

IWCLUL 2024

**The 9th International Workshop on Computational
Linguistics for Uralic Languages**

Proceedings of the Workshop

November 28-29, 2024

©2024 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-128-5

Preface

Welcome to the Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages (IWCLUL), a dedicated forum for researchers, academics and practitioners in the field of computational linguistics with a focus on Uralic languages. This year's workshop builds on the IWCLUL tradition of interdisciplinary collaboration, knowledge sharing, and a common commitment to addressing the linguistic, technical, and social challenges related to Uralic languages in the digital age.

The Uralic language family, spanning across Europe and Asia and encompassing languages as diverse as Finnish, Hungarian and the endangered Udmurt and Khanty languages, presents unique computational challenges. Many of these languages are characterized by complex morphology, agglutinative structures, and unique syntactic and phonological systems, requiring tailored approaches in computational processing and linguistic modeling. Our workshop seeks to bring attention to these challenges and foster the development of innovative solutions that not only support these languages' use in digital contexts but also contribute to their preservation and growth.

This year, IWCLUL received a record number of submissions from researchers worldwide, reflecting the growing interest and engagement in computational approaches to Uralic languages. The selected papers cover a broad spectrum of topics covering both well-studied and lesser-resourced Uralic languages. The diversity of contributions highlights the continuous evolution of the field and the range of challenges being tackled by the computational linguistics community.

We hope that these proceedings inspire continued research and collaboration in computational linguistics for Uralic languages. May the insights, methodologies, and resources shared here contribute to meaningful advances in the field and foster an inclusive future for Uralic languages in the digital landscape.

Sincerely, The IWCLUL 2024 Organizing Committee

Organizing Committee

Organizers

Mika Hämäläinen, Metropolia University of Applied Sciences

Flammie Pirinen, Arctic University of Norway

Melany Macias, Metropolia University of Applied Sciences

Mario Crespo Avila, Complutense University of Madrid

Program Committee

Laszlo Fejes, Hungarian Research Centre for Linguistics

Heiki-Jaan Kaalep, University of Tartu

Gunta Kļava, University of Latvia

Oleg Belyaev, Lomonosov Moscow State University

Trond Trosterud, The Arctic University of Norway

Linda Wiechetek, The Arctic University of Norway

Khalid Alnajjar, F-Secure Oyj

Niko Partanen, University of Helsinki

Jack Rueter, University of Helsinki

Miikka Silfverberg, University of British Columbia

Janne Kauttonen, Haaga-Helia University of Applied Sciences

Michael Rießler, University of Eastern Finland

Aleksei Dorkin, University of Tartu

Jeremy Bradley, University of Vienna

Xinqiao Zhang, UC San Diego

Irina Khomchenkova, Lomonosov Moscow State University

David Dale, Meta

Timofey Arkhangelskiy, University of Hamburg

Viktor Martinović, University of Vienna

Table of Contents

<i>Aspect Based Sentiment Analysis of Finnish Neighborhoods: Insights from Suomi24</i> Laleh Davoodi, Anssi Öörni and Ville Harkke	1
<i>Political Stance Detection in Estonian News Media</i> Lauri Lüüsi, Uku Kangur, Roshni Chakraborty and Rajesh Sharma	12
<i>Universal-WER: Enhancing WER with Segmentation and Weighted Substitution for Varied Linguistic Contexts</i> Samy Ouzerrout	29
<i>DAG: Dictionary-Augmented Generation for Disambiguation of Sentences in Endangered Uralic Languages using ChatGPT</i> Mika Hämäläinen	36
<i>Leveraging Transformer-Based Models for Predicting Inflection Classes of Words in an Endangered Sami Language</i> Khalid Alnajjar, Mika Hämäläinen and Jack Rueter	41
<i>Multilingual Approaches to Sentiment Analysis of Texts in Linguistically Diverse Languages: A Case Study of Finnish, Hungarian, and Bulgarian</i> Mikhail Krasitskii, Olga Kolesnikova, Liliana Chanona Hernandez, Grigori Sidorov and Alexander Gelbukh	49
<i>Towards standardized inflected lexicons for the Finnic languages</i> Jules Bouton	59
<i>On Erzya and Moksha Corpora and Analyzer Development, ERME-PSLA 1950s</i> Jack Rueter, Olga Erina and Nadezhda Kabaeva	67
<i>Towards the speech recognition for Livonian</i> Valts Ernštreits	76
<i>Using Large Language Models to Transliterate Endangered Uralic Languages</i> Niko Partanen	81
<i>Specialized Monolingual BPE Tokenizers for Uralic Languages Representation in Large Language Models</i> Iaroslav Chelombitko and Aleksey Komissarov	89
<i>Compressing Noun Phrases to Discover Mental Constructions in Corpora – A Case Study for Auxiliaries in Hungarian</i> Balázs Indig and Tímea Borbála Bajzát	96
<i>On Erzya and Moksha Corpora and Analyzer Development, ERME-PSLA 1950s</i> Aleksei Dorkin, Taido Purason and Kairit Sirts	104
<i>On the Role of New Technologies in the Documentation and Revitalization of Uralic Languages of Russia in Historical and Contemporary Contexts</i> Alexander Nazarenko	109
<i>Applying the transformer architecture on the task of headline selection for Finnish news texts</i> Maria Adamova and Maria Khokhlova	115

<i>Keeping Up Appearances—or how to get all Uralic languages included into bleeding edge research and software: generate, convert, and LLM your way into multilingual datasets</i>	
Flammie A Pirinen	123
<i>Scaling Sustainable Development Goal Predictions across Languages: From English to Finnish</i>	
Melany Macias, Lev Kharlashkin, Leo Huovinen and Mika Hämäläinen	132
<i>Kola Saami Christian Text Corpus</i>	
Michael Rießler	138

Program

Thursday, November 28, 2024

10:00 - 10:10 *Workshop Opening*

10:10 - 11:00 *Lightning Talks*

11:00 - 12:00 *Oral Session 1*

Aspect Based Sentiment Analysis of Finnish Neighborhoods: Insights from Suomi24

Laleh Davoodi, Anssi Öörni and Ville Harkke

Political Stance Detection in Estonian News Media

Lauri Löösi, Uku Kangur, Roshni Chakraborty and Rajesh Sharma

Scaling Sustainable Development Goal Predictions across Languages: From English to Finnish

Melany Macias, Lev Kharlashkin,, Leo Huovinen and Mika Hämäläinen

12:00 - 13:00 *Lunch*

13:00 - 14:20 *Oral Session 2*

Multilingual Approaches to Sentiment Analysis of Texts in Linguistically Diverse Languages: A Case Study of Finnish, Hungarian, and Bulgarian

Mikhail Krasitskii, Olga Kolesnikova, Liliana Chanona Hernandez, Grigori Sidorov and Alexander Gelbukh

Towards standardized inflected lexicons for the Finnic languages

Jules Bouton

DAG: Dictionary-Augmented Generation for Disambiguation of Sentences in Endangered Uralic Languages using ChatGPT

Mika Hämäläinen

Keeping Up Appearances—or how to get all Uralic languages included into bleeding edge research and software: generate, convert, and LLM your way into multilingual datasets

Flammie A Pirinen

14:20 - 14:40 *Coffee Break*

Thursday, November 28, 2024 (continued)

14:40 - 16:00 *Oral Session 3*

Towards the speech recognition for Livonian

Valts Ernštreits

Using Large Language Models to Transliterate Endangered Uralic Languages

Niko Partanen

*Specialized Monolingual BPE Tokenizers for Uralic Languages Representation
in Large Language Models*

Iaroslav Chelombitko and Aleksey Komissarov

*Leveraging Transformer-Based Models for Predicting Inflection Classes of Words
in an Endangered Sami Language*

Khalid Alnajjar, Mika Hämäläinen and Jack Rueter

Friday, November 29, 2024

10:00 - 11:00 *Keynote*

11:00 - 12:00 *Oral Session 4*

Compressing Noun Phrases to Discover Mental Constructions in Corpora – A Case Study for Auxiliaries in Hungarian
Balázs Indig and Tímea Borbála Bajzát

On the Role of New Technologies in the Documentation and Revitalization of Uralic Languages of Russia in Historical and Contemporary Contexts
Alexander Nazarenko

Applying the transformer architecture on the task of headline selection for Finnish news texts
Maria Adamova and Maria Khokhlova

12:00 - 13:00 *Lunch*

13:00 - 14:20 *Oral Session 5*

Kola Saami Christian Text Corpus
Michael Rießler

On Erzya and Moksha Corpora and Analyzer Development, ERME-PSLA 1950s
Aleksi Dorkin, Taido Purason and Kairit Sirts

Universal-WER: Enhancing WER with Segmentation and Weighted Substitution for Varied Linguistic Contexts
Samy Ouzerrout

On Erzya and Moksha Corpora and Analyzer Development, ERME-PSLA 1950s
Jack Rueter, Olga Erina and Nadezhda Kabaeva

14:20 - 14:40 *Coffee Break*

14:40 - 15:40 *SIGUR Business Meeting*

Friday, November 29, 2024 (continued)

Aspect Based Sentiment Analysis of Finnish Neighborhoods: Insights from Suomi24

Laleh Davoodi and Anssi Öörni and Ville Harkke

Faculty of Social Sciences, Business and Economics, and Law, Åbo Akademi University
first.last@abo.fi

Abstract

This study presents an approach to Aspect-Based Sentiment Analysis (ABSA) using Natural Language Processing (NLP) techniques to explore public sentiment across 12 suburban neighborhoods in Finland. We employed and compared a range of machine learning models for sentiment classification, with the RoBERTa model emerging as the best performer. Using RoBERTa, we conducted a comprehensive sentiment analysis (SA) on a manually annotated dataset and a predicted dataset comprising 32,183 data points to investigate sentiment trends over time in these areas. The results provide insights into fluctuations in public sentiment, highlighting both the robustness of the RoBERTa model and significant shifts in sentiment for specific neighborhoods over time. This research contributes to a deeper understanding of neighborhood sentiment dynamics in Finland, with potential implications for social research and urban development.

1 Introduction

Understanding public sentiment towards specific neighborhoods is crucial for urban planners, sociologists, and policymakers, as it offers insights into the social fabric of communities and the challenges they face. With the advent of social media and online forums, the internet has become a rich source of data reflecting public opinion. Suomi24¹, the largest Finnish online forum, is an ideal platform for studying the sentiments of neighborhoods through user-generated content. Hence, we used Suomi24 as our data source driven by the following factors. First, social media has become a popular platform for expressing opinions, making it valuable for SA on topics such as urban issues, climate change, and healthcare (Huang et al., 2021; Chen and Wei, 2023). Second, online forums provide an efficient way to share information and generate

discussions, offering a wealth of informative comments that can be analyzed for sentiment (Chen and Wei, 2023). Building on this, (Lindén et al., 2023) utilized the Suomi24 data to create a manually annotated dataset² for the SA task, further highlighting the platform’s potential for research in Finnish SA.

The field of urban design and planning increasingly utilizes SA to evaluate urban environments through crowdsourced data (Tas and Sanatani, 2023). Platforms like Point of Interest (POI) databases, social media, and citizen engagement tools have proven valuable for large-scale urban assessments (Martí et al., 2019; Huang and Gartner, 2016; Tas and Sanatani, 2023). There are several techniques in Literature using NLP to extract sentiment from geo-located textual data.

However, most models focus on identifying general positive or negative sentiments within a text, with limited efforts made to train models that can detect specific aspects and their associated sentiments (Tas and Sanatani, 2023). Therefore, to expand the limited research on ABSA in Finnish suburbs, we seek to explore the following research question (RQ): *How do sentiments toward suburbs in Finland, expressed in public online discussions, vary over time, and how effectively can ABSA models capture and predict these sentiments?*

To address this research objective, we investigate public sentiment toward 12 different areas in Finland by utilizing a manually annotated Finnish-language dataset for the ABSA task. For this purpose, each neighborhood is treated as a distinct ‘aspect,’ allowing us to classify sentiments expressed toward these locations rather than identifying specific reasons for likes or dislikes associated with each neighborhood. This approach provides an overview of the sentiment trends for each neighborhood over time, using

¹<https://www.suomi24.fi/>

²<https://clarino.uib.no/comedi/editor/lb-2023012701>

the neighborhood names as predefined aspects for classification. Afterwards, we applied a combination of classical machine learning models such as Support Vector Machines (SVM) and Naive Bayes, alongside modern deep learning models like BERT³(TurkuNLP/bert-base-finnish-cased-v1)(Virtanen et al., 2019) and RoBERTa (Finnish-NLP/roberta-large-finnish-v2)⁴, to analyze sentiment towards these neighborhoods. Our results show that RoBERTa outperforms other models in accurately predicting sentiment across these neighborhoods.

Further, we expanded our analysis by using RoBERTa to predict sentiment on a larger dataset of approximately 32,000 data points for the suburbs of interest in Finland. This allowed us to track trends and shifts in sentiment over time, providing valuable insights into how public perceptions of different neighborhoods evolve.

Our findings contribute to the growing body of research on SA and urban studies, particularly in the Finnish context. The results have practical implications for regional development and policy formulation, as understanding public sentiment is key to fostering better community relations and addressing regional disparities.

2 Sentiment analysis and social media data

In the age of digitalization, the growth of social media has significantly transformed the global flow of information and the organization of social demands. Platforms such as Twitter, Facebook, and Instagram have emerged as vast repositories of real-time opinions from diverse groups (Troya et al., 2021). The vast volume of online-generated data makes manual analysis challenging and hinders the ability to identify trends in a timely manner. As a result, NLP techniques, particularly SA, have become widely used in recent literature for analyzing social media content and user feedback (Benrouba and Boudour, 2023; Xu et al., 2022). Sentiment analysis/classification is a specific type of text categorization, where the classification is based on the author's expressed attitude by understanding of the document and focusing on how the sentiment is conveyed throughout the text (Meena and Prabhakar, 2007).

³<https://huggingface.co/TurkuNLP/bert-base-finnish-cased-v1>

⁴<https://huggingface.co/Finnish-NLP/roberta-large-finn>

SA is usually performed at three levels as follows: document level, sentence level, and Aspect/category level. The objective of document-level sentiment analysis is to assess the overall sentiment of the entire document. Sentence-level sentiment analysis evaluates the overall sentiment of a sentence as a single entity and predicting the opinion expressed within that sentence (Do et al., 2019). Whereas ABSA goes deeper by identifying specific opinion targets (aspects) within the text and determining the sentiment associated with each target (Karimi et al., 2021).

SA techniques are typically divided into supervised, unsupervised, and semi-supervised learning approaches. The supervised method applies labeled data for training and evaluation. The lexicon-based approach is an example of an unsupervised method, where the sentiment of a document is determined by assessing the semantic polarity of individual words by utilizing a pre-existing dictionary that assigns sentiment polarities to words or phrases. The overall sentiment of the text is then determined based on the predominant polarity of the words. (Taboada et al., 2011).

Traditional machine learning techniques, including SVM (Ahmad et al., 2018; Sharma and Sabharwal, 2019) and Naive Bayes (Hariguna and Rachmawati, 2019; Wongkar and Angdresey, 2019), have been extensively utilized in various SA applications. Traditional machine learning methods often rely on Bag-of-Words (BoW) or Term Frequency-Inverse Document Frequency (TF-IDF) for feature extraction. However, these approaches tend to generate sparse, high-dimensional feature vectors and have limited ability to capture the word order, syntactic structure, and semantic meaning of a sentence (Kokab et al., 2022).

Recent advancements in deep learning models have surpassed the performance of traditional methods by addressing the limitations of traditional machine learning methods. Bidirectional Encoder Representations from Transformers (BERT) (Devlin, 2018) is a pre-trained model that utilizes a self-attention mechanism (Vaswani, 2017) to weigh the importance of each word in a sentence, and captures complex word dependencies regardless of their position. Unlike traditional models, BERT processes text bidirectionally to understand the full context, enabling deeper comprehension of syntactic and semantic relationships.

RoBERTa (Robustly Optimized BERT Pretraining Approach)(Liu, 2019) is a variant of the BERT

model, that improves BERT’s performance by modifying its training process and showing better performance on natural language understanding(NLU) tasks (Cortiz, 2022). RoBERTa removes BERT’s next sentence prediction task, increases the training data, and extends the training duration with larger mini-batches and more diverse data sources (Liu, 2019).

2.0.1 Related Work

While numerous studies have explored SA of social media across various domains, such as tourism (Mehra, 2023; Ali et al., 2021) and e-commerce (Vanaja and Belwal, 2018; Davoodi and Mezei, 2022), relatively few have focused on SA in the context of urban environments. As an example, Saeidi et al. (2016) focus on identifying sentiments toward specific aspects of multiple entities within the same text. This work extends traditional ABSA by allowing multiple entities and aspects in a single document and presents the SentiHood dataset, which contains user discussions about urban neighborhoods from a question-answering platform. The dataset includes annotated sentences where various aspects (e.g., safety, price, or transit) of different neighborhoods are discussed. The paper also provides strong baselines using logistic regression and LSTM models and analyzes their performance on this new, more complex SA task.

Another relevant study, Rui (2023) uses machine learning models to reveal the spatial heterogeneity in sentiment, showing a medium-high-low trend from the city center to suburban areas. Factors such as metro route density and walkability were positively correlated with sentiment in formal settlements. The findings provide urban planning recommendations for promoting positive sentiments and achieving sustainable development goals.

Wang (2023) presents a methodology for incorporating public perception into urban planning evaluation using SA. By analyzing comments from social media platforms such as Weibo and Xiaohongshu, the authors develop a public perception-oriented evaluation model. The approach captures sentiment across aspects like urban landscape, ecological environment, and living quality. The results demonstrate that H-CAN outperforms existing models, achieving 84.2% accuracy on the Yelp 2015 dataset. The study by Tas and Sanatani (2023) focuses on evaluating urban environments using geo-located data, particularly through aspect-based analysis. The authors assess different aspects

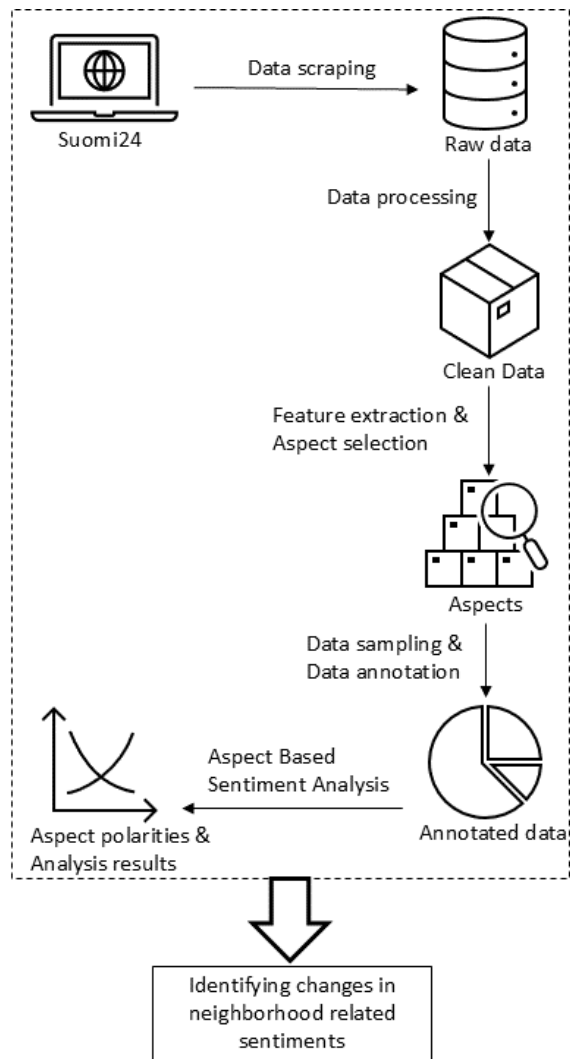


Figure 1: The stages of the research process

of urban living conditions by analyzing various sources, such as user feedback or environmental data. They annotated a dataset consisting of 2,500 crowdsourced reviews of public parks and trained a BERT model with a Local Context Focus (LCF). Their model demonstrates a substantial improvement in prediction accuracy for urban reviews.

3 Methodology

In this section, we outline our methodology for conducting ABSA using user-generated content. The objective is to analyze and understand people’s sentiments toward 12 suburban neighborhoods in Finland. By identifying sentiment changes over time for these areas, we aim to uncover patterns related to public perception and community sentiment changes.

As depicted in Figure 1, our process begins with scraping data from Suomi24, followed by data pro-

cessing. In the feature extraction and aspect selection phase, we identify and tag the specific areas of interest within the text. The data is then sampled and annotated with corresponding sentiment labels, preparing it for ABSA. The output consists of sentiment polarities for each area, which we analyze to detect trends and shifts in public sentiment over time. This methodology helps in understanding how residents and visitors perceive these neighborhoods and track any significant changes in sentiment.

3.1 Data collection and processing

To build our Finnish-language dataset, we collected data from the Suomi24 corpus, available through the Language Bank of Finland⁵. This corpus includes over 84 million messages posted on Suomi24, the largest Finnish online discussion forum (Lagus et al., 2016).

Finnish, being an agglutinative language with complex morphology, presents some challenges in text analysis (Vilkuna, 1989). However, the lemmatized corpus simplifies this by presenting words in their base forms, allowing for easier identification of lexical morphemes. For this study, we focused on twelve suburbs that have undergone systematic urban planning. To collect relevant messages, we used the suburb names as lemmas, retrieving all posts mentioning at least one of these areas from 2001 to 2017. This yielded a dataset of 36491 relevant messages. These areas, in descending order according to the number of messages are Kontula, Varissuo, Leppävaara, Myllypuro, Matinkylä, Maunula, Uittamo, Suikkila, Harittu, Runosmäki, Soukka, and Pihlajamäki. After that, we randomly selected 3183 messages to be annotated manually by 2 native Finnish speakers.

In order to perform ABSA, we opted for a manual sentiment annotation process due to the lack of a domain-specific dataset for this task in the Finnish language. To the best of our knowledge, there is no publicly available dataset in Finnish specifically for the ABSA task. In his Master's thesis, (Hellström, 2022) presents a dataset collected from public reviews on Verkkokauppa.com, focusing on laptops and tablets. The dataset consists of 1,673 sentences, which were manually labeled by the author following the guidelines established by SemEval 2014 for SA tasks (Hellström, 2022). In contrast, our dataset is tailored to analyze sen-

timent across different neighborhoods in Finland, specifically for the ABSA task, with careful resolution of any disagreements during annotation to ensure domain-specific accuracy. In addition to the ABSA dataset, we developed a lexicon-based dictionary by compiling 33742 unique words from our manually annotated dataset assigned with positive and negative scores by a native Finnish speaker annotator.

The aspects are specified as the 12 interest areas mentioned above. Three possible sentiment labels were assigned to each aspect-message pair: positive, neutral, and negative. The two annotators individually assigned sentiment values to each aspect. Every disagreement was marked and discussed to reach the final agreement between the annotators. All the messages of inadequate quality (e.g., the aspect mentioned in the text does not refer to an area) were excluded from further analysis. The final annotated dataset contains 3183 data points.

After finalizing the data, we calculated the most frequent words to find the most frequent Topics discussed in the text. For instance, the most discussed topic in the Varissuo area is immigration, with terms like "maahanmuuttaja" (immigrant) appearing 437 times, alongside related words such as "pakolainen" (refugee) with 79 occurrences and "muslimi" (Muslim) mentioned 154 times. Other terms related to demographics, such as "väestö" (population) and "kantäväestö" (indigenous population), also appear frequently, highlighting the ongoing discussions about immigration and its impact on the area. Another prominent topic is education, with "koulu" (school) being mentioned 356 times. Public transport is also commonly discussed, with terms like "bussi" (bus) at 309 occurrences and "joukkoliikenne" (public transport) appearing 155 times, indicating the importance of infrastructure in the area. Discussions about housing, crime, and services also frequently appear, reflecting various community concerns.

Finally, we performed 4 steps to clean the text: 1) transforming the text into lowercase, 2) removing all non-alpha characters, 3) removing HTML tags and URLs, 4) and eventually tagging specific aspects of interest using custom markers like '<TAG>aspect</TAG>'. Afterwards, to address the imbalance in positive sentiment data, we applied a data augmentation technique using back-translation. Back-translation has emerged as a popular data augmentation technique, where a reverse

⁵<https://www.kielipankki.fi>

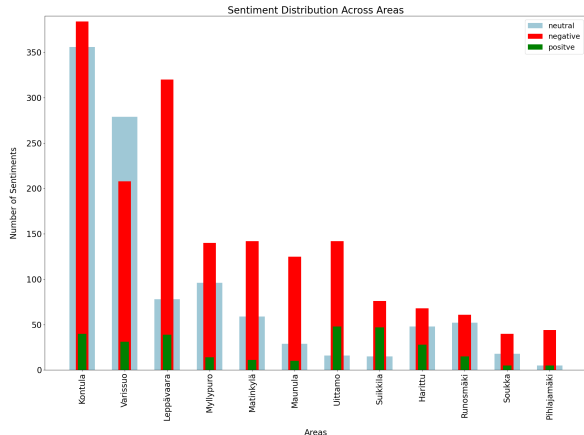


Figure 2: The distribution of Areas in Annotated Dataset

translation system is used to translate monolingual text from the target language back to the source language (Pham et al., 2023). We translated the positive text from Finnish to German and then back to Finnish, generating additional data with subtle variations while maintaining the original sentiment. This augmented data was then combined with the original dataset to achieve a more balanced distribution of positive sentiments for model training. Finally, after translating and tagging the text, we performed downsampling to address class imbalances in the sentiment labels. Since the neutral class was overrepresented, we downsampled the neutral sentiment examples to match the number of negative and positive sentiment examples. This resulted in a more balanced dataset, improving the performance of the model in training and reducing bias towards the majority class.

After performing the data processing, the dataset consists of a total of 2,688 instances. The distribution of sentiment labels within the dataset is as follows in order: Negative Sentiment (1,051 instances), Neutral Sentiment (1,051 instances), and Positive Sentiment (586 instances). The distribution of different areas is presented in Figure 2.

3.2 Machine learning models for sentiment classification

There have been several traditional machine learning (Wagner et al., 2014; Tang et al., 2019) and deep machine learning (Talaat, 2023; Tas and Sanatani, 2023) models that are proposed to solve the ABSA tasks. Based on the results of various studies, deep learning methods tend to show more adequate performance at defining the complex non-linear correlations between the features and the

sentiment polarity. In particular, neural network-based methods have outperformed other methods for ABSA tasks because these methods are trained end-to-end and can comprehend significant features automatically (Jiang et al., 2019). Considering all these observations, in this study, we have chosen models in literature for social media review as: (i) the most commonly used traditional classification models for comparison purposes (Naïve Bayes, Support Vector Machines), (ii) deep learning models that have shown the best performance in prior studies (BERT, and RoBERTa).

3.2.1 Model building

We tested different varieties of text feature extraction techniques and machine learning algorithms to specify the best-performing model. In order to transform the text into a set of understandable features for the traditional classifiers, we utilized BOW and TF-IDF. BOW converts text data into a representation based on the frequency of word occurrences within a document, ignoring the order and syntactic structure of the words (Yan et al., 2020). TF-IDF is a statistic approach that defines the weight for each term (or word) in each document based on frequency and informativeness (Soucy and Mineau, 2005).

For the Naïve Bayes and SVM classifiers, we employed BOW and TF-IDF to generate the numeric features, while for BERT and RoBERTa, we used transformer tokenizers. For the non-deep learning models, we divided the dataset into the train (80% of the data), validation (12%), and test (8%) sets using stratified sampling. The optimal cost parameters on the validation set were identified: (i) by adjusting the n-gram vectorizer parameters and regularization parameters for the linear SVM model, and (ii) by n-gram vectorizer parameters and the α value for Naïve Bayes. After defining the optimal hyperparameters for each model, we selected the best-performing model to evaluate the number of misclassifications on the test set.

In the case of transformer-based models, we used cross-validation to have a realistic estimation of our models’ performance. We split the dataset into training and test sets by using four-fold cross-validation. We utilized transformer tokenizers: AutoTokenizer for BERT, and RobertaTokenizer for RoBERTa. In both cases, we added a dropout layer to control overfitting in addition to the baseline. When training RoBERTa, the model includes the main model with 12 layers and 768 hidden dimen-

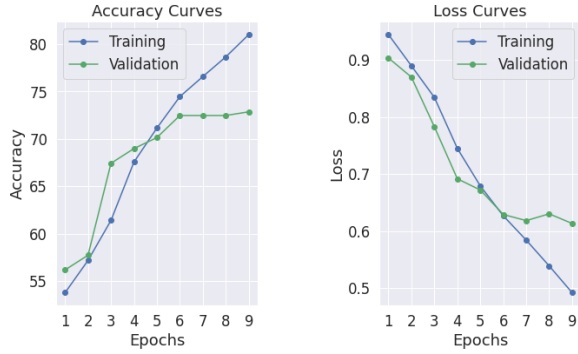


Figure 3: The performance of the best model - RoBERTa

sions. For both BERT and RoBERTa, the optimizer, loss function, and performance metric were defined as Adam, categorical cross-entropy, and accuracy. For the RoBERTa model the initial learning rate was $3e-5$, and the batch size was defined as 8.

3.3 Results

In this section, we present the results of SA, with a detailed comparison of both the performance of different models and the sentiment trends across these areas/aspects considered in the study.

3.3.1 Sentiment classification performance

To quantify the performance of each model for the ABSA task, we utilized weighted F1 score and accuracy. The results are shown in Table 1. Based on the results, RoBERTa slightly outperforms BERT in terms of the F1 score (0.75 vs. 0.74). Since F1 score is crucial for balancing precision and recall, RoBERTa is the best model for this task. Even though RoBERTa has a lower training accuracy compared to other methods (e.g., Naive Bayes or SVM), it generalizes better on the test set.

To build the best model for ABSA, a fine-tuned version of RoBERTa was utilized with a fixed dataset split to ensure robust evaluation. Prior model training, to train the best-performing model, we also utilized a lexicon-based dataset to shorten long texts. For this purpose, we filtered neutral sentences by removing sentences with an overall neutral lexicon score. The overall score was calculated by summing the lexicon scores of all the words in a sentence. If the sum was 0, the sentence was considered neutral; if the score was greater than 0, the sentence was classified as positive, and if it was less than zero, the sentence was classified as negative. Afterwards, the dataset was preprocessed to handle instances where multiple areas (aspects)

were mentioned within a single row. For rows that referenced more than one area, the sentence was split into separate rows, with each row corresponding to a specific area mentioned. After this preprocessing step, the final training dataset comprised 3,678 rows, ensuring that each entry focuses on one area for a more accurate sentiment and aspect-based analysis. Finally, the dataset was divided into fixed splits with 3,318 samples for training, 111 samples reserved for testing, and 258 samples for validation. As shown in Figure 3, the accuracy and loss curves demonstrate effective learning, with training accuracy steadily increasing while validation accuracy stabilizing after epoch 5, suggesting that the model generalizes well without significant overfitting. The number of misclassifications on the test set was 31 out of 111 datapoints. The fine-tuned RoBERTa model was subsequently employed to predict sentiment trends across the rest of the 32,183 data points.

Regarding the impact of aspect tagging, `<TAG>aspect</TAG>`, we evaluated model performance both with and without tagging. The results indicate that tagging aspects in the text improve both training and test accuracy. For instance, the tagged version of the best model, RoBERTa, achieved a test accuracy of 72.07%, compared to 63.37% for the untagged version. This suggests that tagging provides additional contextual cues, helping the model more effectively identify and associate sentiments with specific aspects. Additionally, Finnish morphology presents unique challenges, as aspect names can appear in multiple inflected forms (e.g., 'Uittamo' could appear as 'Uittamolla', 'Uittamolta', or 'Uittamolle'). Without tagging, the model may struggle to recognize these variations as referring to the same aspect, potentially impacting its ability to learn consistent sentiment associations. Tagging standardizes the aspect representation within the text, helping the model to recognize it regardless of morphological variation, which is particularly beneficial in Finnish.

3.3.2 Trend analysis

In order to calculate sentiment trends for each area, we weighted sentiments by assigning values of -2 for negative, 0 for neutral, and +2 for positive sentiments, effectively minimizing the impact of neutral sentiments. Finally, for each area, we summed the weighted sentiment scores by year and normalized the final sum by dividing it by the total

Method	Training set accuracy	Test set accuracy	F1 on test set
Naive Bayes (TFIDF)	0.95	0.67	0.65
Naive Bayes (BOW)	0.94	0.63	0.62
SVM (BOW)	0.99	0.63	0.65
SVM (TFIDF)	0.98	0.64	0.64
BERT	0.79	0.75	0.74
RoBERTa	0.76	0.75	0.75

Table 1: Performance comparison of different models on ABSA tasks

sentiment counts, providing a yearly trend score. This approach emphasizes shifts toward positive or negative sentiment and allows for clearer trend visualization. However, it may also amplify the appearance of consistently declining sentiment if negative posts dominate over time.

Across the 12 areas, several trends emerge. First, annotated datasets tend to show more fluctuation in sentiment, with several areas experiencing noticeable changes over time, either becoming more negative (e.g., Uittamo, Runosmäki) or showing improvement (Soukka, Suikkila). In contrast, the predicted datasets generally show smoother and more stable sentiment trends, often leaning towards neutrality or slight negativity. This smoothing effect may be due to the model’s tendency to generalize sentiment across larger sets of data, failing to capture more subtle changes over time.

As illustrated in Figure 4, the annotated sentiment for Harittu exhibits a general downward trend, beginning with a positive polarity and gradually declining over the years, eventually becoming predominantly negative. A similar trend is observed in the predicted dataset, though the sentiment remains more stable over time, hovering slightly below neutral. Overall, the trend indicates that public sentiment towards Harittu has remained negative over time.

Another interesting area according to the SA results is Kontula as presented in figure 5. The annotated sentiment for Kontula remains negative throughout, with only minor fluctuations, but stays relatively stable. Similarly, the predicted dataset reflects a consistent negative sentiment, showing little variation over time. Both datasets suggest that public sentiment towards Kontula has consistently remained negative with minimal changes. As another example, in the annotated data, sentiment for Varissuo (shown in Figure 6) fluctuates between slightly negative and neutral, with no strong or consistent trends emerging. The predicted sentiment

remains stable, consistently neutral, with only minor fluctuations. Both datasets indicate that sentiment towards Varissuo has remained relatively unchanged over time, with no significant shifts. We recognize that the inherent tendency of users to share more negative experiences than positive ones on social media may indeed influence the sentiment trends observed in this study. Future analyses could benefit from additional controls or normalization methods to account for the naturally higher volume of negative feedback in social media contexts, which might give a more balanced view of sentiment trends over time.

4 Discussion

Naturally occurring online discussions offer an opportunity to study the sentiments of the general public about neighborhoods from a linguistic point of view. Understanding such sentiments is crucial for regional development, urban planning, and public policy. This study aimed to investigate sentiment toward 12 different neighborhoods in Finland by leveraging user-generated content from Suomi24.

Using a combination of classical machine learning models, such as SVM and Naive Bayes, alongside state-of-the-art models like BERT and RoBERTa, we applied ABSA to analyze the sentiment associated with each neighborhood. The RoBERTa model outperformed other methods, demonstrating its superior ability to predict sentiment accurately. We extended our research to analyze sentiment trends for a larger dataset of 32,183 data points, which offers insights into how public opinion evolves over time. Our results contribute to the growing body of knowledge in SA, particularly in the context of regional perceptions, and have practical implications for urban planning and policy making.

We conducted an error analysis using the best-performing model, RoBERTa, to investigate the misclassifications made on the test set. As previ-

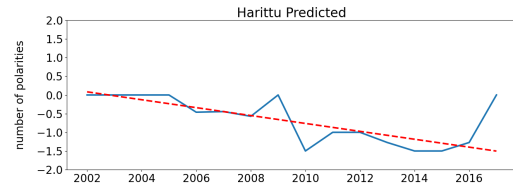
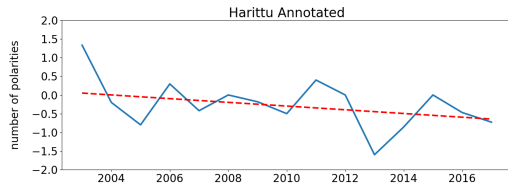


Figure 4: Harittu sentiment change overtime

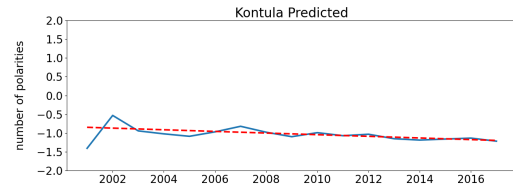
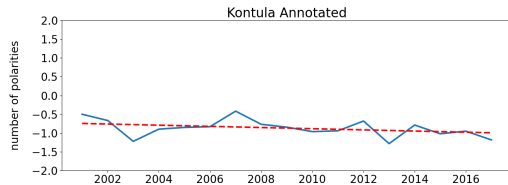


Figure 5: Kontula sentiment change overtime

ously mentioned, the lemmatized version of the text was used for both model training and evaluation. However, in this section, we present examples in their original form to enhance readability and provide clarity on the context.

A significant number of misclassifications stemmed from ambiguous or multi-layered statements where the emotional tone was not immediately clear. The model tended to predict negative sentiment when it encountered critical or confrontational language, whereas human annotators perceived such content as neutral or factual. For instance, the sentence "2 kysymystä mitä sä sitten riehut. vai uhkaako teidän perhettä <TAG>Matinkylä</TAG> syndrooma?" is neutral while the model predicted a negative sentiment. Furthermore, the model struggled with indirect sentiments, such as rhetorical or sarcastic statements, which it frequently misinterpreted as negative when the intended sentiment was neutral or sarcastic. An example is the sentence "Suomalaisen luonteen heikkous näkyy Eikö <TAG>kontula</TAG> ole oikeasti gårdsbacka ?" which criticizes the "weakness" of the Finnish character but does not directly express negativity towards the area (Kontula). Lastly, some factual statements were misclassified, likely because the model identified subtle emotional cues that the annotators did not prioritize. For instance, "Nykyään rakennetaan hometaloja ! Ei <TAG>kontulan</TAG> ostarilla ole vanhoja jykeviä kivitaloja" can be interpreted as factual. The phrase "rakennetaan hometaloja" (meaning "nowadays they build moldy houses") might imply dissatisfaction with construction practices through sarcasm. Increasing the amount of training data

could enhance the model's ability to handle these complexities more effectively.

The analysis of sentiment trends across the 12 neighborhoods reveals a range of patterns, with most neighborhoods showing either stable or slightly declining sentiment over time. Areas such as Leppävaara, Pihlajamäki, and Uittamo exhibit a subtle but consistent decline in sentiment, with annotated data indicating a shift from neutral or slightly positive to more negative sentiment. Matinkylä and Maunula remain largely neutral, with minimal fluctuations in sentiment across both annotated and predicted datasets, indicating a stable public perception. Myllypuro and Runosmäki show a predominantly negative sentiment, with Runosmäki demonstrating greater variability in the annotated data, while predicted data captures a more consistent negative trend. In contrast, Soukka and Suikkila display improvements in sentiment over time according to the annotated data, though the predicted data does not fully reflect this positive shift, suggesting a potential limitation in capturing more complex emotional changes.

5 Conclusion

This study presents a manually annotated dataset for ABSA tasks for studying different areas in Finland. As far as we know, there is no publicly available dataset for this specific domain in the Finnish language. The dataset consists of 3183 sentences taken from Suomi24. The dataset has been designed to perform various research tasks such as aspect extraction, and polarity detection. We believe this dataset would provide considerable significance in the research area of ABSA for social

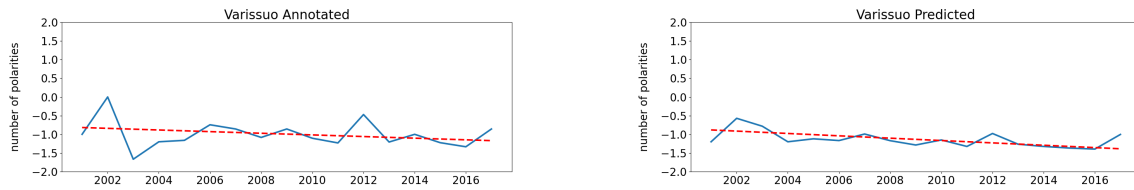


Figure 6: Varissuo sentiment change overtime

media text in the Finnish language. The dataset will be published publicly to be used by researchers in Finland. Moreover, our research addresses a significant gap by using SA to understand public perceptions of different areas in Finland. Utilizing data from Suomi24, one of the largest online forums in Finland, allows for the collection of unfiltered opinions, offering an authentic insight into how residents perceive their living environments. By applying state-of-the-art ABSA models like RoBERTa, our study enhances the analysis of regional sentiment. Additionally, tracking sentiment trends over time using 32,183 data points can provide valuable information for urban planners, policymakers, and social scientists. This combination of large-scale, time-sensitive data analysis and advanced machine learning techniques makes our research highly relevant for improving public policy and regional development. Future research could explore the relationship between sentiment and real estate pricing in Finland by analyzing how changes in public sentiment over time correlate with fluctuations in housing prices. By pairing SA with real estate data, it may be possible to predict price trends in specific areas or explain price shifts based on sentiment dynamics. A similar approach has been applied in other markets, such as the study by (Wang and Hui, 2017) which analyzed the predictive power of sentiment on market indicators like price and transaction volume in the housing market. Applying such a methodology to the Finnish housing market could provide new insights into how local sentiment influences real estate trends, potentially offering predictive power for pricing fluctuations across different areas.

Finally, a limitation of this study is the potential impact of dataset imbalance and prediction errors, which can affect the detection of complex sentiment trends in the larger predicted dataset. An imbalanced training set, with a prevalence of neutral sentiments, may lead the model to favor stable or neutral predictions. Thereby, smoothing out fluctuations that might otherwise reveal slight shifts in

sentiment. Addressing these challenges in future work by employing techniques to balance sentiment classes or reduce prediction errors could improve the model’s sensitivity to evolving sentiment dynamics across neighborhoods.

Moreover, we acknowledge that the number of posts may have varied significantly across the years, particularly as internet and social media usage expanded in Finland. However, while our focus was on examining sentiment trends rather than volume dynamics, future work could delve deeper into how this growth in user base and posting frequency influences sentiment patterns. Incorporating controls for yearly post volumes could offer a clearer view of whether observed sentiment shifts are affected by posting frequency or reflect genuine changes in public sentiment.

Acknowledgments

This research was originally part of a project funded by the Turku Urban Research Program. Its primary goal was to investigate how neighbourhood sentiments expressed in public online discussions affect housing prices in Finland using digital and statistical data, including sources such as Suomi24 discussions.

We would like to extend our sincere gratitude to Professor Jozsef Mezei for providing invaluable support with technical challenges and suggestions in the writing process.

References

- Munir Ahmad, Shabib Aftab, Muhammad Salman Bashir, Noreen Hameed, Iftikhar Ali, and Zahid Nawaz. 2018. Svm optimization for sentiment analysis. *International Journal of Advanced Computer Science and Applications*, 9(4).
- Twil Ali, Bidan Marc, Bencharef Omar, Kaloun Soulaïmane, and Safaa Larbi. 2021. Exploring destination’s negative e-reputation using aspect based sentiment analysis approach: case of marrakech destination on tripadvisor. *Tourism Management Perspectives*, 40:100892.

- Ferdaous Benrouba and Rachid Boudour. 2023. Emotional sentiment analysis of social media content for mental health safety. *Social Network Analysis and Mining*, 13(1):17.
- Kehao Chen and Guiyu Wei. 2023. Public sentiment analysis on urban regeneration: A massive data study based on sentiment knowledge enhanced pre-training and latent dirichlet allocation. *Plos one*, 18(4):e0285175.
- Diogo Cortiz. 2022. Exploring transformers models for emotion recognition: A comparison of bert, distilbert, roberta, xlnet and electra. In *Proceedings of the 2022 3rd International Conference on Control, Robotics and Intelligent System*, pages 230–234.
- Laleh Davoodi and József Mezei. 2022. [A comparative study of machine learning models for sentiment analysis: Customer reviews of e-commerce platforms](#), pages 1–14.
- Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Hai Ha Do, Penatiyana WC Prasad, Angelika Maag, and Abeer Alsadoon. 2019. Deep learning for aspect-based sentiment analysis: a comparative review. *Expert systems with applications*, 118:272–299.
- Taqwa Hariguna and Vera Rachmawati. 2019. Community opinion sentiment analysis on social media using naive bayes algorithm methods. *International Journal of Informatics and Information Systems*, 2(1):33–38.
- Rickard Hellström. 2022. [Aspect based sentiment analysis \(absa\) for finnish language](#). Master’s thesis, Aalto University.
- Haosheng Huang and Georg Gartner. 2016. Using mobile crowdsourcing and geotagged social media data to study people’s affective responses to environments. *European handbook of crowdsourced geographic information*, 385.
- Zhengwei Huang, Jing Ouyang, Xiaohong Huang, Yanni Yang, and Ling Lin. 2021. Explaining donation behavior in medical crowdfunding in social media. *Sage Open*, 11(2):21582440211014520.
- Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. [A challenge dataset and effective models for aspect-based sentiment analysis](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6280–6285, Hong Kong, China. Association for Computational Linguistics.
- Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2021. Adversarial training for aspect-based sentiment analysis with bert. In *2020 25th international conference on pattern recognition (ICPR)*, pages 8797–8803. IEEE.
- Sayyida Tabinda Kokab, Sohail Asghar, and Shehneela Naz. 2022. Transformer-based deep learning models for the sentiment analysis of social media data. *Array*, 14:100157.
- Krista Hannele Lagus, Minna Susanna Ruckenstein, Mika Pantzar, and Marjoriikka Jelena Ylisiurua. 2016. Suomi24: muodonantoa aineistolle.
- Krister Lindén, Tommi Jauhiainen, and Sam Hardwick. 2023. Finnsentiment: a finnish social media corpus for sentiment polarity annotation. *Language Resources and Evaluation*, 57(2):581–609.
- Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Pablo Martí, Leticia Serrano-Estrada, and Almudena Nolasco-Cirugeda. 2019. Social media data: Challenges, opportunities and limitations in urban studies. *Computers, Environment and Urban Systems*, 74:161–174.
- Arun Meena and Tadinada Vankata Prabhakar. 2007. Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis. In *Advances in Information Retrieval: 29th European Conference on IR Research, ECIR 2007, Rome, Italy, April 2-5, 2007. Proceedings 29*, pages 573–580. Springer.
- Payal Mehra. 2023. Unexpected surprise: Emotion analysis and aspect based sentiment analysis (absa) of user generated comments to study behavioral intentions of tourists. *Tourism Management Perspectives*, 45:101063.
- Nghia Luan Pham, Thang Viet Pham, et al. 2023. A data augmentation method for english-vietnamese neural machine translation. *IEEE Access*, 11:28034–28044.
- Jin Rui. 2023. [Exploring the association between the settlement environment and residents’ positive sentiments in urban villages and formal settlements in shenzhen](#). *Sustainable Cities and Society*, 98:104851.
- Marzieh Saeidi, Guillaume Bouchard, Maria Liakata, and Sebastian Riedel. 2016. Sentihood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. *arXiv preprint arXiv:1610.03771*.
- Dipti Sharma and Munish Sabharwal. 2019. Sentiment analysis for social media using svm classifier of machine learning. *Int J Innov Technol Exploring Eng (IJITEE)*, 8(9):39–47.
- Pascal Soucy and Guy W. Mineau. 2005. Beyond tfidf weighting for text categorization in the vector space model. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1130–1135.
- Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

- Amira Samy Talaat. 2023. [Sentiment analysis classification system using hybrid bert models](#). *Journal of Big Data*, 10(1):110.
- Feilong Tang, Luoyi Fu, Bin Yao, and Wenchao Xu. 2019. [Aspect based fine-grained sentiment analysis for online reviews](#). *Information Sciences*, 488:190–204.
- Demircan Tas and Rohit Priyadarshi Sanatani. 2023. [Geo-located aspect based sentiment analysis \(absa\) for crowdsourced evaluation of urban environments](#). *arXiv preprint arXiv:2312.12253*.
- Anina Troya, Reshmi Gopalakrishna Pillai, Dr Cristian Rodriguez Rivero, Dr Zulkuf Genc, Dr Subhradeep Kayal, and Dogu Araci. 2021. [Aspect-based sentiment analysis of social media data with pre-trained language models](#). In *Proceedings of the 2021 5th International Conference on Natural Language Processing and Information Retrieval*, pages 8–17.
- Satuluri Vanaja and Meena Belwal. 2018. [Aspect-level sentiment analysis on e-commerce data](#). In *2018 International conference on inventive research in computing applications (ICIRCA)*, pages 1275–1279. IEEE.
- A Vaswani. 2017. [Attention is all you need](#). *Advances in Neural Information Processing Systems*.
- Maria Vilkkuna. 1989. [Free word order in finnish](#).
- Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. [Multilingual is not enough: Bert for finnish](#). *arXiv preprint arXiv:1912.07076*.
- Joachim Wagner, Piyush Arora, Santiago Cortes, Utsab Barman, Dasha Bogdanova, Jennifer Foster, and Lamia Tounsi. 2014. [Dcu: Aspect-based polarity classification for semeval task 4](#). In *International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland.
- Zixiao Wang. 2023. [Emotion analysis-based decision support system for public perception evaluation in urban planning and design using social media text](#). *Soft Computing*.
- Ziyu Wang and Eddie Chi-man Hui. 2017. [Fundamentals and market sentiment in housing market](#). *Housing, Theory and Society*, 34(1):57–78.
- Meylan Wongkar and Apriandy Angdresey. 2019. [Sentiment analysis using naive bayes algorithm of the data crawler: Twitter](#). In *2019 Fourth International Conference on Informatics and Computing (ICIC)*, pages 1–5. IEEE.
- Qianwen Ariel Xu, Victor Chang, and Chrisina Jayne. 2022. [A systematic review of social media-based sentiment analysis: Emerging trends and challenges](#). *Decision Analytics Journal*, 3:100073.
- Dongyang Yan, Keping Li, Shuang Gu, and Liu Yang. 2020. [Network-based bag-of-words model for text classification](#). *IEEE Access*, 8:82641–82652.

Political Stance Detection in Estonian News Media

Lauri Lüüsi
University of Tartu
lauri.luusi@ut.ee

Uku Kangur
University of Tartu
uku.kangur@ut.ee

Roshni Chakraborty
ABV IITM Gwalior
roshni@iiitm.ac.in

Rajesh Sharma
University of Tartu
rajesh.sharma@ut.ee

Abstract

Newspapers have always remained an important medium for disseminating information to the masses. With continuous access and availability of news, there is a severe competition among news media agencies to attract user attention. Therefore, ensuring fairness in news reporting, such as, politically stance neutral reporting has become more crucial than before. Although several research studies have explored and detected political stance in English news articles, there is a lack of research focusing on low-resource languages like Estonian. To address this gap, this paper examines the effectiveness of established stance-detection features that have been successful for English news media, while also proposing novel features tailored specifically for Estonian. Our study consists of 32 different features comprising of lexical, Estonian-specific, framing and sentiment-related features out of which we identify 15 features as useful for stance detection.

1 Introduction

With the rise of the internet, the information-seeking behavior has undergone a shift such that news media has pivoted away from traditional printed newspapers towards social media and online platforms (Chakraborty et al., 2017, 2019). Furthermore, in today’s interconnected world, online news articles becoming readily available within minutes of an event occurring (Bucy et al., 2007; Chakraborty and Chakraborty, 2023). While users rely on the news media agencies for a fair and high quality reporting of news events, there has been several instances of deviation from journalistic news values in news reporting (Tandoc et al., 2021), such as, deliberately lying or leaving out context, not fact-checking sources, using clickbait, being biased (Spinde, 2021), using politically aligned news reporting (Park et al., 2022; Chakraborty et al., 2020), etc. However, with this

massive growth in news media and shift in news consumption behavior, it has become increasingly challenging and time-consuming to manually verify bias in news articles and ensure that the news articles follows journalistic standards. This is especially true for low-resource languages, where building machine learning based solutions is often more difficult due to the lack of training data. Therefore, it is important to develop and explore techniques, which use automatically extracted text features as a way to gain insight and monitor news media.

Although there are several forms of bias in news media, in this paper, we focus on political leaning or stance in news articles. While few automated stance detection concerning political leaning has been explored for different topics in English news media, such as political elections and candidates, climate change, COVID-19, and abortion rights, more extensive research is needed to enhance understanding and accuracy (ALDayel and Magdy, 2021; Farsi et al., 2024; Mohammad et al., 2016a; Neha et al., 2022; Fisher et al., 2023; Baxi et al., 2022; Chakraborty et al., 2022). However, political stance detection in Estonian news media is a mostly unexplored topic. Estonian, spoken by about a million people, has a much smaller language corpus with around 3 billion words compared to English, which has over 1.4 billion speakers (Dyvik) and a corpus of 800 billion words (Piir, 2023). Meanwhile, the rise of online news media in Estonia is significant. For example, Delfi Meedia, a major Estonian media company, has over 700,000 monthly readers and has amassed over 100,000 paid online subscribers by 2023 (Delfi Meedia; Eesti Meediattevõtete Liit). With such growth, the need for automated systems to verify news articles and detect political stances is essential.

Developing automated approaches for low-resource languages can be challenging, as these smaller languages are particularly affected by the

non-availability of task and domain-specific data (Hedderich et al., 2021). Furthermore, identifying labeled data requires manual annotation, which is time and cost intensive. Compared to English, the data can be of lower quality, which can lead to poorer results and varied performance. Therefore, it is particularly challenging to build automated models, especially train large language models, for political stance detection in Estonian news media.

In this paper, political stance in Estonian news media is analyzed on the target of immigration. Immigration is a concept that encompasses the international movement of people, usually foreign nationals (22, 2019). Immigration is a suitable target for automated stance detection, as stances towards it are varied and can often veer towards extremes (Päll, 2021). Immigrants can be viewed as strong and talented workers with great potential or, conversely, burdens on society who take jobs from locals and will not integrate into the local culture (Kosho, 2016). Media coverage of immigration can influence public opinion, especially when it adopts an overly negative stance. These shifts in attitude can potentially translate to negative treatment of immigrants, fueling racism and social division, and the enactment of discriminatory policies (Vetik, 2000). We explore and identify relevant features and techniques indicative of political stance in Estonian news media. The study identifies 15 significant features out of 32 for detecting political stance in Estonian news media. Sentences opposing immigration are longer, more complex, and used more adjectives and quotes, indicating emotionally charged language. Content analysis shows that anti-immigration texts mentioned destinations like Sweden and Germany, while supportive texts focus on transit countries like Greece and Turkey, highlighting different framing strategies. Estonian-specific features like conditionals and translatives are more prevalent in both supportive and against stances. Framing analysis uncovers distinct language use based on stance: negative terms like *illegaalne immigrant* (*illegal immigrant*) and *neeger* (*nigger*) for opposition, versus more politically neutral ones like *aafrika päritolu* (*African origin*) for support, highlighting contrasting frames in legality and humanity. Sentiment analysis shows that the XLM-RoBERTa model outperforms others, achieving the highest F1-scores across all stances.

The organization of the paper is as follows. Section 2 gives an overview of the dataset and describes the preprocessing step followed by the pro-

posed methodology in Section 3. We discuss the exhaustive analysis of extracted features and their usefulness in political stance detection in Section 4 and finally, conclude in Section 5.

2 Dataset

We use the dataset described by Mets et al. (2023) who collected 266 628 news articles from two Estonian news providers - Ekspress Grupp¹ (the parent company of Delfi Meedia) and Uued Uudised² between 2015 and 2022 on the topic of Immigration. The target is immigration, and the text is a topic-related sentence. The dataset comprises of 3261 sentences out of which 1175 sentences are of against stance, 1597 neutral stance, and 489 supportive stance towards immigration. For our study, we consider only the text of the news article. While additional meta-features, such as the title, author, publication date, and publisher, are available, they are not considered as they require prior outside knowledge about a media outlet or author and their stance on specific issues. An overview of the dataset is illustrated in Table 1. The dataset is publicly accessible on GitHub³. The code and implementation details are available on GitHub⁴.

Stance	Number of Sentences
Against	1175
Neutral	1597
Supportive	489

Table 1: Distribution of sentences in the dataset

Preprocessing Details We employ Estonian language specific preprocessing. For example, two letters with diacritics (š, ž) and sentences in the dataset which had these letters were represented by question marks or other nonsensical symbols. In order to fix this, we used EstNLTK’s SpellCheck-Retagger⁵, a tool that identifies misspellings and adds corrected forms (Laur et al., 2020), i.e., misrepresented letter was replaced by either š or ž and further, validated both by POS tagger and spell check. Additionally, we removed repeated symbols

¹<https://ekspress.delfi.ee/>

²<https://uueduudised.ee/>

³<https://github.com/markmets/immigration-prediction-EST>

⁴<https://github.com/laurilyysi/EstonianStanceDetection>

⁵<https://github.com/estnltk/>

Dataset Sentence	English Translation
aastaga tuli Eestisse 22 000 “ajutist” töolist	in a year, 22,000 “temporary” workers came to Estonia
võttes vastu inimesmugeldajate “ohvreid” aafriklaste ja teiste migrantide näol.	by accepting “victims” of human traffickers in the form of Africans and other migrants.
“Sallivuslased” aitavad neil oma soovitude ja muu “abiga” ennast hädalistena tutvustada ja mõrvarid seavad ennast “pagulastena” Euroopas sisse.	“Tolerance advocates” help them with suggestions and other “assistance” so they could present themselves as sufferers, as murderers establish themselves as “refugees” in Europe.

Table 2: Sentences where quotation marks are used to express doubt or irony.

Estonian Sentence	English Translation
Aga selleks ju migrantipaadid kehvakesed ongi, ja ilmselt lastakse need mõnda laeva märgates meelega vett täis.	But that’s exactly why migrant boats are so flimsy , and presumably they are intentionally filled with water when spotted by a ship.

Table 3: Use of diminutives in a sentence from the dataset to express a stance.

and fixed issues with missing punctuation.

3 Methodology

In this Section, we study different features that can help in identifying the political stance of Estonian news media segregated into lexical features, features specific to the Estonian language, framing-related features, and sentiment features. A complete list of all the features used is shown in Appendix A.

3.1 Lexical features

Lexical features are related to the grammar and construction of words. We consider the following lexical features, such as word count, dependency tree height, Flesch Reading Ease Score (FRES), named entities, noun phrases, adjectives, quotes, and quoted phrases. For calculating lexical features, we use EstNLTK (Tkachenko et al., 2013; Maide, 2020; Laur et al., 2020).

In order to understand sentence complexity and readability, we consider **Dependency tree height** and **Flesch Reading Ease Score (FRES)**. Dependency trees map the grammatical relationships within a sentence and indicate complexity through their height, such as, a taller tree suggests a more complex sentence structure (Nivre, 2010). Subsequently, FRES provides a numerical indication of readability, combining average sentence length and syllable count to generate a score where a higher value signifies easier readability (Zamanian and Heydari, 2012). These metrics help discern

whether sentences are structured in ways that might simplify or complicate the reader’s understanding.

Further, we study **named entities** and **noun phrases** to identify the difference in framing of sentences across different stances. These features are essential for extracting the thematic substance of texts and for understanding the emphases within a narrative (Erelt, 2013). **Adjectives** and **quotes** significantly influence the tone and suggestiveness of sentences, therefore can aid in understanding how the usage of these can impact reader perception. For example, the use of adjectives describe and modify nouns, potentially imbuing them with positive or negative connotations that can subtly influence the reader’s perception of the discussed topics. Quotes, whether marking direct speech or emphasizing irony, can alter the meaning conveyed by sentences. The use of quotes can imply skepticism or irony, potentially shifting the interpreted meaning of the text as shown in Table 2 (Schlechtweg and Härtl, 2023; van den Berg and Markert, 2020).

3.2 Estonian-specific features

We describe features that could be indicative of stance and are specific to the Estonian language. The Estonian language is morphologically complex (Mets et al., 2023) due to the abundance of verb conjugation forms and grammatical cases for nouns and adjectives (Ehala, 2009; Argus, 2009). Although these characteristics can make analyzing Estonian texts challenging, it also aids in identifying features for stance detection. Additionally, to the

Estonian Sentence	English Translation
Massimigratsiooni mahitajad aga ujutaksid kontinendid pigem migrantidega üle ja segaksid ära kogu maailma rahvastiku.	The proponents of mass migration, however, would rather flood continents with migrants and mix up the entire world population.

Table 4: Use of the conditional form in sentences from the dataset to express a stance.

Estonian Sentence	English Translation
Suurim probleem ongi see, et kogu Euroopa on sunnitud migrantidega tegelema [—]	The biggest problem is that the whole of Europe is forced to deal with migrants [—]

Table 5: Use of the superlative form in a sentence from the dataset to express a stance.

best of our knowledge, we could not find any existing research work which analyze Estonian-specific features in detail with respect to stance detection. We discuss the following Estonian-specific features: diminutives, superlatives, conditional form, translative case and indirect speech next.

Diminutives in Estonian are formed by adding the suffixes *-ke* or *-kene* to nouns and adjectives. This can alter the emotional tone of a word to express either affection or belittlement (Liivak, 2023; Kasik, 2015). For instance, the diminutive form of *lollike* (*stupid*) can imply a lack of concern, subtly shifting the stance. English does not have a consistent suffix for diminutive words, unlike Estonian where forming the diminutive is mostly uniform across nouns and adjectives. For example, Liivak (2023) highlights that out of 143 instances of diminutives, 43 were used to express a positive sentiment and 27 were used to express a negative sentiment. An example of diminutives form is shown in Table 3 and conditional form in Table 4 respectively.

In Estonian, the **superlative** is usually denoted by the suffix *-im* (*suurim* – biggest) or by the word *kõige* preceding the comparative form (*kõige kiirem* – fastest) (Erelt et al., 2020). The use of the superlative form can convey extreme opinions or positions, which can indicate stance. **Conditional forms** of verbs, ending with the suffix *-ks*, often imply that the situation being described is hypothetical or unrealistic. This form is used to express exaggerated or implausible scenarios, signaling a stance that suggests skepticism or disapproval. An example of superlative form is shown in Table 5. Similarly, nouns in the **translative case** also end with the suffix *-ks* which can suggest peculiarities or express attitudes (*Pean teda lolliks*, I think he’s stupid) (Pai, 2001) on the basis of the context.

