Proceedings of CONF-MLA 2025 Symposium: Applied Artificial Intelligence Research
DOI: 10.54254/2755-2721/2025.BJ26308

A Transformer-based Approach of Normalization of Historical French Text

Sicong Deng

College of Foreign Languages and Cultures, Sichuan University, Chengdu, China
alphaerdmkll@gmail.com

Abstract. Historical French texts pose unique challenges for modern readers and NLP systems due to highly inconsistent spelling and scarce standardized resources. This paper presents a system for the normalization of early modern French, focusing on enhanced corpus construction and customized sub-word tokenization. Existing aligned corpora were combined with a newly curated dataset based on La Gazette, a rich and stylistically coherent periodical, to provide greater coverage and diversity. To address the orthographic variation typical of the period, a custom sub-word tokenizer was trained to better capture morphological patterns, supporting a Transformer-based sequence-to-sequence model. The approach demonstrates how tailored data preprocessing and tokenization improve the accuracy and robustness of automatic normalization. This work contributes valuable resources and methods for processing historical French and lays the groundwork for broader applications in digital humanities and historical linguistics.

Keywords: Transformer, Corpus, Normalization, Historical French, Sub-word Tokenization

1. Introduction

Until the 18th century, the spelling and grammar of French texts varied widely from today's standardized language. Without fixed rules, writers used personal or phonetic spellings that could differ greatly between regions, authors, or even within the same work. These differences were shaped by local dialects, medieval printing traditions, and the gradual development of official dictionaries such as the Dictionnaire de l'Académie française published in 1694. As a result, early modern French texts pose challenges for modern readers and digital tools that process language.

Textual normalization aims to address this problem. In historical linguistics, normalization means converting archaic or inconsistent spellings and forms into a clear, standard version of the language, while keeping the meaning and structure intact. This is more than just fixing spelling mistakes — it also involves dealing with old word forms, outdated grammar, and variations in punctuation and accents. Normalization not only helps general readers understand old texts but also makes these texts easier to search, index, and analyze using modern natural language processing (NLP) methods. For many digital humanities projects, normalization is a key step that enables further research and the development of useful tools like lemmatizers, parsers, or statistical models.

Earlier approaches to normalization often relied on manually created rules and historical dictionaries. However, recent advances in machine learning make it possible to learn how to

DOI: 10.54254/2755-2721/2025.BJ26308

normalize automatically, by training models on large collections of matching historical and modern texts — similar to how translation systems work. This data-driven approach helps capture variations in spelling and style without needing extensive manual corrections.

Despite this progress, normalizing 17th-century French still faces major challenges. While Latin high-quality parallel corpora (original and normalized versions) are easily accessible, their French counterparts remain rare and expensive to create [1]. Many old texts have only been modernized partially or inconsistently. Another challenge is tokenization: standard byte-level tokenizers work well for large, messy datasets, but they can produce inefficient, overly long sequences for smaller, specialized historical corpora. For this reason, training a custom subword tokenizer can better capture the unique spellings in early modern French while improving model efficiency.

In response to the challenges, this research integrates and extends existing historical French corpora. It proposes a Transformer-based approach, combined with a custom-trained sub-word tokenizer, to improve the normalization of early modern French texts.

2. Literature review

2.1. Existing corpora and resources

The field of machine translation has undergone a major transformation by the Transformer architecture [2]. Transformer architecture which, via self-attention mechanisms, enables greater parallelism and improved long-sequence performance, has become the foundation for systems like BERT, GPT, and MarianMT [3-4]. In parallel, the field of linguistic normalization has gone through the evolution of NLP from traditional rule-based approaches to data-driven methods. However, historical texts—including those in early modern French—remain limited in availability. The FreEM corpus, which focuses on 17th- and 18th-century texts, helps narrow this gap by supporting various NLP normalization tasks such as named entity recognition, morphological tagging, lemmatization, and orthographic normalization [5], thereby facilitating more effective processing of historical language data.

2.2. State of art on OCR

Access to historical texts relies on extraction from digitized sources of scanned images. For such tasks Optical Character Recognition (OCR) offers mature and effective solutions for modern texts. Traditionally, historical OCR has received less attention than its modern counterpart due to the complexity of early print and manuscript materials. But increasing amount of attention has been drawn to the specialized field in recent years. This multidisciplinary effort is enabling the development of more robust tools for treating historical documents [6]. Among options, two open-source OCR frameworks are available and utilized in this project: Kraken and Calamari, both of which are trained for historical prints and manuscripts, supporting archaic glyphs (e.g., long s <ſ>) and noisy documents, with Calamari offering ensemble predictions for higher accuracy and Kraken enabling custom training for domain-specific needs [7,8].

