliser selon Epistemon**
Les lettres	‹u›/‹v› et ‹i›/‹j›	Les majuscules seulement ?	‹u›/‹v› et ‹i›/‹j›
Les chiffres	‹u›/‹v› dans les chiffres romains	?	?
Les fautes	Les coquilles et les caractères illisibles	Les coquilles	Les coquilles et les erreurs manifestes
Les abréviations	Résolution des abréviations (à l'exception des références bibliques)	?	Résolution des abréviations
Les signes auxiliaires	-	Les accents	-
La segmentation de la chaîne graphique	Les apostrophes et séparation des mots agglutinés	Les apostrophes seulement ?	Les apostrophes, séparation des mots agglutinés et agglutination des mots séparés
Les signes fonctionnels	Le signe ‹-› *vs* le signe ‹¬›	?	?
La ponctuation	-	Quelques éléments	-
La syntaxe	-	-	-
Le lexique	-	Les patronymes, les noms propres et les toponymes	-
La grammaire	-	Quelques éléments	-

TABLE 3 – Tableau de comparaison de règles de normalisation semi-diplomatique. Les règles de trois projets numériques sont comparées : notre projet (SETAF), le projet RCnum et la base *Epistemon* des BVH. Pour le projet RCnum et sa « normalisation sans modernisation », nous nous basons sur les deux articles publiés en 2024 par l'équipe du projet (Rubino *et al.*, 2024a,b). Pour la base *Epistemon* et sa « transcription patrimoniale », nous nous basons sur les principes éditoriaux décrits sur le site du projet (Demonet & l'équipe BVH, 2011), cependant les normes éditoriales ultérieures indiquées dans la notice de chaque texte, qui peuvent diverger des règles générales, ne sont pas prises en compte dans ce tableau. Nous avons laissé des points d'interrogation pour les choix de normalisation qui ne sont pas explicitement exprimés dans la documentation d'accompagnement.

Il existe deux autres projets ayant abordé explicitement le problème de la normalisation du français du XVI^{e} siècle : *Epistemon*, déjà cité dans l'introduction, et RCnum (Bouillon *et al.*, 2024). La base *Epistemon*, qui fait partie du programme des *Bibliothèques Virtuelles Humanistes* (Uetani *et al.*, 2016), est un projet de référence pour la diffusion de documents patrimoniaux français du XVI^{e} siècle. L'équipe propose des principes éditoriaux pour une « transcription patrimoniale » (Demonet & l'équipe BVH, 2011), qui est adaptée à un vaste corpus textuel et s'écarte de certaines normes d'édition en usage chez les seiziémistes (Barbiche & Chatenet, 1993; Giraud, 1997). Des normes éditoriales plus spécifiques sont ensuite indiquées dans la notice de chaque document appartenant à cette base. Concernant le projet RCnum, son équipe a publié des travaux sur la normalisation automatique de la langue employée dans les Registres du Conseil (RC) de Genève entre 1536 et 1550. Ces registres ont été transcrits manuellement, et ceux de 1536 à 1544 ont été publiés sous forme de livres imprimés (par ex., voir Chazalon *et al.* 2021). Leur approche, qu'ils qualifient de « normalisation sans modernisation », consiste à ajuster les systèmes graphiques tout en préservant le contenu syntaxique et sémantique des manuscrits originaux. Pour leurs règles, nous nous basons sur les deux articles publiés en 2024 par l'équipe du projet (Rubino *et al.*, 2024a,b). Par rapport à la base *Epistemon* et au projet RCnum, nos règles de normalisation semi-diplomatique présentent peu de différences avec les règles régissant la première, de laquelle nous nous sommes inspirés, mais voient des divergences plus marquées avec celles du second projet (voir table 3).

