Traduction automatique pour la normalisation du français du XVII e siècle

Simon Gabay, Loïc Barrault

Traduction automatique pour la normalisation du français du XVII^{e} siècle

Simon Gabay[1] Loïc Barrault[2]
(1) Université de Neuchâtel, Espace Louis-Agassiz 1, 2000 Neuchâtel, Suisse
(2) University of Sheffield, 211 Portobello, Sheffield S1 4DP, Royaume-Uni
prenom.nom@unine.ch, initiale.nom@sheffield.ac.uk

RÉSUMÉ

L'étude des états de langue anciens se heurte à un double problème : d'une part la distance d'avec l'orthographe actuelle, qui empêche de recourir aux solutions standards de TAL, et d'autre part l'instabilité des systèmes graphiques, qui complexifie l'entraînement de solutions directement sur le texte original. Reprenant ce problème d'un point de vue des humanités numériques, nous partons du raisonnement philologique qui sous-tend la création du corpus d'entraînement, avant de recourir aux méthodes traditionnelles de TAL pour comparer deux systèmes de traduction automatique (statistique et neuronale) et offrir un outil fonctionnel pour la normalisation du français classique qui corresponde aux besoins des philologues.

ABSTRACT

Machine Translation for the Normalisation of 17th c. French

The study of old state of languages is facing a double problem : on the one hand the distance with contemporary spelling prevents scholars from using standard NLP solutions, on the other hand the instability of the *scriptae* complexifies the training of solutions directly on the original source text. Returning to this problem with a DH perspective, we start with the philological reasoning behind the creation of the training corpus, and use traditional NLP methods to compare two machine translation systems (statistical and neural) and offer a functional tool for the normalisation of 17th c. French answering the needs of philologists.

MOTS-CLÉS : Normalisation, XVII^{e} siècle, traduction automatique neuronale, traduction automatique statistique, humanités numériques.

KEYWORDS : Normalisation, 17th c. French, Neural Machine Translation (NMT), Statistical Machine Translation (SMT), digital humanities.

1 Introduction

Le français pré-orthographique est un problème pour deux raisons. D'une part, même si on peut le déplorer (Gabay, 2014), les lecteurs ont pris l'habitude de lire la langue de Molière avec une orthographe contemporaine. D'autre part, la variation graphique complique l'approche computationnelle de la langue : elle altère le rapport token/type (désormais TTR) pour les études stylométriques (Kestemont, 2012 ; Pinche *et al.*, 2019), complique la lemmatisation (Manjavacas *et al.*, 2019) et l'extraction d'information (Pettersson, 2016)… Pour toutes ces raisons, plusieurs solutions ont été testées afin de normaliser les données : les systèmes à base de règles (Bollmann *et al.*, 2011), la traduction

automatique statistique (*Statistical Machine Translation*, désormais SMT) (Pettersson *et al.*, 2013 ; Sánchez-Martínez *et al.*, 2013 ; Scherrer & Erjavec, 2013) ou, plus récemment, la traduction automatique neuronale (*Neural Machine Translation*, désormais NMT)(Bollmann & Søgaard, 2016).

De nombreuses évaluations ont montré l'efficacité des deux dernières solutions (Pettersson *et al.*, 2014 ; Bollmann, 2019), dont le défaut est cependant bien connu : la quantité de données nécessaires pour l'entraînement d'un modèle. De telles données existent pour de nombreuses langues (anglais, allemand, hongrois, islandais, portugais, slovène, espagnol ou suédois), mais pas pour le français. Nous nous proposons donc de reprendre l'approche comparatiste de Bollmann & Søgaard (2016) mais, contrairement à ces derniers, nous travaillons sur la langue française classique, et surtout nous n'opérons pas sur des mots isolés pris dans un dictionnaire mais sur des phrases entières. Il nous est ainsi possible de traiter les mots en contexte, et donc de résoudre des ambiguïtés où des homographes auraient différentes versions modernisées, mais aussi de fournir une évaluation en condition réelle d'utilisation.

