Automatic Normalisation of Middle French and its Impact on Productivity

Rubino, Raphaël; Coram-Mekkey, Sandra; Gerlach, Johanna; Mutal, Jonathan David; Bouillon, Pierrette

Automatic Normalisation of Middle French and its Impact on Productivity

Raphael Rubino*, **Sandra Coram-Mekkey**†, **Johanna Gerlach***, **Jonathan Mutal***, **Pierrette Bouillon***
*TIM/FTI, University of Geneva, 1205 Geneva, Switzerland
{firstName.lastName}@unige.ch
†Fondation de l'Encyclopédie de Genève, Switzerland
coram.mekkey@gmail.com

Abstract

This paper presents a study on automatic normalisation of 16th century documents written in Middle French. These documents present a large variety of wordforms which require spelling normalisation to facilitate downstream linguistic and historical studies. We frame the normalisation process as a machine translation task starting with a strong baseline leveraging a pre-trained encoder–decoder model. We propose to improve this baseline by combining synthetic data generation methods and producing artificial training data, thus tackling the lack of parallel corpora relevant to our task. The evaluation of our approach is twofold, in addition to automatic metrics relying on gold references, we evaluate our models through post-editing of their outputs. This evaluation method directly measures the productivity gain brought by our models to experts conducting the normalisation task manually. Results show a $20+$ token per minute increase in productivity when using automatic normalisation compared to normalising text from scratch. The manually post-edited dataset resulting from our study is the first parallel corpus of normalised 16th century Middle French to be publicly released, along with the synthetic data and the automatic normalisation models used and trained in the presented work.

Keywords: Intralingual diachronic translation, Middle French, Archive, Normalisation, Productivity

1. Introduction

In Switzerland, each canton safeguards administrative, legal and financial documents produced by its successive governments. These large archives contain the oldest publicly released documents about the institutional history of Switzerland. A specific subset of these archives is the focus of our study, namely the Geneva Council Registers (in French: *les Registres du Conseil de Genève*), containing minutes of the council meetings covering local administrative and political decisions. These handwritten registers were held daily and are still being published nowadays as digital documents. They are an invaluable resource for studying the political, legal, economic, social, and religious history of the Geneva canton. More particularly during the 16th century, the Swiss Protestant Reformation took place. During this time, John Calvin played a major role in the Reformation and is considered today as one of the founders of Calvinism, a major branch of Protestantism (Backus and Benedict, 2011). Thus, Geneva Council Registers produced during this time period are interesting for historians, as the local political and religious decisions influenced the Geneva region and other Swiss cantons.

Nowadays, some of the original 16th century Geneva Council Registers (noted RCs hereafter) are available as digitised documents including a few with OCR. The registers produced between 1536 and 1544 entitled *Registres du Conseil de Genève à l'époque de Calvin* (Geneva Council Registers in Calvin's time) are available as hard copies. These archival documents were written in Middle French, a variant of the French language used mostly from the 14th to the 16th century (Buchi et al., 2019).[1] Furthermore, the textual content of these documents is sometimes mixed with Latin. These characteristics make RCs difficult to understand for non experts.

The current effort conducted by historians and palaeographers consists in editing the RCs textual content, mainly focusing on orthographic normalisation of various Middle French wordforms. This variety in spelling is due to the lack of language norms for Middle French during the 16th century, even for patronyms and toponyms. The resulting normalised textual content should follow editorial choices in terms of spelling normalisation and local grammatical modifications but should not contain syntactic alterations. One of the main motivations in normalising Middle French is to make RCs understandable to a wide audience. However, manual normalisation is a challenging and time consuming task.

In this study, we propose to assist the work currently conducted by historians and palaeographers in normalising the various wordforms observed in

[1]The exact time period when Middle French was spoken and written is still subject to debate among experts.

LT4HALA 2024@LREC-COLING-2024, pages 176–189

the RCs in the time of Calvin. Based on recent studies on spelling normalisation for French, we frame the task as a translation task (Bawden et al., 2022) in a very low resource setting. We leveraged pre-trained encoder–decoder large language models (LLMs), fine-tuned them with manually normalised RCs, to constitute a strong baseline. To improve over this baseline, we propose to enrich the available hand-crafted data with automatically generated parallel data, combining generative model prompting and back-translation (Marie and Fujita, 2021; Tonja et al., 2023). Our final model shows improved performances measured by automatic metrics compared to the baseline and to previously released normalisation models for French. Additionally, we present a qualitative analysis which highlights some of the differences between our approach and the baseline model.

To validate these findings, we conduct a manual evaluation to measure the post-editing time and effort spent by experts in correcting the automatically normalised RCs. The results show productivity gains in terms of normalisation throughput when using fine-tuned LLMs compared to manually normalising RCs from scratch. Moreover, our approach relying on synthetic data outperforms the fine-tuned LLMs making use of hand-crafted data only, both in terms of automatic metrics and productivity gain. To summarize, the contribution of our work is twofold: i) we describe a parallel data generation method which was not employed for historical text normalisation in previous work, and ii) we show that fine-tuning LLMs with a small amount of data greatly reduces the manual labour required to normalise Middle French, and can be further improved by using synthetic data.

The remainder of this paper is organised as follows. In Section 2, we introduce the background work for historical text normalisation, focusing on variants of the French language, before motivating our approach. In Section 3, the manual normalisation process is first presented, followed by the description and evaluation of the automatic normalisation process. The productivity gain expertiments based on post-editing is then detailed in Section 4 along with the corresponding results in terms of normalisation throughput. Finally, we conclude our study and present future work in Section 5.

2. Related Work

In this Section, we present the context of our normalisation work which is part of a larger project on Middle French modernisation. Then, previous work on spelling normalisation is introduced, followed by details about available resources relevant to our task. Finally, approaches to leveraging LLMs in low-resource scenario are described, before presenting the current limits of historical texts evaluation methods.

2.1. Context of the Study

The study presented in this paper and focusing on orthographic normalisation of RCs written in 16th century Middle French is part of a larger project which aims at producing a semantic and multilingual online edition of the Geneva Council Registers for the years 1545 to 1550. This project is based on a synergy between two faculties of the University of Geneva, the Centre universitaire d'informatique (CUI) and the Faculty of translation and interpreting (FTI), as well as the Fondation de l'Encyclopédie de Genève. The technical aspect of the project in terms of natural language processing is to automatise the normalisation and modernisation of RCs content, and to develop new functionalities that will make these archival documents accessible to a wide audience. Both normalisation and modernisation steps are leveraging low-resource machine translation techniques to process RCs, including fine-tuning large language models (LLMs) and producing artificial data as presented in this study. Each processing step applied to RCs will result in a version of the corpus, eventually resulting in multiple versions of RCs linked through token alignments. The assumption is that linked normalised and modernised RCs content will provide a useful source of knowledge for further research.

