Le projet FREEM : ressources, outils et enjeux pour l'étude du français d'Ancien Régime

Simon Gabay, Pedro Ortiz Suarez, Rachel Bawden, Alexandre Bartz, Philippe Gambette, Benoît Sagot

Le projet FREEM : ressources, outils et enjeux pour l'étude du français d'Ancien Régime

Simon Gabay[1] Pedro Ortiz Suarez[2, 3] Rachel Bawden[2] Alexandre Bartz[3]
Philippe Gambette[4] Benoît Sagot[2]
(1) Université de Genève, 1211 Genève, France
(2) Inria, 75012 Paris, France
(3) Sorbonne Université, 75005 Paris, France
(4) Université Gustave Eiffel, CNRS, LIGM, 77454 Marne-la-Vallée, France
prenom.nom@unige.ch, prenom.nom@inria.fr,
prenom.nom@sorbonne-universite.fr, prenom.nom@univ-eiffel.fr

RÉSUMÉ

En dépit de leur qualité certaine, les ressources et outils disponibles pour l'analyse du français d'Ancien Régime ne sont plus à même de répondre aux enjeux de la recherche en linguistique et en littérature pour cette période. Après avoir précisément défini le cadre chronologique retenu, nous présentons les corpus mis à disposition et les résultats obtenus avec eux pour plusieurs tâches de TAL fondamentales à l'étude de la langue et de la littérature.

ABSTRACT

The FREEM project: Resources, tools and challenges for the study of *Ancien Régime* French

Despite their undoubted quality, the resources and tools available for the analysis of *Ancien Régime* French are no longer able to meet the challenges of research in linguistics and literature for this period. After having precisely defined the chronological framework, we present the corpora made available and the results obtained with them for several NLP tasks, fundamental to the study of language and literature.

MOTS-CLÉS : Linguistique diachronique, Reconnaissance d'entités nommées, normalisation.

KEYWORDS: Diachronic linguistics, Named-entity recognition, linguistic normalisation.

La communauté de recherche intéressée par la linguistique computationnelle s'attaque depuis déjà plusieurs années aux anciens états de la langue française (Kunstmann & Stein, 2007), longtemps restés parents pauvres du traitement automatique des langues (TAL). Pour l'ancien français, des corpus et des modèles de lemmatisation et d'annotation morpho-syntaxique (Stein, sd; Camps *et al.*, 2021) ou d'analyse syntaxique (Prévost & Stein, 2013; Stein, 2016; Regnault *et al.*, 2019) sont désormais disponibles. Pour le français « moderne », la situation est plus contrastée, les outils développés pour le français contemporain fonctionnant bien sur la partie basse du spectre chronologique (notamment à partir de la fin du XVIII^{e} s.), mais moins sur la partie la plus haute (XVII^{e}, et surtout XVI^{e} s.).

Pour la période de l'Ancien Régime, compris ici comme allant du XVI^{e} à la fin du XVIII^{e} s., le corpus *Épistémon* des Bibliothèques virtuelles humanistes (Uetani *et al.*, 2016), ciblé sur le français de la Renaissance, a conduit à créer des outils de reconnaissance d'entités nommées fondés sur des règles (Maurel *et al.*, 2014). Nous avons aussi pu compter ces dernières années sur l'excellent travail

effectué par les collègues du projet *Presto*, qui ont mis à disposition un riche corpus (Blumenthal *et al.*, 2017) et des premiers outils d'annotation (Diwersy *et al.*, 2017) pour une langue peu retouchée (Gabay, 2014). En ayant recours à un outil utilisant des méthodes probabilistes comme *TreeTagger* (Schmid, 1994) et en n'offrant qu'une annotation limitée (lemmes et parties du discours uniquement), les solutions développées par le projet *Presto* ne répondent néanmoins plus qu'imparfaitement aux besoins des chercheurs en littérature, attirés par des analyses novatrices requérant un nombre croissant d'informations, avec une précision optimale.

Il convient donc de rouvrir le dossier en expérimentant avec de nouveaux corpus, préparés pour cette occasion, les dernières méthodes apparues en TAL. Nous faisons en effet le pari de l'apprentissage profond, exploitant les approches neuronales en plus, le cas échéant, des approches plus traditionnelles par règles ou statistiques. Nos premières expérimentations concernent plusieurs tâches traditionnelles du TAL comme la reconnaissance des entités nommées ou la traduction automatique (utilisée à des fins de normalisation des états anciens de la langue). Outre l'amélioration des outils existants, ces expériences promettent d'ouvrir de nouvelles méthodes d'étude en histoire de la langue, comme pour l'étude des systèmes graphiques (*i.e.* l'orthographe) sur la longue durée et de vastes corpus.