Afin de mener à bien cette tâche, nous avons décidé de créer un corpus parallèle français classique/français contemporain pour entraîner un normaliseur. Ce corpus, en constante évolution, a été pensé pour être aussi représentatif que possible, et tente de couvrir le XVII^{e} siècle dans toute sa diversité. Après une présentation détaillée des enjeux philologiques soulevés par la question de la normalisation, nous explicitons la démarche ayant mené à la création d'un modèle immédiatement fonctionnel pour les chercheurs en humanités numériques spécialisés dans la philologie.

2 Normalisation

Il existe une multitude de possibilités pour normaliser un état de langue pré-orthographique : transcription interprétative (Stutzmann, 2011), canonicalisation phonétique (Jurish, 2011), lemmatisation et étiquetage morpho-syntaxique, alignement avec le français contemporain… Chacune de ces solutions a ses avantages et ses inconvénients, mais nous nous concentrerons ici, pour les raisons évoquées au tout début de cet article, sur la dernière solution.

En pratique, normaliser le français classique implique la correction de plusieurs caractéristiques : la segmentation (*sur tout* → *surtout*), le trait d'union (*long-temps* → *longtemps*), les lettres calligraphiques (*j'ay* → *j'ai*), les archaïsmes alphabétiques (*fecours* → *secours*), les accents (*Ame* → *Âme*), les lettres étymologiques (lat. *Voster* → frm. *voſtre* →*votre*), les lettres ramistes (*vn* → *un*, *Ie* → *Je*), les changements phonétiques (*craignois* → *craignais*)…

Source	Cible
Sur tout ie redout**o**is cette Mélancolie	**Surtout j**e redout**a**is cette Mélancolie
Où j'a**y** v**eu f**i long**-**temps vo**f**tre **A**me enseuelie.	Où j'a**i** v**u s**i longtemps votre **Â**me ensevelie.
Ie craign**o**is que le Ciel, par **v**n cruel ſecours,	**J**e craign**a**is que le Ciel, par **u**n cruel secours,

TABLE 1 – Exemple de normalisation

Il est important de noter que, concernant la normalisation de sources littéraires, il ne peut y avoir un alignement complet avec le français contemporain. Certaines graphies affectant la prosodie ne peuvent être changées, comme *jusques*+voyelle (aujourd'hui *jusqu'*+voyelle) pour conserver la versification (par. exp. "Portez-vous, s'il le faut, juſques à le haïr." où "juſques à" nécessite trois voyelles : [ʒys-kə-za]). Le problème est identique pour *encor(e)*.

Ajoutons que la liste des opérations nécessaires à la normalisation présentée *supra*, comme la fréquence des types de variation dépendent fortement de l'extension diachronique et diatopique des données d'entraînement. L'hétérogénéité du corpus source grandit donc exponentiellement avec l'extension d'un corpus censé être représentatif d'un état de langue.

3 Corpus

La construction d'un outil fonctionnel nécessite un travail important sur la création du corpus. Puisque l'efficacité du normaliseur dépend des données d'entraînement, elles doivent être soigneusement choisies afin de couvrir la langue classique dans toute sa diversité. Trois points ont tout particulièrement attiré notre attention :

— L'extension du lexique est extrêmement importante. Si on peut difficilement espérer traiter les textes littéraires, médicaux, théologiques… avec la même efficacité, il est possible d'éviter de grosses déconvenues en veillant à ce que ces différents genres, et donc leur lexique, soient représentés dans les données d'entraînement.
— Le système graphique est instable tout le XVII^{e} siècle et cette variation évolue dans le temps (Biedermann-Pasques, 2017) : en une centaine d'année, le français passe rapidement d'une multitude de *scriptae* [1] à un état quasi-stable (Bonhomme, 2011). En plus de la variation diachronique, on note une importante polarisation diastratique entre les Anciens et les Modernes, qui défendent des options graphiques totalement opposées (Pellat, 1995b).
— L'utilisation des majuscules est extrêmement différente de la nôtre et ne semble pas suivre de règle claire. S'il est possible de contourner ce problème en abaissant la casse de tous les caractères, agir ainsi nous fait perdre de la lisibilité et de l'information.