2.2. Normalising Spelling Variants

Intralingual diachronic translation aims at modifying textual content to match linguistic features of a time period, as orthographic, grammatical and syntactic features might evolve over time for a given language. A large body of work has been conducted on this task, in particular on normalising spelling variants observed in historical texts. Seminal studies on historical wordforms normalisation relied on distance and rule-based approaches (Hauser and Schulz, 2007; Rayson et al., 2007; Baron et al., 2009; Bollmann et al., 2011; Pettersson et al., 2013a; Bollmann et al., 2014). More recently, Machine Translation (MT) techniques were applied to intralingual diachronic translation, including statistical and neural models (Sánchez-Martínez et al., 2013; Pettersson et al., 2013b; Bollmann and Søgaard, 2016; Korchagina, 2017; Bollmann, 2018; Tang et al., 2018).

2.3. Available Resources

Among recent Neural MT (NMT) architectures, the Transformer (Vaswani et al., 2017) has been used for a variety of Natural Language Processing (NLP) tasks, including for automatic normalisation of Early Modern French (Bawden et al., 2022). The authors of this previous work have shown that Statistical MT (SMT) still outperforms the NMT architectures tested, namely LSTM (Hochreiter and Schmidhuber, 1997) and Transformer, with or without an additional lexicon-based post-processing step. The Transformer model used in their experiments was released and constitutes one of the few pre-trained resources available for spelling normalisation of French. We propose to evaluate their model on our task as a comparison point.

However, this publicly available model, called *ModFR*[2], was trained using the *FreEMnorm* corpus (Gabay and Gambette, 2022) which contains texts taken from French literature of the 17th century.[3] Our task involves the normalisation of Middle French, a language mainly used from the 14th to the 16th century (Buchi et al., 2019), which differs from previous study on French normalisation. While there is no clear consensus among philology and history experts about the beginning and the end dates of Middle French usage, the Dictionary of Middle French (Martin et al., 2020) covers the lexicon used from the year 1330 to 1500. This difference in time period between available resources for French and our task could hinder the straightforward application of previous approaches, but motivates us to leverage pre-trained LLMs to bootstrap our work.

2.4. Leveraging LLMs and Synthetic Data

Recent advances in NLP have been fuelled by the use of LLMs pre-trained on large amounts of data. However, to the best of our knowledge, only a few studies used pre-trained LLMs on tasks involving historical texts. For instance, Klamra et al. (2023) used a generative model to produce synthetic parallel data of archaic to modern Polish. This dataset was then used to fine-tune pre-trained encoder–decoder neural models to perform automatic modernisation of Polish. In our study, we propose to apply existing synthetic data generation techniques to build a parallel corpus. More precisely, inspired by Marie and Fujita (2021) and Tonja et al. (2023), a generative model is used to produce target data further back-translated into source data. In our study, the source data consists in non-normalised text while the target data consists in its normalised version. To the best of our knowledge, this is the first study on Middle French normalisation using fine-tuned LLMs.

2.5. Evaluation

If automatic metrics have been widely used to evaluate automatic normalisation models, the impact of such models on the productivity of experts conducting manual normalisation of historical texts has yet to be measured. In this study, we frame the comparison between normalising from scratch and editing automatically normalised text as a post-editing task. This allows us to perform manual evaluation of NMT-based normalisation models in terms of productivity gain, in addition to reporting results obtained with automatic metrics relying on gold references manually produced.

3. Normalisation of Middle French

This Section describes the normalisation process conducted to reduce spelling variants observed in Middle French contained in RCs written during the 16th century. First, the manual normalisation process is explained, presenting the orthographic modifications applied to the source text and defining the editorial choices made by the experts in terms of normalised wordforms. Second, the training and evaluation of automatic normalisation models are described, along with the synthetic data production method.

3.1. Manual Normalisation

The RCs consist in minutes of meetings held daily by Geneva canton council members. They contain political, administrative and judiciary decisions. They constitute a crucial resource for historical studies of the region for a given time period. The digitisation process of these manuscripts has been an ongoing effort, consisting mostly in scanning physical books. The results is a set of archived documents composed of RCs from 1408 to 1855 being publicly available online.[4] Recently, experts such as historians and palaeographers have been manually transcribing RCs. The work described in this paper is based on the manually transcribed version of RCs, which is the largest relevant dataset for our task.

More precisely, RCs from 1536 to 1550 were manually transcribed by historians and palaeographers. This task also involved slight modifications of the

[2] https://huggingface.co/rbawden/modern_french_normalisation

[3] https://github.com/FreEM-corpora

[4] https://ge.ch/arvaegconsult/ws/consaeg/public/FICHE/AEGSearch

Corpus	Segments	Tokens		Vocabulary		avg. tokens/segment	
		source	target	source	target	source	target
RCs	71.8k	2.7M	–	74.7k	–	37.4	–
RC_pe	2.5k	87.0k	–	7.0k	–	34.9	–
RC_para	2.3k	87.8k	82.4k	7.6k	5.7k	38.4	36.0
RC_synth	1.3M	195.2M	176.5M	0.47M	0.34M	147.2	133.0

Table 1: Number of segments, tokens and vocabulary entries (k for thousands, M for millions) for the transcribed RCs (noted *RCs*), the synthetic data created in our study (*RC_synth*), as well as the RCs subsets of original–normalised parallel text (*RC_para*) and original text used for the post-editing experiments (*RC_pe*). The corpora were normalised, lowercased and tokenised using the scripts released with the Moses toolkit (Koehn et al., 2007) prior to extracting data statistics.