1 Le cadre chronologique

La création d'un corpus cohérent pose la question de la chronologie. N'espérons pas ici résoudre le problème de la périodisation du français, qui est extrêmement complexe, et a amené un nombre incalculable de publications et autant de résumés (Ayres-Bennett & Caron, 2016). Retenons qu'une classification simple retient trois catégories (ancien français, moyen français, français moderne), et que ces trois catégories sont constamment remises en question. Certains font durer le moyen français jusqu'en 1500 (Martin, 1998), et d'autres jusqu'en 1611 (Greimas & Keane, 2007). Le français classique, une sous-catégorie du français moderne, s'est récemment vu séparé d'un français préclassique (Combettes & Marchello-Nizia, 2010) qui s'achèverait au plus tard en 1660 (Brunot, 1905), au plus tôt en 1630 (Amatuzzi *et al.*, 2020). Un « français de la première modernité », pour reprendre la catégorie anglo-saxonne d'*Early Modern* (Badiou-Monferran, 2011), s'achève traditionnellement avec la Révolution française, s'appuyant donc sur un critère externe dont la validité est fortement contestée en linguistique diachronique (Ayres-Bennett & Caron, 2016). Arrive ensuite le français contemporain, courant tout le long des XIX^{e}-XX^{e} s., qui reste, comme le XVIII^{e} s., nettement à l'écart des études de romanistique.

Le problème de la périodisation étant insoluble, écartons-le (pour l'instant) : l'étude du français historique dans la durée, sans même parler de sa périodisation, requiert un empan aussi large que possible pour éviter de se mettre d'emblée des œillères inutiles. Toute étude méritant des bornes chronologiques, trois critères ont été retenus pour les définir :

— L'existence de ressources déjà disponibles. Le rapide exposé *supra* démontre l'existence de données pour toute la période de l'Ancien Régime ;
— Les moyens technologiques à disposition pour le traitement rapide et efficace de documents anciens. Une nouvelle chaîne de traitement permet désormais de garantir une acquisition raisonnable du texte contenu dans (presque) tous les imprimés de langue française, et potentiellement des manuscrits (Gabay *et al.*, 2022c) ;
— Notre hypothèse de travail, à savoir, dans un premier temps, l'étude de la convergence de l'écrit vers un même standard. Cette étude devrait nous permettre de mieux comprendre

l'évolution de la langue et ainsi mieux définir les corpus nécessaires aux futurs outils pour la période d'Ancien Régime.

Le changement linguistique en français est marqué, comme pour de nombreuses langues, par un phénomène de convergence linguistique à l'écrit. Cette standardisation diffère cependant légèrement des autres, car plus qu'une norme, née plus ou moins spontanément et collectivement, le français se voit imposer une règle, dont l'un des principaux promoteurs est évidemment l'Académie française. Dans la mesure où les années 1670 sont encore le théâtre de débats sur la langue (Mézeray, 1951), que le premier dictionnaire date de 1694 (Académie française, 1694), et surtout qu'aucune règle de ce type n'est immédiatement appliquée, le processus de standardisation n'a pu se terminer chez les scripteurs éduqués qu'au cours du XVIII^e s. – le cas de l'orthographe étant sans doute le plus évident (Pellat & Andrieux-Reix, 2006).

Ce français « d'Ancien Régime » pose une série de problèmes spécifiques, notamment du fait de son système graphique instable dans le temps comme dans l'espace. Il change d'un époque à une autre, mais aussi d'un scripteur/copiste à un autre, tout en se distinguant nettement d'états plus anciens comme l'ancien français, ce qui en complexifie son traitement informatique. Cette instabilité, qui tend à se réduire dans le temps, n'a été que très peu étudiée, en dépit des implications fortes que revêt un tel processus pour la linguistique, mais aussi pour le TAL.

Étant donné les trois paramètres retenus, une étude computationnelle raisonnablement ambitieuse du changement linguistique en français, et notamment de sa standardisation, devant retenir les deux dates extrémales du spectre, nous prendrons donc 1500 comme *terminus a quo*, et la fin du XVIII^e s. comme *terminus ad quem*. Cette période renvoyant factuellement à celle de l'Ancien Régime, elle sera désormais définie comme telle, l'utilisation des notions de « (pré)classicisme » (sacralisant idéologiquement un type de littérature) ou de « (première) modernité » (créant une continuité arbitraire avec les siècles suivants) pouvant être utilisées par commodité, mais devant être compris comme s'inscrivant dans le cadre chronologique que nous venons de définir.

2 De nouveaux corpus

Notre approche étant prospective et maximaliste, nous proposons plusieurs corpus répondant à différentes tâches de TAL pour l'étude de la langue : la collection FREEM (pour *French Early Modern*) [1].

