Linguistic change and historical periodization of Old Literary Finnish

Niko Partanen, Khalid Alnajjar, Mika Hämäläinen and Jack Rueter
Faculty of Arts
University of Helsinki & Rootroo Ltd
firstname.lastname@helsinki.fi

Abstract

In this study, we have normalized and lemmatized an Old Literary Finnish corpus using a lemmatization model trained on texts from Agricola. We analyse the error types that occur and appear in different decades, and use word error rate (WER) and different error types as a proxy for measuring linguistic innovation and change. We show that the proposed approach works, and the errors are connected to accumulating changes and innovations, which also results in a continuous decrease in the accuracy of the model. The described error types also guide further work in improving these models, and document the currently observed issues. We also have trained word embeddings for four centuries of lemmatized Old Literary Finnish, which are available on Zenodo.

1 Intoduction

In this study, we investigate linguistic drift and historical periodization of Old Literary Finnish. We use a historical Finnish lemmatizer model trained on the works of Mikael Agricola, and apply the model to the remaining currently available corpus of Old Literary Finnish (Institute for the Languages of Finland, 2013). This allows us to examine both the differences in the model's performance and how the lexicon of Old Literary Finnish has changed and evolved over time.

We hypothesize that the contexts where the model's quality changes significantly correlate, in fact, with changes in the actual form of the literary language. These can be innovations in the orthography, or other kinds of linguistic changes that are known to have happened during the period Finnish has been a written language. Careful error detection should also reveal something about the nature of these changes. As long as the model's quality remains above a specific threshold, we should also be able to monitor the use of specific lexemes over time. We trained the word embeddings for this purpose. The corpus size being limited, and divided to time period of 1543–1809, we concluded that more data is needed to follow the actual semantic changes.

2 Related work

Natural language processing for Old Literary Finnish is still in a very early stage, while extensive work already exists for historical variants of other languages (Dubossarsky et al., 2019; Perrone et al., 2019; Hill and Hengchen, 2019; Degaetano-Ortlieb et al., 2021). Most work has been done with historical newspapers, which represent only later periods of this language variety, starting from 1771. Many studies are connected to improving OCR accuracy, which remains as an important task for old printed materials. Recognizing named entities is another line of research that has been developed relatively far, especially by Kettunen and Ruokolainen (2017), Kettunen et al. (2016a) and Kettunen and Löfberg (2017). This connects to other work in NER of other Finnish varieties (Porjazovski et al., 2020; Ushio and Camacho-Collados, 2021).

Also evaluation and post processing approaches are closely connected to our study. Kettunen and Pääkkönen (2016) and Kettunen et al. (2016b) used a morphological analyser adapted for historical Finnish to evaluate OCR accuracy in these newspapers. Later on, OCR accuracy has been improved through unsupervised post-correction in Finnish newspapers (Duong et al., 2020).

Koskenniemi and Kuutti (2017) studied alignment and analysis of Old Literary Finnish, using a Helsinki Finite-State Transducer (Lindén et al., 2013). Lexical change through neologisms has been studied in historical data by comparing word occurrences in a historical corpus to earliest attestations recorded in dictionaries (Säily et al., 2021).

Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021, pages 21–27
August 6, 2021. ©2021 Association for Computational Linguistics

