Lemmatization of Historical Old Literary Finnish Texts in Modern Orthography

Mika Hämäläinen, Niko Partanen, Khalid Alnajjar

Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles
Lille, France, 28 juin au 2 juillet 2021

Lemmatization of Historical Old Literary Finnish Texts in Modern Orthography

Mika Hämäläinen[1] Niko Partanen[2] Khalid Alnajjar[1]
(1) Department of Digital Humanities
(2) Department of Finnish, Finno-Ugrian and Scandinavian Studies
(1,2) University of Helsinki, Finland
firstname.lastname@helsinki.fi

RÉSUMÉ

Les textes écrits en vieux finnois littéraire représentent la première œuvre littéraire jamais écrite en finnois à partir du XVIe siècle. Il y a eu plusieurs projets en Finlande qui ont numérisé des anciennes collections de textes et qui les ont rendues disponibles pour la recherche. Cependant, l'utilisation de méthodes TAL modernes avec telles données pose de grands défis. Dans cet article, nous proposons une approche pour normaliser et lemmatiser simultanément des textes écrits en vieux finnois littéraire à l'orthographe moderne. Notre meilleur modèle donne une précision de 96,3 % avec les textes écrits par Agricola et de 87,7 % avec d'autres textes contemporains hors du domaine. Notre méthode est publiée gratuitement sur Zenodo et Github.

ABSTRACT

Texts written in Old Literary Finnish represent the first literary work ever written in Finnish starting from the 16th century. There have been several projects in Finland that have digitized old publications and made them available for research use. However, using modern NLP methods in such data poses great challenges. In this paper we propose an approach for simultaneously normalizing and lemmatizing Old Literary Finnish into modern spelling. Our best model reaches to 96.3% accuracy in texts written by Agricola and 87.7% accuracy in other contemporary out-of-domain text. Our method has been made freely available on Zenodo and Github.

MOTS-CLÉS : données historiques, normalisation, lemmatisation.

KEYWORDS: historical data, normalization, lemmatization.

1 Introduction

Finnish language has a long literary history starting from the 16th history. A large portion of the books printed in Finnish is currenlty openly available for research, and especially the bibliographic record they form has already been studied. (Lahti *et al.*, 2019) investigated the document size and language in the bibliographic metadata records, and (Tolonen *et al.*, 2019, 58) took into account a number of other metadata fields including the titles, and recognize the lack of full-text documents as a downside of investigation that is possible. Although the whole body of books printed in Finland is not available as full-text, there are numerous smaller corpora that can already be used in various ways.

Volume 1 : conférence principale, pages 189–198.

Both the National Library of Finland and Institute for the Languages of Finland have produced a large number of digitized material from the era of the Old Literary Finnish. Former has made a large amount of scanned and text recognized publications available (National Library of Finland, 2021), and latter has created large plain text corpora from selected works (Institute for the Languages of Finland, 2014). The Agricola corpus used in this study has been morpho-syntactically annotated, and is to our knowledge the only annotated resource in the Old Literary Finnish (Institute for the Languages of Finland & University of Turku, 2020). Another very central resources is the Dictionary of Old Literary Finnish. [1] Thus far the Old Literary Finnish has largely eluded any attempts of NLP research as the historical written form cannot be processed easily with currently available NLP tools for Finnish, as they are designed for modern Finnish. We present our approach for normalizing and lemmatizing Old Literary Finnish automatically to modern Finnish orthography. As the resources for processing historical Finnish text are scarce, we have released the models presented in this paper on Zenodo [2] and through an easy-to-use Python library [3].

The use of existing NLP tools targeted for modern Finnish on historical materials can be made possible through normalization. Previous work conducted on English data indicates that normalization is a viable way of improving the accuracy of NLP methods such as POS tagging (van der Goot *et al.*, 2017). Another direction of digital humanities study has benefited from normalization of historical data in studying the use of neologisms in old letters (Säily *et al.*, 2018; Säily *et al.*, 2021). In their approach, without normalization, they would have been able to cover only a small subset of the corpus. The same corpus has also been studied without NLP tools (Nevalainen, 2021).