— FREEM$_{LPM}$ (pour « *l*emme, *p*artie du discours, *m*orphologie »). Ce corpus, libre de droits [2], utilise LGeRM comme référentiel de lemmatisation. Il est en trois parties :
  — Un petit corpus (c. 200 000 tokens) annoté avec les lemmes, les parties du discours et la morphologie, entièrement corrigé à la main, tiré des projets *Presto* et *CornMol* (Camps *et al.*, 2020) ;
  — Les textes écrits entre le XVI^e et le XVIII^e c. du corpus "noyau" de *Presto* avec une lemmatisation intégralement corrigée pour correspondre à nos choix de lemmatisation (c. 6 800 000 tokens) ;
  — Le corpus *Frantext Démonstration* avec une lemmatisation intégralement corrigée, là encore pour correspondre à nos choix de lemmatisation (c. 2 400 000 tokens).

1. Pour plus d'informations concernant ces corpus (acquisition, droits, annotation…), voir notre site https://freem-corpora.github.io.
2. Le corpus est disponible à l'adresse suivante : https://doi.org/10.5281/zenodo.6481300.

— FREEM$_{NER}$ (pour « *n*amed-*e*ntity *r*ecognition ») contient la partie "noyau" du corpus *Presto* utilisée par FREEM$_{LPM}$, et est donc aussi libre de droits. Il est annoté à la main avec la typologie de *Quaero* (Rosset *et al.*, 2011), en repérant les éléments imbriqués (comme un nom de lieu dans un nom de personne dans *Louis roi de* ***France***) et en utilisant des étiquettes pour distinguer des composants (*Louis* → `name` et *roi de France* → `title`)[3]. On y trouve un début d'annotation avec les identifiants *Wikidata* pour les lieux.
— FREEM$_{norm}$ (pour « *norm*alisation ») est libre de droit[4], associant des textes à leur version normalisée manuellement (c. 650 000 tokens).
— FREEM$_{max}$ (pour « *max*imal ») qui rassemble le maximum de textes disponibles, normalisés ou non, écrit en français dans la fourchette chronologique retenue (c. 185 643 482 tokens). Une version *open access*, avec moins de texte, est librement disponible[5].

Un manuel d'annotation est maintenu (Gabay *et al.*, 2020) pour documenter en détail les choix effectués et faciliter la réutilisation des données par d'autres. Les tâches concernées sont pour l'instant la lemmatisation, l'étiquetage morpho-syntaxique, la normalisation (*i.e.* alignement avec le français contemporain) et la reconnaissance d'entités nommées (REN).

L'étude de la langue ne nécessitant pas uniquement un seul outil, mais toute une gamme de solutions, nous avons adopté une approche holistique du problème, en misant sur la création d'un modèle de langue dédié au français d'Ancien Régime, D'AlemBERT[6], qui devrait venir en soutien des différentes tâches de TAL envisagées (Gabay *et al.*, 2022b).

3 Reconnaissance des entités nommées

La REN est une tâche d'extraction d'information des corpus, elle consiste plus précisément en la détection des objets textuels catégorisables dans des classes telles que les noms de personnes, les noms des lieux, les organisations, les produits, les évènements et même les quantités ou les fonctions. Si des corpus existent déjà pour les émissions radiophoniques (Galliano *et al.*, 2009) ou la presse des XIX^{e}-XX^{e} s. (Neudecker, 2016; Ehrmann *et al.*, 2020), les ressources manquent pour le français plus ancien (Ehrmann *et al.*, 2021). La seule expérience dont nous avons connaissance pour l'Ancien Régime a été faite sur des données annotées automatiquement, avec peu de types d'entités et sur un corpus concernant l'histoire de la Suisse (Gwerder, 2017).

Les besoins pour un tel outil sont pourtant fort dans la communauté des chercheurs en lettre, en histoire, en linguistique, etc. Il facilite l'étude de ces documents en rendant plus accessible des tâches dérivées telles que la détection des citations ou des mentions, ou en facilitant la création de visualisations (par ex. avec des cartes pour les lieux) à même de rendre compte du contenu d'un texte. Nous avons donc décider de rouvrir le dossier en nous inspirant de précédentes expériences couronnées de succès pour le français contemporain (Ortiz Suárez *et al.*, 2020), en utilisant le modèle de de langue D'AlemBERT (Gabay *et al.*, 2022b) présenté *supra* plutôt que CamemBERT (Martin *et al.*, 2020) et le nouveau corpus FREEM$_{NER}$ plutôt que le *French TreeBank* (Abeillé *et al.*, 2003). La première étape est cependant la création d'une *baseline* que nous présentons ici.