La principale différence avec ces deux projets est que nous intégrons la tâche de normalisation dans une chaîne de traitement qui va de l'OCR jusqu'à l'annotation linguistique, en passant par l'encodage en XML-TEI (voir figure 1 et § 2). D'autres différences entre nos règles et celles d'*Epistemon* concernent l'introduction du signe ‹¬› pour le tiret d'hyphénisation, choix dont nous avons expliqué plus haut les raisons, et la décision de ne pas réunir les mots dans les séquences disjointes qui seraient agglutinées en français contemporain (par ex., *long temps*). Réunir ces mots nous paraît trop chronophage et, dans la perspective de la tâche de lemmatisation, il est souvent possible d'annoter individuellement chaque élément des séquences telles que *long temps*, *en suite*, *lors que*, *puis que*, *aussi tôt*. En revanche, il nous a semblé essentiel de séparer les mots dans les séquences agglutinées qui seraient disjointes en français contemporain (par ex., *tresriche* → *tres riche*).

Les règles de normalisation du projet RCnum ne mentionnent pas la distinction entre les lettres ramistes ‹u›/‹v› et ‹i›/‹j›, ni la résolution des abréviations. Par contre, elles prévoient la normalisation des majuscules et celle des accents. Concernant la ponctuation et la grammaire, quelques éléments font l'objet d'une normalisation. Du point de vue lexical, ce sont les patronymes, les noms propres et les toponymes qui sont normalisés. Ce choix est probablement dû à la présence de nombreuses entités nommées dans les textes sources. En résumé, ces règles partagent avec les nôtres trois caractéristiques : l'ajout des apostrophes, la correction des coquilles et l'absence d'intervention sur la syntaxe.

4 Un modèle pour normaliser le moyen français

4.1 État de l'art

À notre connaissance, les premières recherches sur la normalisation automatique du français se sont d'abord concentrées sur le français du XVII^{e} siècle (Bollmann, 2019; Gabay *et al.*, 2019; Gabay & Barrault, 2020; Bawden *et al.*, 2022). Les auteurs de ces travaux ont présenté *FreEMnorm* (pour *French Early Modern normalisation*), un jeu de données de référence permettant d'évaluer différentes

méthodes de normalisation automatique vers le français contemporain. Plusieurs approches ont été comparées : l'alignement (ABA - *Alignment-Based Approach*), la traduction statistique (SMT - *Statistical Machine Translation*) et la traduction neuronale (NMT - *Neural Machine Translation*), en utilisant dans le dernier cas des architectures *Transformer* et LSTM (*Long Short-Term Memory*). Parmi celles-ci, l'approche par traduction automatique statistique (SMT), avec post-correction lexicale, s'est révélée la plus performante.

Concernant le moyen français, comme nous l'avons déjà mentionné plus haut, l'équipe du projet RCnum a publié des travaux sur la normalisation automatique de la langue employée dans les Registres du Conseil de Genève entre 1536 et 1550. Après avoir exploré plusieurs méthodes permettant d'automatiser ce processus, la méthode qui s'est révélée la plus performante est une approche qui combine le *fine-tuning* de grands modèles de langage (LLM - *Large Language Model*), basés sur une architecture *Transformer*, avec l'ajout de données synthétiques.

4.2 Jeu de données

Notre jeu de données comprend 15 textes, pour un total de 1 437 pages et de 37 935 lignes. Ces lignes ne correspondent pas à des phrases, mais aux lignes de texte telles qu'elles apparaissent dans les documents d'origine, qui sont pour la plupart au format in-8°. Le jeu de données a été réparti comme suit (voir table 4) :

— 31 905 lignes pour l'entraînement (84.1%) ;
— 3 545 lignes pour la validation (9.3%) ;
— 2 485 lignes pour le test (6.6%).

Les données des jeux d'entraînement et de validation ont été réparties de façon aléatoire, tandis que le jeu de test correspond à un texte de 96 pages qui a été choisi pour être représentatif de notre corpus : la *Moralite de la maladie de Chrestiente* ([Neuchâtel] : [Pierre de Vingle], 1533).

