Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Normalisation du moyen français : comparaison de modèles pré-entraînés

Raphael Rubino[1] Sandra Coram-Mekkey[2] Pierrette Bouillon[1]
(1) TIM/FTI, Université de Genève, 1205 Genève, Suisse
(2) Fondation de l'Encyclopédie de Genève, Genève, Suisse
raphael.rubino@unige.ch, coram.mekkey@gmail.com,
pierrette.bouillon@unige.ch

RÉSUMÉ

Pour les humanités numériques, la normalisation de documents historiques est une tâche essentielle qui consiste à réduire les variations orthographiques et à corriger les éventuelles erreurs provenant de la transcription automatique du contenu original. De nombreux travaux l'envisagent comme une tâche de traduction automatique (TA). Cependant, le manque de données parallèles pertinentes pour le domaine et pour la période historique concernée limite les performances des systèmes de TA développés et en fait une tâche comparable à la TA des langues peu dotées. Cette étude vise à évaluer les performances de modèles pré-entraînés afin de déterminer s'ils peuvent être bénéfiques pour la tâche de normalisation de textes en moyen français, pour laquelle les textes source et cible sont des variantes du français contemporain. Pour garantir que nos données ne soient pas connues des grands modèles utilisés, nous exploitons un nouveau corpus parallèle, extrait de documents administratifs rédigés au milieu du XVIe siècle en moyen français et n'ayant jamais été publiés sous leurs formes transcrite ou normalisée. L'étude compare les modèles pré-entraînés populaires en terme d'architecture, de type encodeur–décodeur et décodeur seul, ainsi qu'un modèle Transformer entraîné uniquement pour la tâche de normalisation sur nos données. Les résultats montrent que l'architecture encodeur–décodeur est la plus performante parmi les modèles évalués et soulignent l'utilité du pré-entraînement.

ABSTRACT

Middle French Normalisation : Comparison of Pretrained Models

In the digital humanities, the standardisation of historical documents is a key task that involves reducing spelling variations and correcting any errors arising from the automatic transcription of the original content. Several previous studies have considered this task as a machine translation (MT) task. However, the lack of relevant parallel in-domain data from the relevant time period of our study limits the performance of the developed MT systems and makes this task comparable to low-resource MT. This study aims to evaluate the performance of pre-trained models in order to determine whether they can be beneficial for the task of Middle French normalisation, where the source and target texts are variants of contemporary French. To ensure that our data is not recognised by the models used, we constructed a new parallel corpus drawn from administrative documents written in Middle French during the mid-16th century which have never been published in transcribed or normalised form. The study compares popular pre-trained models in terms of architecture, specifically, encoder–decoder and decoder-only models, as well as a Transformer model trained solely for the normalisation task on our data. The results show that the encoder–decoder architecture performs best among the models evaluated and highlight the value of pre-training.

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : articles scientifiques originaux, pages 386–404.

MOTS-CLÉS : Humanités numériques, normalisation de textes historiques, moyen français, grands modèles de langue.

KEYWORDS: Digital Humanities, Historical Text Normalisation, Middle French, Large Language Models.

1 Introduction

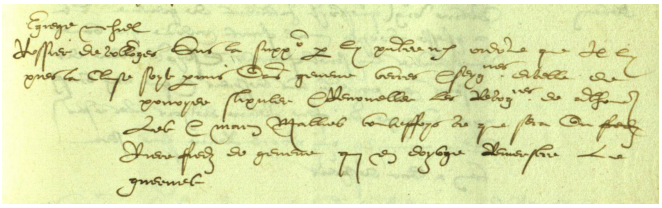
L'objectif principal des recherches en humanités numériques est la préservation de documents historiques. Cette dernière implique une première étape de numérisation du contenu textuel avec des systèmes de reconnaissance optique de caractères, suivie d'une étape de traitement automatique du langage naturel (TALN) pour normaliser les textes. Cette seconde étape, de normalisation, est particulièrement utile pour de grandes collections de documents historiques contenant du texte à numériser. Elle vise à réduire les variantes orthographiques, très fréquentes dans les manuscrits originaux, en raison de l'absence de normes précises à l'époque, et à corriger les erreurs liées à la numérisation. Elle améliore ainsi non seulement la lisibilité des textes par une audience large et non spécialisée, mais facilite aussi les travaux en TALN tels que l'indexation et la recherche d'information (Bollmann, 2013; Sánchez-Martínez *et al.*, 2013).

Les méthodes utilisées jusqu'ici pour la normalisation suivent celles de la traduction automatique (TA) et s'appuient récemment sur les réseaux de neurones artificiels (Bollmann, 2018) et les grands modèles pré-entraînés (Rubino *et al.*, 2024b). Parmi les difficultés liées au traitement automatique des textes historiques (Vilkomir & Herndon, 2024), le manque de données parallèles indispensables pour l'apprentissage supervisé de modèles de normalisation impacte tout particulièrement l'utilisation de réseaux de neurones dont l'entraînement nécessite de grandes quantités de données. De plus, malgré les avancées récentes des grands modèles de langue pré-entraînés et leur application avec succès à une variété de tâches en TALN, le manque de données avec les variantes anciennes des langues modernes ne les rend pas nécessairement utiles pour le traitement de textes historiques.

Dans cet article, nous proposons d'analyser les performances de grands modèles pré-entraînés pour la tâche de normalisation du moyen français du XVIe siècle (1545 et 1546). Cette tâche, combinée au domaine de spécialité de nos données, à savoir des textes administratifs qui reprennent les décisions prises localement dans un canton suisse, s'apparente à une tâche de TA pour les langues peu dotées dans un domaine spécialisé. Ces caractéristiques impliquent des difficultés pour l'apprentissage supervisé de modèles de normalisation automatique, en raison de la rareté des données d'apprentissage. Nous comparons les performances de différentes architectures neuronales, toutes basées sur le Transformer (Vaswani *et al.*, 2017) et communément utilisées en TA. Cette comparaison inclue l'ajustement de modèles pré-entraînés de type encodeur–décodeur et décodeur seul, ainsi qu'une baseline reposant sur l'entraînement d'un Transformer sans pré-entraînement. Afin de conduire ces expériences, nous exploitons un corpus parallèle développé par des experts du domaine composé de la transcription et la normalisation des textes administratifs. [1]

L'organisation de cet article est la suivante : nous décrivons le contexte de notre travail, les objectifs et les questions de recherche en Section 2, suivi d'une présentation des travaux précédents pertinents pour notre étude en Section 3. Puis, dans la Section 4, la méthodologie est introduite, incluant les détails sur les modèles et données utilisés dans nos expériences. Finalement, la Section 5 présente les

1. Les données et modèles issus de notre étude seront mis à disposition librement.



Transcription : (egrege michiel rossier de collonges pres la cluse) — sus la supplication par luy presentee etc. ordonne que il luy soyt permis dans geneve terres et seygneuries dicelle de povoyer stipuler et renoveller les recognoyssances de anthoenne let et marin malliet toutesfoys ce que sera du riere fiedz de geneve qui en doybge remecstre le guernet

Normalisation : (Égrège Michel Rossier, de Collonges, près de L'Écluse) - Sur la supplication par lui présentée etc., il a été ordonné que lui soit permis, à Genève, terres et seigneuries d'icelle, de pouvoir stipuler et renouveler les reconnaissances d'Antoine Lect et de Marin Maillet ; toutefois, de ce qui sera de l'arrière-fief de Genève, qu'il en doive remettre le carnet.