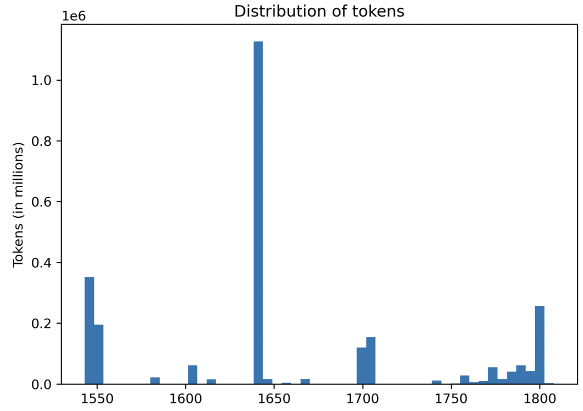

Figure 1: Distribution of tokens in the Corpus of Old Literary Finnish

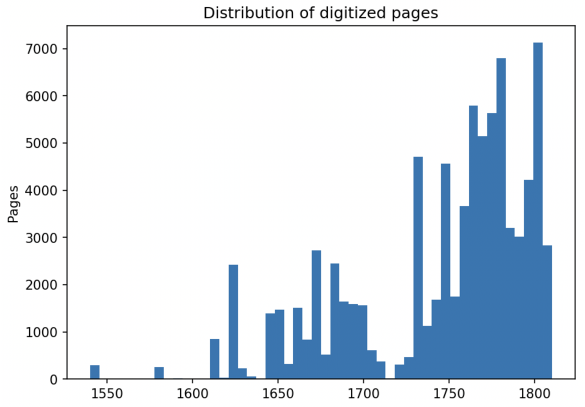

Figure 2: Digitized Finnish pages in the National Library of Finland's Digital collections (May 2021)

3 Data

We use the Old Literary Finnish corpus (Institute for the Languages of Finland, 2013). This is the only proofread corpus of Old Literary Finnish currently available, it aims to be representative, and is created especially for the purposes of lexicography. The current corpus is 4.13 million tokens in size. The distribution of tokens by year is shown in Figure 1. To contextualize the distribution, the Bible translation from year 1642 contains over a million tokens. The corpus has 1.5 million tokens where the year is not defined in the metadata, and thereby were not included in our study.

In order to better understand the relationship this data has to the entire Old Literary Finnish corpus, we can compare it to various adjacent sources. The first logical point of comparison is the national metadata catalogs, which should contain relatively complete information about all books that have ever been printed. This data was already analysed by (Tolonen et al., 2019), and their figures are certainly worth comparing in this context, too.

As text sources, however, these materials are only useful for us if they have been digitized and can be accessed. To understand this context, we examined the number of digitized pages from the same time period in the collections of the National Library of Finland[1]. The distribution of digitized Finnish pages is shown in Figure 2.

This shows that our current sample is still relatively small, and many different sample constellations could be imagined. Comparing to (Tolonen et al., 2019), for example, it seems that the dip in digitized pages we see in Figure 2 in the first half of the 18th century does not seem to correlate with a reduced printing activity in this time period. Similarly the Old Literary Finnish corpus has four larger peaks, representing, presumably, the goal to include all four centuries of this language variety to a comparable degree.

Besides the proofread portion of the corpus, the materials of Agricola have been published as a morphosyntactically annotated version (Institute for the Languages of Finland and University of Turku, 2020). Each resource type we have discussed above is narrower than the one before, as specialized annotation, proofreading and digitization are all resource demanding activities. Our work explores what we can do with the current data, existing annotations, and how we can build NLP solutions around these materials to extend and enrich the available resources. Publishing our word embeddings also contributes to this goal.[2]

For evaluation purposes, we have also created our own manually lemmatized dataset.[3] This ground truth material was created where possible with the Dictionary of Old Literary Finnish (Kotimaisten kielten keskus, 2021). Since the dictionary only currently extends to the word *perstauta* 'to rot; to decay', however, there are instances where we could not consult this resource, and had to decide the evaluated lemma with our own linguistic intuition. For example, one description of metallurgy practices from 1797 contains the segment *jotka makawat palkein ylitze ja wääteillä* 'which lie over the bellows and [unknown word]' [Rin1797-49]. The wordform *wääteillä* is not in the dictionary, and it occurs only in this decade in the currently

[1] https://digi.kansalliskirjasto.fi

[2] https://zenodo.org/record/4906651

[3] https://zenodo.org/record/4906626

available corpus. We have lemmatized this lexeme as *väöde*, with full knowledge that this may be erroneous. As our dataset is openly available, the errors are easily corrected later. This illustrates how extremely complicated tasks normalization and lemmatization of historical texts are, and we approach this question with the goal to evaluate the currently available methods, and to improve our understanding on how to improve our models.