Entraînement + validation		
Identifiant	**Pages**	**Lignes**
CRRPV01	415	12 421
CRRPV03	128	3 473
CRRPV04	29	701
CRRPV08	46	893
CRRPV09	87	1 747
CRRPV12	48	1 206
CRRPV13	48	1 184
CRRPV16	16	405
CRRPV19	48	1 934
CRRPV20	1	70
CRRPV21	78	1 864
CRRPV22	96	2 372
CRRPV23	208	4 907
CRRPV25	93	2 273
Total textes	**Total pages**	**Total lignes**
14	1 341	35 450
Test		
Identifiant	**Pages**	**Lignes**
CRRPV11	96	2 485

TABLE 4 – Répartition des textes dans notre jeu de données.

4.3 Entraînement

Nous avons essayé d'entraîner des modèles NMT avec une architecture LSTM. Les expériences ont été menées en utilisant *Fairseq* (Ott *et al.*, 2019), avec un vocabulaire joint, contenant la source et la cible, et non deux vocabulaires distincts, car elles sont très proches linguistiquement. L'optimisation a été effectuée avec Adam (Kingma & Ba, 2015) et un taux d'apprentissage initial de 0.001. Les données textuelles ont d'abord été segmentées en sous-mots (*Byte Pair Encoding*, voir Sennrich *et al.* 2016) à l'aide de *SentencePiece* (Kudo & Richardson, 2018). Cinq jeux de vocabulaires de tailles différentes (500, 1 000, 2 000, 3 000 et 4 000 unités) ont ensuite été testés. Ces jeux ont été prétraités et binarisés, avant que des modèles de type LSTM ne soient entraînés, en suivant les architectures proposées dans l'étude sur la normalisation du français du XVII^{e} siècle mentionnée plus haut (Bawden

et al., 2022). Les configurations explorées sont présentées dans le tableau 5.

Config.	#enc. layers	#dec. layers	#embed. dim.	#hidden size
XS	1	1	128	256
S	2	2	256	512
M	3	3	384	768

TABLE 5 – Architectures LSTM testées.

4.4 Résultats

Le tableau 6 synthétise les résultats de nos expériences, évalués à l'aide des métriques suivantes : BLEU (*Bilingual Evaluation Understudy*, voir Papineni *et al.* 2002), TER (*Translation Edit Rate*, voir Snover *et al.* 2006) et ChrF (*Character F-score*, voir Popović 2015). Ces métriques ont été calculées à l'aide de l'outil *SacreBLEU* (v2.4.2 [7] ; Post 2018). L'analyse des résultats met en évidence l'importance de la taille du vocabulaire dans la qualité des normalisations obtenues avec les configurations XS, S et M. On observe clairement qu'un vocabulaire de 1 000 sous-mots produit les meilleures performances. À l'inverse, lorsque le nombre de sous-mots augmente au-delà de 1 000, les performances se détériorent considérablement. La configuration S obtient les meilleurs résultats (BLEU = 87.08, TER = 7.35, ChrF = 95.02).

Taille du vocab.	Configuration XS			Configuration S			Configuration M		
	BLEU	TER	ChrF	BLEU	TER	ChrF	BLEU	TER	ChrF
500 sous-mots	81.63	10.12	93.90	82.36	9.54	94.01	80.72	10.32	93.44
1 000 sous-mots	86.64	7.69	94.93	**87.08**	**7.35**	**95.02**	86.18	7.76	94.70
2 000 sous-mots	53.18	29.49	69.01	52.39	30.57	69.28	51.74	31.19	68.22
3 000 sous-mots	43.91	36.32	59.92	43.07	37.70	60.10	42.64	38.11	59.06
4 000 sous-mots	39.20	39.92	54.66	38.44	41.68	54.86	38.11	41.98	53.77

TABLE 6 – Comparaison des configurations XS, S et M (voir table 5). La configuration S obtient les meilleurs résultats avec un vocabulaire de 1 000 sous-mots.