The history of printed written Finnish starts from the 16th century with the works of Mikael Agricola. His primer and religious works were followed by a continuous increase in the amount of the written Finnish materials. The language form used by Agricola is known as Old Literary Finnish (*vanha kirjasuomi*). The majority of the early Finnish publications included religious materials, although the text types started to diversify already in the 18th century. The majority of the oldest texts are translations. The period of Old Literary Finnish is often estimated to have lasted until 1810, after which the written Finnish started to transform into Early Modern Finnish. Eventual changes in printing laws and printing technology, which expanded the amount and variation of the printed materials, and the creation of regular Finnish newspapers, contributed to the stabilization of the written standard.

Linguistically one of the exceptional features of Old Literary Finnish is the variation it displays. The orthography was not yet entirely established, and there was extensive spelling variation. The age of these materials adds also a historical dimension, as there are linguistic features that are not present in the modern Finnish, or exist currently only in the dialects.

Our main contributions in this work are :

— Building the first artificial neural network model for normalizing historical Finnish.
— Conducting an evaluation for assessing the performance of the model on historical data from 1) the same source of the data used in building the model and 2) external out-of-domain historical data. In both cases, a high accuracy is achieved by the model.
— Publishing a user-friendly Python library that permits an instant usability of our model.

The target language form in our work is modern Finnish. Our model primarily lemmatizes, but since the output is harmonized into contemporary orthographic forms, the work is closely connected to the normalization task as well, and the output can be considered as one type of a normalization. The exact

1. https ://kaino.kotus.fi/vks/
2. https ://zenodo.org/record/4734143
3. https ://github.com/mikahama/murre

lemmatization choices and conventions were decided at the level of the original morpho-syntactic database, and we followed those closely also when our own additional test material was created. In later research also the morpho-syntactic annotations present in the corpus could be taken into account to further enrich the analysis. However, at the current stage a successful lemmatization is already a large improvement in available NLP methods.

This paper is structured as follows. We begin by describing the related work. Thereafter, the details of the data used to build the neural model are given, followed by the architecture and hyperparameters of the neural model. Section 5 presents the results and evaluation where we explain the different training strategies we experimented with and their performance against a baseline (historical Omorfi). Lastly, we discuss and conclude our work while highlighting future directions with a potentially great impact on humanities research such as automatic analysis of historical Finnish.

2 Related Work

Historical text normalization has been studied in the past for other languages than Finnish. A recent literature review (Bollmann, 2019) finds that there are five categories in which modern normalization approaches can be divided : substitution lists like VARD (Rayson *et al.*, 2005) and Norma (Bollmann, 2012), rule-based methods (Baron & Rayson, 2008; Porta *et al.*, 2013), edit distance based approaches (Hauser & Schulz, 2007; Amoia & Martinez, 2013), statistical methods and most recently neural methods (Partanen *et al.*, 2019; Duong *et al.*, 2020).

Statistical machine translation (SMT) based methods have been the most successful ones in the past in terms of statistical methods. The key idea behind these methods is to approach the task as a character-level machine translation problem, where a word is translated character by character to its normalized form. These methods have been applied to historical text (Pettersson *et al.*, 2013; Hämäläinen *et al.*, 2018) and dialect normalization (Samardzic *et al.*, 2015).

In the recent years, normalization has been approached as a character-level neural machine translation (NMT) problem similarly to the previous SMT approaches. The additional advantage is that a neural model does not need a separate language model like SMT does. Bollmann & Søgaard (2016) have shown that a bi-directional long short-term memory (bi-LSTM) can be used to normalize historical German texts. The paper presents a so-called multi-task learning setting where auxiliary data is added to improve the performance of the model. Multi-task learning setting generally improved the results. Their system outperformed the existing conditional random fields and Norma based approaches in terms of accuracy.