The OCR models employed in this study were pre-trained on historical French. The output of the OCR process was subsequently subjected to error analysis and correction before further linguistic preprocessing and modeling.

DOI: 10.54254/2755-2721/2025.BJ26308

3. Methodology

3.1. Standardized processing workflow

The complete pipeline for normalization consists of image acquisition, OCR, preprocessing, tokenization, model training, and evaluation. High-resolution scan images are fed into OCR to create parallel datasets. The resulting transcriptions undergo both automatic cleaning and manual correction on a selected subset to ensure linguistic accuracy. The cleaned corpus is then tokenized using a custom-trained sub-word tokenizer. A Transformer-based sequence-to-sequence model is then trained on the tokenized pairs. Finally, performance is evaluated using Bilingual Evaluation Understudy (BLEU) scores and manual inspection on a held-out validation set.

3.2. Corpus construction

A reliable parallel corpus is essential for automatic normalization. In this study, it combined multiple sources to build a more diverse and representative dataset. Existing resources, such as the FrenEMnorm corpus, a manually aligned collection of Early Modern French texts and their modern equivalents, provide high-quality sentence pairs from literary and administrative documents [9]. TheaterLFSV2, another key source, adds more sentence pairs from early modern French literature, further expanding linguistic diversity [10].

To strengthen this foundation, this research combined texts from La Gazette, a 17th-century periodical, chosen for its consistent editorial style and rich vocabulary. As one of France's earliest news publications, La Gazette includes military reports, diplomatic correspondence, and court announcements, offering thematically varied yet stylistically coherent content. High-quality public-source scans from the Bibliothèque nationale de France (BnF) enable effective text extraction [11]. Together, these sources form an enhanced corpus that balances diversity and consistency, supporting the development of robust, data-driven normalization models for historical French.

3.3. Data collection and selection

3.3.1. Automatic cleaning and filtering

Existing corpora often contain non-French text (e.g., Latin passages) or structural data that can harm model performance. Complex layouts and low-quality scans often lead to OCR errors, where visual noise was misread as important characters. Converting original files into binary images reduced some noise but did not fully solve accuracy issues.

To address this, the study developed a dedicated cleaning script. It applies typographic replacements to unify punctuation (e.g., curly quotes, varied dashes) with simple ASCII equivalents. A heuristic filtering stage removes lines likely to be repetitive metadata (like scene headings or speaker names) or lines with too little linguistic content, such as single words or lines dominated by non-Latin characters. This pipeline regularizes input, reduces fragmented tokens, and improves consistency across historical-modern pairs, resulting in a cleaner, more reliable dataset for training.

3.3.2. Manual correction

Alongside automatic scripts, manual review was carried out to further improve accuracy. Special attention was given to lines containing unusual characters, ambiguous abbreviations, or layout artifacts left by the original prints. This extra step ensured that the final parallel corpus maintained a

DOI: 10.54254/2755-2721/2025.BJ26308

high standard of alignment and linguistic integrity, providing a stable base for training and evaluation.

3.3.3. Customized tokenization

A custom sub-word tokenizer was trained to balance efficiency and linguistic sensitivity. Recent studies show that sub-word tokenization respecting morphological components improves semantic generalization compared to arbitrary splits [12]. This is especially relevant for early modern French, which often contains morphologically transparent yet orthographically irregular forms. By preserving meaningful sub-word units, the tokenizer better captures variation and handles out-of-vocabulary forms.

For implementation, the Byte-Pair Encoding (BPE) algorithm was used via the Hugging Face tokenizers library [13,14]. Although the tokenizer is technically monolingual, the coexistence of historical and modern orthographies increases lexical diversity, justifying a larger vocabulary. Empirical tests showed that 15,000 tokens (with four reserved special tokens: <pad>, <unk>, <s>, </s>) offered a good balance between efficiency and coverage. The final tokenizer was serialized and saved in JSON format for reproducible use.

3.4. Model

The normalization system implemented in this project is based on the Transformer architecture with some minor adaptations for improvements. The reliance on self-attention mechanisms allows it to model long-range dependencies both efficiently and in parallel.