Après avoir analysé des prédictions générées par le meilleur modèle obtenu – que nous appelons *FreEM SemiD norm* (*French Early Modern Semi-Diplomatic Normalisation*, Solfrini & Gabay 2025) –, nous avons identifié plusieurs types d'erreurs liées à la segmentation de la chaîne graphique. Par exemple, certaines séquences agglutinées dans le texte source, telles que *descarlate* au lieu de *d'escarlate*, ne sont pas séparées par l'ajout d'une apostrophe. À l'inverse, certains mots dans le texte source qui ne requièrent aucune normalisation, comme *dame*, sont incorrectement modifiés par l'ajout d'une apostrophe : *dame* devient *d'ame*, ce qui altère le sens du terme. L'élargissement du jeu de données, ainsi qu'une attention majeure à la gestion des apostrophes et des séquences agglutinées, pourraient orienter les futures améliorations du modèle afin d'optimiser sa précision sur ce type de données.

7. Signatures des métriques :
   — BLEU : `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.2`
   — chrF : `nrefs:1|case:lc|tok:tercom|norm:no|punct:yes|asian:no|version:2.4.2`
   — TER : `nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.4.2`

.

5 Conclusions et prochaines étapes

Cet article a permis de démontrer la portée et la faisabilité d'une normalisation semi-diplomatique automatique du moyen français, grâce à un ensemble de données d'environ 40 000 lignes et aux règles qui régissent la normalisation de ce corpus. Les performances d'un premier modèle basé sur une architecture LSTM ont confirmé l'intérêt de cette approche.

Dans cette étude, la configuration S avec 1 000 sous-mots apparaît comme un bon point d'équilibre, mais tout changement de taille du corpus ou de langue nécessiterait un ajustement des paramètres. Notre approche est donc généralisable à d'autres projets de normalisation linguistique, mais le vocabulaire optimal dépend de la morphologie de la langue et de sa variabilité orthographique. En ce qui concerne d'autres données du XVI^{e} qui proviendraient de sources différentes, la nature des phénomènes à normaliser pourrait varier, ce qui pourrait affecter les performances du modèle et nécessiter son *fine-tuning* avec un autre corpus normalisé manuellement.

D'autres types de modèles, comme les *Transformers*, mériteraient d'être explorés, notamment en combinant le *fine-tuning* de grands modèles de langue (LLM) avec l'ajout de données synthétiques. Par ailleurs, l'élargissement du corpus permettrait de renforcer la robustesse du modèle. D'une part, la normalisation manuelle de nouveaux textes est en cours dans le cadre de notre projet, et ces textes viendront prochainement enrichir ce premier ensemble de données. D'autre part, nous espérons que d'autres spécialistes du XVI^{e} siècle pourront également produire des données normalisées selon les mêmes règles. Pour conclure, nous souhaitons affiner ces règles dans une perspective de diachronie longue, afin de les adapter à des corpus plus étendus – incluant, par exemple, des textes médiévaux – et de développer des modèles communs de normalisation automatique.

Accès aux données et au modèle

L'ensemble des données utilisées pour les expériences, ainsi que notre modèle, sont accessibles à l'adresse suivante : https://github.com/soniasol/FreEM-SemiD-norm.

Financement

Les recherches ont été menées dans le cadre du projet SETAF (FNS n° 205056), dirigé par Daniela Solfaroli Camillocci.

Remerciements

Les calculs ont été effectués à l'Université de Genève, en utilisant le service HPC. Tous les auteurs ont contribué à la normalisation manuelle des données, tandis que la rédaction de l'article et l'entraînement des modèles ont été réalisés par Sonia Solfrini. Nous remercions Ariane Pinche, Lucence Ing, Nicolas Fornerod et Geneviève Gross pour leurs contributions aux discussions sur les règles de normalisation, ainsi que Maxime Humeau et Dennis Bontempi pour leurs suggestions concernant les scripts. Nous remercions également nos relecteur·trice·s – Hadrien Dami, Brigitte Roux, Nathalie Szczech et Raphaël Rubino – ainsi que les relecteur·trice·s anonymes. Enfin, nous remercions tout particulièrement Simon Gabay pour ses précieux conseils et sa supervision attentive.

Références

AYRES-BENNETT W. & CARON P. (2016). Periodization, Translation, Prescription and the Emergence of Classical French. *Transactions of the Philological Society*, **114**(3), 339–390. DOI : 10.1111/1467-968X.12081.