Finally, **indirect speech** in Estonian, recogniz-

kõik, kõige (all), kunagi, eales (ever), iial (never), alati (always), igavesti (forever), tervenisti, täiesti, üleni (entirely), täitsa, täielikult (completely), üdini, läbinisti (thoroughly), läbini (through and through), absoluutne (absolute), absoluutselt (absolutely), totaalne (total), totaalselt (totally), ainult (only), ainus (sole), kogu (whole)

Table 6: List of words to detect black-and-white thinking. English translation in parenthesis.

able by verbs ending in *-vat*, is used to convey statements heard from others rather than directly from the speaker. This form enables plausible deniability, introduces uncertainty, and allows the speaker to distance themselves from the information, often reflecting a stance of skepticism or disagreement with the reported statements (Teptiuk and Tuuling, 2024).

3.3 Framing Analysis

As framing in news media can indicate towards political bias or stance (Kaukonen, 2022), we discuss next black-and-white thinking, bigram analysis and adjective-based framing that are related to framing.

Black-and-white thinking is a logical fallacy in which a complex situation is simplified into two extremes (Vleet, 2011). When authors use extreme or polarizing language, they often eliminate or do not consider alternate perspectives or possibilities. Black-and-white thinking is detected by word choice. Table 6 contains a list of hyperbolic words that could be considered polarizing. By detecting words from this list, it can be assessed whether a particular stance is being portrayed in a binary matter and lacks a middle ground. An example of a black-and-white thinking is shown in

Estonian Sentence	English Translation
Vahemere paadipõgenike ümber toimuv jätab üha enam mulje, et rändekriis hakkab kõigile närvidele käima, välja arvatud inimõiguslased ja teised sallivuslased, kes ei muutu kunagi .	The events around Mediterranean boat refugees increasingly give the impression that the migration crisis is getting on everyone’s nerves, except for human rights activists and other tolerant individuals who never change.

Table 7: Black-and-white thinking in a sentence from the dataset that expresses a stance.

Table 7.

We additionally employ **bigram analysis** in order to identify any specific word pairs associated with a negative or positive stance. We also study **adjective-based framing** to understand the framing of certain concepts (Morstatter et al., 2018). For example, we observe that the concept of immigration can be referred to as illegal, uncontrollable or unlawful in sentences for the against stance and lawful or controlled are used to frame immigration in the supportive stance.

3.4 Sentiment Analysis

While sentiment analysis focuses on the polarity of the text, stance detection focuses on the viewpoint expressed towards a specific target (Mohammad et al., 2016b). We discuss next how we study and evaluate sentiment analysis on Estonian news media text in order to understand whether it can aid in political stance understanding. For **lexicon-based** sentiment analysis in Estonian, two notable corpora are available (Regita, 2023; Mohammad and Turney, 2013). While Regita (2023) developed a lexicon of 2454 sentiment-annotated words, provided by the Institute of the Estonian Language (EKI), Mohammad and Turney (2013) introduced EmoLex which comprises of 3693 words annotated for positive and negative sentiment. Subsequently, **Emotsioonidetektor** (Pajupuu et al., 2016) classifies a text negative, neutral, or positive directly being trained on Estonian Valence Corpus (Pajupuu et al., 2016). Emotsioonidetektor differs from lexicon-based approaches since it also considers context, such as cases where a positive or negative word was negated, obtaining the opposite sentiment. Inspired by the effectiveness of **BERT** (Bidirectional Encoder Representations from Transformers) in several natural language processing based tasks (Devlin et al., 2019), **EstBERT**, an Estonian-specific BERT model, was trained on the Estonian National Corpus, which contains approximately 1.34 billion words (Tanvir et al., 2021).

This extensive dataset enabled EstBERT to outperform some multilingual BERT models in specific tasks (Tanvir et al., 2021). For sentiment analysis in this paper, both a fine-tuned EstBERT model and a multilingual XLM-RoBERTa model were used.

4 Results

In this Section, we discuss the results for features that provides a significant difference in observations results for the different stances. In total, we confirmed 15 features out of total of 32 features (as shown in Appendix A) to be useful for political stance detection in Estonian news media. All results were tested using a p-test and useful results were confirmed to have a p-value of under 0.01.

4.1 Lexical features

Our observations on comparison of the lexical features across different stances indicate that word count, dependency tree height, Flesch Readability Score, adjectives and quotes are useful whereas named entity counts did not show any difference. For example, our observations indicate sentences with against stance immigration had a higher word count, with a mean of 22.32 which is higher by 15 – 18% compared to supportive and neutral stances, respectively (shown in Appendix B Table 14). Similarly, Dependency tree height of the sentences with against stance are higher by 4 – 10% compared to supportive and neutral stances, respectively (shown in Appendix B Table 15) and Flesch Readability Score indicates that sentences with against stance are more complex by 10 – 15% compared to supportive and neutral stances (Appendix B Table 16). However, we did not observe any difference across the number of named entities used in the sentences irrespective of the stance (Appendix B Table 17). Additionally, we observe that sentences with against stance has a higher usage of adjectives and quotes (Appendix B Tables 19 and 20), thereby highlighting anti-immigration texts use more emotionally charged language. We

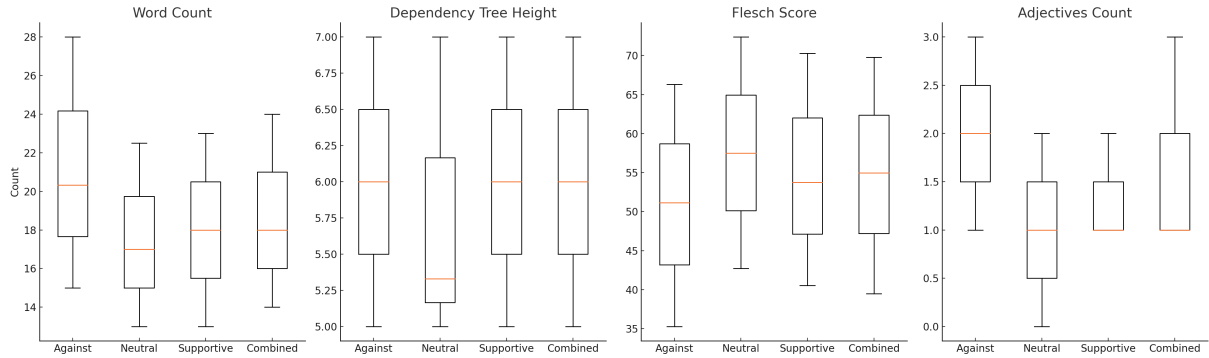


Figure 1: Box plots for Word Count, Dependency Tree Height, Flesch Score, and Adjectives Count across different stances (including a combined class, which includes all sentences).

Stance	Estonian Sentence	English Translation
Against	Rändel on suur demograafiline mõju, mis mõjutab Rootsi rahvuslikku ja kultuurilist identiteeti, samuti hävitav majanduslik mõju Rootsi heaoluriigile.	Migration has a significant demographic impact, affecting Sweden's national and cultural identity, as well as having a devastating economic impact on Sweden's welfare state.
Supportive	See, et inimesed Lesbosel elavad iseteh tud telkides vihma ja külma käes, ei ole Euroopa Liidu vääri line [—].	The fact that people on Lesbos are living in makeshift tents in the rain and cold is not worthy of the European Union [—].

Table 8: Examples of countries (and regions) in sentences for different political stances

show a summary of these results in Figure 1 and the detailed results are provided in Appendix B corresponding to each of these features.

Additionally, we explore the most frequent named entities across different stances. Our observations indicate that the against sentences mentioned Sweden and Germany, which are popular immigration destinations. In contrast, supportive sentences focused more on the immigration transit destinations such as Greece and Turkey (and their associated regions). This can be due to against sentences highlighting troubles in immigration destinations, while supportive sentences focus on the troubles immigrants go through during transit. We highlight few examples in Table 8 and the most the most frequent named entities across different stances is shown in Table 9.

4.2 Estonian-specific features

Diminutives: Diminutives were infrequent in the dataset. Out of 3261 sentences, only 5 contained a word in the diminutive form. However, three of these sentences were annotated as against, and the remaining two were neutral. Although the diminutive is typically used to sound more gentle and pos-

itive, no sentence with supportive stance towards immigration were found with this feature. Two anti-immigration sentences are shown in Table 10 which contained the word *lumehelbeke* (snowflake), a derogatory term used to mock sensitive and delicate young adults who easily take offense and cannot tolerate conflict or criticism. Both sentences also included quoted words, insinuating doubt and judgement. These findings suggest that detecting and analyzing diminutives can be useful in political stance detection. However, as this feature is uncommon in this dataset, we could not make any significant conclusion of its importance as a feature.

Superlatives: Adjectives in the superlative form were uncommon among the sentences, as only 87 sentences contained them. Although the superlative form was slightly more common in the anti-immigration stance, we could not make any significant conclusion due to the lack of data. The most common superlative adjective was *suurim* (biggest), with 24 occurrences, followed by *parim* (best) and *kõige olulisem* (most important). However, no specific superlative adjective was typical for any stance.

	Against		Neutral		Supportive	
3.	Rootsi	51	Kreeka	75	Kreeka	22
4.	Saksamaa	51	Euroopa Liit	67	Türgi	22
5.	EKRE	47	Türgi	62	Saksamaa	20
6.	Euroopa Liit	44	Saksamaa	53	Soome	20
7.	Ungari	42	Rootsi	45	Euroopa Liit	19
8.	Itaalia	31	Itaalia	45	Süüria	16
9.	Prantsusmaa	29	Süüria	41	Prantsusmaa	12
10.	Vahemeri	28	Valgevene	34	Rootsi	12
11.	Kreeka	28	Aafrika	34	Vahemeri	12
12.	Helme	25	Ungari	34	Aafrika	10

Table 9: Top 10 most common named entities per stance, skipping the top 2 for each stance, which were *Eesti* (Estonia) and *Euroopa* (Europe).

Stance	Estonian Sentence	English Translation
Against	Kui Ameerikas tuli võimule Trump, lubasid paljud Hollywoodi näitlejad samuti emigreeruda ja lumehelbeksed akendest välja viskuda, aga jäid siiski kohapeale ussitama – BLM-i suitsulõhnalised meelevaldused lubasid ennast vabalt maha maandada, Portlandis loodi koguni oma anarhistlik “autonoomia”. [—]	When Trump came to power in America, many Hollywood actors promised to emigrate and snowflakes [promised] to throw themselves out of windows, but they still stuck around to nag - the BLM smoke-smelling protests allowed them to calm down, and in Portland, an anarchist ‘autonomy’ was created [—].
Against	Hiljuti lõi “progressiivses maailmas” laineid Rootsi lumehelbeke , kes olevat justkui väljasaadetud afgaani elu päästnud – tegu oli paraku Rootsis juba tuntud kriminaaliga.	Recently, a Swedish snowflake made waves in the “progressive world” for supposedly saving the life of a deported Afghan – who was unfortunately already a known criminal in Sweden.

Table 10: Examples of diminutives in sentences for different political stances

Conditionals and Translatives: The conditional form was present in 380 sentences, as seen in Appendix B Table 21 where 14% of both supportive and against stance based sentences contain a conditional form and only 8% of the neutral stance based sentences use conditionals. Similarly, we observe that while 28% of the against stance based sentences and 26% of supportive stance based sentences use translatives while only 19% in the neutral sentences. The translative case was present in 763 sentences, as seen in Table 11. Therefore, it can be concluded that these features can aid in political stance detection and that there is a statistically significant association between stance and frequency of both conditional form ($\chi^2 = 24.78, p < 0.01$) and translative case ($\chi^2 = 27.31, p < 0.01$). However, the content of the words in both conditional form and translative

case do not reveal much insight about stance. The most common word across stances in the translative case is *näiteks* (for example). Similarly, the two most common words in the translative case are *oleks* (would be) and *peaks* (should be). These words are not indicative of stance solely on their own.

Indirect speech: Although Indirect speech was only present in 38 sentences, it was most prevalent in the against stance based sentences towards immigration (22 occurrences) and was more than in the neutral and supportive stances combined.

4.3 Framing Analysis

On analyzing the most frequently occurring bigrams across different stances for news articles on immigration, we observe that while few bigrams are generic and has been used for both the

Sentences with feature <i>translatives_count</i> not equaling 0.								
Stance	Count	Mean	Std	Min	25%	50%	75%	Max
Against	326	1.27	0.74	0.33	1	1	1.88	6
Neutral	310	1.23	0.60	0.33	1	1	1	4
Supportive	127	1.22	0.60	0.33	1	1	1	5

Table 11: Summary of statistics for feature *translatives_count*.

Model	Against			Neutral			Supportive		
	P	R	F1	P	R	F1	P	R	F1
EKI	0.47	0.39	0.43	0.57	0.47	0.51	0.27	0.51	0.35
EmoLex	0.41	0.40	0.41	0.53	0.30	0.38	0.20	0.49	0.28
Emotsioonidetektor	0.39	0.60	0.47	0.67	0.07	0.13	0.20	0.53	0.29
EstBERT	0.44	0.89	0.59	0.70	0.25	0.37	0.48	0.31	0.38
XLM-RoBERTa	0.53	0.80	0.64	0.68	0.49	0.57	0.50	0.34	0.40

Table 12: Evaluation metrics for each stance class. P – precision, R – recall, F1 – F1-score.

stances, such as *Euroopa Liit (European Union)*, *eesti keel (estonian language)*, there are several bigrams which highlight distinct framing for different stances. For example, while against stance uses *illegaalne immigrant (illegal immigrant)*, *neeger (nigger)*, *araablane (Arab)*, etc., to show their stance, supportive stance uses examples, such as, *aafrika päritolu (African origin)* and *(Eesti Pagulasabi (Estonian Refugee Aid))*. This showcases a distinct difference in tone of the news articles on the basis of stance. We show the top 10 most frequently occurring bigrams for both against and supportive stance in Appendix B Table 23. Adjective-noun pair based understanding of framing similarly reveals contrasting frames (shown in Appendix B Table 24). For example, against immigration sentences focus on illegality, threat and about the massive problem this can lead to whereas supportive stance emphasize on equality and humanity.

4.4 Sentiment analysis

We show the comparative results of sentiment analysis of the lexicon based models, EstBERT and XLM-RoBERTa model in Table 12. Our observations indicate that the XLM-RoBERTa model outperforms by ensuring highest F1-scores across all stances. Lexicon-based models and the Emotsioonidetektor model underperformed, especially on positive sentiment (Appendix B Figures 2, 3 and 4). The fine-tuned *EstBERT128_Sentiment* model, trained on the Estonian Valence Corpus, achieved

an accuracy of 0.74, while the XLM-RoBERTa model slightly outperformed it with an accuracy of 0.76 (Appendix B Figure 5).

5 Conclusion

Automated political stance detection in Estonian is highly challenging due to the lack of existing datasets and Estonian specific language processing tools. In this paper, we study political stance detection with respect to Immigration in detail. Our analysis comprises of 32 features segregated between lexical features, Estonian-specific features, framing-related features and sentiment-related features. These features were exhaustively analyzed to determine their suitability for political stance detection in Estonian news media. Our observations indicate that 15 features were shown to be helpful in political stance detection. Furthermore, to the best of our knowledge, this is the first work that explores novel political stance detection features specific to the Estonian language.

As a future direction, the rich morphology of Estonian could be studied by conducting a comprehensive frequency analysis of all cases and conjugation forms. This could reveal additional features and insights related to stance or sentiment. Additionally, there is a need to develop a more extensive multi-domain dataset focused on political stance detection in Estonian news, which would support the development of automated machine learning models in this language.

Acknowledgments

This work has been funded from the EU H2020 program under the SoBigData++ project (grant agreement No. 871042), ETAg (grant No. SLTAT21096), HAMISON project (PCI2022-135026-2), and PSG grant (PSG855).

References

2019. [Glossary on migration](#). In *International Migration Law No. 34*.
- Abeer ALDayel and Walid Magdy. 2021. [Stance detection on social media: State of the art and trends](#). In *Information Processing & Management*, 58(4):102597.
- Reili Argus. 2009. [Acquisition of estonian: some typologically relevant features](#). *Language Typology and Universals*, 62(1-2):91–108.
- Manmeet Kaur Baxi, Rajesh Sharma, and Vijay Mago. 2022. [Studying topic engagement and synergy among candidates for 2020 us elections](#). *Social Network Analysis and Mining*, 12(1):136.
- Erik Bucy, W. Gantz, and Z. Wang. 2007. [Media technology and the 24 hour news cycle](#). *Communication technology and social change*, pages 143–164.
- Roshni Chakraborty, Srishti Bhandari, Nilotpal Chakraborty, and Ritwika Das. 2020. [Eve2sign: Creating signed networks of news events](#). In *Text2Story@ ECIR*, pages 79–87.
- Roshni Chakraborty, Maitry Bhavsar, Sourav Dandapat, and Joydeep Chandra. 2017. [A network based stratification approach for summarizing relevant comment tweets of news articles](#). In *Web Information Systems Engineering–WISE 2017: 18th International Conference, Puschino, Russia, October 7-11, 2017, Proceedings, Part I 18*, pages 33–48. Springer.
- Roshni Chakraborty, Maitry Bhavsar, Sourav Dandapat, and Joydeep Chandra. 2019. [Tweet summarization of news articles: An objective ordering-based perspective](#). *IEEE Transactions on Computational Social Systems*, PP:1–17.
- Roshni Chakraborty, Maitry Bhavsar, Sourav Kumar Dandapat, and Joydeep Chandra. 2022. [Detecting stance in tweets : A signed network based approach](#). *Preprint*, arXiv:2201.07472.
- Roshni Chakraborty and Nilotpal Chakraborty. 2023. [Twminer: Mining relevant tweets of news articles](#). In *2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing Workshops (CCGridW)*, pages 1–3.
- Delfi Meedia. Delfi meedia. <https://delfimeedia.ee/>. Accessed: July 5, 2024.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#). *Preprint*, arXiv:1810.04805.
- Einar H. Dyvik. [The most spoken languages worldwide](https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/). <https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/>. Accessed: July 5, 2024.
- Eesti Meediaettevõtete Liit. [Tasulised digitellimused 2023](https://meedialiit.ee/statistika/statistika-2023/). <https://meedialiit.ee/statistika/statistika-2023/>. Accessed: July 5, 2024.
- Martin Ehala. 2009. [Linguistic strategies and markedness in estonian morphology](#). *Language Typology and Universals*, 62(1-2):29–48.
- Mati Erelt. 2013. [Nimisõnafraasi sõnajärjest](#). *Oma Keel*, 26:56–60. (Accessed on 15.05.2024).
- Mati Erelt, Tiiu Erelt, and Kristiina Ross. 2020. *Eesti keele käsiraamat*. Eesti Keele Instituut; EKSA, Tallinn.
- Salman Farsi, Asrarul Hoque Eusha, and Mohammad Shamsul Arefin. 2024. [CUET_Binary_Hackers at ClimateActivism 2024: A comprehensive evaluation and superior performance of transformer-based models in hate speech event detection and stance classification for climate activism](#). In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 145–155, St. Julians, Malta. Association for Computational Linguistics.
- Andrew Fisher, Rajesh Sharma, and Vijay Mago. 2023. [Analyzing the trends of responses to covid-19 related tweets from news stations: an analysis of three countries](#). In *International Workshop on Health Intelligence*, pages 273–288. Springer.
- Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. [A survey on recent approaches for natural language processing in low-resource scenarios](#). *Preprint*, arXiv:2010.12309.
- Reet Kasik. 2015. *Sõnamoodustus*. Tartu Ülikooli Kirjastus, Tartu.
- Elisabeth Kaukonen. 2022. [Sooliselt markeeritud sõnad eesti spordiudistes](#). *Keel ja Kirjandus*, 65(6):526–545.
- Joana Kosho. 2016. [Media influence on public opinion attitudes toward the migration crisis](#). *International Journal of Scientific & Technology Research*, 5:86–91.
- Sven Laur, Siim Orasmaa, Dage Särg, and Paul Tammo. 2020. [EstNLTK 1.6: Remastered Estonian NLP pipeline](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7152–7160, Marseille, France. European Language Resources Association.

- Mirjam Liivak. 2023. [Ke\(ne\)-liitelised deminutiivid eesti suulises argisuhtluses](#). Master's thesis, University of Tartu, Institute of Estonian and General Linguistics. (Accessed on 15.05.2024).
- Rasmus Maide. 2020. [Eesti keele nimeolemite märgendaja analüüs ja parandamine](#). (Accessed on 15.05.2024).
- Mark Mets, Andres Karjus, Indrek Ibrus, and Maximilian Schich. 2023. [Automated stance detection in complex topics and small languages: the challenging case of immigration in polarizing news media](#). Preprint, arXiv:2305.13047.
- Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016a. [A dataset for detecting stance in tweets](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3945–3952, Portorož, Slovenia. European Language Resources Association (ELRA).
- Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016b. [SemEval-2016 task 6: Detecting stance in tweets](#). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.
- Saif M. Mohammad and Peter D. Turney. 2013. [Crowdsourcing a word–emotion association lexicon](#). *Computational Intelligence*, 29(3):436–465.
- Fred Morstatter, Liang Wu, Uraz Yavanoglu, Stephen R. Corman, and Huan Liu. 2018. [Identifying framing bias in online news](#). *Trans. Soc. Comput.*, 1(2).
- Kumari Neha, Vibhu Agrawal, Vishwesh Kumar, Tushar Mohan, Abhishek Chopra, Arun Balaji Buduru, Rajesh Sharma, and Ponnurangam Kumaraguru. 2022. [A tale of two sides: Study of protesters and counter-protesters on# citizenshipamendmentact campaign on twitter](#). In *Proceedings of the 14th ACM Web Science Conference 2022*, pages 279–289.
- Joakim Nivre. 2010. [Dependency parsing](#). *Language and Linguistics Compass*, 4(3):138–152.
- Kristina Pai. 2001. [Translatiivne ja essiivne predikatiivadverbiaal eesti kirjakeeles](#). Master's thesis, University of Tartu, Institute of Philosophy.
- Hille Pajupuu, Rene Altrov, and Jaan Pajupuu. 2016. [Identifying polarity in different text types](#). *Folklore: Electronic Journal of Folklore*, 64:125–142.
- Jinkyung Park, Rahul Ellezhuthil, Ramanathan Arunachalam, Lauren Feldman, and Vivek Singh. 2022. [Toward fairness in misinformation detection algorithms](#). In *Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media*. Retrieved from <https://doi.org/10.36190>.
- Rait Piir. 2023. [Finland's chatgpt equivalent begins to think in estonian as well](#). *ERR News*. Accessed: July 5, 2024.
- Richard Päll. 2021. [SisserÄndevastaste hoiakute mõju poliitilisele usaldusele euroopa rÄndekriisi valguses: Ungari ja poola näitel](#).
- Luukas Regita. 2023. [Tartu Ülikooli õppeainete tagasi-side meelsusanalüüs](#).
- Marcel Schlechtweg and Holden Härtl. 2023. [Quotation marks and the processing of irony in english: evidence from a reading time study](#). *Linguistics*, 61(2):355–390.
- Timo Spinde. 2021. [An interdisciplinary approach for the automated detection and visualization of media bias in news articles](#). Preprint, arXiv:2112.13352.
- E Tandoc, RJ Thomas, and L Bishop. 2021. [What is \(fake\) news? analyzing news values \(and more\) in fake stories](#). *media and communication*, 9 (1), 110–119.
- Hasan Tanvir, Claudia Kittask, Sandra Eiche, and Kairit Sirts. 2021. [EstBERT: A pretrained language-specific BERT for Estonian](#). In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 11–19, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.
- Denys Teptiuk and Eda-Riin Tuuling. 2024. [Manner expressions in finnish and estonian: their use in quotative constructions and beyond](#). *Linguistics*, 62(3):577–616.
- Alexander Tkachenko, Timo Petmanson, and Sven Laur. 2013. [Named entity recognition in Estonian](#). In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 78–83, Sofia, Bulgaria. Association for Computational Linguistics.
- Esther van den Berg and Katja Markert. 2020. [Context in informational bias detection](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6315–6326, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Raivo Vetik. 2000. [Eesti elanike hoiakud poliitilise integratsiooniga seoses](#). *Integratsioon Eesti ühiskonnas. Monitoring 2000*.
- Jacob E. Van Vleet. 2011. *Informal Logical Fallacies: A Brief Guide*. University Press of America, Lanham.
- Mostafa Zamanian and Pooneh Heydari. 2012. [Readability of texts: State of the art](#). *Theory and Practice in Language Studies*, 2:43–53.

A Summary of features

	Feature Name	Description
LEXICAL	<i>word_count</i>	The number of words in a sentence.
	<i>dependency_tree_height</i>	The height of a dependency tree as calculated by EstNLTK's Maltparser model.
	<i>flesch_score</i>	The Flesch Reading Ease Score as calculated by EstNLTK's <i>SentenceFleschScoreRetagger</i> .
	<i>named_entities</i>	A list of named entities extracted by EstNLTK's named entity tagger.
	<i>named_entities_count</i>	Number of named entities in a sentence.
	<i>noun_phrases</i>	A list of noun phrases extracted by EstNLTK's experimental noun phrase chunker.
	<i>noun_phrases_count</i>	Number of noun phrases in a sentence.
	<i>adjectives</i>	Lemmas of adjectives used in a sentence.
	<i>adjectives_count</i>	Number of adjectives used in a sentence.
	<i>quotes_count</i>	Number of quotes in a sentence.
	<i>quoted_words</i>	A list of words and short phrases that are between quotes in a sentence.
	<i>quoted_words_count</i>	Number of quoted words and short phrases.
ESTONIAN-SPECIFIC	<i>diminutives</i>	A list of words that are in the diminutive form, noted by the ending <i>-ke</i> or <i>-kene</i> .
	<i>diminutives_count</i>	Number of words in the diminutive form.
	<i>superlatives</i>	A list of adjectives in the superlative form.
	<i>superlatives_count</i>	The number of adjectives in the superlative form.
	<i>conditionals</i>	A list of verbs that are in the conditional form, noted by the suffix <i>-ks</i> .
	<i>conditionals_count</i>	Number of words in the conditional form.
	<i>translatives</i>	A list of nouns that are in the translative case, noted by the suffix <i>-ks</i> .
	<i>translatives_count</i>	Number of words in the translative case.
	<i>indirects</i>	A list of verbs that are indirect, noted by the suffix <i>-vat</i> .
<i>indirects_count</i>	Number of indirect words.	
FRAMING	<i>bw_count</i>	Number of words that insinuate black and white thinking.
	<i>has_against_bigram</i>	A categorical variable indicating whether an against or supportive stance bigram or adjective used for framing was present in the sentence or not.
	<i>has_support_bigram</i>	
	<i>framing_against</i>	
<i>framing_supportive</i>		
SENTIMENT	<i>ekilex_sentiment</i>	A sentiment classification of either negative, neutral, or positive, as determined by the respective model.
	<i>emolex_sentiment</i>	
	<i>eki_emotion</i>	
	<i>estbert_sentiment</i>	
	<i>xlmroberta_sentiment</i>	

Table 13: Summary of features.

B Tables and figures of results

Stance	Count	Mean	Std	Min	25%	50%	75%	Max
Against	1175	22.32	10.19	3	15	20.33	28	94
Neutral	1597	18.31	8.04	3	13	17	22.5	60
Supportive	489	18.95	7.95	4	13	18	23	52
Combined	3261	19.85	9.06	3	14	18	24	94

Table 14: Summary of statistics for feature *word_count*.

Stance	Count	Mean	Std	Min	25%	50%	75%	Max
Against	1175	6.33	1.94	2	5	6	7	18
Neutral	1597	5.72	1.71	2	5	5.33	7	17
Supportive	489	6.06	1.73	2	5	6	7	13
Combined	3261	5.99	1.82	2	5	6	7	18

Table 15: Summary of statistics for feature *dependency_tree_height*.

Stance	Count	Mean	Std	Min	25%	50%	75%	Max
Against	1175	49.46	25.37	-91.73	35.29	51.13	66.30	123.93
Neutral	1597	56.76	22.74	-36.52	42.73	57.50	72.38	134.12
Supportive	489	54.56	23.11	-29.21	40.53	53.76	70.30	129.57
Combined	3621	53.80	24.00	-91.73	39.49	54.96	69.79	134.12

Table 16: Summary of statistics for feature *flesch_score*.

<i>named_entities_count</i> across all sentences								
Stance	Count	Mean	Std	Min	25%	50%	75%	Max
Against	1175	1.54	1.43	0	0	1	2	10
Neutral	1597	1.60	1.56	0	0	1	2	11
Supportive	489	1.55	1.60	0	0	1	2	11
Combined	3621	1.57	1.53	0	0	1	2	11
Sentences with feature <i>named_entities_count</i> not equaling 0.								
Against	897	2.01	1.33	0.33	1	2	3	10
Neutral	1185	2.16	1.45	0.25	1	2	3	11
Supportive	359	2.11	1.52	0.50	1	2	3	11
Combined	2441	2.10	1.42	0.25	1	2	3	11

Table 17: Summary of statistics for feature *named_entities_count*.

<i>noun_phrases_count</i> across all sentences								
Stance	Count	Mean	Std	Min	25%	50%	75%	Max
Against	1175	3.11	1.87	0	2	3	4	17
Neutral	1597	2.83	1.69	0	2	3	4	15
Supportive	489	2.82	1.68	0	2	3	4	11
Combined	3621	2.93	1.76	0	2	3	4	17
Sentences with feature <i>noun_phrases_count</i> not equaling 0.								
Against	1132	3.23	1.80	0.50	2	3	4	17
Neutral	1529	2.95	1.62	0.50	2	3	4	15
Supportive	456	3.02	1.56	1	2	3	4	11
Combined	3117	3.06	1.67	0.50	2	3	4	17

Table 18: Summary of statistics for feature *noun_phrases_count*.

<i>adjectives_count</i> across all sentences								
Stance	Count	Mean	Std	Min	25%	50%	75%	Max
Against	1175	2.18	1.81	0	1	2	3	11
Neutral	1597	1.50	1.47	0	0	1	2	13
Supportive	489	1.64	1.60	0	1	1	2	12
Combined	3621	1.77	1.66	0	1	1	3	13
Sentences with feature <i>adjectives_count</i> not equaling 0.								
Against	1132	3.23	1.80	0.50	2	3	4	17
Neutral	1529	2.95	1.62	0.50	2	3	4	15
Supportive	456	3.02	1.56	1	2	3	4	11
Combined	3117	3.06	1.67	0.50	2	3	4	17

Table 19: Summary of statistics for feature *adjectives_count*.

Stance	Count	Mean	Std	Min	25%	50%	75%	Max
Against	104	1.05	0.60	0.25	0.63	1	1	4
Neutral	53	1.06	0.55	0.25	1	1	1	3
Supportive	27	1.18	0.67	0.25	1	1	1	3
Combined	184	1.07	0.60	0.25	1	1	1	4

Table 20: Summary of statistics for feature *quoted_words_count*.

Sentences with feature <i>conditionals_count</i> not equaling 0.								
Stance	Count	Mean	Std	Min	25%	50%	75%	Max
Against	167	1.20	0.72	0.33	1	1	1	4
Neutral	140	1.15	0.55	0.50	1	1	1	5
Supportive	73	1.14	0.49	0.33	1	1	1	3
Combined	380	1.17	0.62	0.33	1	1	1	5

Table 21: Summary of statistics for feature *conditionals_count*.

Stance	Count	Mean	Std	Min	25%	50%	75%	Max
Against	233	0.99	0.41	0.25	1	1	1	3
Neutral	175	0.99	0.33	0.33	1	1	1	3
Supportive	62	1.02	0.29	0.33	1	1	1	2
Combined	470	0.99	0.37	0.25	1	1	1	3

Table 22: Summary of statistics for feature *bw_count*.

	Against		Supportive	
1.	(euroopa, liit) (european, union)	38	(euroopa, liit) (european, union)	29
2.	(mart, helme) (mart, helme)	18	(eesti, keel) (estonian, language)	11
3.	(eesti, keel) (estonian, language)	18	(euroopa, komisjon) (european, commission)	8
4.	(konservatiivne, rahvaerakond) (conservative, peoples party)	14	(miljon, euro) (million, euro)	7
5.	(illegaalne, immigrant) (illegal, immigrant)	13	(välismaalane, seadus) (foreigner, law)	6
6.	(kogu, euroopa) (whole, [of] europe)	12	(süüria, põgenik) (syrian, refugee)	5
7.	(martin, helme) (martin, helme)	11	(eesti, pagulasabi) (estonian, refugee aid)	5
8.	(tooma, kaasa) (bring, along)	11	(aafrika, päritolu) (african, origin)	5
9.	(eesti, konservatiivne) (estonian, conservative)	10	(sisseränne, piirarv) (immigration, limit)	5
10.	(neeger, araablane) (negro, arab)	10	(globaalne, ränderaamistik) (global, migration framework)	5

Table 23: Top 10 most common bigrams in the against and supportive stances with English translations. Bigrams of interest are bolded.

	Against		Supportive	
1.	(illegaalseid, immigrantide) (illegal, immigrants)	8	(ebaseadusliku, rände) (unlawful, migration)	3
2.	(odava, tööjõu) (cheap, labour)	8	(rahvusvahelist, kaitset) (international, defense)	3
3.	(konservatiivne, rahvaerakond) (conservative, peoples party)	7	(rahvusvahelise, rändekava) (international, migration plan)	2
4.	(massilise, sisserrände) (massive, immigration)	4	(soolise, võrdõiguslikkuse) (gender, equality)	2
5.	(uute, uudiste) (new, news)	4	(avatud, algus) (open, beginning)	2
6.	(illegaalse, immigratsiooni) (illegal, immigration)	4	(salliva, õpikeskkonna) (tolerant, learning environment)	2
7.	(uus, valitsus) (new, government)	3	(kogu, maailmas) ([in the] entire, world)	2
8.	(uued, uudised) (new, news)	3	(suure, panuse) (big, contribution)	2
9.	(illegaalseid, immigrante) (illegal, immigrants)	3	(globaalse, ränderaamistikus) (global, migration framework)	2
10.	(suur, probleem) (big, problem)	3	(rassilise, diskrimineerimise) (racial, discrimination)	2

Table 24: Top 10 most common adjective-noun pairs in the against and supportive stances with English translations. Pairs of interest are bolded.

Against	Terms in both	Supportive
agressiivne, allaheitlik, avantüristlik, efektiivne, elama, isiklik, islamiusuline, jahtiv, järgmine, jätkuv, kahjulik, kogu, konservatiivne, kriminaalne, kuritahtlik, käiv, kõrge, lõtv, ohtlik, paarituhandeline, potentsiaalne, range, rekordkõrge, riiklik, salakaval, sarnane, sealne, senine, seotud, suunduv, suvaline, tark, teisene, toimuv, tugevnev, tuntud, tülikas, valimatu, ähvardav, ühine, üksik, üleeuroopaline	lähtuv, piiramatu, kasvav	esitatud, hiiglaslik, inimlik, laiahaardeline, lubatud, noor, oluline, seaduslik, tõstatatud, vaba, väärikas, üleilmne
aggressive, submissive, adventurous, efficient, living, personal, muslim, hunting, next, ongoing, harmful, entire, conservative, criminal, malicious, ongoing, high, relaxed, dangerous, a few thousand, potential, strict, record high, national, cunning, similar, local, previous, related, heading, arbitrary, smart, secondary, occurring, strengthening, known, troublesome, indiscriminate, threatening, common, single, pan-european	originating, unlimited, growing	submitted, gigantic, humane, extensive, permitted, young, important, legal, raised, free, dignified, global

Table 25: Lemmatized list of unique adjectives in Estonian used to frame immigration, that preceded the stems *immigra* and *rän*. Translation in English is added.

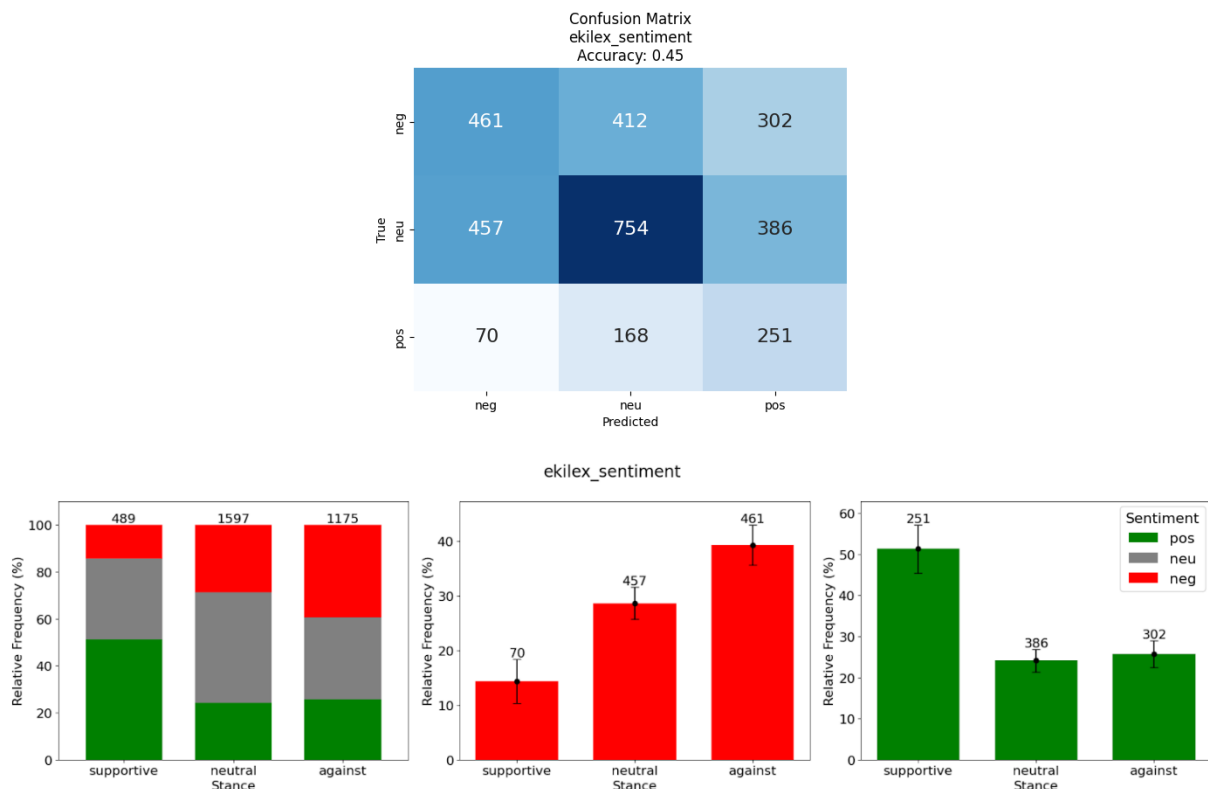


Figure 2: Confusion matrix and relative frequency graphs for sentiment predictions using the lexicon provided by the Institute of the Estonian Language (EKI).

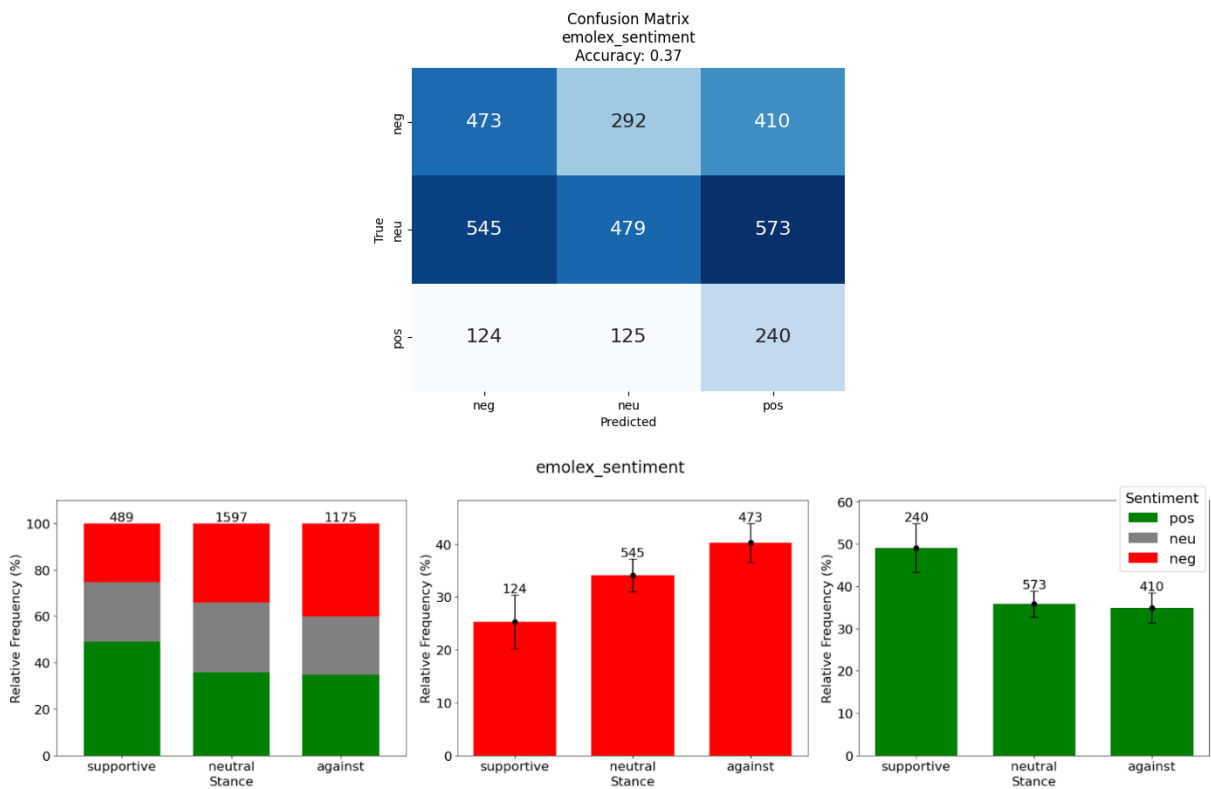


Figure 3: Confusion matrix and relative frequency graphs for sentiment predictions using the EmoLex lexicon.

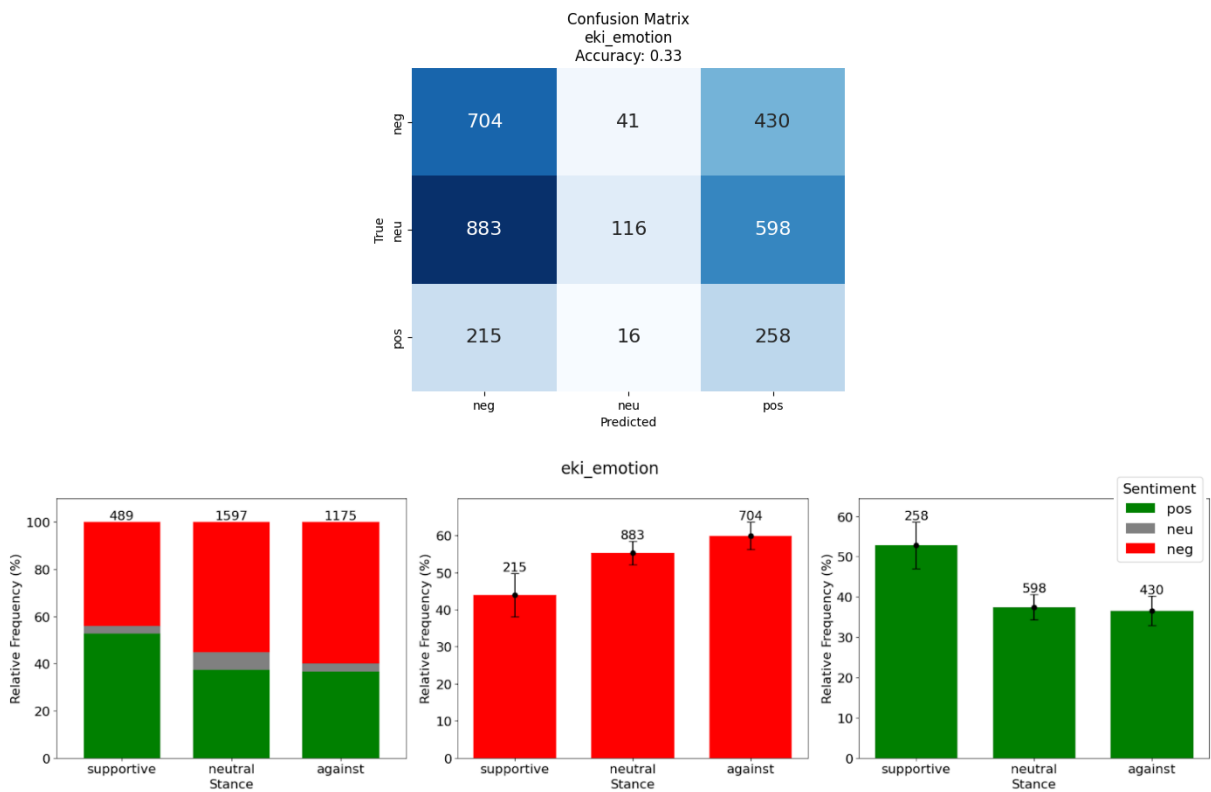


Figure 4: Confusion matrix and relative frequency graphs displaying the results of Emotionidetektor predictions.

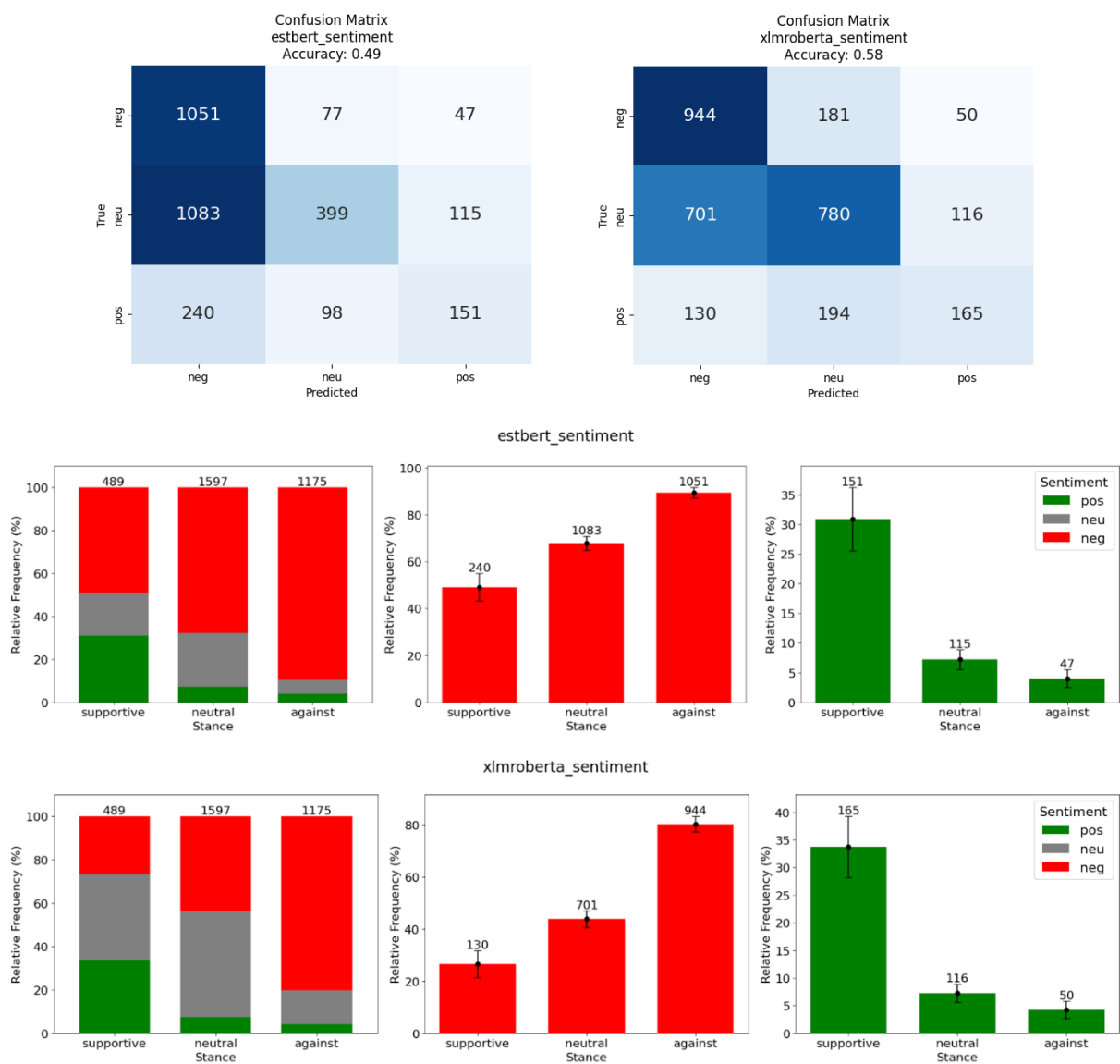


Figure 5: Confusion matrices and relative frequency graphs displaying the results of BERT sentiment model predictions.

Universal-WER: Enhancing WER with Segmentation and Weighted Substitution for Varied Linguistic Contexts

Samy Ouzerrout

University of Orléans

France

samy.ouzerrou@etu.univ-orleans.fr

Abstract

Word Error Rate (WER) is a crucial metric for evaluating the performance of automatic speech recognition (ASR) systems. However, its traditional calculation, based on Levenshtein distance, does not account for lexical similarity between words and treats each substitution in a binary manner, while also ignoring segmentation errors.

This paper proposes an improvement to WER by introducing a weighted substitution method, based on lexical similarity measures, and incorporating splitting and merging operations to better handle segmentation errors.

Unlike other WER variants, our approach is easily integrable and generalizable to various languages, providing a more nuanced and accurate evaluation of ASR transcriptions, particularly for morphologically complex or low-resource languages.

1 Introduction

Automatic speech recognition (ASR) is now ubiquitous in our daily lives, facilitating translation, video transcription, note-taking, and interactions with voice assistants. While advances in deep learning models have significantly improved ASR system accuracy, challenges remain, particularly for underrepresented and morphologically complex languages. Despite these advancements, evaluating the performance of ASR systems remains essential to ensure their accuracy and reliability.

Word Error Rate (WER) is still the benchmark metric for evaluating transcription quality, but it relies on the Levenshtein distance, which does not account for lexical imprecision or segmentation errors, limiting its relevance in the face of linguistic diversity.

Low-resource languages often exhibit complex morphological structures (Lupyan and Dale, 2010),

making them particularly vulnerable to segmentation errors, especially in the case of agglutinative languages.

WER applies a double penalty to these errors, artificially inflating the error rate. Furthermore, these languages are often characterized by high dialectal diversity, leading to inappropriate penalties for variations that are not inherently errors. Lastly, WER treats all lexical substitutions in a binary manner, overlooking minor variations that could be considered acceptable. These limitations highlight the need for a more precise evaluation, better suited to linguistic diversity.

Two main approaches stand out to improve WER calculation: on the one hand, models incorporating weightings based on word meaning, such as the Weighted Word Error Rate (WWER) (Shichiri et al., 2007), and on the other hand, methods like the Phoneme Error Rate (PER) (He and Radfar, 2021), which assess recognition at the phonemic level. Recently, evaluation methods based on language models have also emerged. However, these solutions have limitations, particularly in terms of implementation complexity and generalization to all languages.

Our work proposes an improved version of WER, tailored to the specificities of ASR transcriptions, by introducing weighted substitution based on lexical similarity measures, as well as splitting and merging operations to better handle segmentation errors. This approach aims to ensure adaptability to various languages and different usage contexts.

2 Introducing WER and the Levenshtein Distance

The *Word Error Rate (WER)*, the main metric used to evaluate the performance of ASR systems, cal-

calculates an error rate: the lower the rate (with a minimum of 0), the better the recognition. The maximum rate is unbounded and can exceed 1 (Wikipedia, 2023). The WER formula is given by:

$$WER = \frac{S + D + I}{N}$$

This calculation is based on the *Levenshtein distance*, an algorithm that measures the similarity between two sequences by counting the minimum number of operations required to transform one sequence into another (Levenshtein, 1966). The algorithm recognizes three operations:

- S is the number of **substitutions** (errors where one word is replaced by another),
- D is the number of **deletions** (missing words in the transcription),
- I is the number of **insertions** (extra words added compared to the reference text),
- N is the total number of words in the reference text.

The algorithm works by constructing a matrix where each cell represents the alignment cost (by insertion, deletion, or substitution) of a segment from the input sequence (transcription) with a segment from the target sequence (reference text). The cost calculation is performed iteratively, comparing the elements of the two sequences.

		t	a	c
	0	1	2	3
c	1	1	2	2
a	2	2	1	2
t	3	2	2	2

Figure 1: example of a matrix for aligning the sequences "cat" and "tac".

2.1 Substitution Cost Calculation

The Levenshtein distance calculates the shortest path in the matrix by combining the costs of insertion, deletion, and substitution. The costs of insertion and deletion are fixed at 1. Regarding substitution, the algorithm assigns a cost of **0** if the units being compared are identical and a cost of **1** if they differ.

This mechanism, called *binary substitution*, means that the units are considered either entirely

identical or different. Each cell of the matrix is defined as the minimum between:

$$D(i, j) = \min \begin{cases} D(i-1, j) + 1 & \text{(case of a deletion)} \\ D(i, j-1) + 1 & \text{(case of an insertion)} \\ D(i-1, j-1) + \text{sub_cost}(A[i], B[j]) & \text{(case of a substitution)} \end{cases}$$

where the *substitution cost* is defined as:

$$\text{sub_cost}(A[i], B[j]) = \begin{cases} 0 & \text{si } A[i] = B[j] \\ 1 & \text{si } A[i] \neq B[j] \end{cases}$$

Thus, minimum cost to transform one sequence into another is obtained by following the minimal cost path in this matrix. This mechanism is essential for WER calculation, but it has several limitations due to its application to whole words.

3 Challenges of WER Based on Levenshtein Distance

This cost calculation method is effectively used in the character error rate rate (CER), where the comparison units are individual characters. In this context, each character is compared to another, and the substitution decision is naturally binary: either the units are identical (cost of 0), or they differ (cost of 1).

However, the WER, which compares entire words, has significant limitations, as highlighted by (Shigeki et al., 2023). Due to its binary approach, the WER mainly compares orthographic forms rather than the words themselves, which penalizes minor variations, such as "advisor" and "adviser."

These orthographic variations also include space insertions, as in "doghouse" and "dog house," which are double-counted in WER calculation. This type of situation is treated as a segmentation error.

These limitations affect all languages, but they are particularly pronounced in languages with complex morphology, minority languages, and those with limited resources.

3.1 Weaknesses of Binary Substitution

The binary logic of the Levenshtein distance in substitution cost calculation is problematic in the context of WER, as it treats words as homogeneous entities, without considering their lexical similarity.

For example, the words "hello" and "allo" are phonetically and orthographically closer than "hello" and "sunny." However, Levenshtein distance assigns the same substitution cost (1) to both word pairs, thus failing to distinguish minor errors from major ones.

Traditional WER lacks any mechanism to weight errors based on lexical similarity. As a result, two words differing only by minor variations are treated as if they were significantly divergent.

This approach oversimplifies linguistic errors, significantly limiting WER's ability to accurately assess the performance of ASR systems.

3.2 Segmentation Errors

Levenshtein distance does not account for segmentation errors, such as word splitting or merging, which are common in ASR transcriptions.

For example, if "keyboard" is transcribed as "key board," traditional WER calculation treats this as two distinct errors: a substitution and an insertion. However, this is actually a single splitting error.

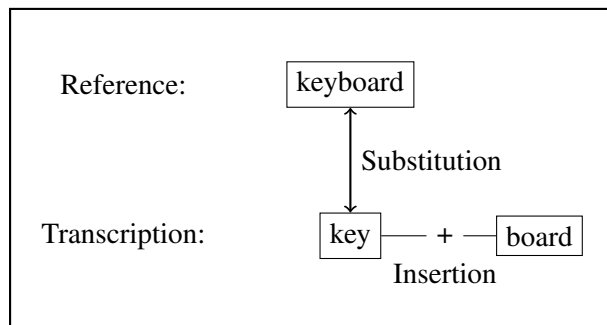


Figure 2: Double counting of segmentation errors.

Similarly, when a compound expression like "ice cream" is transcribed as a single word "icecream," this constitutes a merging error.

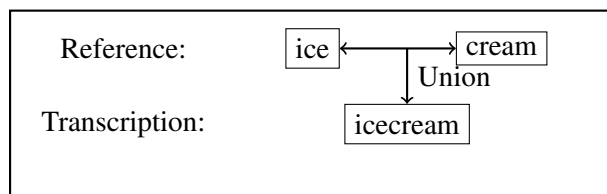


Figure 3: Merging error in transcription

The omission of segmentation errors leads to an inaccurate evaluation of transcriptions, overlooking aspects specific to speech recognition. Moreover, these segmentation issues are often considered less severe than insertions and deletions.

3.3 Morphologically complex languages

Morphological richness poses major challenges for ASR systems, which struggle to handle word inflections (prefixes, suffixes, etc.), thereby increasing the number of lexical forms and creating rare or unseen structures in training data (Morris, 2021).

In highly inflected languages, even small lexical variations can have a disproportionate impact on ASR performance. A simple error in a suffix or internal inflection can significantly increase WER, despite an otherwise accurate transcription. WER, by treating each word as a whole unit, does not account for this morphological variability.

Agglutinative languages, such as Finnish and Estonian, present particular challenges for speech recognition due to their morphological complexities. Words are formed by concatenating roots with numerous affixes, resulting in long lexical units and generating multiple word forms. This presents several difficulties for ASR systems:

- **Vocabulary explosion** : For Finnish, a lexicon of 400,000 words can still lead to a high rate of out-of-vocabulary words (Kurimo et al., 2006). This complicates the accurate transcription of these unknown words, and WER, not accounting for this complexity, severely penalizes variations that might be considered minor in the context of these languages.
- **Segmentation errors**: A poorly trained ASR system might split these elements into multiple words or merge them incorrectly, leading to multiple errors in the Word Error Rate (WER) calculation.

3.4 Minority and Low-Resource Languages

Often characterized by complex morphology, minority languages are subject to the same constraints mentioned earlier.

In evolutionary linguistics, (Lupyan and Dale, 2010) showed that languages spoken by large cosmopolitan communities, with many non-native speakers, tend to simplify their morphology over time. In contrast, minority languages, spoken in smaller communities, generally retain complex morphological structures. Native speakers of these languages share an intuitive understanding of these complex rules, allowing the language to preserve these features.

Similarly, (Lindenfelser, 2020) explains that languages with fewer non-native speakers or those

that have not been significantly influenced as a second language (L2) tend to retain or even develop complex morphological systems, such as elaborate inflection systems for nouns or verbs.

Although low-resource languages are often minority languages, some are also widely spoken. This lack of data imposes various constraints on ASR systems:

- **Lack of diverse data:** Limited and insufficiently diverse corpora affect the ability of ASR models to correctly recognize lexical and linguistic variations.
- **Transcription errors related to data quality:** Errors often stem from poor quality or lack of standardization in transcriptions, rather than an intrinsic weakness of the system.
- **Inability to handle dialectal variations:** The same word or phrase may be pronounced differently depending on the region, dialect, or speaker. Models trained on a standard form (or specific dialect) often fail to recognize variants from other regions.
- **Difficulty in handling accents:** The phonetic diversity is often vast but under-documented, complicating ASR models' ability to accurately process these regional variations or accents.
- **Low phonological standardization:** The lack of formal rules for pronunciation and segmentation makes it difficult for ASR models to manage words effectively.

These training limitations lead to multiple errors, disproportionately increasing the WER, even when the divergences do not reflect actual inaccuracies.

4 Proposed New Method for WER Calculation

4.1 Lexical Measures for Substitution Cost

To overcome the limitations of binary substitution in WER calculation, we introduce the use of lexical similarity measures such as the Jaccard index, CER (character error rate), or cosine similarity. These measures calculate a continuous dissimilarity cost between 0 and 1, reflecting the actual difference between words.

$$sub_cost(A[i], B[j]) = similarity(A[i], B[j])$$

Table 1 presents the error rates (in percentage) for different word pairs. the higher the value, the more dissimilar the words are. The last column indicates the algorithmic complexity of each method.

It is important to note that CER can exceed 100% when the transcription is significantly longer than the reference word, due to its calculation based on Levenshtein distance, which penalizes excessively long transcription sequences.

While these measures can be combined to offer a holistic evaluation, this increases the complexity of the process and, therefore, the execution time.

4.2 Split and Merge Operations

We introduce two new operations for WER calculation: splitting and merging. These operations aim to correct common segmentation errors in transcriptions produced by ASR systems. A word may be incorrectly split into two segments or, conversely, merged into one.

For example, when the word "input" is transcribed as "in put," a single merging operation would suffice to correct this error, rather than treating it as two distinct errors. By incorporating these operations into WER calculation, our approach improves the accuracy of this metric by accounting for word segmentation errors in ASR transcriptions.

To incorporate these operations within the Levenshtein algorithm, we add the following conditions:

```
// Separation
if (j > 1 and (transcript[j] == reference[i-1] +
reference[i])) then
  d[i, j] := min(d[i, j], d[i-2, j-1] +
seg_Cost)

// Union
if (i > 1 and (reference[i] == transcript[j-1] +
transcript[j])) then
  d[i, j] := min(d[i, j], d[i-1, j-2] +
seg_Cost)
```

Splitting and merging errors, being less severe than insertions and deletions, can be given a reduced cost. Moreover, these errors can be treated as adding or removing a character from a word. The cost can thus be calculated using the CER, which is simplified in this configuration:

$$seg_cost = \begin{cases} 1/len(reference) & \text{(cost based on CER)} \\ 1 & \text{(standard cost)} \\ 0.5 & \text{(reduced cost)} \end{cases}$$

Method	hello allo	kitten sitting	intention execution	diner dinner	O(n)
Cosine similarity	36.7%	38.3%	38.5%	5.1%	O(n+m)
Fuzzy Wuzzy	33.0%	38.0%	43.9%	8.9%	O(n*m)
Jaro	21.7%	25.4%	36.3%	5.5%	O(n*m)
Sorensen Dice	42.9%	63.6%	60.0%	11.1%	O(n+m)
CER	50.0%	42.86%	55.55%	16.66%	O(n*m)
LCS similarity	40.0%	42.86%	44.44%	16.66%	O(n*m)
Jaccard LCS	50.0%	55.55%	61.54%	16.66%	O(n*m)

Table 1: Comparison of similarity measures with complexity.

4.3 Experimental Analysis

In this study, we carried out transcriptions of Finno-Ugric languages (Finnish, Meadow Mari, and Hill Mari), as well as Dutch and Afrikaans, using the MMS model. The performance of the transcriptions was evaluated using WER, CER, and our UWER version.