The Transformer employs encoder-decoder as its core architecture. Given a source sequence $\mathrm{X} = (\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{n})$ in historical French, the model produces a corresponding output sequence $\mathrm{Y} = (\mathrm{y}_1, \ldots, \mathrm{y}_\mathrm{m})$ in normalized modern French. Both sequences are first embedded into a high-dimensional vector space:

$$E(x_i) = Embedding(x_i) \cdot \sqrt{d_{model}} \tag{1}$$

where $\mathrm{d}_{\mathrm{model}}$ denotes the embedding dimensionality. Positional encodings are added to preserve word order, defined as:

$$P(i, 2k) = sin(\frac{i}{10000^{2k/dmodel}}),\ P(i, 2k+1) = cos(\frac{i}{10000^{2k/dmodel}}) \tag{2}$$

The encoder comprises N layers, each with two submodules. First, multi-head self-attention computes queries Q, keys K, and values V by linear projections :

$$Q = XW^Q,\ K = XW^K,\ V = XW^V \tag{3}$$

Scaled dot-product attention is then computed as:

$$Attention(Q, K, V) = softmax(\frac{QK^\top}{\sqrt{d_k}})V \tag{4}$$

This is followed by a position-wise feedforward network:

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \tag{5}$$

DOI: 10.54254/2755-2721/2025.BJ26308

Each sublayer uses residual connections and layer normalization to stabilize training:

$$SublayerOutput(x) = LayerNorm(x + Dropout(Sublayer(x))) \tag{6}$$

The decoder is structurally similar but includes cross-attention over encoder outputs. A causal mask:

$$M \in \{0,1\}^{m \times m} \tag{7}$$

prevents access to future tokens during autoregressive generation. Finally, the decoder output is projected to the vocabulary via:

$$P(y_t|y_{<t}, X) = softmax(W_o h_t + b_o) \tag{8}$$

where h_t is the decoder's hidden state at time t.

Training uses cross-entropy loss over the target sequence, with optional label smoothing. It is worth pointing out that though the architecture introduces no innovation beyond the established Transformer formulation, the work done hitherto, i.e., domain-specific tokenizers, cleaned and aligned parallel data, and targeted preprocessing allows the model to learn effective mappings.

4. Experimental design

4.1. Setup

The normalization model is a sequence-to-sequence model following the standard encoder-decoder design. The architecture consists of:

·6 encoder and 6 decoder layers ($N = 6$)

·Model dimensionality of 512 ($d_{model} = 512$)

·Feed-forward layer size of 1024 ($d_{ff} = 1024$)

·8 attention heads ($h = 8$)

·Dropout rate of 0.1

The output layer is a standard linear generator with softmax activation.

The model is trained for 10 epochs using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 22. The loss function is the negative log-likelihood (NLLLoss), with padding tokens ignored during loss calculation. Training was conducted on a single GPU. Batches are prepared by tokenizing and padding both source and target. For the decoder, target sequences are shifted to provide teacher-forcing inputs and targets, and appropriate masks are generated for both source and target to prevent information leakage.

4.2. Validation metrics

To assess the quality and reliability of the model, a rigorous evaluation framework is employed, combining both automated metrics and targeted qualitative review.

Edit Distance Metrics: Further evaluation is done using Levenshtein distance which measures the minimum number of single-character edits required to transform the prediction into the reference [15]. Two related metrics are reported:

·Average Edit Distance, computed as the mean Levenshtein distance per token across the test set.

DOI: 10.54254/2755-2721/2025.BJ26308

·Character Error Rate (CER), which normalizes the total edit distance by the total number of reference characters.

BLEU: The primary metric used is the BLEU score, which originally developed for evaluating machine translation results and measures the overlap of n-grams between the model's output and one or more reference translations. Formally, the score is computed as follows:

Formally, the BLEU score is computed as follows:

$$BLEU = BP \bullet exp(\sum_{n=1}^{N} w_n log\, p_n) \tag{9}$$

where p_n is the modified precision of n-grams, w_n is the weight assigned to each n-gram (typically uniform), and BP is the brevity penalty to penalize overly short outputs. In this project, BLEU is calculated at the sentence level using n-grams up to order four, with smoothing applied to account for data sparsity.

4.3. Results

The normalization system was evaluated using several metrics: corpus BLEU, word-level accuracy, character-level accuracy, character error rate (CER), and average edit distance. On the test set, the system achieved a BLEU score of 58.13, CER of 6.90%, and an average edit distance of 3.98 per sample. The following example presents a case of successful normalization, in which the model accurately restores both orthographic and diacritic forms:

SRC: Ces termes ſont-ils naiz d'vne fille des champs?