Text written in Uyghur language has been normalized with an LSTM and a noisy channel model (NCM) (Tursun & Cakici, 2017). They use a relatively small set of gold annotated data for training (around 200 hand normalized social media sentences). They augment this data by synthetically generating non-normalized text by introducing random changes in normalized text. In the same fashion, another research has used an LSTM model to normalize code-mixed data (Mandal & Nanmaran, 2018).

Recently Hämäläinen *et al.* (2019) have shown that bi-directional recurrent neural networks (BRNN) outperform regular unidirectional recurrent neural networks (RNN) when normalizing historical English data. Interestingly, additional layers and different attention models do not improve the results. Additional data such as time period, social metadata or pronunciation information in IPA characters

makes the results worse. According to them, post-processing can boost the accuracy of a character level NMT model more than changing the network structure. A simple dictionary filtering method improved the results.

Omorfi (Pirinen, 2015) is a popular rule-based tool used to do morphological analysis and lemmatization of modern Finnish. While Omorfi itself is not relevant for our work, there is a GitHub fork of the project known as Historical Omorfi [4]. The fork introduces several improvements to better cater for historical Finnish text. Currently, this tool is the only tool available for lemmatizing historical Finnish. Thereby we compare the results of our model to this tool.

3 Dataset of Historical Finnish

In order to use machine learning methods, data is needed to train a model. The data needs to have text written in Old Literary Finnish and its normalized lemmas. The lemmas should be aligned on a word level with the historical data in order to train the normalization more accurately. Fortunately, such a dataset exists. The corpus we use is *The Morpho-Syntactic Database of Mikael Agricola's Works* (Institute for the Languages of Finland & University of Turku, 2020) that contains 522,237 tokens and 38,222 sentences. The corpus includes all nine Finnish books translated by Mikael Agricola. The corpus is openly available in the Language Bank of Finland. [5] In our testing we also use the Dictionary of Old Literary Finnish [6] (Institute for the Languages of Finland, 2014), from which a small number of sentences has been sampled and manually normalized and lemmatized. Both resources are available under Creative Commons licenses. Since the *The Morpho-Syntactic Database of Mikael Agricola's Works* is licensed under CC BY-ND 4.0 (CLARIN PUB), we do not redistribute the training material ourselves, but it can be accessed in the Language Bank of Finland's concordance service [7]. Naturally, since this data is so old, it is already in Public Domain, and many of the original works are entirely openly accessible. For example, the Agricola's prayer book is available as high quality scans by the National Library of Finland [8].

Although there has not been prior use of this dataset in the computational linguistics, the corpus has been used in the linguistic studies. To illustrate this with few recent examples, Toropainen (2018) investigated the compounds in this variety, and Toropainen (2015) studied nouns that contain an initial adjective. Salmi (2020) discussed recently the German influence in the Agricola's language. Also annotating the corpus into it's current stage has been a long undertaking, and Inaba (2015) investigated the use of two Finnish cases with a prototype of the current database. It is beyond doubt that the materials of Old Literary Finnish still have much to contribute to the linguistic research. We believe that by creating new tools for natural language processing of these and similar materials we can further expand toward these goals. The research concerning Agricola's language and Old Literary Finnish in general is naturally much wider and has a long research history, especially in Finland, and we primarily wanted to illustrate in this section some of the previous studies where the same database was used.

4. https://github.com/jiemakel/omorfi/
5. urn:nbn:fi:lb-2019121804
6. https://kaino.kotus.fi/vks
7. https://korp.csc.fi
8. https://www.doria.fi/handle/10024/43445

4 Neural Normalization

In this section, we describe our artificial neural network model for normalizing and lemmatizing historical Finnish into standard Finnish. We begin by explaining how the dataset is used and how it is preprocessed. Thereafter, we elucidate the architecture of the model along with the technical details of its hyperparameters.

