# Benchmarking Finnish Lemmatizers across Historical and Contemporary Texts

**Emily Öhman**
Waseda University
Tokyo, Japan

**Leo Huovinen and Mika Hämäläinen**
Metropolia University of Applied Sciences
Helsinki, Finland

## Abstract

Lemmatization is crucial in natural language processing (NLP) for languages like Finnish, where complex inflectional morphology significantly affects downstream tasks such as parsing, named entity recognition, and sentiment analysis. This study evaluates the accuracy and efficiency of several Finnish lemmatizers, utilizing the Project Gutenberg corpus, which includes diverse Finnish-language texts from different periods. Notably, this is the first study to employ Trankit for Finnish lemmatization, providing novel insights into its performance. Additionally, the integration of Murre preprocessing has been emphasized, demonstrating substantial improvements in lemmatization results. By comparing traditional and neural-network-based approaches, this paper aims to provide insights into tool selection for NLP practitioners working with Finnish based on dataset characteristics and processing constraints.

## 1 Introduction

Lemmatization, which reduces word forms to their base forms (lemmas), is a critical step in NLP tasks, especially in languages with extensive morphological variation such as Finnish. The morphology of Finnish allows a single root to yield multiple surface forms, each conveying distinct syntactic or semantic nuances. This complexity, compounded by the dialectal (see Hämäläinen et al., 2020) and historical variation of Finnish (see Partanen et al., 2021), presents challenges for lemmatization tools. The Project Gutenberg corpus[1] offers a valuable resource for assessing Finnish lemmatizers, as it includes literature spanning different dialects and historical periods.

Moreover, this study marks the first application of Trankit (Nguyen et al., 2021) in the context of Finnish lemmatization, exploring its capabilities alongside established tools. In this study, we compare the performance of several Finnish lemmatizers on contemporary and historical Finnish texts, assessing their adaptability, accuracy, and processing efficiency. Our findings aim to guide NLP practitioners in choosing the most suitable lemmatizer based on dataset requirements and computational resources.

Additionally, this research contributes to the broader understanding of how lemmatization models handle linguistic diversity within Finnish texts. By evaluating lemmatizers across both standardized and non-standard forms, the study sheds light on their capacity to generalize beyond training data rooted in modern standard Finnish. This aspect is particularly crucial for digital humanities and corpus linguistics, where researchers frequently encounter orthographic and morphological variation in historical or dialectal sources (see Säily et al., 2021; Mäkelä et al., 2020). The analysis not only highlights technical performance metrics but also contextualizes them in terms of linguistic coverage, robustness, and the practical implications for downstream NLP tasks such as part-of-speech tagging, parsing, and information retrieval.

## 2 Related Work

Finnish NLP has seen increasing interest due to the complex morphology of the language and its membership in the Uralic language family (Hämäläinen and Alnajjar, 2021). A range of lemmatization tools have been developed, from traditional rule-based methods to neural-network-driven approaches. Modern tools such as the Turku Neural Parser (Haverinen, 2014) and Murre [Finnish for *dialect*] (Partanen et al., 2019) represent neural advancements, while older tools like Omorfi (Pirinen, 2015) remain foundational resources.

In addition, this study introduces Trankit for Finnish lemmatization, a novel application that has

[1] https://www.gutenberg.org



*Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages*, pages 6–11

not been previously explored in the literature. Previous studies have noted the challenges of lemmatizing Finnish due to its morphological diversity (Öhman and Rossi, 2022; Rossi and Öhman, 2025). Most tools are optimized for contemporary Finnish, limiting their performance on historical dialects that feature unique lexical, orthographic, and morphological characteristics.

As highlighted by Hämäläinen et al. (2021), flexible, open-source resources are essential to support Finnish and other Uralic languages within NLP. Studies on handling morphological richness in NLP include approaches like FinPos (Silfverberg et al., 2016) for unsupervised morpheme segmentation and the broader universal dependency approach (Nivre et al., 2020) for creating a multilingual treebank collection.

## 3 Data

### 3.1 Dataset Description

To comprehensively evaluate lemmatizer performance across different varieties of Finnish, we utilized:

#### 3.1.1 Standard Finnish Corpus (499 sentences)

We selected 499 sentences of contemporary Finnish texts from Project Gutenberg. These texts adhere to current orthographic and morphological standards, serving as a baseline for lemmatizer performance. The corpus was further segmented into sentences and tokenized using the Trankit tokenizer to maintain consistency across experiments. We selected works from different authors and genres (e.g., fiction, essays, and religious texts) to capture stylistic and lexical diversity. Texts were downloaded in UTF-8 format and preprocessed to remove metadata such as licensing information and page headers.