Language	WER	CER	UWER
Finnish	0.691	0.136	0.161
Meadow Mari	0.636	0.151	0.242
Hill Mari	0.922	0.313	0.471
Afrikaans	0.384	0.106	0.141
Dutch	0.477	0.104	0.134

For UWER calculation, the segmentation cost (seg_cost) was adjusted according to the CER. Tests with costs set to 1 and 0.5 showed minimal differences, as illustrated below:

cost = cer	cost = 1	cost = 0.5
0.226	0.22837	0.22734
0.24172	0.2473	0.2441

To better visualize the impact of these error rate differences, here are some examples of reference sentences and their transcriptions, with the two measures compared (table 2).

5 Discussion

5.1 Improvement

- **Acoustic versus linguistic errors:** WER does not distinguish between errors caused by acoustic factors (noise, pronunciation) and those of a linguistic nature, assigning them equal weight in the score calculation.

Reference	Transcription	WER	UWER
ja minä huokasin kevennyksestä	ja mina huokasin kevenyksest	0.50	0.10
kaisa syötteli por-sasta	kaisa syoteli por-sasta	0.33	0.08
oletpa tosiaan lapsellinen	olet pa tosian lapselinen	1.33	0.12
ik ben daar heel blij mee	ik ben dar hel blij me	0.50	0.12
de beatles waren van liverpool	da bitels uaren fan liverpul	1.00	0.36
naaktslakken hebben geen slakkenhuis	naktslaken heben gen slakenhuis	1.00	0.14

Table 2: WER and UWER Comparison

- **Equal penalty for all types of errors:** Although we introduced a dynamic penalty for substitutions, it remains fixed for insertions and deletions. a penalty proportional to the length of inserted or deleted words could, among other things, help mitigate the impact of noise.
- **Combined errors:** Our experimental analyses show that when segmentation and lexical errors are combined, even our metric no longer accurately reflects the transcription quality. For example:

Reference	Transcription	WER	UWER
tervetuloa	tervet tuloa	2.00	1.34
slaapwel	slap wel	2.00	1.43

To address this, segmentation operations should be replaced by *substitution_separation* and *substitution_union*, applied without the requirement for equality.

The cost would then be:

$$\begin{cases} \text{seg_cost} + \text{similarity}(\text{ref}[i-1] + \text{ref}[i], \text{hyp}[j]) \\ \text{(for separation)} \\ \text{seg_cost} + \text{similarity}(\text{ref}[i], \text{hyp}[j-1] + \text{hyp}[j]) \\ \text{(for union)} \end{cases}$$

5.2 Comparison with Other Methods

The Phoneme Error Rate (PER) and Weighted Word Error Rate (WWER) are variants of WER that attempt to address some of its limitations.

PER (Shichiri et al., 2007) focuses on errors at the phoneme level, offering finer granularity than WER. However, it requires phonetic transliteration of both the transcription and the reference text, making generalization more difficult.

WWER (He and Radfar, 2021), on the other hand, assigns different weights to deletion, insertion, and substitution errors, optimized using dictionaries to weigh words based on their importance. However, this approach relies on the creation of specific linguistic resources and does not sufficiently discriminate substitution costs, limiting its effectiveness.

Apple’s "Humanizing WER" method (Apple, 2024) and the work of Hughes (Hughes, 2023) use advanced language models to improve the evaluation of speech recognition systems. HWER weights errors according to their context, offering an evaluation closer to human perception. Despite their potential, these approaches have limitations: complexity of implementation, lack of standardization, potential subjective biases, and difficulty in applying to low-resource languages due to their reliance on language models.

6 Conclusion and Future Directions

This study has highlighted the limitations of WER, especially its inability to account for lexical nuances and segmentation errors, making it unsuitable for morphologically complex or low-resource languages.

We proposed an improved version of WER, which introduces weighted substitution based on lexical similarity, as well as splitting and merging operations. Experimental results show that UWER improves evaluation accuracy across several languages.

Our approach aims to ensure WER’s adaptability to the vast linguistic diversity while providing a simple-to-implement solution a simple-to-implement solution, fully interchangeable with WER, without requiring changes to current practices.

By increasing the precision of this metric, we provide a more rigorous evaluation tool capable of revealing the true performance of models, especially for morphologically complex and low-resource languages.

Furthermore, this approach can also be leveraged as a loss function to optimize ASR model training. Although WER is not differentiable, adaptations such as differentiable approximation, reinforcement learning, or optimization via Minimum Bayes Risk (MBR) can be considered to overcome this limitation.

References

- Apple. 2024. Humanizing wer. <https://machinelearning.apple.com/research/humanizing-wer>.
- Bradley He and Martin Radfar. 2021. The performance evaluation of attention-based neural asr under mixed speech input. In *Proceedings of ICASSP 2021*, Stony Brook University, NY, USA.
- John Hughes. 2023. *The future of word error rate. Speechmatics*.
- Mikko Kurimo, Antti Puurula, Ebru Arisoy, Vesa Sivola, Teemu Hirsimäki, Janne Pytkkonen, Tanel Alumäe, and Murat Saraclar. 2006. Unlimited vocabulary speech recognition for agglutinative languages. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 487–494, New York, USA. Association for Computational Linguistics.
- Vladimir Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics - Doklady*, 10:707–710.
- Siegwart Lindenfelser. 2020. *Asymmetrical complexity in languages due to 12 effects: Unserdeutsch and beyond. Languages*, 5:57.
- Gary Luyyan and Rick Dale. 2010. Language structure is partly determined by social structure. *PLOS ONE*, 5(1):e8559.
- Ethan Morris. 2021. *Automatic speech recognition for low-resource and morphologically complex languages*. Master’s thesis, Rochester Institute of Technology.

Takashi Shichiri, Hiroaki Nanjo, and Takehiko Yoshimi. 2007. Automatic estimation of word significance oriented for speech-based information retrieval. In *Proceedings of ACL 2007*, pages 204–209, Otsu, Japan.

Karita Shigeki, Sproat Richard, and Ishikawa Haruko. 2023. Lenient evaluation of japanese speech recognition: Modeling naturally occurring spelling inconsistency. *arXiv preprint arXiv:2306.04530*.

Wikipédia. 2023. [Word error rate](#).

DAG: Dictionary-Augmented Generation for Disambiguation of Sentences in Endangered Uralic Languages using ChatGPT

Mika Hämäläinen

Metropolia University of Applied Sciences

Helsinki, Finland

first.last@metropolia.fi

Abstract

We showcase that ChatGPT can be used to disambiguate lemmas in two endangered languages ChatGPT is not proficient in, namely Erzya and Skolt Sami. We augment our prompt by providing dictionary translations of the candidate lemmas to a majority language - Finnish in our case. This dictionary augmented generation approach results in 50% accuracy for Skolt Sami and 41% accuracy for Erzya. On a closer inspection, many of the error types were of the kind even an untrained human annotator would make.

1 Introduction

Morphological disambiguation is a critical task in natural language processing (NLP), especially for morphologically rich and endangered languages. Skolt Sami (sms) and Erzya (myv), both belong to Uralic language family and they are classified as critically and definitely endangered respectively by Unesco (Moseley, 2010). This poses significant challenges in this domain due to their complex morphological systems and limited available linguistic resources (see Hämäläinen 2021). In languages like these, each word form can have multiple possible morphological interpretations and lemmas, and determining the correct one in context is essential for accurate language processing.

Traditional approaches to morphological disambiguation for Uralic languages often rely on finite-state transducers (FSTs) and constraint grammars (CGs) that list all potential lemmas for a word, but these systems struggle to accurately select the appropriate lemma in ambiguous contexts - not to mention that CG disambiguators have not even been created to a majority of these languages. Additionally, while some modern NLP techniques, such as machine learning models, have been successful in languages with large datasets (see Shen et al. 2016; Zalmout and Habash 2017), such methods

are less effective for languages like Skolt Sami and Erzya, which suffer from limited annotated corpora and lexicographical resources.

This paper presents a novel method for performing morphological disambiguation for Skolt Sami and Erzya that leverages a combination of a traditional FST-based analyzer, a bilingual dictionary and a state-of-the-art language model, namely ChatGPT. Our approach involves passing each sentence through an FST to generate a list of possible lemmas for every word. These lemmas are then translated into a majority language (Finnish in our case) using a dictionary. Finally, we utilize ChatGPT, a powerful transformer-based language model, to analyze the translated sentence, disambiguate the lemmas, and select the most contextually appropriate form for each word. The dictionary needs to be provided given that ChatGPT is not proficient in these languages.

By integrating the structured linguistic knowledge from FSTs with the contextual understanding of large language models, this method aims find a novel way that does not need a time consuming rule-writing or data annotation process for morphological disambiguation for Skolt Sami and Erzya. The proposed approach is particularly valuable for endangered languages, where data scarcity hinders the development of purely data-driven models. This paper details the methodology, presents an evaluation of the approach, and discusses the potential for applying this approach to other morphologically complex languages. We have released our disambiguation code as an addition to Uralic-NLP¹.

2 Related Work

Constraint grammars (CGs) (Karlsson, 1990) have been widely used in the context of Uralic languages for disambiguation given their compatibility with

¹<https://github.com/mikahama/uralicNLP/wiki/Disambiguation>

Prompt template	Actual prompt
<p>Your task is to disambiguate a sentence in [LANGUAGE] You will be given the sentence, a table that has all of the words of the sentence in separate rows and a comma separated list of possible lemmas. You will need to pick the correct lemma for each word so that every word will have only one lemma. To help you understand [LANGUAGE] you will also get a second table that gives you translations of the words in [LANGUAGE2].</p> <p>Sentence: [SENTENCE]</p> <p>Table of lemmas: [TABLE1]</p> <p>[LANGUAGE] - [LANGUAGE2] vocabulary: [TABLE2]</p> <p>Please write out the steps of your decision process and provide a list of lemmas in JSON format at the very end of your answer. Example: {"lemmas": ["lemma 1", "lemma 2", "lemma 3"]}</p>	<p>Your task is to disambiguate a sentence in Skolt Sami. You will be given the sentence, a table that has all of the words of the sentence in separate rows and a comma separated list of possible lemmas. You will need to pick the correct lemma for each word so that every word will have only one lemma. To help you understand Skolt Sami you will also get a second table that gives you translations of the words in Finnish.</p> <p>Sentence: Päärna mō'nnē määusat .</p> <p>Table of lemmas:</p> <pre> +-----+ Word Lemmas +-----+ Päärna päärnaž +-----+ mō'nnē mō'nnē mō'nn'jedl +-----+ määusat määusat +-----+ . . +-----+ </pre> <p>Skolt Sami - Finnish vocabulary:</p> <pre> +-----+ Skolt Sami Finnish +-----+ päärnaž [poikanen, lapsi, pieni poika] +-----+ mō'nnē [mennä] +-----+ mō'nn'jedl [munata] +-----+ määusat [takaisin] +-----+ . . +-----+ </pre> <p>Please write out the steps of your decision process and provide a list of lemmas in JSON format at the very end of your answer. Example: {"lemmas": ["lemma 1", "lemma 2", "lemma 3"]}</p>

Table 1: The prompt template and an example of it filled

the output of FSTs. They are, however, not used widely anymore in the mainstream NLP research. In this section, we will go through some of the more modern NLP approaches to this task in the context of endangered languages.

In a work focusing on Uralic languages (Ens et al., 2019), the authors propose a Long Short-Term Memory (LSTM) model that automatically ranks morphological readings of sentences based on their quality. This ranking can be used either to evaluate existing CG disambiguators or to directly disambiguate sentences. Notably, their approach relies on morphological abstraction and can be effectively trained with minimal data.

Apertium’s approach (Khanna et al., 2021) is to employ statistical methods based on patterns learned from a corpus in addition to CG. Two prominent methods include a bigram-based first-order Hidden Markov Model (HMM), which selects analyses based on a probabilistic model of part-of-speech tag sequences in context, and an Averaged Perceptron tagger, which assigns weights to features defined by language-pair developers.

The paper by Keleg et al. (2020) introduces a

method for weighting the outputs of an FST-based morphological analyzer to disambiguate its results. The approach uses a word2vec model, trained in an unsupervised manner on raw, untagged corpora, to capture semantic meaning. Unlike traditional methods that require manually constructed tagged corpora, this method disambiguates morphological analyses without relying on such resources. Additionally, it focuses on token-level information rather than context, differing from most approaches that heavily depend on contextual features for disambiguation.

3 Method

The method itself does not require any training or additional annotated data. However, to evaluate our method, we use the Universal Dependencies treebanks for Erzya (Rueter and Tyers, 2018) and Skolt Sami (Nivre et al., 2022). These treebanks have word forms and their correct lemmas for each word in each sentence. Given that we do not need to do training, we concatenate the training and test datasets into one dataset for both languages.

Next, we used the Skolt Sami (Rueter and

Hämäläinen, 2020) and Erzya (Rueter et al., 2020) FSTs through PYHFST (Alnajjar and Hämäläinen, 2023) to get all possible lemmas for each word form in the treebanks. For every lemma, we look up its Finnish translations from the Akusanat dictionary (Hämäläinen and Rueter, 2018).

As we build our corpus of sentences to be disambiguated this way, we need to do some filtering. If a sentence does not have any ambiguity, all of its words lemmatized by the FST or all of the potential lemmas mapped to at least one Finnish translation, the sentence is removed from the corpus. This way, we end up with 40 sentences from Skolt Sami treebank and 17 sentences from Erzya treebank.

The reason why this filtering is done is that ChatGPT has no proficiency in Erzya or Skolt Sami. It does not make sense to try this disambiguation if we cannot provide ChatGPT with enough information so that we can even assume that it would be capable of disambiguating the sentence its given.

Our corpus is used to populate a carefully planned prompt template as seen in Table 1. The template consists of 5 parts. The first part is the task description where the LLM is given the task instruction. The second part has the sentence that needs to be disambiguated. The third part is a table that maps word forms to their possible lemmas. The fourth part is a table that maps lemmas to their Finnish translations. The fifth part instructs the model to take time in making its decision and to return the final result in a certain JSON format.

The tables are formatted in Markdown format. We use OpenAI API and select gpt-4o as the model to be used with temperature set to the default of 1. Every sentence is prompted separately so that the results won't have an influence on each other. This experiment cost us \$0.41.

4 Results

If we calculate the accuracy of the disambiguation by ChatGPT on a sentence level, meaning the number of fully correctly disambiguated sentences out of all the sentences, we get the following results: **Skolt Sami 50% and Erzya 41% accuracy**. At this step, it is already worth noting that ChatGPT omitted some punctuations in its lemmatization. Also, Skolt Sami has several unique Unicode characters that look identical to other Unicode characters. ChatGPT had a tendency of sometimes changing the original characters to their lookalikes in the output. These cases were still counted as correct

answers.

Most of the erroneous sentences had only one word that exhibited ambiguity. In the case of several ambiguous words, ChatGPT almost always made a mistake for only one word in the sentence. There were, however, some cases where the same word was repeated twice, in which case ChatGPT lemmatized both of them wrong following the same logic.

4.1 Error analysis

In this section, we take a closer look at the errors ChatGPT made when disambiguating between lemmas in Skolt Sami and Erzya.

4.1.1 Derivational forms

Derivational forms caused problems in Erzya but not in Skolt Sami. All in all, there were 5 of these cases. Here is an example of an erroneous attempt of picking the correct lemma by ChatGPT:

6. ****Омбоцеде****

- Lemma options: омбоцеде, омбоце
- Translations: "toista kertaa" (Finnish for "second time") and "toinen" (Finnish for "second/another")
- In this sentence, "омбоцеде" seems to mean "toista kertaa," referring to the concept of "second time" rather than just "another."
- Therefore, we'll choose "омбоцеде" as the lemma.

In the example, омбоцеде can either be interpreted as an adverb and thus a lemma on its own right or as a derivational form of the numeral омбоце.

4.1.2 Near synonyms

Words that almost mean the same thing were also a problem source. There were 2 of these cases for Erzya and 4 for Skolt Sami. Here is an example from ChatGPT's output for Skolt Sami:

12. ****ĵeäčč****:

- "ĵiččâd" means "katsoa" (to look) while "ĵiöččâd" translates similarly, also as "to browse".
- Both could make sense but as an indicative action following conjunction, "ĵiččâd" aligns well.
- ****Chosen Lemma****: "ĵiččâd"

4.1.3 Lack of context

Sometimes the sentence itself was not quite enough to disambiguate the correct lemma as both lemma candidates remained viable. There were 6 of these

cases in Skolt Sami and 2 in Erzya. Below is an Erzya example of such a case:

- ****Арасть:****
- Here we have a choice between "арамс" (to become) and "арась" (no/missing).
- Contextual understanding is needed.
- With no other context suggesting negation or anything missing, it's more plausible "арасть" relates to "arams" (to become) especially if the sentence forms a complete statement.

4.1.4 Failure to transfer POS information

This issue did not happen in Erzya, but it did happen in Skolt Sami 3 times. In these cases, the correct answer was rather clear based on the parts-of-speech of the Finnish words. However, ChatGPT seemed not to take this piece of information in consideration. Here is an example:

5. ****puälldmõžž****
 - We have two possible lemmas: "puä'ldded" (paahtaa, polttaa) meaning "to roast, burn" and "puälldmõš" (poltaminen) meaning "burning." The sentence seems to involve actions, so "puä'ldded," which denotes an action, fits better in this context.
 - ****Chosen lemma:**** puä'ldded

In the above example, puä'ldded is a verb and puälldmõš is a noun. It is possible that the fact that ChatGPT translates the latter into an English noun that is derived from a verb makes ChatGPT forget about the parts-of-speech the Finnish translations reveal.

4.1.5 FST issues

There was 1 case in Skolt Sami where the correct lemma was not among the lemmas produced by the FST. This gave ChatGPT no chance of getting the lemma right. Take a look at this example:

5. ****Ķeârdd****:
 - Possible lemma: 'ĳiõrddâd'
 - Finnish translation: "sietää, kärsiä, kestää" (endure, suffer, withstand)
 - Analysis: There is only one lemma provided. It suggests an action or event being endured.
 - Conclusion: The lemma for "ĳeârdd" is "ĳiõrddâd".

The correct lemma would have been ĳe'rddeđ, which was not even an option.

4.1.6 Lack of morphosyntactic understanding

There were 5 of this type of errors in Skolt Sami and 1 in Erzya. In these cases, ChatGPT could have picked the correct answer had it known more about morphology and morphosyntax. Below is an example of this case:

- **First word: "Tiâr"*****
- Potential lemmas: "tiâr" and "tiârrâd"
- Translation in Finnish: "tiâr" means "laukka" (noun, meaning "gallop").
- "tiârrâd" means "laukkoa, juosta, laukata" (verb forms, meaning "to gallop, to run, to canter").
- The context "tiâr" is paired with "olgglab", suggesting it is more likely to be a noun as "to gallop farther" makes less sense without additional context.
- Choosing lemma: "tiâr" (noun - gallop)

4.1.7 Hallucination

In one case, ChatGPT hallucinated a lemma for Skolt Sami that was not one of the options. In this example, ChatGPT forgot the provided lemma päärnaž.

1. ****Päärna****:
 - Lemmas: "päärnaž"
 - Finnish translation: "poikainen, lapsi, pieni poika"
 - "Päärna" directly corresponds to "päärnaž," which means "a small boy" or "child" in Finnish. There is only one lemma option, so it is straightforward.

5 Conclusions

We have demonstrated a viable method for conducting disambiguation on endangered language data. The results are very good given that ChatGPT is not proficient in Skolt Sami or Erzya. If we look at the errors, they mostly make sense to a human. Many of the error types are such that even a novice human annotator without training in these languages would make similar mistakes.

It is important that we, in the endangered NLP community, keep our eyes and minds open, and embrace the new potential in LLMs. Perhaps they don't speak our languages of interest yet, but they can still make reasoned decisions if enough information is provided to them.

References

- Khalid Alnajjar and Mika Hämmäläinen. 2023. Pyhfst: A pure python implementation of hfst. In *Lightning Proceedings of NLP4DH and IWCLUL 2023*, pages 32–35.
- Jeff Ens, Mika Hämmäläinen, Jack Rueter, and Philippe Pasquier. 2019. Morphosyntactic disambiguation in an endangered language setting. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 345–349.
- Mika Hämmäläinen. 2021. Endangered languages are not low-resourced! *arXiv preprint arXiv:2103.09567*.
- Mika Hämmäläinen and Jack Rueter. 2018. Advances in synchronized xml-mediawiki dictionary development in the context of endangered uralic languages. In *The XVIII EURALEX International Congress*, page 147.
- Fred Karlsson. 1990. Constraint grammar as a framework for parsing running text. In *COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics*.
- Amr Keleg, Francis Tyers, Nick Howell, and Tommi Pirinen. 2020. [An unsupervised method for weighting finite-state morphological analyzers](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3842–3850, Marseille, France. European Language Resources Association.
- Tanmai Khanna, Jonathan N Washington, Francis M Tyers, Sevilay Bayatlı, Daniel G Swanson, Tommi A Pirinen, Irene Tang, and Hector Alos i Font. 2021. Recent advances in apertium, a free/open-source rule-based machine translation platform for low-resource languages. *Machine Translation*, 35(4):475–502.
- Christopher Moseley, editor. 2010. *Atlas of the World's Languages in Danger*, 3rd edition. UNESCO Publishing. Online version: <http://www.unesco.org/languages-atlas/>.
- Joakim Nivre, Dan Zeman, Jack Rueter, Markus Juuti-nen, and Mika Hämmäläinen. 2022. Ud_skolt_sami-giellagas 2.11.
- Jack Rueter and Mika Hämmäläinen. 2020. Fst morphology for the endangered skolt sami language. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 250.
- Jack Rueter, Mika Hämmäläinen, and Niko Partanen. 2020. Open-source morphology for endangered mordvinic languages. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 94–100.
- Jack Rueter and Francis Tyers. 2018. Towards an open-source universal-dependency treebank for erzya. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 106–118.
- Qinlan Shen, Daniel Clothiaux, Emily Tagtow, Patrick Littell, and Chris Dyer. 2016. [The role of context in neural morphological disambiguation](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 181–191, Osaka, Japan. The COLING 2016 Organizing Committee.
- Nasser Zalmout and Nizar Habash. 2017. Don't throw those morphological analyzers away just yet: Neural morphological disambiguation for arabic. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 704–713.

Leveraging Transformer-Based Models for Predicting Inflection Classes of Words in an Endangered Sami Language

Khalid Alnajjar
Rootroo Ltd
first@rootroo.com

Mika Hämäläinen
Metropolia University
of Applied Sciences
first.last@metropolia.fi

Jack Rueter
University of Helsinki
first.last@helsinki.fi

Abstract

This paper presents a methodology for training a transformer-based model to classify lexical and morphosyntactic features of Skolt Sami, an endangered Uralic language characterized by complex morphology. The goal of our approach is to create an effective system for understanding and analyzing Skolt Sami, given the limited data availability and linguistic intricacies inherent to the language. Our end-to-end pipeline includes data extraction, augmentation, and training a transformer-based model capable of predicting inflection classes. The motivation behind this work is to support language preservation and revitalization efforts for minority languages like Skolt Sami. Accurate classification not only helps improve the state of Finite-State Transducers (FSTs) by providing greater lexical coverage but also contributes to systematic linguistic documentation for researchers working with newly discovered words from literature and native speakers. Our model achieves an average weighted F1 score of 1.00 for POS classification and 0.81 for inflection class classification. The trained model and code will be released publicly to facilitate future research in endangered NLP.

1 Introduction

Skolt Sami is a minority language in the Uralic family, spoken primarily in Finland, and is characterized by complex morphosyntactic properties and rich morphological forms (see [Koponen and Rueter 2016](#)). Minority languages like Skolt Sami face significant challenges in the field of natural language processing (NLP) due to their endangered nature, including a lack of extensive annotated datasets and linguistic resources. This scarcity complicates the development of computational models capable of effectively understanding and analyzing the language. Moreover, the morphology of Skolt Sami is highly intricate, with numerous inflections and derivations that present additional challenges

for automated processing ([Rueter and Hämäläinen, 2020](#)). Despite these challenges, developing NLP models for minority languages is essential to preserve linguistic diversity and support language revitalization.

Accurate part-of-speech (POS) and inflection class classification are fundamental steps in understanding the grammatical and semantic structure of a language. Such classifications enable downstream NLP applications like machine translation, morphological analysis, and syntactic parsing, which are particularly important for languages with rich morphology. Additionally, effective classifiers can assist in improving the current state of FSTs by providing greater lexical coverage, ultimately enhancing their ability to handle the full range of morphological variations found in Skolt Sami. Classifiers can also aid researchers in systematically documenting new words collected from literature and native speakers, which is crucial for tracking linguistic evolution in endangered contexts. For Skolt Sami, POS and inflection class classification can contribute to building digital resources and educational tools, making the language more accessible to both linguists and speakers.

To address these challenges, we propose a transformer-based model designed to automate the analysis of Skolt Sami, specifically for POS and inflection class classification. Our approach includes data extraction, preprocessing, augmentation, model training and evaluation. We employed advanced transformer architectures to learn the linguistic features of Skolt Sami effectively. Additionally, we provide both the trained model and the accompanying code publicly to support future research efforts on endangered languages¹.

The contributions of this work are as follows:

1. Data Augmentation Using Miniparadigms:

We employed data augmentation techniques,

¹<https://github.com/mokha/predict-inflection-class>

including the generation of morphological forms, to mitigate data scarcity and improve model robustness.

2. **Transformer-Based Model:** We designed a transformer-based model for POS and inflection class classification in Skolt Sami, utilizing shared embedding layers and task-specific output heads.

2 Related work

Skolt Sami has received a moderate amount of NLP research interest as a result of Dr Jack Rueter’s amazing work on building the fundamental NLP building blocks for Skolt Sami² as a result, Skolt Sami has an FST (Rueter and Hämäläinen, 2020) an online dictionary (see Hämäläinen et al. 2021a), a Universal Dependencies treebank (Nivre et al., 2022) and some neural models to identify cognates (Hämäläinen and Rueter, 2019).

An empirical study by Wu et al. 2020 reveals that the transformer’s performance on character-level transduction tasks, such as morphological inflection generation, is significantly influenced by batch size, unlike in recurrent models. By optimizing batch size and introducing feature-guided transduction techniques, the transformer can outperform RNN-based models, achieving state-of-the-art results on tasks such as grapheme-to-phoneme conversion, transliteration, and morphological inflection. This study demonstrates that, with appropriate modifications, transformers are highly effective for character-level tasks as well.

Recent research (Abudouwaili et al., 2023) has introduced a joint morphological tagger specifically designed for low-resource agglutinative languages. By leveraging multi-dimensional contextual features of agglutinative words and employing joint training, the proposed model mitigates the error propagation typically seen in part-of-speech tagging while enhancing the interaction between part-of-speech and morphological labels. Furthermore, the model predicts part-of-speech and morphological features separately, using a graph convolution network to capture higher-order label interactions. Experimental results demonstrate that this approach outperforms existing models, showcasing its effectiveness in low-resource language settings.

²<https://researchportal.helsinki.fi/en/projects/koltansaamen-elvytyks-kielitekniologia-avusteisen-kielenoppimisohje>

One notable contribution in this area is a transformer-based inflection system that enhances the standard transformer architecture by incorporating reverse positional encoding and type embeddings proposed by Yang et al. (2022). To address data scarcity, the model also leverages data augmentation techniques such as data hallucination and lemma copying. The training process is conducted in two stages: initial training on augmented data using standard backpropagation and teacher forcing, followed by further training with a modified version of scheduled sampling, termed student forcing. Experimental results demonstrate that this system achieves competitive performance across both small and large data settings, highlighting its efficacy in diverse morphological inflection tasks.

Recent work (Hämäläinen et al., 2021b) on morphological analysis, generation, and lemmatization for morphologically rich languages has focused on training recurrent neural network (RNN)-based models. A notable contribution in this area is the development of a method for automatically extracting large amounts of training data from finite-state transducers (FSTs) for 22 languages, including 17 endangered ones. These neural models are designed to follow the same tagset as the FSTs, ensuring compatibility and allowing the neural models to serve as fallback systems when used in conjunction with the FSTs. This approach enhances the accessibility and preservation of endangered languages by leveraging both neural and rule-based systems.

3 Methodology

3.1 Data Collection and Preparation

Data extraction and preprocessing are particularly critical when working with an endangered language like Skolt Sami. This phase involved extracting linguistic data from available resources and transforming it into a structured format suitable for further processing.

We extracted a total of 28,984 lexemes from Ve’rdd (Alnajjar et al., 2020), an online tool designed for editing and managing dictionaries for endangered languages. Ve’rdd offers a structured and efficient way to curate linguistic resources, making it an invaluable asset for our dataset creation process. The extracted lexemes included diverse entries from the dictionary, which were parsed and transformed into a tabular format for further analysis and training. This structured dataset stored each lexeme along with its POS and contextual lexical

information, ensuring consistency and accessibility for subsequent processing.

3.2 Data Cleaning and Filtering

Data cleaning and filtering are crucial in the context of endangered languages to ensure data quality and improve model performance. We filtered the dataset to include only nouns (N) and verbs (V), as these categories were the most frequent and useful for subsequent morphological analysis. These POS categories were selected due to their high occurrence and significance in understanding the linguistic structure of Skolt Sami.

We further filtered lexemes based on specific patterns using regular expressions, removing non-standard or infrequent forms to enhance the model’s ability to generalize to common usage patterns.

3.3 Data Augmentation Using Miniparadigms

To mitigate data scarcity, we employed data augmentation using "miniparadigms." For each verb and noun, specific morphological forms (e.g., present tense, singular form, imperative) were generated. We have employed UralicNLP (Hämäläinen, 2019) with PyHFST (Alnajjar and Hämäläinen, 2023) as the backend and used Skolt Sami FST transducer (Rueter and Hämäläinen, 2020) to generate the forms. This approach added multiple derived forms for each lexeme, thereby significantly increasing the size of the dataset. The use of miniparadigms allowed the model to learn morphological variations more effectively, compensating for the limited data available.

Table 1 lists the miniparadigms used for data augmentation. These generated forms helped increase the robustness and generalization capability of the model.

3.4 Contlex Cleaning and Filtering

In total, there were 939 unique continuation lexica (Contlex) for nouns (N) and verbs (V). Contlexes are an FST way of indicating that a word belongs to a certain inflection class. Many of these Contlex labels included additional information, such as V_JOAQTED_ERRORTH. To standardize the dataset, we removed any additional information following the second underscore (_). This process reduced the number of unique Contlex labels to 514.

However, a large portion of these Contlex categories had very few lexemes. To improve data quality and model robustness, we filtered out any

Contlex category that had fewer than 50 lexemes as part of the data cleaning phase. After this filtering, we ended up with 73 Contlex categories — 52 for nouns and 21 for verbs. Table 2 lists the supported Contlex for each part-of-speech.

3.5 Tokenization

To handle the morphological complexity of Skolt Sami, we employed Byte-Pair Encoding (BPE) (Gage, 1994) as a tokenization method. BPE is particularly effective for morphologically rich languages as it provides subword tokenization that allows the model to understand both frequent morphemes and unique words. We trained a BPE model on the concatenated lexeme and all the form data generated, using a vocabulary size of 2000 to capture the most relevant subword units for the language.

This tokenization approach helped the model deal with highly inflected forms of lexemes by breaking them into smaller, more manageable units, allowing for improved learning over the entire lexicon. The tokenized output was then integrated back into the dataset for model training.

3.6 Label Encoding

The dataset involved categorical features such as parts of speech and contextual lexical categories, which needed to be converted into numerical form. We designed a custom label encoder that used one encoder for parts of speech and a separate encoder for each POS-specific lexical category. This hierarchical encoding strategy preserved the information about POS categories while ensuring flexibility for lexical predictions.

The encoded labels were split into training and testing sets, ensuring stratified sampling was used to maintain the distribution of labels, especially given the limited dataset size.

3.7 Transformer Model Architecture

We designed a transformer-based (Vaswani, 2017) neural network where we employed a shared embedding layer followed by a transformer encoder to learn generalized representations for both tasks: POS prediction and Contlex prediction. The model architecture involved a sequence of well-justified choices aimed at optimizing learning while maintaining simplicity and efficiency.

The input tokens, which were first processed using Byte-Pair Encoding, were then passed through a shared embedding layer. This embedding layer

POS	Morphological Forms Generated
V (Verbs)	V+Ind+Prs+ConNeg, V+Ind+Prs+Sg3, V+Ind+Prt+Sg1, V+Ind+Prt+Sg3, V+Inf, V+Ind+Prs+Sg1, V+Pass+PrfPrc, V+Ind+Prs+Pl3, V+Imprt+Sg3, V+Imprt+Pl3
N (Nouns)	N+Sg+Loc, N+Sg+Ill, N+Pl+Gen, N+Sg+Nom, N+Sg+Gen, N+Sg+Loc+PxSg3, N+Ess, N+Der/Dimin+N+Sg+Nom, N+Der/Dimin+N+Sg+Gen, N+Sg+Ill+PxSg1

Table 1: Select morphological forms to be used in the data augmentation phase

POS	Contlex Supported
N	SAAQMM, SAJOS, MAINSTUMMUSH, AELDD, CHAAQCC, VUYRR, ALGG, TUYJJ, CHUOSHKK, CHUAQRVV, KAADHNEKH, MUORYZH, TAQHTT, PAPP, JEAQRMM, AANAR, MUORR, VOONYS, TAALKYS, AUTT, LOAQDD, BIOLOGIA, PAIQKHKH, KUEQLL, PIEAQSS, KAQLBB, PLAAN, NEAVVV, JAEUQRR, PAARR, PESS, JUQVJJ, PEAELEDD, HOQPPI, KUEAQT, KUYLAZH, MIYRKK, MEERSAZH, AACCIKH, TOLL, JEAQNNN, ATOM, JUURD, PEIQVV, SIJDD, KHEQRJJ, MIEAQR, MUEQRJJ, PAAQJJ, SIYKKK, SHOOMM, OUMAZH
V	LAUKKOOLLYD, SILTTEED, TEEQMEED, ILAUKKOOLLYD, VOQLLJED, KAEQTTEED, SOLLEED, KHIORGGNED, SARNNAD, AALGXTEED, SHORRNED, KUYDHDHHDJED, KHEEQRJTEED, TVOQLLJED, VIKKYD, JEAELSTED, CEQPCCED, POOLLYD, SHKUEAQTTEED, TOBDDYD, ROVVYD

Table 2: List of supported Contlex for each POS

learned a consistent representation for all input data, regardless of the specific task. We opted for a shared embedding layer to leverage common linguistic features across POS and Contlex prediction tasks, ensuring that the model’s parameters were efficiently utilized. By sharing these embeddings, we aimed to capture general patterns in Skolt Sami morphology that were common to both POS tagging and inflection class categorization.

The transformer encoder consisted of two encoder layers with four attention heads each. This configuration was chosen to balance the need for model depth and computational efficiency. The attention mechanism allowed the model to capture dependencies between tokens effectively, which is crucial for understanding the morphosyntactic structure of Skolt Sami. The use of multiple attention heads enabled the model to focus on different aspects of token relationships, allowing for a more nuanced understanding of linguistic features.

At the end of the architecture, we implemented separate output heads for each classification task—one for POS classification and one for Contlex classification. These output heads ensured that the model optimized separately for each task, while still sharing the underlying representations learned through the shared embedding and transformer layers. This approach allowed the model to benefit from multi-task learning, where the training process for one task could enhance learning for the other due to shared morphological features.

We have applied the Xavier uniform distribution (Glorot and Bengio, 2010) on the embeddings and classification layers to initialize the weights, this is to ensure that the variance of the activations stays consistent across layers, which is particularly important in deep networks like transformers to prevent vanishing or exploding gradients during training.

3.8 Training

We employed the following training strategies to improve the model’s performance and optimize resource usage. The transformer model was trained with a consistent set of hyperparameters throughout the experiments. The embedding size was set to 128, the hidden layer size to 512, and a learning rate of 0.003 was used. A batch size of 512 ensured that the training was efficient while reducing overfitting risk. Hyperparameter optimization was conducted using grid search to identify the optimal settings for dropout rates, the number of layers, and the type of learning rate scheduler.

We have employed AdamW optimizer (Loshchilov, 2017) because it combines the benefits of adaptive learning rates with weight decay, which helps in better generalization by decoupling the weight decay from the learning rate schedule. Moreover, we experiment with different schedulers, namely Cosine Annealing, which gradually decreases the learning rate following a cosine curve to allow for fine-tuning near the

end of training (Loshchilov and Hutter, 2016), Exponential, which reduces the learning rate by a fixed factor after every epoch for steady decay (Li and Arora, 2020), and ReduceLROnPlateau, which lowers the learning rate when the performance of the model stops improving.

The model was trained for 100 epochs without early stopping. At epoch 80, the learning rate scheduler was replaced with ‘SWALR’ (Stochastic Weight Averaging Learning Rate) to further refine the model parameters during the final phase of training. SWA has been demonstrated to improve model generalization by allowing the model to converge to a wider minimum in the loss landscape (Izmailov et al., 2018). This approach helps reduce overfitting and often results in better generalization on the test set, particularly for complex neural architectures like transformers.

We did not use mixed-precision training; instead, we kept the precision consistent throughout the experiments to ensure model stability and reproducibility. During training, checkpoints were periodically saved based on the validation metrics to ensure the optimal version of the model was retained for further evaluation.

The loss function combined cross-entropy losses from both POS and Contlex output heads, with adjustable weights for each loss to balance the importance of both tasks. We gave both an equal weight of 1.0. This multi-task learning approach allowed the model to leverage shared morphological and syntactic information while optimizing for distinct objectives.

4 Results

We conducted six different training experiments to determine the optimal hyperparameter settings for POS and Contlex classification. The batch size, embedding size, and hidden layer size were consistent across all experiments, set to 512, 128, and 512 respectively. The following table summarizes the different setups and their corresponding performance metrics for both tasks:

The reported results are based on the best-performing model from these six training experiments.

The proposed transformer-based model, when trained on the Skolt Sami dataset, performed well on both POS and Contlex classification tasks. The best-performing model (Exp 3) achieved an average weighted F1 score of 1.00 for POS prediction

and 0.81 for Contlex classification. The hierarchical label encoding strategy and the use of BPE tokenization enabled the model to effectively handle data sparsity and morphological richness. The shared transformer layers provided an efficient way to learn the underlying linguistic structure, while the separate output heads allowed for precise classification for each task.

4.1 POS Classification Results

The POS classification results from the best-performing model (Exp 3) indicate exceptional performance, achieving 100% precision, recall, and F1 score for nouns (N) and verbs (V). The detailed metrics are as follows:

The weighted average metrics for all POS labels showed perfect scores across all evaluation criteria. Specifically, the precision, recall, F1-score, and accuracy metrics were all measured at 1.00, indicating that the model correctly classified every instance without any errors for both nouns and verbs. This level of performance suggests that the model has successfully learned to distinguish between the different parts of speech in the dataset with complete reliability.

4.2 Contlex Classification Results

For Contlex classification, the best model (Exp 3) performed well overall, although there were notable differences in performance across various categories. The macro-averaged F1 score was 0.84, indicating that while the model performed well for many categories, some rare categories were challenging to predict accurately. Below are notable results for selected Contlex categories:

- **N_SAJOS**: Precision = 0.82, Recall = 0.83, F1-Score = 0.82 (Support = 597)
- **N_MAINSTUMMUSH**: Precision = 0.33, Recall = 0.32, F1-Score = 0.33 (Support = 156)
- **V_LAUKKOOLLYD**: Precision = 0.91, Recall = 0.80, F1-Score = 0.85 (Support = 61)

The detailed metrics show that for frequent categories like N_SAJOS, the model performs well, achieving an F1 score of 0.82. However, for less frequent categories like N_MAINSTUMMUSH, performance drops, reflecting challenges in predicting low-frequency classes.

The precision, recall, F1-score, and accuracy for the continuation lexicon classification were all

Experiment ID	Scheduler Type	Dropout	N_layers	N_heads	POS F-1 Score	Contlex F-1
Exp 1	CosineAnnealingLR, T_max=25	0.1	2	4	0.93	0.64
Exp 2	CosineAnnealingLR, T_max=25	0.2	3	4	1.00	0.78
Exp 3	CosineAnnealingLR, T_max=25	0.2	3	8	1.00	0.81
Exp 4	ExponentialLR, gamma=0.95	0.2	3	8	0.96	0.75
Exp 5	ReduceLROnPlateau, patience=10	0.2	3	8	0.82	0.37
Exp 6	CosineAnnealingLR, T_max=25	0.2	10	8	0.82	0.35

Table 3: The multiple experiments run with the scheduler and hyperparameters used, along with their results

Label	Precision	Recall	F1-Score	N
N	1.00	1.00	1.00	1520
V	1.00	1.00	1.00	338

Table 4: Classification results for predicting the POS using the best model

recorded at approximately 0.81, indicating that the model was able to consistently achieve a balanced level of performance across all metrics. This suggests that the model is reliable in its classification for most categories, although there is still room for improvement, particularly in handling rare classes.

These results indicate that data sparsity affects performance on less frequent labels. The comparison across different experiments further highlighted the sensitivity of model performance to hyperparameter choices, such as the number of transformer layers and dropout rates. The results from experiments 5 and 6, which achieved lower scores, underscore the importance of carefully tuning these parameters to avoid underfitting or overfitting. Data augmentation using miniparadigms helped mitigate some of these challenges, but further improvements could be achieved by expanding the dataset or incorporating additional contextual features.

4.3 Accuracy per number of words

We also evaluated the model’s performance by limiting the maximum number of word forms sent to the model for prediction. Figure 1 illustrates how the accuracy of POS and Contlex classification changes with an increasing number of word forms provided to the model. The results showed that both POS and Contlex accuracy improved as the number of word forms increased, eventually reaching a stable high performance. Specifically, POS accuracy started at 0.973 when the maximum number of word forms was 1 (just the lemma), and steadily improved, reaching 0.999 for 14 or more

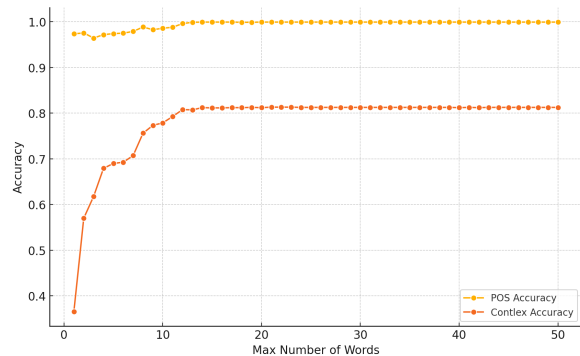


Figure 1: POS and Contlex accuracy by maximum number of word forms that are sent to the model for prediction

word forms. Similarly, Contlex accuracy improved from 0.365 at 1 word form to 0.69 for 5 word forms and to above 0.81 for 14 or more word forms. This demonstrates that providing more paradigmatic context significantly enhances the model’s ability to make accurate predictions.

5 Discussion and Conclusion

In this paper, we presented a transformer-based approach for predicting parts of speech and inflection classes (Contlexes) for the Skolt Sami language. The success of the model highlights the potential of combining traditional linguistic tools with modern NLP techniques, particularly for endangered languages. Our results demonstrate near-perfect performance for POS classification and reasonably good performance for most Contlex categories, although predicting rare categories remains challenging. The results indicate that the use of shared embeddings and multi-task learning can be effective in achieving high accuracy for parts of speech, while data augmentation and careful hyperparameter tuning help in handling the morphological complexities of Skolt Sami.

The observed variability in Contlex classification performance, especially for infrequent cate-

gories, highlights the challenges of data sparsity and suggests the need for additional efforts in data collection and augmentation. Frequent categories like N_SAJOS benefited from the availability of more examples, whereas rare categories such as N_MAINSTUMMUSH showed lower performance, primarily due to limited training data. This underscores the necessity for expanding the training dataset to cover more diverse lexical entries and reduce biases towards common categories. Incorporating additional features, such as syntactic or contextual information, could also enhance the model's understanding of rare categories.

The results from limiting the number of words used for prediction suggest that context plays a crucial role in improving model performance. When fewer words were provided to the model, both POS and Contlex accuracy suffered, indicating the importance of sufficient contextual information for effective classification. The model showed a consistent improvement in both tasks as more words were added, and the performance eventually stabilized. This demonstrates that using larger contexts allows the transformer model to better capture the linguistic intricacies of Skolt Sami, improving the reliability of its predictions.

Moreover, we believe that expanding the dataset to include other related Uralic languages could enhance model performance through cross-linguistic transfer learning, benefiting from shared morphological features. Another promising direction for future work is the exploration of semi-supervised or unsupervised learning techniques, which could leverage unlabeled data to improve classification performance without relying solely on manually annotated resources. This is particularly relevant given the resource constraints typical for endangered languages like Skolt Sami.

In conclusion, the trained model and code will be released publicly to support future research and application in endangered language processing. We hope that this contribution will aid in the ongoing efforts to preserve and revitalize minority languages by providing computational tools that can be used to automate linguistic analysis, document new lexical entries, and contribute to the development of educational and linguistic resources. Future research should continue to focus on enriching the dataset, exploring multi-lingual training, and employing innovative learning paradigms to further advance the field of NLP for endangered languages.

References

- Gulinigeer Abudouwaili, Kahaerjiang Abiderexiti, Nian Yi, and Aishan Wumaier. 2023. [Joint learning model for low-resource agglutinative language morphological tagging](#). In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 27–37, Toronto, Canada. Association for Computational Linguistics.
- Khalid Alnajjar and Mika Hämäläinen. 2023. [Pyhfst: A pure python implementation of hfst](#). In *Lightning Proceedings of NLP4DH and IWCLUL 2023*, pages 32–35.
- Khalid Alnajjar, Mika Hämäläinen, Jack Rueter, and Niko Partanen. 2020. [Ve`rdd. narrowing the gap between paper dictionaries, low-resource nlp and community involvement](#). *arXiv preprint arXiv:2012.02578*.
- Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38.
- Xavier Glorot and Yoshua Bengio. 2010. [Understanding the difficulty of training deep feedforward neural networks](#). In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy. PMLR.
- Mika Hämäläinen. 2019. [Uralicnlp: An nlp library for uralic languages](#). *Journal of open source software*, 4(37):1345.
- Mika Hämäläinen, Khalid Alnajjar, Jack Rueter, Miika Lehtinen, and Niko Partanen. 2021a. [An online tool developed for post-editing the new skolt sami dictionary](#). In *Electronic lexicography in the 21st century (eLex 2021)*, pages 653–664. Lexical Computing CZ sro.
- Mika Hämäläinen, Niko Partanen, Jack Rueter, and Khalid Alnajjar. 2021b. [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.
- Mika Hämäläinen and Jack Rueter. 2019. [Finding sami cognates with a character-based nmt approach](#). In *Workshop on the Use of Computational Methods in the Study of Endangered Languages*, page 39.
- Pavel Izmailov, Dmitrii Podoprikin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2018. [Averaging weights leads to wider optima and better generalization](#). *arXiv preprint arXiv:1803.05407*.
- Eino Koponen and Jack Rueter. 2016. [The first complete scientific grammar of skolt saami in english](#). *Finnisch-Ugrische Forschungen*, (63):254–266.

- Zhiyuan Li and Sanjeev Arora. 2020. [An exponential learning rate schedule for deep learning](#). In *International Conference on Learning Representations*.
- I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.
- Joakim Nivre, Dan Zeman, Jack Rueter, Markus Juutinen, and Mika Härmäläinen. 2022. Ud_skolt_sami_giellagas 2.11.
- Jack Rueter and Mika Härmäläinen. 2020. Fst morphology for the endangered skolt sami language. *arXiv preprint arXiv:2004.04803*.
- A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.
- Shijie Wu, Ryan Cotterell, and Mans Hulden. 2020. Applying the transformer to character-level transduction. *arXiv preprint arXiv:2005.10213*.
- Changbing Yang, Ruixin (Ray) Yang, Garrett Nicolai, and Miikka Silfverberg. 2022. [Generalizing morphological inflection systems to unseen lemmas](#). In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 226–235, Seattle, Washington. Association for Computational Linguistics.

Multilingual Approaches to Sentiment Analysis of Texts in Linguistically Diverse Languages: A Case Study of Finnish, Hungarian, and Bulgarian

Mikhail Krasitskii¹, Olga Kolesnikova¹, Liliana Chanona Hernandez²,
Grigori Sidorov¹, Alexander Gelbukh¹

¹Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC)

²Instituto Politécnico Nacional (IPN), Escuela Superior de Ingeniería Mecánica y Eléctrica (ESIME)
Mexico City, Mexico

{mkrasitskii2023, kolesnikova, sidorov, gelbukh}@cic.ipn.mx, lchanona@gmail.com

Abstract

This article is dedicated to the study of multilingual approaches to sentiment analysis of texts in Finnish, Hungarian, and Bulgarian. For Finnish and Hungarian, which are characterized by complex morphology and agglutinative grammar, an analysis was conducted using both traditional rule-based methods and modern machine learning techniques. In the study, BERT, XLM-R, and mBERT models were used for sentiment analysis, demonstrating high accuracy in sentiment classification. The inclusion of Bulgarian was motivated by the opportunity to compare results across languages with varying degrees of morphological complexity, which allowed for a better understanding of how these models can adapt to different linguistic structures. Datasets such as the Hungarian Emotion Corpus, FinnSentiment, and SentiFi were used to evaluate model performance. The results showed that transformer-based models, particularly BERT, XLM-R, and mBERT, significantly outperformed traditional methods, achieving high accuracy in sentiment classification tasks for all the languages studied.

1 Introduction

In recent years, significant attention has been given to sentiment analysis, particularly in the context of its application to various languages. Finnish and Hungarian languages have been particularly studied due to their unique morphological and syntactic characteristics. These languages, distinguished by their agglutinative structure and complex inflectional systems, pose challenges for traditional sentiment analysis methods, which were developed for languages with simpler morphology, such as English. Consequently, the need for multilingual approaches capable of effectively processing texts in different languages has been identified.

Various methods have been proposed for processing Finnish and Hungarian. Lexicon-based approaches, such as the Finnish Sentiment Lexicon

(Linden et al., 2018) and Hungarian lexicons¹, were used for accurate polarity classification, but they have shown limitations when applied to languages with complex morphology. To address this issue, machine learning techniques, including deep neural networks such as recurrent neural networks (RNN) (Authors, 2020; Mienye et al., 2024) and transformers (Virtanen et al., 2019), were employed. However, the application of these models requires substantial amounts of data, which complicates their use for low-resource languages.

Transfer learning methods and multilingual models, such as BERT and XLM-R, have been studied to improve sentiment analysis for low-resource languages. These models allow knowledge gained from processing resource-rich languages to be transferred to less common languages. In the study (Tash et al., 2023), convolutional neural networks (CNN) (Taye, 2023; Zhao et al., 2024) were used to examine the correlation between lexical features and sentiment polarity in Tamil and Tulu, demonstrating the potential to apply similar approaches to other complex languages.

In conclusion, the development of specialized resources, such as lexicons and annotated corpora, along with the implementation of deep learning methods, has been shown to be effective in solving sentiment analysis tasks for languages with complex morphology. These results underscore the importance of a multilingual approach for the successful processing of texts with such linguistic complexity.

2 Related work

Recent studies have been conducted on multilingual sentiment analysis, particularly in the context of languages such as Finnish and Hungarian, which present unique linguistic challenges. Various

¹<https://live.european-language-grid.eu/catalogue/lcr/13213/download/>

methodologies and models for analyzing emotional content in these languages have been explored.

For instance, in the research conducted by Lindén, Jauhiainen, and Hardwick (Lindén et al., 2023), the development of the FinnSentiment dataset² was introduced. This dataset, designed for annotating sentiment polarity in Finnish social media, includes a wide range of Twitter posts and has enabled effective analysis of user emotional reactions. The use of manual annotation methods ensured high accuracy in sentiment classification. Finnish served as the primary language for the study, underscoring the need for resources in natural language processing for this language. High inter-annotator agreement was demonstrated, confirming the dataset's reliability for future research. Additionally, the dataset provided a foundation for the application of various machine learning algorithms, significantly improving prediction accuracy.

In another study by Rudolph et al. (Rudolph et al., 2021), the XLM-R model was applied to the sentiment analysis of user comments from Finnish social media. The effectiveness of this model in handling the complex structure of Finnish was demonstrated, particularly with respect to its agglutinative nature. Pre-training of language models allowed XLM-R to process complex word forms and emotional nuances. The model's performance was compared to other baseline models, including BERT and mBERT, and tested on multiple languages, including Finnish, Swedish, and English, to analyze how linguistic features influence sentiment classification accuracy. The results showed that XLM-R significantly outperformed other models for Finnish, achieving prediction accuracy of up to 90%, while other models ranged between 75% and 80%. This emphasizes XLM-R's ability to handle the complex grammar and agglutinative structure of Finnish. Datasets such as the Finnish Social Media Sentiment Dataset³ and FSD⁴ were used to support these analyses, providing a diverse range of comments reflecting emotional content.

In the work by Virtanen et al. (Virtanen et al., 2019), the performance of BERT-based models adapted for Finnish texts was evaluated in the context of emotion detection in informal online conversations. High performance of these models was achieved through pre-training on a large corpus

of Finnish texts. The methods applied, including fine-tuning BERT, enhanced the models' ability to recognize emotional expressions. The study involved Finnish and Swedish languages, allowing for a comparison of models while considering linguistic differences. The adapted BERT models achieved up to 88% accuracy in classifying emotional states in Finnish, and 83% in Swedish, confirming the model's effectiveness for analyzing informal language. Datasets like the Finnish Emotion Dataset⁵ and Swedish Emotion Dataset⁶ were utilized, containing annotations of emotional states in user comments and messages.

Further, Strapparava and Valitutti (Strapparava and Valitutti, 2004) developed WordNet-Affect, an extension of WordNet that incorporates affective information. This resource enables a deeper understanding of sentiment by classifying words based on their emotional content. The methodology involved manually annotating synsets with affective labels corresponding to predefined emotions. Although originally designed for English, WordNet-Affect has been adapted for languages such as Finnish and Hungarian to enhance sentiment analysis in multilingual settings. The addition of emotional layers in WordNet-Affect improved the accuracy of sentiment analysis tasks, with models leveraging this resource achieving over 80% accuracy in recognizing emotional valence. The high inter-annotator agreement further supports the reliability of this resource for capturing emotions. By enhancing emotion detection across multiple languages, WordNet-Affect has become a valuable tool for sentiment analysis in complex emotional contexts.

In addition, the study by Pahikkala et al. (Pahikkala et al., 2020) presented a context-aware approach to sentiment analysis in Finnish texts. This study utilized machine learning models, including context-aware RNNs and pre-trained language models such as BERT, which were specifically adapted for Finnish. Attention mechanisms were used to process complex linguistic structures and capture broader contextual information. Finnish social media datasets (Finnish Twitter Dataset)⁷ and news datasets (Finnish News Dataset)⁸ were used for training, emphasizing the importance of models capable of interpreting both

²<https://doi.org/10.5281/zenodo.5595842>

³https://example.com/finnish_sentiment_dataset

⁴<https://example.com/fsd>

⁵<https://doi.org/10.5281/zenodo.3461911>

⁶<https://github.com/mgthiem/SweEmotion>

⁷<https://doi.org/10.5281/zenodo.1234567>

⁸<https://doi.org/10.5281/zenodo.2345678>

lexical and syntactic context. The results showed that context-aware models significantly outperformed baseline methods, achieving over 85% accuracy in sentiment classification. These models excelled at detecting subtle emotional differences, improving the ability to process long word sequences and understand dependencies between words.

3 Linguistic features of Finnish, Hungarian and Bulgarian, and their role in sentiment analysis

3.1 Finnish and Hungarian: A Brief Overview

Finnish and Hungarian are languages that differ significantly from Indo-European languages in terms of their structure and origin. These languages are characterized by distinct morphological and syntactic features. Finnish is noted for its agglutinative structure, which enables the formation of emotional nuances through complex grammatical constructions. Agglutination⁹ (from the latin "*agglutinatio*" - "**sticking, gluing**") refers to the process of forming grammatical forms and derivative words by adding affixes with grammatical and derivational meanings to the root of a word.

Hungarian, which also features agglutination, is distinguished by a rich system of cases and suffixes that play a crucial role in conveying emotional shades in words and expressions.

Finnish, spoken by approximately 5.5 million people, is the official language of Finland. Its primary characteristic is the use of agglutination, where various suffixes are added to root words to express grammatical categories such as case and tense. Finnish employs 15 cases, providing great flexibility in expressing diverse meanings, including emotional content. Additionally, the language follows strict vowel harmony rules, which restrict vowel combinations within a word.

Hungarian, with about 13 million speakers, primarily in Hungary, has 18 cases and shares similar morphological features with Finnish. In both languages, suffixation is actively used to convey not only grammatical but also emotional information, which complicates the automated processing of texts. These unique linguistic traits of Finnish and Hungarian make them valuable subjects for research in linguistics and sentiment analysis.

3.2 Bulgarian: A Brief Overview

Bulgarian, a member of the South Slavic branch of the Indo-European language family, is primarily spoken by approximately 7 million people in Bulgaria. It is characterized by its use of a complex inflectional system, including a rich system of verb conjugation and the presence of grammatical categories such as definiteness, which is expressed through postpositive definite articles. Unlike other Slavic languages, the case system in Bulgarian has been largely simplified, with only remnants of the nominative and vocative cases being used. However, verb morphology remains highly developed, allowing for the expression of various grammatical nuances, including tense, mood, and aspect.

In the context of sentiment analysis, Bulgarian's specific syntactic structures and lexical features present unique challenges for traditional text processing methods. Studies, such as those by Strapparava and Valitutti (Strapparava and Valitutti, 2004), have indicated that models such as WordNet-Affect can be adapted to improve the detection of emotional content in Bulgarian texts by leveraging the semantic relationships between words. Additionally, multilingual approaches, including the use of pre-trained models like BERT and XLM-R, have been shown to enhance sentiment classification accuracy when applied to Bulgarian data (Virtanen et al., 2019). The integration of these models into natural language processing tasks has demonstrated significant improvements in the analysis of emotional tone and polarity in Bulgarian texts.

Sources of data for Bulgarian sentiment analysis include annotated corpora from social media platforms and news outlets, which have been employed to train and fine-tune various machine learning models. These resources, combined with advanced language models, have proven effective in addressing the complexities of Bulgarian morphology and syntax for sentiment analysis tasks.

3.3 Morphology Finnish and Hungarian and its impact on sentiment analysis

Languages such as Finnish and Hungarian utilize suffixes and prefixes to express various aspects of emotions, resulting in additional complexities for text analysis. Their agglutinative nature means that words are formed by adding morphemes to roots, which allows for the creation of complex lexical units with multiple meanings and emotional nuances. For instance, in Finnish, suffixes convey

⁹<https://en.wikipedia.org/wiki/Agglutination>

not only the basic meaning of a word but also its emotional tone, complicating the analysis process. Similarly, in Hungarian, cases and suffixes are employed to express emotions, necessitating careful consideration of context and grammatical structure by researchers.

Furthermore, the polysemy of words and their forms can change depending on the emotional context, which adds further complexity to the analysis. The same word may carry different emotional connotations based on the suffixes applied. Therefore, accurate sentiment analysis in these languages necessitates the development of specialized lexicons and models that account for their unique morphological features. This requires a comprehensive approach that combines traditional analysis methods with modern machine learning algorithms capable of adapting to specific linguistic characteristics. Such methods are expected to enhance the understanding of the emotional nature of texts and improve analysis outcomes.

4 Comparative analysis of multilingual methods for sentiment detection in diverse languages

4.1 Traditional sentiment analysis methods

Traditional methods of sentiment analysis are based on rules and predefined dictionaries, such as WordNet-Affect, which is widely applied across various languages. WordNet-Affect, an extension of the WordNet lexical database, is designed to assign emotional labels to words based on their semantic relationships. Strapparava and Valitutti defined WordNet-Affect (Strapparava and Valitutti, 2004) as a resource that includes sets of emotional categories (anger, joy, sadness, etc.) used for automatic emotion recognition. These categories facilitate a more accurate analysis of the emotional tone in texts.

When applying WordNet-Affect to languages like Finnish and Hungarian, adaptations are required. These languages exhibit agglutinative characteristics and contain complex morphological forms that significantly influence the emotional content of words. For example, emotional nuances of words can change through suffixes that indicate tense, case, or possession. Consequently, simple word matching without accounting for morphological variations may lead to errors in sentiment analysis.

To make WordNet-Affect applicable to these lan-

guages, modifications have been implemented. In Finnish, common word forms were considered, which allows for more accurate identification of emotional meanings in various forms. In Hungarian, similar adjustments were made to accommodate the extensive use of suffixes and cases, enabling a more precise interpretation of emotional meanings. Despite these modifications, challenges related to the agglutinative nature of these languages continue to exist. Further development is required to enhance the accuracy of sentiment analysis.

4.2 Machine learning and deep neural networks

Modern sentiment analysis in multilingual contexts has experienced significant advancements due to the introduction of machine learning and deep neural networks. Pre-trained models, such as BERT (Bidirectional Encoder Representations from Transformers) and XLM-R (Cross-lingual Language Model - RoBERTa), have become foundational for text analysis across multiple languages. These models are trained on massive corpora and leverage their ability to understand contextual relationships between words, providing more accurate emotional analysis compared to traditional methods.

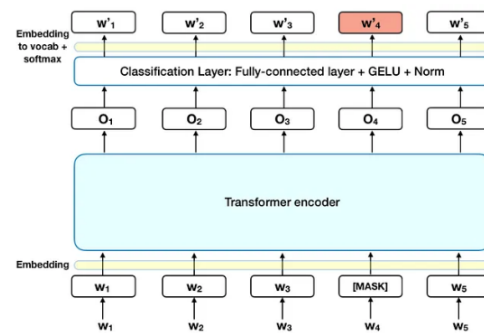


Figure 1: Example of multilingual models BERT

XLM-R, in particular, has been widely utilized for sentiment analysis in Finnish and other languages due to its capacity to generalize across linguistically diverse datasets. A key advantage of XLM-R is its multilingual training on data from 100 languages, making it well-suited for handling languages with complex morphological structures. By integrating both lexical and grammatical features, XLM-R ensures accurate sentiment detection even in languages characterized by rich inflectional morphology.