Pour évaluer notre corpus d'entraînement FREEM$_{NER}$ et pour établir un résultat de base avec lequel

3. Le corpus est disponible à la même adresse que FREEM$_{LPM}$, dont il constitue une sous-partie.
4. Le corpus est disponible à l'adresse suivante : `https://doi.org/10.5281/zenodo.5865428`.
5. Le corpus est disponible à l'adresse suivante : `https://doi.org/10.5281/zenodo.6481135`.
6. Référence à l'architecture BERT et au célèbre encyclopédiste Jean Le Rond d'Alembert (1717-†1783).

de futures études pourront se comparer, nous entraînons un modèle Bi-LSTM-CRF (Lample *et al.*, 2016). Pour cela, nous utilisons la librairie *Flair*[7] qui nous fournit une distribution de ce modèle classique simple à utiliser et à entrainer. Pour les plongements lexicaux nous utilisons la distribution française de FastText (Grave *et al.*, 2018), pré-entrainée avec le corpus web *CommonCrawl*[8]. Nous utilisons les hyper-paramètres standard proposés par Flair pour ce modèle[9], sauf pour la taille du *batch*, pour laquelle nous utilisons une valeur de 256, l'idée étant de maximiser la quantité d'entités moins fréquentes (`org`, `prod`, et `event` dans notre cas) par itération de l'entraînement. Le corpus est divisé en 3 parties (entraînement, développement et test). Pour cette partition nous prenons le 90% des phrases de chaque document pour la partie d'entraînement, 5% pour la partie de développement et 5% pour le la partie de test.

Classe	Précision	Rappel	Score-f1	Support
pers	0,8808	0,8435	0,8617	2734
loc	0,8109	0,8707	0,8397	1384
amount	0,9040	0,9040	0,9040	250
time	0,9604	0,9237	0,9417	236
func	0,8872	0,8429	0,8645	140
org	0,8824	0,6122	0,7229	49
prod	0,9231	0,4444	0,6000	27
event	0,7273	0,6667	0,6957	12
micro avg	0,8640	0,8533	0,8586	4832

TABLE 1 – Résultats par classe pour la partie *test* du corpus FREEM$_{NER}$ du modèle Bi-LSTM-CRF pour l'annotation d'entités nommées.

Les résultats de nos expériences, présentés dans la table 2, montrent que notre modèle obtient de très bons scores pour les catégories *quantité*, *temps* et *fonction*. La proximité forte entre les français contemporain et historique pour ces catégories pourrait expliquer la qualité de ces résultats, qui démontreraient donc un transfert de connaissances entre deux états de langue – le plongement de mots ayant été entraîné avec une énorme quantité de contenu textuel contemporain. Concernant les autres catégories, le transfert de connaissances grâce à FastText serait plus limité, sans néanmoins affecter la qualité de résultats qui restent particulièrement bons pour les catégories *personne* et *lieu*, abondamment présentes dans FREEM$_{NER}$.

4 Normalisation

La tâche de normalisation consiste à transformer un texte qui ne correspond pas à un standard choisi au préalable en un texte équivalent qui respecte ce standard. Elle a donc un double effet : celui d'aplanir les variations présentes dans le texte d'origine, par exemple les incohérences graphiques, et celui d'opérer une conversion vers le standard retenu. Dans ce travail, nous avons utilisé le français contemporain comme standard cible. Les deux vers de Marie-Catherine de Villedieu de la figure 1 illustrent divers types de transformations résultant d'une telle opération de normalisation. Cela va

7. https://github.com/flairNLP/flair
8. https://commoncrawl.org
9. https://github.com/flairNLP/flair/blob/master/resources/docs/TUTORIAL_7_TRAINING_A_MODEL.md

de changements graphématiques (*ſ* → *s*, *e* → *é*, *e* → *è*, *y* → *i*) à des changements de segmentation (*bien-toſt* → *bientôt*), en passant par des changements de digrammes vocaliques pour la conjugaison de l'imparfait ou du participe passé (*pouvoit* → *pouvait*, *seroit* → *serait*), ou dans la notation des voyelles longues (*ſ* diacritique dans *bien-toſt* remplacé par un accent circonflexe dans *bientôt*).

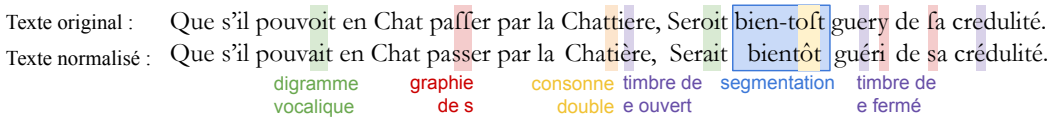

FIGURE 1 – Deux vers tirés de la fable « Le Chat, & le Grillon » de Marie-Catherine de Villedieu (*Fables, ou Histoires allégoriques, dédiées au Roy*, Paris : Claude Barbin, 1670, p. 45).