The corpus contains nine distinct works. In order to test the model's accuracy realistically, we selected seven books into the training data, and left the remaining two into the test set. Thereby the training data was 393,779 tokens and the test set 128,294 tokens. The books in the test set were specifically 'Messu eli Herran echtolinen' from 1549 and 'Rucouskiria' from 1544. Additionally 15% of the training data was used in the validation set. We considered it important not to select the test set from the entire corpus, as this would not give a clear picture about how much the model generalizes into new use cases, which would be the other books written in the same language variety. With the current sparsity of manually annotated data, the Agricola works in the currently used corpus, but which we kept unseen for the model, were the best option we had. We also extended the testing into other examples of Old Literary Finnish, but in a smaller scale. The data has the original written form of each sentence and a token level normalized lemma for each word. We use this parallel data to train our models.

We model the problem as a character level NMT problem. In practice, we split words into characters separated by white space and mark actual spaces between words with an under score (_). This allows to pass word boundary information to the model while the characters themselves are separated by spaces. We train the model to predict from the Old Literary Finnish word forms to the lemmas. As previous research (Partanen *et al.*, 2019) has found that using chunks of words instead of full sentences at a time improves the results, we train different models with different chunk sizes. This allows direct comparison of different chunk sizes. This way we train the models to predict one word at a time, two words at a time all the way to five words at a time. An example of the data can be seen in Table 1. This example is within the sentence A-II-369 in the corpus, which is located in the New Testament translation.

Size	**Source**	**Target**
Chunk 1	s y d h e m e n	s y d ä n
Chunk 3	p a l u e l l e n _ h e r r a _ c a i k e n	p a l v e l l a _ h e r r a _ k a i k k i

TABLE 1 – An example of the training data for chunk sizes 1 and 3.

We train all models using a bi-directional long short-term memory (LSTM) based model (Hochreiter & Schmidhuber, 1997) by using OpenNMT-py (Klein *et al.*, 2017) with the default settings except for the encoder where we use a BRNN (bi-directional recurrent neural network) (Schuster & Paliwal, 1997) instead of the default RNN (recurrent neural network), since BRNN based models have been shown to provide better results in a variety of tasks. We use the default of two layers for both the encoder and the decoder and the default attention model, which is the general global attention presented by Luong *et al.* 2015. The models are trained for the default of 100,000 steps. All models are trained with the same random seed (3435) to ensure reproducibility and to make their intercomparison possible.

5 Results and Evaluation

Our initial evaluation results were very good as seen in Table 2 where the accuracies are reported on a word level. The best model was the chunk of 3. The quality we reach is very high and on par with other comparable normalization tasks, such as in (Partanen *et al.*, 2019), which also corroborate the best performance at the chunk of three tokens. However, we are also interested in seeing how well our model works with other Old Literary Finnish texts. At any rate, we can see that our models outperform Historical Omorfi, which is the only tool publicly available for historical Finnish. As Omorfi produces all the possible lemmas for a given word, we count the accuracy based on if the correct lemma is in the list of the lemmas Omorfi produced for each word.

	Chunk 1	**Chunk 2**	**Chunk 3**	**Chunk 4**	**Chunk 5**	**Omorfi**
Accuracy	96.1%	96.2%	96.3%	96.2%	95.9%	40.5%

TABLE 2 – Token level accuracy of each model in the test data.

Because there is no other dataset freely available that would both be written in Old Literary Finnish and lemmatized to modern orthography, we take randomly 50 sentences from the example sentences of the dictionary of Old Literary Finnish [9] that is available online. Altogether these sentences have 562 words (excluding punctuation). We lemmatize these sentences with the model that has been trained with chunks of 3 as it worked the best out of the models and verify the lemmatization by hand. The results of this experiment are seen in Table 3.