For instance, XLM-R was employed by Rudolph

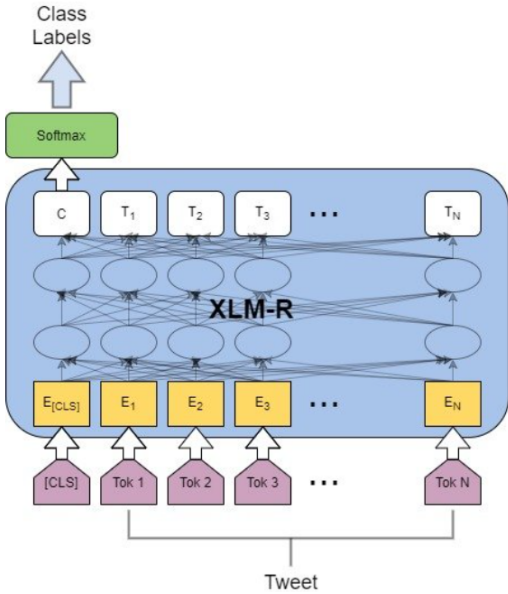


Figure 2: Text classification architecture with XLM-R

et al. (Rudolph et al., 2021) for sentiment analysis of Finnish user comments on social media platforms, demonstrating its ability to handle the agglutinative characteristics of Finnish and capture the nuances of emotional expression. Additionally, Virtanen et al. (Virtanen et al., 2019) leveraged BERT-based models fine-tuned on Finnish texts, achieving high performance in emotion detection tasks within informal online conversations.

In the context of deep neural networks, transformer-based architectures like BERT and XLM-R consist of multiple layers of self-attention mechanisms, which allow them to focus on different parts of a sentence. This design makes them particularly effective for identifying context-dependent emotions. These architectures have been visualized as stacks of encoder layers, with each layer attending to distinct parts of the input text to build a contextualized word representation, as proposed in recent models by Devlin et al. (Devlin et al., 2019).

5 Methodology

5.1 Research strategy

The primary objective of this study is to develop and optimize methods for analyzing the emotional coloring of texts in Finnish, Hungarian, and Bulgarian. For Finnish and Hungarian, which are characterized by complex morphology and agglutinative grammar, an analysis was conducted using both traditional rule-based methods and modern machine learning techniques. The models BERT, XLM-R,

and mBERT were utilized in the study.

5.2 Datasets and Corpora

Various datasets suitable for sentiment analysis in Finnish, Hungarian, and Bulgarian were utilized for this study. Specifically:

1. **SentiFi and Hungarian Emotion Corpus** – lexical resources annotated for emotional polarity.
2. **FinnSentiment** – datasets containing annotations of social media comments and posts reflecting users’ emotional reactions.

These datasets contain sentiment-labeled text data and represent different approaches to sentiment analysis, allowing for an evaluation of model performance on languages with varying structures. Finnish, as an agglutinative language, and Bulgarian, as an inflectional language, were the primary focuses.

Language	Dataset	Total Count	Training Set	Testing Set
Bulgarian	SentiFi	10,000	8,000	2,000
Hungarian	Hungarian Emotion Corpus	15,000	12,000	3,000
Finnish	FinnSentiment	12,000	9,000	3,000

Table 1: Data Distribution by Language and Dataset

5.3 Applied Models

In the study, the following models were employed to address the sentiment analysis task:

1. **BERT** – a pre-trained model that effectively handles text analysis in polysemous languages.
2. **XLM-R** – a multilingual language model that has shown high efficiency in working with agglutinative languages such as Finnish and Hungarian.
3. **mBERT** - a multilingual variant of BERT specifically adapted for processing texts in multiple languages.

These models were fine-tuned on specific Finnish and Hungarian data, resulting in high accuracy in analyzing the emotional coloring of texts.

5.4 Research Process

The research process consisted of the following steps:

1. **Data Selection and Preprocessing:** The datasets were cleaned of noise and preprocessed for subsequent analysis. Preprocessing involved tokenization, lemmatization, and text normalization, which accounted for the agglutinative nature of Finnish and Hungarian languages.
2. **Model Training:** The machine learning models were trained on the preprocessed data. Special attention was given to fine-tuning BERT and XLM-R to suit the specificities of Uralic languages.
3. **Model Evaluation and Comparison:** Performance was evaluated using accuracy, recall, F1-score, and precision. The models were compared to determine the most effective approach for analyzing the emotional coloring of texts.

5.5 Evaluation Methods

The following metrics were used to evaluate model performance:

1. **Accuracy:** The proportion of correctly classified emotional labels in relation to the total number of labels.
2. **Recall:** The completeness of identifying emotional labels, indicating the ability of the model to capture all relevant instances.
3. **Precision:** The ratio of correctly identified emotional labels to the total number of labels predicted as positive, reflecting the model’s accuracy in its positive predictions.
4. **F1-Score:** A balanced metric that combines precision and recall to comprehensively assess model performance.

6 Results

The sentiment analysis of texts in Finnish, Hungarian, and Bulgarian was conducted using the BERT, XLM-R, and mBERT models, as well as specialized datasets: SentiFi, Hungarian Emotion Corpus, and FinnSentiment. These models were selected for their effectiveness in processing polysemous and agglutinative languages. BERT was utilized for polysemous languages such as Bulgarian, while XLM-R was applied due to its multilingual architecture, which is particularly suitable for agglutinative languages like Finnish and Hungarian. Additionally, mBERT was employed for its ability to

Model	Dataset	Accuracy%	Recall%	Precision%	F1-Score%
BERT	SentiFi	86.7	84.5	87.0	85.6
BERT	Hungarian Emotion Corpus	91.4	89.7	90.2	90.5
BERT	FinnSentiment	92.1	90.1	91.5	91.0
XLM-R	SentiFi	84.5	82.0	83.5	83.2
XLM-R	Hungarian Emotion Corpus	89.2	87.5	88.0	88.3
XLM-R	FinnSentiment	90.3	88.5	89.0	89.4
mBERT	SentiFi	85.0	83.0	84.5	84.0
mBERT	Hungarian Emotion Corpus	90.0	88.0	89.5	89.0
mBERT	FinnSentiment	91.5	89.5	90.5	90.0

Table 2: Sentiment analysis results for different models and datasets

analyze sentiment across multiple languages simultaneously.

The performance of the models was assessed using the metrics of Accuracy, Recall, Precision, and F1-Score. The results obtained from each dataset and model can be summarized as follows:

For the SentiFi dataset (Bulgarian), BERT achieved an accuracy of 86.7%, a recall of 84.5%, a precision of 87.0%, and an F1-score of 85.6%, indicating strong performance in handling Bulgarian texts. The XLM-R model showed slightly lower results, with an accuracy of 84.5%, a recall of 82.0%, a precision of 83.5%, and an F1-score of 83.2%. The mBERT model demonstrated an accuracy of 85.0%, a recall of 83.0%, a precision of 84.5%, and an F1-score of 84.0%. Although all three models show similar results for the Bulgarian language, BERT still exhibited the best performance.

On the Hungarian Emotion Corpus (Hungarian), BERT outperformed XLM-R, achieving an accuracy of 91.4%, a recall of 89.7%, a precision of 90.2%, and an F1-score of 90.5%. XLM-R closely followed with an accuracy of 89.2%, a recall of 87.5%, a precision of 88.0%, and an F1-score of 88.3%. The mBERT model showed an accuracy of 90.0%, a recall of 88.0%, a precision of 89.5%, and an F1-score of 89.0%. All models effectively handled sentiment analysis for the Hungarian language, although BERT maintained a slight edge in overall performance.

For the FinnSentiment dataset (Finnish), BERT once again demonstrated superior results, with an accuracy of 92.1%, a recall of 90.1%, a precision of 91.5%, and an F1-score of 91.0%, confirming its effectiveness in analyzing Finnish texts. XLM-R, while also producing strong results with an accuracy of 90.3%, a recall of 88.5%, a precision of 89.0%, and an F1-score of 89.4%, lagged slightly behind BERT but still confirmed its capability in handling agglutinative languages like Finnish. The mBERT model achieved an accuracy of 91.5%, a recall of 89.5%, a precision of 90.5%, and an

F1-score of 90.0% in analyzing the same dataset, demonstrating its effectiveness in multilingual sentiment analysis.

Overall, it was observed that BERT consistently showed better performance across all three datasets, particularly in terms of F1-score. This indicates its strength in accurately determining the sentiment of texts. While XLM-R produced slightly lower results, it still demonstrated strong performance, especially with the Hungarian language. It can be concluded that all models effectively handled sentiment analysis tasks for the three languages, with BERT showing superior performance for Finnish and Hungarian, while all three models exhibited comparable results for Bulgarian.

7 Challenges in sentiment analysis across linguistically diverse languages

7.1 Lack of data

A significant challenge faced in the sentiment analysis of these languages is the insufficiency of large data corpora. While Finnish and Hungarian, the most widely spoken languages in this group, are supported by relatively extensive language databases and annotated datasets, smaller languages experience a severe scarcity of linguistic resources. The limited amount of publicly available text data for these languages, resulting from insufficient academic and technological focus, hinders the development of reliable sentiment analysis tools.

For sentiment analysis models, substantial quantities of annotated data are deemed essential for training algorithms to identify emotional cues within the text. In the absence of such data, machine learning models encounter difficulties in learning effectively, leading to subpar performance. Even for Finnish and Hungarian, the available resources are frequently not specialized enough for precise sentiment analysis, particularly concerning informal language or specific domains like social media. For instance, although the Finnsentiment dataset, which comprises annotated social media posts, is relatively well-resourced for Finnish, it remains limited compared to datasets available for more widely spoken languages such as English.

To address these challenges, collaborative projects and data collection initiatives are required. One potential solution involves engaging native speakers in the creation of text corpora via crowdsourcing platforms or community-driven

projects, similar to Mozilla's Common Voice initiative aimed at gathering voice data for lesser-spoken languages. Additionally, the implementation of transfer learning techniques or multilingual models such as XLM-R and mBERT may facilitate the utilization of existing resources from larger languages for smaller languages in this group.

7.2 Morphological complexity

Another significant challenge for sentiment analysis is posed by the morphological complexity of these languages. Known for their agglutinative nature, words in these languages are constructed by appending multiple suffixes to a root. Finnish and Hungarian, for example, possess extensive case systems, with Finnish having 15 cases and Hungarian having 18, as well as numerous derivational suffixes, resulting in a multitude of word forms derived from a single root. This complexity complicates tasks such as lemmatization (the process of reducing words to their base or root form) and normalization, both of which are crucial for accurate text processing in sentiment analysis.

In languages with simpler morphology, basic word forms are often directly matched to sentiment lexicons. However, in these languages, grammatical modifications such as possessive suffixes or case endings can alter the emotional content of words. For example, in Finnish, the addition of possessive or diminutive suffixes can impact the tone and sentiment of a word, rendering direct word matching ineffective for sentiment analysis. Without appropriate lemmatization and morphological analysis, the emotional context of a word may be misinterpreted by machine learning models, resulting in incorrect sentiment classification.

Recent advancements in models like XLM-R, BERT, and mBERT have demonstrated improvements in handling morphologically rich languages, owing to attention mechanisms that capture context-dependent information across multiple word forms. Studies focusing on Hungarian text processing have shown that adjusting these models to account for morphological factors leads to improved sentiment detection results. However, further efforts are needed to refine these models for various languages, ensuring accurate handling of the inherent morphological complexity.

7.3 Need for multimodal models

To achieve greater accuracy in sentiment analysis for these languages, a growing need has emerged

for the development of multimodal models that can analyze not only textual data but also additional inputs such as audio or visual data. Emotions in communication are frequently conveyed through non-verbal cues, including tone, intonation, and facial expressions, which may be overlooked when relying solely on text. This aspect is particularly relevant for these languages, where emotions can be expressed through intonation or contextual elements, making text-based analysis less effective.

For instance, in Finnish, the meaning of a sentence can be altered by intonation, consequently affecting the sentiment. The incorporation of audio data into sentiment analysis systems could enhance the model's capacity to capture emotional nuances that are not explicitly present in written text. Multimodal models, such as Speech2Text or Deep Audio-Visual Networks, are capable of processing both text and audio data, allowing for a more accurate interpretation of a speaker's tone, mood, and emotions.

Research on multimodal sentiment analysis has indicated that transformer-based models can be adapted to handle audio-visual data alongside textual input. The architecture of models like XLM-R and mBERT can be expanded to include additional data modalities, improving their performance in sentiment analysis tasks by considering factors such as intonation and facial expressions. The integration of these multimodal elements has proven particularly effective in languages where tone plays a crucial role in conveying emotions.

8 Possible directions for future research

8.1 Expansion of multilingual corpora

In future research, the expansion of multilingual corpora will be deemed critical for enhancing the quality of sentiment analysis in these languages. Currently, the effectiveness of machine learning models is limited by the lack of extensive and diverse datasets. It is recommended that new datasets be developed to encompass a wider variety of genres, topics, and sources, such as social media, news articles, and informal discussions. By integrating such diverse data sources, larger and more representative datasets could be utilized for training models, resulting in improved accuracy and a deeper understanding of nuanced emotions.

8.2 Adaptation of Existing Models

The potential for the adaptation of existing multilingual models, such as XLM-R and mBERT, has been highlighted as a promising approach. These models, originally trained on large multilingual corpora, can be fine-tuned specifically for this linguistic group using domain-specific and language-specific data. It is anticipated that fine-tuning will allow for better capture of the unique lexical, grammatical, and contextual structures found in Finnish, Hungarian, and other related languages. Through this process, it is expected that sentiment analysis in these languages will become more contextually aware and accurate in detecting sentiment nuances across different text types.

8.3 Development of new tools for analysis

Future advancements may also focus on the development of entirely new tools and algorithms specifically designed for these languages. This focus is particularly important given the agglutinative nature of many of these languages, wherein word formation involves adding suffixes to a root, rendering simple word matching ineffective. By creating specialized algorithms that address these unique linguistic features, significant improvements can be achieved in both the accuracy and speed of sentiment analysis. Such tools would further support the broader goal of automating sentiment detection across these languages, thereby facilitating its application in more diverse and complex scenarios.

9 Conclusion

The study demonstrated the effectiveness of multilingual models, specifically BERT, XLM-R, and mBERT, in performing sentiment analysis on linguistically diverse languages such as Finnish, Hungarian, and Bulgarian. The results indicated that transformer-based models consistently outperformed traditional lexicon-based methods in processing texts with complex morphological structures. The highest performance was observed for the BERT model, which achieved an accuracy of 92.1% on the FinnSentiment dataset, with a recall of 90.1%, precision of 91.5%, and an F1-score of 91.0%. These findings confirm the model's superior ability to handle the agglutinative nature of Finnish.

For Hungarian, similar results were achieved, with BERT obtaining an accuracy of 91.4%, a recall of 89.7%, precision of 90.2%, and an F1-score

of 90.5% on the Hungarian Emotion Corpus. The XLM-R model followed closely behind, achieving an accuracy of 89.2%, a recall of 87.5%, precision of 88.0%, and an F1-score of 88.3%. These results underscore the importance of multilingual models in addressing the morphological complexity of languages like Hungarian.

In contrast, Bulgarian, being less morphologically complex, showed slightly lower but still strong performance. BERT achieved an accuracy of 86.7%, a recall of 84.5%, precision of 87.0%, and an F1-score of 85.6% on the SentiFi dataset. The performance of mBERT was comparable, with an accuracy of 85.0%, a recall of 83.0%, precision of 84.5%, and an F1-score of 84.0%. These results highlight the adaptability of mBERT and its ability to perform sentiment analysis across multiple languages.

Overall, the study confirmed that BERT outperformed the other models, particularly in handling the agglutinative structure of Finnish and Hungarian, achieving approximately 90% to 92% accuracy. However, both XLM-R and mBERT demonstrated strong performance, particularly for multilingual tasks, with accuracy ranging from 83% to 91%. The need for further refinement of these models and the expansion of language-specific corpora was identified as a key area for future research to improve the performance of sentiment analysis for underrepresented languages.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONAHCYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaria de Investigacion y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONAHCYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercomputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

Additionally, we acknowledge the invaluable feedback and guidance provided by our peers during the review process. We are also grateful to the Instituto Politécnico Nacional for providing the necessary infrastructure and resources to carry out this research.

Finally, we extend our thanks to the developers

of open-source tools and libraries, whose work significantly facilitated the technical aspects of our project.

References

- Various Authors. 2020. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404.
- K. Bøhm et al. 2021. [Sentiment analysis in finnish with xlm-r](#). In *Proceedings of the 2021 International Conference on Language and Technology*. Association for Computational Linguistics.
- K. Dahlgren et al. 2020. [Deep learning models for sentiment analysis in uralic languages](#). In *Proceedings of the 2020 Conference on Computational Linguistics*. Association for Computational Linguistics.
- J. Devlin et al. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#). *Proceedings of the 2019 NAACL-HLT Conference*.
- Henri Konno et al. 2020. Cross-lingual transfer learning for sentiment analysis in uralic languages. *Proceedings of the Workshop on Computational Linguistics*.
- Krister Linden, Tommi Jauhiainen, and Sam Hardwick. 2018. Finnish sentiment lexicon for opinion mining. *Proceedings of the LREC 2018*.
- Krister Lindén, Tommi Jauhiainen, and Sam Hardwick. 2023. Finnsentiment: a finnish social media corpus for sentiment polarity annotation. *Language Resources and Evaluation*, 57(2):581–609.
- Ibomoiye Domor Mienye, Theo G Swart, and George Obaido. 2024. Recurrent neural networks: A comprehensive review of architectures, variants, and applications. *Information*, 15(9):517.
- T. Pahikkala et al. 2020. [Context-aware sentiment analysis for finnish texts](#). In *Proceedings of the 14th International Conference on Language Resources and Evaluation (LREC)*.
- Orsolya Ring, Martina Katalin Szabó, Csenge Guba, Bendegúz Váradi, and István Úveges. 2024. Approaches to sentiment analysis of hungarian political news at the sentence level. *Language Resources and Evaluation*, pages 1–29.
- K. Rudolph et al. 2021. [Ensemble language models for multilingual sentiment analysis](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.
- Bence Sass, Attila Novak, and Gabor Oros. 2020. Hungarian sentiment lexicon and annotated corpora for polarity detection. *Journal of Language Technology*.

- C. Strapparava and A. Valitutti. 2004. [Wordnet-affect: an affective extension of wordnet](#). In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*.
- Moein Tash, Jesus Armenta-Segura, Zahra Ahani, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023. [Lidoma@dravidianlangtech: Convolutional neural networks for studying correlation between lexical features and sentiment polarity in tamil and tulu languages](#). In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 180–185, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- Mohammad Mustafa Taye. 2023. Theoretical understanding of convolutional neural networks: Concepts, architectures, applications, future directions. *Computation*, 11(3):52.
- S. Tzeng et al. 2021. [Multilingual sentiment analysis with transformer-based models](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.
- A. Virtanen et al. 2019. [Multilingual bert-based model for sentiment analysis in finnish texts](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.
- A. Virtanen et al. 2020. Multilingual bert for low-resource languages. *Proceedings of ACL 2020*.
- Xia Zhao, Limin Wang, Yufei Zhang, Xuming Han, Muhammet Deveci, and Milan Parmar. 2024. A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, 57(4):99.

Towards standardized inflected lexicons for the Finnic languages

Jules Bouton

Université Paris-Cité, LLF, CNRS
jules.bouton@u-paris.fr

Abstract

We introduce three richly annotated lexicons of nouns for Livonian, standard Finnish and Livvi Karelian. Our datasets are distributed in the machine-readable Paralex standard, which consists of linked CSV tables described in a JSON metadata file. We built on the morphological dictionary of Livonian, the VepKar database and the Omorfi software to provide inflected forms. All noun forms were transcribed with grapheme-to-phoneme conversion rules and the paradigms annotated for both overabundance and defectivity. The resulting datasets are usable for quantitative studies of morphological systems and for qualitative investigations. They are linked to the original resources and can be easily updated.

1 Introduction

1.1 Rationale

Over recent years, the amount of morphological resources available for the Uralic languages has strongly increased. Reasons for this are (a) the efforts of Finno-Ugrists to provide dictionaries and translation tools for minority languages; (b) the interest of typologists for computational approaches to linguistic diversity. However, these resources are scattered across different standards and do not necessarily fit the needs of morphologists. Although recent researches in computational morphology rely on various approaches (e.g. Malouf, 2017; Baayen et al., 2019; Beniamine et al., 2021), they all share the need for high quality morphological data in phonemic transcription.

Several projects strive to provide good coverage of the numerous Finnic languages. Recently, lexicons following the UniMorph format have flourished: Finnish (Kirov et al., 2016); Estonian and Northern Sami (Kirov et al., 2018); Livvi, Livonian and several other (McCarthy et al., 2020); Võro (Batsuren et al., 2022). Despite its increasing size, Malouf et al. (2020) have shown the pit-

falls of UniMorph when it comes to linguistically informed studies of morphological variation. Semantic information, inflectional classes or frequencies are hard to extract and wordforms are provided in orthographical representations. The GiellaLT infrastructure (Pirinen et al., 2023) also provides access to dozens of morphological rule-based parsers. However, they are intended to enhance language-learning tools and they are not meant for morphological investigation either.

On the other hand, scholars and language institutes have developed their own resources, providing both inflected forms and rich annotation. Such resources are invaluable, but there is few of them. As a result of their dispersal, they are provided in different formats and through idiosyncratic infrastructures which make them less accessible for large scale comparative studies. Still, efforts for interoperability exist: in UniMorph 3.0, resources for Karelian languages are directly extracted from the VepKar database (McCarthy et al., 2020), although a lot of information is lost in the conversion, due to the limits of the UniMorph format.

Our lexicons in phonemic transcription are designed to fill this gap. We selected valuable, well-curated and rich resources for three Uralic languages from the Finnic group with very different backgrounds. Standard Finnish is the national language of Finland, spoken by around five million people in Finland.¹ Livvi Karelian is a southern Karelian language spoken by 25,000 individuals in Russia, near lake Ladoga. Courland Livonian is a minority language spoken until the end of the 20th century on the coast of Courland. Although our pipeline can in theory be extended to verbs, this release only covers nouns. As our main contribution, we enriched the datasets with phonemic transcriptions and linguistic annotations.

¹Statistics are from the corresponding chapters of Bakró-Nagy et al. (2022).

Dataset	ISO	Licence	DOI	Cells	Lexemes	Forms	
<i>ParaLiv</i>	1.0	liv	CC BY-SA 4.0	10.5281/zenodo.11391421	16	6,769	110,449
<i>ParaKar</i>	1.0	olo	CC BY-SA 4.0	10.5281/zenodo.13736171	33	4,975	196,555
<i>ParaFin</i>	1.0	fin	GNU GPL v3	10.5281/zenodo.13736132	151	5,000	879,117

Table 1: Main properties of the three datasets

1.2 The Paralex format

Beniamine et al. (2023) introduced the Paralex standard², which provides a structured way of representing morphological data. A Paralex dataset is a relational database constituted of CSV files linked together by relations. Beniamine et al. (2024) provide a detailed presentation of the structure of such a dataset. Thanks to the underlying Frictionless framework (Fowler et al., 2017), a Paralex dataset is adaptable to one’s needs but also machine-readable. Thus, the Paralex standard puts good data management practices (FAIR: Wilkinson et al. 2016 ; DEAR: Beniamine et al. 2023) at the core of the dataset development.

Paralex datasets are intended for morphologists. As such, they offer two crucial improvements over other formats: phonemic representations and rich annotations. Since orthographic representations of words often obfuscate crucial features, the inflected forms are provided both in orthographic and phonemic writing. The phonemic transcriptions are checked on development sets to avoid regressions and cover most of the morphologically meaningful contrasts. Allophony is left out when it doesn’t affect morphology. Paralex takes into account morphological diversity and has built-in methods to tag variants or defectivity (see below).

Our datasets follow these principles. They are made available on Zenodo under the names *ParaKar*, *ParaFin* and *ParaLiv* (see Table 1). The pipelines used to build the lexicons are available on Gitlab and ensure replicability of the results. Changes in the upstream sources can easily lead to updated versions of the datasets thanks to Zenodo’s versioning system. They are distributed under open-source licences.

2 Building the lexicons

2.1 Lexemes and forms

For Livonian, we relied on the morphological dictionary of the Livonian Institute (Ernštreits et al.,

2024), which itself builds on the Livonian dictionary by Viitso and Ernštreits (2012). In the absence of reliable frequency information, we provide support for all the nouns in the dictionary. We extracted the inflected wordforms and their properties as a JSON file and controlled the quality of the forms. A dozen of lexemes required upstream corrections and were ruled out. All the cells available in the dictionary were retained, which does not include lexicalized external local case forms. In compounds, the boundary between the components is marked. For phonological reasons, the derivatives ending in *-nikā* were treated as compounds, following Posti (1942, 301).

Similarly to Paralex datasets, the VepKar corpus used for Karelian (VepKar, 2009/2024; Boyko et al., 2022) is a relational database with annotated tables. Thus, converting the extracted tables was rather straightforward, despite the difference in the data structure (resp. CSV and MySQL). As for Livonian, the VepKar database provided pre-inflected forms for Livvi (Novak et al., 2020; Krizhanovskaya et al., 2024). Since VepKar has a better support for New Written Livvic, we focused on this variety of Livvi and excluded forms from other dialects. We retained all the lexemes that were attested at least once in the corpus. We replaced the accusative cell used in VepKar by genitive and nominative labels, depending on the form in question.³ We additionally filtered the database and corrected a few forms. VepKar features an affix column, which made it possible to insert a boundary in wordforms after the immutable part of the stem. Compounds are segmented.

The situation of Finnish is different as we did not use a database of wordforms. We selected the 5000 most frequent nouns from the frequency dataset provided with the LASTU software (Itkonen

²<https://paralex-standard.org/>

³With respect to the accusative, the situation in Karelian is similar to that in Finnish. Bielecki (2009) shows that older descriptive grammars introduced an accusative while recent accounts only feature nominative and genitive. While syntacticians tend to agree in favour of an accusative (Holmberg and Nikanne, 1993), we adopt here a morphological perspective.

et al., 2024), which in turn relies on the Finnish Parsebank (Luotolahti et al., 2015). We then matched those nouns with the internal resources of the Omorfi HFST (Pirinen, 2015; Pirinen et al., 2017) and used the generator to produce inflected forms. Although the interaction of clitics, case and number markers and personal suffixes leads to a large amount of paradigm cells, we decided to only retain the combination of case, number and possessive suffixes. In our dataset, this already amounts to 151 cells.⁴ Compounds and immutable stem boundaries are marked as well.

Table 1 summarizes the quantitative properties of the extracted datasets. All have around 5000 lexemes, which is a standard size for such resources (Beniamine et al., 2024).

2.2 Phonemic transcriptions

Grapheme-to-phoneme (G2P) transcription was performed with the Epitran software (Mortensen et al., 2018). Epitran requires a mapping of graphemes to phonemes and a set of pre- and post-processing regular expressions. For our datasets, we used a bundle of custom and modified rules.

For Livonian, we used a heavily modified version of the Estonian rules built for the Eesthetic package (Beniamine et al., 2024). Traditional accounts of Livonian phonology (Posti, 1942; Viitso, 2007) introduced numerous distinctions which are not always crucial for a phonemic description. For our transcription we relied on Tuisk’s (2016) analysis and complemented it with previous accounts. We review the most crucial design choices.

Traditional accounts distinguish between short phonemes, long phonemes, short geminates and long geminates. We decided to keep a three-fold distinction for consonants and a two-fold opposition for vowels (ex 1). Due to the existence of feet isochrony (Viitso, 2007, 49), we mark the length of the first syllable coda when the second syllable is short (ex 2). Livonian is known for its tonal opposition (broken or plain) which affects accented syllables (Tuisk, 2015). We transcribe the broken tone as a property of vowels and polyphthongs and mark it with a superscript glottal stop ^ʔ (ex 3). We insert glides where required before orthographic <ž>, <j> and <v> (ex 3). Finally, Livonian displays a wide range of polyphthongs which were all documented. Table 2 showcases the triphthongs.

⁴For possessives, the values 3SG and 3PL are treated as syncretic. For instance, the cell NOM.SG.3 covers singular and plural possessors.

	Front-back		Back-front	
	Plain	Broken	Plain	Broken
All short	ieu	ie ^ʔ u	uoi	uo ^ʔ i
Last long	ieu:		uoi:	
First long			u:oi	u: ^ʔ oi

Table 2: Inventory of Livonian triphthongs found in our dataset

- | | | | | |
|-----|----|------------------------------|-----------|---------|
| (1) | a. | kik → kik: | ‘rooster’ | NOM.SG |
| | b. | kikīd → kik ^ʔ i:d | | NOM.PL |
| | c. | kikkō → kik:u | | PART.SG |
| (2) | a. | mustā → mus ^ʔ ta: | ‘black’ | NOM.SG |
| | b. | mustō → mus:tu | | PART.SG |
| (3) | | ke’ž → ke ^ʔ jʒ | | ‘flea’ |

For Finnish, we used a modified version of the Finnish G2P converter introduced in Epitran 1.25. We don’t mark the allophones of /h/, /s/, /l/, /m/, /n/ (ex 4), but we added additional rules to distinguish diphthongs from vowel sequences (ex 5) in conformity with Suomi et al. (2008, 49-51). We marked as a glottal stop the stop that alternates with intervocalic /k/ during gradation (ex 6). Following Karlsson’s (1983, 349) view, morphs triggering boundary lengthening were not considered in the phonemic transcription, but we documented them in the analysed orthographic and phonemic transcriptions with the superscript symbol ^x (ex 6).

- | | | | | |
|-----|----|-------------------------------|--|-------------|
| (4) | a. | vihko → vihko | | ‘notebook’ |
| | b. | kohta → kohta | | ‘place’ |
| (5) | a. | hyötyä → hyötyæ | | ‘benefit’ |
| | b. | aie → gje (gje ^x) | | ‘intention’ |
| (6) | | vaa’an → va: ^ʔ an | | ‘scale’ |

The Karelian G2P is a slightly modified version of the Finnish one. It is based on Pyöli (2011), but was extended with more detailed sources (Novak et al., 2022; Arhimaa, 2022). The Livvi transcription covers the digraphs and affricates specific to Karelian (ex 7) and introduces support for the contextual palatalization of /l/, /n/, /r/, /d/ and /t/ (ex 8) following the principles described by Novak et al. (2022, 58). We included palatalized and voiced geminates and we took into account the existence of six triphthongs, although they do not occur in our dataset as they are limited to verbs.

- | | | | | |
|-----|--|-----------------------------------|--|----------|
| (7) | | čondžoi → tʃondʒoi | | ‘flea’ |
| (8) | | ellendys → el ^ʃ :endys | | ‘wisdom’ |

–	Our datasets	ill.sg
–	UD	N; IN+ALL; SG
–	UniMorph	Case=Ill Number=Sing
fin	Omorfi	[NUM=SG][CASE=ILL]
olo	VepKar ID	10
liv	Liv. Institute	IllSg
liv	Tartu	sg.ill.

Table 3: Mapping of the ILL.SG cell to other dialects

3 Rich annotations

Phonemes and graphemes For each dataset, we provided a grapheme inventory to ensure consistency in our orthographical sources. All three datasets also contain a machine-readable phoneme inventory with contrasting articulatory features.

Features-values To ensure compatibility with external resources, we linked our features and values to other standards. All datasets contain mappings to UniMorph (Sylak-Glassman et al., 2015) and Universal Dependencies (Nivre et al., 2016) dialects. Additionally, *ParaLiv* maps to the referential used by the Livonian Institute and the University of Tartu dialect corpus (Lindström et al., 2022), *ParaFin* maps to the Omorfi encoding and *ParaKar* to the VepKar unique identifiers. These mappings have proven valuable in extracting token frequencies (see below). An overview of the mappings offered in the three datasets is provided for the illative singular cell in Table 3.

Overabundance and defectivity In the Paralex format, each wordform is assigned a record. If two forms are available for a given cell, a case of overabundance (Thornton, 2019), two records are created. If a cell has no known form, a record is still created with the label #DEF#. For such non-canonical phenomena, we provide semantic annotations to distinguish overabundant forms and to make explicit the reason for defectivity. For instance, in Finnish, the third person possessive suffix takes two forms: *-nsA* or *-Vn*. Such forms are tagged *poss_nsA* and *poss_Vn*. A record can have several tags.⁵ Concerning defectivity, Omorfi and VepKar tend to provide extensive paradigms.⁶

⁵Some forms follow idiosyncratic patterns and are not tagged. The percentage of untagged forms is: 1.27% in *ParaFin*, 4.24% in *ParaLiv* and 5.08% in *ParaKar*.

⁶In Omorfi, only *pluralia tantum* appear as defective.

This can partly be explained by the difficulty of assessing the defectivity of a given form, due to low frequency effects (Nikolaev and Bermel, 2023).

Frequencies Paralex lexicons can optionally store frequencies at three different levels: cells, forms and lexemes. As for our lexicons, we provide all frequencies for Finnish and Livvi, but only cell frequencies for Livonian.

The frequencies were extracted from the Finnish dataset provided with the LASTU software (Itkonen et al., 2024), which in turn relies on the Finnish Parsebank (Luotolahti et al., 2015). We used the frequency table for forms occurring at least 10 times in the parsebank. We matched the universal dependency features used in the original dataset with our own cells and ruled out all inconsistent annotations. In further versions of the dataset, we plan to introduce frequencies directly extracted from the parsebank. For lexemes, we use the cumulated lexeme frequencies already provided by the LASTU dataset. For Karelian, we used the annotated VepKar corpus to extract form, lexeme and cell frequencies. For Livonian, we extracted word frequencies from the Estonian Dialects Corpus (Lindström et al., 2019, 2022) and grouped them by cell. This corpus was too small to assign a frequency to the lexemes or to the forms.

4 Conclusion

We introduced inflected lexicons for three Finnic languages: Livonian, Finnish and Livvi. We reviewed current practices in Uralic language resources and emphasized the importance of rich, machine-readable formats to facilitate cross-linguistic studies of morphological systems. We presented the design choices for our datasets and introduced our linguistically motivated grapheme-to-phoneme rules. We outlined the annotations that we performed. Appendix A showcases the main tables of one of the resulting datasets.

Although we did our best to manually check the transcriptions by evaluating random samples of forms and by carrying out targeted verifications, it is very likely that some mistakes remain, especially for loanwords. In addition to improved transcriptions, further versions should include more morphological annotations (e.g. information on stem gradation according to traditional descriptions) and reference other sources of frequencies (especially for Finnish). The datasets could also be extended to verbal inflection.

Acknowledgments

We are thankful to the Livonian Institute, the Karelian Research Center and the developers of Omorfi for providing access to their morphological resources. The help of Tuuli Tuisk and Valts Ernreits was precious in answering our questions about Livonian phonology and the resources of the Livonian Institute. Olivier Bonami provided comments on this paper and Sacha Beniamine answered questions related to the datasets.

References

- Anneli Arhima. 2022. *Karelian*. In Marianne Bakró-Nagy, Johanna Laakso, and Elena Skribnik, editors, *The Oxford Guide to the Uralic Languages*, pages 269–290. Oxford University Press, Oxford.
- R. Harald Baayen, Yu-Ying Chuang, Elnaz Shafaei-Bajestan, and James P. Blevins. 2019. *The Discriminative Lexicon: A Unified Computational Model for the Lexicon and Lexical Processing in Comprehension and Production Grounded Not in (De)Composition but in Linear Discriminative Learning*. *Complexity*, 2019:39.
- Marianne Bakró-Nagy, Johanna Laakso, and Elena Skribnik. 2022. *The Oxford Guide to the Uralic Languages*. Oxford Guides to the World’s Languages. Oxford University Press, Oxford.
- Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóga, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marc Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud’hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, pages 840–855, Marseille, France. European Language Resources Association (ELRA).
- Sacha Beniamine, Mari Aigro, Matthew Baerman, Jules Bouton, and Maria Copot. 2024. Eesthetic: A Paralex Lexicon of Estonian Paradigms. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5526–5537, Torino, Italia. European Language Resources Association (ELRA) and International Committee on Computational Linguistics (ICCL).
- Sacha Beniamine, Cormac Anderson, Mae Carroll, Matías Guzmán Naranjo, Borja Herce, Matteo Pellegrini, Erich Round, Helen Sims-Williams, and Tiago Tresoldi. 2023. Paralex: a DeAR standard for rich lexicons of inflected forms. In *International Symposium of Morphology (ISM0 2023)*, Nancy, France. <https://www.paralex-standard.org>.
- Sacha Beniamine, Olivier Bonami, and Ana R. Luís. 2021. The fine implicative structure of European Portuguese conjugation. *Isogloss. Open Journal of Romance Linguistics*, 7:1–35.
- Robert Bielecki. 2009. On the Nature of the Accusative in Finnish. *Lingua Posnaniensis*, 51(1):19–38.
- Tatyana Boyko, Nina Zaitseva, Natalia Krizhanovskaya, Andrew Krizhanovsky, Irina Novak, Nataliya Pellinen, and Aleksandra Rodionova. 2022. The Open Corpus of the Veps and Karelian Languages: Overview and Applications. *KnE Social Sciences*, 7(3):29–40.
- Valts Ernštreits, Tiit-Rein Viitso, and Milda Kurpniece. 2024. Livonian morphology database. <http://www.livonian.tech>.
- Dan Fowler, Jo Barratt, and Paul Walsh. 2017. Frictionless Data: Making Research Data Quality Visible. *International Journal of Digital Curation*, 12(2):274–285.
- Anders Holmberg and Urpo Nikanne, editors. 1993. *Case and Other Functional Categories in Finnish Syntax*. Number 39 in Studies in Generative Grammar. De Gruyter Mouton.
- Sami Itkonen, Tuomo Häikiö, Seppo Vainio, and Minna Lehtonen. 2024. LASTU: A psycholinguistic search tool for Finnish lexical stimuli. *Behavior Research Methods*, 56(6):6165–6178.

- Fred Karlsson. 1983. *Suomen kielen äänne- ja muotorakenne*. Werner Söderström, Porvoo, Finland.
- Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. Very-large Scale Parsing and Normalization of Wiktionary Morphological Paradigms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3121–3126, Portorož, Slovenia. European Language Resources Association (ELRA).
- Natalia Krizhanovskaya, Irina Novak, and Nataliya Pellinen. 2024. *Pravila generacii imennyh slovoform po minimizirovannomu šablonu dlâ novopis'mennyh variantov sobstvenno karel'skogo i livvikovskogo narečij*.
- Liina Lindström, Pärtel Lippus, and Tuuli Tuisk. 2019. The online database of the University of Tartu Archives of Estonian Dialects and Kindred Languages and the Corpus of Estonian Dialects. *Uralica Helsingiensia*, (14):327–350.
- Liina Lindström, Triin Todesk, and Maarja-Liisa Pilvik. 2022. *Corpus of Estonian Dialects*.
- Juhani Luotolahti, Jenna Kanerva, Veronika Laippala, Sampo Pyysalo, and Filip Ginter. 2015. Towards Universal Web Parsebanks. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 211–220, Uppsala, Sweden. Uppsala University.
- Robert Malouf. 2017. *Abstractive morphological learning with a recurrent neural network*. *Morphology*, 27(4):431–458.
- Robert Malouf, Farrell Ackerman, and Arturs Semenuks. 2020. Lexical databases for computational analyses: A linguistic perspective. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 446–456, New York, New York. Association for Computational Linguistics (ACL).
- Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miiikka Silfverberg, Timofey Arkhangelskiy, Natalya Krizhanovskaya, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*, pages 3922–3931, Marseille, France. European Language Resources Association (ELRA).
- David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. EpiTran: Precision G2P for Many Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2711–2714, Miyazaki, Japan. European Language Resources Association (ELRA).
- Alexandre Nikolaev and Neil Bermel. 2023. *Studying negative evidence in Finnish language corpora*. *Word Structure*, 16(2-3):206–232.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).
- Irina Novak, Natalia Krizhanovskaya, Tat'jana Bojko, and Nataliya Pellinen. 2020. *Development of rules of generation of nominal word forms for new-written variants of the Karelian language*. *Bulletin of Ugric studies*, 10(4):679–691.
- Irina Novak, Martti Penttonen, Alekski Ruuskanen, and Lea Siilin. 2022. *Karelian in Grammars : A study of phonetic and morphological variation*. Karelian Research Centre of the Russian Academy of Sciences, Petroskoi.
- Flammie Pirinen, Sjur Moshagen, and Katri Hiovain-Asikainen. 2023. GiellaLT — a stable infrastructure for Nordic minority languages and beyond. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 643–649, Tórshavn, Faroe Islands. University of Tartu Library.
- Tommi A. Pirinen. 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.
- Tommi A. Pirinen, Inari Listenmaa, Ryan Johnson, Francis M. Tyers, and Juha Kuokkala. 2017. *Open morphology of Finnish*. University of Helsinki.
- Lauri Posti. 1942. *Grundzüge der livischen Lautgeschichte*. Number 75 in *Mémoires de la Société Finno-Ougrienne*. University of Helsinki, Helsinki.
- Raija Pyöli. 2011. *Livvinkarjalan kielioppi*. Karjalan kielen seura, Helsinki.
- Kari Suomi, Juhani Toivanen, and Riikka Ylitalo. 2008. *Finnish sound structure: phonetics, phonology, phonotactics and prosody*. Number 9 in *Studia Humaniora Ouluensia*. University of Oulu, Oulu.

- John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. 2015. **A Language-Independent Feature Schema for Inflectional Morphology**. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 674–680, Beijing, China. Association for Computational Linguistics (ACL).
- Anna Maria Thornton. 2019. **Overabundance: A Canonical Typology**. In Franz Rainer, Francesco Gardani, Wolfgang U. Dressler, and Hans Christian Luschützky, editors, *Competition in Inflection and Word-Formation*, number 5 in *Studies in Morphology*, pages 223–258. Springer, Cham.
- Tuuli Tuisk. 2015. Acoustics of Stød in Livonian. In *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, UK. University of Glasgow.
- Tuuli Tuisk. 2016. **Main features of the Livonian sound system and pronunciation**. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 7(1):121–143.
- VepKar. 2009/2024. Open Corpus of Veps and Karelian languages. <http://dictorpus.krc.karelia.ru/>.
- Tiit-Rein Viitso. 2007. **Livonian Gradation : Types and Genesis**. *Linguistica Uralica*, 43(1):45.
- Tiit-Rein Viitso and Valts Ernštreits. 2012. *Līvõkīel-ēstikīel-leŕkīel sōnārōntōz = Liivi-eesti-läti sōnaraamat = Lībiešu-īgaunu-latviešu vārdnīca*. University of Tartu, Tartu.
- Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. **The FAIR Guiding Principles for scientific data management and stewardship**. *Scientific Data*, 3(1):160018.

A Appendix: Sample tables from the *ParaLiv* dataset

(a) The forms table

form_id	lexeme	cell	orth_form	phon_form	analysed_orth_form	analysed_phon_form	defectiveness_tag	overabundance_tag
võrböz_22999-dat.pl	võrböz_22999	dat.pl	võrbödön	v u : r b u d u n	võrbödön	v u : r b u d u n		
irī_13393-ill.sg-1	irī_13393	ill.sg	irī	ir i :	irī	ir i :		illsg_without_z
ērškōmōrapōzō_12800-ine.sg	ērškōmōrapōzō_12800	ine.sg	ērškōmōrapōzōs	er f k u m o : p a p' u i : z u s	ērškōmōrapōzōs	er f k u m o : p a p' u i : z u s		
passōr_18233-dat.pl	passōr_18233	dat.pl	passōrdön	p a s : u r d u n	passōrdön	p a s : u r d u n		
vikāt_22643-ine.pl	vikāt_22643	ine.pl	vikāis	v i k' a : t i s	vikāis	v i k' a : t i s		
nōrkōz_17668-ine.pl	nōrkōz_17668	ine.pl	nōrkōzīs	n u : p k u z i s	nōrkōzīs	n u : p k u z i s		
silmadkōp_20193-Imm.pl	silmadkōp_20193	Imm.pl	silmadkōpōdōks	s i : l m a d k o : p' u d u k s	silmadkōpōdōks	s i : l m a d k o : p' u d u k s		
kōrtami_15156-gen.pl	kōrtami_15156	gen.pl	kōrtamizt	k u : r t a m i s t	kōrtamizt	k u : r t a m i s t		
suōm_20758-nom.pl	suōm_20758	nom.pl	sūoimōd	s u o i p m u d	sūoimōd	s u o i p m u d		
saländöm_19842-ela.pl	saländöm_19842	ela.pl	saländōmist	s a l a : n d u m i s t	saländōmist	s a l a : n d u m i s t		

(b) The lexemes table

lexeme_id	label	inflection_class	POS
armäiga_11913	armäiga	101	noun
skūolsoppjij_20388	skūolsoppjij	286	noun
ministrij_16883	ministrij	199	noun
kūjabulā_15356	kūjabulā	83	noun
pūlēd_19079	pūlēd	234	noun
azūmsōnā_11992	azūmsōnā	83	noun

(c) The sounds table

sound_id	CLTS_id	syllabic	stress	long	half:long	consonantal	...
j:	j:	0		1		0	...
k'	k'	0		0	1	1	...
i'ʔu		1	1	0		0	...
p'		0		0	1	1	...
ieu:		1	1	0		0	...
o:ʔ		1	1	1		0	...

(d) The graphemes table

grapheme_id	comment	canonical_order
a		3
ā		4
k		18
o		24
r		29
z		39

(e) The values table

value_id	label	POS	feature	unimorph	ud	livonian_tech	tartu	canonical_order
nom	nominative	noun	case	NOM	Case=Nom	Nom	nom	1
gen	genitive	noun	case	GEN	Case=Gen	Gen	gen	2
prt	partitive	noun	case	PRT	Case=Par	Prt	part	3
dat	dative	noun	case	DAT	Case=Dat	Dat	dat	4
Imm	instrumental-comitative	noun	case	INS	Case=Ins	Imm	tr	5
ill	illative	noun	case	IN+ALL	Case=Ill	Ill	ill	6

(f) The cells table

cell_id	POS	unimorph	ud	livonian_tech	tartu	frequency
nom.sg	noun	N;NOM;SG	Case=Nom Number=Sing	NomSg	sg.nom.	5410
gen.sg	noun	N;GEN;SG	Case=Gen Number=Sing	GenSg	sg.gen.	3941
dat.sg	noun	N;DAT;SG	Case=Dat Number=Sing	DatSg	sg.dat.	762
prt.sg	noun	N;PRT;SG	Case=Prt Number=Sing	PrtSg	sg.part.	1890
Imm.sg	noun	N;INS;SG	Case=Ins Number=Sing	ImmSg	sg.tr.	177
ill.sg	noun	N;IN+ALL;SG	Case=Ill Number=Sing	IllSg	sg.ill.	877

(g) The tags table

tag_id	tag_column_name	comment
defective	defectiveness_tag	defective for unknown reasons
pluralia_tantum	defectiveness_tag	defective in singular because pluralia tantum
illsg_without_z	overabundance_tag	a parallel form for illatives; without z final consonant
illsg_with_z	overabundance_tag	a parallel form for illatives; with z final consonant
elasg_without_ö	overabundance_tag	a parallel form for consonantal words; with ö final vowel
elasg_without_õ	overabundance_tag	a parallel form for consonantal words; without ö final vowel

Table 4: Excerpts from the forms, lexemes, cells, sounds, tags, graphemes, features tables from the *ParaLiv* package. Primary keys have a grey shading.

On Erzya and Moksha Corpora and Analyzer Development, ERME-PSLA 1950s

Jack Rueter¹, Olga Erina² and Nadezhda Kabaeva³

¹University of Helsinki

²Independent researcher

³Mordovian State University

¹first.last@helsinki.fi

²first.jerina@gmail.com

Abstract

This paper describes materials and annotation facilitation pertinent to the «Erzya-Moksha Electronic Resources and Linguistic Diversity» (EMERALD) project. It addresses work following the construction of finite-state analyzers for the Mordvin languages, the gathering of test corpora, and the development of metadata strategies for descriptive research.

In this paper, we provide three descriptors for a set of new Erzya and Moksha research materials at the Language Bank of Finland. The descriptors illustrate (1) a low-annotation subcorpora set of the «Electronic Resources for Moksha and Erzya» (ERME); (2) the state of the open-source analyzers used in their automatic annotation, and (3) the development of metadata documentation for the «EMERALD» project, associated with this endeavor.

Outcomes of the article include an introduction to new research materials, an illustration of the state of the Mordvin annotation pipeline, and perspectives for the further enhancement of the annotation pipeline.

1 Introduction

The Mordvin languages, Erzya and Moksha, are spoken in settlements scattered throughout the Volga Basin (see [Rueter 2013](#)), but there have been settlements beyond that as well (see [Sarv 2002](#)).

Work with the description of these languages dates from the late 1600s in the form of word lists [Witsen 1705](#), but it was not until the end of the 1830s that the first attempts at grammars were made for Moksha [Ornatov 1838](#) and Erzya [Gabelentz 1838–1839](#).

The Erzya grammar by Herr Conon von der Gabelentz is a striking study in that it illustrates the author’s meticulous parallel-corpus-type knowledge of the Biblical texts. With

this knowledge, Hhe was able to identify irregularities in the text and draws relatively accurate conclusions with regard to the meaning of morphological items¹. The text of the Erzya Gospel is partially available now in the «Parallel Bible Verses for Uralic Studies» (PaBiVUS) corpus version 1 (see [Helsingin yliopisto, FIN-CLARIN et al. 2020-06-07](#)), but the next version of PaBiVUS will see the entire New Testament in Erzya from 1827, which can be aligned with translations into several other minority Uralic languages.

In this paper, we describe a new portion of the Electronic Resources Moksha-Erzya (ERME) and its annotation, which is soon to be introduced on the Language Bank of Finland Korp server.

The original intent of the ERME corpora was to provide the language community and researchers with citeable materials that distinguish writers and other language sources both geographically and in temporal space.

This meant the establishment of metadata features to served the purposes of the «EMERALD» project: provide parallel, consistent metadata for (1) Fieldwork and Early Literary Texts (FELT), with a focus on language materials collected in the Pre-Soviet Era, such as the Mordvin fieldwork collections by Heikki Paasonen and others (cf. [Finno-Ugrian-Society, Suihkonen 2003](#)); (2) non-central publications in the minority languages of the Soviet Union between the two World Wars², and (3) work

¹The grammar in full can be accessed here: https://rueter.github.io/emerald/historical-mordvin-grammars/docs/gabelentz_hcvonder-versuch-einer-mordwinischen-grammatik-1838-39.html

²Here is a collection of openly licensed printed media including those from the era of minority-language popularization in the USSR. Outcomes of the Kindred Language Digitization pilot at the National Library of Finland with funding from the Kone Foundation «Language Programme» <https://fennougrica.kansalliskirjasto.fi/>

with modern collection of the Mordvin language.

The metadata was seen as a means to align research and fieldwork documentation of the languages with texts from an era of language popularization, subsequent fieldwork conducted with these languages and their modern state. Thus, the metadata was organized to describe materials from the Heikki Paasonen collections of folklore, the Fenno-Ugrica collections at the National Library of Finland, the MORMULA corpora in Turku, the dialect archives at the Mordovian State University in Saransk (see Rueter 2020; Kabaeva 2021; Agafonova and Râbov 2021), and the ERME corpora at the Language Bank of Finland. This work was conducted in Turku, Helsinki and Saransk (2018–2022) in close association with corpora lemmatization, Constraint Grammar development and testing in Turku, dialect archive development in Saransk, and subsequent work with corpora in Helsinki.

In addition to metadata strategies for research citation, the annotation of the ERME corpora can be described as a combination of work with optical character recognition (OCR) and the development of finite-state descriptions for the morphology and syntax of the languages.

Over the years, attempts have been made to bring these two more tasks closer together. In the 1990s, OCR meant accurate recognition of individual characters. Since then attempts have been made to include finite-state work in OCR with hopes of word form recognition Silfverberg and Rueter 2015 for improved accuracy. Simple examples of finite-state approaches also include plain word lists or regular expressions used to represent numerous Uralic languages in OCR work in pilot projects conducted at the National Library of Finland³.

Subsequent work with Mordvin lexica, morphology and syntax have played a major role in finding a purpose and collaboration beyond these languages. Work in lexica has meant collaboration in Erzya and Moksha with NorthEuraLex together with Mordovian State University staff⁴, but it has also meant

³https://www.doria.fi/bitstream/handle/10024/101915/Hakkarainen_Tallinn_19112014.pdf?sequence=2

⁴Erzya <http://www.northeuralex.org/languages/myv> and Moksha <http://www.northeuralex.org/>

the development of dictionary editing (see Hämäläinen et al. 2021) and the enhancement of these dictionaries (see Alnajjar et al. 2022). Work with morpho-syntax, in turn, has opened connections to collaboration with specialist in the Komi languages in Syktyvkar (see Rueter et al. 2021) and the Universal Dependencies project with contributions to languages beyond Erzya and Moksha Zeman and et al. 2024. All of these together have contributed to utilizing the Giellatekno/Divvun⁵ infrastructure GielLaLT (an infrastructure for Saami Language-Technology research and facilitation) and the open-source, shallow-transfer machine translation infrastructure Apertium⁶. The Apertium concept of shallow transfer makes it possible to draw parallels between lexicon, morphology, syntax and phraseology for the inspection of language diversity among closely related languages (see Rueter and Hämäläinen 2020, Rueter 2022). It also allows for a better comparison of closely related languages in linguistic research (cf (Rueter, 2023; Rueter and Kabaeva, 2024)).

The article proceeds by discussing the materials, metadata and methods, the state of the individual analyzers and prospects for future development.

2 Materials and methods

Despite a history of over 200 years of published texts in Erzya and Moksha, there is a dearth of searchable Mordvin text corpora consistently annotated for morphology, metadata and openly accessible. For this reason the Erzya-Moksha Electronic Resources And Language Diversity (EMERALD) project continues to augment metadata enriched materials available to the research communities (Rueter 2024).

The Electronic Resources for Moksha-Erzya (ERME) corpora versions one and two have been made available through Fin-CLARIN on the Language Bank of Finland Korp server since 2017. The first version of ERME focused merely on providing necessary metadata to facilitate a better alignment of text and authors, such that time-line and geographical plotting could

languages/mdf

⁵<https://giellalt.github.io/>

⁶https://wiki.apertium.org/wiki/Main_Page

be made consistent with analogous information available for Erzya-language fieldwork. In the second version of ERME, however, the materials were extended to include both Mordvin literary languages – Erzya and Moksha – with metadata for individual publications, and automatic annotation made possible with finite-state analyzers for the two languages.

Morpho-syntactic annotation has been accomplished using finite-state transducers describing the two languages (Lindén et al., 2013; Rueter et al., 2020). These, in turn, have been followed in the same pipeline by Constraint Grammar (CG) disambiguation, functions and dependencies tools⁷. The coding is open-source and facilitated in the Giellalt⁸ infrastructure (cf. Moshagen et al. 2014) where the Erzya and Moksha languages share the progress in mutually applicable code that serves over 90 languages. At Giellalt, the code is reused wherever possible, i.e., analyzer code is flipped to make compatible generators, pedagogic descriptions are filtered to provide standard and descriptive models of the language for linguists, while normed filtering provides for spellers.

2.1 Materials

In the 2023, it was decided at the Language Bank of Finland that more Erzya and Moksha texts could be published on the Korp server if it involved less annotation. To this end, texts were selected from the Moksha-language journal «Mokša» and Erzya-language journals «Suran’ tolt» and «Sätko». The texts represent original and translated writings from the late 1920s to the beginning of the 2000s. The right to use these texts in searchable corpora had been secured in Saransk as the beginning of the new millennium. The new portion of corpora was called ERME-Paragraph Segmentation Low Annotation (ERME-PSLA).

The majority of the texts was scanned in the 2010s with funding from the Finno-Ugrian Society and the University of Helsinki. Optical recognition was then conducted from 2017, and in 2024 the recognition changed from ABBYY Finereader to Transkribus⁹, which would allow

greater access to the OCR engine and models for recognition.

As the size of this material became apparent, it was decided that the corpora might further be divided according to decade and, of course, language. The first portion of the corpora came from the last four years of the 1950s (1956–1959), hence the name ERME-PSLA 1950s.

2.1.1 Figures

The size of the corpora can be measured as twenty-two issues in each language, i.e., four issues from 1956 and six issues from each of the subsequent years. The yield of those four years is 831 pieces written in Moksha, and 707 pieces in Erzya. This equates to approximately 91,017 sentences of Erzya and 92,432 sentences of Moksha, which is 803,406 and 902,518 words in Erzya and Moksha, respectively. The number of words might be compared with the analogous figures for ERME version 2 (Rueter and Erina 2023-03-23) – Moksha 855,435 and Erzya 2,041,196.

For the four years of publications in twenty-two issues for each language, there were 163 authors with pieces in Erzya, and 185 with pieces in Moksha. If we count the number of authors with four or more pieces, we arrive at thirty-five writers in Erzya, and forty-three writers in Moksha.

The genres include poetry, story, short story, novel, essay, parody, critique, etc.

2.1.2 Errors

As is the case with most of the ERME materials, all of the ERME-PSLA corpora have been acquired through Optical Character Recognition (OCR). It goes without saying, some of the words in the corpora will be broken or unrecognized, which might lessen the value of the automatically annotated text. This shortcoming in the texts is one reason why the metadata includes page numbers and sentence enumeration; ORC errors might be located and corrected for future enhanced publications.

The probability of OCR errors exists. This can be seen in a simple comparison of ERME-PSLA 1950s figures against the number of word forms attested in the digitally transferred Erzya New Testament (NT) 2006 and Moksha New Testament 2016. In Table 1, the unique number of word forms is shown in parallel with word

⁷This site provides extensive information on Constraint Grammar https://edu.visl.dk/constraint_grammar.html

⁸<https://github.com/giellalt/>

⁹<https://app.transkribus.org/>

forms consisting of one to four letters in length followed by their ratio. The lower the ratio, the larger the number of short word forms in the language. Of the 7,648 unique Moksha word forms of four or less letters in length, 3,800 were not recognized by the Moksha analyzer. Likewise, of the 6,654 unique Erzya word forms of four or less letters in length, 3,082 were not recognized by the Erzya analyzer.

corpus	unique	unique 1–4	ratio
NT Erzya	18,439	874	21.10:1
1950s Erzya	99,287	6,654	14.92:1
NT Moksha	17,902	1,036	17.18:1
1950s Moksha	118,121	7,648	15.44:1

Table 1: OCR error statistics

2.2 Metadata

When ERME version two was published in the spring of 2023, it had twenty metadata features. Some of the attribute values were required by the Korp system, while others were introduced by the ERME documentation. The system required six attributes – a unique identifier, an ISO 639 language code, and timestamps – date from, date to, time from, time to. The remaining fourteen were optional.

The optional features were basically bibliographical. These consisted of <author>, <genre>, <number of pages>, <page range>, <publication place>, <publication name>, <publication year>, <publisher> and <bibliography> (an entire segment dedicated to bibliographic citation). After the information necessary for citation, came information which might help to explain linguistic characteristics of the text, namely, <corrector>, <electronic corrector> and <geographic origin of author> (this information is documented separately complete with coordinates). This was followed by two bits of statistics: <word count> and <character count>. All information was explicitly available in the corpora sources or it was readily derived. All texts were deemed original-language materials, and no extra information was needed for translation.

In the ERME-PSLA corpora, the twenty features shown above are elaborated upon. The <bibliography> segment is now written in Cyrillics, and it has a twin <bibliography iso9>, where all information in Cyrillics is con-

verted according to the International Library Convension in ISO-9¹⁰.

The challenge presented by journals is that individual pieces do not always explicitly indicate metadata important for locating a text in time and space. Thus, all important metadata, such as information on authors, titles, genres, and even correctors, electronic correctors are not readily available for all pieces. In fact, there are only three bits of information that can serve as consistent key identifiers: (1) the publication; (2) issue, and (3) page ranges. Information on authors, titles, genres and correctors are given whenever specifically stated in the publications or their sibling issues. Poetry and lyrics lacking titles are named using the first line of piece.

Journals are collections of pieces, such that a distinction must be made between the concepts of publication the container and the individual piece. In previous iterations of ERME corpora, this distinction has not been necessary, but the necessity for a feature <title> has already been encountered and facilitated in short stories by the Erzya author Pëtr Klûčagin 1997.

Journals are also collections of pieces representing both translations and original-language writing. This information is described in with the addition of features for <translator> and <translated>. While the former might readily be associated with a human actor, the latter requires explanation.

When approaching the concept of translation into Erzya and Moksha, we must all agree that anything written by Longfellow, Heine or Lenin has obviously be translated. Hence, the attribute <translated> in such pieces can easily be assigned the value <yes>.

A problem arises, however, when a piece by an Erzya or Moksha writer is indicated as having been translated from a different language by another native language writer, e.g., the play Ульнесь истямо тейтерь... ‘There was such a girl...’ by Ivan Antonov, 1957 was translated from the Russian by Aleksandr Šeglov. For this piece, the attributes translated and translator are given to attest certainty of translation, on the one hand, and knowledge of translator, on the other. Further investigation must be conducted before the value <no> can be assigned

¹⁰https://en.wikipedia.org/wiki/ISO_9

to the <translated> attribute.

Both journals present elements of folklore, such as riddles, lyrics and poetry. These require an additional human actor, the <collector>. Thus, materials collected by M. E. Evsev'ev might have subsequently been corrected by a <corrector>, while the language informant is provided with the <author> attribute.

Finally, it is important that versions and translations of a text be associated with each other for comparative studies, and that segments of larger texts be associated with each other. The concept of translation parallel is expressed in the feature <has parallel>. The <version> feature is initially used to indicate original version, but this may be altered in further development. Segment pointers are expressed with the attributes <continues from> and <continues to>. The attribute <comment> is reserved as a miscellaneous container, and the <content> attribute helps determine whether the piece will be contained in the corpora; the value <text> entails selection for use in the corpora.

2.3 Method

The low-annotation pipeline entails the same components found in the annotation pipelines for PaBiVUS version two and ERME version two. As these two sets of corpora that presuppose sentence-level annotation where the texts have been broken down to allow for sentence-level commenting, page numbers, indication of temporary change in genre and even language. ERME-PSLA makes the assumption that manual annotations are only made in the root element.

ERME-PSLA paragraph segmentation proceeds directly from the root element. Whereas previous work with ERME has presupposed sentence-level annotation, the PSLA pipeline utilizes features of the optical character recognition machines. These features include page and paragraph breaks. This has facilitated the numbering of pages and the recognition of paragraphs, which have simply been set off by double line breaks, and sentences have been recognized automatically with the help of sentence-final punctuation marks, such as full stop, question mark, exclamation mark and colon.