FIGURE 1 – Extrait des registres manuscrits du 19 mai 1545 accompagné de sa transcription et de sa normalisation manuelle.

résultats expérimentaux ainsi que leur analyse, avant de conclure en Section 6.

2 Objectif et questions de recherche

L'étude porte sur la normalisation orthographique de registres manuscrits, composés de rapports et résumés de conseils rédigés en moyen français du XVIe siècle. Un échantillon extrait des données originales, ainsi que la transcription et la normalisation du contenu textuel, sont présentés par la Figure 1. La normalisation de ces documents, après leur transcription manuelle, s'inscrit dans un projet plus vaste visant à produire une édition sémantique et multilingue en ligne de ces registres, sur la période de 1545 à 1550 (Bouillon *et al.*, 2024). Ce projet transversal repose sur une synergie entre plusieurs facultés et domaines d'expertise, incluant l'histoire, et plus particulièrement la connaissance pointue des évènements relatifs à la localisation et la période temporelle, la paléographie, le TALN, la traduction et la linguistique.

L'aspect technique du projet en termes de TALN consiste à automatiser la normalisation du contenu transcrit manuellement des registres, et à développer de nouvelles fonctionnalités qui rendront ces documents d'archives accessibles à un large public. La normalisation automatique de documents historiques, pour lesquels peu de données parallèles existent, est considérée ici comme une tâche de TA pour des langues peu dotées. Ainsi, nous proposons d'évaluer sur cette tâche des modèles ajustés sur nos données, notamment avec l'ajustement de plusieurs types de modèles pré-entraînés, en comparaison avec des modèles entrainés uniquement sur des données de normalisation.

Les questions de recherche sont les suivantes :

— Le pré-entraînement améliore-t-il la normalisation automatique d'une langue sous-représentée, comme le moyen français ?
— Les tâches de pré-entraînement, pouvant varier selon les modèles, impactent-elles les performances de normalisation malgré un ajustement identique ?
— Quelle architecture, encodeur–décodeur ou décodeur seul, permet d'atteindre la meilleure normalisation ?

3 Travaux précédents

De manière similaire à la TA, la normalisation des textes historiques s'est d'abord appuyée sur des règles (Baron *et al.*, 2009; Bollmann *et al.*, 2011), puis des approches statistiques (Pettersson *et al.*,

2013, 2014). Plus récemment, plusieurs auteurs ont proposé des méthodes fondées sur les réseaux neuronaux (Bollmann & Søgaard, 2016; Korchagina, 2017; Tang *et al.*, 2018), avec des résultats qui peuvent encore rester inférieurs aux approches fondées sur des règles. Par exemple, Zilio *et al.* (2024) comparent une approche fondée sur des règles à trois modèles de type encodeur–décodeur pré-entraînés et ajustés pour la tâche de normalisation de textes en portugais du XVIIIe siècle. Les résultats montrent qu'un seul modèle neuronal surpasse les performances d'un ensemble de règles.

Les approches statistiques obtiennent également des résultats comparables à ceux des modèles neuronaux, comme le montrent Bawden *et al.* (2022). Dans cette étude, les modèles neuronaux sont surpassés par un modèle de TA statistique combiné à une étape de post-traitement basée sur un lexique. Ce dernier peut être considéré comme un ensemble de règles appliquées *a posteriori* pour corriger la sortie du modèle de normalisation automatique. Cette étude est particulièrement pertinente pour notre travail, car elle porte sur une variante historique du français, à savoir le français moderne précoce, tandis que notre propre étude se concentre sur le moyen français.

À notre connaissance, peu d'études se sont concentrées sur le moyen français, une variante de la langue française utilisée entre le XIVe et le XVIe siècle (Smith, 2002). Solfrini *et al.* (2025) ont notamment développé des modèles fondés sur l'architecture LSTM (*Long-short Term Memory* (Hochreiter & Schmidhuber, 1997)), avec des résultats qui motivent l'utilisation de réseaux neuronaux pour la tâche de normalisation du moyen français pour des textes datant du XVIe siècle. De plus, Rubino *et al.* (2024a) ont mesuré les performances d'un système Transformer (Vaswani *et al.*, 2017), de type encodeur–décodeur pré-entraîné, sur une tâche similaire à Solfrini *et al.* (2025), la normalisation du moyen français sur la même période, mais sur un autre jeu de données. Les résultats de cette étude montrent que la normalisation automatique contribue à réduire le travail manuel d'experts au travers de la post-édition des hypothèses de normalisation, en comparaison à la normalisation manuelle complète des textes historiques.

4 Méthodologie

Les sous-sections suivantes présentent le jeu de données élaboré manuellement et utilisé dans nos expériences, ainsi que les modèles entraînés ou ajustés pour la tâche de normalisation.

4.1 Données

La source de nos données est la transcription manuelle de manuscrits administratifs rédigés en moyen français entre janvier 1545 et avril 1546 en Suisse. La transcription des manuscrits originaux et leur normalisation ont été effectuées manuellement par des historiens et des paléographes spécialisés de cette période et de la région géographique en question, ce qui assure une connaissance pointue des événements politiques et sociaux locaux. L'ensemble des *règles* suivies par les experts pour la transcription manuelle des manuscrits est présenté en Annexe A. Aucun usage de la reconnaissance automatique des caractères n'est prévu dans notre étude présenté dans cet article.

Un ensemble de *règles* de normalisation est appliqué au texte original, incluant notamment des corrections orthographiques, l'ajout de la ponctuation et des majuscules, l'accord en genre et nombre des noms et verbes, la modification de l'ordre des mots dans les séquences nom–adjectif et nom–verbe, entre autres. Une liste exhaustive des opérations de normalisation est présentée en Annexe B.

Contrairement à la plupart des travaux précédents sur la normalisation des textes historiques, nous ne nous appuyons pas sur la reconnaissance optique de caractères pour transcrire le contenu textuel. Nous garantissons ainsi qu'aucune erreur liée à une transcription automatique ne se trouve dans le texte source et pouvons nous focaliser sur la normalisation des variantes orthographiques et structures syntaxiques, ainsi que sur l'évaluation des différents modèles présentés dans la Section 4.2. [2]

Nous divisons le corpus parallèle par mois. Ce découpage nous permet de considérer chaque mois comme corpus de test tandis que le reste du corpus est utilisé pour l'entraînement. Il permet également de mener des expériences sur un corpus relativement petit, composé de 8 373 segments parallèles, de manière comparable à une validation croisée avec 16 sous-ensembles d'entraînement et de test, tout en permettant une étude diachronique fine. Les détails du corpus sont présentés dans le Tableau 2, indiquant que chaque mois contient un nombre variable de segments et que le nombre de mots est plus élevé dans le texte normalisé (cible) que dans l'original (source). Nous mesurons la similarité entre la source et la cible normalisée grâce à la métrique chrF (Popović, 2015) (un score élevé indique une similarité élevée) et observons que la majorité des mois composant notre jeu de données obtient une similarité source–cible d'environ 48,3 à 49,7 chrF, sauf juillet, septembre et novembre 1545. Une similarité moindre entre la source et la cible de ces mois pourrait indiquer un plus grand nombre d'opérations de normalisation nécessaire en comparaison aux autres mois.

4.2 Modèles

Tous les modèles utilisés reposent sur l'architecture Transformer (Vaswani *et al.*, 2017) ; les modèles pré-entraînés sont disponibles via le pôle d'échange de modèles hébergé par *Hugging Face* [3].