BARBICHE B. & CHATENET M. (1993). *L'Édition des textes anciens. XVIe - XVIIIe siècle*. Volume 1 de Documents & Méthodes. Paris : Inventaire général - ELP. 2e éd. - avec le concours scientifique de l'École nationale des Chartes, HAL : hal-02270825.

BAWDEN R., POINHOS J., KOGKITSIDOU E., GAMBETTE P., SAGOT B. & GABAY S. (2022). Automatic Normalisation of Early Modern French. In *LREC 2022 - 13th Language Resources and Evaluation Conference*, p. 3354–3366, Marseille, France : European Language Resources Association. HAL : hal-03540226.

BERTHOUD G. (1980). Les Impressions genevoises de Jean Michel (1538-1544). In J.-D. CANDAUX & B. LESCAZE, Éds., *Cinq siècles d'imprimerie genevoise*, volume 1, p. 55–88. Genève : Droz.

BOLLMANN M. (2019). A large-scale comparison of historical text normalization systems. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 3885–3898, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1389.

BOUILLON P., CHAZALON C., CORAM-MEKKEY S., FALQUET G., GERLACH J., MARCHAND-MAILLET S., MOCCOZET L., MUTAL J., RUBINO R. & SORBI M. (2024). RCnum : A Semantic and Multilingual Online Edition of the Geneva Council Registers from 1545 to 1550. In C. SCARTON, C. PRESCOTT, C. BAYLISS, C. OAKLEY, J. WRIGHT, S. WRIGLEY, X. SONG, E. GOW-SMITH, M. FORCADA & H. MONIZ, Éds., *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)*, p. 21–22, Sheffield, Royaume-Uni : European Association for Machine Translation (EAMT).

CARBONNIER-BURKARD M. (2014). Salut par la foi, salut par la lecture : les nouveaux abécédaires en français au XVIe siècle. In *Protestantisme et éducation dans la France moderne*. LARHRA.

CATACH N. (1968). *L'Orthographe française à l'époque de la Renaissance : auteurs, imprimeurs, ateliers d'imprimerie*. Genève : Droz.

CHAGUÉ A. & CLÉRICE T. (2023). "I'm Here to Fight for Ground Truth"' : HTR-United, a Solution Towards a Common for HTR Training Data. In *Digital Humanities 2023 : Collaboration as Opportunity*, Graz, Autriche : Alliance of Digital Humanities Organizations and University of Graz. HAL : hal-04094233.

CHAGUÉ A., CLÉRICE T. & ROMARY L. (2021). HTR-United : Mutualisons la vérité de terrain ! In *DH Nord 2021 - Publier, partager, réutiliser les données de la recherche : les data papers et leurs enjeux*, Lille, France : MESHS. HAL : hal-03398740.

CHAZALON C., CORAM-MEKKEY S., DAMI H., GROSS G., ISOZ A., SANTSCHI C. & VERNHES RAPPAZ S., Éds. (2021). *Registres du Conseil de Genève à l'époque de Calvin. Tome VIII : Du 1er janvier au 31 décembre 1543*, volume 615 de *Travaux d'Humanisme et Renaissance*. Genève : Droz. Préface de François Longchamp.

CLÉRICE T., JANÈS J., PILLA J., CAMPS J.-B., PINCHE A., GILLE-LEVENSON M. & JOLIVET V. (2024). Pyrrha, a Language Independant Post Correction App for POS and Lemmatization. DOI : 10.5281/zenodo.2325427.

COMBETTES B. & CARLIER A. (2022). Du Moyen français au français préclassique : aspects de l'énoncé complexe. *Le Français préclassique (1500-1650)*, **24**, 65–96. HAL : hal-03950029.

COMBETTES B. & MARCHELLO-NIZIA C. (2010). La Périodisation en linguistique historique : le cas du français préclassique. In *Le Changement en français : études de linguistique diachronique*, volume 89 de Sciences pour la communication, p. 129–134. Berne : P. Lang.