Wordforms	Meaning
embossiou, enbosseu, entonnoir	a funnel
faulccry, faulxcry, foulcry, forcri	an alarm call
lause, lauze, loze, lose	a flat stone, a tile
maysoner, maisonner	to build
treul, true, trué, truez	a press
Toponyms	
Allemagne, Allamaignie, Allemagnyes, Allemaigne, etc.	Germany
Genève, Genefe, Genesve, Genevez, Genff, etc.	Geneva
Strasbourg, Estrabour, Estrapurg, Extrabourg, Strasburg, etc.	Strasburg

Figure 1: Examples of various wordforms encountered in the 16th century RCs, their normalised form is underlined, along with their meaning. Variants of toponyms are presented in the bottom part while general nouns and verbs are in the top part. The lists of toponym variants are truncated due to the large amount of wordforms observed.

textual content for increased readability by non-experts. The resulting corpus is a digital version of the RCs for the given time period covering 15 years. Furthermore, RCs from 1536 to 1544 were published as hard copy books. Both the digital and the hard copy versions of this corpus were not orthographically normalised and still contain a variety of wordforms, as illustrated in Figure 1. In addition to the manual transcription task, experts are currently conducting the manual orthographic normalisation of RCs content, starting from the transcription already done.

Due to the lack of spelling norms for Middle French during the 16th century, a large variety of wordforms were used compared to modern French. The manual normalisation consists in applying local orthographic and grammatical modifications to the original RCs content while leaving potentially archaic syntactic structures untouched. The normalisation guidelines defined by experts are described in Appendix A. This process differs from the historical text *modernisation* task, as it does not aim at transforming Middle French texts into their contemporary version. The objective is to reduce the spelling variations observed in RCs by selecting single wordforms. The latter are decided by experts conducting the manual normalisation task and follow editorial guidelines. We illustrate the normalisation process in Figure 2.

The main motivation behind conducting the orthographic normalisation of RCs is to improve the readability of texts difficult to understand while preserving the original structure. This will facilitate research in the historical, geographical and genealogical fields, among many others, by replacing various spelling variants with a single one. The orthographic normalisation will also serve as the basis for the syntactic normalisation of the text, which will in turn lead to its modernisation in current French. The latter two objectives are planned as future work but are out of scope of the presented study.

As a result of the manual normalisation, we currently have at our disposal a parallel set of RCs published over six months, one month per year from 1545 to 1550 (noted *RC_para*). This dataset is a subset of the non-normalised RCs manually transcribed from 1536 to 1550 (noted *RCs*). Details about these two corpora, along with the synthetic data described in Section 3.2 (noted *RC_synth*) and the RCs subset dedicated to post-editing (noted *RC_pe*), are presented in Table 1. Due to the small size of our hand-crafted parallel corpus, we will perform 5-fold cross-validation for all our automatic normalisation experiments presented in Section 3.2.

3.2. Automatic Normalisation

The aim of automatic normalisation is to assist historians and palaeographers in their task of manual normalisation and ultimately reduce their workload. We first propose to compare the performances of a publicly available pre-trained normalisation model for Early Modern French to

Le mardy 9e de octobre 1548 – L'on fasse respondre aut president de sadicte lectre
Le mardi 9e d'octobre 1548 – L'on fasse répondre au président de sadite lettre
Tuesday, October 9, 1548 – We answer to the president about his letter

---

(Les marchandz de Geneve) - Lesquieulx hont presenté une supplication par laquelle ilz prient
(Les marchands de Genève) - Lesquels ont présenté une supplication par laquelle ils prient
(The merchants of Geneva) - Who have presented a supplication by which they pray

---

Et dempuys a esté resoluz qui soyt liberé publiquement, à voex de trompe, et aut tribunal ordinayre.
Et depuis a été résolu qui soit libéré publiquement à voix de trompe et au tribunal ordinaire.
And it has since been resolved that he be released publicly and in ordinary court.

Ledictz jour, vendredy 28 octobrix 1547, en l'Evesché
Ledit jour vendredi 28 octobris 1547 en l'Évêché
Said day Friday October 28 1547 in the bishop's house[a]

Ayme Richard, habitant et ferratier, filz de feu Thivent Richard, de Sonzier
Aimé Richard habitant et ferratier fils de feu Thivent Richard de Scionzier
Aimé Richard inhabitant and ironworker son of the late Thivent Richard of Scionzier

[a]The bishop's house, translation of *Evesché* in this example, refers to the house inhabited by the previous bishop which was converted into a prison.

Figure 2: Segments sampled from the RCs original–normalised parallel corpus in Middle French, with segments in their original form (top, colored), their normalised version (middle, in black, normalised words underlined) and a possible English translation (bottom, *italic*).

an out-of-the-box pre-trained LLM. We then make use of our parallel data (*RC_para*) consisting of manually transcribed RCs as source and their normalised version as target.

3.2.1. LLM Setup

Our preliminary experiments showed that *m2m100* (Fan et al., 2021) outperforms other pre-trained MT models when fine-tuned with our data. Thus, we decided to conduct all our experiments using this model in its *base* version (418M parameters). We used the publicly released checkpoint available with the HuggingFace Transformers library (Wolf et al., 2019).[5] The fine-tuned version of this model using our parallel data is the *baseline* in our study. The fine-tuning procedures employed in our experiments are detailed for all models in Appendix B.

3.2.2. Synthetic Data

Due to the lack of parallel data relevant to the RCs and written in Middle French, we generated synthetic parallel data with a two-step process: 1) generative model prompting for target data generation, followed by 2) normalised-to-non-normalised back-translation to obtain a parallel corpus (Marie and Fujita, 2021; Tonja et al., 2023). The generative model used was *Bloomz* with 560M parameters (Muennighoff et al., 2022). This model was fine-tuned with the target side of our parallel corpus written in normalised Middle French. As this fine-tuning step relies on the training data taken from *RC_para*, it was conducted individually for each of the 5 folds. The motivation behind fine-tuning the generative model is to increase the relevancy of automatically generated data for the task at hand. Once the fine-tuning step was done, we proceed with prompting the model to produce synthetic data. The prompting method consisted in inputting sequences composed of consecutive tokens taken from the target side of *RC_para*, the same corpus used to fine-tune the generative model.[6]

The resulting target-side corpus automatically generated was then back-translated into the non-normalised source side of the synthetic parallel corpus. The back-translation model was trained on the combination of the *RC_para* corpus with the automatic translation of the *RCs* corpus.[7] The resulting parallel corpus, presented in Table 1 and noted *RC_synth*, was used to perform continued training of the pre-trained LLM (model noted *synthetic*) (Gururangan et al., 2020).[8] The average

[5]https://huggingface.co/facebook/m2m100_418M

[6]We used between 8 and 12 tokens as prompts to obtain different results and combine all the generated data.