Les modèles du type encodeur–décodeur non pré-entraînés se fondent sur l'implémentation Marian (Junczys-Dowmunt *et al.*, 2018). Nous évaluons deux profondeurs du Transformer, avec 2 et 3 couches pour l'encodeur et pour le décodeur, donnant lieu à deux architectures de Transformer non pré-entraînés notés 2x2 et 3x3. Nous limitons le nombre de couches à 3x3. Les modèles plus profonds n'ont en effet pas montré de meilleures performances dans nos exériences préliminaires, ce qui rejoint aussi les résultats obtenus dans des travaux précédents sur la traduction automatique de langues peu dotées (Rubino *et al.*, 2020). Afin de mesurer l'impact du pré-entraînement de modèles encodeur–décodeur sur la tâche de normalisation, en comparaison avec des modèles entraînés uniquement sur les données produites pour notre étude, nous testons également deux types de modèles exposés à des données multilingues pendant leur pré-entraînement : M2M100 (Fan *et al.*, 2021), avec les variantes M2M100_418M et M2M100_1.2B ayant 418 millions et 1,2 milliards de paramètres respectivement, et mT5 (Xue *et al.*, 2021) dans sa version avec 580 millions de paramètres noté mT5_580M.

Les modèles de type encodeur–décodeur sont parfois considérés comme redondant pour la tâche de traduction automatique (Gao *et al.*, 2022). Ainsi, nous proposons d'évaluer des modèles de type décodeur seul et de les comparer aux modèles de type encodeur–décodeur. Ces modèles ont été sélectionnés en fonction de leur taille en nombre de paramètres pour une comparaison équitable entre les différentes architectures employées dans notre étude. Ainsi, seuls des modèles relativement petits ont été retenus. Cela a pour avantage de permettre leur ajustement sur du matériel grand public (eg. des GPUs dotés d'une mémoire inférieure à 40Go). Nous avons pris en compte des modèles capables de gérer plusieurs langues et avons retenu les modèles suivants : Gemma 3 (Gemini Team *et al.*, 2023) avec leurs variantes Gemma_270M et Gemma_1B possédant respectivement 270 millions et 1

2. Le glossaire et le corpus parallèle résultant de la normalisation manuelle seront publiés avec cet article.
3. https://huggingface.co/models

milliard de paramètres, LFM2 [4] avec LFM2_350M, LFM2_700M et LFM2_1.2B avec 350 millions, 700 millions et 1,2 milliards de paramètres, et finalement Qwen 3 (Yang *et al.*, 2025) dans sa version avec 600 millions de paramètres notée Qwen_600M.

Tant l'entraînement des modèles encodeur–décodeur que l'ajustement des modèles pré-entraînés ont été réalisés à l'aide du corpus d'entraînement, tandis que l'évaluation a été effectuée en utilisant le corpus de test (ces deux corpus sont présentés dans la section 4.1). Le Tableau 1 reprend les différents modèles testés dans cette étude. Les hyper-paramètres, comme la configuration des modèles Transformer non pré-entraînés ou les taux d'apprentissage testés pendant l'entraînement ou l'ajustement, sont présentés en Annexe C.

nom	architecture	#params.
modèles non pré-entraînés		
Transformer_2x2	enc–dec	18,8M
Transformer_3x3	enc–dec	26,2M
encodeurs–décodeurs pré-entraînés		
M2M100_418M	enc–dec	418M
mT5_580M	enc–dec	580M
M2M100_1.2B	enc–dec	1,2B
décodeurs seuls pré-entraînés		
Gemma_270M	decoder	270M
LFM2_350M	decoder	350M
Qwen_600M	decoder	600M
LFM2_700M	decoder	700M
Gemma_1B	decoder	1B
LFM2_1.2B	decoder	1,2B

TABLE 1 – Modèles évalués dans notre étude sur la tâche de normalisation automatique du moyen français. *#params.* indique le nombre total de paramètres du modèles, M et B signifient respectivement millions et milliards.

		#tokens		
mois & année	#phrases	source	cible	identité
janvier 1545	497	15,4k	16,5k	48,7
février 1545	903	15,8k	16,4k	49,6
mars 1545	502	15,2k	16,0k	48,7
avril 1545	458	13,2k	14,1k	48,3
may 1545	486	14,7k	15,7k	48,6
juin 1545	511	14,8k	15,9k	48,6
juillet 1545	512	15,8k	16,7k	50,2
août 1545	383	12,4k	13,2k	49,7
septembre 1545	347	10,8k	11,4k	51,6
octobre 1545	475	13,6k	14,5k	48,8
novembre 1545	485	14,8k	15,8k	50,1
décembre 1545	608	17,1k	18,2k	49,4
janvier 1546	476	13,1k	14,0k	49,3
février 1546	590	9,0k	9,3k	49,5
mars 1546	605	16,7k	17,7k	48,1
avril 1546	535	15,0k	16,1k	49,1
moyenne	523,3	14,2k	15,1k	49,3
total	8,4k	227,6k	241,7k	

TABLE 2 – Détails des données produites pour notre étude, incluant le nombre de phrases et de mots sources (moyen français) et de mots cibles (normalisés), sur la période de janvier 1545 à avril 1546. (*k* signifie milliers.) La colonne *identité* indique les scores chrF lorsque la source est comparée à la cible.

5 Résultats et analyse

Cette section présente tout d'abord les résultats obtenus par les modèles inclus dans notre étude, suivi par une analyse des erreurs commises par le meilleur modèle d'après notre évaluation (M2M100_1.2B).

4. https://huggingface.co/LiquidAI

5.1 Résultats

modèle	chrF moyenne	chrF σ^2	BLEU moyenne	BLEU σ^2
Transformer_2x2	79,7	4,24	63,8	12,61
Transformer_3x3	79,9	2,98	64,2	6,60
M2M100_418M	90,5	1,41	78,3	4,65
mT5_580M	87,9	2,05	73,5	6,00
M2M100_1.2B	**90,9**	1,41	**79,2**	4,26
Gemma_270M	84,0	1,71	66,8	4,93
LFM2_350M	86,4	1,82	69,5	5,09
Qwen_600M	86,4	0,95	68,7	4,12
LFM2_700M	87,3	1,59	71,0	5,54
Gemma_1B	87,6	1,43	72,0	5,41
LFM2_1.2B	88,1	1,31	72,1	5,38

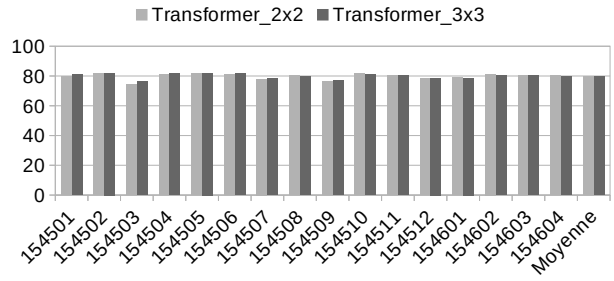
TABLE 3 – Résultats de normalisation moyens sur 16 mois avec variances (σ^2) mesurés par les métriques automatiques chrF et BLEU.