DEMONET M.-L. & L'ÉQUIPE BVH (2011). Epistemon - Principes éditoriaux. https://www.bvh.univ-tours.fr/Epistemon/principeseditoriaux.asp. Université de Tours, Les Bibliothèques Virtuelles Humanistes. Consulté le 21 mars 2025.

DUVAL F., Éd. (2006). *Pratiques philologiques en Europe*. Études et rencontres. Paris : Publications de l'École nationale des chartes.

FRAGONARD M.-M. & KOTLER E. (1994). *Introduction à la langue du XVIe siècle*. Paris : Nathan.

GABAY S. & BARRAULT L. (2020). Traduction automatique pour la normalisation du français du XVIIe siècle. In C. BENZITOUN, C. BRAUD, L. HUBER, D. LANGLOIS, S. OUNI, S. POGODALLA & S. SCHNEIDER, Éds., *Actes de JEP-TALN-RECITAL 2020*, p. 213–222, Nancy, France : ATALA. HAL : hal-02784770.

GABAY S., CAMPS J.-B. & CLÉRICE T. (2022a). Manuel d'annotation linguistique pour le français moderne (XVIe -XVIIIe siècles). Pré-publication / document de travail, HAL : hal-02571190.

GABAY S., CLÉRICE T. & REUL C. (2023). OCR17 : Ground Truth and Models for 17th c. French Prints (and hopefully more). *Journal of Data Mining and Digital Humanities*. DOI : 10.46298/jdmdh.6492, HAL : hal-02577236.

GABAY S., ORTIZ SUAREZ P., BAWDEN R., BARTZ A., GAMBETTE P. & SAGOT B. (2022b). Le Projet FreEM : ressources, outils et enjeux pour l'étude du français d'Ancien Régime. In Y. ESTÈVE, T. JIMÉNEZ, T. PARCOLLET & M. ZANON BOITO, Éds., *Actes de TALN-RECITAL 2022*, p. 154–165, Avignon, France : ATALA. HAL : hal-03701524.

GABAY S., PINCHE A., CHRISTENSEN K. & CAMPS J.-B. (2024). SegmOnto : A Controlled Vocabulary to Describe and Process Digital Facsimiles. *Journal of Data Mining and Digital Humanities*. DOI : 10.46298/jdmdh.12689, HAL : hal-04343404.

GABAY S., RIGUET M. & BARRAULT L. (2019). A Workflow For On The Fly Normalisation Of 17th c. French. In *DH2019*, Utrecht, Pays-Bas : ADHO. HAL : hal-02276150.

GARNIER I. (2019). Traduire la poésie de la Renaissance en français moderne : translation partielle commentée du Dialogue en forme de vision nocturne de Marguerite de Navarre (1524). *Revue italienne d'études françaises*, (9). DOI : 10.4000/rief.3877.

GIRAUD Y. (1997). Protocole pour l'édition de textes imprimés en moyen français. *Bulletin de liaison de la Société Française d'Etude du Seizième Siècle*, **42**, 37–40.

GOUGENHEIM G. (1973). *Grammaire de la langue française du seizième siècle*. Volume 8 de Connaissance des langues. Paris : Picard.

GUILLOT C., PRÉVOST S. & LAVRENTIEV A. (2013). *Principes d'annotation Cattex09*. Rapport interne, École normale supérieure de Lyon, Lyon.

GUYOTJEANNIN O. & VIELLIARD F. (2014). *Conseils pour l'édition des textes médiévaux. Fascicule I : conseils généraux*. Paris : Comité des travaux historiques et scientifiques - avec le concours scientifique de l'École nationale des chartes.

HUCHON M. (1988). *Le Français de la Renaissance*. Volume 2389 de Que sais-je ? Paris : Presses Universitaires de France.

KEMP W. (2004). La Redécouverte des éditions de Pierre de Vingle imprimées à Genève et à Neuchâtel (1533-1536). In J.-F. GILMONT & W. KEMP, Éds., *Le Livre évangélique en français avant Calvin*, p. 147–177. Turnhout : Brepols.