Concernant le lexique, nous avons décidé de créer un corpus primaire et un corpus secondaire. Le corpus primaire se concentre sur les textes littéraires avec la poésie (Viau), le roman (La Fayette), la tragédie (Racine), la comédie (Molière), la correspondance (Guez de Balzac). Le corpus secondaire étend le lexique en ajoutant des documents traitant de physique (Pascal), de médecine (Ellain), de théologie (Sales, Bossuet), de philosophie (Descartes).

Parmi tous les textes, nous avons délibérément augmenté le nombre de pièces de théâtre pour deux raisons. La première est que les textes dramatiques utilisent plus la majuscule (acte, scène, tours de parole) que les autres genres. La seconde est que le théâtre est l'un des genres les mieux connus en ce qui concerne l'histoire des livres au XVII^{e} siècle (Riffaud, 2009) et que l'on connaît donc le nom de l'imprimeur pour chaque livre – une information importante pour équilibrer au mieux le corpus.

Concernant les systèmes graphiques, les textes de ce double corpus sont distribués chronologiquement, avec au moins deux imprimés par décennie. Si la plupart sont imprimés à Paris, certains proviennent des Flandres (Bussy-Rabutin à Bruxelles) ou de Hollande (Descartes à Leyde), où le système graphique est différent (Pellat, 1995a). Cette distribution chronologique et géographique n'est pas parfaite car elle dépend de documents OCRisés avec un modèle en cours de construction (Gabay, 2019).

1. Une *scripta* est une koinè graphique, *i.e.* une langue écrite partagée par un large groupe de scripteurs d'une même langue.

4 Entraîner et évaluer les modèles

La création de données d'entraînement a été faite avec l'aide d'un système à base de règles et d'expressions régulières pour accélérer le processus (Gabay *et al.*, 2019) : chaque phrase est pré-normalisée automatiquement avant d'être manuellement corrigée pour garantir la qualité des résultats. Ainsi, un corpus de référence de c. 140 000 tokens a pu être produit rapidement pour effectuer des premiers tests.

Deux formats de sortie sont disponibles : tsv, mais aussi XML-TMX. Ce dernier permet de conserver les données dans un format aisément manipulable, mais aussi de réutiliser les données avec de logiciels acceptant des mémoires de traduction comme MateCat (Federico *et al.*, 2014) ou des outils de textométrie comme TXM (Heiden *et al.*, 2010). Pour l'entraînement, chaque texte a été segmenté en lignes, constituées dans la majorité des cas de phrases, mais aussi de propositions dans le cas où les phrases se sont avérées trop longues (plus de soixante mots). 10% des lignes de chaque texte a été prélevé aléatoirement pour évaluer le modèle, et le reste a été utilisé pour l'entraînement.

	Lignes	Tokens	Caractères
Train	8 962	128 146	571 290
Test	814	9 068	40 253
Total	9 776	137 342	611 543

TABLE 2 – Constitution du corpus

Tous les résultats sont évalués avec l'exactitude de mots (*word accuracy* ou *Wacc*) selon les recommandations de M. Bollmann (Bollmann, 2018), mais nous fournissons aussi les scores BLEU (Papineni *et al.*, 2002) et METEOR (Denkowski & Lavie, 2014) car ces mesures sont toujours utilisées par d'autres collègues (Domingo & Casacuberta, 2018b) et qu'elles permettent donc des comparaisons avec d'autres publications. Il doit aussi être noté que, dans notre corpus d'entraînement, nous avons délibérément augmenté l'hétérogénéité des données, ce qui ne peut qu'abaisser les scores finaux.

Nous avons décidé de comparer les traductions automatiques neuronale et statistique de niveau caractère (*character level Statistical Machine Translation*, désormais cSMT).