Cette tâche a un double intérêt : d'une part l'amélioration des résultats d'autres tâches de TAL par l'application de modèles et outils adaptés au français contemporain sur le texte normalisé et d'autre part l'édition de textes anciens, pour laquelle une normalisation manuelle partielle est parfois réalisée et où un outil facilitant cette normalisation manuelle ne peut être que le bienvenu. Des outils de normalisation automatique ont été développés depuis les années 80 (Fix, 1980) pour d'autres langues (Scherrer & Erjavec, 2013; Bollmann & Søgaard, 2016; Hämäläinen *et al.*, 2018) et un peu plus récemment pour le français (Riguet, 2019; Gabay *et al.*, 2019; Gabay & Barrault, 2020). Ils s'appuient sur des approches à base de règles (Baron & Rayson, 2009; Porta *et al.*, 2013; Riguet, 2019), sur des distances d'édition et des ressources linguistiques (Mitankin *et al.*, 2014) ou encore sur des approches de traduction automatique (TA) statistique (Scherrer & Erjavec, 2013; Domingo & Casacuberta, 2018) ou neuronale (Bollmann & Søgaard, 2016; Hämäläinen *et al.*, 2018).

Pour l'évaluation de telles approches, un corpus, PARALLEL17, avait été développé (Gabay *et al.*, 2019; Gabay & Barrault, 2020). Il a depuis été complété et partitionné en sous-corpus d'entraînement (17 930 phrases), développement (2 433 phrases) et test (4 706 phrases), en prenant soin d'inclure dans les deux derniers des textes de genres littéraires différents de ceux présents dans le corpus d'apprentissage, portant par exemple sur la médecine ou la physique. Les choix de partitionnement, effectués en fonction des genres littéraires des textes, s'appuient sur des métadonnées associées au corpus. Suite à cette importante refonte, PARALLEL17 prend désormais la forme d'un *benchmark* nommé FREEM$_{norm}$ (Gabay, 2022; Bawden *et al.*, 2022), réutilisable par d'autres équipes souhaitant proposer de nouvelles approches de normalisation du français d'Ancien Régime.

Nous proposons deux nouvelles approches pour cette tâche. La première, appelée ABA, pour *Alignment-Based Approach*, combine l'application d'un lexique de traduction appris automatiquement par alignement au niveau des mots sur un corpus d'entraînement, avec l'application de règles conçues manuellement mais en tenant compte de l'alignement de ce même corpus au niveau des caractères. La seconde est fondée sur l'apprentissage automatique, et plus précisément sur des modèles de TA statistique et neuronale entraînées sur ce même corpus (Bawden *et al.*, 2022), en retenant le modèle conduisant aux meilleures performances sur le corpus de développement. De telles approches peuvent être suivies d'un post-traitement s'appuyant sur un lexique, dans notre cas le lexique Le*fff* (Sagot, 2010), approche qui consiste à remplacer tout mot inconnu du lexique par un mot connu qui n'en diffère que par certaines correspondances prédéfinies (p. ex les caractères accentués vs. non accentués), dès lors qu'il n'existe qu'un seul mot connu qui satisfait cette contrainte.

Avec une évaluation sur la précision de mots sur le jeu de test de FREEM$_{norm}$, nous atteignons des scores de 94,70% avec la méthode ABA, contre 96,83%, 96,13% et 96,03% pour les trois méthodes

Modèle	Précision (%)	Précision OOV (%)
Fonction d'identité	72,40	41,89
ABA	94.70	67,51
SMT	**96,83**±0,06	73,41±0,16
LSTM	96,13±0,06	**74,52**±0,64
Transformer	96,03±0,06	74,02±0,77
Fonction d'identité + Le*fff*	85,80	63,08
ABA + Le*fff*	95,00	71,54
SMT + Le*fff*	**97,00**±0,00	**76,04**±0,18
LSTM + Le*fff*	96,27±0,06	**76,29**±0,29
Transformer + Le*fff*	96,17±0,06	75,56±0,32

TABLE 2 – Précision au niveau des mots (%) avec et sans post-traitement *Précision OOV* représente la précision uniquement sur les mots n'apparaissant pas dans le corpus d'entraînement.

de TA testées : SMT (TA statistique (Koehn *et al.*, 2007)), LSTM (TA neuronale de type LSTM (Bahdanau *et al.*, 2015)) et Transformer (TA neuronale de type transformer (Vaswani *et al.*, 2017)) respectivement [10]. Toutes les méthodes bénéficient de l'ajout de l'étape de post-traitement, surtout les modèles statistiques (ABA et SMT), qui ont tendance à être plus conservateurs que les modèles neuronaux, illustrés par la précision plus basse sur les mots dits OOV (*out of vocabulary* ; ceux qui n'apparaissent pas dans les données d'entrainement). Le meilleur modèle final est donc le modèle de TA statistique, qui n'est pas l'état de l'art en TA, mais qui est parfois performant lorsque le volume des données d'entrainement est petit (Fourrier *et al.*, 2021).