KIESSLING B., TISSOT R., STOKES P. & STÖKL BEN EZRA D. (2019). eScriptorium : An Open Source Platform for Historical Document Analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2. DOI : 10.1109/ICDARW.2019.10032.

KINGMA D. P. & BA J. (2015). Adam : A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations (ICLR)*, San Diego, California, États-Unis.

KUDO T. & RICHARDSON J. (2018). SentencePiece : A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In E. BLANCO & W. LU, Éds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 66–71, Bruxelles, Belgique : Association for Computational Linguistics. DOI : 10.18653/v1/D18-2012.

MALINGRE M. (1533). *Moralite de la maladie de Chrestiente a VIII personnages*. [Neuchâtel] : [Pierre de Vingle]. DOI : 10.3931/e-rara-563.

MEYER P. (1910). Instructions pour la publication des anciens textes français. *Bibliothèque de l'École des chartes*, **71**, 224–233.

OTT M., EDUNOV S., BAEVSKI A., FAN A., GROSS S., NG N., GRANGIER D. & AULI M. (2019). Fairseq : A Fast, Extensible Toolkit for Sequence Modeling. In W. AMMAR, A. LOUIS & N. MOSTAFAZADEH, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, p. 48–53, Minneapolis, Minnesota, États-Unis : Association for Computational Linguistics. DOI : 10.18653/v1/N19-4009.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : A Method for Automatic Evaluation of Machine Translation. In P. ISABELLE, E. CHARNIAK & D. LIN, Éds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 311–318, Philadelphia, Pennsylvania, États-Unis : Association for Computational Linguistics. DOI : 10.3115/1073083.1073135.

PINCHE A., CHRISTENSEN K. & GABAY S. (2022). Between Automatic and Manual Encoding. In *TEI 2022 conference : Text as data*, Newcastle, Royaume-Uni. DOI : 10.5281/zenodo.7092214, HAL : hal-03780302.

PINCHE A. & CLÉRICE T. (2022). Cortado. v. 2.0.0, DOI : 10.5281/zenodo.6818057.

PINCHE A., CLÉRICE T., CHAGUÉ A., CAMPS J.-B., VLACHOU-EFSTATHIOU M., GILLE LEVENSON M., BRISVILLE-FERTIN O., BOSCHETTI F., FISCHER F., GERVERS M., BOUTREUX A., MANTON A. & GABAY S. (2023). CATMuS Medieval. v. 1.0.0, DOI : 10.5281/zenodo.10066218.

POPOVIĆ M. (2015). ChrF : Character n-gram F-score for Automatic MT Evaluation. In O. BOJAR, R. CHATTERJEE, C. FEDERMANN, B. HADDOW, C. HOKAMP, M. HUCK, V. LOGACHEVA & P. PECINA, Éds., *Proceedings of the Tenth Workshop on Statistical Machine Translation*, p. 392–395, Lisbonne, Portugal : Association for Computational Linguistics. DOI : 10.18653/v1/W15-3049.

POST M. (2018). A call for clarity in reporting BLEU scores. In O. BOJAR, R. CHATTERJEE, C. FEDERMANN, M. FISHEL, Y. GRAHAM, B. HADDOW, M. HUCK, A. J. YEPES, P. KOEHN, C. MONZ, M. NEGRI, A. NÉVÉOL, M. NEVES, M. POST, L. SPECIA, M. TURCHI & K. VERSPOOR, Éds., *Proceedings of the Third Conference on Machine Translation : Research Papers*, p. 186–191, Bruxelles, Belgique : Association for Computational Linguistics. DOI : 10.18653/v1/W18-6319.

PRÉVOST S., GUILLOT C., LAVRENTIEV A. & HEIDEN S. (2013). *Jeu d'étiquettes morphosyntaxiques CATTEX2009*. Rapport interne, École normale supérieure de Lyon, Lyon. Version 2.0.

ROQUES M. (1926). Règles pratiques pour l'édition des anciens textes français et provençaux. *Bibliothèque de l'École des chartes*, **87**, 453–459.