Although errors may have occurred in this part of the segmentation, the sentence-level

texts are now ready for annotation.

Annotation at the sentence level involves minimal additional human input. The recognized sentences are automatically annotated with unique identifiers and page numbers for reference to the source texts.

3 FST models

The finite-state description of the two Mordvin literary languages started nearly fifteen years apart. Work on the Erzya analyzer was begun in the late 1990s, whereas work on Moksha was part of the «Creation of Morphological Parsers for Minority Finno-Ugrian Languages» project funded by the Kone Foundation «Language Programme», 2013–2014 (see Rueter 2014; Rueter et al. 2020). Despite attempts to make the descriptions as parallel as possible, there are still statistical differences in their coverage and accuracy afforded by the Erzya and Moksha analyzers.

Both analyzers are used in the analysis of corpora texts on different platforms and projects¹¹, system-wide spell checkers¹². Their development is part of a collaboration with the «Experimental Treebanking for the Minority Moksha Language and Finite-State Descriptions» and «Experimental Treebanking for the Minority Erzya Language and Finite-State Descriptions» projects. The analyzers have been used in the annotation of treebanks on the Korp servers of both the Language Bank of Finland¹³ in Helsinki, Finland, and Giellatekno/Divvun¹⁴ in Tromsø, Norway.

Evaluation and enhancement of these analyzers is important for improving community language facilitation. Brief statistical descriptions of the Erzya and Moksha analyzers have been given online through the Language Bank of Finland¹⁵. These provide annotational statistics based on materials of the upcoming version of PaBiVUS.

Below we offer an enhanced evaluation of the PaBiVUS annotation for comparison with figures for the outcome of the ERME-PSLA 1950s corpora. While notions of size are imme-

¹¹<https://universaldependencies.org/>

¹²<https://divvun.org/>

¹³kielipankki.fi/korp/

¹⁴gtweb.uit.no/u_korp/

¹⁵https://www.kielipankki.fi/tools/giellalt_language_models/

diately obvious from the statistics, one must also bear in mind the cleanliness of the corpora. As noted above, in materials, ERME-PSLA 1950s materials are likely to have OCR errors, this might be observed in the a comparison of unique forms to unique misses in the modern New Testament texts and the ERME-PSLA materials.

3.1 The Erzya FST

The Erzya finite-state model has a relatively large lexical base of 176,832 lemma-stem pairs, and 1,370 continuation lexica. Together, these provide for a variety of inflectional patterns in verbs and words in the nominal categories.

Recently, a simple analysis of the Erzya model was published on the Language Bank of Finland website¹⁶. It provided statistics based on testing with the upcoming version two of Parallel Biblical Verses for Uralic Studies (PabiVUS). Modified statistics on the analyzer are given here, in which a distinction is drawn between results for the 1821–1827 version of the New Testament and the 2006 version. These results are aligned with results for the 1950s portion of ERME-PSLA (ERME-Paragraph Segmentation Low Annotation).

The three corpora can be distinguished in many ways. While notions of size are immediately obvious from the statistics, one must also bear in mind the cleanliness of the three. As noted above, in materials, ERME-PSLA 1950s materials are likely to have OCR errors, this might be observed in a comparison of unique forms to unique misses in the New Testament texts from 2006 and the ERME-PSLA materials. The New Testament (NT) figures for unique forms over unique misses renders a ratio of 48.53:1, whereas the correlating figures for ERME-PSLA are 5.13:1. The ratio for NT 1821–1827 is 2.54:1, which may be attributed to high variation in spelling in the older version of the Erzya New Testament.

The sizes of the three corpora illustrate a difference between two versions of the New Testament in Erzya and the text content of twenty-two issues of «Suran’ tolt», which is five and a half times the size of the New Testament. Their automated annotation is illustrated in

¹⁶[urlhttps://www.kielipankki.fi/tools/giellalt_language_models/erzya/](https://www.kielipankki.fi/tools/giellalt_language_models/erzya/)

Table 2.

corpus	NT 1821–1827	NT 2006	ERME-PSLA
words total	128,245	140,942	803,406
characters total	711,716	857,812	5,003,429
unique forms	22,569	18,439	99,287
unique misses	8,899	380	19,342
lines before hapax	1943	58	2,917
ambiguous PoS	8,943	449	30,352
unique amb. PoS	8,899	399	26,890
ambiguous dep.	29,120	9,999	91,625
unique amb. dep.	9,151	692	20,269

Table 2: Erzya annotation statistics

The Erzya analyzer and disambiguation do a better job than the dependency parser and subsequent conversion scripts from Constraint Grammar to Universal Dependencies presentation.

In the Erzya New Testament, 2006, the ratio for unique forms to unique forms with ambiguous parts of speech is 46.21:1, while the ratio for unique forms to unique forms with ambiguous or unrecognized dependencies is 26.65:1. The correlating figures for NT 1821–1827 and ERME-PSLA 1950s are: 2.54:1, 2.47:1 and 3.69:1, 4.90:1, respectively. Words total over ambiguous dependencies were 14.1:1 (NT 2006), 4.4:1 (NT 1821–1827), and 8.77:1 (ERME-PSLA 1950s).

3.2 The Moksha FST

The Moksha finite-state analyzer is younger than the Erzya model. It has been under construction since 2012, as part of the «Creation of Morphological Parsers for Minority Finno-Ugrian Languages» project, 2013–2014. Since then the analyzer has grown with approximately 189,476 lemma-stem pairs in the lexicon and 852 continuation lexica for facilitating complex morphology in the nominal and verbal categories.

A simple evaluation of the Moksha analyzer was recently published on the Language Bank of Finland website¹⁷, in which test results for the model’s performance on the New Testament texts in PaBiVUS version two were described.

Below, we provide a modified version of those statistics, where we draw only on the texts published in 2016, and compare them to analogous results for the ERME-PSLA 1950s corpus.

¹⁷[urlhttps://www.kielipankki.fi/tools/giellalt_language_models/moksha/](https://www.kielipankki.fi/tools/giellalt_language_models/moksha/)

The two corpora can be distinguished by cleanliness, size and accuracy. As noted above, in materials, ERME-PSLA 1950s materials are likely to have OCR errors, this might be observed in the a comparison of unique forms to unique misses in the New Testament texts from 2016 and the ERME-PSLA materials. The ERME-PSLA 1950s corpus for Moksha is over six and a half times the size of the New Testament materials from 2016. In fact, it is slightly larger than the ERME version two Moksha corpus from 2023. The NT figures for unique forms over unique misses renders a ratio of 21.99:1, whereas the correlating figures for ERME-PSLA are 3.56:1.

corpus	NT 2016	ERME-PSLA 1950
words total	136,718	902,518
characters total	793,393	5,559,553
unique forms	17,902	118,121
unique misses	814	33,195
lines before hapax	129	4,436
ambiguous PoS	1,075	46,361
unique amb. PoS	856	33,195
ambiguous dep.	10,853	117,888
unique amb. dep.	1,238	34,775

Table 3: Moksha annotation statistics

The Moksha analyzer and disambiguation perform better than subsequent function and dependency parsing followed by conversion scripts.

In the Moksha New Testament 2016, the ratio for unique forms to unique forms with ambiguous parts of speech is 20.91:1, while the ratio for unique forms to unique forms with ambiguous or unrecognized dependencies is 14.46:1. The correlating figures for ERME-PSLA 1950s are: 3.56:1 and 3.4:1. Total word forms divided by ambiguous dependencies give us the figures 12.6:1 (NT 2016), and 7.66:1 (ERME-PSLA 1950s).

3.3 FST retrospect

A Korp Vertical structure (VRT) validator under continuous development at the Language Bank of Finland is used to determine validity of the XML files, and additional scripts are run to assess the number of word forms lacking recognition, word forms with ambiguous part-of-speech readings and word forms with ambiguous dependencies. These figures have then been used in the evaluation of the individual analyzers, disambiguators, annotation

for function and dependency, and conversion scripts.

It has been noted that over half of the unique missing word forms occur but once. The following assumptions have been made: High frequency of missing word forms would indicate need for lexical inspection and enhancement. High frequency of part-of-speech ambiguity may point to homonymy. And high frequency of dependency ambiguity may actually point to shortcomings in the CG-to-UD conversion scripts.

The annotation stops when the lexica or morphology are lacking or blocked. No non-described annotations are make.

4 Discussion and Conclusions

The statistics for the Erzya and Moksha analyzers were drafted before the present article was begun. During the course of writing the article, a number of problems were noted with regard to the CG-to-UD-format transfer and the quality of the OCR.

It was decided that enhancement and evaluation ought to be included in the release protocol of each new korpus for either of the languages.

In a brief inspection of the ERME-PSLA OCR quality, a lists of Erzya and Moksha words four letters or less in length were extracted from New Testament texts. These words had been human inspected and digitally transferred, which would guarantee their quality. In the future, however, is was decided that separate lists of this kind should be drafted and curated to be used in inspection of text validity.

Finally, more work should be allotted to the development of function and dependency parsing, as these along with conversion strategies would immensely improve the usability of ERME corpora.

Ethics statement

When dealing with an endangered language it is important to make sure that the research also contributes to the language community. This is the reason why we open-source our FST analyzers and components. We also work with data licensed to us by speakers of Moksha and Erzya with the intention of contributing to morpho-syntactic descriptions, tools and meta-data practices for the languages. This means

that we are not conducting our research with no regard to the language community.

Acknowledgments

This research is supported by FIN-CLARIAH and Academy of Finland (grant 358720 Kielivarojen ja kieliteknologian tutkimusinfrastruktuuri).

References

- Nina Agafonova and Ivan N. Râbov. 2021. Ulânovskoj oblasten novomalyklinskoj rajonon Ėrzân velen kortavkstnèšè azorkščin nevtycâ suffikstnén baška ênksost. In Niko Partanen, Mika Hämäläinen and Khalid Alnajjar, editors, *Multilingual Facilitation*, page 263–274. University of Helsinki Library. This book has been authored for Jack Rueter in honor of his 60th birthday.
- Khalid Alnajjar, Mika Hämäläinen, Niko Tapio Partanen, and Jack Rueter. 2022. [Using graph-based methods to augment online dictionaries of endangered languages](#). In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 139–148, Dublin, Ireland. Association for Computational Linguistics.
- Ivan Antonov. 1957. Ul’nes’ istâmo tejter’... Suran’ tolt, (№ 1):3–32. Genre = p’esa, translated from Russian = Šeglov, Aleksander.
- Finno-Ugrian-Society. [Suomalais-Ugrilaisen Seuran kenttäyökorpus](#).
- Herr Conon von der Gabelentz. 1838–1839. Versuch einer mordwinischen grammatik. *Zeitschrift für die Kunde des Morgenlandes.*, II(2–3):235–284, 383–419.
- Mika Hämäläinen, Khalid Alnajjar, Jack Rueter, Miika Lehtinen, and Niko Partanen. 2021. [An online tool developed for post-editing the new skolt sami dictionary](#). In *Electronic lexicography in the 21st century (eLex 2021)*. Proceedings of the eLex 2021 conference, Electronic lexicography in the 21st century (eLex 2021). Proceedings of the eLex 2021 conference, pages 653–664, Czech Republic. Lexical Computing CZ s.r.o. Electronic lexicography in the 21st century (eLex 2021) ; Conference date: 05-07-2021 Through 07-07-2021.
- Helsingin yliopisto, FIN-CLARIN, Jack Rueter, and Erik Axelson. 2020-06-07. [Raamatun jakeita uralilaisille kielille, rinnakkaiskorpus, Korp](#).
- Nadežda Kabaeva. 2021. Fonetičeskie osobennosti govora sela adaševo ũgo-vostočnogo dialekta mokšanskogo ũzyka. In Ksenia Shagal with Heini Arjava, editor, *Mordvin Languages in the Field*, volume 10 of *Uralica Helsingiensia*, page 171–186. University of Helsinki Finno-Ugric Language Section and the Finno-Ugrian Society.
- Pëtr Andreevič Ključagin. 1997. Garec’. In *Cëkan’ka*, page 58–62. Mordovskoj knižnoj izdatel’stvas’, Saransk.
- 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 *Systems and Frameworks for Computational Morphology: Third International Workshop, SFCM 2013, Berlin, Germany, September 6, 2013 Proceedings 3*, pages 53–71. Springer.
- Sjur Moshagen, Jack Rueter, Tommi Pirinen, Trond Trosterud, and Francis M Tyers. 2014. Open-source infrastructures for collaborative work on under-resourced languages. *Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era*, pages 71–77.
- Pavel Ornatov. 1838. *Mordovskaâ grammatika*. Moskva: V” Sinodal’noj tipogrffii. Sostavlennaâ narečij mordvy mokši Tambovskoj seminarii professorom”, magistrom”, Pavlom” Ornatomvym”.
- Jack Rueter. 2013. [The erzya language, where is it spoken?](#) *Études finno-ougriennes*, 45.
- Jack Rueter. 2014. [The livonian-estonian-latvian dictionary as a threshold to the era of language technological applications](#). *Journal of Estonian and Finno-Ugric Linguistics*, 5(1):251–259. ESUKA – JEFUL 2013, 5–1: 253–261.
- Jack Rueter. 2020. Корпус национальных мордовских языков: принципы разработки и перспективы функционирования/ действия. In *Финно-угорские народы в контексте формирования общероссийской гражданской идентичности и меняющейся окружающей среды*, pages 118–127, Russia. Издательский центр Историко-социологического института. Conference date: 08-10-2020 Through 09-10-2020.
- Jack Rueter. 2022. [Shallow-transfer problems in erzya-moksha conjunctive-preterite2 syncretism](#). *The Journal of Brief Ideas*.
- Jack Rueter. 2024. [On searchable mordvin corpora at the language bank of finland, emerald](#). *Journal of Data Mining and Digital Humanities*.
- Jack Rueter and Olga Erina. 2023-03-23. [ERME Ersän ja mokšan laajennettu korpus versio 2, Korp](#).
- Jack Rueter and Mika Hämäläinen. 2020. Prerequisites For Shallow-Transfer Machine Translation Of Mordvin Languages: Language Documentation With A Purpose, pages 18–29. Ижевск:

- Институт компьютерных исследований, Russian Federation.
- Jack Rueter, Mika Härmäläinen, and Niko Partanen. 2020. [Open-source morphology for endangered mordvinic languages](#). In Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS), pages 94–100, Online. Association for Computational Linguistics.
- Jack Rueter and Nadezhda Kabaeva. 2024. On quantification and the ablative in erzya and moksha. In Hajner Réka, Kubínyi Kata, Dóra Pődör, and Tamm Anne, editors, European partitives in comparison. L’Harmattan Kiadó, Károli Gáspár Református Egyetem, Budapest, Magyarország. Budapest, Magyarország : L’Harmattan Kiadó, Károli Gáspár Református Egyetem (2024).
- Jack Rueter, Niko Partanen, Mika Härmäläinen, and Trond Trosterud. 2021. [Overview of open-source morphology development for the Komi-Zyrian language: Past and future](#). In Proceedings of the Seventh International Workshop on Computational Linguistics of Uralic Languages, pages 29–39, Syktyvkar, Russia (Online). Association for Computational Linguistics.
- Jack Michael Rueter. 2023. [Moksha Mordvin](#), 2nd edition edition, Routledge Language Family Series. Routledge, International.
- Heno Sarv. 2002. Indigenous Europeans East of Moscow, Population and Migration Patterns of the Largest Finno - Ugrian Peoples in Russia from the 18th to the 19th Centuries., volume 17. Geographicae Universitatis Tartuensis. Dissertation.
- Miikka Silfverberg and Jack Rueter. 2015. [Can morphological analyzers improve the quality of optical character recognition?](#) In First International Workshop on Computational Linguistics for Uralic Languages, volume 2 of Septentrio Conference Series, pages 45–56, Norway. Septentrio Academic Publishing. Proceeding volume: 2; International Workshop on Computational Linguistics for Uralic Languages ; Conference date: 16-01-2015 Through 16-01-2016.
- P. M. Suihkonen. 2003. [Metadata descriptions for combining information on multimodal data located at the university of helsinki language corpus server](#). In S. Darányi, editor, Proceedings of the Higher Order Morphologies’ Observer 2003 Conference on Information Society: Cultural Heritage and Folklore Text Analysis Budapest University of Technology and Economics. L’Harmattan Kiadó, Károli Gáspár Református Egyetem, Budapest, Hungary.
- Nikolaes Witsen. 1705. Noord en Oost Tartarye, Ofte Bondig Ontwerp Van eenig dier Landen en Volken Welke voormaels bekend zijn geweest. Benneffens verscheide tot noch toe onbekende, en meest nooit voorheen beschreeve Tartersche en Nabuurige Gewesten, Landstreeken, Steden, Rivieren, en Plaetzen, in de Noorder en Oosterlykste Gedeelten Van Asia En Europa Verdeelt in twee Stukken, Met der zelviger Landkaerten: mitsgaders, onderscheide Afbeeldingen van Steden, Drachten, enz. Zedert naeuwkeurig onderzoek van veele Jaren, door eigen onderzondinge ontworpen, beschreven, geteekent, en in t licht gegeven, Door Nicolaes Witsen. Jan van der Deyster, Amsterdam By François Halma, Boekverkooper op de Nieuwendyk.
- Daniel Zeman and et al. 2024. [Universal dependencies 2.14](#). Universal Dependencies 2.14, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Towards the speech recognition for Livonian

Valts Ernštreits
University of Latvia Livonian Institute
valts.ernstreits@lu.lv

Abstract

This article outlines the path toward the development of speech synthesis and speech recognition technologies for Livonian, a critically endangered Uralic language with around 20 contemporary fluent speakers. It presents the rationale behind the creation of these technologies and introduces the hypotheses and planned approaches to achieve this goal. The article discusses the four-stage approach of leveraging existing data and multiplying voice data through speech synthesis and voice cloning to generate the necessary data for building and training speech recognition for Livonian.

1 Introduction

In October 2024 University of Latvia Livonian Institute has launched a new project aimed at developing speech synthesis for Livonian. This project has a broader goal of laying the groundwork for future speech recognition technology for the language. This article outlines the rationale behind the creation of speech synthesis and speech recognition technology for Livonian, and presents the hypotheses and approaches intended to achieve this goal.

Recently, as technological advancements progress, significant efforts have been made to develop speech-recognition technologies for underrepresented and critically under-resourced languages, including Uralic languages (e.g. [Partanen et al 2020](#)). This work is also being undertaken by two key partners of the UL Livonian Institute in this project – AILAB¹ and the University of Tartu NLP group.

It is worth noting that the current article focuses on the broader context, data production and

handling approaches, and hypothetical methods for acquiring data and achieving the desired quality. The article does not cover the actual development of the technologies (speech synthesis and recognition) themselves, which will be conducted later in the project by the UL Livonian Institute's partners.

2 Broader context

The world's linguistic and cultural diversity is in urgent danger. To safeguard endangered languages and intangible cultural heritage, major policy documents and programmes have been introduced, e.g., the UNESCO Convention for the Safeguarding of the Intangible Cultural Heritage (2003), the UN International Decade of Indigenous Languages (2022–2032; further – IDIL) and others. A large part of this work in the digital era has been finding ways to narrow the digital gap between languages and cultures with extensive resources and data and the – mostly endangered – ones lacking resources and especially the data needed for creating digital technologies.

With every new technology, this gap only grows, and closing it is crucial for keeping endangered languages competitive and vital. Thus, in 2024, the Ad-Hoc Group on Digital Equality and Domains was created by the Global Task Force of the IDIL to tackle these issues.

Another crucial aspect of this work is ensuring access to existing data and archives, especially materials only in analogue format. Such archives may be a great resource for academia and communities themselves, e.g., for language acquisition or expanding the language environment, but not all analogue materials can be easily digitised and utilised, as some are scattered

¹ Artificial Intelligence Laboratory at the University of Latvia Institute of Mathematics and Computer Science.

or require considerable skilled human resources to be made digitally accessible.

As endangered languages enter the digital world, this access only becomes more important, as these populations are increasingly exposed to majority populations and subject to free movement and the rapid expansion of the information space – also through recently developed AI-based tools for generating language content (Wang 2024; Houboud, Ortega 2023, 287, 298).

Emerging technologies including AI offer new challenges, but also new possibilities, for tackling these issues. Due to a lack of sufficient data and other aspects specific to critically under-resourced languages and cultures, rule-based and AI technologies cannot and should not be used in the same way as when “big data” is available. Specific methods must be invented to meet the specific needs of languages and cultures with extremely limited resources. Obtaining knowledge and developing such approaches, methods, and tools is a speciality and one of the cornerstones of the research and resources developed by the University of Latvia Livonian Institute for Livonian language and culture.

3 Creating digital resources for Livonian

Like many other Uralic languages, Livonian with around 20 contemporary fluent speakers (Druviete, Kļava 2018) is highly constrained in terms of its domains of use and available resources. While many digital tools developed in the 21st century can potentially offer support, they first require large amounts of digital data. This presents a significant challenge not only for Livonian but also for most Uralic languages and the majority of the world’s languages.

Since its founding in 2018, the University of Latvia’s Livonian Institute has been developing a range of digital resources and methods for Livonian, focusing on data acquisition, efficient data use, resource-saving workflows, and improved accessibility.

What began as a digital dictionary has now expanded into a cluster of interconnected databases, accessible through Livonian.tech (LT). This includes an up-to-date lexical database with translations of Livonian lemmas and example

sentences in Estonian, Latvian, and English; a morphology database; a partially annotated corpus of written Livonian with translations; and a geospatially linked place-name database. New databases focused on documentation sources and individuals are currently in development.

To meet the specific data extraction and management needs of extremely low-resource languages, tailored data processing approaches have been implemented in the construction of this cluster. In collaboration with partners, the institute has also explored opportunities to introduce various digital technologies for Livonian.

However, Livonian continues to face significant obstacles in obtaining new data, particularly due to the extremely small number of contemporary speakers and the limited domains in which the language is actively used. The language is also nearing a “glass ceiling” of available written data², much of which has already been digitized.

4 Data extraction or data synthesis?

At the core of every technological advancement in the digital field lies the availability of sufficient data. For languages like Livonian, which have very few contemporary speakers, extremely limited domains, and is entirely overshadowed by dominant language, the natural production of new data is an enormous challenge. Therefore, to obtain the additional data necessary for developing language technologies, one must either rely on extracting data from earlier documentation (if available) or turn to synthetic data produced through the use of technology. However, both approaches come with their own flaws and risks.

In terms of using earlier documentation, written records are generally more accessible. Converting analog documentation into digitally usable form requires only basic technologies (such as OCR for printed materials) or staff with basic language skills, knowledge of orthography, and an understanding of transcription methods used in the original documentation. However, the amount of pre-existing written documentation is usually quite limited, as is the case with Livonian.

Transliterating spoken language poses a much greater challenge. Not only is it significantly more time-consuming, but it also requires full

² The Livonian corpus representing written Livonian sources currently consists of ca 500 000 tokens; its full extent can be estimated at ca 750 000 tokens.

proficiency in the language to accurately capture what is being spoken. Consequently, manual transliteration can deplete the limited human resources proficient in the language for an indefinite period of time.

When it comes to synthesizing language data with the help of technology, a major risk is that the data produced may be of poor quality, due to the lack of sufficient data for building or training the necessary models. Poor-quality synthetic data may even pollute the digital environment of the language (Ernštreits, Fišel et al., 2022, p. 24; Trosterud, 2009), particularly given the ability of many technologies to generate large amounts of data in a short time—amounts that could far exceed the natural language data available for languages like Livonian.

A relevant example is machine translation, which is often one of the first technologies mentioned in discussions about endangered and low-resource languages. As with human translation, the quality of machine translation depends on its ability to generate accurate output in the target language, which, in turn, relies on the availability of training data and the methods used. Consequently, it is easier—and more beneficial—to develop machine translation from the endangered language to provide access to texts written in it, rather than to create machine translation into the endangered language that can synthesize high-quality data for further use.

Experiments with machine translation have been conducted also for Livonian (e.g., Rikters, Tomingas et al., 2022; Ernštreits, Fišel et al., 2022), but the evaluation of these results has shown that even with the use of all available aligned and monolingual corpora and databases, the data was still insufficient to build a high-quality machine translation model capable of synthesizing quality Livonian data.

However, these two approaches—data extraction and data synthesis—can be combined to generate new language data and significantly enhance the data production capabilities for languages like Livonian.

5 Choosing speech synthesis and recognition

Analyzing technologies that can be most beneficial for Livonian from the perspective of obtaining data and best serving the community, it has been concluded that speech recognition may actually be

the most advantageous (Ernštreits, Fišel et al. 2022, 31) at this stage.

As mentioned, Livonian corpus building is nearing the point where no major written sources remain to be added, but there are extensive Livonian speech recordings in several archives and private collections. The contents of these recordings could significantly expand the Livonian language corpus, but to access information beyond phonetic features (to get to vocabulary, morphology, syntax, etc.), these recordings need to be transcribed.

Currently, the only way to achieve this is through the involvement of fully proficient speakers, which is neither viable nor effective. Speech recognition, on the other hand, would significantly accelerate this process. Moreover, it would be a powerful tool for rapidly expanding the Livonian information space (and corpus) by enabling speakers to record and transcribe large volumes of information in Livonian. However, speech recognition is quite challenging to develop and requires a significant amount of data for training.

Conversely, speech synthesis—a technology that is much easier to develop compared to speech recognition—is another technology urgently needed by the Livonian community. The need for speech synthesis arises from the fact that a natural Livonian-speaking environment no longer exists, and most of the data available to the general public and those learning the language is in written form. This means that Livonian speakers and learners primarily read Livonian rather than hear it.

To address this issue, from 2022 to 2024, the Livonian lexicographic database (LT) has been supplemented with audio recordings of lemmas and example sentences spoken by contemporary Livonian speakers, giving users the opportunity to hear Livonian. Thus, the collection of audio data necessary for developing speech technologies has already begun.

Successful voice synthesis would give the general public, and especially the Livonian community, the opportunity to hear all digitized Livonian texts immediately, although having golden-standard level voice recordings should clearly be preferable.

Speech synthesis would also greatly expand the Livonian audio information space, offering the community opportunities to create content or even develop audio and video media, thereby allowing

Livonian to be encountered in new language domains. Additionally, speech synthesis could provide the conditions and data necessary for creating and training speech recognition technology.

6 How is it planned?

The approach proposed by the researchers of the UL Livonian Institute for developing speech recognition for Livonian involves four stages and is based on using existing data from the Livonian.tech database cluster (LT) and both aligned and non-aligned speech recordings from various informants found in archives.

In the first stage, initial aligned datasets are created using pre-existing audio data from the Livonian.tech databases and the 'gold standard' natural speech data from contemporary Livonian speakers reading texts from written corpora. Subsequently, based on this data, speech synthesis is created.

In the second stage, voice cloning takes place, creating additional synthetic voices by using the natural speech data of previously recorded informants from archives. Initially, the focus is on informants who have been recorded more extensively and whose speech has also been transcribed and published as language samples.

In the third stage, Livonian written data is used to generate a synthetic voice corpus using all available voices (both contemporary voices and those from archive recordings), in order to obtain voice data from all existing written sources. This process effectively multiplies the corresponding data by the number of synthetic voices. In the fourth stage, both natural and synthesized speech data are used to train the speech recognition model.

In the final stage, voice data manipulations are planned, including techniques such as voice merging, adding disturbances based on recording quality and phonetic peculiarities, and using non-aligned data from both living and deceased speakers for training. This point is particularly important considering the varying time periods and quality of many Livonian recordings.

As the current project is relatively short (a little over a year), only the first stage will be fully completed, which involves creating speech synthesis (text to speech), along with some data multiplication and initial isolated speech recognition (speech to text) experiments. However, this will lay the groundwork for a continuation

project, which will focus more specifically on speech recognition.

The tasks to be completed within this stage are:

1. Restructuring the existing aligned (text and voice) corpus and expanding it by using contemporary Livonian speakers;
2. Creating a standard Livonian pronunciation guide to be used for referencing and assessing data quality;
3. Creating speech synthesis using the highest quality "gold standard" data from living speakers participating in the creation of the aligned corpus (text and voice);
4. Expanding the aligned corpus by synthesizing available written text data using the voices of living speakers, thus multiplying the audio data;
5. Assessing the quality and making necessary corrections to improve the synthesized data;
6. Cloning voices of both contemporary speakers and those recorded earlier (for both "gold standard" example data and synthesized data) and building a speech synthesizer using altered "gold standard" data;
7. Conducting a data assessment of archive speaker data and making necessary updates;
8. Performing tests on speech recognition using all created and available datasets and reviewing preliminary findings;
9. Integrating speech synthesis into the Livonian.tech database cluster and making it available for other applications.

The project is expected to conclude with the first preliminary results of the experiments on the creation of speech recognition in February 2026.

7 Final notes

All the necessary preconditions exist for the proposed approach to be successful, at least to some extent. Even a partial acceleration of the transcription process would offer significant benefits for extracting data from audio sources. Gradually expanding the capabilities of Livonian in the sound environment would undoubtedly strengthen the language and make it more competitive in the long run.

Moreover, the methods and approaches developed during this research could serve as a foundation for other research teams seeking opportunities to apply speech technologies to extremely low-resource languages.

If methods for providing high-quality speech synthesis and recognition to languages with extremely limited data are discovered, this would be beneficial to all endangered languages, particularly those lacking domains for everyday language use (speech synthesis) or without a written tradition, where the language is primarily oral (speech recognition). Furthermore, this would provide the varied and abundant data essential for developing other technologies, such as machine translation, chatbots, caption generation, and other solutions powered by advancing AI technologies.

Acknowledgments

This study has been performed as part of the project “Improving access to a critically under-resourced language: AI-based approaches for producing and obtaining Livonian content” financed by the Recovery and Resilience Facility / NextGeneration EU (LU-BA-PA-2024/1-0056).

References

- Druviete, Ina & Gunta Kļava. 2018. *The role of Livonian in Latvia from a sociolinguistic perspective*. *Eesti Ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 9(2), 129–146. <https://doi.org/10.12697/jeful.2018.9.2.06>
- Ernštreits, Valts, Fišel, Mark, Riktors, Matīss, Tomingas, Marili & Tuuli Tuisk. 2022. Language resources and tools for Livonian. *Eesti Ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 13(1), 13–36. <https://doi.org/10.12697/jeful.2022.13.1.01>
- Haboud, Marleen & Fernando Ortega. 2023. *Linguistic diversity endangered: the Waotodedo language and the effects of intense contact*. In: Eda Derhemi and Christopher Moseley (eds.). *Endangered Languages in the 21st Century*. Routledge: London and New York.
- LT = Ernštreits, Valts (ed. in chief), Vāvere, Signis, Viitso, Tiit-Rein, Damberg, Pētõr, Kurpniece, Milda, Kļava, Gunta, Balodis, Uldis, Tuisk, Tuuli, Kūla, Gita, Tomingas, Marili, Soosaar Sven-Erik, Sedláčková, Anna & Jurgenovskis, Toms. 2024. *Livonian language and culture resource platform “Livonian.tech”*. Riga: University of Latvia Livonian Institute. <https://livonian.tech/>
- Riktors, Matīss, Tomingas, Marili, Tuisk, Tuuli, Ernštreits, Valts & Mark Fishel. 2022. *Machine Translation for Livonian: Catering to 20 Speakers*. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Volume 2: Short Papers, pages 508–514 May 22-27. Available: <https://aclanthology.org/2022.acl-short.55.pdf>
- Trosterud, Trond. 2009. *Developing Prototypes for Machine Translation between Two Sámi Languages*. *Proceedings of the 13th Annual Conference of the European Association of Machine Translation*, EAMT09. Allschwil: European Association for Machine Translation.
- Wang, Luyi. 2024. *Artificial intelligence's role in the realm of endangered languages: Documentation and teaching*. *Applied and Computational Engineering, Proceedings of the 4th International Conference on Signal Processing and Machine Learning*, March 2024, 48(1):123-129. DOI: [10.54254/2755-2721/48/20241249](https://doi.org/10.54254/2755-2721/48/20241249)
- Partanen, Niko, Hämäläinen, Mika & Klooster Tiina. 2020. *Speech Recognition for Endangered and Extinct Samoyedic languages*. *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*. Association for Computational Linguistics. Available: <https://aclanthology.org/2020.paclic-1.60>

Using Large Language Models to Transliterate Endangered Uralic Languages

Niko Partanen

Department of Finnish, Finno-Ugrian and Scandinavian Studies
University of Helsinki
niko.partanen@helsinki.fi

Abstract

This study investigates whether the Large Language Models are able to transliterate and normalize endangered Uralic languages, specifically when they have been written in early 20th century Latin script based transcription systems. We test commercially available closed source systems where there is no reason to expect that the models would be particularly adjusted to this task or these languages. The output of the transliteration in all experiments is contemporary Cyrillic orthography. We conclude that some of the newer LLMs, especially Claude 3.5 Sonnet, are able to produce high quality transliterations even in the smaller languages in our test set, both in zero-shot scenarios and with a prompt that contains an example of the desired output. We assume that the good result is connected to the large presence of materials in these languages online, which the LLM has learned to represent.

1 Introduction

There is a long tradition of publishing transcribed texts on various Uralic languages using Finno-Ugric Transcription system, also known as Uralic Phonetic Alphabet. Recently also IPA has become more commonly used, but various transcription systems and their interpretations are widely used. Thousands of pages of linguistic materials are printed or archived that use these transcription systems. When the materials are being digitized, we need to address several questions that relate to how the texts are ideally treated for contemporary use. This includes both the use within the scientific community and the language communities, which may have different needs. Yet there is a shared starting point that the older transcription systems are not ideal for all modern applications.

The use of orthographies in contemporary language documentation has been seen preferable to phonemic or detailed phonetic transcription (Gerstenberger et al., 2016, 32), and also in this study

the primary question is how to transform the transcriptions into currently used orthographies. The use of some other systematic transcription system could be likewise suggested, and it seems reasonable to expect that our approach could be extended to other comparable transformations. The target chosen here, contemporary orthography, comes with its own problems that some other choices would have avoided. There is inevitably more subjectivity in the details of the result than what i.e. phonemic representation would entail.

Transforming a transcription into orthography moves along the blurry boundary of transliteration and normalization. The task is not only a simple transliteration. Often the exact language variety in the transcription matches only partially with the contemporary written standard. This makes the desired normalization vaguely defined, as we would like to keep some amount of dialectal features that are present in the text, but not necessarily everything. Especially the dialectal morphology and lexicon should remain recognizable, if possible.

We define the desired output as something comparable to what contemporary language documentation material would reasonably be expected to look like, when collected from the same dialect today using the current orthography. Another expectation would be that the result would be close enough to the contemporary orthography that the modern language community members can read it, and the modern language technology tools are able to process it, possibly with minor modifications. For both of these purposes the methods to mark dialectal features that are already attested and used in the literary language would be ideal choices.

Ideally different versions would be stored together, possibly with alignment at word or sentence level. Thereby the issue of information loss in the orthographic version is not a problem for the entire dataset.

2 Related work

It seems there has not been extensive research on automatic transliteration of endangered languages within the language documentation or natural language processing communities. The task is fairly relevant in this context. It is common in many areas of the world that the contemporary language community members cannot read the older language documentation materials produced in scientific transcription systems. For example, [Siegl and Rießler \(2015, 211-212\)](#) point out that for the Enets language this extends to most of the texts published during the Soviet period. At the same time the old transcriptions are often complicated to use even for specialists. Combining diacritics can be hard to type systematically, and variations of the transcription system used in different publications make comparable searches challenging.

[Bradley \(2017\)](#) have worked extensively with transliteration of different languages spoken in Russia. Also [Bradley and Skribnik \(2021\)](#) discuss the problems specific to the Mansi orthographies, and provide a rule-based toolset for transliterating across different writing conventions. They highlight some of the problems in their approach, mainly that there is ambiguity in different writing systems that cannot be captured by rule-based models. For example, vowel length may not be marked at all which requires the model to have knowledge of a wider context outside the source text. Additional issue is that they have transliterated between different systems that have roughly the same phonemic accuracy, whereas the text we are working with include very extensive diacritics at varying levels of details. Very relevantly, [Bradley and Blokland \(2023\)](#) also discuss in detail the situation of Unicode development and use in the Uralic context.

With larger Uralic languages, earlier work has been done on transliteration of dialectal Finnish texts. [Partanen et al. \(2019\)](#) showed that detailed dialectal transcriptions can be accurately converted to modern literary Finnish using neural networks. Still in the Finland’s context, [Hämäläinen et al. \(2020\)](#) extended this work to the Swedish spoken in Finland. These studies connect closely to the current work as the goal has been to process scientific transcriptions. [Partanen et al. \(2022\)](#) describe in detail the wider workflow into which transliterations connects to. This includes OCR or HTR, transliteration and also audio processing.

Grapheme-to-phoneme conversion can be seen as a sister task to what is undertaken here, as the aim is to convert orthographic text into phonemic or phonetic realization. [Suvarna et al. \(2024\)](#) performed a complex evaluation of various related tasks, including grapheme-to-phoneme conversion, syllable counting, and rhyme word generation. In their test no single model outperformed in all tasks. [Fetrat Qharabagh et al. \(2024\)](#) tested Persian grapheme-to-phoneme conversion with LLMs and reported better performance than the traditional methods.

3 Data and Experiment Design

Our dataset contains digitized transcriptions in Unicode characters and their paragraph level correspondences in the contemporary Cyrillic orthographies. The languages included are Komi-Zyrian, Udmurt, Northern Mansi and Kildin Saami. The data comes mainly from the publications of the Finno-Ugrian Society and represent mostly early 20th century fieldwork written in Finno-Ugric Transcription. In the case of Kildin Saami the selected text is a religious translation, but we believe it was still suitable for this experiment. These materials represent a very prominent and potentially underused data source in the Uralic studies.

The linguistic sources include: Udmurt texts published in [Munkácsi \(1952\)](#), edited by D. R. Fuchs. The Northern Mansi example is published in [Kannisto \(1956\)](#), which is part of a large collection of Mansi texts in several volumes. The Komi-Zyrian example is published in [Uotila \(1986\)](#). These are published by the Finno-Ugrian Society¹. As stated, the Kildin Saami text differs from the other examples. It was published in [Genetz \(1879\)](#) by a Hungarian publisher, it represents a religious genre and, given its age, it is clearly in the public domain. The latter appears to be true for all of the earliest contributors to these materials.

The materials used in this study have been published in GitHub², which allows replicating the results and makes it easy to test the outcome with different preferences and choices. The Table 1 shows the size of the current dataset. See the appendices A and B for example of how the prompts were formatted.

¹The author of this work is the librarian and archivist of the Finno-Ugrian Society, and the processing of these materials is part of the larger digitization initiative of the Society.

²<https://github.com/nikopartanen/finno-ugric-transliteration-examples>

Table 1: Data size, displayed with character and word count by language, transcription and orthography

	kpv	udm	mns	sjd
Chars (trans)	2101	2257	3445	2399
Chars (ortho)	1781	1977	3015	2328
Words (trans)	263	259	441	328
Words (ortho)	254	304	442	328

The selected languages and their speech communities exhibit significant linguistic, historical and sociolinguistic differences from one another. Komi and Udmurt are relatively large and widely used Uralic languages, closely related to one another, and used in different domains. Northern Mansi is a much smaller Uralic language with only thousands of speakers, and has a still standardizing, yet used, orthography. Kildin Saami is even smaller than Northern Mansi, with hundreds of speakers, and no continuous press, but still some publishing activities and an on-going orthography development. All these languages utilize Cyrillic orthographies with some characters differing from the Russian orthography, and Russian is the main contemporary contact language. This is seen in the presence of the Russian vocabulary in these texts.

When different large language models are tested, it becomes clear that most of them are not able to process Finno-Ugric transcriptions. There are hundreds of models, and evaluating most of them would be entirely unnecessary in this task. The output is usually unintelligible with clearly no understanding about these languages. At the same time we see that there are individual models that do perform above the average.

Four different models were selected and tested further. These are Claude 3 and 3.5 Sonnet³, Gemini 1.5 Pro⁴ and ChatGPT 4o⁵. We conduct two different experiments, first is a zero-shot scenario where the model is asked to transliterate the text with no additional information. This is not an ideal setting, since as explained, it is not obvious what kind of representation is actually wanted. However, it gives information about the model’s capabilities. It can be questioned why a zero-shot experiment is needed, as it is clear that example data should improve the result, but at the same time

³<https://www.anthropic.com/news/claude-3-5-sonnet>

⁴<https://deepmind.google/technologies/gemini/pro/>

⁵<https://openai.com/index/hello-gpt-4o/>

zero-shot scenario does give valuable information about what the models are capable to do without any further guidance. In the second experiment the same prompt is used, but there is an additional example provided, which illustrates the style and type of transformation that is desired. This example is taken from the same text as the transliterated sample, so that the dialect is certainly the same. The test text is one paragraph and the added example little bit longer.

4 Results

When comparing the tested models, Claude 3.5 Sonnet performs distinctly from the others. As shown in Table 2 this model produces results that are far beyond what is produced by the other models. The difference is large enough that the results of Claude 3.5 Sonnet are already close to being possibly integrated into different research tasks. The remaining mistakes are fairly nuanced as well, and many could be considered acceptable depending from how we define the wanted output and variation allowed. At the moment we measure CER and WER against only one ground truth version, which is a methodological limitation.

Some of the models display high word and character error rates. However, close to 5% character error rates we see in Komi and Udmurt are certainly useful already, and the result for Mansi is in the same category.

In order to get a more concrete overview, let’s compare some of the transliterated sentences. The mistakes are underlined.

• Transcription

- Komi-Zyrian: me tšuzi tša tša ok-miš-šo vited voin ėiked jule đerevna p u st u: n ajn i ž - v a rajjonin.
- Mansi: kit ıörn ölèy. ıanjı ıumite nėn, mān ıumite nėtāl nėtāl ıos öls, βāt öls, nė βis. nė βis, üs^on minās, βetrā βinā βis.
- Udmurt: ud-murt kišnojos nil-pi vajon-đirjazi, nilı-pizı tširkkām-kā vordkono, tuž kaptšiān vajo.
- Kildin Saami: A mañña Vavilon’ vealhtetmužest Iexonia saııj Salafuil;

• Ground Truth

- Komi-Zyrian: Ме чужи тысяча өкмыссё витод воын өтикод юльö деревня Пустыняын Изьва районын.

Table 2: Zero-shot results

Language	Tool	CER	WER
Komi	Claude-3-Sonnet	0.2092	0.7273
Komi	Claude-3.5-Sonnet	0.0492	0.2576
Komi	ChatGPT 4o	0.2626	0.8561
Komi	Gemini 1.5 Pro	0.1810	0.6591
Mansi	Claude-3-Sonnet	0.3857	0.9388
Mansi	Claude-3.5-Sonnet	0.0807	0.3469
Mansi	ChatGPT 4o	0.7848	1.0204
Mansi	Gemini 1.5 Pro	0.3572	0.8367
Udmurt	Claude-3-Sonnet	0.2265	0.5067
Udmurt	Claude-3.5-Sonnet	0.0571	0.2933
Udmurt	ChatGPT 4o	0.2571	0.8533
Udmurt	Gemini 1.5 Pro	0.1939	0.7867
Kildin Saami	Claude-3-Sonnet	0.9140	1.0000
Kildin Saami	Claude-3.5-Sonnet	0.3393	0.7731
Kildin Saami	ChatGPT 4o	0.2581	0.7395
Kildin Saami	Gemini 1.5 Pro	0.3118	0.6050

Table 3: Extended prompt results

Language	Tool	CER	WER
Komi	Claude-3-Sonnet	0.1726	0.7121
Komi	Claude-3.5-Sonnet	0.0523	0.2652
Komi	ChatGPT 4o	0.2123	0.7576
Komi	Gemini 1.5 Pro	0.1778	0.6970
Mansi	Claude-3-Sonnet	0.1854	0.6224
Mansi	Claude-3.5-Sonnet	0.0987	0.4898
Mansi	ChatGPT 4o	0.2825	0.8265
Mansi	Gemini 1.5 Pro	0.3274	0.7857
Udmurt	Claude-3-Sonnet	0.1531	0.4667
Udmurt	Claude-3.5-Sonnet	0.0429	0.2533
Udmurt	ChatGPT 4o	0.2082	0.7067
Udmurt	Gemini 1.5 Pro	0.2041	0.6400
Kildin Saami	Claude-3-Sonnet	0.2963	0.7731
Kildin Saami	Claude-3.5-Sonnet	0.2139	0.7227
Kildin Saami	ChatGPT 4o	0.3118	0.7479
Kildin Saami	Gemini 1.5 Pro	0.2867	0.7311

- Mansi: Кит ёрн олэг. Яныг хумитэ нэң, мань хумитэ нэтāl нэтāl хос олыс, вать олыс, нэ вис.
- Udmurt: Удмурт кышноёс нылпи ваён дыръязы, нылзы-пизы чырккем ке вордконо, туж капчиен ваё.
- Kildin Saami: А маңна Вавилон вьллтэдтмужэсьт Иехония соāгэ Салафйл;

• Claude 3.5 Sonnet

- Komi-Zyugian: Ме чужи тысяча өкмьсьсё витөд воын отикөд июлө деревня Пустойын Изьва районын.
- Mansi: кит ёрн олэг. яныг хумитэ нэң, мань хумитэ нэтāl нэтāl хос олс, вать олс, нэ вис.
- Udmurt: удмурт кышноос ныл-пи ваён дыръязы, нылзы-пизы чырккем ке вордсконо, туж капчиен ваё.

– Kildin Saami: А маңҕа Вавилонь
веалхтэтмуҕэст Иехония сӕгый
Салафиил;

• **Gemini 1.5 Pro**

– Komi-Zyrian: Ме тшужы тіся окмисӧ
витӧд воин ӧтикед юльӧ деревня
пустуныын Изьва районын.

– Mansi: Кити йӧнн ӧлӧх. Йӧнӧх
хумитӧ нӧнӧ, мань хумитӧ
нӧтӧл-нӧтӧл хӧс ӧлс, вӧт ӧлс, нӧ
вис.

– Udmurt: Удмурт кизноӧс нилпи ваён-
дыръязи, нилзи-пизи чирккам-ка
вордсконо, туж капчиян ваӧ.

– Kildin Saami: А мӧнньа Вавилон
веалхтӧтмуҕэст Иехония сӕйий
Салафиил;

It is noticeable that in many cases the mistakes are matters of interpretation, especially for Claude 3.5 Sonnet. For Udmurt the model returns dialectal forms such as чирккем, but then also normalizes forms such as вордконо into more literary вордсконо. This illustrates that part of the error connects to the line between dialectal representation and orthography.

In the Komi example one mistake appears in the settlement name Пустыняын, which is replaced by Пустӧйын. Interestingly, the Claude 3.5 Sonnet model did return a valid but wrong word in Komi. This would hint that this model is somehow more aware about the Komi lexicon than the others.

With the Kildin Saami the results are much worse than with the other languages. Especially the diacritics seem to cause problems with all the models. Kildin Saami is the only language where Claude 3.5 Sonnet performs worse than the other models in the first experiment, although it outperforms others in the second experiment.

Compounds are another regular issue. Many transcriptions seem to use hyphen in possible compounds where the current literary language does not. Partially the use in transcriptions seems to be guided by etymological assumptions, i.e., parts of the word that could be analysed as etymologically distinct words are differentiated with a hyphen. Reanalysis of the word boundaries has inevitable impact to the word error rate as well. This could also make the word level alignment of different versions a challenge.

Between the first and second experiment, it seems that especially Northern Mansi diacritical marks in the Cyrillic orthography improved significantly when an additional example was provided. With Kildin Saami similar phenomena could have been expected, as the language has similarly complex macron usage, but in our Kildin Saami example the improvement still left the result way worse than with the other languages and the use of diacritics did not become very close to what is expected in the current orthography.

How difficult this task is in general should be separately evaluated, but it clearly is far from trivial. Converting these transcriptions into contemporary orthographies can be a challenge even for a specialist in these languages. Especially so if we want to take the dialectal features somehow into account. We have not yet tried to estimate whether some of the transliteration tasks are objectively harder than the others. It is possible that the phonetic representation of the transcription is more complicated in case of some languages, and the dialects in these examples may differ to varying degrees from the literary languages. Also, in some languages there may be conventions to show the dialectal features in the orthographical texts.

One reason why the Komi result is so good may be connected to extensive scanning and digitization work carried out in the Komi Republic, which has made Komi materials widely available online.⁶ Similarly it could be reasonable to assume that the Kildin Saami results were worse than the others because the amount of text in this language available online is likely much smaller than on the others. There are no clear and up to date statistics about this, but there have been projects that have collected texts from the internet in different languages. Jauhiainen et al. (2020) report their results from 2017, where they have 59 sentences for Kildin Saami, 825 for Northern Mansi, 18,966 for Komi-Zyrian and 42,545 for Udmurt. Again, this is certainly not entirely reflective, as the amount of Komi materials is certainly much larger, and there is a Northern Mansi newspaper Лӱимӧ сӧрипос with already a decade of online presence and articles in HTML format. Horváth (2019, 170) measured that between 2013 and 2019 the size of the corpus produced already by these newspaper articles is over half a million tokens.

⁶Кomi кыв корпус by FU-Lab Team contains over 85 million tokens in October 2024: <https://komicorpora.ru>

Individual corpus creators may have had different well considered reasons to include and exclude various sources. However, with the modern LLMs the assumption must be that all materials that have been placed online may be collected and used to create these models. With the case of Claude 3.5 Sonnet model this data collection must have been particularly successful and wide, including numerous minority languages. Also other recent studies have indicated that Claude 3.5 Sonnet has returned very proficient translations between Russian, Azerbaijani and Lezgian (Asvarov and Grabovoy, 2024), which matches well with our results with the Uralic languages spoken in Russia. Similarly Shandilya and Palmer (2024) had the best results with Claude 3.5 Sonnet in glossing endangered languages with Retrieval-Augmented Generation.

There is the possibility that many minority language materials that are online contain various issues and different solutions with the character encoding. Whether the authors of the large language models have systematized this type of problems, or found ways to harmonize them otherwise, may have a large impact to the final result. There are also large amounts of text online that are missing some of the officially used characters, as inputting them is not always possible, which may impact the ability of the models to output them in the correct positions when needed.

5 Conclusion

We are starting to see Large Language Models that are able to process endangered Uralic languages at a very advanced level. The superiority of an individual model is fleeting, and in some months we expect to see new models with similar and even better capabilities. Still, the high accuracy of Claude 3.5 Sonnet is beyond the results we see at the moment with other models that can be publicly tested to any extent.

As other models develop similar capabilities, it would be important to evaluate them accurately against transliteration and other related tasks. Our results show that in a language specific task such as transliteration the differences between different models can be surprisingly large, and some are capable of producing close to a correct output.

As far as we have been testing these models in last years, this is the first time an LLM has been able to process this proficiently smaller Uralic lan-

guages. This is in itself a major development, and their capabilities should be extensively tested against different tasks. These could include machine translation, interlinear glossing, disambiguation or dependency parsing, just to mention a few usually manual work phases that the researchers of Uralic languages have been engaging with very regularly, and where automatization could have a major impact. Transliteration and normalization in themselves are tasks that the researchers may not have performed that often before. We have only recently started to receive high quality Unicode versions of the older transcriptions, and thereby the need may have not been acute yet either.

Ethics statement

The work discussed here has been done with materials that are almost a century old, and do not contain identifiable personal information about living individuals. They represent cultural heritage of different indigenous groups living in Russia, and processing these materials into writing systems that are currently in use by the language communities can be seen as a community oriented and beneficial task. These approaches take loosely place in the context of cultural repatriation. Making already existing materials available and more suitable for the contemporary scientific use may also lessen the need for new fieldwork and language documentation, which also can be a stress for the communities in question. At least this can enhance the contemporary fieldwork by providing larger transcribed corpora.

It must be noted that when texts are processed and eventually made available online, attention to the high quality and accuracy is necessary, as it is very likely these materials will in turn be scraped and used in new Large Language Models. If we release very large amounts of texts in our own dialectally adapted orthographies, there is a risk these will not remain separated from the materials that the language users themselves create, and there may be problems with the future language models returning varieties that are not in real use and are not desired in production.

We have used in this study proprietary models that allow limited free testing. This may create conditions where we are too reliant on commercial actors. However, as the field is advancing fast, it seems likely that similar results can eventually be repeated with open source models as well.

Acknowledgments

I would like to thank Nikolai Anisimov, Csilla Horváth, and Michael Rießler for providing proof-read orthographic examples. All errors remain my own. The feedback from two anonymous referees, whose thorough reviews were invaluable, is gratefully acknowledged.

References

- Alidar Asvarov and Andrey Grabovoy. 2024. Neural machine translation system for Lezgian, Russian and Azerbaijani languages. *arXiv preprint arXiv:2410.05472*.
- Jeremy Bradley. 2017. [Transcribe.mari-language.com](https://transcribe.mari-language.com) Automatic transcriptions and transliterations for ten languages of Russia. *Acta Linguistica Academica*, 64(3):369–382.
- Jeremy Bradley and Rogier Blokland. 2023. Mansi et al. in Print before and under Unicode. *Linguistica Uralica*, 59(4):243–257.
- Jeremy Bradley and Elena Skribnik. 2021. The many writing systems of Mansi: challenges in transcription and transliteration. *Multilingual Facilitation*, page 12.
- Mahta Fetrat Qharabagh, Zahra Dehghanian, and Hamid R Rabiee. 2024. LLM-powered grapheme-to-phoneme conversion: Benchmark and case study. *arXiv e-prints*, pages arXiv–2409.
- Arvid Genetz. 1879. Orosz-lapp nyelvmutatványok (máté evangélioma és eredeti textusok). *Nyelvtudományi közlemények*, 15(1):74–152.
- Ciprian Gerstenberger, Niko Partanen, Michael Rießler, and Joshua Wilbur. 2016. Utilizing language technology in the documentation of endangered Uralic languages. *Northern European Journal of Language Technology*, 4:29–47.
- Mika Härmäläinen, Niko Partanen, and Khalid Alnajjar. 2020. Normalization of different Swedish dialects spoken in Finland. In *Proceedings of the 4th ACM SIGSPATIAL Workshop on Geospatial Humanities*, pages 24–27.
- Csilla Horváth. 2019. The ‘extraordinary thing’: the only Mansi newspaper on online presence and social media practices. In *Digitalne medijske tehnologije i društveno-obrazovne promene 8*, pages 165–176, Serbia. Univerzitet u Novom Sadu. International scientific conference The Bridges of Media Education ; Conference date: 14-09-2018 Through 15-09-2018.
- Tommi Jauhiainen, Heidi Jauhiainen, Niko Partanen, and Krister Lindén. 2020. Uralic language identification (ULI) 2020 shared task dataset and the Wanca 2017 corpora. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 173–185, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).
- Artturi Kannisto. 1956. *Wogulische Volksdichtung. III*. Number 111 in *Mémoires de la Société Finno-Ougrienne*. Finno-Ugrian Society, Helsinki. Gesammelt und übersetzt von Artturi Kannisto, bearbeitet und herausgegeben von Matti Liimola.
- Bernhard Munkácsi. 1952. *Volksbräuche und Volksdichtung der Wotjaken*. Number 102 in *Mémoires de la Société Finno-Ougrienne*. Finno-Ugrian Society, Helsinki. Aus dem Nachlasse von Bernhard Munkácsi herausgegeben von D. R. Fuchs.
- Niko Partanen, Rogier Blokland, Michael Rießler, and Jack Rueter. 2022. Transforming archived resources with language technology: From manuscripts to language documentation. In *The 6th Digital Humanities in the Nordic and Baltic Countries 2022 Conference, Uppsala, Sweden, March 15-1, 2022.*, volume 3232, pages 370–380. CEUR-WS.
- Niko Partanen, Mika Härmäläinen, and Khalid Alnajjar. 2019. Dialect text normalization to normative standard Finnish. In *Workshop on Noisy User-generated Text*, pages 141–146. The Association for Computational Linguistics.
- Bhargav Shandilya and Alexis Palmer. 2024. Boosting the Capabilities of Compact Models in Low-Data Contexts with Large Language Models and Retrieval-Augmented Generation. *arXiv preprint arXiv:2410.00387*.
- Florian Siegl and Michael Rießler. 2015. Uneven steps to literacy: The history of the Dolgan, Forest Enets and Kola Sámi literary languages. *Cultural and linguistic minorities in the Russian Federation and the European Union: Comparative studies on equality and diversity*, pages 189–230.
- Ashima Suvarna, Harshita Khandelwal, and Nanyun Peng. 2024. Phonologybench: Evaluating phonological skills of Large Language Models. *arXiv preprint arXiv:2404.02456*.
- T. E. Uotila. 1986. *Syrjänische Texte. Band II. Komi-Syrjänisch: Ižma-, Pečora- und Vym-Dialekte*, volume 193 of *Mémoires de la Société Finno-Ougrienne*. Finno-Ugrian Society. Übersetzt und herausgegeben von Paula Kokkonen.

A Prompt example 1

Transform the following text into contemporary Northern Mansi orthography. Keep the dialectal features if there are existing conventions to retain them.

יעֶפֶרַפֶּאעֶ םֹנֶפֶאעֶ לִינֶשֶׁפֶ תִּסְתֶּפֶ. תֶּאֱבֶ תֶּאֶ כְּחֻיִי.
כְּחֹסֶאֶ כְּחֻיִיאַס, בֶּאֱתִי כְּחֻיִיאַס, נֹנְחֶכְּחֶאֱלֶ סֻנְסִי: יעֶפֶרַפֶּאעֶ
םֹנֶפֶאעֶ לִינֶשֶׁפֶ תִּסְתֶּפֶ. נֹמְסִי: »אֶמֶאֱנֶרִיֶּי לִינֶשֶׁפֶ?»
נֹכְּחֶסֶאִיקֶאֶתֶאֶכְּטֶס. כִּיתִיֶּלִיֶּאעֶ: »מֶאֱנֶרִיֶּי לִינֶשֶׁיֶנֶ?»
לֶאֱבֶפֶ: »נֶאֱנֶ נֶעֶ־נֶ אֶלֶסֶ(LN).» תֶּאֱבֶ לֶאֱבִי: »אֶמֶ מֶאֶ־נֶ
פֶּרֶלֶ אֶלֶסֶ(LM)?» »נֶאֱנֶ נֶעֶ־נֶ םֹסֶמֶלֶ פִּנֶפֶנְכֶ לֶאֱבֶסֶלֶמֶעֶ.»
»תֶּעֶ!» לֶאֱבִי.

B Prompt example 2

Transform the following text into contemporary Northern Mansi orthography. Keep the dialectal features if there are existing conventions to retain them.

יעֶפֶרַפֶּאעֶ םֹנֶפֶאעֶ לִינֶשֶׁפֶ תִּסְתֶּפֶ. תֶּאֱבֶ תֶּאֶ כְּחֻיִי.
כְּחֹסֶאֶ כְּחֻיִיאַס, בֶּאֱתִי כְּחֻיִיאַס, נֹנְחֶכְּחֶאֱלֶ סֻנְסִי: יעֶפֶרַפֶּאעֶ
םֹנֶפֶאעֶ לִינֶשֶׁפֶ תִּסְתֶּפֶ. נֹמְסִי: »אֶמֶאֱנֶרִיֶּי לִינֶשֶׁפֶ?»
נֹכְּחֶסֶאִיקֶאֶתֶאֶכְּטֶס. כִּיתִיֶּלִיֶּאעֶ: »מֶאֱנֶרִיֶּי לִינֶשֶׁיֶנֶ?»
לֶאֱבֶפֶ: »נֶאֱנֶ נֶעֶ־נֶ אֶלֶסֶ(LN).» תֶּאֱבֶ לֶאֱבִי: »אֶמֶ מֶאֶ־נֶ
פֶּרֶלֶ אֶלֶסֶ(LM)?» »נֶאֱנֶ נֶעֶ־נֶ םֹסֶמֶלֶ פִּנֶפֶנְכֶ לֶאֱבֶסֶלֶמֶעֶ.»
»תֶּעֶ!» לֶאֱבִי.

Use this text and the following orthographic representation as an example in the task.

kit jörn olèy. janiŷ k̄xumite nēŋ, mān k̄xumite
nētāl nētāl k̄xos ōls, βāt ōls, nē βis. nē βis, ūs^on
minās, βētrā βinā βis. aijun̄k tūltk̄ātas. aijnēte
palit^l jōl a^t pāti. βināte k̄xōlas i pa^lds. rāŷŷā^lts
ta k̄xuii. iεp̄p̄aε ēn̄p̄aε lāβēŷ mānēn nūp^ol: »ōs-
mal pinēln, lūl̄l̄ssāŋ pa^lds.» ēkβā ōsmā pinun̄k
tuβ kβāls. tāβ kāt nōŋk̄ā^l tōtj̄^lls, kā^ltnā k̄xōiβ^os,
ēkβā šāmraŷŷā^lds.

Кит ёрн олэг. Яныг хумитэ нэң, мань хумитэ
нэтэл. Нэтэл хос олыс, вать олыс, нэ вис.
Нэ вис, ұсын минас, вётра вина вис. Аюңкве
түлтхатас. Айнэтэ палытыл ёл ат паты. Винатэ
холас и патыс. Рагатас та хуи. Ягпыгаге-
оньгаге лавёг маньнэ нупыл: «Осмал пинэлен,
лбольшаң патыс.» Эква осма пинуңкве тув
квалыс. Тав кат нонхаль тотыгылыс, катна
хойвес, эква щам-рагатас.

Specialized Monolingual BPE Tokenizers for Uralic Languages Representation in Large Language Models

Iaroslav Chelombitko
Neapolis University Pafos
Paphos, Cyprus
i.chelombitko@nup.ac.cy

Aleksey Komissarov
Neapolis University Pafos
Paphos, Cyprus
ad3002@gmail.com

Abstract

Large language models show significant inequality in language representation, particularly for Uralic languages. Our analysis found that existing tokenizers allocate minimal tokens to Uralic languages, highlighting this imbalance. To address this, we developed a pipeline to create clean monolingual datasets from Wikipedia articles for four Uralic languages. We trained Byte Pair Encoding (BPE) tokenizers with a vocabulary size of 256,000 tokens, though Northern Sami had only 93,187 due to limited data.

Our findings revealed most tokens are unique to each language, with 8,102 shared across all four, and 25,876 shared among Estonian, Finnish, and Hungarian. Using the Compression Ratio metric, our tokenizers outperformed popular ones like LLaMA-2 and Gemma 2, reducing Finnish’s compression ratio from 3.41 to 1.18.