#### 3.1.2 Non-Standard Finnish Corpus (189 sentences)

We selected 189 sentences from Finnish-language texts from Project Gutenberg, encompassing:

- **1950s Finnish Texts**: Mid-20th-century works that reflect Finnish language conventions from a transitional period.
- **Historical Finnish Texts**: Older works, including Old Literary Finnish, which exhibit archaic vocabulary and distinct morphological variations.

The corpus was further segmented into sentences and tokenized using the Trankit tokenizer to maintain consistency across experiments. For each temporal category, we selected works from different authors and genres (e.g., fiction, essays, and religious texts) to capture stylistic and lexical diversity. This stratified selection process enabled us to examine how temporal and stylistic variation influences lemmatization accuracy, providing a more comprehensive evaluation of the tools' robustness across linguistic and historical dimensions.

To assess lemmatizer performance on historical texts, we manually annotated a test set derived from Old Literary Finnish materials. This annotated set captures unique features such as archaic vocabulary and morphological patterns absent in modern Finnish, providing a reliable reference for evaluating context-sensitive lemmatization. Despite its modest size, this set serves as a valuable benchmark for identifying the strengths and limitations of each tool when applied to historical language data.

### 3.2 Ground Truth Lemma Annotation

Ground truth lemmas were manually annotated by a native Finnish speaker with expertise in Finnish linguistics. The annotation process followed standard Finnish morphological conventions:

- Converting all verbs to infinitive forms (ending in -a/-ä)
- Reducing nouns to nominative singular forms
- Normalizing pronouns to base forms (e.g., "mun" $\rightarrow$ "minä")
- Handling dialectal forms by first normalizing orthography, then lemmatizing

The annotation process was conducted by a native Finnish speaker with sufficient expertise in historical linguistics, ensuring consistency and linguistic accuracy. Annotations were performed following established Finnish morphological and orthographic conventions, with special attention given to variant spellings, obsolete inflectional forms, and compounds that deviate from contemporary usage. The resulting dataset thus not only functions as a gold standard for evaluating lemmatization tools but also contributes to the broader effort of building linguistically grounded resources for historical Finnish NLP research.

## 4 Method

We evaluated the lemmatizers on F1 score comparing their output to the annotated gold standard. Special attention was paid to context-sensitive lemmatization, where words assume different lemma forms depending on sentence context. The tools assessed include:

- **Turku Neural Parser**: A neural model known for high accuracy in contemporary Finnish lemmatization (Haverinen, 2014; Kanerva et al., 2020).
- **Murre**: Handles dialectal variation, designed specifically for dialectal Finnish (Partanen et al., 2019).
- **spaCy Experimental Models**: Neural models for Finnish lemmatization within spaCy's framework (Pires et al., 2019).
- **Omorfi**: A rule-based morphological database, foundational in Finnish NLP (Pirinen, 2015).
- **Trankit-FTB and Trankit-TDT**: For the first time, we incorporate Trankit models tailored for Finnish, specifically the FinnTreeBank 1 (FTB) and Turku Dependency Treebank (TDT) variants from Universal Dependencies (Zeman et al., 2020), to evaluate their performance against established lemmatizers.

We trained two Trankit models, Trankit-FTB and Trankit-TDT, using the Finnish Universal Dependencies (UD) Treebanks: FinnTreeBank 1 (FTB) and the Turku Dependency Treebank (TDT). These treebanks provide syntactically annotated Finnish sentences, each containing gold-standard lemma annotations suitable for supervised learning. The FTB corpus primarily represents more formal, edited Finnish, while the TDT contains a broader range of contemporary written texts, including journalistic and web-based material. Both corpora were split into training, development, and test sets following the UD conventions to ensure reproducibility.

The Trankit models were fine-tuned on these datasets using their respective UD splits, employing the default multilingual pre-trained weights as initialization. Training was performed for multiple epochs until convergence, with early stopping based on development set performance. This setup allowed us to evaluate how well Trankit generalizes across different Finnish language varieties and annotation schemes. By training separately on both FTB and TDT, we aimed to capture potential differences in domain-specific morphological patterns and assess the transferability of Trankit's lemmatization capabilities to historical and dialectal data in the Project Gutenberg corpus.

We tested the aforementioned tools on both the standard Finnish corpus (n=499) and non-standard Finnish corpus (n=189) with and without preprocessing using Murre. F1 scores were calculated to evaluate the accuracy of each tool by comparing predicted lemmas against ground truth annotations. This evaluation allows us to assess whether Murre preprocessing provides consistent benefits across different varieties of Finnish or specifically targets non-standard variation.