5 Conclusion

Nos premiers résultats sont extrêmement encourageants, mais appellent encore un important travail. Ainsi, l'utilisation d'un modèle transformer doit être plus abondamment testée pour les tâches envisagées par notre projet (comme la normalisation), et la granularité de l'annotation n'a pas toujours été exploitée à son maximum (comme pour la REN). Le travail accompli permet néanmoins d'ouvrir de nouvelles possibilités pour l'étude de la langue française et de son évolution, mais aussi de la littérature. Il devient possible, en utilisant des grandes masses textuelles, d'interroger géographiquement les textes (Kogkitsidou & Gambette, 2020; Gabay *et al.*, 2021), ou encore de comparer les versions originales et normalisées des sources pour en étudier la graphie (Gabay *et al.*, 2022a). De telles considérations relevant moins du TAL que d'autres disciplines, comme la philologie, nous reviendrons ailleurs sur ces pistes.

Données et modèles

Les données et modèles distribués par le projet FREEM sont disponibles à l'adresse `https://freem-corpora.github.io`.

10. L'entrainement se fait avec Moses (Koehn *et al.*, 2007) pour SMT et Fairseq (Ott *et al.*, 2019) pour LSTM et Transformer. Pour les trois, les textes sont segmentés en sous-mots (Sennrich *et al.*, 2016) avec SentencePiece (Kudo & Richardson, 2018).

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). *French TreeBank (FTB)*. Université Paris-Diderot.

ACADÉMIE FRANÇAISE (1694). *Dictionnaire de l'Académie françoise*. Veuve Jean-Baptiste Coignard et Jean-Baptiste Coignard.

AMATUZZI A., AYRES-BENNETT W., GERSTENBERG A., SCHØSLER L. & SKUPIEN-DEKENS C. (2020). Changement linguistique et périodisation du français (pré)classique : deux études de cas à partir des corpus du RCFC. *Journal of French Language Studies*, p. 1–26. Publisher : Cambridge University Press, DOI : 10.1017/S0959269520000058.

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

BADIOU-MONFERRAN C. (2011). Le « francais préclassique » et l'early modern french. *Diachroniques - Revue de linguistique française diachronique*, **1**, 83–109.

BAHDANAU D., CHO K. & BENGIO Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*, San Diego, CA, USA.

BARON A. & RAYSON P. (2009). Automatic standardisation of texts containing spelling variation : How much training data do you need ? In *Proceedings of the Corpus Linguistics Conference : CL2009*, University of Liverpool, UK.

BAWDEN R., POINHOS J., KOGKITSIDOU E., GAMBETTE P., SAGOT B. & GABAY S. (2022). Automatic Normalisation of Early Modern French. In *Proceedings of the 13th Language Resources and Evaluation Conference*, Marseille, France. working paper or preprint, HAL : hal-03540226.

BLUMENTHAL P., DIWERSY S., FALAISE A., LAY M.-H., SOUVAY G. & VIGIER D. (2017). Presto, un corpus diachronique pour le français des XVIe-XXe siècles. In *Actes de la 24ème conférence sur le Traitement Automatique des Langues Naturelles - TALN'17* : Association pour le traitement automatique des langues. HAL : halshs-01585010.

BOLLMANN M. & SØGAARD A. (2016). Improving historical spelling normalization with bi-directional LSTMs and multi-task learning. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, p. 131–139, Osaka, Japan.

BRUNOT F. (1905). *Histoire de la langue française, des origines à 1900*. Armand Colin.

CAMPS J.-B., CLÉRICE T., DUVAL F., ING L., KANAOKA N. & PINCHE A. (2021). Corpus and models for lemmatisation and POS-tagging of old french. HAL : halshs-03353125.

CAMPS J.-B., GABAY S., FIÈVRE P., CLÉRICE T. & CAFIERO F. (2020). Corpus and models for lemmatisation and POS-tagging of classical french theatre. *Journal of Data Mining & Digital Humanities*.

COMBETTES B. & MARCHELLO-NIZIA C. (2010). La périodisation en linguistique historique : le cas du français préclassique. In *Le Changement en français. Études de linguistique diachronique*, p. 129–142. Peter Lang.

DIWERSY S., FALAISE A., LAY M.-H. & SOUVAY G. (2017). Ressources et méthodes pour l'analyse diachronique. *Langages*, **N° 206**(2), 21–44. DOI : 10.3917/lang.206.0021.