[7]The automatic translation of the *RCs* corpus was obtained using the *baseline* model.

[8]A few samples of the produced synthetic data are presented in Appendix C.

model	BLEU	chrF	TER	WER	acc.
identity	24.2	65.8	45.0	42.5	13.4
m2m100	23.1	57.5	54.1	66.0	1.5
ModFR	32.3	71.1	38.5	38.8	11.9
baseline	79.7	91.1	11.9	6.7	47.4
synthetic	81.8*	92.2*	11.4	11.6	36.2
synthetic+ft	83.5*	93.6*	9.0*	5.7	47.8

Table 2: Averaged test results (5-fold cross-validation) measured by automatic metrics for the orthographic normalisation task of RCs, comparing the identity function (copy of the source) to previously released models (top part) and to our approach (bottom part). For BLEU, chrF and segment-level accuracy (*acc.*), the higher the better, while the lower the better for TER and WER. Results marked with * are significantly better than previous rows with $p < 0.01$, based on the paired bootstrap resampling technique with 1000 resamples.

number of tokens per segment for *RC_synth* is larger than for *RCs* and *RC_para* because we do not truncate the generated sequences, but instead let the generative model produce the end of sequence token. Finally, we fine-tune the resulting model using RCs parallel corpus (model noted *synthetic+ft*).

3.2.3. Automatic Metrics

The automatic evaluation was conducted using popular MT metrics, namely BLEU (Papineni et al., 2002), chrF (Popović, 2015) and TER (Snover et al., 2006), implemented in the SacreBLEU toolkit (Post, 2018).[9] For these three metrics, significance testing using paired bootstrap resampling with 1000 resamples was conducted to compare the *baseline*, *synthetic* and *synthetic+ft* models (Koehn, 2004). In addition, we measured the word error rate (WER) and the segment-level accuracy reached by the evaluated models. We believe that the latter metrics allow to grasp the manual effort required to produce publishable normalised text.

3.2.4. Quantitative Analysis

The 5-fold cross-validation test results measured by automatic metrics are presented in Table 2. We averaged results over the 5 runs, each run consisting in 60% of *RC_para* used as training set, 20% as validation and 20% as test (roughly 1.4k, 450 and 450 segments for the train, validation and test sets respectively). We evaluated a previously released normalisation model for Early Modern French (noted *ModFR*) (Bawden et al., 2022), along with a non-fine-tuned pre-trained LLM (*m2m100*). As an additional comparison point, we also considered the identity function, i.e. leaving the source non-normalised text untouched and comparing it to the normalised reference (noted *identity*). Finally, three fine-tuned versions of the *m2m100* model were also evaluated, namely the *baseline* model which was fine-tuned using the *RC_para* corpus only, the *synthetic* and *synthetic+ft* models which were fine-tuned using the *RC_synth* and *RC_synth+RC_para* respectively. The latter model was trained following a two-step process: continued training with *RC_synth* followed by fine-tuning with *RC_para*.

The results obtained with the segment-level automatic metric (*acc.*) show that previously released models do not outperform the identity function. The three MT-oriented metrics, namely BLEU, chrF and TER, as well as WER, show that *ModFR* outperforms both the identity function and out-of-the-box *m2m100*. Both the baseline and our final model (*synthetic+ft*) outperform the previously released model for Early Modern French according to the five metrics used. Adding synthetic data to the hand-crafted parallel corpus improves normalisation performances at the n-gram (BLEU), token (TER) and character (chrF) levels. However, when using synthetic data only without the final fine-tuning (model noted *synthetic*), a 11.2pts drop in terms of segment-level accuracy is observed, while gains are observed with MT metrics. This indicates that synthetic data improves normalisation at the n-gram, token and character levels, but introduce errors which lower the number of correctly normalised full segments. Finally, we see that our final model reaches the best scores overall, validating our synthetic data generation approach and confirming the need to eventually fine-tune the model using hand-crafted parallel data.

3.2.5. Coverage Analysis

As an additional experiment to help analyse the automatic normalisation results, we computed the rates of source-side out-of-vocabulary (OOV) tokens between the test set of each fold and the training sets used in our experiments, namely *RC_para* and *RC_synth*. We also included the *FreEMnorm* (Gabay and Gambette, 2022) corpus in the OOV rates calculation as it was used to train the *ModFR* (Bawden et al., 2022) model. We lowercased and tokenised all datasets prior to computing these rates, using the scripts released with the *Moses* toolkit (Koehn et al., 2007). We present the OOV results in Figure 3 for each fold in

[9] SacreBLEU signatures: version:2.3.1|nrefs:1 case:mixed|eff:no|tok:13a|smooth:exp case:mixed|eff:yes|nc:6|nw:0|space:no case:lc|tok:tercom|norm:no|punct:yes|asian:no

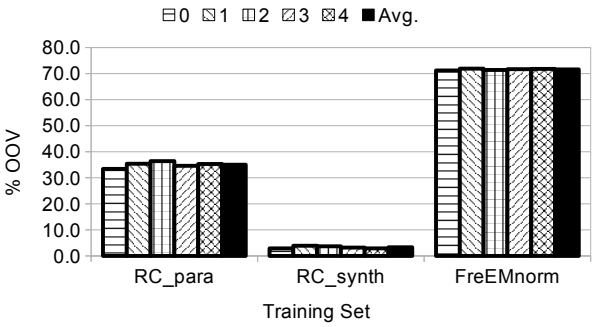

Figure 3: Out-of-vocabulary rates (%) for test tokens wrt. the training sets used in our experiments and the *FreEMnorm* (Gabay and Gambette, 2022) corpus. Hatched bars represent each fold individually (from 0 to 4) and the solid bar represents the averaged rate over 5 folds.

order to show that no particular fold suffered from a lower token-level coverage compared to the other folds.

The OOV rates clearly indicate that the *FreEMnorm* corpus provides a lower vocabulary coverage compared to the *RC_para* training set (71.6% vs. 35.0% OOV rates respectively). The low coverage of the Early Modern French corpus could partially explain the normalisation performances reached by the *ModFR* model on our Middle French data. Surprisingly, the source side of the synthetic data (resulting from the back-translation of the generative model output) reaches an average OOV rate of 3.3%, a 31.7pts absolute decrease compared to the average OOV rate obtained with the *RC_para* training sets. This particular result validates the use of synthetic data for vocabulary coverage in a low-resource scenario. However, while synthetic data is relatively cheap to produce, this approach still requires a small amount of well-formed target data to fine-tune the generative model.