	Chunk 3	**Omorfi**
Accuracy	87.7%	47.9%

TABLE 3 – Accuracy in the Old Literary Finnish dictionary sample.

The results drop in this evaluation, but it is only to be expected given that out-of-domain performance is typically lower for neural models. Nevertheless, we see clearly that the model does well in out-of-domain data and beats the current state of the art. The fact that Historical Omorfi gets better results in this dataset is a good indication that the text is very different from what is in the Agricola dataset. However, it clearly is not entirely distinct when we consider how well the model still performed.

If we look at the results more closely, analyzing the errors, we can see that the model usually does not predict non-words but rather words that are a part of the Finnish vocabulary. This indicates that the model has learned a good target representation. There are several errors in which the model has normalized the historical word correctly, but it has not lemmatized it, for example *runsast* was normalized to *runsaasti* although the lemma would be *runsas* 'plenty'. Another example is *ulosteon* that was already written as in the modern orthography was normalized unchanged to *ulosteon*, while the lemma is *ulosteko*. This example also illustrates how there is variation and historical change in spelling, as we find in the Agricola corpus comparable words spelled with the initial *v*, whereas in the later materials we find the variant above that is spelled closely the current standard. Analysing how this relates to the changes in characters used in different centuries is beyond the scope of our study, but it illustrates well the kind of variation we can find in the historical texts and their digitized versions.

9. https ://kaino.kotus.fi/vks

To analyze the errors further, the most typical source of problems are verbs that get lemmatized into nouns that look similar and vice versa. Thereby *olan* was lemmatized as *olla* 'to be' while it should have been *olka* 'shoulder'. Also, *nuole* was lemmatized as *nuoli* 'arrow', while the correct lemma is *nuolla* 'to lick'. Normalization to a wrong lemma within the same part-of-speech is also possible e.g. *kaipanne* to *kaivaa* 'to dig' instead of *kaivata* 'to miss'. Improving the recognition of such instances is a very important task for the future work, but the current accuracy also appears to be useful and satisfactory for many tasks, and is without doubt an improvement to the existing methods.

6 Conclusions and Future Work

Our results have a clear indication, both with in-domain and out-of-domain test data, of working successfully in lemmatizing Old Literary Finnish in the modern orthography. The models have been released on Zenodo and in a Python library [10]. By sharing the models we are making NLP research on historical Finnish data more widely accessible for the research community, as the currently available Historical Omorfi does not work well for texts that are this old. Our study also creates a benchmark into which the further work can easily be compared.

Having lower accuracy in another dataset is a reminder of the importance of evaluating normalization models on data that comes from a different distribution. This is something seldom seen in the previous work on historical spelling normalization. Despite this, the accuracy has remained relatively high and we have identified several possibly problematic phenomena in the test data that are more prone to errors. This error analysis helps in understanding the biases that using our model might introduce in historical data when it is used to lemmatize a corpus completely new to the model. Further research is needed to evaluate how the error rate varies when the distance grows to the materials of Agricola. It is beyond doubt that more diverse training material is needed to successfully process the entire corpus of Old Literary Finnish, but our study certainly has improved the position to initiate and continue such work.

The work we presented here also makes it possible for the current research on analyzing historical Finnish newspapers, such as (Jean-Caurant & Doucet, 2020; Kettunen *et al.*, 2020), to standardize post-OCR historical Finnish, which in turn permits employing state-of-the-art Finnish NLP methods and tools on such data (e.g. sentiment and semantic analysis (Hämäläinen & Alnajjar, 2019)).

When we work with the currently available resources, we must also remain aware that the digital versions have been edited in various ways, and do not contain necessarily all features of the original printed text (Toropainen, 2016, 175). Now when we increasingly have access also to the original prints as high quality scans, it is important to think how these different resources can be connected to one another. This will need a combination of both text recognition and NLP tools.

In the future, we are interested in conducting work on semantic change on historical data. This should be greatly facilitated by the fact that we can now considerably reliably lemmatize historical text. This means that training word embeddings models will become more accurate as the model is trained on lemmas instead of inflectional forms.