4 Experiment design

We used an Old Literary Finnish lemmatizer (Hämäläinen et al., accepted) trained with manually lemmatized corpus form Agricola (Institute for the Languages of Finland and University of Turku, 2020). The lemmatization model reached 96.3% accuracy in texts written by Agricola, and 87.7% accuracy in out-of-domain data (Hämäläinen et al., accepted). The model follows the same LSTM architecture that has been found useful both for modern Finnish normalization (Partanen et al., 2019) and dialectalization (Hämäläinen et al., 2020).

Our hypothesis is that if we evaluate the errors the model makes with the texts originating from different periods, we can use the errors as a proxy for progressing changes. These results can be later verified and dated more accurately with larger corpora as such resources become available.

For our error analysis, we have selected 25 sentences from different decades, and manually lemmatized them. If a decade had a smaller number of sentences, then we took all the sentences available. The manual annotations are used as the gold standard against which the model's predictions are compared. This results in a manually corrected dataset of 476 sentences.

5 Result

We find that the word error rate (WER) of the lemmatization model fluctuates between 1–23% in our test dataset. The WER, however, increases gradually when measured by the decade, and our hypothesis is that this change represents the linguistic distance that increases when new vocabulary and conventions are added to the written standard.

This can be tested through a detailed error classification and analysis, which we conduct in the next section. Whenever possible, we aim to provide estimations of when different features emerge, which hopefully allows to detect various periods that can be distinguished from one another.

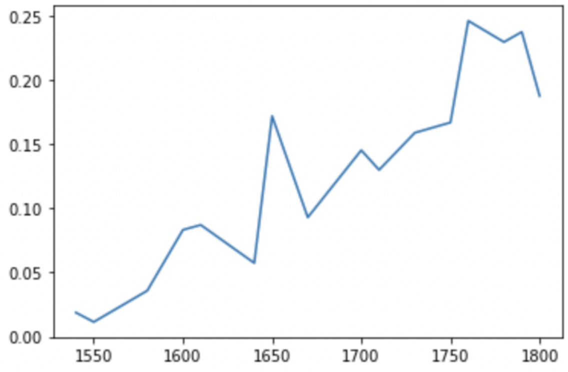

Figure 3: Word Error Rate per decade

6 Error analysis

6.1 Agricola texts

Although the model was trained with the Agricola data, there are still individual errors even in this material. These relate often to personal names such as *Ziphi* and *Zipheis*, which in the original corpus were normalized as *Sifi*. In the vicinity of these words the normalization is very good, and generally we do not find that lemma level errors would impact a more extensive sentence.

We presume these are words which have not occurred in the original training data, where part of the Agricola corpus was used as the test data, or then forms are simply too rare or exceptional. Needless to say, as the Figure 3 shows the accuracy is almost flawless in the earliest portion of the corpus. We can illustrate the accuracy with an example from Agricola's Prayer book. The original sentence is *Mine rucolen sinua sinun poias cautta* 'I pray for you through your son' [rk1544-647]. The correct lemmatization is *minä rukoilla sinä sinä poika kautta*, which is exactly what the model outputs.

6.2 Emerging changes

It appears that the errors are strongly connected to new types of linguistic content and writing conventions. For example, Agricola never used the pronoun form *sä* 'you', opting for full forms instead, as illustrated in example above. Once the shorter form starts to appear, the model is not always able to normalize them correctly. For example, from 1616 there is an example *Sä quin ryövärille jaoid Paradiisin perimisen* 'As if you shared the inheritance of Paradise with the robber' [Hemm1616-50], where the first pronoun is normalized with the verb *säätää* 'to ordain', which is entirely incorrect. The model is clearly extremely sensitive to small differences in the spelling conventions. We can see this well in spelling variants that are used only by individual authors. For example, *neidzö* appears to be used only in texts by Jacobus Petri Finno from 1583 (for example, see [FinnoVk-4:15-3a]). We presume the actual distributions of different variants are larger, just not yet visible in this corpus.