3.2.6. Qualitative Analysis

To assess the strengths and weaknesses of the *baseline* and *synthetic+ft* models on specific elements to be normalised, we conduct a qualitative analysis of the automatically normalised segments. While the *baseline* model reaches relatively high performances compared to the other models, the *synthetic+ft* model is better at normalising the spelling of proper nouns and verbs, as presented by the examples in Appendix D. In the first example, the spelling of the proper noun *Pregnier* in the source segment should be normalised as *Pregny* but the *baseline* failed to do so while the *synthetic+ft* normalised it correctly. Similarly, in the third example, *Dolle* is normalised as *Dole* with the model using synthetic data. In terms of verb spelling, in the second example, *Doygbe* is correctly normalised as *Doive* by *synthetic+ft*, and in the third example, *requesté* is normalised as *requêté*.

Both models, however, introduce errors for source tokens which should not be modified, i.e. when no normalisation is necessary according to the gold reference. For instance, in Appendix D, the second example shows that the verb *levés* is correctly spelled in the source and reference segments while both the *baseline* and *synthetic+ft* models remove the plural form and rewrite it as *levé*. Overall, at the segment-level according to the *chrF* metric on the validation set, *synthetic+ft* is better than *baseline* for approx. 30% of the segments and both models are equal for approx. 55% of the segments. These results show that for approx. 15% of the segments, *baseline* is better than *synthetic+ft*.

4. Post-editing and Productivity Gain

One of the aims of this study is to measure the productivity gain achieved by using automatic normalisation followed by post-editing, compared to manually normalising from scratch. Moreover, we would like to validate the results obtained with automatic metrics in our previous experiments (cf. Table 2). Our post-editing experiments make use of a subset taken from the non-normalised source corpus, covering 5 months of the year 1545, which consists in approx. 2500 segments. Details about the dataset used are presented in Table 1 and noted *RC_pe*. The segments contained in *RC_pe* were normalised by our systems, namely *baseline* and *synthetic+ft*, or kept as is (i.e. the identity function), before being randomly presented to a human expert for post-editing.[10] We removed target segments which were identical between the two normalisation models and the identity function (approx. 500 segments were removed). The post-editing platform used in our experiments is *COPECO* (Mutal et al., 2020).

To conduct the post-editing task, we relied on a single historian who is an expert in 16th century Middle French texts and has participated in the manual transcriptions of RCs. The time spent on each segment, as well as the number of keystrokes for each segment, were measured during the post editing task. Due to the difficulty of this task even for trained experts, the set of segments to be post-edited was split in subtasks of approx. 100 segments. In order to limit the impact of normalising short and long segments on the final results, we kept segments containing between 2 and 128

[10] The same post-editing platform was used to post-edit all segments, including the source segments in case of the identity function and the automatically normalised segments as well. The post-editing interface is presented in Appendix E.

model	segments	tokens	keystrokes/token	time/token (sec)	token/minute	segment/hour
identity	385	12.1k	1.86	2.55	23.5	45.0
baseline	833	27.1k	0.42	1.33	45.3	83.4
synthetic+ft	834	28.1k	0.36	1.19	50.5	90.0

Table 3: Manual post-editing of RCs, comparing the identity function (copy of the source) to our approach (*baseline* and *synthetic+ft*) in terms of number of keystrokes per token, the time in seconds spent per token, the number of tokens processed per minute and the number of segments processed per hour. A larger number of segments were post-edited for the *baseline* and *synthetic+ft* systems compared to *identity* as we noticed a smaller gap in productivity gains between the outputs coming from the two former models.

model	BLEU	chrF	HTER	WER	acc.
identity	21.2	63.3	47.7	50.7	0.8
baseline	79.5	91.3	10.1	9.5	31.5
synthetic+ft	84.9	94.3	6.5	7.9	36.5

Table 4: Automatic metrics scores obtained when evaluating automatically normalised outputs using their manually post-edited version as reference, comparing the identity function (copy of the source) to our approach. For BLEU, chrF and segment-level accuracy (*acc.*), the higher the better, while the lower the better for HTER and WER.

tokens. Furthermore, we removed segments for which the post-editing time exceeded 5 minutes. Finally, segments for which 0 keystrokes were recorded but with a post-editing time exceeding 0.5 seconds were removed.

The results obtained in terms of normalisation productivity are presented in Table 3. The post-editing results show that both the *baseline* and the *synthetic+ft* models lead to increased normalisation productivity compared to normalising RCs from scratch. This is clearly shown by an increase in normalised token per minute and segment per hour. The number of keystrokes per token decreases with automatic normalisation compared to fully manual normalisation. Between the *baseline* and the *synthetic+ft* models, we observe a processed token per minute rate of 45.3 and 50.5 respectively. When measuring the number of segments processed per hour, an increase of 6.6 segments is reached by the model using synthetic data compared to the baseline.

These findings corroborate the results obtained in Table 2 with automatic metrics. We conducted further evaluations with the latter metrics to measure the distance between the models' outputs and their manually post-edited version. The results in terms of automatic metrics using the post-edited target as gold reference are presented in Table 4. We observe with these results that our final model (*synthetic+ft*) outperforms the baseline by 5.4pts BLEU and 5.0pts segment-level accuracy. Comparing results presented in Table 4 to results presented in Table 2, we see a decrease in performances when normalising the *RC_pe* corpus compared to normalising the *RC_para* corpus. This could be due to the lack of vocabulary coverage in the RCs from 1545, as *RC_para* is a mix of RCs covering one month per year from 1545 to 1550. We noticed that the RCs content vary from one year to another in terms of vocabulary, which is due to the various topics of discussion changing over time.

5. Conclusion

This paper presented a study on 16th century Middle French spelling normalisation. We compiled a dataset taken from the publicly available Geneva Council Registers which were manually transcribed, before manually normalising a subset of this corpus to build a parallel normalisation corpus. A strong baseline based on a pre-trained LLM, fine-tuned on the hand-crafted parallel corpus, was shown to outperform a previously released model trained for the normalisation of Early Modern French, as indicated by automatic metrics. Further experiments with synthetic data generation improved over this baseline at the segment, n-gram, token and character levels.