10. https ://github.com/mikahama/murre

Références

AMOIA M. & MARTINEZ J. M. (2013). Using comparable collections of historical texts for building a diachronic dictionary for spelling normalization. In *Proceedings of the 7th workshop on language technology for cultural heritage, social sciences, and humanities*, p. 84–89.

BARON A. & RAYSON P. (2008). VARD2 : A tool for dealing with spelling variation in historical corpora. In *Postgraduate conference in corpus linguistics*.

BOLLMANN M. (2012). (Semi-)automatic normalization of historical texts using distance measures and the Norma tool. In *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*, Lisbon, Portugal.

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. & 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 : The COLING 2016 Organizing Committee.

DUONG Q., HÄMÄLÄINEN M. & HENGCHEN S. (2020). An unsupervised method for OCR post-correction and spelling normalisation for Finnish. *arXiv preprint arXiv :2011.03502*.

HÄMÄLÄINEN M. & ALNAJJAR K. (2019). Let's FACE it. Finnish poetry generation with aesthetics and framing. In *Proceedings of the 12th International Conference on Natural Language Generation*, p. 290–300, Tokyo, Japan : Association for Computational Linguistics. DOI : 10.18653/v1/W19-8637.

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.

HÄMÄLÄINEN M., SÄILY T., RUETER J., TIEDEMANN J. & MÄKELÄ E. (2019). Revisiting NMT for normalization of early English letters. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, p. 71–75, Minneapolis, USA : Association for Computational Linguistics. DOI : 10.18653/v1/W19-2509.

HAUSER A. W. & SCHULZ K. U. (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*, p. 1–6.

HOCHREITER S. & SCHMIDHUBER J. (1997). Long short-term memory. *Neural computation*, **9**(8), 1735–1780.

INABA N. (2015). *Suomen datiivigenetiivin juuret vertailevan menetelmän valossa*, volume 272 de *Mémoires de la Société Finno-Ougrienne*. Société Finno-Ougrienne.

INSTITUTE FOR THE LANGUAGES OF FINLAND (2014). Corpus of Old Literary Finnish. http ://urn.fi/urn :nbn :fi :lb-201407165.

INSTITUTE FOR THE LANGUAGES OF FINLAND & UNIVERSITY OF TURKU (2020). The Morpho-Syntactic Database of Mikael Agricola's Works version 1.1, Korp.

JEAN-CAURANT A. & DOUCET A. (2020). Accessing and investigating large collections of historical newspapers with the NewsEye platform. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, p. 531–532, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3383583.3398627.

KETTUNEN K., KOISTINEN M. & KERVINEN J. (2020). Ground truth OCR sample data of Finnish historical newspapers and journals in data improvement validation of a re-OCRing process. *Liber quarterly*. DOI : 10.18352/lq.10322.

KLEIN G., KIM Y., DENG Y., SENELLART J. & RUSH A. M. (2017). OpenNMT : Open-Source Toolkit for Neural Machine Translation. In *Proc. ACL*. DOI : 10.18653/v1/P17-4012.

LAHTI L., MARJANEN J., ROIVAINEN H. & TOLONEN M. (2019). Bibliographic data science and the history of the book (c. 1500–1800). *Cataloging & Classification Quarterly*, **57**(1), 5–23.

LUONG M.-T., PHAM H. & MANNING C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv :1508.04025*.

MANDAL S. & NANMARAN K. (2018). Normalization of transliterated words in code-mixed data using Seq2Seq model & Levenshtein distance. In *Proceedings of the 2018 EMNLP Workshop W-NUT : The 4th Workshop on Noisy User-generated Text*, p. 49–53, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-6107.

NATIONAL LIBRARY OF FINLAND (2021). Digital Collections. https ://digi.kansalliskirjasto.fi.