These results demonstrate the importance of specialized tokenizers for underrepresented languages, improving model performance and lowering costs. By sharing our tokenizers and datasets, we provide crucial resources for further research, emphasizing the need for equitable language representation.

1 Introduction

Large language models have rapidly integrated into our daily lives, yet they exhibit significant inequality in language representation due to issues of digital vitality. This imbalance is particularly pronounced in the Uralic language family, which includes well-represented languages like Finnish, Estonian, and Hungarian, as well as underrepresented ones like Northern Sami. Our initial assessment revealed that existing tokenizers inadequately represent Uralic languages, with a minimal number of tokens allocated to them.

However, there is significant inequality in language representation at all stages due to the issue

of digital vitality (Acs et al., 2017; Zaugg et al., 2022), where there is an insufficient amount of digitized text available (Arkhangelskiy, 2019). As a result, training datasets are dominated by texts in popular languages, leading to an imbalance (Choi et al., 2023). It’s not just the training datasets; the same problem exists with tokenizers (Petrov et al., 2023), as was shown earlier. The efficiency of a tokenizer directly impacts performance in that language and, consequently, the cost of tasks, since pricing is tightly linked to the number of tokens.

On the other hand, new methods are emerging that allow adjustments to the tokenizer’s vocabulary during fine-tuning, and even better, future models could pay closer attention to token representation in tokenizers during training (Downey et al., 2024). However, to achieve this, a resource is needed that enables such adjustments (Alnajjar et al., 2023; Paul et al., 2024). The approach we propose for creating monolingual tokenizers serves as exactly this kind of resource. Uralic languages provide a useful model, as the group is relatively small and includes three languages that are comparatively well represented on Wikipedia, one that is poorly represented, and a significant number of languages that are not represented at all.

First, we decided to assess how many Uralic language tokens are present in existing tokenizers. After evaluating the token representation, we realized that these languages are almost entirely absent. This led us to the question: can we train a BPE tokenizer both for individual Uralic languages and for all Uralic languages combined?

We encountered the challenge that detecting Uralic languages, beyond Finnish, Estonian, and Hungarian, is a complex task. For the tokenizer, we used Wikipedia articles from four languages: in addition to the three mentioned, we also included Northern Sami. We developed a pipeline to extract the cleanest monolingual texts for these four languages, and based on these texts, we trained four

monolingual tokenizers with a size of 256K, as well as one tokenizer for the combined languages.

Although the initial goal was to create monolingual tokenizers, the tokenizers themselves turned out to be an interesting new tool for comparative analysis. They provide an additional perspective for approaching many classic tasks (Toraman et al., 2023).

We have made both the tokenizers and the clean monolingual Wikipedia datasets publicly available. These datasets can also serve as a foundation for other research projects.

Finally, we evaluated how much more effectively our proposed tokenizers perform compared to existing ones. We also assessed the impact of tokenizer size on tokenization efficiency and proposed methods for determining how many additional tokens need to be added to a tokenizer for it to effectively handle Uralic languages.

2 Results and Methods

2.1 Assessing the Representation of Uralic Languages in Existing Tokenizers

Our first task was to assess how well Uralic languages are represented in existing tokenizers. To study the representation of Uralic languages within the tokenizers of popular LLMs, we first classified each token according to the language it belongs to. Table 1 presents the Uralic language encodings from the key standards.

The ISO 639-1 standard encompasses three languages: Estonian, Finnish, and Hungarian. Expanding to broader classifications, the ISO 639-2 standard includes nine languages, while ISO 639-3 covers twenty-seven languages. Additionally, the Common Locale Data Repository (CLDR) incorporates sixteen Uralic languages. This hierarchical structuring across multiple standards highlights the varying levels of granularity and coverage, underscoring the importance of comprehensive language encoding for effective NLP applications in the Uralic language family. While our findings are promising, we are currently unable to accurately assess the representation of these languages in the Common Crawl dataset, which constitutes a substantial portion of the training data for foundation models. The seemingly large number of Uralic languages represented in standards like ISO 639-3 is offset by the fact that most language classifiers do not use this standard. To compare the classifications, we use two popular libraries for language identification

Tokenizers	llama-2	llama-3	gemma
fi	415	875	3518
et	429	1082	3335
hu	711	1359	4456

Table 1: Langid numbers of tokens for llama-2, llama-3, and gemma tokenizers across three languages.

Tokenizers	llama-2	llama-3	gemma
fi	202	779	1905
et	251	1145	2066
hu	403	1115	2570

Table 2: Cld3 numbers of tokens for llama-2, llama-3, and gemma tokenizers across three languages.

— langid.py (Lui and Baldwin, 2012) and pylcl3 (Ooms, 2024) The langid.py classifier is based on the ISO 639-1 standard, which significantly limits the classification of Uralic languages, narrowing it down to Finnish, Estonian, and Hungarian. The pylcl3 classifier uses the CLDR standard, which offers an apparent advantage over langid.py and ISO 639-1 when it comes to Uralic language classification. However, none of the existing classification tools are comprehensive, as they have considerable biases when classifying short tokens and underrepresented languages (Chelombitko et al., 2024). To highlight this issue, we provide a comparison of the number of Uralic language tokens in the tokenizers of top models with publicly available tokenizers in Table 1 and 2.

2.2 Creating Monolingual Datasets Based on Wikipedia Articles

We downloaded Wikipedia dumps in ZIM format for four languages—Estonian, Finnish, Hungarian, and Northern Sami. The links are organized by language and further classified by size, such as "nopic" or "maxi," and date. Specifically, links related to the "all" topic are added to the data structure. For each language, we prioritize the "nopic" size if it is available. If the "nopic" size is not available, the "maxi" size is used instead. Within each size category, the links are sorted by date to identify the most recent links. If a link from May 2024 is available, it is given priority. For Finnish, no link from May 2024 was available. Therefore, we used the most recent link from May 2023.

Downloaded zim files were processed with a custom C++ program available in github. It iterates through each entry in the archive by path. If

an entry is identified as a redirect, it is skipped. The script parses the HTML content using an `HtmlParser` library. The script looks for elements with the class name “mw-parser-output” within the parsed HTML document. If such elements are found, the script further extracts all HTML paragraph elements within the “mw-parser-output” element. For each paragraph element found, the script outputs the HTML content of the paragraph. Finally we used a custom python script to clean HTML paragraphs of the text. The script reads input from the standard input, which is usually provided via a pipeline or redirection. The input text is passed through the ‘clean_text’ function, and the cleaned text is printed to the standard output.

To eliminate contamination from other languages, we performed two iterations. In the first iteration, we removed all paragraphs containing non-Latin characters, which we identified based on their Unicode values. In addition, we removed all paragraphs containing fewer than ten words, which is particularly relevant for Wikipedia. Wikipedia articles often include not just text but also various links and other irrelevant content that we wanted to eliminate. The ten-word filter effectively removed such unnecessary paragraphs. In the second iteration, we used CLD2, a tool commonly used for language detection in the Common Crawl dataset.

We removed all paragraphs that CLD2 did not identify as at least 90 percent belonging to the target language in order to ensure the dataset remained monolingual. The exception was Northern Sami, as CLD2 does not support this language. For Northern Sami, we retained all paragraphs without applying language detection.

Within Hungarian language content, there exists both modern Hungarian and historical Hungarian (Old Hungarian) texts. In our research, we only include modern Hungarian content, as Wikipedia dumps are categorized using ISO 639-1 language codes, which only includes the code ‘hu’ for Hungarian. While Old Hungarian exists as a distinct language variety in the more comprehensive ISO 639-3 standard (with code ‘ohu’), the current Wikipedia’s infrastructure does not distinguish between historical and modern variants of Hungarian.

The significant reduction in the number of paragraphs—almost 35 percent—is due to the fact that a Wikipedia paragraph is far from being clean text. It contains a large amount of metadata, and many paragraphs include links to other languages, frag-

ments of other languages, and quotes in different languages. We removed all of this to make our datasets as monolingual as possible. Comprehensive statistics for the dataset are provided in Table 3.

2.3 Training and Characterization of Tokenizer

Typically, a tokenizer training dataset uses only a portion of the available data. However, we decided to use the entire monolingual Wikipedia dataset we created to minimize any bias from dataset sampling on the tokenizer. By using all available data, we aim to produce a tokenizer that is as close as possible to an ideal tokenizer for the given language.

For training the tokenizer, we used a custom program based on the `Tokenizers` library from `HuggingFace`. The program is available on our `GitHub`. The only change we made to the reference code was setting the vocabulary size to 256,000, as we aimed to capture the maximum number of tokens. It’s clear that such a large tokenizer is not practical for use, and the actual tokenizer will be a subset of these 256,000 tokens.

Interestingly, for Northern Sami, the dataset was insufficient to train a tokenizer with 256,000 tokens. We were only able to generate 93,187 tokens for this language. It is important to emphasize that even with some amount of text available, a minimum dataset size is required to effectively train a tokenizer. Nevertheless, the tokenizer for Northern Sami remains a valuable outcome.

2.4 Comparative Tokenology of Uralic Languages

With the trained tokenizers in hand, we asked how many common tokens exist between them. Since we created a monolingual dataset for training and ensured that the tokenizers were as monolingual as possible, we were curious to see how many tokens are shared among the Uralic languages, as well as how many tokens are specific to each language. When we used our tokenizer and examined the total number of unique tokens across all tokenizers, we found that there were 785,115 unique tokens. Not surprisingly, the majority of tokens are unique to a single tokenizer. We were particularly interested in finding how many tokens are shared across tokenizers, but we found no fully common tokens. The number of tokens that appear in only one tokenizer is 692,081. Only 8,102 tokens are shared across all four languages. After excluding Northern Sami,

Language Code	Estonian et	Finnish fi	Hungarian hu	Northern Sami se
Content pages	248,159	581,930	548,859	7,892

Table 3: Language and number of content pages in Wikipedia.

Language Code	Estonian et	Finnish fi	Hungarian hu	Northern Sami se
Raw Paragraphs	1,342,090	3,141,975	3,932,082	18,652
HQ paragraphs	926,628 (69%)	2,081,601 (66%)	2,385,635 (61%)	6,720 (36%)
HQ Words	33,140,238	86,033,550	122,512,745	249,444
HQ Chars	265,243,383	765,211,790	967,144,007	1,938,899

Table 4: Language and number of paragraphs, words, and characters.

we examined the common tokens between Estonian, Finnish, and Hungarian, and found 25,876 shared “core” tokens. It’s important to note that, since the data comes from Wikipedia, the number of shared tokens is slightly overestimated. This is due to the presence of similar names, toponyms, and terms. While we tried to exclude non-Latin scripts, some countries and terms appear in English on Wikipedia, leading to an inflated count. In reality, the number of truly shared tokens is likely lower.

2.5 Evaluation of Tokenizer Efficiency for Tokenizing Texts in Uralic Languages

Our next task was to assess how effectively our tokenizers, trained on individual languages, tokenize their corresponding datasets. We compared their performance with widely used open tokenizers, including Mistral, LLaMA-2, LLaMA-3, and Gemma 2. We aimed to determine whether our language-specific tokenizers outperform these general-purpose tokenizers, which inevitably include tokens from multiple languages. To evaluate tokenizer efficiency, we used the Compression Rate metric. This metric assesses how effectively the tokenizer performs during training. The smaller the resulting text after tokenization, the better the tokenizer’s performance. Overall, this metric can be interpreted as ‘how many tokens are needed to represent a single word.’ The closer this value is to one, the more efficient the tokenizer. The results of the analysis are presented in Table 5.

The figure 1 shows tokenization of parallel text samples in English and Finnish using different tokenizers. The Finnish text is processed by GPT4o (149 tokens, compression ratio 2.76), fi BPE (77 tokens, compression ratio 1.43), and uralic BPE (86 tokens, compression ratio 1.59). The English tokenization ratio (1.96) can be considered as a ref-

	urBPE	llama2	llama3	Gemma2
et	1.13	3.09	2.88	2.45
fi	1.18	3.41	3.12	2.49
hu	1.11	2.69	2.83	2.18
se	1.32	3.53	3.27	3.04
uralic	1.24	3.00	2.94	2.33

Table 5: Comparison of tokenizer efficiency between our tokenizer and three popular open-source tokenizers for four individual languages and the combined Uralic dataset. The metric used is the Compression Ratio, which is the ratio of the number of tokens after tokenization to the number of words in the dataset. The lower the ratio, the better the performance.

erence point for expected tokenizer performance, showing that current general-purpose tokenizers handle Finnish significantly worse than English (2.76 vs 1.96). However, specialized Finnish and Uralic tokenizers achieve even better compression ratios than English, demonstrating that proper language-specific tokenization can be highly efficient.

This visualization effectively demonstrates the efficiency gains achieved by language-specific tokenizers over general-purpose ones, with fi BPE showing nearly 50% reduction in token count compared to GPT4o for the same content. This reduction in token count directly translates to a proportional decrease in processing costs - effectively halving the cost of working with Finnish text. While our current work focuses on building separate specialized tokenizers, future research might explore whether similar efficiency gains could be achieved by selectively adding essential language-specific tokens to existing tokenizers, potentially offering a more practical path to improving multilingual performance while maintaining compatibility with existing models.

GPT4o: 149 (2.76)

Yhdessä volgalaiskielten kanssa ryhmä on muodostanut uralilaisten kielten perinteisessä binäärisessä sukupuuluokittelussa suomalais-volgalaisien kielten ryhmän, mutta nykyinen tutkimus ei enää pidä sen paremmin volgalaisia kuin suomalais-volgalaisiakaan kieliä yhteisen kantakielen muodostaneena kieliryhminä vaikka käsitteitä voidaankin käyttää alueellisina nimityksinä. Itämerensuomalaisten kielten yhteyden volgalaiskieliin arvellaan katkenneen viimeistään vuoden 1000 paikkeilla, kun venäläisasutus levittäytyi pohjoiseen.

fi BPE: 77 (1.43)

Yhdessä volgalaiskielten kanssa ryhmä on muodostanut uralilaisten kielten perinteisessä binäärisessä sukupuuluokittelussa suomalais-volgalaisien kielten ryhmän mutta nykyinen tutkimus ei enää pidä sen paremmin volgalaisia kuin suomalais-volgalaisiakaan kieliä yhteisen kantakielen muodostaneena kieliryhminä vaikka käsitteitä voidaankin käyttää alueellisina nimityksinä. Itämerensuomalaisten kielten yhteyden volgalaiskieliin arvellaan katkenneen viimeistään vuoden 1000 paikkeilla, kun venäläisasutus levittäytyi pohjoiseen.

GPT4o: 106 (1.96)

Together with the Volga languages, the group has formed the Finno-Volga group in the traditional binary genealogical classification of Uralic languages, but current research no longer considers either the Volga or Finno-Volga languages as a language group that formed a common ancestral language, although the terms can be used as regional designations. It is thought that the connection between the Baltic Finnish languages and the Volga languages was broken at the latest around the year 1000, when the Russian settlement spread northwards.

uralic BPE: 86 (1.59)

Yhdessä volgalaiskielten kanssa ryhmä on muodostanut uralilaisten kielten perinteisessä binäärisessä sukupuuluokittelussa suomalais-volgalaisien kielten ryhmän mutta nykyinen tutkimus ei enää pidä sen paremmin volgalaisia kuin suomalais-volgalaisiakaan kieliä yhteisen kantakielen muodostaneena kieliryhminä vaikka käsitteitä voidaankin käyttää alueellisina nimityksinä. Itämerensuomalaisten kielten yhteyden volgalaiskieliin arvellaan katkenneen viimeistään vuoden 1000 paikkeilla, kun venäläisasutus levittäytyi pohjoiseen.

Figure 1: Visual comparison of tokenization patterns across different tokenizers. Each tokenizer’s output is color-coded to show individual token boundaries, illustrating how the same text is segmented differently. Parallel text is shown in English (top-right panel, GPT4o) and Finnish (other panels) using different tokenizers.

3 Discussion

3.1 The Importance of Creating Well-Designed Multilingual Tokenizers

As shown in Table 5, if a tokenizer compresses texts efficiently, this not only holds fundamental importance but also has practical implications for the use of large language models, whose costs increase with each new iteration and generation. Perhaps we should rethink how we approach training tokenizers—not just by processing large amounts of text, but by assembling tokenizers from more specialized ones. This would ensure that tokens and languages are represented more equally in each tokenizer, allowing them to tokenize a wider variety of texts rather than primarily texts in the most popular languages, such as English, Russian, Chinese, German, and Japanese.

3.2 Significant Challenges with Low Digital Vitality Languages

For languages with low representation in textual data, none of the available tools for language analysis work reliably (Zaugg et al., 2022). We encountered difficulties even with relatively larger languages, such as Estonian. This issue is critical not only for endangered languages but also for any underrepresented languages in online texts (Hjortnaes et al., 2021). The most accurate count of existing Uralic languages is around 39-40 living and extinct languages (including recently extinct ones like Kamassian). Currently, 14 Uralic languages have

their own Wikipedia editions, and only four have automatic Wikipedia dumps for computer-friendly processing.

The complete absence of Wikipedia editions for the most Uralic languages actually reinforces our paper’s main point about the severe digital divide within the Uralic family. While other text sources might exist, using them would require entirely different methods for data cleaning and quality control.

The issue is not just the availability of text; it also lies in the tools themselves, which are trained on specific datasets. The performance of these tools varies greatly across different languages. This needs to be considered, and we believe that providing a high-quality monolingual dataset is a crucial resource for both evaluating existing tools and, in cases where they perform poorly, improving them or developing new tools tailored specifically to languages for which standard tools are ineffective.

In addition, we observed that language detection proved to be quite challenging, especially for Estonian. Estonian Wikipedia articles often contain more English words than other languages, which led to many false negatives and false positives in the detection process.

One potential approach for low-resource languages could be the artificial transfer of tokenizers from one language to another. This might be a solution for such languages, as word fragments can still be derived from relatively small amounts of text,

as we have demonstrated here with Northern Sami. Despite the limited amount of text, we were able to extract the basic tokens quite effectively; however, the challenge lies with full words. Perhaps enriching tokenizers with additional dictionaries could be a viable solution for languages that lack sufficient textual resources.

3.3 Monolingual BPE tokenizer as promising new tool for comparative linguistic

One of the surprising properties of BPE is its ability to find all repeating substrings in a language, making it a potentially valuable tool for comparative linguistics, especially when working with large text corpora. It can serve as an additional resource alongside existing linguistic tools (Hämäläinen, 2019; Silfverberg and Tyers, 2019).

This tool can be used, for example, for language detection, potentially more effectively than probability-based models. In essence, BPE is also a type of probabilistic model, but it has the advantage of identifying tokens that are more characteristic of a language, even if they are relatively short. By analyzing these short tokens, we can estimate the likelihood that a given text belongs to a specific language. Currently, no such tools exist Language Modeling and Perplexity Reduction

3.4 Texts Authorship Quality in Low-Resource Uralic Languages

For low-resource languages, particularly within the Uralic language family, large-scale datasets remain scarce. A critical consideration when working with extremely low-resource languages is the assessment of dataset quality and representativeness. This is particularly relevant for languages with a small number of native speakers yet maintain some presence in digital corpora. A fundamental methodological challenge arises regarding the linguistic authenticity of the collected data, specifically concerning the authorship of corpus texts. The distinction between native speakers and proficient L2 (second language) users becomes crucial, as the latter may possess sufficient competency to produce texts while potentially introducing subtle non-native patterns. In low-resource scenarios, where the corpus size is inherently limited, such authorship variations can significantly impact the linguistic quality of the dataset, subsequently affecting downstream tasks such as tokenization and model training.

Future research will explore methodologies for assessing dataset authenticity through perplexity-

based analysis. We hypothesize that texts authored by native and non-native speakers will demonstrate measurably different perplexity patterns, though the exact nature of these differences requires empirical investigation. We plan to validate this hypothesis by developing a systematic approach to perplexity-based authorship analysis, which could provide a quantitative tool for dataset quality assessment in low-resource scenarios, particularly valuable for Uralic languages where high-quality digital data is crucial yet scarce.

4 Conclusion

In this work, we addressed the inequality in language representation within large language models, focusing on the Uralic language family. Using high-quality monolingual datasets from Wikipedia for Estonian, Finnish, Hungarian, and Northern Sami, we trained specialized BPE tokenizers. Our analysis showed that existing tokenizers poorly represent Uralic languages, leading to inefficient tokenization and increased computational costs. Our monolingual tokenizers outperformed widely used open-source tokenizers with lower compression ratios, improving both performance and cost-efficiency for underrepresented languages. Additionally, these tokenizers offer valuable tools for comparative linguistic analysis, highlighting shared and unique features of Uralic languages. By sharing our tokenizers and datasets, we provide key resources for further research in natural language processing for low-resource languages. This work underscores the need for equitable tokenization, especially for languages with low digital vitality.

5 Availability

All resources from this research are publicly available:

- Source Code: Complete analysis tools and scripts are available in our GitHub repository <https://github.com/nup-csai/uralicBPE>
- Reproducibility: A Jupyter notebook containing step-by-step analysis reproduction, including all tables and figures
- Data: Processed Wikipedia datasets and pre-trained tokenizers are hosted on Hugging Face <https://huggingface.co/datasets/nup-csai/uralicBPE>

References

- Judit Acs, Katalin Pajkossy, and András Kornai. 2017. [Digital vitality of Uralic languages](#). *Acta Linguistica Academica*, 64(3):327–345.
- Khalid Alnajjar, Mika Härmäläinen, and Jack Rueter. 2023. [Sentiment analysis using aligned word embeddings for Uralic languages](#). In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 19–24, Tórshavn, the Faroe Islands. Association for Computational Linguistics.
- Timofey Arkhangelskiy. 2019. [Corpora of social media in minority Uralic languages](#). In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 125–140, Tartu, Estonia. Association for Computational Linguistics.
- Iaroslav Chelombitko, Egor Safronov, and Aleksey Komissarov. 2024. [Qtok: A Comprehensive Framework for Evaluating Multilingual Tokenizer Quality in Large Language Models](#). *arXiv preprint arXiv:2410.12989*.
- Dami Choi, Derrick Xin, Hamid Dadkhahi, Justin Gilmer, Ankush Garg, Orhan Firat, Chih-Kuan Yeh, Andrew M. Dai, and Behrooz Ghorbani. 2023. [Order Matters in the Presence of Dataset Imbalance for Multilingual Learning](#). *arXiv preprint arXiv:2312.06134*.
- C. M. Downey, Terra Blevins, Dhvani Serai, Dwija Parikh, and Shane Steinert-Threlkeld. 2024. [Targeted multilingual adaptation for low-resource language families](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15647–15663, Miami, Florida, USA. Association for Computational Linguistics.
- Nils Hjortnaes, Niko Partanen, and Francis M. Tyers. 2021. [Keyword spotting for audiovisual archival search in Uralic languages](#). In *Proceedings of the Seventh International Workshop on Computational Linguistics of Uralic Languages*, pages 1–7, Syktyvkar, Russia (Online). Association for Computational Linguistics.
- Mika Härmäläinen. 2019. [UralicNLP: An NLP Library for Uralic Languages](#). *Journal of Open Source Software*, 4(37):1345.
- Marco Lui and Timothy Baldwin. 2012. [langid.py: An off-the-shelf language identification tool](#). In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.
- Jeroen Ooms. 2024. [cld3: Google’s Compact Language Detector 3](#).
- Ravi Paul, Himanshu Buckchash, Shantipriya Parida, and Dilip K. Prasad. 2024. [Towards a More Inclusive AI: Progress and Perspectives in Large Language Model Training for the Sámi Language](#). *arXiv preprint arXiv:2405.05777*.
- Alexander Petrov, Enrico La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. [Language Model Tokenizers Introduce Unfairness Between Languages](#). *arXiv preprint arXiv:2305.15425*.
- Miikka Silfverberg and Francis Tyers. 2019. [Data-driven morphological analysis for Uralic languages](#). In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 1–14, Tartu, Estonia. Association for Computational Linguistics.
- Cansu Toraman, Emine Hande Yilmaz, Firat Şahinuç, and Osman Ozelik. 2023. [Impact of Tokenization on Language Models: An Analysis for Turkish](#). *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–21.
- Isabelle A. Zaugg, Anushah Hossain, and Brendan Molloy. 2022. [Digitally-disadvantaged languages](#). *Internet Policy Review*, 11(2).

Compressing Noun Phrases to Discover Mental Constructions in Corpora – A Case Study for Auxiliaries in Hungarian

Balázs Indig and Tímea Borbála Bajzát

Doctoral School of Linguistics

National Laboratory for Digital Heritage

Eötvös Loránd University Department of Digital Humanities

indig.balazs@btk.elte.hu, bajzat.timi9696@gmail.com

Abstract

The quantitative turn in functional linguistics has emphasised the importance of data-oriented methods in describing linguistic patterns. However, there are significant differences between constructions and the examples they cover, which need to be properly formalised. For example, noun chains introduce significant variation in the examples, making it difficult to identify underlying patterns. The compression of noun chains into their minimal form (e.g. as they appear in abstract constructions) is a promising method for revealing linguistic patterns in corpora through their examples. This method, combined with identifying the appropriate level of abstraction for the additional elements present, allows for the systematic extraction of good construction candidates. A pilot has been developed for Hungarian infinitive structures, but is adaptable for various linguistic structures and other agglutinative languages.

1 Introduction

Functional construction grammars (Goldberg, 1995; Langacker, 2005) have recognised that human language consists of a network of symbolic form-meaning pairs (Langacker, 2008), which are influenced by frequency of use (Bybee, 1995). As a result, several methods have been developed and applied to support empirical exploration, using corpus-based and corpus-driven methods to identify linguistic patterns by considering frequency data, rather than relying on introspection, such as collocation metrics and collostructional analysis (Glynn and Robinson, 2014; Gries and Stefanowitsch, 2007; Luodonpää-Manni et al., 2017). The trend, which has led much of functional linguistic research to adopt a data-driven and quantitative approach, is referred to as the *quantitative turn* (Janda, 2013). In linguistic description, collecting datasets of sufficient quantity and quality, and processing them in an unbiased manner has

become a clear challenge for linguists. However, without a well-structured technological apparatus, intuition distorts the objectivity of the query and the analysis used to identify patterns. The ability to validate patterns derived from theory using corpus-driven methods has, therefore, become a pressing issue. It is not sufficient for the results precisely match the theory, each step and the chain of reasoning must be examined to correct intuition.

Extracting constructions from corpora requires the processing of a large number of individual examples. We have found that the combinatorial explosion mainly arises from noun phrases that vary greatly in length and internal structure. However, the constructions we are looking for are not primarily concerned with the noun phrases they contain and, therefore, do not define their form precisely. To simplify entries, we compressed the noun phrases into their minimal form, which allowed us to extract shorter, more schematic patterns that better aligned with our theoretical expectations.

To test our hypothesis, we developed a rule-based method for compressing noun chains using practically POS tags, which we applied to the example clauses. The remaining elements in the constructions are replaced by either a word form, lemma or POS tag, depending on the expectations and their statistical variations, to achieve the best coverage. In this paper, we focus on the noun phrase compression method in light of the constructions found.

2 Data Sources

We used two corpora for the measurements, allowing us to compare and validate our method against overfitting. The first corpus, the *Hungarian Gigaword Corpus 2.0.5* (HGC2) (Oravecz et al., 2014), with 1.04 billion words, contains texts from six stylistic and five regional varieties. The second corpus, the *Hungarian Webcorpus 2.0*

Type of the Auxiliary Verbs	POS	Original (HGC2)	Filtered (HGC2)	% (HGC2)	Original (HW2)	Filtered (HW2)	% (HW2)
Akar [‘want to’]	verb	610 836	419 324	68.65	650 000	518 123	79.71
Bír [‘can’, ‘endure’]	verb	22 191	15 387	69.34	179 846	112 112	62.34
Hajlandó [‘prone to’]	adj.	48 267	36 334	75.28	272 806	179 330	65.74
Képes [‘able to’]	adj.	134 843	86 833	64.40	650 000	462 001	71.08
Képtelen [‘incapable to’]	adj.	48 036	14 424	30.03	164 909	104 274	63.23
Kíván [‘wish to’]	verb	192 678	139 498	72.40	650 000	391 413	60.22
Mer [‘dare to’]	verb	63 729	39 177	61.47	473 966	278 887	58.84
Szeret(ne) [‘(would) like (to)’]	verb	484 448	278 834	57.56	650 000	448 324	68.97
Tud [‘can’]	verb	675 000	466 863	69.16	650 000	540 175	83.10

Table 1: The distribution of samples in the two corpora is as follows: From the HGC2, a total of 2 097 149 instances were collected (1 496 674 remaining after automatic filtering, constituting 71.37% of the total sample). From the HW2, 3 691 527 instances were collected (3 034 639 remaining after automatic filtering, constituting 82.21% of the total sample).

(HW2) (Nemeskey, 2020), has approximately 9 billion words and was obtained from the Web (Common Crawl). Both corpora are accessible through the *NoSketch Engine* corpus query framework (Kilgarriff et al., 2007). However, the amount of data that can be exported is limited, so the aim was to obtain as comprehensive a sample as possible within the system’s constraints.

We used the samples from an experiment conducted by Indig and Bajzát (2023) and extended them to include the most common modal auxiliaries with infinitives in Hungarian, as well as patterns of adjective + infinitive associated with modal meaning (Van linden, 2010), to get a more comprehensive picture. Table 1. summarises the sample sizes. The extracted concordances were re-analysed to ensure consistent and up-to-date annotation (Indig et al., 2019; Novák et al., 2016).

3 The Mosaic Method

The mosaic n-gram method can model linguistic data at different levels of abstraction, such as word, lemma, and POS tag, simultaneously (Indig, 2017). This concept aligns with the usage-based approach, as linguistic schemas become entrenched at various levels of semanticity, and patterns are conventionalised based on their frequency of use (Bybee, 1995). The mosaic n-gram method efficiently generates and ranks all possible abstractions of linguistic data, thereby reducing the reliance on linguistic intuition to identify statistically sound construction candidates that have the appropriate level of abstraction for each element to maximise coverage.

The method includes a classification step to iden-

tify *inferior mosaic n-grams* (i.e. subset relation): any less frequent mosaic n-gram that generalises from a set of examples that are a subset of those covered by a more frequent mosaic n-gram. Conversely, the more frequent n-gram is labelled as major compared to the less frequent one if the latter only covers examples that are also covered by the former. Among entries of equal frequency covering the same set of examples, all but the least abstract are deleted as redundant. In addition, by setting a frequency threshold to discard entries that are rare despite their abstraction, the processing time and the number of entries to be manually checked can be further reduced. This approach allows for a high degree of customisation, and the generated mosaic n-gram patterns can be easily converted into a query expression (e.g. the CQL in Sketch Engine (Kilgarriff et al., 2007) to check them in the corpus and explore linguistic data matching the pattern). In an abstraction of the method to additionally handle free word order n-grams are substituted for bag of words (Indig and Bajzát, 2023).

4 Compressing Noun Phrases

The formal definition of constructions typically includes only the bare minimum of noun phrases, as the focus is on the whole structure. When clauses that may contain elements at various levels of abstraction are automatically compared (e.g. element by element), it becomes difficult to match patterns of different lengths without introducing additional measures. However, noun phrases with modifiers can separate functionally and structurally similar samples (see the first three rows of Table 2).

N	Frequency	Example				
3	1688			[/N][Acc]	lemma:akar	[/V][Inf]
4	1103		[/Adj][Nom]	[/N][Acc]	lemma:akar	[/V][Inf]
5	1665	[/Adj][Nom]	[/Adj][Nom]	[/N][Acc]	lemma:akar	[/V][Inf]
5	1365	[/Det Art.Def]	[/Adj][Nom]	[/N][Acc]	lemma:akar	[/V][Inf]
5	997	[/Det Art.Def]	[/N][Nom]	[/N][Poss][Acc]	lemma:akar	[/V][Inf]

Table 2: All five entries could be reduced to ‘[/N][Acc] lemma [/V][Inf]’ without violating syntax. (*akar* ‘want to’)

On the other hand, sequences of the same length with different modifiers in the noun phrases present a different issue. The last three rows of Table 2. show that, although the modifiers differ, the head noun appears in the same grammatical case, making the sequences practically analogous from our perspective. In summary, based on the examples presented above, it can be assumed that the majority of the found examples originate from such partially abstracted (i.e. not simplified) sequences.

We chose a rule-based approach because, in Hungarian, apart from a few well-separated cases, noun phrases can be trivially compressed using word order, morphology, and POS tags with simple regular expressions, and we can retain the property of converting the resulting patterns to CQL. A challenging case is possessives, where there are two ways of expressing the genitive function, both of which are homonymous. The first is marked by the nominative case, followed immediately by its property (e.g. *a* [/Det][Art|Def] *kutya* [/N][Nom] *háza* [/N][Poss.3Sg][Nom] ‘the dog’s house’). The second is expressed with the dative case (e.g. *a* [/Det][Art|Def] *kutyának* [/N][Nom] *a* [/Det][Art|Def] *háza* [/N][Poss.3Sg][Nom] ‘the house of the dog’), which allows for flexible word order and even interruption (e.g. *a kutyának* [/N][Dat] *lefestette* [/V][Pst.Def.3Sg] *a* [/Det][Art|Def] *házát* [/N][Poss.3Sg][Acc] ‘he/she painted the house of the dog’). While nominative homonymy is easy to handle because of its word order (Ligeti-Nagy et al., 2019) (see the last row in Table 2.), examples with the dative variant were excluded from our sample. This concerns 0.36% of instances (across 11 types) in HGC2, whereas 0.59% of instances (across 27 types) in HW2.

The rules are iterated in two steps, as running them simultaneously would produce incorrect results due to the aforementioned ambiguity of the nominative case. The first iteration specifically compresses cases where nouns with possessive suffixes are preceded by unmarked

genitive cases (e.g. *János* [/N][Nom] *könyvét* [/N][Poss.3Sg][Acc] *olvasták* [/V][Pst.Def.3Pl] ‘(They) read the book of John’ compared to *János* [/N][Nom] *a* [/Det][Art|Def] *könyvét* [/N][Poss.3Sg][Acc] *olvasta* [/V][Pst.Def.3Sg] ‘John read his book’). Then the remaining cases are processed to avoid overlaps. Finally, the aggregated frequencies of identical entries are calculated and classified according to their new lengths.

The procedure could be trivially refined and applied similarly to other agglutinative languages (e.g. Uralic languages). However, such rule-based transformations rely heavily on morphological markedness, and the results depend on both the language analysis tool and the quality of the corpus used (e.g. the amount of noise present).

5 Evaluation

We evaluated the results in two ways. First, we examined the changes in the distribution of frequencies and the total number of resulting entries. Next, we analysed how the most frequent and general patterns are reflected in the usage patterns found, which enables further analysis.

5.1 Changes in the top candidates

Our first concern was to see how the number of different patterns (types) associated with each length had changed. First, the longest patterns (N = 7 and above) have disappeared, likely because they contain noun phrases with two or more elements. There is also a noticeable decrease in the number of types for the shorter constructions, but the rate flattens out as the length decreases (e.g. N = 7 to 5). However, the opposite trend can be seen for N = 3 and 4 as a result of the compression. The number of types increases significantly for 4 grams, while the increase for 3 grams is more limited. Most auxiliaries exhibit these trends (see Table 3.).

In cases where significant differences were observed between the two corpora (e.g. *Hajlandó* N = 4), the discrepancy arises from the fact that the sam-

	N = 3		N = 4		N = 5		N = 6		N = 7	
	H	W	H	W	H	W	H	W	H	W
Akar	103.26	103.06	117.22	109.87	52.79	49.05	33.33	34.67	0	0.99
Bír	100	100	115.38	122.78	57.14	63.73	0	41.67	0	65.28
Hajlandó	100	112.50	200	163.41	87.10	60.55	85.71	76.47	0	5.00
Képes	108.33	103.92	268.42	229.63	63.33	47.64	5.26	21.95	0	3.26
Képtelen	125.00	107.14	150	116.67	25.00	52.94	0	5.88	-	0
Kíván	148.15	116.95	225.49	217.56	23.76	29.64	29.41	37.82	0	0.83
Mer	100	110.94	107.41	107.83	74.42	68.07	0	30.77	-	0
Szeret	111.32	103.40	119.21	119.29	28.92	34.34	6.25	10.80	0	0
Tud	111.25	109.52	137.65	137.46	53.92	50.47	40.25	39.34	2.04	0.88

Table 3: Change in samples compared to the initial type distribution (%) from HGC2 (H) and HW2 (W)

ples drawn from HGC2 were considerably smaller than those from HW2. This resulted in a proportionally greater increase in the number of entries as they were compressed. Excluding the three adjective samples, the average difference between the two corpora is 2.56%, indicating relatively similar values across both corpora.

Overall, the number of types decreased to 54.43% (2593 types) for HGC2, while HW2 showed a reduction to 60.83% (5573 types). This indicates that, on average, between 289 and 622 different types per auxiliary/adjective remain. The resulting entries are construction candidates that can be easily validated manually due to their limited quantity (cf. the large number of examples they cover). Thus, the expected reduction in the number of individual entries, aimed at enhancing the analysis, has been successfully achieved, leaving only the qualitative evaluation.

5.2 Constructional similarities

The quality of the remaining patterns can best be observed by looking at common types and their variations across several auxiliaries. To illustrate this, we selected the ten most frequent entries of the 3- and 4-gram types (see Table 4 in the Appendix.) as they cover 64.69% of the instances for 3-grams and 34.37% for 4-grams in HGC2, while 65.97% for 3-grams and 29.44% for 4-grams in HW2. Longer n-grams (≥ 5 , which account for only 25.92% and 27.38% of the examples respectively) would require a different approach due to data sparsity, as we cannot be sure why they are missing for individual auxiliaries.

The present study (Table 4.) validates the previous findings on Hungarian auxiliary verb structures (Indig and Bajzát, 2023; Bajzát, 2022;

Kálmán et al., 1989) using a corpus-driven approach and extends the scope of the investigation to predicative adjectives with modal meanings. The verbs *akar* ('want'), *tud* ('can') and *szeret(ne)* ('(would) like (to)') were omitted from Table 4., because they were found for all the patterns listed (i.e. they do indeed show frequent prototypical patterns for auxiliaries as expected). Beyond that, one can identify patterns in Table 4 that are specific to auxiliaries (e.g. insertion [/Prev] L [/V][Inf], which indicates a greater semantic integration between the infinitive and the auxiliary (Imrényi, 2013; Modrián Horváth, 2020; Bajzát, 2020) and serves as a key criterion for auxiliary (Kálmán et al., 1989)). While the insertion of certain predicative adjectives (*be* [/Prev] *képes* [/Adj][Nom] *menni* [/V][Inf] 'he/she is able to go into') is theoretically possible and could represent the next step in their grammaticalisation process (Langacker, 2006; Heine and Nargog, 2012), we found no evidence of such behaviour. However, frequent patterns (e.g. [/N][Acc] L [/V][Inf]) do not exclusively specify auxiliary structures, as they also often occur with other verbs as well. A more sophisticated and systematic analysis of construction patterns (e.g. using this method) could reveal auxiliaries and predicative adjectives with similar modal functions but belonging to different word classes. The characterisation of such relationships makes it possible to draw a network of auxiliaries, verbs and compatible predicates through their usage patterns.

The most frequent word order pattern in Table 4. covers instances of grammatical focus (e.g. [/N][Acc] L [/V][Inf]), where the speaker emphasises a particular component of the event rather than the event as a whole. Negative contexts are also typical for all modal auxiliaries and pred-

icative adjectives, except for *képtelen*, which already contains the negation (the *-tAlAn* suffix, e.g. *Mari képes/képtelen aludni* 'Mary is able/unable to sleep'). classification (Imrényi, 2013; Kálmán et al., 1989). For three auxiliary verbs (*kíván*, *mer*, and *bír*), not all of the top 10 patterns were identified. For *kíván*, only one pattern was missing from the HGC2 sample, likely due to limited occurrence of special structures. A similar issue arose with *mer*, although all relevant patterns were found in the HW2 sample. From this we can conclude that the auxiliary verb *bír* is the one that actually shows distinct patterns. The verb *bír* typically occurs in contexts with negative polarity (e.g. *Nem bírja felemelni azt* 'He/She cannot lift it'), which inherently increases the number of words in the patterns. Additionally, *bír* often appears in structures with restrictive adverbs (e.g. *Alig bírja felemelni azt* 'He/She can hardly lift it'). Furthermore, this phenomenon is also evident in the fact that when certain grammatical cases of a noun are placed in the 3-grams, these patterns do not appear with the *bír*. Based on its specific patterns, it can be assumed that it is less advanced on the grammaticalisation path to a typical auxiliary verb than, for example, the more abstract auxiliary *tud* ('can'). Among the predicative adjectives (*képes*, *hajlandó* and *képtelen*), *képes* ('able to') most closely matches the patterns of general modal auxiliary structures (i.e. it has instances for most of the patterns listed). This may be functionally because it is quasi-synonymous with the auxiliary *tud* ('can') in contexts expressing ability.

It can be seen that the ten most frequent n-grams from HGC2 and HW2 differ slightly at the same length. This could be due to differences in the size of the two corpus, but it also raises further questions about the correlation between the patterns used and the text types or genres. Additionally, the still high number of variations (see Table 3.) is partly due to the flexibility of Hungarian word order and partly to the presence of idiomatic patterns as separate entries. However, this aligns with our aim of preserving structural diversity, reflecting both the variety and similarities among auxiliaries.

6 Conclusion and Future Work

We have presented the steps¹ for reducing examples in the corpus to a form a form that is nearly identi-

¹The full source code is available under copyleft licence at <https://github.com/bajzattimi/Research-of-infinitive-structures-related-to-the-modal-semantic-domain>

cal to theoretical constructions in a corpus-driven manner through a case study. The semi-automatic clause extraction and feature reduction are covered by Indig and Bajzát (2023), from which we used the samples, while the selection of the optimal level of abstraction is obtained automatically with Mosaic methods (Indig, 2017). This work focuses on the precise compression of the noun phrases (cf. NP chunking), which halved the number of the remaining candidates, yielding manually comparable results that still need further evaluation and reduction. The method is currently under development, so the latter steps are not yet in their final form.

The limitations of the compression include the handling of interrupted possessive structures. However, we have shown that their small number is more of a usage pattern than an oddity, opening up new directions for research. Free argument order poses another challenge, which we plan to address by using bag of words instead of n-grams after the compression, allowing for the classification of otherwise similar neutral and focused structures. The idiomatic structures identified (e.g. entries with specific elements kept) are not included in this evaluation and should be examined separately.

We have used rule-based components (cf. LLMs) to maintain maximum control over the workflow and to be able to examine and validate the chain of reasoning at each step in order to develop a more correct intuition from which theories can benefit. For example, the need for a rigorous formalisation of the constructions sought and the development of tools to achieve such reduction of examples. This approach has revealed unexpected behaviours (e.g. the absence of interrupted possessive structures) that would otherwise remain hidden. If studied further, they could provide more insight into the reasoning of the language users through their usage patterns and would shape theoretical thinking. On the other hand, the empirical validation of the revealed theoretical results based on human cognition is essential to support the cognitive aspect of the approach. Therefore, in the next step of the research we plan to validate the identified patterns with cloze tests in the form of a language game (Indig and Lévai, 2023)), as this approach also supports the testing of LLMs for the same task.

The proposed steps are loosely language and task dependent as they can be easily adapted to other languages and phenomena. They are tested on Hungarian, so they are particularly suitable for other Uralic and morphologically rich languages.

Acknowledgments

Supported by the EKÖP-24-3 University Research Fellowship Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund.

References

- Tímea Bajzát. 2020. de akarám kdnek is tudtára adni' - az akar, a tud és a mer segédige + főnévi igenév kompozitumszerkezetek szintaktikai vizsgálata az időbeliség perspektívájában. In Gábor Simon and Gábor Tolcsvai Nagy, editors, *Nyelvtan, diskurzus, megismerés*, pages 119–149. ELTE Eötvös Kiadó, Budapest.
- Tímea Bajzát. 2022. A premodális tartományokkal összekapcsolódó segédige/melléknév + főnévi igenév konstrukciók mondatszintű szemantikai vizsgálata. In Szilárd Tátrai and Gábor Tolcsvai Nagy, editors, *A magyar mondat és kontextuális környezete*, pages 141–194. ELTE Eötvös Kiadó, Budapest.
- Joan Bybee. 1995. *Frequency of Use and the Organization of Language*. Oxford University Press, Oxford.
- Dylan Glynn and Justyna A. Robinson, editors. 2014. *Corpus Methods for Semantics. Quantitative studies in polysemy and synonymy*. John Benjamins, Amsterdam–Philadelphia.
- Adele E. Goldberg. 1995. *Constructions: A construction grammar approach to argument*. University of Chicago Press, Chicago.
- Stefan Th. Gries and Anatol A. Stefanowitsch, editors. 2007. *Corpora in Cognitive Linguistics*. Mouton de Gruyter, Berlin.
- Bernd Heine and Heiko Narrgog. 2012. *The Oxford Handbook of Grammaticalization*. Oxford University Press, Oxford.
- András Imrényi. 2013. A beférkőző segédigés szerkezetek függőségi nyelvtani elemzéséhez. *Magyar Nyelv*, 109(3):291–308.
- Balázs Indig, Bálint Sass, Eszter Simon, Iván Mittelholcz, Noémi Vadász, and Márton Makrai. 2019. *One format to rule them all – the emtsv pipeline for Hungarian*. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 155–165, Florence, Italy. Association for Computational Linguistics.
- Balázs Indig. 2017. Mosaic n-grams: Avoiding combinatorial explosion in corpus pattern mining for agglutinative languages. In *Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 147–151, Poznań. Adam Mickiewicz University.
- Balázs Indig and Tímea Borbála Bajzát. 2023. Bags and mosaics: Semi-automatic identification of auxiliary verbal constructions for agglutinative languages. In *Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 111–116, Poznań. Adam Mickiewicz University.
- Balázs Indig and Dániel Lévai. 2023. I'm smarter than the average bert! – testing language models against humans in a word guessing game. In *Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 106–111, Poznań. Adam Mickiewicz University.
- Laura A. Janda, editor. 2013. *The Quantitative Turn: The Essential Reader*. De Gruyter Mouton, Berlin.
- Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2007. *The sketch engine: ten years on. Lexicography*, 1(1):7–36.
- C. György Kálmán, Lászkó Kálmán, Ádám Nádasdy, and Gábor Prószéky. 1989. A magyar segédigék rendszere. In Zsigmond Telegdi and Ferenc Kiefer, editors, *Általános Nyelvészeti Tanulmányok XVII. Tanulmányok a magyar mondatköréből*, pages 49–103. Akadémiai Kiadó, Budapest.
- Ronald W. Langacker. 2005. *Construction grammars: cognitive, radical, and less so*. In Francisco J. Ruiz de Mendoza Ibáñez and Sandra M. Peña-Cervel, editors, *Construction Grammars: cognitive, radical, and less so*, 32, pages 101–159. Mouton de Gruyter, Berlin–New York.
- Ronald W Langacker. 2006. Subjectification, grammaticalization, and conceptual archetypes. In Angeliki Athanasiadou, Costas Canakis, and Bert Cornillie, editors, *Subjectification. Various paths to subjectivity*, pages 49–103. Mouton de Gruyter, Berlin–New York.
- Ronald W. Langacker. 2008. *Cognitive Grammar: A Basic Introduction*. Oxford University Press, Oxford.
- Noémi Ligeti-Nagy, Andrea Dömötör, and Noémi Vadász. 2019. *What does the nom say? an algorithm for case disambiguation in Hungarian*. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 27–41, Tartu, Estonia. Association for Computational Linguistics.
- Milla Luodonpää-Manni, Esa Penttilä, and Johanna Viimaranta, editors. 2017. *Empirical Approaches to Cognitive Linguistics*. Cambridge Scholars Publishing, Cambridge.
- Bernadett Modrián Horváth. 2020. Beférkőzés és keretképzés a magyar nyelvben. In Géza Balázs, András Imrényi, and Gábor Simon, editors, *Hálózatok a nyelvben*, pages 281–296. Magyar szemiotikai társaság, Budapest.

Dávid Márk Nemeskey. 2020. *Natural language processing methods for language modeling*. Ph.D. thesis, Eötvös Loránd University.

Attila Novák, Borbála Siklósi, and Csaba Oravecz. 2016. [A new integrated open-source morphological analyzer for Hungarian](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1315–1322, Portorož, Slovenia. European Language Resources Association (ELRA).

Csaba Oravecz, Tamás Váradi, and Bálint Sass. 2014. [The Hungarian Gigaword corpus](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1719–1723, Reykjavik, Iceland. European Language Resources Association (ELRA).

An Van linden. 2010. [From premodal to modal meaning: Adjectival pathways in english](#). *Cognitive Linguistics*, 21:537–571.

A Appendix

TOP	C	N	l	kíván	mer	bír	hajl	képes	képt
[/N][Acc] L [/V][Inf]	H	35273	3	H W	H W	0 W	H W	H W	0 0
nem L [/V][Inf]	H	8683	3	H W	H W	H W	H W	H W	0 0
[/Cnj] [/V][Inf] L	H	6972	3	H W	H W	H W	0 0	0 W	0 0
[/N][Nom] L [/V][Inf]	H	6052	3	H W	H W	H W	0 W	H W	0 W
[/Prev] L [/V][Inf]	H	5197	3	H W	H W	H W	0 0	0 0	0 0
[/N][Ins] L [/V][Inf]	H	5181	3	H W	H W	0 0	0 W	H W	0 0
[/N][Subl] L [/V][Inf]	H	3936	3	H W	H W	0 0	0 0	0 W	0 0
L [/V][Inf] [/N][Acc]	H	3468	3	H W	H W	0 0	0 W	H W	0 0
[/N][Ine] L [/V][Inf]	H	2721	3	H W	0 W	0 0	0 W	H W	0 0
[/N][Nom] [/V][Inf] L	H	1768	3	H W	H W	0 0	0 0	0 0	0 0
[/N][Acc] L [/V][Inf]	W	66688	3	H W	H W	0 W	H W	H W	0 0
nem L [/V][Inf]	W	12775	3	H W	H W	H W	H W	H W	0 0
[/N][Ins] L [/V][Inf]	W	12134	3	H W	H W	0 0	0 W	H W	0 0
[/N][Nom] L [/V][Inf]	W	10435	3	H W	H W	H W	0 W	H W	0 W
[/Cnj] L [/V][Inf]	W	10202	3	H W	H W	H W	H W	H W	0 W
L [/V][Inf] [/N][Acc]	W	9349	3	H W	H W	0 0	0 W	H W	0 0
[/Cnj] [/V][Inf] L	W	8093	3	H W	H W	H W	0 0	0 W	0 0
[/N][Ine] L [/V][Inf]	W	6639	3	H W	0 W	0 0	0 W	H W	0 0
[/N][Subl] L [/V][Inf]	W	6520	3	H W	H W	0 0	0 0	0 W	0 0
[/Prev] L [/V][Inf]	W	5721	3	H W	H W	H W	0 0	0 0	0 0
[/N][Nom] nem L [/V][Inf]	H	9930	4	H W	H W	H W	H W	H W	0 0
[/Cnj] nem L [/V][Inf]	H	8323	4	H W	H W	H W	H W	H W	0 0
[/Cnj] [/N][Acc] L [/V][Inf]	H	8274	4	H W	H W	0 W	H W	H W	0 W
[/N][Acc] nem L [/V][Inf]	H	8107	4	H W	H W	H W	H W	H W	0 0
[/N][Nom] [/N][Acc] L [/V][Inf]	H	7984	4	H W	0 W	0 W	0 W	H W	0 0
[/N][Acc] [/Prev] L [/V][Inf]	H	7028	4	H W	0 W	0 W	0 0	0 0	0 0
[/Cnj] [/Prev] L [/V][Inf]	H	5742	4	H W	H W	H W	0 0	0 0	0 0
[/N][Nom] [/Prev] L [/V][Inf]	H	4410	4	H W	H W	0 W	0 0	0 0	0 0
[/Prev] L [/V][Inf] [/N][Acc]	H	4130	4	H W	H W	0 0	0 0	0 0	0 0
[/N][Nom] [/Post] L [/V][Inf]	H	3343	4	H W	0 W	0 W	0 W	H W	0 W
[/Cnj] [/N][Acc] L [/V][Inf]	W	18950	4	H W	H W	0 W	H W	H W	0 W
[/N][Nom] nem L [/V][Inf]	W	16290	4	H W	H W	H W	H W	H W	0 0
[/N][Nom] [/N][Acc] L [/V][Inf]	W	16130	4	H W	0 W	0 W	0 W	H W	0 0
[/Cnj] nem L [/V][Inf]	W	15768	4	H W	H W	H W	H W	H W	0 0
[/N][Acc] nem L [/V][Inf]	W	13869	4	H W	H W	H W	H W	H W	0 0
[/N][Acc] [/Prev] L [/V][Inf]	W	9242	4	H W	0 W	0 W	0 0	0 0	0 0
[/Cnj] [/Prev] L [/V][Inf]	W	8663	4	H W	H W	H W	0 0	0 0	0 0
nem L [/V][Inf] [/N][Acc]	W	7127	4	H W	H W	H W	H W	H W	0 0
[/Cnj] L [/V][Inf] [/N][Acc]	W	6738	4	0 W	H W	0 W	H W	H W	H W
[/N][Nom] [/Prev] L [/V][Inf]	W	5100	4	H W	H W	0 W	0 0	0 0	0 0

Table 4: The most frequent mosaic 3-gram and 4-gram types found in the samples of the two corpora. The auxiliary verb lemmas *akar*, *tud* and *szeret(ne)* were detected in all the patterns in the table and are therefore not shown individually. Abbreviations: C = the original corpus which the pattern was derived, N = Number of occurrences, l = length of the pattern, L = lemma (of the auxiliary); hajl = *hajlandó*, képt = *képtelen*; H = Hungarian Gigaword Corpus 2.0., W = Hungarian Webcorpus 2.0., 0 = Not present. Glossary: *nem* ‘not’;)

Prune or Retrain: Optimizing the Vocabulary of Multilingual Models for Estonian

Aleksei Dorkin and Taido Purason and Kairit Sirts

Institute of Computer Science

University of Tartu

{aleksei.dorkin, taido.purason, kairit.sirts}@ut.ee

Abstract

Adapting multilingual language models to specific languages can enhance both their efficiency and performance. In this study, we explore how modifying the vocabulary of a multilingual encoder model to better suit the Estonian language affects its downstream performance on the Named Entity Recognition (NER) task. The motivations for adjusting the vocabulary are twofold: practical benefits affecting the computational cost, such as reducing the input sequence length and the model size, and performance enhancements by tailoring the vocabulary to the particular language. We evaluate the effectiveness of two vocabulary adaptation approaches—retraining the tokenizer and pruning unused tokens—and assess their impact on the model’s performance, particularly after continual training. While retraining the tokenizer degraded the performance of the NER task, suggesting that longer embedding tuning might be needed, we observed no negative effects on pruning.

1 Introduction

Adapting multilingual pretrained language models to specific languages can enhance both their efficiency and performance (Kuratov and Arkhipov, 2019; Mroczkowski et al., 2021). The adaptation generally involves continuously training the full model on language-specific data. This approach can be expensive and unsuitable for less-represented languages.

In this study, we explore how modifying the vocabulary of a multilingual language model to better suit the Estonian language affects its downstream performance. Compared to the previous works (Gee et al., 2022; Csaki et al., 2024; Tejaswi et al., 2024), we focus on training newly initialized embeddings rather than the specific initialization approaches. The motivations for adjusting the vocabulary are twofold:

1) Practical Benefits: A language-specific vocabulary can reduce the length of tokenized sequences, leading to more efficient training and fine-tuning. Meanwhile, the vocabulary of a multilingual model has to accommodate for all the languages it supports, which results in a significant amount of unused tokens in a monolingual use case. Consequently, adapting the vocabulary to a single language either by pruning the tokenizer or training a new one leads to the decrease in vocabulary size. In turn, decreasing the vocabulary size reduces the overall model size, which can improve computational efficiency.

2) Performance Enhancement: Tailoring the vocabulary to a particular language may improve the model’s ability to understand and process text in that language, potentially boosting performance on language-specific tasks.

Our goal is to evaluate the effectiveness of two vocabulary adaptation approaches—retraining the tokenizer and pruning unused tokens—and assess their impact on encoder models’ performance in the Estonian language, particularly after continual training, which is evaluated by fine-tuning on the named entity recognition task. In the retraining approach, we train a new tokenizer on the Estonian National corpus (ENC),¹ and use the resulting vocabulary to replace/adapt the encoder-based multilingual DeBERTa model (He et al., 2023). We first train the new embeddings with other model parameters frozen, and then continue training the full model with the masked language modeling loss. Finally, we fine-tune the model with new/adapted vocabulary on the Estonian NER dataset (Sirts, 2023) to evaluate the effect of vocabulary optimization. In the second approach, we simply prune the model’s initial vocabulary to only keep the tokens that are

¹<https://doi.org/10.15155/3-00-0000-0000-0000-08C04M>

present in the ENC, and experiment with continuing the training of either only the embeddings or all model parameters.

2 Related Work

Adapting a tokenizer to a new domain or language has been mainly done in two ways: modifying an existing tokenizer or training a new tokenizer on the domain data. The main focus of previous works has been on the embedding initialization methods for new or extended vocabulary, which is not needed in cases of vocabulary pruning.

The vocabulary pruning has been previously explored by [Abdaoui et al. \(2020\)](#). The main motivation was that mBERT ([Devlin et al., 2019](#)), for instance, allocates more than 51% of its parameters to the embeddings layer, yet only a fraction of the vocabulary is used for a single given language. Accordingly, the proposed approach is to create monolingual models from multilingual mBERT by shrinking the vocabulary of the original model. To select the tokens to keep for a given monolingual model, the authors collected token frequency statistics from the Wikipedia of the target language, and used these frequencies to filter out the tokens that appeared in less than 0.05% lines. As a result, the monolingual models retained up to 23.8% of the original vocabulary.

Tokenizer extension for BPE models is usually done by first training a new tokenizer and then adding non-overlapping tokens to the existing tokenizer’s vocabulary ([Csaki et al., 2024](#); [Gee et al., 2022](#), etc). [Csaki et al. \(2024\)](#) investigated extending an existing tokenizer and found that a correctly implemented vocabulary extension does not negatively affect downstream performance. [Tejaswi et al. \(2024\)](#) also studied the vocabulary extension of LLMs, finding that a larger extension requires more pre-training data for optimal results.

[Gee et al. \(2022\)](#) introduced a method for fast vocabulary transfer (FVT) to train a domain-specific tokenizer. The embeddings for tokens shared between the new and original tokenizers were copied. The embeddings for new tokens were obtained by averaging their sub-token embeddings from the original tokenizer. The model was then further pre-trained on in-domain data using the masked language modeling (MLM) loss before fine-tuning on target tasks. [Dagan et al. \(2024\)](#) applied the FVT ([Gee et al., 2022](#)) for LLMs and found that the tokenizer choice impacts the effectiveness and down-

stream performance of LLMs. Specifically, they found that adapting the model to a new tokenizer requires tens of billions of tokens of retraining to outperform the original tokenizer. While our training ENC training corpus is far smaller, containing only few billions of tokens, we are experimenting with encoder models that are much smaller than LLMs.

3 Methodology

The overall methodology of optimizing a model’s vocabulary to Estonian entails two main steps: 1) modifying the content of the vocabulary and adjusting the embeddings accordingly, and 2) continuing the training of the whole model to align it better with the new vocabulary. As the base multilingual model we select mDeBERTa v3—the multilingual version of DeBERTa V3 ([He et al., 2023](#))—the model which is considered the SOTA encoder model at the time of the writing.

We compare two methods to modify the vocabulary of the mDeBERTa v3 multilingual language model for Estonian. The first method involves training a new tokenizer on the Estonian National Corpus (ENC), while in the second method we simply prune the model’s original vocabulary to remove tokens that are not used in the tokenized ENC.

Retraining the Tokenizer We retained all the original settings (such as special tokens and pre- and post-processing steps) from the base mDeBERTa v3 model and retrained the underlying SentencePiece tokenizer. For new tokens introduced by the retrained tokenizer, we initialized their embeddings using the mean of the original embedding matrix, while for tokens present in both the original and new tokenizers, we copied the existing embeddings. We also adjust the token to id mapping and resize the embedding matrix.

To align the newly initialized embedding vectors with the rest of the model, we first train the model on the training corpus with only the embeddings unfrozen using the masked language modeling (MLM).

Tokenizer Pruning The pruning process starts with applying the existing tokenizer to the training data. Then we observed what tokens in the vocabulary never appear in our training data and removed them from the vocabulary. After that the token to id mapping was adjusted and the embedding matrix was rearranged and resized. Since no new tokens

were added, we retained the original embeddings for the remaining tokens.

In our experiments, we use an approach similar to [Abdaoui et al. \(2020\)](#) with two key differences. Firstly, we do not use a frequency threshold, but rather keep all the tokens that do appear in the language-specific data. This results in our model retaining approximately 67% of the original vocabulary. Secondly, we employ a larger data source (that also includes Wikipedia)—the Estonian National Corpus. Both differences are aimed at maximizing the vocabulary coverage.

Continuous Training with LoRA To simulate continual training and assess the model’s adaptability after vocabulary modification, we applied Low-Rank Adaptation (LoRA) ([Hu et al., 2021](#)) training with MLM objective to both models.

4 Experimental Setup

Training Data For training the new tokenizer, and training and validation of the MLM objective, we employed the Estonian National Corpus (ENC).² The corpus contains approximately 16M documents with documents coming from different domains such as old and contemporary literature, academic texts, Wikipedia pages and discussions, as well as crawled web pages. We performed light deduplication on the corpus resulting in ca 3.4B tokens and randomly split it into train, validation, and test, with both validation and test sets containing 1% of the documents.

Models Developed In our experiments we employ the base version of mDeBERTa V3 as our base model, and apply the previously described approaches—tokenizer retraining and pruning—to it. For tokenizer retraining, we settle for 32K tokens in the final vocabulary, and train the new tokenizer using the train split of the ENC. Meanwhile, for pruning we collect the statistics on the appearance of tokens in the base model vocabulary in the ENC train split, then we remove all tokens that never appear in the data. This results in the vocabulary size of approximately 169K tokens. For both approaches we resize the embedding matrix and rearrange the corresponding vectors, while for tokenizer retraining we initialize the vectors that has not previously appeared in the base model vocabulary using the mean of the embedding matrix.