To validate these findings, we conducted a manual evaluation based on a post-editing task, comparing normalisation from scratch to the proposed approach. We show that fine-tuning a multilingual pre-trained LLM with a small amount of normalised parallel data increases the productivity of human experts by a relative gain of 92.8% in terms of normalised tokens per minute, compared to manually normalising text from scratch. Furthermore, adding synthetic data to the LLM fine-tuning increases productivity compared to the baseline by 5.2 tokens per minute, a 114.9% gain relative to full manual normalisation. It is, to the best of our knowledge, the first study on productivity gain measured through post-editing of 16th century Middle French archival documents normalisation.

As future work, we plan to run our approach iteratively, making use of the manually post-edited data to improve the performances of our

automatic normalisation model. The next step in the ongoing Middle French modernisation project is to conduct normalisation at the syntactic level, in addition to the current local orthographic and grammatical normalisation. In addition, we will explore various prompting techniques in order to obtain more relevant synthetic data from generative models. Finally, due to the change in topics discussed during Council meetings depending on the local events, we will conduct a diachronic study, measuring the impact of using temporally-related training and test data, compared to randomly sampling segments from the whole RCs content as we did in this study.

Limitations

We recognize the following limitations of this work.

First, the experiments were conducted on a variant of the Middle French language from the 16th century. Middle French has evolved over time and our work is considering a relatively narrow time frame in the history of this language.

Second, only a few pre-trained language models were tested during our preliminary experiments relatively to the large number of models currently publicly available. Some of these models were pre-trained on Modern or Early Modern French language, while other models were trained jointly on several languages, including languages relevant to our work such as Latin. Therefore, the models selected in our study may not be representative of all publicly released pre-trained models in terms of languages, number of parameters, training objectives nor architectures.

Third, the hand-crafted corpus produced in our work is relatively small in terms of number of tokens and vocabulary size compared to commonly used corpora in natural language processing experiments. This is mainly due to the high cost of producing such dataset for which the expertise of historians and palaeographers is required, while following strict editorial guidelines.

Finally, the post-editing experiments conducted in our work involves a single human expert. This is due to the nature of the task itself, requiring strong expertise in 16th century history, geographical knowledge of the Geneva canton, as well as a solid philological background to allow for Middle French normalisation and local grammatical alterations.

Ethical Considerations

The dataset hand-crafted in our study is based on publicly available archives from the 16th century (non-license, public domain). We reviewed the content of the documents selected for manual normalisation and we believe that this resource represents accurate historical events. However, some textual elements of this corpus could be considered as toxic and harmful, or disrespectful of the privacy of the people and places mentioned in these archives. We thus made sure that all data used in our work and to be released as part of our parallel datasets are in the public domain and already freely available. Consequently, no increased risks or harm is caused by our dataset. Instead, it serves as a resource for historical studies and digital humanities.

The fine-tuned models to be released with our work are based on publicly released and licensed pre-trained models (MIT License). We respect the permissions to use, modify and distribute the models. We will release the fine-tuned models under the MIT License.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments.

This work was partially funded by the FNS (Fonds national suisse), SSH 2022 grant n. 215733, for the project entitled *Une édition sémantique et multilingue en ligne des registres du Conseil de Genève (1545-1550)*, acronym *RCnum*.

All experiments were conducted on the University of Geneva computing cluster HPC *Baobab* and *Yggdrasil*.

Bibliographical References

Irena Backus and Philip Benedict. 2011. *Calvin and his influence, 1509-2009*. Oxford University Press.

Alistair Baron, Paul Rayson, and DE Archer. 2009. Automatic standardization of spelling for historical text mining. *Proceedings of Digital Humanities 2009*.

Rachel Bawden, Jonathan Poinhos, Eleni Kogkitsidou, Philippe Gambette, Benoît Sagot, and Simon Gabay. 2022. Automatic normalisation of early modern french. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3354–3366.

Marcel Bollmann. 2018. *Normalization of historical texts with neural network models*. Ph.D. thesis, Dissertation, Bochum, Ruhr-Universität Bochum, 2018.

Marcel Bollmann, Florian Petran, and Stefanie Dipper. 2011. Rule-based normalization of historical texts. In *Proceedings of the Workshop*

on Language Technologies for Digital Humanities and Cultural Heritage, pages 34–42.

Marcel Bollmann, Florian Petran, and Stefanie Dipper. 2014. Applying rule-based normalization to different types of historical texts—an evaluation. In *Human Language Technology Challenges for Computer Science and Linguistics: 5th Language and Technology Conference, LTC 2011, Poznań, Poland, November 25–27, 2011, Revised Selected Papers 5*, pages 166–177. Springer.

Marcel Bollmann and Anders Søgaard. 2016. Improving historical spelling normalization with bi-directional lstms and multi-task learning. *arXiv preprint arXiv:1610.07844*.

Eva Buchi, Bernard Combettes, Veronika Lux-Pogodalla, Béatrice Stumpf, Gilles Toubiana, and Delphine Barbier-Jacquemin. 2019. Histoire du français–frise chronologique.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Simon Gabay and Philippe Gambette. 2022. FreEM-corpora/FreEMnorm: FreEM norm Parallel (original vs. normalised) corpus for Early Modern French.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Andreas W Hauser and Klaus U Schulz. 2007. Unsupervised learning of edit distance weights for retrieving historical spelling variations. In *Proceedings of the first workshop on finite-state techniques and approximate search*, pages 1–6.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Cezary Klamra, Katarzyna Kryńska, and Maciej Ogrodniczuk. 2023. Evaluating the use of generative llms for intralingual diachronic translation of middle-polish texts into contemporary polish. In *International Conference on Asian Digital Libraries*, pages 18–27. Springer.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180. Association for Computational Linguistics.

Natalia Korchagina. 2017. Normalizing medieval german texts: from rules to deep learning. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pages 12–17.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Benjamin Marie and Atsushi Fujita. 2021. Synthesizing monolingual data for neural machine translation. *arXiv preprint arXiv:2101.12462*.