We can also point out that the model leaves numbers untouched, but larger years are usually restructured, so that 1761 becomes 1161. In Agricola's materials there are no larger years than 1551, which is when his Book of Psalms was published.

6.3 Challenge of multilinguality

Another type of errors comes from materials in languages other than Finnish. Currently, these are left out from the accuracy count, but they are present in our balanced sample. These include, for example Latin phrases such as: *Magi de longe veniunt Aurum Thus Myrrham offerunt / Intrantes domum in vicem Natum salutant homines.* [FinnoVk-51:0-46b]. What the model returns is *magi de loki vene auru tus myrha oferu / intrante tuomus ja viedä natus saluttaa huomines.* As we see, it has tried to normalize Latin into Finnish, which obviously fails. In the future it will be important to investigate whether the same model can be used with different languages, or if we can teach the model to ignore non-Finnish content, so that they could be processed with more suitable tools. Another instance like this is the Greek phrase *kyrie eleison* 'Lord, have mercy', which is not used by Agricola.

There are also multiple instances of foreign names that the model cannot process. Names such as *Küttleri*, *Sinclair* and *Gezelius* are not processed correctly. The ideal behaviour for the model would be to leave proper nouns unnormalized, other than lemmatizing them into the nominative singular. Currently the model often returns a close approximation of this, but names such as *Gezelius* are slightly normalized to *geselius*. Similarly *Stockholmin* is lemmatized into *tokkolma*. This is a common problem with many neural models: the number of potential new or foreign proper nouns that can occur in the text is enormous, and they regularly contain characters and character sequences that have not been seen before. However, similar issues are also met with Finnish toponyms such as *Tammela* and *Jokiainen*, so the problems are not exclusively related to foreign names.

6.4 Evolving punctuation & conventions

The use of comma was not yet characteristic for Agricola's materials. Interestingly, in the contemporary handwritten Westh Codex the comma is regularly used. We find increasing use of the comma from 1640, and after 1740 it appears to be fully established alongside other modern punctuation. The change has been gradual, and deserves further investigation. For the periodization the use of modern punctuation would be an obvious candidate, as we could possibly split the material into sections before and after the emergence of this practice. It seem that the process has been gradual. For example, Petraeus in 1656 has already begun using rather modern punctuation, including regular use of the comma. However, not exclusively, and / can also be seen to have a function. 1700 is the last decade when / is regularly used in writing. More comprehensive corpus would certainly allow more nuanced analysis. This is also a decade in which we see a massive increase in the use of hyphens to separate elements in the compounds. Still, the use of the comma is entirely new to the model, and these are regularly returned as numbers or individual letters.

Another distinction that emerges in the 18th century is the use of the section sign, §. Our first occurrence is in an almanac from 1705, after which they become common: especially so in almanacs and legal texts. Thereby this also connects to the differentiation of text genres. In 1640s we see that the accuracy improves in relation to previous decades. Since most of the data from that period comes from a Bible translation, we believe there is a domain match with Agricola's data, which improves the performance.

We can also point out the increased use of abbreviations separated from case marking with a colon, such as the word 'majesty' in *Cosca Kuningallisen Maj:tin uscollinen Mies* 'Because of the man loyal to the his majesty the King' [ZLith1718-1]. Agricola doesn't yet use this convention, so the model has never encountered it, and cannot normalize these instances correctly. In the current corpus this convention is used in other texts but not in Agricola, which makes it impossible to date more exactly when it has started to be used. For the future work, we would suggest to train the model so that abbreviations are expanded automatically.