DOMINGO M. & CASACUBERTA F. (2018). A Machine Translation Approach for Modernizing Historical Documents Using Back Translation. In *Proceedings of the 15th International Workshop on Spoken Language Translation (IWSLT 2018)*, p. 38–47, Bruges, Belgium.

EHRMANN M., HAMDI A., PONTES E. L., ROMANELLO M. & DOUCET A. (2021). Named entity recognition and classification on historical documents : A survey. *arXiv :2109.11406 [cs]*.

EHRMANN M., ROMANELLO M., FLÜCKIGER A. & CLEMATIDE S. (2020). Overview of CLEF HIPE 2020 : Named entity recognition and linking on historical newspapers. *Experimental IR meets multilinguality, multimodality, and interaction. 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings*. DOI : 10.1007/978-3-030-58219-7_21.

FIX H. (1980). *Automatische Normalisierung - Vorarbeit zur Lemmatisierung eines diplomatischen altisländischen Textes*, In *Teil 3 Beiträge zum dritten Symposion Tübingen 17. - 19. Februar 1977*, p. 92–100. Max Niemeyer Verlag. DOI : doi :10.1515/9783111438788.92.

FOURRIER C., BAWDEN R. & SAGOT B. (2021). Can cognate prediction be modelled as a low-resource machine translation task ? In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 847–861, Online. DOI : 10.18653/v1/2021.findings-acl.75.

GABAY S. (2014). Pourquoi moderniser l'orthographe ? principes d'ecdotique et littérature du XVIIe siècle. *Vox Romanica*, **73**(1), 27–42. HAL : hal-01903832.

GABAY S. (2022). FreEM-corpora/FreEMnorm : FreEM norm Parallel corpus. DOI : 10.5281/zenodo.5865428.

GABAY S. & BARRAULT L. (2020). Traduction automatique pour la normalisation du français du XVIIe siècle. In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, p. 213–222, Nancy, France. HAL : hal-02784770.

GABAY S., BAWDEN R., GAMBETTE P., POINHOS J., KOGKITSIDOU E. & SAGOT B. (2022a). Le changement linguistique au XVIIe s. : nouvelles approches scriptométriques. In *Actes du 8e Congrès Mondial de Linguistique Française*, Orléans, France.

GABAY S., CAMPS J.-B. & CLÉRICE T. (2020). Manuel d'annotation linguistique pour le français moderne (XVIe -XVIIIe siècles). HAL : hal-02571190.

GABAY S., ORTIZ SUAREZ P., BARTZ A., CHAGUÉ A., BAWDEN R., GAMBETTE P. & SAGOT B. (2022b). From FreEM to D'AlemBERT. In *Proceedings of the 13th Language Resources and Evaluation Conference*, Marseille, France : European Language Resources Association. 8 pages, 2 figures, 4 tables, HAL : hal-03596653.

GABAY S., ORTIZ SUÁREZ P. J. & VITALI G. P. (2021). Chercher la frontière : le théâtre classique en cartes. In *Chercher la frontière*.

GABAY S., PINCHE A., CARBONI N., CAMPS J.-B., BARTZ A., VIACCOZ C. & JOLIVET V. (2022c). Towards the fourth paradigm : From digital facsimiles of historical documents to highly annotated data. In *à paraître*.

GABAY S., RIGUET M. & BARRAULT L. (2019). A workflow for on the fly normalisation of 17th c. French. In *Proceedings of the 2019 Digital Humanities Conference*, Utrecht, Netherlands. HAL : hal-02276150.

GALLIANO S., GRAVIER G. & CHAUBARD L. (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Interspeech 2009 - Proceedings of the 10th Annual Conference of the International Speech Communication Association*, p. 2583–2586.

GRAVE E., BOJANOWSKI P., GUPTA P., JOULIN A. & MIKOLOV T. (2018). Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan : European Language Resources Association (ELRA).

GREIMAS A. J. & KEANE T. M. (2007). *Moyen français : la langue de la Renaissance de 1340 à 1611*. Grand dictionnaire. Larousse.

GWERDER Y. (2017). *Named Entity Recognition in Digitized Historical Texts*. University of Zurich.

HÄMÄLÄINEN M., SÄILY T., RUETER J., TIEDEMANN J. & MÄKELÄ E. (2018). Normalizing early English letters to present-day English spelling. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, p. 87–96, Santa Fe, New Mexico.

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R., DYER C., BOJAR O., CONSTANTIN A. & HERBST E. (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*, p. 177–180, Prague, République tchèque.

KOGKITSIDOU E. & GAMBETTE P. (2020). Normalisation of 16th and 17th century texts in French and geographical named entity recognition. In *Proceedings of the 4th ACM SIGSPATIAL Workshop on Geospatial Humanities*, p. 28–34, Seattle, Washington, USA.