Robert Martin, Sylvie Bazin, and Pierre Cromer. 2020. Dictionnaire du moyen français. *ATILF-CNRS & Université de Lorraine*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Jonathan Mutal, Pierrette Bouillon, Perrine Schumacher, and Johanna Gerlach. 2020. Copeco: a collaborative post-editing corpus in pedagogical context. In *Proceedings of 1st Workshop on Post-Editing in Modern-Day Translation*, pages 61–78.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Eva Pettersson, Beáta Megyesi, and Joakim Nivre. 2013a. Normalisation of historical text using context-sensitive weighted Levenshtein distance and compound splitting. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 163–179, Oslo, Norway. Linköping University Electronic Press, Sweden.

Eva Pettersson, Beáta Megyesi, and Jörg Tiedemann. 2013b. 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, pages 54–69.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Paul Rayson, Dawn Archer, Alistair Baron, and Nicholas Smith. 2007. Tagging historical corpora-the problem of spelling variation. In *Dagstuhl Seminar Proceedings*. Schloss Dagstuhl-Leibniz-Zentrum fr Informatik.

Felipe Sánchez-Martínez, Isabel Martínez-Sempere, Xavier Ivars-Ribes, and Rafael C Carrasco. 2013. An open diachronic corpus of historical spanish: annotation criteria and automatic modernisation of spelling. *arXiv preprint arXiv:1306.3692*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 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*, pages 223–231.

Gongbo Tang, Fabienne Cap, Eva Pettersson, and Joakim Nivre. 2018. An evaluation of neural machine translation models on historical spelling normalization. *arXiv preprint arXiv:1806.05210*.

Atnafu Lambebo Tonja, Olga Kolesnikova, Alexander Gelbukh, and Grigori Sidorov. 2023. Low-resource neural machine translation improvement using source-side monolingual data. *Applied Sciences*, 13(2):1201.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

A. Appendix: Normalisation Guidelines

The normalisation guidelines were defined by the historian in charge of manually normalising RC content. This person is an expert in 16th century Middle French, in the Geneva region and in the political landscape in Calvin's time. The same expert was in charge of post-editing the automatically normalised content produced by our models. The same guidelines were used when manually normalising RC content from scratch and when post-editing our models' output.

The normalisation applied to the source textual content is focused on local orthographic and grammatical elements while leaving syntactic structures unchanged. This normalisation process is part of a larger normalisation and modernisation effort, as well as lexical enrichment and indexing, as described in Section 2.1. The normalisation guidelines were the following:

- First characters are uppercased at the start of sentences, but also for patroyms and toponyms.
- Limit the use of ponctuation marks:
  - semicolons in lemmas only to separate different items,
  - commas before decisions, e.g. (regarding) ordered/stopped/solved,
  - periods at the end of sentences.
- Use of diacritical marks (apostrophes) except for cases where *que*, followed by a vowel, actually stands for *qui*, e.g. *sont survenues quelques lettres que attouchaient à Genève* (in English: a few letters about Geneva appeared).
- Extended emphasis and accentuation based on modern usage
- Gender and number of past participle agreement, e.g. *de celui qui les a baillé* becomes *de celui qui les a baillés*, *sus la supplication qui a présenté* becomes *sus la supplication qui a présentée*, except when there is a doubt such

as *lui soit baillé trois écus* **not** to be corrected in *lui soient baillés trois écus* because it is an ambiguous case: *trois écus* could be the object or the subject.

- Verb agreement, e.g. *ordonné que lesdits six écus lui soit délivrés* becomes *ordonné que lesdits six écus lui soient délivrés* (in English: ordered that the said six écus be delivered to him)
- Modernisation of patronyms, first names and toponyms.
- Correction of genders according to modern usage, e.g. *la dimanche* (in English: the Sunday) becomes *le dimanche*, *la reste* (in English: the rest) becomes *le reste*.
- Singular feminine possessive determiner replacement, e.g. *ma* (my), *ta* (your), *sa* (his, her, their), for nouns starting with a vowel or with a silent *h*, by the masculine forms *mon*, *ton*, *son*. For instance, *à sa humble requête* becomes *à son humble requête* (in English: to his/her/their humble request).

B. Appendix: LLM Fine-tuning Procedure

All models fine-tuned and evaluated in this work relied on the HuggingFace Transformers library (Wolf et al., 2019) with the Pytorch backend (Paszke et al., 2019). Models fine-tuning were conducted on single Nvidia RTX A5000 and 3090 GPUs with 24GB memory during a maximum of 100k steps (maximum of 12h) with early stopping if convergence is reached. We used batch sizes between 4 and 16 segments depending on training and testing phases. The optimizer used was AdamW (Loshchilov and Hutter, 2017), measuring BLEU scores on the validation set every 500 steps for the *baseline* and *synthetic+ft* models, and every 5000 steps for the *synthetic* model. The back-translation and normalisation models based on *m2m100* with 418M parameters were using the configuration released with the checkpoint, except for the learning rate. For the latter hyper-parameter, we searched for the best learning rate in a given range by monitoring performances obtained on the validation set. The learning rate search ranges were:

- *baseline* model: between $1e^{-6}$ and $2e^{-5}$
- *synthetic* model: between $8e^{-7}$ and $2e^{-5}$
- *synthetic+ft* model: between $8e^{-7}$ and $2e^{-6}$

The generative models were fine-tuned for 100k steps with a batch size of 4 using the AdamW optimizer. Three learning rates were used leading to three fine-tuned models: $8e^{-7}$, $1e^{-6}$ and $5e^{-6}$. The resulting models were finally averaged to compose the final generative model used to produce synthetic target data through prompting.

C. Appendix: Synthetic Parallel Data

The segments below are sampled from the *RC_synth* corpus, with the target side (in black, with differences underlined) produced by prompting a fine-tuned generative model before being back-translated to produce the source side (non-normalised, colored).

Accord passé entre Jehan Cuvat, ancien admodiataire du revenuz de l'Hospital, et François Beguin, consellier des comptes
Accord passé entre Jean Cuvat ancien amodiataire du revenu de l'hôpital et François Béguin conseiller des comptes

---

M. Morel, le tressorier Corne, disant qui ont remercié Dieu et la Ville de ne fere poyé aulchongs droys ny aulcunes retenues de ce qui a esté adjugé à l'Hospitall.
M. Morel le trésorier Corne disant qui ont remercié Dieu et la ville de ne faire payer aucuns droits ni aucunes retenues de ce qui a été adjugé à l'hôpital.

Deviser et conferir ensemble que ilz puissent aussi avoir conseilz de ceulx qui serontz expers.
Deviser et conférer ensemble qu'ils puissent aussi avoir conseil de ceux qui seront expers.