6.5 Expanding domains & vocabulary

Especially in the newer data we see the domain difference growing. For example, in Frosterus's 1791 work *Hyödyllinen Huwitus Luomisen Töistä* 'Beneficial pastime in the work of creation' among the topics discussed are planets and other modern scientific concepts, which include terminology the model has never seen. Yet, we can see that the model has some internal logic also here. Word *planetit* 'planets (modern spelling *planeetat*)' is lemmatized as *planetti*, which is not the contemporary singular form *planeetta*, but still a reasonable guess from the old spelling. This can be compared to Lissander's 1793 publication *Maa-Pärunain Kasswattamisesta* 'On the growing of potatoes'. Again, the model is not able to handle the entirely new type of terminology, including plants. As this terminology is often borrowed, it is even more difficult to normalize. The task we are performing is also somewhat more challenging than just lemmatization, as we have combined it with the normalization to modern Finnish. Thereby the correct lemma for word *soldati* 'soldier' would be *sotilas*, and not *soltatti* the model currently proposes. Similarly normalizing the word *phasianus* 'Common pheasant' as *fasaani* would probably require information the model currently cannot have. Naturally, it is another question in itself how these words should be lemmatized, and whether the contemporary Finnish should even be used as the desired target.

In a recent study that investigated neural morphological models for different languages one of the found error types were the unknown and foreign words that were phonotactically or orthographically unusual (Hämäläinen et al., 2021). We believe this process is present also here, when neural model fails to generalize to the input that contains innovations that are beyond the patterns in the training data, even though there is some generic capacity to deal with unseen material.

6.6 What about the periodization?

We believe that detecting and delineating different periods when the features emerge and become established is important, as the process how they have spread and become adapted may be very relevant for both historical and linguistic studies. By understand how the material differentiates we can also design our tools in more systematic and appropriate manner. However, as illustrated in Figure 1, the currently used corpus is not temporally perfectly representative, as there are several periods with no data available. Our analysis also suggests that the change is gradual and complex, and very clear cut periods cannot necessarily be found. In this point we refrain from presenting more definite numbers, as those are necessarily connected to conventions in individual works in our small sample, and the wider relationship between the texts cannot be seen.

In order to do periodization successfully more data is needed. However, many of these materials have been digitized (See Figure 2), and are in Public Domain. The path toward such a task is thereby open, and we hope our methodological demonstration in this study also contributes into this work.

7 Conclusion

We show that analysing the errors produced by a neural network that is trained for one task in one specific material serves as a good indicator for salient and emerging differences between the texts. The methodological contribution of our study is that we can use neural networks effectively to track these changes. We could not successfully split the material into distinct periods, but we propose this can be done. Still, we were able to trace the changes in some phenomena, especially the punctuation conventions. We see more of a gradual process than clear phases, which also indicates that our initial goal of periodization may not be ideal.

The most important finding of our study is that the proposed method works. As the error rate of the neural network increases linearly with newer material, we are convinced that this signals the increasing differentiation of the data in these periods when compared to the texts written by Agricola.

Although in reality there is no need to process Old Literary Finnish materials with the data from Agricola alone, besides the fact that only this material is available for training, we think the experiment design also has relevance for NLP research more generally. The language changes also in our day, and the models we train should be able to handle innovations that are only currently emerging. Therefore the test setting, although artificial, asks a question that is worth presenting.

Very importantly, our study provides a clear roadmap for the further development of normalization and lemmatization of Old Literary Finnish. As we published our models and materials openly in Zenodo, our analysis is easy to reproduce, and our initial benchmark can be improved.

References

Stefania Degaetano-Ortlieb, Tanja Säily, and Yuri Bizzoni. 2021. Registerial adaptation vs. innovation across situational contexts: 18th century women in transition. *Frontiers in Artificial Intelligence*, 4:56.

Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470.

Quan Duong, Mika Hämäläinen, and Simon Hengchen. 2020. An unsupervised method for OCR post-correction and spelling normalisation for Finnish. *arXiv preprint arXiv:2011.03502*.

Mika Hämäläinen, Niko Partanen, Khalid Alnajjar, Jack Rueter, and Thierry Poibeau. 2020. Automatic dialect adaptation in Finnish and its effect on perceived creativity. In *11th International Conference on Computational Creativity (ICCC'20)*. Association for Computational Creativity.