²<https://doi.org/10.1515/3-00-0000-0000-0000-08C04M>

The models were trained on the University High-Performance Cluster ([University of Tartu, 2018](#)) using up to two A100 80GB GPUs.

Embedding training For both approaches, we tuned the embeddings for a single epoch on sequences of 128 tokens in half-precision. The number of devices, per device batch size, and gradient accumulation steps were configured so that the global batch size was 3092. The warm-up ratio was set to 0.05.

Continuous training with LoRA Most of the training parameters remain the same for LoRA continuous training, except for the learning rate which we set to be 1e-3. The LoRA itself was configured with a rank of 4 for the update matrices, using a scaling factor (α) of 32. A dropout rate of 0.1 was applied to the LoRA layers to prevent overfitting. The adaptation was applied to the attention mechanism components (query, key, and value matrices) as well as the dense feed-forward layers. No bias parameters were updated during training.

Fine-tuning on NER Intuitively, a downstream task where the model has to produce classification scores for individual tokens in the input is affected the most by the vocabulary modification. The most common type of such task is likely Named Entity Recognition (NER). For the Estonian language EstNER ([Sirts, 2023](#)) is the most comprehensive NER dataset. It contains 46K sentences annotated with 11 entity classes. To assess the performance of the modified models, we fine-tuned models on EstNER for 50 epochs in half-precision with a global batch size of 64. For each model version we repeated the process three times and recorded the highest achieved F1 score in each run. We report the mean and the standard deviation over the three runs.

Model	# Params	Vocab Size	Tok per Word
EstBERT	124M	50K	1.82
XML-RoBERTa base	278M	250K	2.04
TartuNLP Est-RoBERTa	278M	250K	2.04
EMBEDDIA Est-RoBERTa	116M	40K	1.69
mDeBERTa base	279M	250K	2.23
mDeBERTa base Tuned	110M	32K	1.75
mDeBERTa base Pruned	215M	169K	2.23

Table 1: Statistics on vocabulary size, number of parameters, and tokens per word (estimated on the validation split of the ENC) for related models.

Model	F1 Score (Mean \pm Std)	MLM Accuracy
EstBERT	75.72 \pm 0.19	-
XLM-RoBERTa base	80.66 \pm 0.37	-
TartuNLP EstRoBERTa	81.37 \pm 0.28	-
Embeddia EstRoBERTa	83.77 \pm 0.24	-
mDeBERTa-base	80.96 \pm 0.19	-
mDeBERTa-base \rightarrow Tuned Embeddings	76.40 \pm 0.23	15.86
mDeBERTa-base \rightarrow Tuned Embeddings \rightarrow LoRA	77.58 \pm 0.47	29.74
mDeBERTa-base \rightarrow Pruned	80.62 \pm 0.12	-
mDeBERTa-base \rightarrow Pruned \rightarrow Tuned Embeddings	80.45 \pm 0.22	25.84
mDeBERTa-base \rightarrow Pruned \rightarrow LoRA	80.62 \pm 0.10	38.42

Table 2: EstNER Evaluation F1 and ENC MLM Accuracy scores (excluding baseline models).

Evaluation Metrics We evaluate the tokenization efficiency by calculating the token per word ratio for different tokenizers. We measure the performance of the models on MLM objective using word prediction accuracy. To evaluate the downstream NER task we use the F1 score.

5 Results

In addition to the DeBERTa baseline, we also compare with various other models, including both Estonian-specific EstBERT (Tanvir et al., 2020), XLM-RoBERTa base (Conneau et al., 2020), an Estonian-specific EstRoBERTa finetuned from the XLM-RoBERTa³ and another EstRoBERTa model trained from scratch.⁴

Tokenizer efficiency We first present the impact on the models’ size and tokenizer efficiency in Table 1. We observe that adopting the smaller 32K language-specific vocabulary (mDeBERTa base Tuned) leads to approximately 60% reduction in the number of parameters and 20% reduction in tokens per word. Meanwhile, simply pruning the vocabulary (mDeBERTa base Pruned) results in ca 23% reduction in the number of parameters.

Tokenizer Optimization Results The models with optimized vocabulary MLM accuracy and the downstream NER task F1-scores are shown in Table 2. The top part shows the results for the baseline mDeBERTa base and the comparison models. The baseline mDeBERTa is in line with the multilingual XLM-RoBERTa, but little bit worse than Estonian-specific RoBERTa models. The middle section of the Table 2 shows the results on the models with newly created 32K vocabulary both only after the embedding tuning and then after training

continuation with LoRA. While training continuation with LoRA substantially improves the MLM accuracy, replacing the tokenizer led to a substantial decrease in NER performance, with the model average F1 score being below both the baseline multilingual models and two out of three language specific models. The bottom section of the Table 2 shows the results for the models with vocabulary pruning. Again, continuation with the LoRA training improves the MLM accuracy, while the NER results are in the same range with the baseline. The embedding tuning and LoRA training took approximately 120 GPU hours each with the pruned model taking longer due to the large vocabulary size.

6 Discussion

While replacing the vocabulary of the mDeBERTa model with a smaller Estonian-specific vocabulary led to more efficient input tokenization, the results on the downstream NER task suffered even after both embedding layer training and subsequent full model training with LoRA. First, suboptimal embedding initialization approach likely played a role in the observed outcome. Secondly, it is likely that a single epoch of embedding tuning was insufficient to match the performance of the base model. The subsequent LoRA MLM training resulted in slightly reducing the gap between the base model and the model with the retrained tokenizer, however it also remained insufficient to recover the original model’s performance. We presume that training for longer, both the embeddings and the LoRA parameters, would further reduce that gap.

In contrast, we observe that the vocabulary pruning has no observable negative effect on the downstream task. Meanwhile, tuning of the embeddings appears to have little to no effect on the downstream task, which suggests that such tuning is redundant

³<https://huggingface.co/tartuNLP/EstRoBERTa>

⁴<https://huggingface.co/EMBEDDIA/est-roberta>

in case of pruning. Surprisingly and similarly to embeddings tuning, continued training with LoRA had no observable benefit for the pruned model, despite the gains in the MLM accuracy.

Finally, we observed that the relation between the MLM accuracy and F1 on NER is not transparent. While we acknowledge that MLM accuracy scores with different vocabularies are not directly comparable, the absence of the effect on the NER result in the presence of a notable improvement in the MLM accuracy in the pruned model is puzzling.

7 Conclusion

In this study, we explored two options for optimizing the vocabulary of a multi-lingual model for the Estonian language. In summary, we found that replacing the tokenizer with a retrained language-specific version noticeably degrades model performance on the downstream NER task, and one epoch of embedding layer training on a 3.4B word corpus did not suffice to restore it. While LoRA offers efficient way for further training continuation, a single epoch was insufficient to mitigate the negative impact of the tokenizer replacement. On the other hand, pruning unused tokens proved to be an effective method to reduce vocabulary size without compromising performance.

Acknowledgments

This research was supported by the Estonian Research Council Grant PSG721 and Estonian Language Technology Program Grant EKTB104.

References

- Amine Abdaoui, Camille Pradel, and Grégoire Sigel. 2020. [Load What You Need: Smaller Versions of Multilingual BERT](#). In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 119–123, Online. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised Cross-lingual Representation Learning at Scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Zoltan Csaki, Bo Li, Jonathan Li, Qiantong Xu, Pian Pawakapan, Leon Zhang, Yun Du, Hengyu Zhao, Changran Hu, and Urmish Thakker. 2024. [Sambalingo: Teaching large language models new languages](#). *arXiv preprint arXiv:2404.05829*.
- Gautier Dagan, Gabriel Synnaeve, and Baptiste Roziere. 2024. [Getting the most out of your tokenizer for pre-training and domain adaptation](#). In *Forty-first International Conference on Machine Learning*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](#). 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Leonidas Gee, Andrea Zugarini, Leonardo Rigutini, and Paolo Torrioni. 2022. [Fast Vocabulary Transfer for Language Model Compression](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 409–416, Abu Dhabi, UAE. Association for Computational Linguistics.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing](#). *Preprint*, arXiv:2111.09543.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *Preprint*, arXiv:2106.09685.
- Yuri Kuratov and Mikhail Arhipov. 2019. [Adaptation of deep bidirectional multilingual transformers for Russian language](#). *arXiv preprint arXiv:1905.07213*.
- Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. [HerBERT: Efficiently Pretrained Transformer-based Language model for Polish](#). In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10.
- Kairit Sirts. 2023. [Estonian Named Entity Recognition: New Datasets and Models](#). In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 752–761, Tórshavn, Faroe Islands. University of Tartu Library.
- Hasan Tanvir, Claudia Kittask, and Kairit Sirts. 2020. [EstBERT: A Pretrained Language-Specific BERT for Estonian](#). *Preprint*, arXiv:2011.04784.
- Atula Tejaswi, Nilesch Gupta, and Eunsol Choi. 2024. [Exploring Design Choices for Building Language-Specific LLMs](#). *arXiv preprint arXiv:2406.14670*.
- University of Tartu. 2018. [UT Rocket](#).

On the Role of New Technologies in the Documentation and Revitalization of Uralic Languages of Russia in Historical and Contemporary Contexts

Alexander Nazarenko

Independent researcher and enthusiast of Uralic languages,
amateur database and software developer
aleksanterinazarenko@gmail.com

Abstract

The Uralic languages spoken in Russia face significant challenges due to historical and sociopolitical factors, resulting in their endangered status. While only Finnish, Estonian, and Hungarian enjoy solid support as official languages, most Uralic languages suffer from limited resources and declining speaker populations. This paper examines the development of written Uralic languages, the impact of Russian language and its writing system to them, and the consequences of the lack of state interest in these languages for preservation efforts. Despite these challenges, technological advancements present valuable opportunities for revitalization. Existing projects, such as dictionaries and language corpora, highlight both the potential and shortcomings of current linguistic resources. Innovative approaches, including AI-based applications and user-driven platforms, can enhance engagement among people. By emphasizing the importance of high-quality linguistic data, this study advocates for a more proactive and collaborative effort in the preservation and promotion of Uralic languages.

1 Outline of the Problem

Only three Uralic language-speaking nations have succeeded in establishing their own states where the Uralic language has official status and is utilized in all aspects of life. The emergence of autonomous and independent political entities facilitated the development of sophisticated literary languages and supported the establishment of strong national identities and diverse cultures. Regrettably, the fate of other Uralic peoples has been less fortunate.

The earliest texts written in Finnish and Estonian are relatively late compared to many European languages; moreover, they are not significantly later than the early texts of Uralic languages spoken in Russia. It is known that the Komi people had their own writing system, *Važ Perym gižöm*, as early as the 15th century, which was used until the 18th

century. The Old Permic script, like Cyrillic, is not entirely original; it incorporates features from Cyrillic, Greek, and Komi tamga signs, which are ultimately of Turkic descent. However, it emerged at a time when, according to current knowledge, no other Uralic peoples, except for the Hungarians, had any writing system. The first known sentence in the Baltic-Finnic languages was written in Cyrillic in the 13th century (*Birch bark letter no. 292*), while subsequent sentences date back to the second half of the 15th century and were written using the Latin script.

In most cases, the formation and development of written Uralic languages and their writing systems can be attributed to the expansion of various branches of Christianity, and their nature is closely tied to the church. Nevertheless, the influence of primary languages, such as Russian, and their characteristics, including writing and phonetic systems, was often more substantial in later stages. Essentially, most modern spellings of Uralic languages are based on the Russian variant of the Cyrillic script, which was designed and adapted to meet the needs of Russian and other Slavic languages but may not necessarily fulfill the requirements of the Uralic languages. Let's take a closer look at some of these impacts.

One example of this is the modern Moksha spelling, where the sound /ə/ can be represented by the letters a, o, e and the sound /æ/ can be represented by the letters я, e, э, depending on various factors. This variability can cause problems even for people who are familiar with the language. Another example is the Erzya sound /æ/, which has been entirely eliminated from the standard language, despite its presence in many dialects where it serves to differentiate meanings. For instance, /'kədʲ/ ('hand') (< Proto-Mordvinic *käd') and /'kædʲ/ ('skin') (< Proto-Mordvinic *ked'). In the modern language created in 1922, the word кеДь has been used for both meanings, likely due

to the desire to simplify the language norms and maximize their alignment with Russian language standards. It is also known that, with the Bolshevik takeover, the idea of creating a common literary language for the Erzya and Moksha people was conceived, which ultimately ended in failure, as these two languages are not mutually intelligible and do not even form a clear continuum.

Texts composed in the 19th century and earlier display a relatively high level of linguistic quality, which is due to the fact that their authors were proficient in the languages, and the influence of Russian, particularly on grammar, was comparatively moderate during this period. The writing systems of that time vary, yet it is generally evident that an effort was made to create Cyrillic-based systems that prioritized phonetic accuracy. Compare the first three verses of Nikolai Barsov's Moksha translation of the Gospel of John from 1901:

1. Первай ульсь Вал, Валськæ Шкайсълъ, Шкайсъкæ Валълъ.
2. Сон первай кигæ Шкайсълъ.
3. Сонь вельденза сембæ ушъдъзь улемаснън, а Соньфтемънза улемаснън ущъдыхънень ёткста мезямътка исъть ушъдуфт.

with the same fragment, translated by Institute for Bible translation Helsinki, in 2003:

1. Ушетксса ульсь Валсь, Валсь Шкайтъ мархтоль, и Валсь ульсь Шкай.
2. Сон ушеткссостокиге Шкайтъ мархтоль.
3. Сембось вельденза тиевсь, Соньфтомонза мезевок ашезъ тиев.

The system used by Barsov has a distinct letter æ for the sound /æ/ and consistently employs the letter ъ to represent the sound /ə/ in all cases, ensuring that each sound corresponds to a specific letter. In contrast, in the modern language, the letter ъ no longer serves that function. The 1993 language reform was intended to reintroduce its use at the beginning and in the first syllable of words, for example, вѣрѣгаз /vɛrʲ'gʌz/ ('wolf') (< Proto-Mordvinic *vɛrgas) and тѣрѣва /tɛ'rʲva/ (< Proto-Mordvinic *tɛrva) instead of вѣрѣгаз, тѣрѣва, but it was rejected.

There were also quite a few attempts to create a Latin script for Uralic languages, which are now written exclusively in Cyrillic. An Estonian linguist, Ferdinand Johann Wiedemann, possibly drawing on the earlier work of Hans Conon von der Gabelentz, *Versuch einer Mordwinischen Grammatik*, published in 1839, used a Latin writing system for the Erzya language in his work *Grammatik Der Erza-Mordwinischen Sprache* and in his transliteration of the *Gospel of Matthew* (*Das Evangelium des Matthäus ersamordwinisch*) from 1865. Although this system is somewhat irregular, it represents a clear initiative to establish a Latin script for the language. Below is a short example from the 22nd chapter of the Gospel:

37. Jisus jovtaž tenze: vetškik es pazot vese sädeiset i vese oimset i vese prävset toít.
38. Te uli ikel'tse i vesemedede pokš zapoved.
39. Ombotse že teń kond'amo: vetškik es malavikset koda es práat.

Note that *Novum Testamentum Mordvinice litt. cyrill.* from 1821, which served as the basis for Wiedemann's adaptation, contains some characteristics of the ä-dialects.

In spite of these efforts, all literary Uralic languages of Russia and their respective writing systems were established only during the Soviet era. Many native words and structures, as well as widely occurring dialectal features, were discarded. Simultaneously, a vast number of words and structures were borrowed from Russian, completely overlooking the possibility of creating new words based on existing ones, as was done in Finnish in the 18th and 19th centuries. Compulsory education was also introduced, delivered almost entirely in Russian, apart from a brief period of "Korenizatsiia" in the 1920s and early 1930s. The introduction and implementation of new writing systems, particularly those based on the Latin alphabet, were abandoned in the 1930s and have been legally prohibited since the early 2000s.

Currently, all Uralic languages, except for Finnish, Estonian, and Hungarian, are considered endangered, and the speaker populations in Russia are experiencing a dramatic decline each year. The actual situation may be even more concerning than the statistics suggest. Proficiency in these languages among individual minority groups is low

and continues to decline. The ongoing advancement of technology and media has further diminished the role of smaller Uralic languages and contributed to the deeper assimilation of their speakers. Most users of these languages are older individuals who may not be well-versed in modern technology, which somewhat slows the process of assimilation, but conversely, means that the limited language resources available do not effectively meet their needs. Younger generations, despite having easier access to technology, often show little interest in these languages, and existing solutions fail to counter this trend.

However, technological evolution presents a valuable opportunity to address the challenges faced by these languages. The collection, organization, and visual presentation of high-quality linguistic data would not only allow enthusiastic individuals to learn and study independently but also bolster the efforts of scientists, researchers, teachers, and activists concerned about the future of endangered languages, thereby raising public awareness of their fate.

2 Main challenges in Light of Existing Solutions

The main issue in the collection and digitization of data related to smaller Uralic languages lies in its significant dispersion and inconsistency, along with the limited quantity and quality of available materials. Therefore, existing databases frequently suffer from inaccuracies, incompleteness, and a lack of refinement that makes them less suitable for interactive language projects. A constructive approach could involve enhancing and refining these databases through qualitative improvements, such as incorporating native vocabulary and grammatical structures. This would include integrating archaic and less commonly used terms, creating new words based on the existing linguistic material, and minimizing reliance on Russian loanwords whenever possible. Sometimes it leads to a dilemma between prioritizing the ‘accuracy’ of language data and the mostly subjective concept of language purity, which might not always align with the preferences of native speakers.

Another issue is that the materials essential for learning and researching Uralic languages are primarily in Russian, which makes them hard to analyze directly for non-Russian speakers. The quantity of materials available in English and the three

major Uralic languages is decidedly insufficient, likely due to the fact that translation and direct data collection is time- and money-consuming.

Despite these challenges, several remarkable projects have emerged, including:

- dictionaries created by The Institute of the Estonian Language (*Eesti Keele Instituut*)¹
- the Giellatekno dictionaries and Oahpa! tools managed at UiT The Arctic University of Norway (*UiT Norges arktiske universitet*)²
- Korp and other text corpora available in The Language Bank of Finland (*Kielipankki*)³
- dictionaries and materials created by the Institute for the Languages of Finland (*Kotimaisen kielten keskus, Kotus*)⁴
- Udmurt and Komi languages in Google Translate⁵

This list is not exhaustive; however, it effectively illustrates the landscape. Let’s take a brief look at some of the projects mentioned above. The dictionaries presented by The Institute of the Estonian Language are notable for their relatively extensive vocabulary, numerous examples, and a clear effort to find suitable equivalents for terms missing in the target languages. This was achieved in part by assigning new meanings to words with closely related meanings, creating calques, and adding “descriptive equivalents”, such as the Erzya word ардомапель (‘vehicle’), derived from ардомс (‘to go, travel’) and the suffix -пель (used to form object names). While the dictionaries remain a highly reliable and innovative resource, their audience is understandably restricted to Estonian speakers.

The absence of corresponding dictionaries in the reverse direction, the lack of direct references to sources and literary examples, and the omission of transliteration are common issues found in many online dictionaries. The dictionaries available on the UiT The Arctic University of Norway website, while allowing bidirectional translations and containing an impressive amount of collected material and fairly extensive grammatical data, almost entirely lack usage examples for individual words

¹<https://eki.ee/keeleinfo/sonastikud/>

²<https://dicts.uit.no/>

³<https://www.kielipankki.fi/korp/>

⁴<https://www.kotus.fi/sanakirjat>

⁵<https://translate.google.com/>

in the English section, and some translations may even appear rather unusual. On a positive note, some provide very interesting alternatives to the increasingly common Russicisms, even though they lack any confirmation in literature.

It is necessary to emphasize that citing sources is absolutely critical in the development of linguistic resources. Providing information about the authors of cited language data and innovations, such as neologisms, enhances the reliability of the documented information.

The role of language corpora in the documentation of endangered languages is indispensable, but they are also invaluable in the creation of language projects, such as online dictionaries and learning tools. According to the list on the FID FINNUG site⁶, the Korp platform on the Language Bank of Finland website is the only tool that allows for the simultaneous display of the same texts in multiple Uralic languages, thus considerably facilitating their interpretation and comparative analysis. The size of the database and the number of available languages are distinctly unique within their field.

The support of Udmurt and Komi on Google Translate deserves special mention. Although the quality of the translation is not perfect (e.g. the phrase *Good night!* is translated into Udmurt as Бур уен! instead of the correct Ёеч кӧл! or Ёеч кӧлӧ!), this is undoubtedly a significant step towards promoting these languages. Hopefully, in time, translation into other Uralic languages will be launched, with a particular emphasis on restoring their original forms.

For the purpose of comparison, the following is a list of several notable projects that have been developed in Russia:

- MarlaMuter Mari-Russian and Erzya-Russian dictionaries⁷
- FU-Lab dictionaries, primarily focused on Permic and Mari languages⁸
- Sámi dictionaries, Saamskije slovari (Саамские словари)⁹
- the Open corpus of Veps and Karelian languages VepKar¹⁰

⁶<https://fid.finnug.de/en/language-corpora/>

⁷<https://marlamuter.com/en/>

⁸<https://dict.fu-lab.ru/>

⁹<https://slovari.saami.su/>

¹⁰<http://dictorpus.krc.karelia.ru/en>

- the Erzya corpus¹¹ and the Erzya social media corpus¹²
- the Moksha corpus¹³
- the National Corpus of the Udmurt Language (Национальный корпус удмуртского языка) with a dictionary¹⁴
- the LANGO.TO translator, which supports Erzya, Finnish and Estonian languages¹⁵

Comparing projects developed outside of Russia with those emerging within the country, it is regrettable to acknowledge that, in certain respects, the former demonstrate higher quality. This disparity is not surprising, as these projects often rely on existing works in Russian and are, in fact, digitized versions of books without any modifications. For example, MarlaMuter includes five digitized Mari dictionaries, an Erzya-Russian dictionary, and offers very useful features such as the ability to report typographical errors and buttons corresponding to letters with diacritical marks not present in the Russian alphabet. Additionally, it provides interfaces in both English and Russian.

The FU-Lab website contains 42 digitized dictionaries primarily focusing on Permic languages. Despite the extensive amount of gathered data, has a somewhat chaotic structure and lacks an interface in any language other than Russian, presenting an additional obstacle for individuals outside of Russia wish to study these languages. A similar issue is found with the Sámi dictionaries website.

VepKar, or the Open Corpus of Veps and Karelian Languages, is an example of a well-constructed website with many valuable materials, such as a speech corpus, an audio map with recordings, and a corpus-based dictionary that provides information about the specific region of Karelia from which each word originates, along with grammatical categories and relevant examples accompanied by Russian translations.

One of the advantages of the Erzya language corpora site is its capability for automatic transliteration of text according to the Uralic Phonetic Alphabet. It contains an extensive collection of

¹¹https://erzya.web-corpora.net/index_en.html

¹²https://erzya.web-corpora.net/erzya_social_media

¹³https://moksha.web-corpora.net/index_en.html

¹⁴<https://udmcorpus.udman.ru/>

¹⁵<https://lango.to/>

linguistic material, including, importantly, examples of colloquial language used in contexts such as online forums. It also includes translations for most terms in Russian.

The site of National Corpus of the Udmurt Language, in turn, includes an autonomous Russian-Udmurt and Udmurt-Russian dictionary, featuring usage examples and some audio recordings of pronunciations. This addition certainly enhances the usability of the corpus. A drawback is once again the lack of transcription and an English interface.

Finally, LANGO.TO offers an effective AI-based translator for the Erzya language and several other non-Uralic minority languages of Russia. It would not be overstatement to say that this represents one of the more intriguing initiatives of recent years, as AI has not been widely applied to the revitalization of endangered Uralic languages. The accuracy of translations between Russian and Erzya is quite impressive, especially considering the limited resources and the relatively undeveloped state of the language. In addition to Russian, it also supports Finnish and Estonian languages.

3 Summary and Example Solutions

Projects aimed at documenting and revitalizing endangered languages should, on one hand, include as much data as possible and reference specific sources, while, on the other hand, analyzing this data in terms of its quality and usefulness, and supplementing it with new information, such as grammatical categories, inflection, and usage examples. Websites and applications should feature a simple and accessible interface, offer multiple language versions, provide translations into English and major Uralic languages, and include transliterations or phonetic transcription for languages using the Cyrillic alphabet. Pronunciation recordings are invaluable for preserving the original pronunciation of the languages. Introducing new solutions, including experimental ones, with a particular emphasis on AI, is essential. At the same time, it is worth exploring how existing language corpora and “raw” lexical databases, such as those available on the Giellatekno Webdict¹⁶, can be effectively utilized.

The implementation of the data does not have to be a difficult task, as demonstrated by my website Learn Erzya¹⁷, where I utilized some databases

from the aforementioned Giellatekno for testing purposes. I also incorporated an alternative Latin spelling of the Erzya language, as presented in the book by linguists Ksenija Djordjević and Jean-Léo Léonard, *Parlons mordve: erzya et mokša*, with minor modifications. The dictionaries feature a switch between the Latin spelling and the original Cyrillic spelling. Ultimately, I intend to replace these with databases containing verified data and additional elements such as phonetic transcription, transliteration, grammatical categories, automatic inflection, common phrases, usage examples, and examples from literature, possibly sourced directly from corpora.

The experimental transliteration of the databases from Giellatekno was carried out using a transliteration tool¹⁸ that was partly based on the one the transliteration modules used in Wiktionary. The optimization of the code was facilitated largely by ChatGPT. It also has a phonetic variant¹⁹.

One more initiative underway is creating language maps. Currently, there are three simple websites featuring maps that display specific words in various European²⁰, Uralic²¹, and Mordvinic languages²². The latter pulls data directly from Wikisource. Moreover, it would be an exciting prospect to create a similar map using data from *The Dialect Dictionary of the Mordvin Languages Based on the Heikki Paasonen Materials*.

Another interesting option is creating open databases using MediaWiki.²³

In summary, although the situation of the smaller Uralic languages is, to put it mildly, far from ideal, we are equipped with tools today that offer us almost limitless possibilities. The accumulation of accumulated knowledge and technological resources at our disposal is unprecedented, yet much of its potential remains untapped. Many sources are awaiting digitization and thorough analysis, without which the development of interactive tools is not possible. Of course, this is also a matter of

learnerzya/

¹⁸<https://aleksanterinazarenko.github.io/learnerzya/transliteration-tool.html>

¹⁹<https://aleksanterinazarenko.github.io/transliterator/>

²⁰<https://aleksanterinazarenko.github.io/interactivemap-europe/>

²¹<https://aleksanterinazarenko.github.io/interactivemap-uralic/>

²²<https://aleksanterinazarenko.github.io/interactivemap-mordvinic-wiktionary/>

²³https://uralowiki.unaux.com/index.php/Main_Page?i=1

¹⁶<https://gtweb.uit.no/webdict/>

¹⁷<https://aleksanterinazarenko.github.io/>

funding, which is allocated to these goals in a very limited capacity, shifting the entire burden onto enthusiasts and amateurs, along with their financial and time constraints. This has a direct impact on the results. Nevertheless, even this barrier can be overcome if social awareness and engagement are increased, and the only way to achieve this is by providing concrete, ready-made solutions. What has contributed to the decline of the smaller Uralic languages should be used as an instrument for their revitalization. The role of new technology in this process is not only underestimated but is also absolutely crucial.

Acknowledgments

Acknowledgements to Jack Rueter for his valuable comments on the article.

References

1821. *Novum Testamentum Mordvinice litt. cyrill.* Soc. bibl. Russica, Saint Petersburg.
2002. Law on the unified graphic basis. Federal Law of the Russian Federation No. 165-ФЗ, enacted on 11.12.2002.
2016. *The New Testament in the Mordvin-Moksha language.* Institute for Bible translation Helsinki, Helsinki.
- N. P. Barsov. 1901. *Ot Ioanna svjatoe evangelie.* Helsinki.
- A. P. Feoktistov. 2008. *Očerki po istorii formirovanija mordovskix pis'menno-literaturnyx jazykov.* Saransk.
- Herr Conon Gabelentz. 1839. Versuch einer mordwinischen grammatik. *Zeitschrift für die Kunde des Morgenlandes*, 2:235–419.
- J. S. Jelisejev. 1961. Vanhin itämerensuomalainen kielimuistomerkki. *Virittäjä-lehti*, 65(1):134.
- A. N. Kelina. 2003. *Mordovija. Ènciklopedija*, volume 1. Mordovskoe kn. izd-vo, Saransk.
- A.N. Kelina and O.E. Poljakov. 2024. *Orfoèpičeskij slovar' mokšanskogo jazyka.* Saransk.
- László Keresztes. 1986. *Geschichte Des Mordwinischen Konsonantismus.* Szeged.
- E.A. Kosminskij. 1943. *Srednej pingetnen' istorijas'.* Mordovskoj gosudarstvennoj izdatel'stvas', Moscow.
- V. I. Lytkin. 1952. *Drevnepermiskij jazyk: čtenie tekstov, grammatika, slovar'.* Moscow.
- Rein Taagepera. 2013. *The Finno-Ugric Republics and the Russian State.* Routledge, New York.
- Ferdinand Johann Wiedemann. 1865a. *Gospel of Matthew (Das Evangelium des Matthäus ersamordwinisch).* London.
- Ferdinand Johann Wiedemann. 1865b. Grammatik der ersa-mordwinischen sprache: nebst einem kleinen mordwinisch-deutschen und deutsch-mordwinischen wörterbuch. *Impériale des Sciences de St.-Pétersbourg: VIIe Série.*
- B. Š. Zaguljaeva. 1991. *Russko-udmurtiskij razgovornik.* Udmurtija, Izhevsk.

Applying the transformer architecture on the task of headline selection for Finnish news texts

Maria Adamova

St Petersburg State University
Universitetskaya emb., 7-9-11
199034 St Petersburg, Russia
mariaa.gorokhova@ya.ru

Maria Khokhlova

St Petersburg State University
Universitetskaya emb., 7-9-11
199034 St Petersburg, Russia
m.khokhlova@spbu.ru

Abstract

The paper evaluates the possibilities of using transformer architecture in creating headlines for news texts in Finnish. The authors statistically analyse the original and generated headlines according to three criteria: informativeness, relevance and impact. The study also substantiates for the first time the effectiveness of a fine-tuned text-to-text transfer transformer model within the task of generating headlines for news articles in Finnish. The results show that there is no statistically significant difference between the scores obtained by the original and generated headlines on the mentioned criteria of informativeness, relevance and impact.

1 Introduction

The headline of any text plays one of the most important roles. Today, online media are ahead of their paper counterparts in terms of popularity, timeliness and mass appeal. The diversity of Internet media emphasises the importance of the task of creating unique and attractive headlines. More and more attention is paid to search engine optimisation, the main goal of which is to attract new users and increase website traffic. Since electronic media is a sphere of accumulation of huge arrays of text data, the question of its optimisation and automation is among the problems of modern computational linguistics. Creating headlines specially adapted to search engines can help increase the visibility of news articles and, consequently, increase traffic. Moreover, creating headlines manually is a labour-intensive and time-consuming process, which does not meet the requirements of the responsiveness of modern electronic media.

The aim of this paper is to critically evaluate the effectiveness of transformer architecture in creating headlines for news texts in Finnish. The paper statistically analyses original and generated headlines according to three criteria: informativeness, relevance and impact. It is also the first attempt to substantiate the effectiveness of a fine-tuned Text-to-Text Transfer Transformer model in the task of generating headlines for news articles in Finnish.

2 Related Work

The popularity of transformers has led to the development of a wide range of pre-trained language models based on this architecture. Transformers' strengths, such as its deep understanding of natural language and its ability to focus on particularly meaningful information in input data, have made models based on it effective tools, including for headers generation. Applying fine-tuning to pre-trained models on specific tasks allows the knowledge gained from pre-training to be extended to a new task.

The results of fine-tuning pre-trained models to generate news headlines in Russian were demonstrated in (Bukhtiyarov & Gusev, 2020) using the mBART and BertSumAbs models as examples. The Multilingual BART model (mBART) is a language model based on the transformer architecture, consisting of an encoder and an autoregressive decoder, and pre-trained on large-scale monolingual corpora covering 25 languages (Liu et al., 2020). The training process is built on reconstructing a document from its noisy version, which leads to significant improvements in machine translation quality at both sentence level and whole document level (ibid).

The BertSumAbs model uses Bidirectional Encoder Representations from Transformers (BERT) as the encoder and a randomly initialised 6-layer transformer as the decoder (Liu & Lapata, 2019). The decoder is pre-trained and the decoder is trained from scratch, so the tuning may be unstable. To overcome the mismatch, the optimisers of the encoder and decoder are separated (ibid).

The fine-tuning of the selected models led to a significant improvement in the results: the ROUGE metric scores increased on average by 2.9 points compared to the previous leading edge performance of the Phrase-Based Attentional Transformer model (Sokolov, 2019) and by 2.0 points compared to CopyNet (Gusev, 2019). The performance of BertSumAbs was also shown to be higher within the task of generating headlines in Russian; moreover, the BertSumAbs model produced headlines of a more abstract nature, while mBART was more prone to copying (Bukhtiyarov & Gusev, 2020). Human evaluation confirmed the effectiveness of the developed models: headlines produced by BertSumAbs were selected by five or more experts in 32% of cases, while original headlines were selected by five or more experts in 28% of cases (ibid).

In (Koppatz et al., 2022), the authors discuss Finnish news headline generation using GPT-2. The key issue in developing the model was the possibility of using the proposed system as an auxiliary tool in real journalistic practice. The generated headlines were expertly evaluated by journalists from a Finnish publishing house. As a result, the generated headlines were very close to being practically usable, and although the concrete implementation is not yet ready to become a fully automated headline generation system, as it still needs human control, the algorithm may well be applicable to potential needs.

3 Models

As part of our work, it was decided to use fine-tuning of the pre-trained model. During the training, experiments were conducted on different variations of the GPT-2 and T5 models for Finnish, provided by the Finnish-NLP community and publicly available on Hugging Face (Finnish-NLP, Hugging Face). In the first stages, both GPT-2 and T5 demonstrated the ability to generate coherent results for the task at hand, but

in order to increase the novelty of the study it was decided to continue working with the T5 models.

T5 (Text-to-Text Transfer Transformer) is a model introduced in 2020 by the Google AI team, which has fundamentally the same standard structure as the original transformer, consisting of 12 pairs of encoder-decoder blocks, using the self-awareness mechanism, direct communication network and encoder-decoder attention described in (Raffel, 2020: 11). One of the characteristics of the structure of T5 models is the use of relative scalar embeddings, which are a type of positional coding. Unlike absolute positional encodings, which assign a unique encoding to each position in a sequence, relative positional encodings encode the relative position between two tokens. To do this, the difference between the positions of the two tokens is computed, which is later used to compute a scalar value that is added to the embedding of the token.

The language model for Finnish, *Finnish T5*, was pre-trained on a combination of 6 datasets: the Finnish language subset of the mC4 dataset and Wikipedia, Yle Finnish News Archive 2011-2018 and 2019-2020, Finnish News Agency Archive (STT) and The Suomi24 Sentences Corpus. The raw datasets were automatically cleaned to improve quality and weed out non-Finnish examples. The result was a raw dataset containing about 76 GB of text. The texts were not lower-cased, so the model retained case sensitivity. A masked language modelling (*MLM*) task was used in the pre-training process.

All variations of the T5 model presented by the Finnish-NLP community were studied and the following ones were selected as a result of preparatory experiments:

1. t5-large-nl36-finnish
2. t5-mini-nl8-finnish
3. byt5-base-finnish

ByT5 is an extension of the T5 model. Compared to other known models (*BERT*, *T5*, *GPT*) that rely on learned vocabulary, the ByT5 model works with UTF-8 bytes, thus eliminating the need for text preprocessing. The underlying assumption is that text data is typically stored as a sequence of bytes, the passing of which to the model will allow arbitrary text sequences to be processed (Xue, 2022: 291). In the context of the ByT5 model, Text-to-Text framework reduces to the task of generating a byte sequence based on

Model	Parameters	Number of layers	Dimension of embedding vector (output vector of transformers block)	Dimension of intermediate vector within transformer block (size of feed-forward projection matrix)	Dimension of key/value projection matrix	Number of attention heads
t5-mini	72M	8	384	1536	64	8
byt5-base	582M	18	1536	3968	64	12
t5-large	1425M	36	1024	4096	64	16

Table 1: Main characteristics of the selected models.

some input bytes. Models of this kind are more robust to the presence of noise in the data, since they do not depend on the preprocessing step and are also free from the problems that arise when processing words that are absent in the dictionary. In addition, the parameters that account for the word matrix in large dictionary models can be allocated in byte models for other purposes, such as increasing the number of layers. The characteristics of the models are summarised in Table 1.

4 Methodology

4.1 Data collection and preprocessing

A key element in the fine-tuning process is collecting and organising the data needed to train the model to perform a new task. The study collected a relatively small corpus of 1,600 examples of news text-headline pairs. The news text was not represented in all cases by the whole news article; in most cases the first paragraphs were extracted.

Six online sources of Finnish-language news were selected for data collection (due to their popularity and accessibility):

1. selkosanomat.fi;
2. aamulehti.fi;
3. suomenutiset.fi;
4. iltalehti.fi;

5. is.fi;

6. mtvuutiset.fi.

The corpus included articles from 2021 to 2024 on various topics. The principle of thematic division did not coincide in all cases in different sources. From each resource we selected from 1 to 6 of the most widely presented thematic headings, for which we then randomly selected about 70 articles. The exceptions were the section on culture (Kulttuuri) from suomenutiset.fi with 51 articles, as there were no more news items in this category, and the political section (Politiikka) from mtvuutiset.fi with 79 articles, as it was necessary to complete the planned number of examples.

Pre-processing of texts included removing paragraph and line breaks, adding the prefix ‘header:’ before the text of the news article, putting dots in headings to indicate the end of a sentence, if necessary.

The text of the news article did not include technical information such as captions to illustrations, information about the author and/or hero of the article, date of publication, subject tags and links. It is typical for aamulehti.fi and is.fi to put the first word of the first paragraph of the article in upper case. Such cases were brought to the standard spelling of words in a sentence. The whole text was not reduced to a single lower case, as the model is case sensitive as a result of pre-training.

As a result, a table in the csv format was generated, consisting of two columns: ‘input’, which contains the news article excerpt with the task prefix, and ‘target’, which contains the corresponding headline with a full stop or other terminating punctuation at the end of it.

4.2 Evaluation

The ROUGE series of metrics is considered to be the baseline for evaluating the performance of the text summarisation task (Maples, 2017: 2), so the metrics are reported in this study. The ROUGE metrics package is based on counting the number of matched units in human-generated and generated texts. The counted units are combinations of n words and the longest matched word sequence (Cohan & Goharian, 2016: 807). ROUGE is calculated as follows:

$$ROUGE[n] - recall = \frac{N_{grampred} \cap N_{gramref}}{N_{gramref}}$$

$$ROUGE[n] - precision = \frac{N_{grampred} \cap N_{gramref}}{N_{grampred}}$$

$$ROUGE[n] - F1 = 2 \times \frac{recall \times precision}{recall + precision}$$

where $N_{gram\ pred.}$ — n-grams in the generated text;
 $N_{gram\ ref.}$ — n-grams in the original text.

We used the *Rouge* package to calculate the scores for pairs of original and generated titles and the *FilesRouge* package to calculate the mean when comparing all original and all generated titles.

The ROUGE metric is easy to compute and versatile as it can be applied to data in any language. However, ROUGE considers n-gram matches without considering semantics and grammar, so the results of such metrics cannot give a complete picture of the suitability of the generated materials.

4.3 Questionnaire parameters

There are subjective indicators that are particularly important when it comes to the quality of headlines. The key characteristics of headlines are informativeness, which is not always expressed by the number of N-grams matched, and potential attractiveness, which is entirely based on human perception. For these reasons, human judgement is still considered to be the most reliable way to assess the quality of the generated text. So in this paper, the main reference point for assessing the quality of the model’s performance is the results of an expert questionnaire.

The experts were non-first year undergraduate students with a relevant major in linguistics or philology, on the basis of which an assumption is made about the adequacy of their Finnish language proficiency. In the questionnaire given to the participants, they are first asked to read an excerpt of a news article of the same size that the model receives as input, then each of a pair of headlines (original and generated) is evaluated according to three criteria: *informativeness*, *relevance* and *impact*. It is assumed that these are the parameters that are fundamental to the

creation of quality headlines and cannot be reliably assessed using metrics.

The criterion of informativeness refers to the extent to which the headline reflects the content of the news article. The informativeness of the headline is considered as “the ability to provide the reader with a relatively adequate representation of the main topic or idea” of the text (Chekut’, 2015: 81).

The relevance criterion is understood as the degree of actual correspondence of the proposed headline to the content of the text.

The criterion of impact implies the degree of produced emotional impact on the reader, the degree of “attracting the reader’s attention to the subject of the message” (ibid: 82).

To ensure unbiasedness, no indication of which headline was composed by a human and which was suggested by the model is provided before the questionnaire is run. Evaluating each headline against multiple criteria seems more appropriate as it allows for a more nuanced comparison and indicates areas for further work to improve the model. Each headline for each criterion is scored on a five-point scale, where

- 1 - not expressed at all;
- 2 - insufficiently expressed;
- 3 - weakly expressed;
- 4 - well expressed;
- 5 - strongly expressed.

5 Experiments

For the experiments, it was decided to select 10 non-training articles from each of the six sources that were included in the fine-tuning corpus. 10 news articles from the news resource Yle were also added, as this company is one of the leading news providers in Finland: according to Yle’s annual report 2023, the weekly reach of Yle’s online and mobile services (including Yle Areena) was 81% of the population (Yle’s annual reports, 2023).

The fine-tuning did not emphasise learning any particular topic category, so articles for the experiments were also selected from different topics. To fully analyse the model, headlines were generated using the most sophisticated implementation of Finnish T5 (t5-large-nl36-finnish) to see the results of the untuned model. The following hyperparameters were used for this network (see Table 2).

Hyperparameter	Value
max_length (tokenization)	512
truncation	True
max_length (generation)	40
length_penalty	0.2
num_beams	5

Table 2: Hyperparameters for t5-large-nl36-finnish

A case of correct formation of a compound word deserves a positive assessment. In the input text there were the words “*ravintolamoguli*”, consisting of two “*ravintola*” (‘restaurant’) and “*moguli*” (‘tycoon’), and “*yökerhojen*”, where “*yökerho*” (‘nightclub’), ‘-jen’ is the plural ending of genitive). Taking these words as a basis, the model generated a new compound word in the title “*yökerhomoguli*”.

It is also worth noting the case where the model attempted to capture the headline formation style characteristic of Finnish news outlets. For the resources participating in the study, the following headline structure is typical: a word or phrase summarising the essence of the message, containing a key named entity or indicating the source of information, followed by ‘:’ and the main part of the headline, e.g. “*Ranskalaistutkijat: Kännykkäkielto ei riitä - rajat tarvitaan kaikkeen ruutu-aikaan, myös television*” (‘French researchers: banning mobile phones is not enough — we need restrictions on all screen time, including TV’) or “*Ennuste sen kuin paranee: Luvassa jopa 24 astetta*” (‘The forecast is improving: up to 24 degrees is expected’). The model generated the following example headline: “*Tiktok: Nuoret etsivät tietoa ja seuraavat uutisia useimmiten*” (‘Young people are more likely to seek information and follow the news’).

For the generated and original headlines we calculated the average ROUGE (see Table 3).

	recall	precision	F-measure
Rouge 1	0.08	0.09	0.07
Rouge 2	0.03	0.03	0.02
Rouge 3	0.08	0.09	0.07

Table 3: Evaluation of the selected models.

However, the overall quality of the generated material is far from satisfactory: in most cases, the generated headline was accompanied by superfluous tokens that do not carry any semantics, for example: “*Moni suomalainen europa-parlamentissa eli mepit luopuvat*

paikastaan EU-parlamentissa. query: query: query:”. In this case, the sequence can be seamlessly cleaned of unwanted tokens without affecting the main part of the generated header.

This resulted in very low scores, indicating low efficiency, but one should keep in mind the already mentioned lack of indicativeness of the evaluation metrics. The rouge_2 values are significantly lower than rouge_1, which is quite expected for the special case of abstract summarisation. The values of rouge_1 completely coincide with those of rouge_1, which suggests that the longest common sequence consists of a single word.

5.1 Model fine-tuning: parameters

In the next step, the t5-mini-nl8-finnish network was fine-tuned using the following hyperparameters (see Table 4).

Hyperparameter	Value
test_size	0.2
num_train_epochs	5
per_device_train_batch_size	16
per_device_eval_batch_size	16
eval_steps	40
warmup_steps	50
max_len	100
max_length	20
num_beams	3

Table 4: Hyperparameters for the t5-mini-nl8-finnish network

The training loss (training_loss), which reflects how well the model performs on the training data, was 4.29. The loss on the test data (eval_loss) decreased from 16.92 (before training) to 1.04 (after training).

No further attempts were made to reduce the loss rates, as the error observed in the generated data made it meaningless to continue the search for more optimal training parameters.

Hyperparameter	Value
test_size	0.2
num_train_epochs	5
batch_size	28
max_input_length (TTTrainArgs)	50
max_output_length (TTTrainArgs)	50

max_length (TTSettings)	50
num_beams	8
do_sample	True
top_k	0
top_p	0.8

Table 5: Hyperparameters for byt5-base-finnish

Fine-tuning of the byt5-base-finnish network was carried out with the following values of hyperparameters (see Table 5).

5.2 Model fine-tuning: headlines analysis

As a result of training, the training loss (training_loss) decreased from 2.12 to 1.22, the loss on the test data (eval_loss) from 2.04 (before training) to 1.09 (after training).

Nominatives in this study refer to headings containing a proper name, e.g. “*Koskinen tunnusti syyllistyneensä törkeään talousrikokseen*” (‘Koskinen pleaded guilty to aggravated financial crime’) or “*Windows95man kertoo, että esiintymisasuun ei tule muutoksia*” (‘Windows95man reports that there will be no changes to the suit for the speech’).

Dotted headlines are slightly less common. Punctuated headlines were considered to be those that indicate the subject of the news item but do not reveal it in full, e.g. “*Dramaattiset tapahtumat saivat alkunsa lapsen syntymästä*” (‘Dramatic events caused by the birth of a child’) or “*Demokratian tulevaisuus on turvattava kaikilla tasoilla*” (‘The future of democracy must be protected at all levels’).

Most of the headlines were of the predicative type, in which the subject of speech and the predicate are included, thus forming an extended thesis, e.g. “*Sateet tulevat maan etelä- ja keskiosassa*” (‘Rains will take place in the south and centre of the country’) or “*Mestaruusjuhlat alkavat Tampereella maanantaina*” (‘Championship celebrations start in Tampere on Monday’).

Approximately half of all the headlines received can be categorised as noun headlines. They contain an indication of the general topic of the news article and actively fulfil the function of attracting attention, as the main details of the question asked are not disclosed. For example, “*Millainen on paras leivonnainen?*” (‘Which baked goods are the most delicious?’) or “*7. tammikuuta 2023 kuolleen naisen lähiomainen*

kertoo, miten se toimi” (‘A relative of the woman who died on 7 January 2023 tells how it happened’).

Slightly fewer headlines can be called transitive, characterised by a direct statement of the main facts. For example, “*Tulppaanifestivaali järjestetään Amsterdamissa*” (‘Tulip Festival will be held in Amsterdam’) or “*Tappara voitti Suomen mestaruuden*” (‘Tappara won the Finnish championship’).

A small proportion of headlines were categorised as opinion. These headlines consist of a reference to a famous person or expert and a subject heading. For example, “*Riikka Purra on huolissaan siitä, mitä hallitus tekee*” (‘Riikka Purra is concerned about what the government is doing’) or “*Laurence des Cars toivoo, että Mona Lisa saisi oman huoneen*” (‘Laurence de Carse hopes Mona Lisa will have her own room’).

We also found a few examples of clickbait headlines that aim to evoke emotions in readers as much as possible: “*Italian hallituksella ei ole mitään tekemistä*” (‘The Italian government has nothing to do’) or “*Maahanmuuttopolitiikka on karannut käsistä*” (‘Immigration policy is out of control’).

The network also demonstrated the ability to generate new compound words. For example, there were two words “*taitouinnin maajoukkue*” (‘synchronised swimming team’), which in the generated headline merged into one – “*taitouintimaajoukkue*” with a similar meaning. Special attention should be paid to the observance of the rules of alternation of the steps ‘*nt-nn*’. An example of the actual use of the word ‘*taitouintimaajoukkue*’ was found in the news section of the Finnish-language resource (uimaliitto.fi).

Nevertheless, the occurrence of factual errors could not be avoided. In the headline “*Suomen taitouintimaajoukkue kilpailee EM-kisoissa*” (‘The Finnish national team will compete at the European Championship’), the information was distorted, as the news item stated that the Finnish national team would qualify for the European Championship if they had enough points, which had not happened yet. In another case, the second key person was omitted and the meaning of the message was not fully disclosed: “*Lepistö pääsi opiskelemaan musiikkia*” (‘Lepistö has been accepted to study music’). In reality, it was about the directors of Sastamala Music College, Sini-Mari Lepistö and Tuomas Honkkila, who are also

students of the institution. The last example points to a surname error: “*Johanna ja Samuel Glassar ovat olleet yhdessä jo vuosia*” (‘Johanna and Samuel Glassar have been together for many years’), when in fact the news story refers to “*Johanna Puhakka*” and “*Samuel Glassar*”.

There was no difference in the generation of news for Yle and other sources whose materials were used in the training. The only thing that can be noted is that Yle is characterised by shorter headlines with simple constructions, so the generated headlines seem to be closer to the original ones. Equally important is that all generated headlines are characterised by varying degrees of extractiveness, and no cases of direct copying of a fragment from a news article were found. The results of calculating the mean ROUGE score were as follows (see Table 6).

	recall	precision	F-measure
Rouge 1	0.09	0.10	0.09
Rouge 2	0.02	0.03	0.02
Rouge 3	0.09	0.10	0.09

Table 6: Evaluation of the selected models.

We can note a slight increase in rouge_1 scores compared to the results of the t5-large-nl36-finnish network. Otherwise, the conclusions remained unchanged: rouge_2 is expectedly lower than rouge_1, the longest common sequence presumably consists of one word. The results of byt5-base-finnish were included in the expert questionnaire.

6 Results

The simpler network t5-mini-nl8-finnish as a result of fine-tuning stably generated the token ‘*Äijä*’, recognised as problematic because it does not carry any semantic load, but nevertheless fulfills the role of a full member in the generated headers.

The most productive was the byt5-base-finnish network fine-tuning, which is based on processing text directly at the byte level. This is most likely the reason why the network does not allow the generation of a problematic token. For the previously discussed Finnish T5 implementations, it is assumed that there were errors in the training of the model or tokenisers.

The results obtained with the pre-trained byt5-base-finnish network are diverse: by content, the generated headings were categorised into classes

such as nominative, predicative and punctuated; by the techniques used, they were categorised as transitive, nominative, opinion and clickbait headings. Among the 70 headings generated, only three factual errors were found.

The questionnaire included 5 randomly selected news articles that participated in the study. A total of 20 experts participated in the survey. For each criterion, the maximum amount of points a headline could receive if each of the 20 experts gave a score of 5 was 100 points.

The average score on a five-point scale for original headlines on the informativeness was 3.90, for generated headlines reached 3.57. The Student’s t-test resulted in a p-value above 0.05, so the null hypothesis of no statistically significant differences between the mean scores on the informativeness is accepted.

For the relevance, the original headlines received a mean score of 3.86 on a five-point scale, while the generated headlines received a mean score of 3.75. As in the case of testing the previous criterion, the t-test confirmed the null hypothesis that there are no statistically significant differences between the sample scores on the relevance criterion.

For original headlines, the mean score on a five-point scale on the impact is 3.69, while for generated headlines it is 3.67. For the t-test, the null hypothesis is also confirmed, namely that there is no statistically significant difference between the sample scores on the impact criterion.

Thus, the hypothesis put forward in this study is confirmed: there is no statistically significant difference between the scores obtained by the original and generated headlines according to the criteria of informativeness, relevance and impact. This means that the results of neural network work are close in quality to the results of human work.

7 Conclusion

In this paper, the results of a practical application of the Transformer-based Finnish T5 model were studied in a task of generating Finnish headlines from input news text. A critical evaluation was carried out, noting both the strengths and promising aspects of the different implementations of the model, as well as points still in need of improvement.

The byt5-base-finnish network performed the best. The experts’ evaluations indicate that the

network shows sufficiently high potential in all the criteria considered (informativeness, relevance and impact) to be useful, for example, as an auxiliary tool for creating headlines in Finnish. News content authors can use the headline variant proposed by the network as a basis for further work.

To be fully self-sufficient, the network still needs to improve its reliability, namely to get rid of factual errors. Such a problem can be solved by learning from a larger example dataset, but this also requires more powerful computational resources.

Acknowledgments

Maria Khokhlova acknowledges St Petersburg State University for a research project 124032900006-1 (inner ID 95435961).

References

- Bukhtiyarov A., Gusev I. 2020. Advances of Transformer-Based Models for News Headline Generation. In *Communications in Computer and Information Science*. P. 54-61
- Chekut' E.P. 2015. Functions of the headline as an actualiser of textual categories 'Funkcii zagolovka kak aktualizatora tekstovykh kategorij'. In *Actual questions of Germanic philology and methods of teaching foreign languages. XIX International Scientific and Practical Conference 'Aktual'nye voprosy germanskoj filologii i metodiki prepodavanija inostrannyh jazykov. XIX Mezhdunarodnaja nauchno-prakticheskaja konferencija'*. P. 81-83.
- Cohan A., Goharian N. 2016. Revisiting Summarization Evaluation for Scientific Articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. P. 806-813. <https://aclanthology.org/L16-1130.pdf>.
- Finnish-NLP. Hugging Face. <https://huggingface.co/Finnish-NLP>.
- Gusev I. O. 2019. Importance of copying mechanism for news headline generation. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019"*. P. 228-236.
- Koppatz M., Alnajjar Kh., Hämäläinen M., Poibeau Th. 2022. **Automatic Generation of Factual News Headlines in Finnish**. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 100–109, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.
- Liu Y., Gu J., Goyal N., Li X., Edunov S., Ghazvininejad M., Lewis M., Zettlemoyer L. 2020. **Multilingual Denoising Pre-training for Neural Machine Translation**. In *Transactions of the Association for Computational Linguistics*. Vol. 8. P. 726-742.
- Liu Y., Lapata M. 2019. **Text Summarization with Pretrained Encoders**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, P. 3730–3740, Hong Kong, China. Association for Computational Linguistics.
- Maples S. 2017. The ROUGE-AR: A Proposed Extension to the ROUGE Evaluation Metric for Abstractive Text Summarization. 10 p. <https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2761938.pdf>.
- Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., Zhou Y., Li W., Liu P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. In *Journal of Machine Learning Research* 21. P. 1-67. <https://arxiv.org/pdf/1910.10683.pdf>.
- Sokolov A. M. 2019. Phrase-Based Attentional Transformer For Headline Generation. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019"*. P. 615-621
- Xue L., Barua A., Constant N., Al-Rfou R., Narang S., Kale M., Roberts A., Z Raffel C. 2022. ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models. In *Transactions of the Association for Computational Linguistics* 10(01). P. 291-306.
- Yle's annual reports 2023. <https://drive.google.com/file/d/1wkBa5zWLGc3hh2FUwWHfFG9Vd8jHXMbg/view>.

Keeping Up Appearances—or how to get all Uralic languages included into bleeding edge research and software: generate, convert, and LLM your way into multilingual datasets

Flammie A Pirinen

Divvun

UiT—Norgga árkatalaš universitehta

Tromsø, Norway

flammie.pirinen@uit.no

Abstract

The current trends in natural language processing strongly favor large language models and generative AIs as the basis for everything. For Uralic languages that are not largely present in publically available data on the Internet, this can be problematic. In the current computational linguistic scene, it is very important to have representation of your language in popular datasets. Languages that are included in well-known datasets are also included in shared tasks, products by large technology corporations, and so forth. This inclusion will become especially important for under-resourced, under-studied minority, and Indigenous languages, which will otherwise be easily forgotten. In this article, we present the resources that are often deemed necessary for digital presence of a language in the large language model -obsessed world of today. We show that there are methods and tricks available to alleviate the problems with a lack of data and a lack of creators and annotators of the data, some more successful than others.

1 Introduction

In recent years, the landscape of language technology has changed quite rapidly, mainly with the advent large language models, but the overarching shift towards big data has been ongoing for longer. The problem with this shift is, that it is based on the big data for large majority languages, the inclusion of all the smaller languages, including all of the Uralic languages, has come as an afterthought if at all.

The expected solution for the continued sustainability of minority Uralic languages in the landscape of modern languages in the time of large language models is to “generate” more data. Ideally, by ‘generate’, the engineers in large language model contexts mean, that authentic written (or spoken) data needs to be created by native writers who should not make too many spelling or grammar

errors and write the most current normative form. This can be an unreachable goal for a language that has fewer than million speakers and writers who are not L1, as while the requirements for large language models are going down over time, they are still orders of magnitude larger that can plausibly be created by limited amount of writers and speakers in limited amount of time.

What we suggest in this paper is to carefully organise the initial work of corpus curation and creation around materials that are of high importance to the contemporary language technology community. We leverage existing resources and language technologies to minimise unnecessary and repetitive work by linguists and language professionals on the language data that is being worked on; automating what can be automated and re-using linguists annotation efforts is a key to efficient development of high-quality human verified gold data.

Our *research question* is, going from existing language technology resources: which tools are best suitable for launching and bootstrapping which resources. If language has usable electronic dictionaries, morphological analysers and generators, spell-checkers and so on, what can be used to effectivise the dataset creation and corpus curation. The question is especially interesting now, as there is a possibility to use contemporary multilingual large language models, as well as traditional rule-based, statistical and hybrid language models to perform various pre-processing and processing tasks.

Our *key contributions* from this article are: *the experimental framework* for others to compare and combine methods of gold data annotation for smaller languages, the *pipelines* from traditional rule-based annotations and LLM generations into concrete target formats, and the results of comparing some of the approaches for a low resource Uralic language along with recommendations of what is currently the most effective approach. As a side product we have created, curated and an-

notated beginnings of *several new datasets* for an under-resourced Uralic language.

We have laid out experimental computational linguistics data creation and annotation system that can use both existing rule-based tools as well as large language models to aid the process. One of the goals of this experiment and the approach is that we want to promote inclusion of more Uralic languages in all of the common language technology datasets. We are considering three separate approaches to help creation of annotated gold data:

1. rule-based generators and generative language models to generate a starting point for a data set, to be proof-read and re-annotated by humans,
2. rule-based analysers creating annotated dataset in legacy and ad hoc formats that are converted and organised into a starting point for human re-annotation, and
3. generative language models providing human annotators with starting points or improvements during annotation process

There are of course other possibilities as well, these are based on our previous experience and iterations with different datasets and projects. It must be noted that the goal here is to generate something comparable to human annotated gold corpus, so we are not planning to automate data generation or annotation. This has to be also contrasted to the reality of limited human resources for working with smaller Uralic languages, we do not necessarily have a possibility to hire 5 annotators to work on data full hours for several months, but to ask if the language experts who have other main jobs as language experts can use hours or two here and there on the task, this is one of the motivations of our experiment as well.

2 Background

The Uralic languages, especially besides the bigger national languages, are relatively under-resourced; the size of freely available texts is measured in millions of tokens or less. However, Uralic languages do have strong traditions of rule-based language technology. Also, lately, the large language model-based language technology has showed itself as a viable option for some use cases. Our approach to resource creation to overcome some of the under-resourcedness problem is thus to see if we can

leverage the existing technology to supplement the well-planned tactical selection of language dataset resources. In this article, we suggest curating and creating data that are highly relevant for the large language model building industry and also for the researchers of languages in language technology and linguists as well. While majority of industry and researchers concern themselves with basically English and maybe handful of commercially plausible majority languages of the world, we have discovered some related research both from the industry and the researchers who specialise in minority and under-resourced languages.

As one reference point, we study what technology companies and central research groups in LLM-based language technology have said about support for smaller language in the recent years; One reason for writing this article and its experiments is also inspired by these works: Meta and FAIR research group (Facebook’s AI Research) have released resources and studies under the moniker of *No language left behind* (NLLB) (Costa-jussà et al., 2022), also known for datasets and evaluation schemes under *Scaling neural machine translation to next 200 languages* (FLoRES) (Team et al., 2024). Unsurprisingly, this data set has so far included only Finnish, Estonian, and Hungarian when it comes to Uralic language inclusion. Alphabet and Google research have also been active on extending the range of languages supported under the name of *next 1000 languages* (Bapna et al., 2022). They have also published several research papers listing exactly the sources they use to gather information and data on the languages (Ritchie et al., 2024), this is directly useful information to know that, if you want to be included in Google’s considerations list of languages that might be supported or relevant, perhaps you want to have data in the resources and datasets they use.

The resources that we use in this articles experiments here have also been used for several years now in the academic community as the go-to resource to measure if your tool works with the given language. For example, the *Universal Dependencies* (UD) treebanks (Zeman et al., 2024), are used in a huge number of papers investigating computational linguistic methods in a large number of languages, including the annual shared tasks in syntactic parsing. It would thus appear that UD as a resource has passed the test of time. Secondly we have seen the *Unimorph* dataset, that concerns morphology of languages, has been used widely in the

research and applications. Namely with research of morphophonology and machine learning there have been regular shared tasks. We have explicitly left out parallel corpora and machine translations from this article for two reasons: firstly it is already a main focus of the large corporations and research groups working on the natural language engineering tasks and secondly our corpus selection is based on aiming to have a large subset of professionally human-translated texts as the source texts in these datasets, we find these are much more valuable than machine translated or post-edited texts, for the early phases of big data building we are in.

For the experimentation of this article I have chosen Inari Sámi as a target language; Inari Sámi is a Uralic language, that does not as of now have many of the resources that we are about to create. It is a low-resource Indigenous language with limited amount of speakers and written resources available, but an active speaker community that writes new texts. We have existing tools in rule-based language technology available from the well-known free and open source repository¹. Furthermore, the most recent versions of large language model -based systems have been seen to support Inari Sámi (instead of just refusing to handle it and deferring to professionals as earlier versions did). Finally, we have a computational linguist who is not a native speaker but is capable of working with the language and has contacts to language experts, we find this is sufficient for initial experimentation, but of course for serious language data building, more expert knowledge is needed.