4.1 cSMT

L'efficacité de la traduction statistique de niveau caractère ayant été plusieurs fois démontrée (Ljubešić *et al.*, 2016), nous avons décidé de ne pas tester différents degrés de granularité (mot, sous-mot, caractère…) comme avec la NMT (cf. *infra*) mais de tenter l'ajout de données avec un modèle de langue (Scherrer & Ljubešić, 2016) et la rétro-traduction (Domingo & Casacuberta, 2018a). Concernant les modèles de langue, nous avons utilisé deux corpus différents :

— Un premier constitué de 3 151 778 tokens tirés de textes classiques normalisés proches de ceux des données d'entraînement : du théâtre (Molière, Racine, Corneille), des recueils (*Nouvelles nouvelles* de Donneau de Visé), des romans (*Histoire amoureuse* de Bussy), des nouvelles (*Historiettes* de Tallemant des Réaux), des sermons (Bossuet), des essais philosophiques (les œuvres complètes de Descartes et de Pascal), de la correspondance (Sévigné, La Fayette) et des textes religieux (La Bible de Lemaître de Sacy, Saint Augustin et Saint Thomas) pour anticiper de possibles références religieuses.

— Un second corpus fait de 89 972 791 tokens en ajoutant au premier corpus la totalité des données en français du projet Gutenberg.

Nous avons aussi essayé la rétro-traduction : les données normalisées deviennent la source et les données originales la cible. Un nouveau jeu de données de 121 012 tokens (8 905 lignes) composées de pièces (Corneille), d'essais (La Rochefoucauld), de sermons (Bossuet) et textes variés (Donneau de Visé) ont ainsi été convertis en textes pseudo-anciens. Pour ce faire un troisième modèle de langue a été créé avec des transcriptions non-normalisées provenant de wikisource (La Fayette, Racine, La Fontaine, Tristan l'Hermite) et de deux grosses éditions publiées en ligne (*L'Astrée* d'Honoré D'Urfé et *Artamène* des Scudéry).

	BLEU (4-grammes)	METEOR	wAcc
Normal	77,667	87,891	86.68496
+ML	77,108	87,308	86.4022
+trad. inv.	76,680	87,024	86.06121
+trad. inv.+ML	75,766	86,257	85.57052

TABLE 3 – Evaluation des modèles avec cSMT

Les résultats sont bons, mais il semble que les différentes techniques utilisées pour améliorer les scores aient un effet neutre, voire négatif sur le score final.

4.2 NMT

En ce qui concerne la traduction automatique neuronale, nous avons décidé d'utiliser NMTPy-Torch (Caglayan *et al.*, 2017). Le modèle de base est composé d'un encodeur GRU bidirectionnel (Cho *et al.*, 2014) et d'un décodeur (GRU conditionnel à deux couches (Sennrich *et al.*, 2017)) avec mécanisme d'attention de type perceptron multicouche (Bahdanau *et al.*, 2015). L'encodeur et le décodeur ont tous les deux 400 unités cachées et leur état initial caché est initialisé à zéro. Les plongements lexicaux ont une taille fixée à 200.

Trois versions du système ont été entraînées : une première au niveau mot, une seconde avec des unités *byte pair encoding* ou *BPE* (Sennrich *et al.*, 2015) opérant au niveau du sous-mot, et une troisième au niveau du caractère. Le tableau 4 montre le résultats de tels pré-traitements. Chaque système a été entraîné quatre fois avec une initialisation différente, deux fois avec *drop out*, deux fois sans. Seuls les modèles avec les meilleurs scores BLEUS sont conservés.

Les résultats sont particulièrement encourageants, car ils améliorent de précédents tests (BLEU de 82.96) malgré un doublement du nombre de données d'entraînement (c. 60 000 vs 140 000 tokens) et une nette augmentation de leur diversité (2 imprimés vs 30) (Gabay *et al.*, 2019).

Granularité	Version	Phrase
Mot	Source	Cherchons avec empreſſement
	Target	Cherchons avec empressement
BPE	Source	Ch@@ er@@ ch@@ ons avec em@@ preſ@@ ſement
	Target	Ch@@ er@@ ch@@ ons avec em@@ pr@@ ess@@ ement
Caractères	Source	C h e r c h o n s • a v e c • e m p r e ſ ſ e m e n t
	Target	C h e r c h o n s • a v e c • e m p r e s s e m e n t

TABLE 4 – Exemple de phrase pré-traitée à différentes granularités. '@@' après une unité sous-mot spécifie que l'unité ne termine pas le mot. '•' représente un espace du texte initial pour le traitement au niveau des caractères.