L'évaluation est réalisée en comparant les sorties des modèles aux références normalisées manuellement. Nous utilisons des métriques automatiques couramment employées dans le domaine de la TA, à savoir BLEU (Papineni *et al.*, 2002) et chrF (Popović, 2015), implémentées dans SacreBLEU (Post, 2018). Ces métriques de surface, bien que contestées dans le contexte de la TA (Marie *et al.*, 2021), se justifient ici : la tâche de normalisation est en effet très contrainte et ne présente pas la même variabilité que la traduction de textes modernes. Les deux métriques employées dans notre étude ont des niveaux de granularité différents et ne mesurent donc pas les mêmes phénomènes linguistiques. Le score BLEU évalue le recouvrement de n-grammes de mots et le score chrF mesure le F-score (combinaison de précision et rappel) au niveau des caractères. Les scores mesurés par ces deux métriques permettent de comparer les sorties de plusieurs modèles : plus le score est élevé, plus la similarité entre la sortie d'un modèle et la référence manuellement produite est élevée.

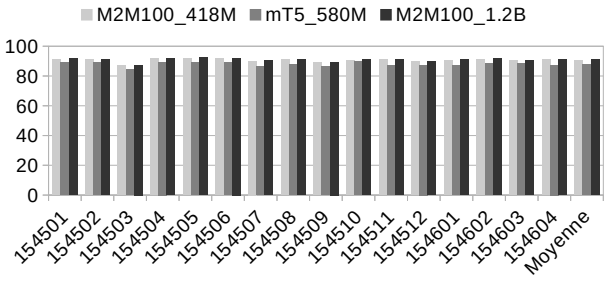
Les scores moyens pour les 16 mois sont présentés dans le Tableau 3 et montrent que le modèle M2M100_1.2B atteint les scores de normalisation les plus élevés, avec 90,9 chrF et 79,2 BLEU, en comparaison aux autres modèles évalués dans notre étude. Afin de comparer les deux modèles aux scores les plus élevés (M2M100_418M et M2M100_1.2B), nous effectuons un test permutationnel basé sur la randomisation approximative pairée (Riezler & Maxwell, 2005), et ce pour chaque mois individuellement. Le modèle M2M100_1.2B obtient de meilleurs scores chrF et BLEU pour 9 mois sur 16 avec $p < 0.05$, en comparaison avec M2M100_418M.

Les modèles non pré-entraînés, quant à eux, obtiennent les scores les plus bas, avec 79,7 chrF et 63,8 BLEU pour le modèle Transformer_2x2, et 79,9 chrF et 64,2 BLEU pour le Transformer_3x3. Il est intéressant de noter que **les modèles pré-entraînés surpassent les modèles non pré-entraînés**, ces derniers n'utilisant que nos données. Ces résultats motivent ainsi l'approche reposant sur l'ajustement de grands modèles pré-entraînés, malgré la faible représentation du moyen français dans les données de pré-entraînement. C'est donc la présence de français contemporain dans les modèles pré-entraînés, ainsi que potentiellement d'autres langues, qui permet d'atteindre les performances de normalisation reportées selon les métriques automatiques utilisées. De plus, parmi les modèles pré-entraînés,

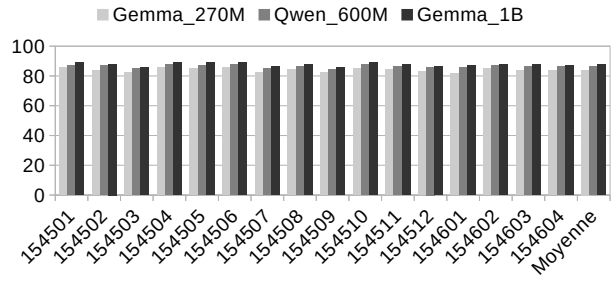
l'architecture encodeur–décodeur surpasse les modèles de type décodeur seul, ce qui souligne l'importance du type d'architecture, et potentiellement des objectifs de pré-entraînement, pour la tâche de normalisation du moyen français.



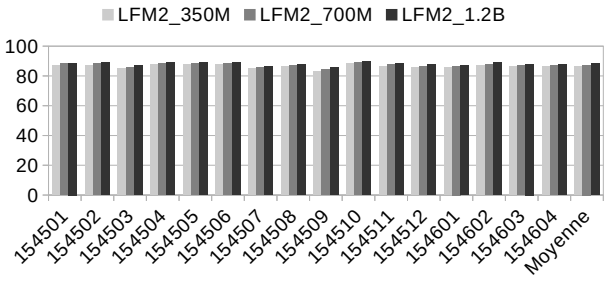
(a) Résultats obtenus par les modèles Transformer de type encodeur–décodeur non pré-entraînés.



(b) Résultats obtenus par les modèles de type encodeur–décodeur pré-entraînés.



(c) Résultats obtenus par les décodeurs seuls pré-entraînés de la famille *Gemma* et *Qwen*.



(d) Résultats obtenus par les décodeurs seuls pré-entraînés de la famille *LFM2*.

FIGURE 2 – Résultats par mois et moyennes obtenues par les modèles évalués dans notre étude, mesurés par la métrique chrF.

Les résultats obtenus avec les mêmes modèles, selon la métrique chrF, sont présentés pour chaque mois individuellement sur la Figure 2. [5] Ces quatre figures reprennent les résultats respectifs obtenus par les modèles Transformer non pré-entraînés, les modèles pré-entraînés de type encodeur–décodeur et ceux de type décodeur seul. Nous proposons de comparer les modèles de la famille Gemma à un modèle de type Qwen sur la Figure 2c et les modèles de la famille LFM2 sur la Figure 2d.

Parmi les modèles pré-entraînés de type encodeur–décodeur (Figure 2b), la famille de modèles M2M100 permet d'atteindre des scores chrF et BLEU plus élevés comparé à mT5_580M sur chaque mois représenté dans notre jeu de données. Même le modèle avec moins de paramètres, M2M100_418M, dépasse les performances de mT5_580M. Ceci est probablement dû aux objectifs de pré-entraînement différents, bien que ces deux modèles soient multilingues. En effet, mT5_580M a été pré-entraîné selon l'approche du modèle de langue masqué tandis que les modèles de la famille M2M100 ont été pré-entraînés sur la tâche de TA directement. Ces résultats soulignent l'importance des objectifs de pré-entraînement dans la sélection d'un modèle à ajuster pour la tâche de normalisation du moyen français et montrent qu'il est **approprié d'utiliser des modèles spécialisés dans la TA** pour notre tâche.

Nous observons aussi que **l'augmentation du nombre de paramètres permet d'améliorer les scores de normalisation**, mesurés par les métriques automatiques. C'est le cas pour le Transformer non pré-entraîné, pour les modèles de type encodeur–décodeur pré-entraînés M2M100 et pour les

5. Les résultats mesurés par la métrique BLEU pour chaque mois sont présentés en Annexe D.

décodeurs seuls comme Gemma ou LFM2. Notamment, sur la moyenne des 16 mois, l'ajustement de M2M100_418M conduit à un score BLEU de 78,3 et à un score chrF de 90,5, tandis que celui de M2M100_1.2B permet d'arriver à un score BLEU de 79,2 (+0,9 points) et chrF de 90,9 (+0,4 points). Ces améliorations sont, cependant, relativement coûteuses en terme de puissance de calcul. Environ 800 millions de paramètres supplémentaires sont effet nécessaires pour obtenir ces gains mesurés par les métriques automatiques. À noter que le modèle M2M100_1.2B obtient le plus faible taux d'erreur au niveau des caractères, comme indiqué par la métrique chrF, ce qui a un impact direct sur l'effort de post-édition humaine nécessaire pour obtenir une normalisation correcte de la source. Finalement, ces résultats motivent non seulement l'utilisation de modèles pré-entraînés, mais aussi de modèles larges de type encodeur–décodeur avec la TA comme objectif de pré-entraînement.