(Le seigneur Curteti, de Jussier) - Lequel a prier luy faire aulmone de ce que possede et des biens qui sera expirer, et l'a faict poyer.
(Le seigneur Curtet de Jussy) - Lequel a prié lui faire aumône de ce que possède et des biens qui sera expiré et l'a fait payer.

Et sur ce, ordonné qui soit faict ung prisonnier et que le chastellain se doibge enquerré de la verité du faict, et sus luy l'on fera justice.
Et sur ce ordonné qui soit fait un prisonnier et que le châtelain se doive enquerre de la vérité du fait et sur lui l'on fera justice.

Leur a esté par cy-devant imposé. Sur quoy, Messieurs du Petit Conseyl, il ont refferuz que hier, il furent informés que le seigneur Amyed Perrin, jadix ministre de Loys Bernard, lequelt avoyt malle servente avecque Claude Du Pan, lequell ont palliarder et ce que il avient fayct, ce ont estés chastiés, et maentenant il en ont

pour leur responces . . .
Leur a été par ci-devant imposée sur quoi messieurs du petit conseil ils ont référé que hier ils furent informés que le seigneur Ami Perrin jadis ministre de Louis Bernard lequel avait maille servante avec Claude Dupan lequel ont paillardé et ce que ils avaient fait ce ont été châtiés et maintenant ils en ont pour leurs réponses . . .

D. Appendix: Qualitative Analysis

The segments below are extracted from the validation set where the *synthetic+ft* model outperforms the *baseline* on verbs and proper nouns spelling.Underlined tokens are correctly normalised by *synthetic+ft* and erroneous with *baseline*.

source
(Les admodiataires et dismier de Pregnier; Michiel Mallet) - Lequel a requis qui plaise à Messieurs avoir regard sus la tempeste tombee sur leurs diesme etc.
baseline
(Les amodiataires et dîmeurs de Périgny Michel Malet) - Lequel a requis qui plaise à messieurs avoir regard sur la tempête tombe sur leur dîme etc.
synthetic+ft
(Les amodiataires et dîmeurs de Pregny Michel Malet) - Lequel a requis qui plaise à messieurs avoir regard sur la tempête tombée sur leur dîme etc.
reference
(Les amodiataires et dîmeur de Pregny Michel Maillet) - Lequel a requis qui plaise à messieurs avoir regard sur la tempête tombée sur leur dîme etc.
translation
(The lessees and the tithe stewards of Pregny Michel Maillet) - Who requested that the councillors consider the storm that fell on their tithe etc.

---

source
(Jacque-Nycolas Vulliet) - Doybge rendre les gages levés à cause qui ne cria pas.
baseline
(Jacques-Nicolas Vulliet) - Doyge rendre les gages levé à cause qui ne criera pas.
synthetic+ft
(Jacques-Nicolas Vulliet) - Doive rendre les gages levé à cause qui ne cria pas.
reference
(Jacques-Nicolas Vulliet) - Doive rendre les gages levés à cause qui ne cria pas.
translation
(Jacques-Nicolas Vulliet) - Must return the guarantees because he did not auction.

---

source
(La Guygona) - Laquelle a requesté luy oultroyer une lectre de faveur, affin avoir pour son mary detenuz en prison à Dolle etc.
baseline
(La Guygona) - Laquelle a requéré lui octroyer une lettre de faveur afin avoir pour son mari détenu en prison à Dolle etc.
synthetic+ft
(La Guyonay) - Laquelle a requêté lui octroyer une lettre de faveur afin avoir pour son mari détenu en prison à Dole etc.
reference
(La Guigone) - Laquelle a requêté lui octroyer une lettre de faveur afin avoir pour son mari détenu en prison à Dole etc.
translation
(La Guigone) - Who requested to grant her a letter of favor in order to have for her husband detained in prison in Dole etc.

E. Appendix: Post-editing Interface

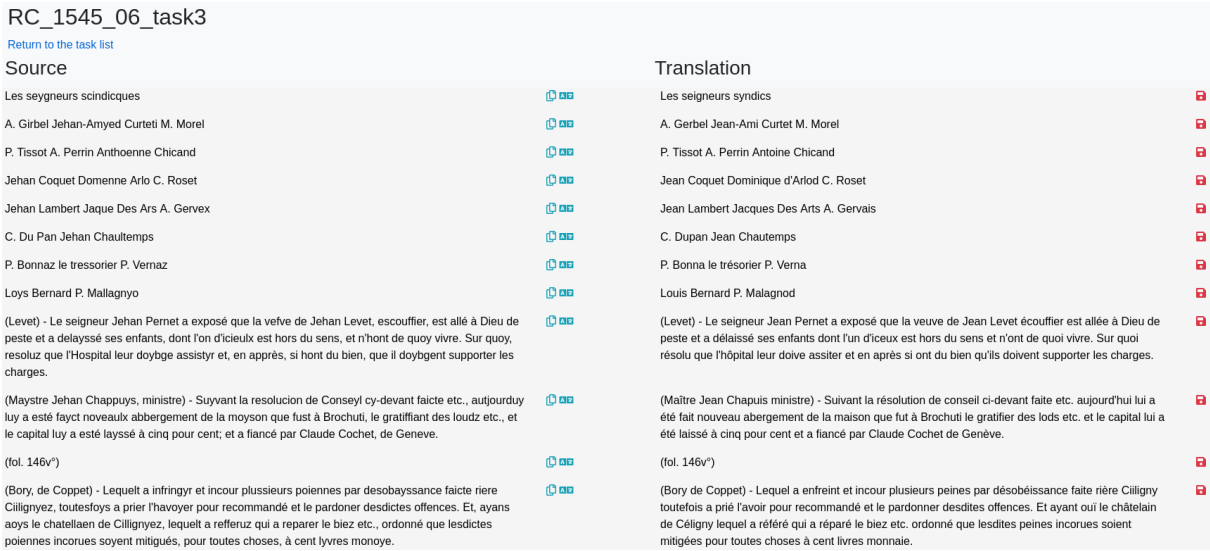

Figure 4: Post-editing interface used in our experiments to measure the productivity gain brought by automatic normalisation models compared to manually normalising RCs from scratch. The source text is presented on the left side while the normalised hypothesis is presented on the right side. Each editable block contains a segment as it appears in the original manuscript.