	BLEU (4-grammes)	METEOR	wAcc
Word	83,647	90,783	91.41786
BPE	60,200	71,836	71.11808
characters	71,283	76,816	77.86646

TABLE 5 – Évaluation des modèles avec NMT

5 Analyse des résultats

En plus des scores précédemment présentés, revenons concrètement sur le type de résultat produit par chacun des systèmes. Malgré des scores similaires, on peut noter quelques différences assez importantes :

Version	Exemple
Original	En cet eſtat **la**, Monſeigneur, il n'y a point d'apparence de ſonger à la Valtoline, ny de ietter les yeux ſur le bien d'autruy cependant qu'on nous **diſpute** le noſtre, & qu'il faut que nous le **tenions** auecque les deux mains de peur qu'il ne nous **eſchappe**.
NMT	en cet état **la**, Monseigneur, il n' y a point d'apparence de songer à la Valteline, ni de jeter les yeux sur le bien d'autrui cependant qu'on nous **dispute** le nôtre, et qu' il faut que nous le **commun** avec les deux mains de peur qu'il ne nous **apportée**.
cSMT	en cet état, monseigneur, il n'y a point d'apparence de songer à la valteline, ni de jeter les yeux sur le bien d'autrui cependant qu'on nous **disputer** le nôtre, et qu'il faut que nous le **contentions** avec les deux mains de peur qu'il ne nous **échappe**.

TABLE 6 – Comparaison des résultats

Phénomème attendu, les deux systèmes commettent des fautes sensiblement sur les mêmes passages. Les erreurs de NMTPyTorch (*commun*) sont néanmoins beaucoup plus éloignées de l'original (*tenions*) que celles de cSMTiser (*contentions*), dont les bévues, plus nombreuses, sont souvent peu conséquentes (*dispute* vs *disputer*). Ce dernier système a en revanche tendance à "oublier" (plus rarement "rajouter") des mots (*la*).

Ajoutons que le passage du français classique au français contemporain réduit la richesse lexicale (passage d'un TTR de 0.11 à 0.092) : la question des homographes est donc peu importante. On remarque cependant dans cet exemple que les deux systèmes ont su, comme partout ailleurs, distinguer sur la base du contexte (présence du déterminant *le*) le pronom *nôtre* de l'adjectif *notre* en dépit d'une forme unique en français classique (*noftre*).

6 Conclusion

Il est donc relativement simple d'approcher les 90% de *wAcc* avec les systèmes actuels, sans configuration spéciale. Malgré des résultats plus que satisfaisants, le cSMT est moins performant que le NMT, et cela même avec un corpus d'entraînement de petite taille. Le score de 90%, obtenu uniquement par le NMT, est proche de l'état de l'art (Bollmann, 2019) malgré des données d'entraînement

d'une grande hétérogénéité, ce qui démontre l'extrême robustesse du système.

Cette grande hétérogénéité, volontairement introduite, permet de garantir l'efficacité du modèle pour les philologues. La performance n'est en effet pas évaluée sur des listes de mots uniques, comme cela se fait souvent pour les bancs d'essai, mais sur des données proches de celle que rencontrent les philologues au quotidien.

Du point de vue informatique, les futurs recherches doivent s'orienter vers les solutions neuronales, dont l'efficacité sera clairement améliorée par les techniques que nous avons utilisées dans cet article avec le cSMT. Il conviendra donc, parallèlement à une augmentation significative des données d'entraînement, de tester la rétro-traduction et l'utilisation d'un modèle de langue pré-entraîné comme CamemBERT (Martin *et al.*, 2019) pour améliorer encore plus les résultats.

Du point de vue philologique, il serait intéressant de tenter un partitionnement des données d'entraînement afin de tester l'efficacité de modèles entraînés sur des jeux de données plus restreints, mais plus homogènes. Les cadrages chronologiques qui produiraient les modèles les plus efficaces seraient porteurs d'une information linguistique permettant de repenser la périodisation de la langue moderne sur d'autres critères que l'histoire des idées.