## 5 Results

The F1 scores for each lemmatizer, shown in Figures 1 and 2, reveal a striking contrast between standard and non-standard Finnish. Our results demonstrate that **Murre preprocessing provides substantial benefits for non-standard Finnish while showing minimal effect on standard Finnish**, highlighting its specific utility for non-standard language varieties.

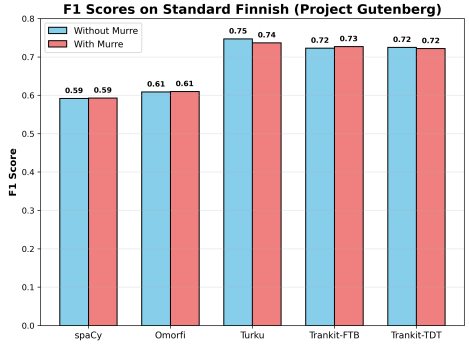


Figure 1: Standard Finnish (n=499) with and without Murre Preprocessing.

### 5.1 Standard Finnish Results

On the Project Gutenberg corpus (n=499), all lemmatizers achieved high baseline performance, and Murre preprocessing showed minimal impact:

- **spaCy**: 0.592 → 0.593 (+0.2%)
- **Omorfi**: 0.609 → 0.610 (+0.2%)
- **Turku**: 0.747 → 0.737 (-1.3%)

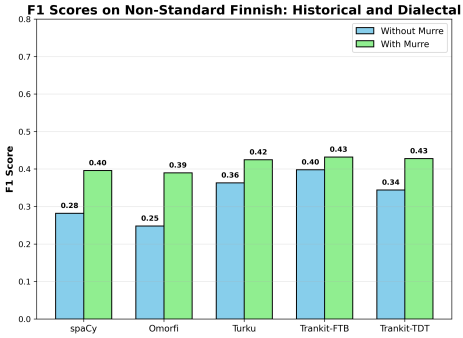


Figure 2: Non-Standard Finnish (n=189) with and without Murre Preprocessing.

- **Trankit-FTB**: 0.723 → 0.727 (+0.6%)
- **Trankit-TDT**: 0.725 → 0.722 (-0.4%)

The negligible changes (mostly ±0-1%, with Turku at -1.3%) indicate that Murre provides little benefit for well-formed standard Finnish, with Turku even showing slight degradation.

### 5.2 Non-Standard Finnish Results

In stark contrast, the non-standard Finnish corpus (n=189) showed dramatic improvements with Murre preprocessing across all models:

- **spaCy**: 0.282 → 0.396 (+40.3%)
- **Omorfi**: 0.248 → 0.390 (+57.4%)
- **Turku**: 0.363 → 0.425 (+17.2%)
- **Trankit-FTB**: 0.398 → 0.432 (+8.5%)
- **Trankit-TDT**: 0.344 → 0.428 (+24.5%)

Notably, Omorfi exhibited the largest relative improvement (+57%), while Trankit-FTB achieved the highest absolute F1 score (0.432) after Murre preprocessing. The consistently large improvements (8-57%) validate Murre's effectiveness specifically for dialectal variation, where orthographic and morphological normalization bridges the gap between non-standard forms and lemmatizer expectations.

## 6 Analysis & Discussion

The integration of Murre preprocessing has been pivotal in enhancing the performance of all evaluated lemmatizers on non-standard Finnish, while showing minimal impact on standard Finnish. The evaluation of Finnish lemmatization tools reveals several insights into the effectiveness of neural versus rule-based approaches, particularly when combined with normalization using Murre. The ability of Murre to standardize non-standard Finnish forms allows neural tools like spaCy to capitalize on their deep learning architectures by focusing on morphology within standardized contexts. This preprocessing step effectively bridges the gap between dialectal language forms and modern NLP tools, demonstrating the value of Murre in enhancing lemmatization accuracy on non-standard datasets (Bollmann, 2019).

Furthermore, the introduction of Trankit marks a significant advancement in Finnish lemmatization, as this study is the first to explore its capabilities in this context. Neural approaches such as spaCy and the Turku Dependency Parser outperformed traditional tools on both standard and non-standard texts after Murre normalization, with Trankit-FTB achieving the highest F1 score (0.432) on non-standard Finnish. However, when Murre normalization was applied to non-standard data, both rule-based and neural tools saw substantial improvements, with Omorfi showing the largest relative gain (+57%). This suggests that while neural lemmatizers excel with standardized data, preprocessing with tools like Murre remains a critical step for maximizing performance on dialectal forms.

The error analysis highlighted that compound words and dialectal or archaic spellings posed challenges for all lemmatizers without normalization. Typical errors included incorrect segmentation of compounds and failure to map archaic forms to their standard lemmas. Murre normalization alleviated these issues significantly, although it introduced occasional inaccuracies by altering foreign terms or named entities—a limitation worth addressing in future tool development (Piotrowski, 2012).