5.2 Analyse

Entitée nommée			**Genre**		
référence	*sortie du modèle*	*fréquence*	*référence*	*sortie du modèle*	*fréquence*
Saint-Victor	Saint-Vicoltor	6	lesdites	lesdits	6
Bonna	Bonne	10	lesquels	lesquelles	9
Curteti	Curtet	10	fait	faite	12
Céligny	Céligy	13	'ils	'elles	13
Peney	Peneys	42	la	le	24
Nombre			**Casse**		
référence	*sortie du modèle*	*fréquence*	*référence*	*sortie du modèle*	*fréquence*
détenu	détenus	13	Résolu	résolu	11
au	aux	15	Plus	plus	12
aux	au	15	Toutefois	toutefois	29
ils	il	23	il	Il	40
seigneurs	seigneur	34	Il	il	46
il	ils	65	ville	Ville	58
soit	soient	91	et	Et	193
			Et	et	244
Préposition			**Divers**		
référence	*sortie du modèle*	*fréquence*	*référence*	*sortie du modèle*	*fréquence*
par	pour	26	seigneur	seignur	17
pour	par	26	'Hôpital	'Hhôpital	23
du	de	33	lui	le	26
à	de	47	'on	'ont	37
dans	en	78	il	'il	41
			que	qu	53
			-	—	192

TABLE 4 – Exemples d'erreurs extraites de la sortie de M2M100_1.2B, selon la métrique chrF, sur toute la période temporelle couverte, avec leur fréquence et la référence correspondante.

Afin d'analyser les erreurs commises par le meilleur modèle évalué dans notre étude, nous extrayons les mots erronés présents dans les sorties de M2M_1.2B. Nous catégorisons les erreurs les plus

fréquentes et les présentons dans le Tableau 4. Un grand nombre d'erreurs commises par le meilleur modèle sont liées à la ponctuation, par exemple, le modèle génère « — » au lieu de « - » dans 192 occurrences, ou à la casse, les plus fréquentes concernant la génération de « et » au lieu de « Et », et vice-versa, avec plus de 400 occurrences au total. Ce type d'erreurs est relativement trivial à corriger avec un ensemble de règles applicables *a posteriori*, une méthode employée dans les travaux de Bawden *et al.* (2022).

Le modèle normalise aussi des mots qui ne doivent pas être changés selon nos règles d'édition, par exemple la préposition « dans » en « en » :

> *source* : fire les recommandations a cieulx compryns dans leur charge
> *référence* : firent les recommandations à ceux compris dans leur charge
> *modèle* : **faire** les recommandations à ceux compris **en** leur charge

Ou encore « de » en « à » dans l'exemple suivant :

> *source* : lequelt par plussieurs foys a perseverer de joyer
> *référence* : Lequel, plusieurs fois, a persévéré à jouer
> *modèle* : Lequel, plusieurs fois, a persévéré **de** jouer

Le pré-entraînement du modèle sur du français contemporain influence ici les choix de prépositions en contexte. L'ajustement de ce modèle sur plus d'exemples en moyen français permettrait de conserver les prépositions utilisées en source tout en normalisant les variantes orthographiques.

D'autres catégories d'erreurs nécessiteraient cependant l'ajout de ressources spécifiques afin de les corriger. C'est le cas pour les entités nommées, pour lesquelles des bases de connaissances et glossaires apporteraient les informations, comme montré dans des travaux précédents (Fischer & Volk, 2025). Par exemple, pour le patronyme « Bonna », le système génère « Bonne », ou encore pour « Curteti », « Curtet ». Le mois de mars 1545 contient un grand nombre d'erreurs liées aux entités nommées, par exemple :

> *source* : claudaz de cres detenuz ; guiellaume beney
> *référence* : Clauda Decrest, détenue ; Guillaume Benoît
> *modèle* : Clauda **Decre**, détenue ; **Gueillot** Benoît

Ou encore :

> *source* : urban besson loys dunant lallamandaz la curtaz et aultres detenuz
> *référence* : Urbain Besson, Louis Dunant, l'Allemande, la Curta et autres détenus
> *modèle* : **Urban** Besson, Louis Dunant, l'**Allemanda**, **La** Curta et autres détenus

Ce mois est particulièrement difficile à normaliser, si on se fonde sur les résultats obtenus par les modèles Transformer non pré-entraîné : 74,2 pour Transformer_2x2 et 76,4 pour Transformer_3x3. Les scores pour ces deux modèles sont les plus bas pour les 16 mois évalués. Cependant, les modèles pré-entraînés permettent d'améliorer les résultats de ce mois en particulier. C'est en revanche le mois de septembre qui leur pose le plus de problèmes. Celui-ci, dans sa version normalisée, est le plus distant de la source, selon la fonction identité mesurée avec chrF et reprise dans le Tableau 2.

6 Conclusion

Cet article présente une étude sur la normalisation de textes historiques écrits au XVIe siècle en moyen français. Nos contributions sont doubles. D'une part, l'étude porte sur un nouveau corpus

parallèle de haute qualité composé de textes en moyen français en source et leur version normalisée en cible. Ce corpus unique de textes administratifs sera mis à disposition de la communauté. D'autre part, nous avons comparé différents modèles pour la normalisation, avec et sans pré-entraînement et de type encodeur–décodeur et décodeur pour les modèles pré-entraînés. Les résultats mesurés à l'aide de métriques automatiques montrent que les modèles de type encodeur–décodeur pré-entraînés surpassent tous les autres modèles évalués dans notre étude. Les modèles non pré-entraînés atteignent les performances de normalisation les plus faibles. Ces résultats mettent en évidence l'utilité de l'ajustement de modèles pré-entraînés, même pour des langues sous-représentées comme le moyen français du XVIe siècle dans un domaine de spécialité. Ils montrent aussi que les objectifs de pré-entraînement ont un impact sur les performances de normalisation, pour des modèles ayant la même architecture et une taille comparable en termes de nombre de paramètres.

L'analyse des erreurs commises par le meilleur modèle indique que des connaissances externes, comme des bases de connaissances, et des données d'entraînement supplémentaires seront utiles pour améliorer la normalisation des entités nommées. De plus, accroître la quantité de données d'ajustement pourrait permettre la conservation de certaines caractéristiques de la source étant actuellement normalisées de manière erronée, tout en réduisant les variantes orthographiques, ce qui est un des objectifs de la normalisation de textes anciens. Ces éléments d'analyse vont guider nos travaux futurs, notamment pour la construction et l'utilisation de ressources relatives aux toponymes, patronymes et autres entités nommées, comme des glossaires et lexiques.

Remerciements

Nous tenons à remercier les relecteurs anonymes pour leurs commentaires, suggestions et remarques pertinentes. Ce travail a été partiellement financé par le FNS (Fonds National Suisse), subvention SSH 2022 n° 215733, pour le projet intitulé « Une édition sémantique et multilingue en ligne des registres du Conseil de Genève (1545-1550) » (acronyme RCnum). Toutes les expériences ont été réalisées sur les clusters de calcul de l'Université de Genève.

Références

BARON A., RAYSON P. & ARCHER D. (2009). Automatic Standardization of Spelling for Historical Text Mining. *Proceedings of Digital Humanities*.

BAWDEN R., POINHOS J., KOGKITSIDOU E., GAMBETTE P., SAGOT B. & GABAY S. (2022). Automatic Normalisation of Early Modern French. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 3354–3366, Marseille, France : European Language Resources Association.