NEVALAINEN T. (2021). Time's arrow reversed ? the (a)symmetry of language change. In M. HÄMÄLÄINEN, N. PARTANEN & K. ALNAJJAR, Éds., *Multilingual Facilitation*. Rootroo Ltd.

PARTANEN N., HÄMÄLÄINEN M. & ALNAJJAR K. (2019). Dialect text normalization to normative standard Finnish. In W. XU, A. RITTER, T. BALDWIN & A. RAHIMI, Éds., *The Fifth Workshop on Noisy User-generated Text (W-NUT 2019)*, p. 141–146, United States : The Association for Computational Linguistics.

PETTERSSON E., MEGYESI B. & TIEDEMANN J. (2013). An SMT approach to automatic annotation of historical text. In *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013 ; May 22-24 ; 2013 ; Oslo ; Norway. NEALT Proceedings Series 18*, volume 087, p. 54–69 : Linköping University Electronic Press.

PIRINEN T. A. (2015). Development and use of computational morphology of Finnish in the open source and open science era : Notes on experiences with Omorfi development. *SKY Journal of Linguistics*, **28**, 381–393.

PORTA J., SANCHO J.-L. & GÓMEZ J. (2013). Edit transducers for spelling variation in Old Spanish. In *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013 ; May 22-24 ; 2013 ; Oslo ; Norway. NEALT Proceedings Series 18*, volume 087, p. 70–79 : Linköping University Electronic Press.

RAYSON P., ARCHER D. & SMITH N. (2005). VARD versus WORD : a comparison of the UCREL variant detector and modern spellcheckers on english historical corpora. *Corpus Linguistics 2005*.

SÄILY T., MÄKELÄ E. & HÄMÄLÄINEN M. (2018). Explorations into the social contexts of neologism use in early English correspondence. *Pragmatics & Cognition*, **25**(1), 30–49. DOI : 10.1075/pc.18001.sai.

SALMI H. (2020). German influence on the Finnish in Mikael Agricola. *Finnish-German Yearbook of Political Economy, Volume 2*, p. 135.

SAMARDZIC T., SCHERRER Y. & GLASER E. (2015). Normalising orthographic and dialectal variants for the automatic processing of Swiss German. In *Proceedings of the 7th Language and Technology Conference*. ID : unige :82397.

SCHUSTER M. & PALIWAL K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, **45**(11), 2673–2681.

SÄILY T., MÄKELÄ E. & HÄMÄLÄINEN M. (2021). From plenipotentiary to puddingless : Users and uses of new words in early english letters. In M. HÄMÄLÄINEN, N. PARTANEN & K. ALNAJJAR, Éds., *Multilingual Facilitation*. Rootroo Ltd.

TOLONEN M., LAHTI L., ROIVAINEN H. & MARJANEN J. (2019). A quantitative approach to book-printing in Sweden and Finland, 1640–1828. *Historical methods : a journal of quantitative and interdisciplinary history*, **52**(1), 57–78.

TOROPAINEN T. (2015). Adjektiivialkuiset yhdyssubstantiivit Mikael Agricolan teoksissa. *Sananjalka*, **57**(1), 54–85.

TOROPAINEN T. (2016). Typografian vaikutus yhdyssubstantiivien oikeinkirjoitukseen agricolan teoksissa. *Sananjalka*, **58**(1), 175–198.

TOROPAINEN T. (2018). *Yhdyssanat ja yhdyssanamaiset rakenteet Mikael Agricolan teoksissa*. Thèse de doctorat, University of Turku. Turun yliopiston julkaisuja. Sarja C, Scripta lingua Fennica edita.

TURSUN O. & CAKICI R. (2017). Noisy Uyghur text normalization. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, p. 85–93, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/W17-4412.

VAN DER GOOT R., PLANK B. & NISSIM M. (2017). To normalize, or not to normalize : The impact of normalization on Part-of-Speech tagging. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, p. 31–39.