The findings underscore the practical implications for NLP practitioners: for datasets containing historical or dialectal language, preprocessing steps like Murre normalization are beneficial, especially when paired with high-performing neural lemmatizers such as Trankit. This study thus provides actionable recommendations for optimizing Finnish lemmatization accuracy based on dataset characteristics and offers a clear direction for integrating normalization as a preprocessing standard in Finnish NLP.

## 7 Conclusions

This evaluation demonstrates the substantial impact of Murre normalization in improving lemmatization accuracy for non-standard Finnish texts across both rule-based and neural lemmatizers. By enhancing the effectiveness of neural lemmatizers on non-standard Finnish, Murre normalization supports more accurate lemmatization across the diverse language variations present in non-standard corpora, while showing minimal impact on standard Finnish. Additionally, the novel application of Trankit in this study opens new avenues for Finnish lemmatization research, showcasing its potential alongside established tools. For future work, developing lemmatizers specifically trained on historical Finnish could further reduce reliance on normalization, allowing even greater adaptability for morphologically rich languages like Finnish.

Moreover, the results highlight the complementary nature of normalization and neural modeling in tackling Finnish's morphological complexity. While normalization mitigates surface-level variation, models like Trankit leverage contextual embeddings to capture deeper syntactic and semantic relations, suggesting that a hybrid pipeline combining these strengths yields the most robust outcomes. Future research should explore joint training approaches that integrate normalization directly within lemmatization architectures, allowing the model to learn from both standardized and non-standard forms simultaneously. Expanding training data to include diachronic and dialectal corpora will be essential for building lemmatizers capable of handling Finnish's full linguistic spectrum without extensive preprocessing.

## References


M. Bollmann. 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)*, pages 3885–3898. Association for Computational Linguistics.

Mika Hämäläinen and Khalid Alnajjar. 2021. The current state of finnish nlp. *arXiv preprint arXiv:2109.11326*.

Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar. 2021. Lemmatization of historical old literary finnish texts in modern orthography. *arXiv preprint arXiv:2107.03266*.

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.

Marko Haverinen. 2014. Turku neural parser. *Unpublished Manuscript*.

Jenna Kanerva, Filip Ginter, and Tapio Salakoski. 2020. Universal lemmatizer: A sequence to sequence model for lemmatizing universal dependencies treebanks. *Natural Language Engineering*, pages 1–30.

Eetu Mäkelä, Krista Lagus, Leo Lahti, Tanja Säily, Mikko Tolonen, Mika Hämäläinen, Samuli Kaislaniemi, and Terttu Nevalainen. 2020. Wrangling with non-standard data. In *Digital Humanities in the Nordic Countries*, pages 81–96. CEUR-WS. org.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*.

Joakim Nivre et al. 2020. Universal dependency parsing. *Journal of Computational Linguistics*.

Emily Öhman and Riikka Rossi. 2022. Computational Exploration of the Origin of Mood in Literary Texts. *NLP4DH 2022@Asian Association for Computational Linguistics*, page 8.

Niko Partanen, Mika Hämäläinen, and Khalid Alnajjar. 2019. Dialect text normalization to normative standard finnish. In *The Fifth Workshop on Noisy User-generated Text (W-NUT 2019)*. The Association for Computational Linguistics.

Niko Partanen, Jack Rueter, Khalid Alnajjar, and Mika Hämäläinen. 2021. Processing ma castrén's materials: Multilingual historical typed and handwritten manuscripts. In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 47–54.

J. Piotrowski. 2012. Natural language processing for historical texts. *Journal of Computational Linguistics*.

Thomas Pires et al. 2019. Multilingual language models for zero-shot cross-lingual transfer and language understanding. *Transactions of the Association for Computational Linguistics*.

Tommi A. Pirinen. 2015. Omorfi—free and open source morphological lexical database for finnish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 313–315.

Riikka Rossi and Emily Öhman. 2025. Combining qualitative and computational approaches for literary analysis of Finnish novels. *Scandinavian Studies*, 97(3):27–51.

Tanja Säily, Eetu Mäkelä, and Mika Hämäläinen. 2021. From plenipotentiary to puddingless: Users and uses of new words in early english letters. In *Multilingual Facilitation*, pages 153–169. University of Helsinki.

Miikka Silfverberg, Teemu Ruokolainen, Krister Lindén, and Mikko Kurimo. 2016. Finnpos: an open-source morphological tagging and lemmatization toolkit for finnish. *Language Resources and Evaluation*, 50(4):863–878.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Željko Agić, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Angelina Aquino, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, and Flavio Massimiliano Cecchini. 2020. Universal dependencies 2.6. http://hdl.handle.net/11234/1-3226. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.