BOLLMANN M. (2013). POS Tagging for Historical Texts with Sparse Training Data. In A. PAREJA-LORA, M. LIAKATA & S. DIPPER, Éds., *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, p. 11–18, Sofia, Bulgaria : Association for Computational Linguistics.

BOLLMANN M. (2018). *Normalization of Historical Texts with Neural Network Models*. Thèse de doctorat, Dissertation, Bochum, Ruhr-Universität Bochum.

BOLLMANN M., PETRAN F. & DIPPER S. (2011). Applying Rule-based Normalization to Different Types of Historical Texts – An Evaluation. In *Language and Technology Conference*, p. 166–177 : Springer. DOI : 10.1007/978-3-319-08958-4_14.

BOLLMANN M. & SØGAARD A. (2016). Improving Historical Spelling Normalization with Bi-directional LSTMs and Multi-task Learning. In Y. MATSUMOTO & R. PRASAD, Éds., *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, p. 131–139, Osaka, Japan : The COLING 2016 Organizing Committee.

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, UK : European Association for Machine Translation (EAMT).

FAN A., BHOSALE S., SCHWENK H., MA Z., EL-KISHKY A., GOYAL S., BAINES M., CELEBI O., WENZEK G., CHAUDHARY V., GOYAL N., BIRCH T., LIPTCHINSKY V., EDUNOV S., GRAVE E., AULI M. & JOULIN A. (2021). Beyond English-centric Multilingual Machine Translation. *Journal of Machine Learning Research*, **22**(107), 1–48. DOI : 10.5555/3546258.3546365.

FISCHER D. P. & VOLK M. (2025). Name Consistency in LLM-based Machine Translation of Historical Texts. In P. BOUILLON, J. GERLACH, S. GIRLETTI, L. VOLKART, R. RUBINO, R. SENNRICH, A. C. FARINHA, M. GAIDO, J. DAEMS, D. KENNY, H. MONIZ & S. SZOC, Éds., *Proceedings of Machine Translation Summit XX : Volume 1*, p. 204–219, Geneva, Switzerland : European Association for Machine Translation.

GAO Y., HEROLD C., YANG Z. & NEY H. (2022). Is Encoder-Decoder Redundant for Neural Machine Translation ? In Y. HE, H. JI, S. LI, Y. LIU & C.-H. CHANG, Éds., *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 562–574, Online only : Association for Computational Linguistics. DOI : 10.18653/v1/2022.aacl-main.43.

GEMINI TEAM, ANIL R., BORGEAUD S., ALAYRAC J.-B., YU J., SORICUT R., SCHALKWYK J., DAI A. M., HAUTH A., MILLICAN K. *et al.* (2023). Gemini : a family of highly capable multimodal models. *arXiv preprint arXiv :2312.11805*.

HOCHREITER S. & SCHMIDHUBER J. (1997). Long Short-term Memory. *Neural Computation*, **9**(8), 1735–1780. DOI : 10.1162/neco.1997.9.8.1735.

JUNCZYS-DOWMUNT M., GRUNDKIEWICZ R., DWOJAK T., HOANG H., HEAFIELD K., NECKERMANN T., SEIDE F., GERMANN U., FIKRI AJI A., BOGOYCHEV N., MARTINS A. F. T. & BIRCH A. (2018). Marian : Fast Neural Machine Translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, p. 116–121, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-4020.

KORCHAGINA N. (2017). Normalizing Medieval German Texts : From Rules to Deep Learning. In G. BOUMA & Y. ADESAM, Éds., *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, p. 12–17, Gothenburg : Linköping University Electronic Press.

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, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-2012.

LOSHCHILOV I. & HUTTER F. (2019). Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

MARIE B., FUJITA A. & RUBINO R. (2021). Scientific Credibility of Machine Translation Research : A Meta-Evaluation of 769 Papers. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 7297–7306, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.566.

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, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073135.

PETTERSSON E., MEGYESI B. & NIVRE J. (2014). A Multilingual Evaluation of Three Spelling Normalisation Methods for Historical Text. In K. ZERVANOU, C. VERTAN, A. VAN DEN BOSCH & C. SPORLEDER, Éds., *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, p. 32–41, Gothenburg, Sweden : Association for Computational Linguistics. DOI : 10.3115/v1/W14-0605.

PETTERSSON E., MEGYESI B. & TIEDEMANN J. (2013). An SMT Approach to Automatic Annotation of Historical Text. In *Proceedings of the Workshop on Computational Historical Linguistics at NODALIDA 2013, NEALT Proceedings Series*, volume 18, p. 54–69.

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, Lisbon, Portugal : Association for Computational Linguistics. DOI : 10.18653/v1/W15-3049.

POST M. (2018). A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation : Research Papers*, p. 186–191, Belgium, Brussels : Association for Computational Linguistics. DOI : 10.18653/v1/W18-6319.

RIEZLER S. & MAXWELL J. T. (2005). On Some Pitfalls in Automatic Evaluation and Significance Testing for MT. In J. GOLDSTEIN, A. LAVIE, C.-Y. LIN & C. VOSS, Éds., *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, p. 57–64, Ann Arbor, Michigan : Association for Computational Linguistics.

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, Torino, Italia : 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 City, Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2024.findings-naacl.215.

RUBINO R., MARIE B., DABRE R., FUJITA A., UTIYAMA M. & SUMITA E. (2020). Extremely Low-resource Neural Machine Translation for Asian Languages. *Machine Translation*, **34**(4), 347–382. DOI : 10.1007/s10590-020-09258-6.

SÁNCHEZ-MARTÍNEZ F., MARTÍNEZ-SEMPERE I., IVARS-RIBES X. & CARRASCO R. C. (2013). An Open Diachronic Corpus of Historical Spanish. *Language Resources and Evaluation*, **47**(4), 1327–1342.

SMITH J. C. (2002). Middle French : When ? What ? Why ? *Language Sciences*, **24**(3), 423–445. DOI : 10.1016/S0388-0001(01)00042-0.

SOLFRINI S., DEJOUY M., MARQUES OLIVEIRA A. & BEAULNES P.-O. (2025). Normaliser le moyen français : du graphématique au semi-diplomatique. In F. BECHET, A.-G. CHIFU, K. PINEL-SAUVAGNAT, B. FAVRE, E. MAES & D. NURBAKOVA, Éds., *Actes des 18e Rencontres Jeunes Chercheurs en RI (RJCRI) et 27ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL)*, p. 239–252, Marseille, France : ATALA & ARIA.

TANG G., CAP F., PETTERSSON E. & NIVRE J. (2018). An Evaluation of Neural Machine Translation Models on Historical Spelling Normalization. In E. M. BENDER, L. DERCZYNSKI & P. ISABELLE, Éds., *Proceedings of the 27th International Conference on Computational Linguistics*, p. 1320–1331, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is All you Need. *Advances in neural information processing systems*, **30**.

VILKOMIR K. & HERNDON N. (2024). Challenges of Automatic Document Processing with Historical Data. In *Proceedings of the 2024 ACM Southeast Conference*, p. 50–59.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M., DAVISON J., SHLEIFER S., VON PLATEN P., MA C., JERNITE Y., PLU J., XU C., LE SCAO T., GUGGER S., DRAME M., LHOEST Q. & RUSH A. (2020). Transformers : State-of-the-art natural language processing. In Q. LIU & D. SCHLANGEN, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 38–45, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-demos.6.