Mika Hämäläinen, Niko Partanen, Jack Rueter, and Khalid Alnajjar. 2021. Neural morphology dataset and models for multiple languages, from the large to the endangered. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 166–177, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Mark J Hill and Simon Hengchen. 2019. Quantifying the impact of dirty ocr on historical text analysis: Eighteenth century collections online as a case study. *Digital Scholarship in the Humanities*, 34(4):825–843.

Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar. accepted. Lemmatization of historical old literary Finnish texts in modern orthography. In *Proceedings of Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*.

Institute for the Languages of Finland. 2013. Corpus of Old Literary Finnish. Saatavilla: https://kaino.kotus.fi/korpus/vks/meta/vks_coll_rdf.xml.

Institute for the Languages of Finland and University of Turku. 2020. The Morpho-Syntactic Database of Mikael Agricola's Works version 1.1.

Kimmo Kettunen and Laura Löfberg. 2017. Tagging named entities in 19th century and modern Finnish newspaper material with a Finnish semantic tagger. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 29–36.

Kimmo Kettunen, Eetu Mäkelä, Juha Kuokkala, Teemu Ruokolainen, Jyrki Niemi, et al. 2016a. Modern tools for old content-in search of named entities in a Finnish OCRed historical newspaper collection 1771-1910. In *LWDA*, pages 124–135.

Kimmo Kettunen and Tuula Pääkkönen. 2016. Measuring lexical quality of a historical Finnish newspaper collection—analysis of garbled OCR data with basic language technology tools and means. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 956–961.

Kimmo Kettunen, Tuula Pääkkönen, and Mika Koistinen. 2016b. Between diachrony and synchrony: Evaluation of lexical quality of a digitized historical Finnish newspaper and journal collection with morphological analyzers. In *Baltic HLT*, pages 122–129.

Kimmo Kettunen and Teemu Ruokolainen. 2017. Names, right or wrong: Named entities in an OCRed historical Finnish newspaper collection. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, pages 181–186.

Kimmo Matti Koskenniemi and Pirkko Kuutti. 2017. Indexing old literary Finnish text. *K+ K= 120 Papers dedicated to László Kálmán and András Kornai on the occasion of their 60th birthdays*.

Kotimaisten kielten keskus. 2021. *Vanhan kirjasuomen sanakirja*. Number 38 in Kotimaisten kielten keskuksen verkkojulkaisuja. Päivitettävä julkaisu. Päivitetty 20.5.2021 [viitattu 7.6.2021]. Available https://kaino.kotus.fi/vks/.

Krister Lindén, Erik Axelson, Senka Drobac, Sam Hardwick, Juha Kuokkala, Jyrki Niemi, Tommi A Pirinen, and Miikka Silfverberg. 2013. HFST a system for creating NLP tools. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 53–71. Springer.

Niko Partanen, Mika Hämäläinen, Khalid Alnajjar, et al. 2019. Dialect text normalization to normative Standard Finnish. In *The Fifth Workshop on Noisy User-generated Text (W-NUT 2019) Proceedings of the Workshop*. The Association for Computational Linguistics.

Valerio Perrone, Marco Palma, Simon Hengchen, Alessandro Vatri, Jim Q Smith, and Barbara McGillivray. 2019. Gasc: Genre-aware semantic change for ancient Greek. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 56–66.

Dejan Porjazovski, Juho Leinonen, and Mikko Kurimo. 2020. Named entity recognition for spoken Finnish. In *Proceedings of the 2nd International Workshop on AI for Smart TV Content Production, Access and Delivery*, AI4TV '20, page 25–29, New York, NY, USA. Association for Computing Machinery.

Tanja Säily, Eetu Mäkelä, and Mika Hämäläinen. 2021. From plenipotentiary to puddingless: Users and uses of new words in early English letters. In Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar, editors, *Multilingual Facilitation*. Rootroo Ltd.

Mikko Tolonen, Leo Lahti, Hege Roivainen, and Jani Marjanen. 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.

Asahi Ushio and Jose Camacho-Collados. 2021. T-NER: An all-round Python library for transformer-based named entity recognition. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 53–62, Online. Association for Computational Linguistics.