Du point de vue ecdotique enfin, il conviendrait de réfléchir à l'utilité d'autres types de normalisation, car il n'est pas certain que l'alignement sur l'orthographe actuelle soit le seul choix souhaitable. La création de modèles permettant une normalisation plus légère, comme les dissimilations *i* vs *j* et *u* vs *v*, permettrait de produire facilement des textes aisément lisibles qui ne perdraient pas toute leur richesse linguistique.

Remerciements

Merci à A. Baillot, organisatrice de la *Human-Machine Translation German-French Summer School*, et à Y. Scherrer pour son aide dans l'optimisation des résultats de cSMTiser.

Références

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 (ICLR 2015)*, San Diego, CA.

Biedermann-Pasques L. (2017). *Les grands courants orthographiques au XVII^e siècle et la formation de l'orthographe moderne, Impacts matériels, interférences phoniques, théories et pratiques (1606–1736)*. Berlin, Boston : De Gruyter, reprint 2017 édition. doi : 10.1515/9783110938593.

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

Bollmann M. (2019). A Large-Scale Comparison of Historical Text Normalization Systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 3885–3898, Minneapolis, Minnesota : Association for Computational Linguistics. doi : 10.18653/v1/N19-1389.

Bollmann M., Petran F. & Dipper S. (2011). Rule-Based Normalization of Historical Texts. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage (DigHum 2011)*, p. 34–42, Hissar, Bulgaria.

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 : Technical Papers*, p. 131–139, Osaka, Japan. Anthologie ACL : C16-1013.

Bonhomme M. (2011). La standardisation du français au XVII^e siècle. Le cas des observations sur la langue françoise de ménage. In *Du système linguistique aux actions langagières : Mélanges en l'honneur d'Alain Berrendonner*, Champs linguistiques. Bruxelles : De Boeck Supérieur.

Caglayan O., García-Martínez M., Bardet A., Aransa W., Bougares F. & Barrault L. (2017). NMTPY : A Flexible Toolkit for Advanced Neural Machine Translation Systems. *The Prague Bulletin of Mathematical Linguistics*, **109**(1), 15–28. arXiv : 1706.00457, doi : 10.1515/pralin-2017-0035.

Cho K., van Merrienboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H. & Bengio Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1724–1734, Doha, Qatar. arXiv : 1406.1078.

Denkowski M. & Lavie A. (2014). Meteor universal : Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Domingo M. & Casacuberta F. (2018a). 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. 39–47, Bruges, Belgium.

Domingo M. & Casacuberta F. (2018b). Spelling Normalization of Historical Documents by Using a Machine Translation Approach. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation (EAMT 2018)*, p. 129–137, Alicante, Spain.

Federico M., Bertoldi N., Cettolo M., Negri M., Turchi M., Trombetti M., Cattelan A., Farina A., Lupinetti D., Martines A., Massidda A., Schwenk H., Barrault L., Blain F., Koehn P., Buck C. & Germann U. (2014). The MateCat Tool. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics : System Demonstrations*, p. 129–132, Dublin, Ireland : Dublin City University and Association for Computational Linguistics.

Gabay S. (2014). Pourquoi moderniser l'orthographe ? Principes d'ecdotique et littérature du XVII^e siècle. *Vox Romanica*, **73**(1).

Gabay S. (2019). OCRising 17th French prints. *E-ditiones*.

Gabay S., Riguet M. & Barrault L. (2019). A Workflow For On The Fly Normalisation Of 17th c. French. In *DH2019*, Utrecht, Netherlands : ADHO.

Heiden S., Magué J.-P. & Pincemin B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. In *10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*, volume 2, p. 1021–1032, Rome, Italy : Edizioni Universitarie di Lettere Economia Diritto.

Jurish B. (2011). *Finite-state canonicalization techniques for historical German*. Thèse de doctorat, Universität Potsdam, Potsdam.

Kestemont M. (2012). Stylometry for Medieval Authorship Studies : An Application to Rhyme Words. *Digital Philology : A Journal of Medieval Cultures*, **1**(1), 42–72. doi : 10.1353/dph.2012.0002.