XUE L., CONSTANT N., ROBERTS A., KALE M., AL-RFOU R., SIDDHANT A., BARUA A. & RAFFEL C. (2021). mT5 : A Massively Multilingual Pre-trained Text-to-Text Transformer. In K. TOUTANOVA, A. RUMSHISKY, L. ZETTLEMOYER, D. HAKKANI-TUR, I. BELTAGY, S. BETHARD, R. COTTERELL, T. CHAKRABORTY & Y. ZHOU, Éds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 483–498, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.41.

YANG A., LI A., YANG B., ZHANG B., HUI B., ZHENG B., YU B., GAO C., HUANG C., LV C. *et al.* (2025). Qwen3 Technical Report. *arXiv preprint arXiv :2505.09388*.

ZILIO L., LAZZARI R. R. & FINATTO M. J. B. (2024). Can Rules Still Beat Neural Networks ? The Case of Automatic Normalisation for 18th-century Portuguese Texts. In P. GAMALLO, D. CLARO, A. TEIXEIRA, L. REAL, M. GARCIA, H. G. OLIVEIRA & R. AMARO, Éds., *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 2*, p. 83–92, Santiago de Compostela, Galicia/Spain : Association for Computational Lingustics.

A Annexe : règles de transcription manuelle

Nous transcrivons le texte des registres du Conseil de Genève au plus proche de l'original, avec ses erreurs (sans les relever), et en respectant les espaces ou leur absence entre les mots : « de Berne » est fréquemment écrit d'un seul tenant et est donc transcrit « deberne ». Il n'y a ni ponctuation, ni accents, ni cédilles, ni apostrophes, ni majuscules. Nous reviendrons sur ces dernières ultérieurement. Il y a deux exceptions à la transcription au plus proche :

— La première concerne la distinction entre deux lettres, à la graphie aujourd'hui distincte, mais parfois identique au XVIe siècle. Dans le cas du « i/j », nous respectons l'usage actuel : « iniustice » est transcrit « injustice » et « subiect » est transcrit « subject ». De même pour le « u/v » : « trouue » est transcrit « trouve », « souuent » est transcrit « souvent » et « vne » est transcrit « une ».

— La seconde est la résolution des abréviations, selon l'usage le plus courant de chacun des scribes concernés. Ainsi « 9missre » est transcrit « commissa**y**re » pour le scribe A et « commissa**i**re » pour les scribes B et C, tandis que « se », « srie », « seige » et « seigrie » sont transcrits « seign**eu**rie » pour les scribes A et B, et « seign**o**rie » pour le scribe C.

Les abréviations numérales, en revanche, ne sont pas résolues, pour rester proche du texte original, mais elles sont peu nombreuses et facilement compréhensibles : « 4or » (quatuor) et « 200um » (ducentenorum).

Nous rencontrons, pour un même scribe, différentes façons d'écrire une même lettre. Souvent, la graphie d'une lettre dépend de sa place dans un mot (au début, au milieu ou à la fin) ainsi que des lettres qui l'entourent. Le « e » de « de » (final) est différent du « e » de « les » (au milieu). Cependant, même à une place identique, la graphie peut varier. Dans le cas du « u/v », le « u » se rencontre généralement à l'intérieur d'un mot, comme dans « seruiteur », « caluin » et « priuaulte », tandis que le « v » se trouve le plus souvent en début de mot, comme dans « vigne », « venu », « vescu » et « vne » (une). Mais ce n'est pas toujours le cas. On rencontre aussi « advys », « reveuhe » et « privaulte ».

Au sujet des majuscules, si certaines lettres ressemblent à nos majuscules actuelles, nous doutons qu'elles aient été utilisées comme telles, raison pour laquelle nous n'en tenons pas compte et nous les transcrivons comme des minuscules. Selon nous, elles ne sont qu'une variante d'une même lettre, le plus souvent employée en début de mot, comme le « R » : « Requerant », « Riere », « Recongnoissance », etc., peu importe la place du mot dans la phrase.

Concernant le « i/j », le « i » ne se trouve que sous forme d'une minuscule, tandis que le « j » ne se trouve que sous forme d'une majuscule, aussi bien au début qu'au milieu d'un mot : « ilz », mais « Jlz », « iniustice », mais « inJustice », « subiect », mais « subJect ». Il y a toutefois une lettre qui ressemble à un « j » minuscule, mais qui, en réalité, est un « i » final allongé : « maij », « messerij », etc. Nos interventions, réduites au strict minimum, sont signalées au moyen de parenthèses carrées.

B Annexe : opérations de normalisation manuelle

Les directives de normalisation ont été définies par l'expert historien et paléographe chargé de transcrire manuellement les registres manuscrits et de normaliser le contenu textuel. L'objectif de la normalisation est d'améliorer la compréhension du texte, sans trop bouleverser sa structure, tout en conservant le vocabulaire ancien et en réduisant les variantes orthographiques. Les modifications

apportées concernent l'orthographe, la conjugaison et dans une certaine mesure la syntaxe, sauf lorsque la structure de la phrase aurait entraîné un remaniement trop important. Dans ces cas, la structure a été conservée au détriment de la syntaxe. Les toponymes, patronymes et prénoms ont été modernisés, mais les spécificités des prénoms ont été maintenues, par exemple : *Thyvent* devient *Thivent* et non *Étienne*.

Ci-dessous, la liste des diverses modification opérées lors de la normalisation manuelle du contenu original des manuscrits : Les emplois :

- emploi des majuscules (débuts de phrase + patronymes + toponymes)
- emploi des accents
- emploi de la ponctuation
- emploi des apostrophes
- emploi des cédilles

Les corrections :

- correction de l'orthographe
- correction des genres : « la dimanche » devient « le dimanche » ; « la reste » devient « le reste »
- correction des déterminants possessifs singuliers féminins « ma », « ta », « sa » des noms commençant par une voyelle ou par un « h » muet par les formes « mon », « ton », « son » : « à sa humble requête » devient « à son humble requête »
- correction des singuliers/pluriels
- correction des prépositions incorrectes : « en l'hôpital » devient « à l'hôpital » ; « en Genève » devient « à Genève »
- correction de la conjugaison : temps, auxiliaires (« a été », « ont été » devient « est allé », « sont allés »), genre et nombre des accords verbaux (« de celui qui les a baillé » devient « de celui qui les a baillé(e)s »)
- correction de « qui » en « qu'il(s) » ou en « que » et, inversement, de « que » en « qui »
- correction de « il(s) » en « y »
- correction de « si » en « s'il »
- correction des COD/COI (lui/le)

Les suppressions :

- suppression des « que » superflus (ordonné que avant de venir que il fasse)
- suppression des conjonctions et prépositions inutiles (et + ou + pour + en)
- suppression de « par » devant « par plusieurs fois »
- suppression de « avec » précédant les jours de la semaine : « qu'il soit ouï avec mercredi prochain »
- suppression de « dernier » avec « passé » : dimanche dernier passé
- suppression de « un » avant « chacun » : un chacun
- suppression des doubles négations : défendre de ne pas

Les ajouts :

- ajout des « que » introduisant un verbe
- ajout des articles manquants (le, la, un, ce (avant que))
- ajout des pronoms manquants (il, en)
- ajout des prépositions manquantes (notamment « de » + « à » (infinitif + lieux)
- ajout des « y » manquants (il y a)
- ajout de « après » (avant un infinitif, par exemple : « avoir ouï » devient « après avoir ouï »)
- ajout de « à ce que » après « jusque »
- ajout des pronoms réfléchis

— ajout du « pas » dans les phrases négatives (ne. . .pas)
— ajout de mots manquants entre parenthèses carrées []