RUBINO R., CORAM-MEKKEY S., GERLACH J., MUTAL J. D. & BOUILLON P. (2024a). Automatic Normalisation of Middle French and Its Impact on Productivity. In R. SPRUGNOLI & M. PASSAROTTI,

Éds., *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, p. 176–189, Turin, Italie : ELRA and ICCL.

RUBINO R., GERLACH J., MUTAL J. & BOUILLON P. (2024b). Normalizing without Modernizing : Keeping Historical Wordforms of Middle French while Reducing Spelling Variants. In K. DUH, H. GOMEZ & S. BETHARD, Éds., *Findings of the Association for Computational Linguistics : NAACL 2024*, p. 3394–3402, Mexico, Mexique : Association for Computational Linguistics. DOI : 10.18653/v1/2024.findings-naacl.215.

SENNRICH R., HADDOW B. & BIRCH A. (2016). Neural Machine Translation of Rare Words with Subword Units. In K. ERK & N. A. SMITH, Éds., *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1715–1725, Berlin, Allemagne : Association for Computational Linguistics. DOI : 10.18653/v1/P16-1162.

SNOVER M., DORR B., SCHWARTZ R., MICCIULLA L. & MAKHOUL J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas : Technical Papers*, p. 223–231, Cambridge, Massachusetts, États-Unis : Association for Machine Translation in the Americas.

SOLFRINI S. & GABAY S. (2024). Catmus Gothic Print. v. 1.0.0, DOI : 10.5281/zenodo.10599911.

SOLFRINI S. & GABAY S. (2025). FreEM SemiD norm model. v. 1.0.0, DOI : 10.5281/zenodo.15551750.

SOLFRINI S., GABAY S., GROSS G., BEAULNES P.-O., MARQUES OLIVEIRA A. & SOLFAROLI CAMILLOCCI D. (2023a). Guide de transcription pour les imprimés français du XVIe siècle en caractères gothiques. Pré-publication / document de travail, HAL : hal-04281804.

SOLFRINI S., GABAY S., HUMEAU M., PINCHE A., BEAULNES P.-O., MARQUES OLIVEIRA A., GROSS G. & SOLFAROLI CAMILLOCCI D. (2024). Océriser les imprimés du XVIe siècle en langue française : le cas d'un corpus romand en caractères gothiques. In *Humanistica 2024*, Meknès, Maroc : Association francophone des humanités numériques. HAL : hal-04555002.

SOLFRINI S., GROSS G., ROUX B., SZCZECH N., BEAULNES P.-O., MARQUES OLIVEIRA A. & SOLFAROLI CAMILLOCCI D. (2023b). Étudier le « groupe de Neuchâtel » : de l'édition des Faits à un corpus numérique de la première Réforme romande. In *Humanistica 2023*, Poster, Genève, Suisse : Association francophone des humanités numériques. HAL : hal-04097381.

SOUVAY G. & PIERREL J.-M. (2009). LGeRM Lemmatisation des mots en Moyen Français. *Revue TAL : traitement automatique des langues*, **50**(2), 21. HAL : halshs-00396452.

STUTZMANN D. (2011). Paléographie statistique pour décrire, identifier, dater... normaliser pour coopérer et aller plus loin ? In F. FISCHER, C. FRITZE & G. VOGELER, Éds., *Kodikologie und Palä̈o-graphie im digitalen Zeitalter = Codicology and Palaeography in the Digital Age*, volume 2 de *Schriften des Instituts für Dokumentologie und Editorik*, p. 247–277. BoD.

UETANI T., PORTE G., BREUIL S. & DUBOC M. (2016). The BVH in Tours : Digital Library of Image, Text, and Data. In *TEI Conference 2016*, Vienne, Autriche : TEI Consortium.

VACHON C. H. (2010). *Le Changement linguistique au XVIe siècle. Une étude basée sur des textes littéraires français*. Strasbourg : Éditions de linguistique et de philologie.

VOGELER G. (2023). Proto-editions : Historians and the "something between digital image and digital scholarly edition". In *Digital Humanities 2023 : Collaboration as Opportunity*, Graz, Autriche. DOI : 10.5281/zenodo.8107922.