REF: Ces termes sont-ils naiz d'une fille des champs?

GEN: Ces termes sont - ils naiz d'une fille des champs?

5. Discussion

5.1. Summary

Despite achieving a moderate edit distance of 3.98, which indicates that several edits are required, the model's corpus BLEU score of 58.13 is decent, indicating that the overall performance is of good quality. CER of 6.90% suggests that many predictions contain minor errors—such as missing diacritics or slight orthographic mismatches—even when the general structure and content are preserved.

Qualitative examples reveal that the model successfully handles frequent and regular orthographic changes, such as substituting archaic glyphs, but often fails to recover diacritics, leading to errors at both the word and character levels.

5.2. Limitations

Several aspects of the current system offer room for future improvement. Although the model is a faithful implementation of the original Transformer architecture, future iterations may benefit from newer architectures or training strategies developed since then, particularly those optimized for low-resource or stylistically variable inputs. Additionally, while the evaluation relies on standard automatic metrics such as BLEU, the cultural and interpretive dimensions in this task may be better validated through expert linguistic review or task-oriented benchmarks.

DOI: 10.54254/2755-2721/2025.BJ26308

6. Conclusion

This study investigated the effectiveness of a Transformer-based neural model for the normalization of historical French texts, utilizing a large parallel corpus and a custom sub-word tokenizer. Evaluation was conducted using multiple metrics, including BLEU and average edit distance, to provide a comprehensive assessment of model performance.

The results indicate that the model effectively captures good orthographic patterns and systematic glyph substitutions, yielding moderate performance on BLEU and character-level metrics. However, word-level accuracy remains suboptimal, suggesting that fully accurate normalization is still challenging, particularly in the presence of diacritics and irregular forms.

These findings reveal the strengths and current limitations of neural models for historical text normalization. Future work should consider expanding the diversity of training data, incorporating explicit linguistic features, and developing post-processing strategies for fine-grained error correction. The integration of external lexical resources, as well as the exploration of semi-supervised or multi-task learning approaches, may further enhance the model's ability to generalize to rare or ambiguous forms.

References

[1] Reul, C., Wick, C., Nöth, M., Büttner, A., Wehner, M., & Springmann, U. (2021). Mixed model OCR training on historical Latin script for out-of-the-box recognition and finetuning. In: Proceedings of the 6th International Workshop on Historical Document Imaging and Processing. pp. 7-12.
[2] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. Advances in neural information processing systems, 30: 5998-6008.
[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In: 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). pp. 4171-4186.
[4] Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., et al. (2018). Marian: Fast neural machine translation in C++. arXiv preprint arXiv: 1804.00344.
[5] Gabay, S., Gambette, P. (2022) FreEM-corpora/FreEMnorm: FreEM norm Parallel corpus (1.0.1). Zenodo, Geneva.
[6] Eyharabide, V., Likforman-Sulem, L., Orlandi, L. M., et al. (2023). Study of historical Byzantine seal images: the BHAI project for computer-based sigillography. In: Proceedings of the 7th International Workshop on Historical Document Imaging and Processing. pp. 49-54.
[7] Wick, C., Reul, C., Puppe, F. (2020) Calamari - A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition. Digital Humanities Quarterly 14(1).
[8] Kraken OCR GitHub repository. (2024) https: //github.com/mittagessen/kraken.
[9] Gabay, S., Gambette, P. (2022) FreEM-corpora/FreEMnorm: FreEM Norm Parallel (Original vs. Normalised) Corpus for Early Modern French. Zenodo, Geneva. https: //doi.org/10.5281/zenodo.6481179
[10] Gabay, S., Clérice, T. (2024) DEFI-COLaF – Theatre-17e (Version 1.0). Geneva University; INRIA Paris.
[11] Gallica Digital Library. (2024) https: //gallica.bnf.fr/ark: /12148/cb32780022t/date& rk=21459; 2.
[12] Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. arXiv preprint arXiv: 1508.07909.
[13] Batsuren, K., Vylomova, E., Dankers, V., et al. (2024) Evaluating subword tokenization: Alien subword composition and oov generalization challenge. arXiv preprint arXiv: 2404.13292.
[14] Hugging Face Tokenizers GitHub repository.(2024). https: //github.com/huggingface/tokenizers.
[15] Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In Soviet Physics Doklady. New York. pp. 707–710.