Les remplacements :

— remplacement de « ordonné » par « il a été ordonné »
— remplacement, dans les phrases négatives, de « et aussi » par « ni »
— remplacement de « non + verbe» par « ne pas + verbe » (« non faire » devient « ne pas faire »)
— remplacement de l'infinitif par sa forme conjuguée précédée de « que » (« il a requis divorce être fait » devient « il a requis que le divorce soit fait » ; « il se conste "avoir"/"être" » devient « il se conste "qu'il/qu'elle a"/"qu'il/qu'elle est" »)
— remplacement de « par » par « pour » et vice-versa
— remplacement de « comment » par « comme » et vice-versa
— remplacement de « n'est, n'ont » en « en est, en ont »
— remplacement de « de quoi » par « dont »
— remplacement de « un sien/une sienne » par « son/sa » ou par « un/une de ses » (« une sienne femme » devient « sa femme » ; « une sienne servante » devient « une de ses servantes »)

Les inversions :

— inversion des pronoms et des auxiliaires (« se doive enquérir » devient « doive s'enquérir », « le doive suivre » devient « doive le suivre »)
— inversion des noms et des adjectifs

Pour terminer, les sommes d'argent sont résolues, par exemple : « huit vingt et dix écus » devient « cent septante écus ».

C Annexe : Configurations et hyper-paramètres des modèles

C.1 Modèles Transformer non pré-entraînés

Les deux architectures de modèles Transformer (Vaswani *et al.*, 2017) non pré-entraînés sont de type encodeur-décodeur et reposent sur l'implémentation de Marian (Junczys-Dowmunt *et al.*, 2018) avec soit 2 couches d'encodeur et 2 couches de décodeur (architecture 2x2), soit 3 couches (architecture 3x3). Pour ces modèles, les plongements lexicaux ont 512 dimensions, tout comme la dimensionnalité du modèle, avec des couches à action directe (*feed-forward layers*) ayant 2 048 dimensions. Les couches d'attention de l'encodeur et du décodeur ont 8 têtes. Le vocabulaire est partagé entre l'encodeur et le décodeur, et donc les plongements lexicaux, et le vocabulaire est limité à 12 000 sous-mots selon la méthode *SentencePiece* (Kudo & Richardson, 2018).

L'entraînement est effectué en utilisant l'algorithme d'optimisation AdamW (Loshchilov & Hutter, 2019) et la fonction de coût d'entropie croisée. Un échauffement du taux d'apprentissage est effectué au début de l'entraînement de manière linéaire pendant 4 000 itérations, pour atteindre un maximum de 0,003, avant de diminuer linéairement jusqu'à 50 000 itérations. Le taille de batch est fixée à 8 paires d'échantillons, ce qui permet d'entraîner ce type de modèles sur des GPUs avec 12Go de vRAM. Un taux de *dropout* est fixé à 0,1 pendant toute la durée de l'entraînement.

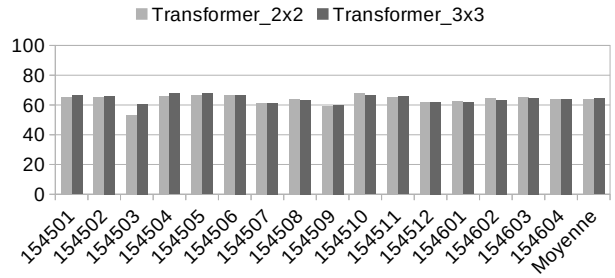
C.2 Modèles de type encodeur–décodeur pré-entraînés

Les modèles de type encodeur–décodeur pré-entraînés incluent M2M100_418M, M2M100_1.2B et mT5_580M. Ils sont ajustés à l'aide de nos données pendant 20 000 itérations avec une taille de batch fixée à 4 paires d'échantillons. Le taux d'entraînement choisi est sélectionné selon les performances atteintes sur le corpus d'entraînement (entropie croisée). Les valeurs testées sont entre 0,0001 et 0,0005 avec les modèles de la famille M2M100, et entre 0,00009 et 0,0004 pour mT5, avec un échauffement linéaire de 100 itérations, suivi d'une décroissance linéaire jusqu'à la fin de l'ajustement pour tous les modèles. L'algorithme d'optimisation employé est AdamW (Loshchilov & Hutter, 2019). Tous les autres hyper-paramètres sont laissés à leurs valeurs spécifiées dans le fichier de configuration associé à chaque modèle distribué par Hugging Face (Wolf *et al.*, 2020).

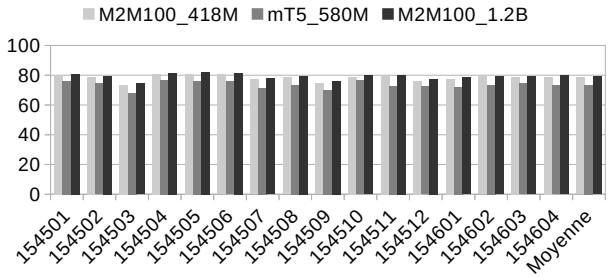
C.3 Décodeurs pré-entraînés

Les décodeurs pré-entraînés incluent Qwen_600M, Gemma_270M et Gemma_1B, LFM2_350M, LFM2_700M et LFM2_1.2B. Ils sont ajustés pendant 10 000 itérations avec un batch de 4 paires d'échantillons. Nos expériences préliminaires ont montré une stagnation de la fonction de coût sur les données d'entraînement au delà de ce nombre d'itérations. Le taux d'apprentissage est choisi selon les performances pendant l'entraînement (entropie croisée) avec des valeurs allant de 0,00007 à 0,0001. Les autres hyper-paramètres sont par défaut ceux spécifiés dans la configuration de chaque modèle distribué par Hugging Face (Wolf *et al.*, 2020).

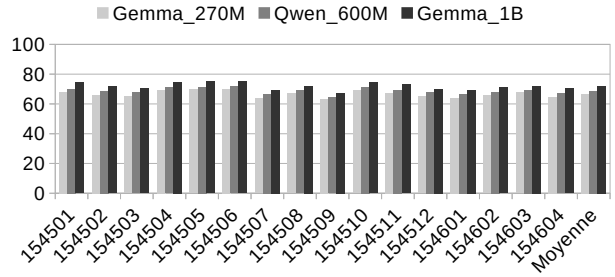
D Annexe : résultats expérimentaux mesurés par BLEU



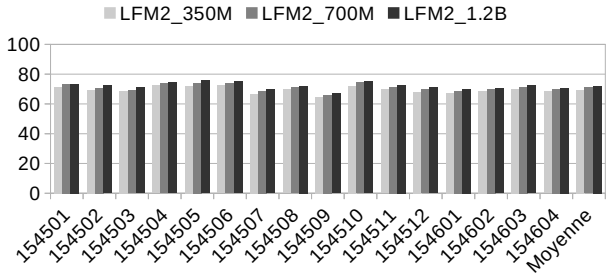
(a) Résultats obtenus par les modèles de type Transformer encodeur–décodeur non pré-entraînés.



(b) Résultats obtenus par les modèles de type encodeur–décodeurs pré-entraînés.



(c) Résultats obtenus par les décodeurs seuls pré-entraînés de la famille *Gemma* et *Qwen*.



(d) Résultats obtenus par les décodeurs seuls pré-entraînés de la famille *LFM2*.

FIGURE 3 – Résultats par mois et moyennes obtenues par les modèles évalués dans notre étude, mesurés par la métrique BLEU.