KUDO T. & RICHARDSON J. (2018). SentencePiece : A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 66–71, Brussels, Belgium. DOI : 10.18653/v1/D18-2012.

KUNSTMANN P. & STEIN A., Éds. (2007). *Le Nouveau Corpus d'Amsterdam : actes de l'atelier de Lauterbad, 23-26 février 2006*. Volume 34 de Zeitschrift für französische Sprache und Literatur. F. Steiner.

LAMPLE G., BALLESTEROS M., SUBRAMANIAN S., KAWAKAMI K. & DYER C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 260–270, San Diego, California : Association for Computational Linguistics. DOI : 10.18653/v1/N16-1030.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

MARTIN R. (1998–). *Dictionnaire du Moyen Français*. ATILF - CNRS & Université de Lorraine.

MAUREL D., FRIBURGER N. & ESHKOL-TARAVELLA I. (2014). Enrichment of Renaissance texts with proper names. *INFOtheca : Journal of Information and Library Science*, **15**(1), 15–27. HAL : hal-01174733.

MITANKIN P., GERDJIKOV S. & MIHOV S. (2014). An approach to unsupervised historical text normalisation. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, DATeCH '14, p. 29–34, Madrid, Spain. DOI : 10.1145/2595188.2595191.

MÉZERAY F. E. D. (1951). *Observations sur l'orthographe de la langue françoise : transcriptions, commentaire et fac-similé du manuscrit de Mézeray, 1673, et des critiques des commissaires de l'Académie, précédés d'une histoire de la gestation de la 1e éd. du dictionnaire de l'Académie françoise (1639-1694)*. Volume 198 de Bibliothèque de l'Ecole des Hautes Etudes. Champion. Charles Beaulieux (éd.).

NEUDECKER C. (2016). An open corpus for named entity recognition in historic newspapers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 4348–4352 : European Language Resources Association (ELRA).

ORTIZ SUÁREZ P. J., DUPONT Y., MULLER B., ROMARY L. & SAGOT B. (2020). Establishing a new state-of-the-art for French named entity recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, p. 4631–4638, Marseille, France : European Language Resources Association.

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

PELLAT J.-C. & ANDRIEUX-REIX N. (2006). Histoire d'É ou de la variation des usages graphiques à la différenciation réglée. *Langue française*, **151**(3), 7–24. DOI : 10.3917/lf.151.0007.

PORTA J., SANCHO J.-L. & GOMEZ J. (2013). Edit Transducers for Spelling Variation in Old Spanish. In *Proceedings of the Workshop on Computational Historical Linguistics (NoDaLiDa 2013)*, p. 70–79, Oslo, Norway.

PRÉVOST S. & STEIN A. (2013). *Syntactic Reference Corpus of Medieval French (SRCMF)*. Lyon/Stuttgart : ENS de Lyon ; Lattice, Paris ; ILR University of Stuttgart.

REGNAULT M., PRÉVOST S. & VILLEMONTE DE LA CLERGERIE É. (2019). Challenges of language change and variation : towards an extended treebank of Medieval French. In *TLT 2019 - 18th International Workshop on Treebanks and Linguistic Theories*, Paris, France. HAL : hal-02272560.

RIGUET M. (2019). Normalisa, Script à base de règles pour normaliser les textes français du XVI^{e} au XIX^{e} siècle. https://github.com/mriguet/Normalisa/.

ROSSET S., GROUIN C. & ZWEIGENBAUM P. (2011). *Entités nommées structurées : guide d'annotation Quaero*. LIMSI-CNRS.

SAGOT B. (2010). The Le*fff*, a Freely Available and Large-coverage Morphological and Syntactic Lexicon for French. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. HAL : inria-00521242.

SCHERRER Y. & ERJAVEC T. (2013). Modernizing historical Slovene words with character-based SMT. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, p. 58–62, Sofia, Bulgaria.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the Conference on New Methods in Language Processing, Manchester, UK*. DOI : 10.1.1.28.1139.

SENNRICH R., HADDOW B. & BIRCH A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1715–1725, Berlin, Germany. DOI : 10.18653/v1/P16-1162.

STEIN A. (2016). Old French dependency parsing : Results of two parsers analysed from a linguistic point of view. In *Proceedings of the Tenth International Conference on Language Resources and*

Evaluation (LREC'16), p. 707–713, Portorož, Slovenia : European Language Resources Association (ELRA).

STEIN A. (s.d.). Parameters for Old French. https://sites.google.com/site/achimstein/research/resources.

UETANI T., PORTE G., BREUIL S. & DUBOC M. (2016). The BVH in Tours : digital library of image, text and data. In *TEI Conference 2016*, Vienne, Austria : TEI Consortium. HAL : hal-01459324.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. U. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, **30**, 5998–6008.