Ljubešić N., Zupan K., Fišer D. & Erjavec T. (2016). Normalising Slovene data : historical texts vs. user-generated content. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, p. 146–155, Bochum, Germany.

Manjavacas E., Kádár Á. & Kestemont M. (2019). Improving Lemmatization of Non-Standard Languages with Joint Learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 1493–1503, Minneapolis, Minnesota : Association for Computational Linguistics. doi : 10.18653/v1/N19-1153.

Martin L., Muller B., Ortiz Suárez P. J., Dupont Y., Romary L., Villemonte de la Clergerie É., Seddah D. & Sagot B. (2019). CamemBERT : a Tasty French Language Model. *arXiv e-prints*, p. arXiv :1911.03894. arXiv preprint : 1911.03894.

Papineni K., Roukos S., Ward T. & Zhu W.-J. (2002). BLEU : A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, p. 311–318, Philadelphia, USA. doi : 10.3115/1073083.1073135.

Pellat J.-C. (1995a). L'évolution de l'orthographe des imprimés au XVIIe s. (libraires français et hollandais). In *« Ces mots qui sont nos mots ». Mélanges d'Histoire de la Langue Française, de Dialectologie et d'Onomastique offerts au professeur Jacques Chaurand*, p. 83–96. Institut Charles Bruneau.

Pellat J.-C. (1995b). Norme et variation orthographique au XVII^{e} siècle. In *Rencontres linguistiques en pays rhénan 5/6*, volume 3 de Sciences Cognitives, Linguistiques & Intelligence Artificielle, p. 245–260, université des sciences humaines de Strasbourg : Université des sciences humaines Strasbourg.

Pettersson E. (2016). *Spelling Normalisation and Linguistic Analysis of Historical Text for Information Extraction*. Studia Linguistica Upsaliensia. Acta Universitatis Upsaliensis.

Pettersson E., Megyesi B. & Nivre J. (2013). 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)*, p. 163–179, Oslo, Norway.

Pettersson E., Megyesi B. & Nivre J. (2014). A Multilingual Evaluation of Three Spelling Normalisation Methods for Historical Text. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, p. 32–41, Gothenburg, Sweden. doi : 10.3115/v1/W14-0605.

Pinche A., Camps J.-B. & Clérice T. (2019). Stylometry for Noisy Medieval Data : Evaluating Paul Meyer's Hagiographic Hypothesis. In *Digital Humanities Conference 2019 - DH2019*, Utrecht, Netherlands : ADHO and Utrecht University.

Riffaud A. (2009). *Répertoire du théâtre français imprimé entre 1630 et 1660 - Librairie Droz*. Travaux du Grand Siècle. Genève : Droz.

Sánchez-Martínez F., Martínez-Sempere I., Ivars-Ribes X. & Carrasco R. C. (2013). An open diachronic corpus of historical spanish. *Language resources and evaluation*, **47**(4), 1327–1342. doi : 10.1007/s10579-013-9239-y.

Scherrer Y. & Erjavec T. (2013). Modernizing Historical Slovene Words with Character-Based SMT. In *4th Biennial Workshop on Balto-Slavic Natural Language Processing (BSNLP 2013)*, p. 58–62, Sofia, Bulgaria.

Scherrer Y. & Ljubešić N. (2016). Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation. In *KONVENS*.

Sennrich R., Firat O., Cho K., Birch A., Haddow B., Hitschler J., Junczys-Dowmunt M., Läubli S., Barone A. V. M., Mokry J. & Nădejde M. (2017). Nematus : a Toolkit for Neural Machine Translation. In *Proceedings of the EACL 2017 Software Demonstrations*, p. 65–68, Valencia, Spain. Anthologie ACL : E17-3017.

Sennrich R., Haddow B. & Birch A. (2015). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, p. 1715–1725, Berlin, Germany.

Stutzmann D. (2011). Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin ? *Kodikologie und Paläographie im digitalen Zeitalter 2 - Codicology and Palaeography in the Digital Age 2*, p. 247–277.