For some the work on dataset creation there has been previous works, for example in Universal Dependencies and rule-based analyser there are existing methods that have been used for other existing uralic dependencies treebanks, such as the North Sámi (Tyers and Sheyanova, 2017) and Karelian treebanks (Pirinen, 2019). For generation of the UniMorph data, some of the datasets are generated based on rule-based generators (Batsuren et al., 2022), strictly speaking Wiktionary can also be considered as rule-based morphological generation, however, we have not found this mentioned explicitly in existing articles about unimorph.

3 Methods

Our experimentation concerns the use of existing language technology tools to help the creation of

¹<https://giellalt.github.io/lang-smn>

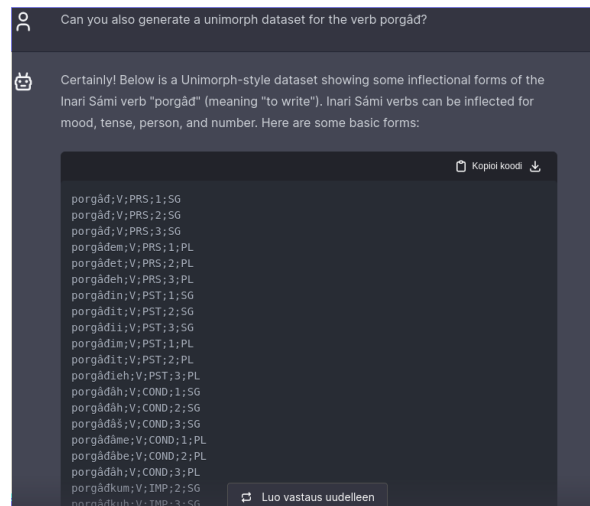


Figure 1: ChatGPT generating data for Inari Sámi UniMorph dataset.

the datasets while following the rules and ideals behind the given datasets. For example, when Universal Dependencies guidelines dictates that the dependency annotation must be manual or human made, we do not use the tools to generate unchecked 1-best annotations that would pollute the dataset. The most common strategy here is to give all plausible hypotheses from the automatic analysis to the linguist to post-edit, but another option is that the post-edited analyses are verified to be plausible analyses of the system (our end goal is to have a gold standard that agrees with the analyser and linguistic expertise).

For the existing rule-based systems, we have downloaded and installed well-known GiellaLT softwares, which are freely available from the GitHub with an open source licence (Pirinen et al., 2023).² The LLM experimentation is performed using a ChatGPT, the state-of-the-art chatbot interface to a closed-source, commercial neural network.³ We have chosen ChatGPT since it is the most popular one, it has freely usable version available for most Uralic language researchers even without expensive AI budget. An example of ChatGPT performing UniMorph dataset generation task can be seen in Figure 1.

When working with a preexisting computational linguistic, rule-based system, one of the main engineering efforts lies on the conversion. Although

²<https://giellalt.github.io>

³The version tested at the time of writing identifies itself as GPT-4, which was the newest model at the time we began experimenting but has probably been outdated by the time of the publication.

it sounds trivial, there is a lot of linguistic and engineering work to be taken into account here: the actual format of the analyses is rarely exactly the same, so a mapping needs to be devised, for example, converting “noun” analyses from +N to N; or NOUN. The mappings can also be 1:n or m:1, merging and joining ‘tags’, as well as more involved re-writings. There are a lot of other technical minor details related to such generations and conversions that are beyond the scope of this article, for example, we needed an algorithm that could remove duplicate forms that is aware of Unicode normalisation forms and folding to avoid having the linguist read word forms that look exactly the same several times. The topic of conversions in itself is large enough to deserve its own article,⁴ for the purposes of this article we will point the readers to our github repositiorium containing freely available scripts.⁵ Some examples of conversions are given in the Figure 2.

The experiments with LLMs are based on the currently available free ChatGPT interface prompted in English. We begin prompting with the most straightforward requests, e.g. “can you generate a unimorph annotated list of all word-forms Inari Sámi noun táálu?”, “create a CONLL-U annotated version of this sentence”, etc.

It might be noteworthy, that since our goal is inclusion of our Uralic languages in the relevant datasets, there is also a component of social engineering involved in all of the dataset creations. Merely producing text files that contain acceptable data is only a first step. The datasets we have selected to experiment with, the selection has been also based on the openness and documentation of the contribution process; all of the given datasets exist on GitHub, and the contribution process is detailed in the documentation and happens largely over GitHub only. This is in contrast to the commercially backed datasets mentioned earlier; while it would be very valuable to have all Uralic languages in the *No Languages Left Behind* and *Next Thousand Languages*, the way to contribute here is not immediately so obvious and available to larger audiences.

⁴we have attempted to write one such article, even at very condensed format it easily exceeds 8 pages that is the maximum for average conference article in language technologies.

⁵anonymised

4 Corpora and Data Selection

The corpora available for low-resource Uralic languages are scarce and limited. The whole corpora of publically available web crawl data is typically less than the millions of tokens that is often advertised as minimum requirement of large language models. Furthermore, the data that is available is limited by licences, quality, and genres: While some argue that all data that can be crawled is free to use for language technologies, in practice ethical use requires selecting only the data that has explicitly been licenced with a suitable licence, such as Wikipedia or data coming from governmental public domain records—or that has been personally licenced with the author for the specific use. That furthermore limits both quality—wikipedia data is written by language learners—and genres—government’s publication are mainly politics, healthcare and such.

In this experiment we have used primarily freely licenced data from Saami international corpora (SIKOR), (SIKOR, 2021) but we have also performed a short experiment on self-created and self-translated data that large language model should not contain from beforehand.

5 Experimental results

The main results of our experiment will be the actual datasets we can produce. To quantify the usefulness of the language technology tools we have measured post-edit distances. We have also performed a linguistic error analysis to quantify the errors made, the effect on the time/effort tradeoff is further discussed in the Section 6.

In our experiment in creating datasets for Unimorph, we used both the rule-based system and the LLM to generate the full datasets, that can be read and corrected by a human. The results of generating are shown in the table 1. The expected forms is based on the linguistic grammars we have available (Morottaja and Olthuis, 2023). We have measured the numbers of forms generated, Coverage counted as proportion of generated unique forms out of expected and Accuracy as proportion of fully correct forms and analyses of all generated. In general rule-based approach is close to the gold standard, which is expected from rule-based systems, the LLM has also generated a smaller subset of forms with lower accuracy.

In our experiments in Universal Dependencies annotation, we used the rule-based system to gen-

E.g. *Finite State Morphology to Unimorph*

táálu	táálu+N+Sg+Nom	<->	táálu	táálu	N;SG;NOM
táálust	táálu+N+Sg+Loc	<->	táálust	táálu	N;SG;LOC
tálustân	táálu+N+Sg+Loc+PxSg1	<->	tálustân	táálu	N;SG;LOC;PSS1S

E.g. *VISL CG 3 to Universal Dependencies*

```
"<mun>"
  "mun" Pron Pers Sg1 Nom @SUBJ> #1->2
:
"<juuhim>"
  "juuhâd" <mv> V TV Ind Prt Sg1 @FMV #2->0
:
"<vuolâ>"
  "vuolâ" N Sem/Drink Sg Acc @<OBJ #3->2
          ^^^
          |||
          vvv

# textid = example.1
# text = mun juuhim vuolâ
1 mun mun PRON Pron Pers Case=Nom|Number=Sing|Person=1|PronType=Pers 2 nsubj _ _
2 juuhim juuhâd VERB V TV Mood=Ind|Number=Sing|Person=1|Tense=Past 0 root _ _
3 vuolâ vuolâ NOUN N Sem/Drink Case=Acc|Number=Sing 2 obj _ _
```

Figure 2: Conversions between traditional rule-based analyses and target dataset formats

POS	Expected forms	RB forms	RB Cov %	RB Acc %	LLM forms	LLM Cov %	LLM Acc %
Nouns	58	100*	100 %		14	15 %	21 %
Verbs	57	55	96 %	99 %	22	39 %	0 %
Adjectives	51	61	100 %		14	20 %	10 %

Table 1: Unimorph dataset creation statistics. Expected forms is number of forms based on the grammar, RB from rule-based generator and LLM from large language model, Coverage and Accuracy measured in % units. * Some extra forms in rule-based model are due to allomorphy which was not accounted for expected forms.

System	Full WER	Dep WER
Rule-Based	0.47	0.22
LLM	1.00	0.52

Table 2: Caption

erate ambiguous listing of all potential readings of the sentence with annotations, according to the guidelines in previous works by Pirinen (2019), and asked LLM to generate similar hypotheses likewise. In Table 2 we measure the post edit distance of the sentences fixed and re-annotated, the error rates are calculated as $E = \frac{S+I}{N}$, where E is the error rate, S is number of substitutions made, I is the insertions made, and N number of readings (i.e. N is number of CONLL-U lines with an index). We do not have D for deletions since both methods generated correctly generated one token per token in the input and there are so far no retokenisation requirements (multi-word tokens, multi-token words etc.), however LLM missed some punctuation tokens causing an insertion to be required. The full error rate basically counts whole lines of CONLL-U when making matches and dep error rate just the dep field.

6 Discussion

We have tested rule-based and LLM-based annotations as a help in linguistic work. Currently, for morphology we get clearly better results with the rule-based tools and the results are good enough that it makes work on dataset creation more effective. If we analyse the errors that the systems make, we see that rule-based system includes some results with linguistically motivated potential errors, like wrong stem alternation or missing accent in a suffix. The errors in LLM generated version are that it just uses seemingly random suffixes with unchanged stem, it also uses some forms like cases that do not exist in Inari Sámi (but for example exist in Finnish), all in all cleaning this data would possible even be slower than writing the data by hand. When we error-analyse the dependency analysis the results get more interesting, like both starting points require quite a significant amount of work to get to gold-standard state, but this is also to be expected if reference the past experiences of UD annotation from converted or machine analysed starting point. What is interesting is that the LLM can sometimes generate quite accurate de-

pendency subgraphs of certain expressions, for example personal names, we assume this is due to them appearing in very similar form in existing English documentations, where high level dependency structure is the same even if there are slight variations in the morphological level.

There are a large number of different large language models and generative artificial intelligence that could possibly be used to experiment this and that is a common feedback we get. We are using a version of a popular LLM that is available to us, without excessive extra costs. This is also available to most researchers who are the target audience of this paper.

A common feedback we get, that there are various techniques that should be used for low resource setup, like fine-tunings, transfer learnings, in context learnings, prompting techniques and so on. We are experimenting in a situation where we start with zero data for the fine-tuning task, we are the ones who will create these data initially, so the use of such data will generally be a future research topic, after we have done the initial data creation. As the methodology here is extremely fast moving and outdates itself in matter of months, we try to begin by only importing either approaches that have been proven and stabilised, perhaps in majority language context, into out lesser resourced languages, or we can perform experimentation that does not tie up too much valuable and scarce resources. Another interesting future research question would be whether it is more beneficial and time-effective to fine-tune early or on-goingly, given the constraints in data and human resources we face in the processing of smaller Uralic languages. We have not found an easy enough recipe to do transfer learning that would not take us more time than actually working on the data creation as described by the approach of this article. Our impression is furthermore that there is currently ongoing research on this topic that we hope will yield some answers that are relevant to us as well.

It is exciting to see that, even if the large language models have rather disappointing accuracy in generating and annotation of smaller Uralic languages, they are able to generate something that is relevant to the task and occasionally some word-forms or annotations are even correct. This suggests that maybe with further fine-tuning, prompting, in-context learning, transfer learning, and so forth, there could be a usable version of LLM-aided language data annotation and generation in the fu-

ture.

One question for future work is of course how to integrate these findings to a workflow and softwares for annotation. In this experiment we used normal text editors and raw data formats for data annotation, which is suitable for programmers and short experiments, for the full scale linguistic annotation this would be integrated to a specific editor. And that raises the question of if the ideal way to help linguistic jobs would bear a user interface similar to what we get in the email post writing programs, office tools and programming editors today with a so-called *co-pilot*?

7 Conclusion

We performed several experiments to find out an efficient way of creating NLP datasets for smaller Uralic languages. We have found that using both existing rule-based technology and large language models can help rapid creation of the data, but neither approach is without its caveats. The gold standard remains fully human annotated data, but in lack of that it should be considered if we can achieve reasonable amounts of resources with computer-aided annotation modes.

Limitations

The experimentation on large language models is done using one closed source commercial system and is not reproducible at all, however, this is a common practice in the science of natural language processing in 2024.

The experiments were performed by language learner instead of native speaker or expert, the qualitative results may differ when language experts are working on the same pre-processed data.

Ethics

The large language models used in this experimentation have wasted an estimated several hundreds of litres of drinking water⁶ and not insignificant amount of energy (Strubell et al., 2019).⁷ If LLM method is taken in to use in the development of annotated gold corpora and data sets, this needs to be taken into consideration until the providers of LLMs resolve the excessive use of natural resources.

⁶<https://www.thetimes.com/uk/technology-uk/article/thirsty-chatgpt-uses-four-times-more-water-than-previously-thought-bc0pqsdr>

⁷<https://disconnect.blog/silicon-valley-is-sacrificing-the-climate-for-ai/>

No underpaid crowd-sourcers were involved in performing the linguistic tasks, all annotations and evaluations were made by fully paid colleagues.

References

- Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. *Building machine translation systems for the next thousand languages*. *Preprint*, arXiv:2205.03983.
- Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, et al. 2022. Unimorph 4.0: Universal morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855.
- Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Petter Morottaja and Marja-Liisa Olthuis. 2023. *Inarin-saamen taivutusoppi*. Sámediggi.
- Flammie Pirinen, Sjur Moshagen, and Katri Hiovain-Asikainen. 2023. Giellalt—a stable infrastructure for nordic minority languages and beyond. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 643–649.
- Tommi A Pirinen. 2019. Building minority dependency treebanks, dictionaries and computational grammars at the same time—an experiment in karelian treebanking. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 132–136.
- Sandy Ritchie, Daan van Esch, Uche Okonkwo, Shikhar Vashishth, and Emily Drummond. 2024. *LinguaMeta: Unified metadata for thousands of languages*. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10530–10538, Torino, Italia. ELRA and ICCL.
- SIKOR. 2021. SIKOR uit norgga árkatalaš universitehta ja norgga sámédikki sámi teakstačoakkáldat, veršuvdna 06.11.2018. <http://gtweb.uit.no/korp>. Accessed: 2024-10-01.
- Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. *Energy and policy considerations for deep*

learning in NLP. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

NLLB Team et al. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841.

Francis M. Tyers and Mariya Sheyanova. 2017. *Annotation schemes in North Sámi dependency parsing*. In *Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages*, pages 66–75, St. Petersburg, Russia. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Salih Furkan Akkurt, Gabrielè Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Lene Antonsen, Tatsuya Aoyama, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas Arican, Hórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Katla Ásgeirsdóttir, Deniz Baran Aslan, Cengiz Asmazoğlu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Mariana Avelãs, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Shabnam Behzad, Juan Belieni, Kepa Bengoetxea, İbrahim Benli, Yifat Ben Moshe, Ansu Berg, Gözde Berk, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnè Bielinskienė, Esmá Fatıma Bilgin Taşdemir, Kristín Bjarnadóttir, Verena Blaschke, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Johnatan Bonilla, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, António Branco, Kristina Brokaitė, Aljoscha Burchardt, Marisa Campos, Marie Candito, Bernard Caron, Gauthier Caron, Catarina Carvalho, Rita Carvalho, Lauren Cassidy, Maria Clara Castro, Sérgio Castro, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Liyanage Chamila, Shweta Chauhan, Yifei Chen, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Bermet Chontaeva, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Claudia Corbetta, Daniela Corbetta, Francisco Costa, Marine Courtin, Benoît Crabbé, Mihaela Cristescu, Vladimir Cvetkoski, Ingerid Løyning Dale, Philemon Daniel, Elizabeth Davidson, Leonel Figueiredo de Alencar, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Roberto Antonio Díaz Hernández, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Hoa Do, Kaja

Dobrovoljc, Caroline Döhmer, Adrian Doyle, Timothy Dozat, Kira Droganova, Magali Sanches Duran, Puneet Dwivedi, Christian Ebert, Hanne Eckhoff, Masaki Eguchi, Sandra Eiche, Roald Eiselen, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Soudabeh Eslami, Farah Essaidi, Aline Etienne, Wograiné Evelyn, Sidney Facundes, Richárd Farkas, Federica Favero, Jannatul Ferdousi, Marília Fernanda, Hector Fernandez Alcalde, Amal Fethi, Jennifer Foster, Theodoros Fransen, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Edith Galy, Federica Gamba, Marcos Garcia, Moa Gärdenfors, Tanja Gaustad, Efe Eren Genç, Fabrício Ferraz Gerardi, Kim Gerdes, Luke Gessler, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökürmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Kirian Guillier, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Takahiro Harada, Sam Hardwick, Kim Harris, Naïma Hassert, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Diana Hoefels, Petter Hohle, Yidi Huang, Marivel Huerta Mendez, Jena Hwang, Takumi Ikeda, Inessa Iliadou, Anton Karl Ingason, Radu Ion, Elena Irimia, Olájdé Ishola, Artan Islamaj, Kaoru Ito, Federica Iurescia, Sandra Jagodzińska, Siratun Janat, Tomáš Jelínek, Apoorva Jha, Katharine Jiang, Mayank Jobanputra, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritván Karahóga, Andre Kåsen, Tolga Kayadelen, Sarveswaran Kengatharaiyer, Václava Kettnerová, Lilit Kharatyan, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Petr Kocharov, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Alexey Koshevoy, Natalia Kotsyba, Barbara Kovačić, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Adrian Kuqi, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Kris Kyle, Käbi Laan, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Lauren Levine, Cheuk Ying Li, Josie Li, Keying Li, Yixuan Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Yi-Ju Jessica Lin, Krister Lindén, Yang Janet Liu, Nikola Ljubešić, Irina Lobzhanidze, Olga Loginova, Lucelene Lopes, Stefano Lusito, Anne-Marie Lutgen, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Ilya Makarchuk, Aibek Makazhanov, Francesco Mambrini, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărânduc, David Mareček, Katrin Marheinecke, Stella Markantonatou, Héctor Martínez Alonso, Lorena Martín Rodríguez, André

Martins, Cláudia Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Maitrey Mehta, Pierre André Ménard, Gustavo Mendonça, Tatiana Merzhevich, Paul Meurer, Niko Miekka, Emilia Milano, Aaron Miller, Karina Mischenkova, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhle, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Lng Nguyễn Thị, Huyèn Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Victor Norrman, Alireza Nourian, Maria das Graças Volpe Nunes, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Hulda Óladóttir, Adédayo Olúòkun, Mai Omura, Emeka Onwuegbuzia, Noam Ordan, Petya Osenova, Robert Östling, Annika Ott, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Teresa Paccosi, Alessio Palmero Aprosio, Anastasia Panova, Thiago Alexandre Salgueiro Pardo, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Giulia Pedonese, Angelika Peljak-Łapińska, Siyao Peng, Siyao Logan Peng, Rita Pereira, Sílvia Pereira, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Andrea Peverelli, Jason Phe-lan, Claudel Pierre-Louis, Jussi Piitulainen, Yuval Pinter, Clara Pinto, Rodrigo Pintucci, Tommi A Pirinen, Emily Pitler, Magdalena Plamada, Barbara Plank, Alistair Plum, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Rigardt Pretorius, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Robert Pugh, Tiina Puolakainen, Christoph Purschke, Sampo Pyysalo, Peng Qi, Andreia Querido, Andriela Rääbis, Alexandre Rademaker, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Joana Ramos, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Arij Riabi, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Ivan Roksandic, Mykhailo Romanenko, Rudolf Rosa, Valentin Roşca, Davide Rovati, Ben Rozonoyer, Olga Rudina, Jack Rueter, Paolo Ruffolo, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Aleksí Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Xulia Sánchez-Rodríguez, Manuela Sanguinetti, Ezgi Sanyar, Dage Särg, Marta Sartor, Albina Sarymsakova, Mitsuya Sasaki, Baiba Saulīte, Agata Savary, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Emmanuel Schang, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Sven Sellmer, Mojgan Seraji, Syeda Shahzadi, Mo Shen,

Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Maria Shvedova, Janine Siewert, Einar Freyr Sigurðsson, João Silva, Aline Silveira, Natalia Silveira, Sara Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Haukur Barri Símonarson, Kiril Simov, Dmitri Sitchinava, Ted Sither, Aaron Smith, Isabela Soares-Bastos, Per Erik Solberg, Barbara Sonnenhauser, Shafi Sourov, Rachele Sprugnoli, Vivian Stamou, Steinhór Steingrímsson, Antonio Stella, Abishek Stephen, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Daniel Swanson, Zsolt Szántó, Chihiro Taguchi, Dima Taji, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Mirko Tavoni, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Tarik Emre Tıraş, Sara Tonelli, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sveinbjörn Hórfarson, Vilhjálmur Hörsteinsson, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Elena Vagnoni, Sowmya Vajjala, Socrates Vak, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Uliana Vedenina, Giulia Venturi, Eric Villemonte de la Clergerie, Veronika Vincze, Anishka Vissamsetty, Natalia Vlasova, Eleni Vligouridou, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, John Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Shira Wigderson, Sri Hartati Wijono, Vanessa Berwanger Wille, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Qishen Wu, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Enes Yilandiloğlu, Olcay Taner Yıldız, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, He Zhou, Hanzhi Zhu, Yilun Zhu, Anna Zhuravleva, and Rayan Ziane. 2024. [Universal dependencies 2.14](#). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Scaling Sustainable Development Goal Predictions across Languages: From English to Finnish

Melany Macias, Lev Kharlashkin, Leo Huovinen, Mika Hämäläinen

Metropolia University of Applied Sciences

Helsinki, Finland

first.last@metropolia.fi

Abstract

In this paper, we leverage an exclusive English dataset to train diverse multilingual classifiers, investigating their efficacy in adapting to Finnish data. We employ an exclusively English classification dataset of UN Sustainable Development Goals (SDG) in an education context, to train various multilingual classifiers and examine how well these models can adapt to recognizing the same classes within Finnish university course descriptions. It's worth noting that Finnish, with a mere 5 million native speakers, presents a significantly less-resourced linguistic context compared to English. The best performing model in our experiments was mBART with an F1-score of 0.843.

1 Introduction

The list of 17 sustainable development goals (SDGs) established by the United Nations (UN) has gained significance in assessing the societal, humanitarian, and environmental impact of companies in EU. This is particularly relevant for large companies compelled to include robust sustainability reporting in their annual reporting to authorities¹. As a response to this growing importance, numerous universities and educational institutions have incorporated the UN SDGs into their academic curricula. This development prompts a crucial inquiry into how educational institutions, at a higher administrative level, can ascertain which specific SDGs are being integrated into different degree programs (see [Kopnina 2020](#); [Chankseliani and McCowan 2021](#)).

To address this concern, adopting smaller local models —specifically distil-mBERT, mBERT, mBART, and XLM-RoBERTa— tailored to the content of course descriptions proves beneficial for

universities. Such an approach facilitates compliance with the General Data Protection Regulation (GDPR)², ensuring the preservation of data privacy more than reliance on commercial large language models. The selection of the models stems from a need to accommodate linguistic diversity and the intricacies of academic content, ensuring accurate SDG classification while respecting data confidentiality. Each model brings distinct advantages: distil-mBERT and mBERT for their efficiency and language coverage, and mBART and XLM-RoBERTa for their superior cross-lingual and contextual understanding capabilities, making them well-suited for analyzing Finnish and English course descriptions in the context of SDGs.

In our work, we experiment with the viability of producing SDG classification data automatically using a large language model (LLM) in English. We use this English only data to train several multilingual classifiers and study the scalability of these models to Finnish data. Finnish, with only 5 million native speakers, is considerably less resourced than English. This is not to say that Finnish would be particularly under-studied in the context of NLP, given the vast amount of different NLP applications available for the language (see [Hämäläinen and Alnajjar 2021](#)).

2 Related work

The examination of sustainable development within the realm of NLP has been approached from various perspectives, including investigations into fairness in NLP ([Hessenthaler et al., 2022](#)), exploration of poverty and societal sustainability through interviews ([van Boven et al., 2022](#)), analysis of argumentation mining ([Fergadis et al., 2021](#)), and community profiling ([Conforti et al., 2020](#)), among other aspects. Our approach distinguishes itself by striving to encompass all UN sustainable devel-

¹https://finance.ec.europa.eu/capital-markets-union-and-financial-markets/company-reporting-and-auditing/company-reporting/corporate-sustainability-reporting_en

²<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679>

opment goals and implementing them within an educational framework.

The effect of multilingual models has been studied before in several contexts such as sentiment analysis (Hämäläinen et al., 2022) and persuasion detection (Pöyhönen et al., 2022) using parallel data in several languages. The findings suggest that translation strategies have a huge impact on the performance of the models.

Regarding the educational aspect of our study, there exists ample prior research on integrating Sustainable Development Goals (SDGs) into teaching methods (Collazo Expósito and Granados Sánchez, 2020; Rajabifard et al., 2021; Kwee, 2021). However, this previous research is non-computational, and as far as we are aware, there is no existing computational research on this subject from the standpoint of Natural Language Processing (NLP).

3 SDG Data

Our dataset, containing 5988 entries from Metropolia University, was accessed via their API³, a resource available to staff and researchers upon formal request and approval by the university’s IT department. In acquiring and handling this data, we observed strict ethical standards, including anonymization of identifiable information and adherence to the university’s data use policies, ensuring the preservation of data confidentiality. This dataset, spanning from 2010 to 2023, encompasses a diverse range of courses across various departments and majors, offered in both Finnish and English. The time frame was chosen to provide a comprehensive collection of course materials, aiding in the robustness of our model training.

After the initial data collection, we utilized the Vertex AI API⁴ to conduct batch processing, a pivotal step in annotating each course description with the corresponding Sustainable Development Goals (SDGs). The API facilitated the automation of this task by allowing us to process large volumes of text data and generate labels that indicate the relevance of specific SDGs to the course content.

Our study focuses on the following SDGs: 3 (Good Health and Well-being), 7 (Affordable and Clean Energy), 8 (Decent Work and Economic Growth), 9 (Industry, Innovation, and Infrastructure), and 10 (Reduced Inequalities), selected for

³<https://wiki.metropolia.fi/display/opendata/REST-rajapinnat>

⁴<https://cloud.google.com/vertex-ai/docs/reference/rest>

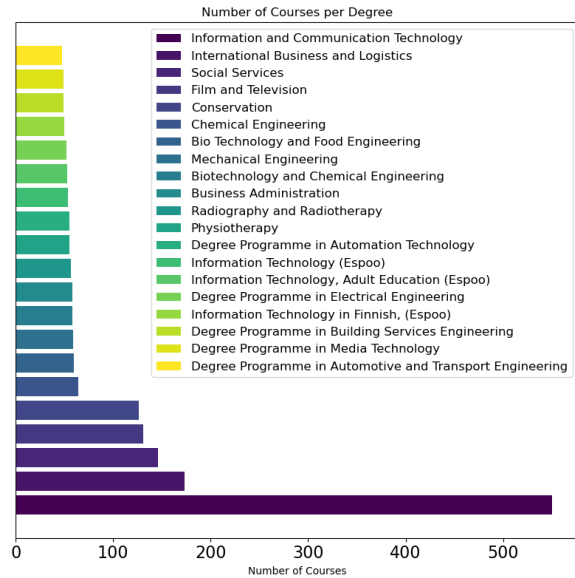


Figure 1: Distribution of courses per degree after the initial cleaning step.

their significant relevance to the university curriculum, as other goals were less represented, leading to potential data imbalance. The data is formatted as JSON objects for multilabel classification with multilingual models; one typical entry would look as follows: "input": "Heat Distribution Systems in Buildings, the student learns: The building’s heating power demand and its calculation. Different heat distribution methods and devices. Ways of adjusting the heating system. Water-circulating radiator and floor heating system. Dimensioning of pipework and selection of radiator, circulation pump, expansion vessel and safety devices. The student can calculate the heating power demand of the building, can dimension the pipework, radiators, circulation pump and expansion vessel and safety devices.", "labels": [0, 1, 0, 0, 0], where the input text is related to goal 7 Affordable and clean energy.

Therefore, the entries are structured as "input" - a detailed course description and "labels" - binary encoding indicating the course’s relevance to the selected SDGs.

Figure 2 displays the distribution of the Sustainable Development Goal (SDG) mentioned in the training dataset.

4 Cross-lingual Models for SDG Prediction

In this study, we utilized four advanced multilingual models: Distil-mBERT (Sanh et al., 2019),

Model	Strengths	Weaknesses
Distil-mBERT	Efficient with significant language understanding	Nuance understanding may be limited
mBERT	Comprehensive language comprehension	Finnish-specific tuning may be less
mBART	Strong in deep contextual understanding	High computational requirements
XLM-RoBERTa	Excellent in cross-lingual tasks	Possible compromise in Finnish depth

Table 1: Comparative Analysis of Multilingual Models

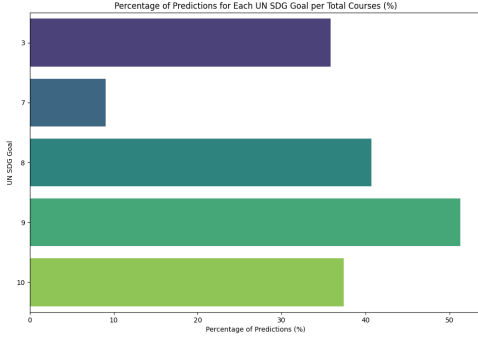


Figure 2: Distribution of SDG mentions within the training dataset.

mBERT (Devlin et al., 2019), mBART (Tang et al., 2020), and XLM-RoBERTa (Conneau et al., 2020), each uniquely suited for processing Finnish, a language with limited NLP resources. These models were selected for their balance between computational efficiency and linguistic depth, crucial for handling Finnish.

Our dataset, split into 70% training, 15% validation, and 15% testing, primarily consisted of English for training and validation, with Finnish reserved for testing. This approach was intended to test the models’ transfer learning capabilities from English to Finnish.

Each model underwent fine-tuning for multilabel SDG classification using a PyTorch-based (Paszke et al., 2019) framework. Key steps in the training process included tokenization, encoding, and employing a BCEWithLogitsLoss function, as shown in Equation (1).

$$\text{BCELoss} = -\frac{1}{N} \sum_{i=1}^N [y_i \cdot \log(\sigma(x_i)) + (1 - y_i) \cdot \log(1 - \sigma(x_i))] \quad (1)$$

where $\sigma(x_i)$ is the sigmoid function applied to the model’s output for the i^{th} sample, y_i is the true label, and N is the number of samples.

This study not only demonstrates the effectiveness of these models in a multilingual context but also sheds light on the scaling behavior of LLMs, particularly in adapting from high-resource to low-

resource languages. The findings provide valuable insights into the adaptability of multilingual models, with a special focus on Finnish, illustrating the broader applicability of these models in diverse linguistic settings.

The exploration of these multilingual models in predicting SDGs in Finnish highlights significant insights into the scaling behavior and adaptability of LLMs. Our strategic data split, along with the tailored training and architecture of each model, demonstrates our approach in tackling the challenges of language representation within NLP. This methodology is particularly pertinent in understanding how LLMs perform across languages with varying resource levels.

5 Results and Evaluation

Table 2 provides an overview of the model’s performance based on micro-average scores. The F1-Score, which balances precision and recall, showcases the models’ effectiveness in multilabel classification across Sustainable Development Goals (SDGs). Notably, the mBART outperforms the others with an F1-Score of 0.843, indicating its robustness in handling diverse SDGs.

Model	Precision	Recall	F1-Score
distil-mBERT	0.749	0.557	0.547
mBERT	0.798	0.678	0.716
mBART	0.825	0.867	0.843
XLM-RoBERTa	0.842	0.824	0.829

Table 2: Models performance based on the micro scores

In addition, figure 3 depicts the F1 scores, which represent the harmonic mean of precision and recall for each SDG. Our in-depth analysis of the classifiers’ performance on a per-label (per SDG) basis provides subtle insights into their prediction capability and limitations.

These per-label findings highlight the importance of model architecture and training corpus diversity in addressing the unique linguistic issues given by each SDG. The variations in performance

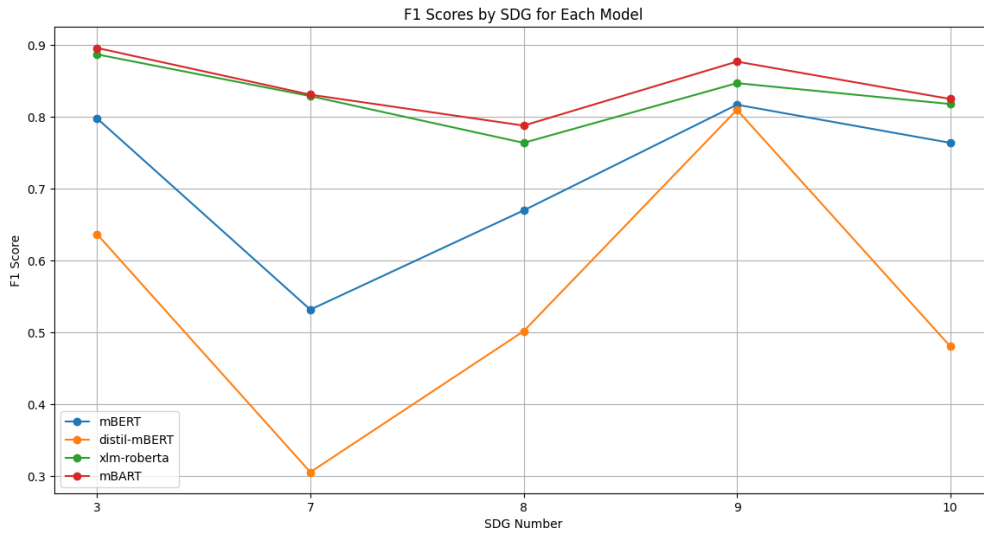


Figure 3: F1 Scores by SDG for Each Model

across objectives indicate that, while certain SDGs are well-represented and easier to forecast with current NLP models, others require additional research and targeted data enrichment to improve model performance.

6 Conclusions

This study takes a novel approach to assessing the integration of the United Nations Sustainable Development Goals (SDGs) into university curricula. We investigated the effectiveness of multilingual classifiers in adapting to Finnish data, a language with significantly fewer resources in the field of Natural Language Processing (NLP). Our study sought to ascertain the scalability of these models in recognizing SDG-related content within Finnish university course descriptions.

The findings show that multilingual models can bridge the language gap effectively, with notable success in identifying SDG-related content across multiple languages. The performance varied across SDGs, with some models excelling in some areas while struggling in others. This variation emphasizes the significance of model selection based on the target language’s specific characteristics and linguistic nuances.

Our work, which aligns with workshop talks on scaling behavior across linguistic settings, advances the NLP community’s understanding of multilingual model applicability in resolving language resource inequities by utilizing Finnish as a case

study. The study opens the door to further investigation into the use of multilingual models for other low-resource languages.

In conclusion, the study successfully demonstrates the feasibility of using English-trained multilingual models to process and analyze data in Finnish, a language with limited resources. This method not only provides a useful tool for educational institutions’ sustainability reporting, but it also contributes to a better understanding of multilingual NLP applications.

7 Limitations

While our study provides insights into the use of multilingual models for Sustainable Development Goal (SDG) prediction in Finnish, it is crucial to consider several limitations:

- **Subset of UN SDGs:** Our research focused on a subset of the UN SDGs. Extending our approach to encompass all SDGs would provide a more comprehensive understanding of the models’ capabilities across a broader range of sustainability topics.
- **Model Size and Performance Trade-offs:** We employed models like ‘mBART-large-50’, ‘distilbert-base-multilingual-cased’, ‘bert-base-multilingual-cased’, and ‘xlm-roberta-base’, each varying significantly in size and architecture. A detailed comparative analysis of these models reveals notable trade-offs

between their sizes and their precision and F1 scores. Larger models tend to offer better performance but at the cost of increased computational resources and complexity. This aspect is particularly relevant in the context of scaling behavior in LLMs.

- **Language Resource Limitations:** While Finnish is considered a low-resourced language in NLP, it is still better represented than many of the world’s approximately 7,000 languages. Our findings for Finnish may not directly translate to other low-resource languages, especially those with very limited digital presence or NLP tools.
- **Domain Specificity:** Our study was confined to the academic context of Metropolia University of Applied Sciences. The models’ performance may not generalize to other educational institutions, especially those offering a different range of academic disciplines.
- **Potential Model Biases:** Our classifiers, while effective in a controlled environment, may have learned an oversimplified version of the problem domain. There is also a risk of unknown biases when these models are applied in real-world settings.

In light of these limitations, our study should be viewed as a stepping stone towards understanding the scalability and adaptability of multilingual models in handling low-resourced languages, particularly in the domain of educational sustainability reporting.

8 Ethics statement

Our study’s focus on Finnish and English in multilingual models raises concerns about their performance and potential biases in other languages, especially those underrepresented in NLP research. While we ensured adherence to data privacy and consent in our methodology, the limited scope, centered on one university’s curriculum, may not fully represent other educational contexts or disciplines.

The findings highlight the need for broader linguistic representation in NLP models to ensure fairness and mitigate biases. Future research should extend to diverse languages and educational settings, adhering to ethical research standards and

prioritizing equitable representation in NLP applications.

References

- Maia Chankseliani and Tristan McCowan. 2021. Higher education and the sustainable development goals. *Higher Education*, 81(1):1–8.
- Leslie Mahe Collazo Expósito and Jesús Grados Sánchez. 2020. Implementation of sdgs in university teaching: a course for professional development of teachers in education for sustainability for a transformative action. *Sustainability*, 12(19):8267.
- Costanza Conforti, Stephanie Hirmer, Dai Morgan, Marco Basaldella, and Yau Ben Or. 2020. [Natural language processing for achieving sustainable development: the case of neural labelling to enhance community profiling](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8427–8444, Online. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Aris Fergadis, Dimitris Pappas, Antonia Karamolegkou, and Haris Papageorgiou. 2021. [Argumentation mining in scientific literature for sustainable development](#). In *Proceedings of the 8th Workshop on Argument Mining*, pages 100–111, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Mika Hämäläinen and Khalid Alnajjar. 2021. The current state of finnish nlp. In *Proceedings of the Seventh International Workshop on Computational Linguistics of Uralic Languages*, pages 65–72.
- Mika Hämäläinen, Khalid Alnajjar, and Thierry Poibeau. 2022. Video games as a corpus: Sentiment analysis using fallout new vegas dialog. In *Proceedings of the 17th International Conference on the Foundations of Digital Games*, pages 1–4.
- Marius Hessenthaler, Emma Strubell, Dirk Hovy, and Anne Lauscher. 2022. [Bridging fairness and environmental sustainability in natural language processing](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages

7817–7836, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Helen Kopnina. 2020. Education for the future? critical evaluation of education for sustainable development goals. *The Journal of Environmental Education*, 51(4):280–291.

Ching Ting Tany Kwee. 2021. I want to teach sustainable development in my english classroom: A case study of incorporating sustainable development goals in english teaching. *Sustainability*, 13(8):4195.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. [Pytorch: An imperative style, high-performance deep learning library](#). In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Teemu Pöyhönen, Mika Hämmäläinen, and Khalid Alnajjar. 2022. Multilingual persuasion detection: Video games as an invaluable data source for nlp. In *Proceedings of the 2022 DiGRA International Conference*. DiGRA.

Abbas Rajabifard, Masoud Kahalimoghadam, Elisa Lumantarna, Nilupa Herath, Felix Kin Peng Hui, and Zahra Assarkhaniki. 2021. Applying sdgs as a systematic approach for incorporating sustainability in higher education. *International Journal of Sustainability in Higher Education*, 22(6):1266–1284.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. [Multilingual translation with extensible multilingual pretraining and finetuning](#).

Goya van Boven, Stephanie Hirmer, and Costanza Conforti. 2022. [At the intersection of NLP and sustainable development: Exploring the impact of demographic-aware text representations in modeling value on a corpus of interviews](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2007–2021, Marseille, France. European Language Resources Association.

Kola Saami Christian Text Corpus

Michael Rießler
University of Eastern Finland
michael.riessler@uef.fi

Abstract

Christian texts have been known to be printed in Kola Saami languages since 1828; the most extensive publication is the Gospel of Matthew, different translations of which have been published three times since 1878, most recently in 2022. The Lord’s Prayer was translated in several more versions in Kildin Saami and Skolt Saami, first in 1828. All of these texts seem to go back to translations from Russian. Such characteristics make these publications just right for parallel text alignment.

This paper describes ongoing work with building a Kola Saami Christian Text Corpus, including conceptual and technical decisions. Thus, it describes a resource, rather than a study. However, computational studies based on these data will hopefully take place in the near future, after the Kildin Saami subset of this corpus is finished and published by the end of 2024.

In addition to computation, this resource will also allow for comparative linguistic studies on diachronic and synchronic variation and change in Kola Saami languages, which are among the most endangered and least described Uralic languages.

1 Religious text production in Kola Saami languages

Religious texts constitute a significant part of the earliest documented data for all four Kola Saami languages, chiefly translations of Christian texts which started to be created in the same period of time for Akkala Saami, Kildin Saami, and Skolt Saami. (No similar Christian texts are known to exist for Ter Saami, though.) They include the Lord’s Prayer and the complete Gospel of Matthew, each in different languages and versions, but also several other texts. The oldest text is from 1828, the

youngest from 2022; the references of this paper include a full list of sources.

Notable are the recent texts created by Alexandra Antonova, in particular her Kildin Saami translation of Arapović’s *Jesus Friend of Children* – a shorter version of *Children’s Bible* including Lord’s Prayer – and her completely new Kildin Saami translation of the Gospel. This text includes two different new translations of Lord’s Prayer printed at the end of the book together with a translation of Apostles’ Creed.

This book also includes a preface written in Kildin Saami by a non-Saami author. This text is relevant too because its language uses Christian metaphors and Christian symbols are explained, while using biblical terminology. Another relevant text is a prayer written by Saami author Jekaterina Korkina in Kildin Saami and Russian, with which she introduced a literary publication of hers (Korkina 2005).

These new texts not only add data to the corpus in terms of quantity, but allow for interesting comparative linguistic studies into various dimensions. This is particularly true because the idiolect of the recent translator Antonova (born 1932) is more than 100 years younger than that of Arvid Genetz’s native speaker informant Parfenty Pyanov (born 1821), while both speakers seem to have the same dialectal background due to their family ties in the original *Kiillt siida*. Furthermore, the comparison between Antonova’s two translations of different New Testament texts – created within a period of about a decade, but including interesting deviations in spelling and terminology – may potentially unravel some linguistic mysteries around her own derivation of the Kildin Saami orthography standard. Note also that the fragment of an intermediate manuscript version of Antonova’s transla-

Table 1: Currently included texts

Year	Text	Language	Word tokens	Status
1826	(1828) Lord’s Prayer	Kildin	59	Finished
1826	(1828) Lord’s Prayer	Skolt		Planned
1826	(1828) Lord’s Prayer	Skolt		Planned
1876	(1878) Lord’s Prayer	Kildin	60	Finished
1876	(1878) Matthew (1–22)	Kildin	13,114	Not proofread
1876	(1878) Matthew (23–28)	Akkala	5,014	Not proofread
1876	(1879) Lord’s Prayer	Kildin	62	Finished
1876	(1879) Matthew (1–22)	Kildin	13,149	Not proofread
1876	(1879) Matthew (23–28)	Akkala	5,001	Not proofread
≤1894	(1894) Matthew (1–28)	Skolt		Planned
≤1895	(1895) Lord’s Prayer	Skolt		Planned
≤1895	(1895) Primer	Skolt		Planned
≤1996	(1996) Jesus Friend of	Kildin	8,180	Finished
≤1996	(1996) Lord’s Prayer	Kildin	63	Finished
1999	(1999) Orthodoxy	Skolt		Planned
≤2008	(2010) Matthew (1)	Kildin	322	Finished
2005	(2005) Prayer	Kildin	76	Finished
≤2009	(2022) Lord’s Prayer	Kildin	60	Finished
≤2009	(2022) Matthew (1–28)	Kildin	18,215	Not proofread
≤2014	(2022) Apostles’ Creed	Kildin	71	Finished
≤2014	(2022) Lord’s Prayer	Kildin	58	Finished
2022	(2022) Preface	Kildin	559	Finished

tion (including the complete first chapter) is also available to corpus studies because it has been published in a research paper (Jermolajeva 2010).¹

Furthermore, the Kildin Saami subcorpus includes a small amount of data relevant to study Kildin Saami learners’ language. In addition to the abovementioned preface, written by L2 learner Scheller, lines 16:22 through 16:28 of Gospel of Matthew were translated by Scheller (born 1977) and Elvira Galkina (born 1965).² The first is a language researcher and language activist. The latter has become known as poet and author of children’s books and song lyrics in Kildin Saami language, although she describes herself as having full L1

¹Approx. 50 individual words from Antonova’s manuscript are also listed in a paper by Bakula (2016, pp. 18–19) and could potentially be used for comparison.

²Both have published a relatively significant amount of other texts in Kildin Saami, which are not Christian but are available and could potentially be compared too. See, e.g. the Wikidata Query Service (<https://query.wikidata.org/>), where relevant metadata for titles with Galkina resp. Scheller as author or translator can be found easily.

speaker proficiency only in Russian.³

Since literacy for the Akkala Saami has never been established and no newer written language published, the six chapters from the Gospel are the only existing orthographic texts in Akkala Saami.

Regarding Skolt Saami, the corpus is incomplete. New relevant texts have been produced in contemporary Skolt Saami, but at the current state of this research I have not had the time and resources to identify and catalogue all existent Skolt Saami texts. The only exception is a small pictorial dictionary named *The What, Why, and How of Orthodoxy* (Kasala 1999). Since this work has a parallel version in Finnish, it is perfectly suited for the current project.

In addition to incomplete coverage of Skolt Saami, I’ve also not yet systematically searched for secondary or tertiary reprints of original texts to include them in the cor-

³See the archived version of her professional CV at <http://web.archive.org/web/20240404155219/https://www.masu.edu.ru/special/fip-saami/files/CVГалкина.pdf> (2024-04-04).

pus. But plenty of them exist, first of all Genetz’s own reprints of his transcripts of the Gospel in Genetz (1879b) and Genetz (1891). But also his Lord’s Prayer was reprinted in Bergholtz (1894). Sjögren’s Lord’s Prayers were reprinted in Dalton (1870). Note also the more recent reprint of all these earlier versions in Németh (1991).

2 Corpus data

All mentioned Saami text sources from the 19th century are in the public domain. The same is true for versions in other languages which are all potentially useful for text parallelization but not included here.

Some of the Kola Saami texts were available in digital form earlier, others were digitized and proofread by me. Also Markus Juutinen (University of Oulu) – with whom I exchanged significant parts of these data – digitized and proofread texts for his abovementioned study. In addition, I worked together with Sergey Nikolaev (a Saami from Russia, today living in Oulu, Finland). Later I started uploading texts to Wikisource,⁴ where proofreading and indexing has since been continued with the help of collaborators, who I don’t know personally.

Parts of the 1878 edition of the Gospel – currently including chapters 3 through 10 – have been structured and made available as a corpus by the Lingvodoc project led by Julia Normanskaja (ILS RAS, Moscow). This corpus is structured in XML (at the levels of chapter, verse, word, and bound morpheme) and includes the original orthography, a Russian translation, tokenization of the original orthography and a translation of each token in contemporary Kildin Saami, and a morphological interlinearization with glosses.⁵

The user rights for Jesus Friend of Children were cleared by the Language Bank of Finland already in 1989, when the printed book was digitized in a project led by Pirkko Suihkonen

⁴See, for instance, the index for Kildin Saami: https://wikisource.org/wiki/Category:КӀЛЛТ_САМБ_КӀЛЛ (2024-10-11).

⁵See https://lingvodoc.ispras.ru/corpora_all (2024-10-11). The resource consists of one single file and does neither include metadata about its origin or any specification of a user license. But the sole originator seems to be Viktoria Bakula, professor and specialist of Kola Saami languages at Murmansk Arctic University.

(University of Helsinki).⁶ The digital data was stored in pre-processed form – including text files with pre-Unicode encoding and including OCR errors – when I got in contact with the Language Bank in 2015. I was allowed to create a working copy of the repository for my own research. After having fixed the encoding and rebuilt Antonova’s spelling with the help of a Pearl script and additional manual correction, and gave my improved version of the corpus back to the colleagues in Helsinki. But unfortunately, the Kildin Saami data has still not been published by the Language Bank.

The user rights for the new translation of the Gospel have yet to be cleared,⁷ but the copyright laws of Finland and the European Union principally allow the use of such data as research material – including communicating it as part of research activities – even without a specific agreement with the copyright holders. This includes the typical processes for text and data mining of printed texts: digitizing as well as digital storing and processing.⁸ It is also legal to publishing fragments as data illustrations for the purpose of teaching or in scientific publications, like in this paper.

However, more specifically defined user rights for Antonova’s translation of the Gospel will hopefully lead to an open corpus publication in the future. Ideally, this can be done using the functional user interface Korp, for instance at GiellaLT in Tromsø, which promotes Open Science and with whom Scheller has been collaborating for several years.⁹ But also the Korp platform at the Language Bank

⁶The metadata in the repository, dated July 10. 1998, specify that The texts of the computer corpus of Kildin Sámi have been donated to the University of Helsinki by the Institute for Bible Translation (Stockholm, Sweden) to be used as research material. Reference to the corpus has to be made in papers in which it is used as a source.

⁷According to the publisher, i.e. the Stockholm branch of Institute for Bible Translation, copyright is held by the correctors of the text (researchers Elisabeth Scheller at the Arctic University of Tromsø and Elvira Galkina at the Arctic University of Murmansk) and the legal heir of the translator (Antonova’s son Sergey Antonov from Lovozero).

⁸This refers to the exceptions in the EU Directive on Copyright in the Digital Single Market, which apply to text and data mining in academic research. National laws in EU countries follow the Directive. The name of the relevant Finnish law is (in Swedish) Upphovs-rättslag, see <https://www.finlex.fi/sv/laki/ajantasa/1961/19610404> (2024-11-01).

⁹See <https://sanj.oahpa.no/about/> (2024-10-11).

Table 2: Parallel text fragments from Lord’s Prayer in Kildin Saami; the two versions from 1876 origin from the same spoken recording, which was first transcribed and later represented in Cyrillic orthography; all versions but the first represent the one and the same dialect.

Speaker	Dialect	Text
1828 (unknown, b. \leq 1800)	Arsjogk	[...] Pafs lãndfj tono namme. [...] Amin.
1876 (Pyanov, b. 1821)	Kiillt	[...] а,nn pa,zxuv tøn' nomm; [...] Amin.
1876 (Pyanov, b. 1821)	Kiillt	[...] анъ пазьхув тонэ нэм, [...] АМИН
1996 (Antonova, b. 1932)	Kiillt	[...] святэ ляннч нэмм Төн; [...] Зоӓбэль
2014 (Antonova, b. 1932)	Kiillt	[...] Я пассьлувант нэмм Төн; [...] АМИНЬ.
2022 (Antonova, b. 1932)	Kiillt	[...] Анъ пассьювв нэмм Төн; [...] АМИНЬ.

of Finland – where similar parallel corpora for other Uralic languages are already available¹⁰ – would be a logical option.

With the exception of Schekoldin’s primer, all texts easily allow for alignment to parallel versions. These versions exist in between the Kola Saami languages: a) Lord’s Prayer in Skolt Saami (currently 2 versions) vs Kildin Saami (5 versions, plus one orthographic derivation), b) Gospel of Matthew in Skolt Saami vs Kildin Saami (chapter 1; one version in Skolt Saami and three versions in Kildin Saami), c) Gospel of Matthew in Skolt Saami vs Kildin Saami (chapters 2–23; one version in Skolt Saami and two versions in Kildin Saami), d) Gospel of Matthew in Skolt Saami vs Akkala Saami (chapters 24–28, one version each), and e) Gospel of Matthew in Kildin Saami vs Akkala Saami (chapters 24–28, one version each). But all of them can also easily be aligned with other language versions of the same texts, first of all to the Russian sources of the Saami translations.

Another dimension for parallel alignment results from the fact that the Akkala Saami and Kildin Saami translations of the Gospel published by Genetz were first documented in phonemic script (first published 1879) and later normalized by Genetz in Cyrillic orthography (first published 1878).

The overview in Table (1) lists the subparts of the corpus and the current state of their completion (year refers to the date of origin (\leq marks a terminus ante quem), the data of first publication is shown in parenthesis; word tokens may be due to corrections later).

It seems that the very existence of these par-

allel Christian texts has been known in general, but not in detail by all researchers in the field. For instance, a set of phonological studies by Bakula (2016) and Normanskaja (2016)¹¹ ignores the existence of Pyanov/Genetz’s 1876 translation of the Gospel as a phonemically exact transcript and builds on the orthographic version instead. This is an omission which made the results significantly less useful.

Also the work with the new translation of the Gospel would likely have profited from a more complete overview of earlier texts. Deducing from the description of the translation and edition process in Scheller (2022) the two text correctors (Scheller and Galkina) were not aware of all different earlier versions of the Lord’s Prayer, not even Antonova’s own. And Scheller doesn’t mention in the preface that Antonova’s earlier translation of New Testament texts would potentially be related to her new translation of Gospel of Matthew. See, for instance the Sermon on the Mount, which Antonova translated in two different versions. This may be counterintuitive for readers, even if both versions are idiomatic Kildin Saami.¹²

3 Technical procedures and conventions

Building this corpus has been carried out for two decades already as part of the author’s

¹¹These papers were reprinted with minor modifications as Normanskaja and Bakula (2022) without reference to the original work.

¹²The 1996 translation by Antonova was published by the Helsinki branch of Institute for Bible Translation, which specializes in the Uralic languages of Russia and supported by an editorial team. The 2022 translation, published by the Stockholm branch, lacked resources for thematic editorial checks. They had to rely on the competence of the text originators and could only support typography and typesetting (Brane Kalcevic, email 2024-10-08).

¹⁰See <https://clarino.uib.no/comedi/editor/lb-2020021121> (2024-10-11).

work with the Kola Saami Documentation Project (KSDP)¹³ but did not aim at more than a convenient corpus of interesting data samples until very recently. Work on this corpus has also never been funded by means of a specific project grant.

At present, the corpus is stored and versioned in a private GitHub repository,¹⁴ because parts of it are protected by copyright and can only be shared with research collaborators.

All original texts have either been digitally copied from other repositories or digitized by means of OCR by me before being modelled in XML. The data is encoded in UTF-8.

XLM markup follows the conventions of KSDP (cf. Blokland et al. 2015, pp. 12–14). There are other, more common formats available for modelling corpus data nowadays than XML (e.g. JSON). But XML has been the format of choice for KSDP because its data already includes a large amount of speech recordings and even video recordings, all of which are annotated and time-aligned in XML with the help of the tool ELAN.¹⁵ Adding written corpus data in the same structure (even though time-alignment is not relevant for written data) makes cross-corpus searches very simple. On the other hand, the used XML structure is consistent and well documented and can therefore easily be converted in other formats if future users prefer to do so.

Since the original intention of this project was different from digitization projects run by archival institutions or libraries, original pagination is not modelled in these corpus data. Also, all non-textual graphical details on the original pages are ignored because this corpus is aimed to serve linguistic research.

All texts are first chunked at the chapter level (if they are longer than one chapter). This chunking resulted in separate files which can be called “corpus sessions” (and which are conceptually equal to corpus sessions consisting of one continuous speech recording, e.g. an interview or a procedural, in the case of mul-

timedia corpus data for Kola Saami). In the case of the Gospel, each corpus session is chunked for verses, in order to keeping the original indexes for parallelization.

The different versions of Lord’s Prayer are not printed in one consistent verse structure, but manual alignment is simple for this short text and done based on verses Matthew 9:6–13 throughout all versions. Thus, parallel locations in the different versions of the Gospel – including Lord’s Prayer – are linked to each other by means of a pointer to chapters and verses.

Whereas the verses in the Gospel are relatively long and often include several sentences, other texts are chunked for sentences. This is how I typically also chunk my other written corpus sessions because sentences are conceptually equal to utterances in my spoken corpus data.

Textual structure at larger levels (headers, empty lines, paragraphs, etc.) is modelled by means of additional markup, added by me in the text if needed.

No further lexical, morphological, or syntactic tagging of the corpus has been carried out so far. Currently, I focus on the consistent and complete structuring of the Kildin Saami parts and complete proofreading of the Akkala Saami and Skolt Saami parts. But inspection and even systematic filtering of many morphosyntactic forms is already possible using RegEx and lists of bound and free grammatical markers.

4 Preliminary linguistic observations

The parallel data in Tables (2) and (3) – illustrating language use in worlds almost 200 years apart from each other – clearly show that the Kildin Saami language has not changed substantially since 1826.

These versions are relatively similar in terms of syntax, morphology, and lexicon. But there are also differences, some of them may indicate language change, others are due to different choices by the translators, or perhaps translation errors. For instance, Antonova’s syntax is clearly more involved than the older translation. Perhaps this is because it tries to reproduce the underlying Russian constructions.

Antonova had been translating very produc-

¹³A description of the early stages of this project is found in Rießler and Wilbur (2007).

¹⁴[https://github.com/langdoc/KSCTC/\(2024-11-21\)](https://github.com/langdoc/KSCTC/(2024-11-21))

¹⁵ELAN was originally created for building, annotating, and searching multimedia corpora, see <https://archive.mpi.nl/tla/elan> (2024-10-18).

Table 3: Parallel text of Matthew 1:1 published 2022 (originators Antonova/Scheller/Galkina), 2010 (Antonova), and 1878 (Pyanov/Genetz) – compared with a North Saami translation from 1998.

Speaker	Language	Text (Matthew 1:1)
1878 (Pyanov)	Kildin	Isus Xristos, Dəvid a ₁ lk', Ābram a ₁ lk' pūldəy s̄ā'n'.
2010 (Antonova)	Kildin	Авраам Альк Давид Альк Иисус Христос пуллдэгк.
2022 (Antonova)	Kildin	Йисус Христос, Авраам Альк, Давид Альк пуллдэгк.
1998	North	Dát lea Jesus Kristusa, Dávveda bártni ja Abrahamá bártni, sohka.

tively since the beginning of her writing in the 1980s. Her work is clearly based on the intuition of a fully proficient and active L1 speaker. But it is not much informed by earlier literary work, not even her own work. This can be seen in her different variants of Lord’s Prayer (Table 2). Note, for instance the creative translation of “Amen” with a Saami discourse marker in her 1996 version. This seems to originate from a sudden inspiration but was revoked again later and instead the Russian form of this declaration is used.¹⁶

Interesting is also the order of the possessive pronoun. Antonova puts it after the head noun like in the Russian original, even though the constituent order in Saami is much stricter than in Russian and would normally not allow this (see Table 2).

Also the comparison of the different versions of the Gospel reveals interesting findings. Already the very first sentence (1:1) is recorded in three different versions, including a fragment of the unpublished manuscript by Antonova which was mentioned by Scheller (2022). This sentence describes Jesus Christ’s descent after David and Abraham, thus in chronological relation to the Babylonian captivity. Syntactically, this sentence consists only of a noun phrase in all three versions (see Table 3). But it can be observed in these examples that constituent order is different in the old translation compared to Antonova’s former version. Whereas Antonova uses a strict head-final order even for all intermediate constituents (which seems consistent with archaic Saami and reconstructed Uralic syntax), Pyanov/Genetz put only the lexical nouns in head-final position. The proper nouns in

the intermediate noun phrases are head initial (this syntax looks closer to Russian). Interestingly, in Antonova’s second version, the order of constituents is scrambled in a completely new way, which does not follow the logical content of the original biblical genealogy. This change in the constituent order may be due to a translation error, because Jesus Christ descends from David’s lineage (after the exile), who in turn descends from Abraham’s lineage (before the exile).

Thus, already in its current form, the Kola Saami Christian Text Corpus allows interesting studies on diachronic and synchronic variation and change in Kildin Saami. The next step will be the complete inclusion of the mentioned Akkala Saami and Skolt Saami texts. The availability of this resource will hopefully prompt new qualitative and quantitative linguistic studies on these Uralic languages in the future.

References

- Arapović, Borislav (1996). *Iiisus – paarne kaann’c*. Stockholm: Institute for Bible Translation.
- Bakula, Viktorija B. (2016). “Vokalizm p’ervogo sloğa v kil’dinskom dialekte saamskogo jazyka po dannym Jevang’elija ot Matfeja (1878)”. In: *Uralo-altajskije issledovanija* 3.22, pp. 13–33.
- Bergholtz, Gustaf Fredrik, ed. (1894). *The Lord’s Prayer. In the principal languages, dialects and versions of the World*. Chicago.
- Blokland, Rogier et al. (2015). “Language documentation meets language technology”. In: *IWCLUL 2015*. Ed. by Tommi A. Pirinen et al. Tromsø: The University Library of Tromsø, pp. 8–18.

¹⁶The spelling of this word not as АМИНЬ – with the so-called half-palatalization sign – clearly indicates this. In Kildin Saami, нь marks the voiced palatal nasal /ɲ/ which doesn’t occur in this word.

- Dalton, Hermann, ed. (1870). *Das Gebet des Herrn in den Sprachen Russlands*. Kaiserliche Akademie der Wissenschaften.
- Genetz, Arvid (1879a). “Orosz-lapp nyelvmutatványok. Máté evangélioma és eredeti textusok”. In: *Nyelvtudományi közlemények* 15.1, pp. 74–152.
- (1879b). *Orosz-lapp nyelvmutatványok (Máté evangélioma és eredeti textusok)*. Budapest: Magyar Tudományos Akadémia.
- (1891). *Wörterbuch der Kola-Lappischen Dialekte nebst Sprachproben*. Helsingfors: Finska Vetenskaps-Societeten.
- Gospoda mij Iisusa Christa Pas’ Jevangelie Matveest (1894). Archangel’sk.
- Jermolajeva, A. S. (2010). “Sv’aščennye teksty v p’er’evod’e na saamskij jazyk (istoričeskij obzor)”. In: *Bogoslovije, istorija i praktika misij*. Ed. by A. B. Jefimov and L. N. Ivanova. Moskva: Izdatel’stvo PSTGU, pp. 111–117.
- Kasala, Kalevi (1999). *Ortodokslažvuõd mâi’d, mōõzz, mä’httceerkavteâd keârjjaž. Ortodoklaž Noõri Lett*.
- Korkina, Jekaterina N. (Mar. 2005). “Jímmel!” In: *Saa’m. Vaalt k rr’j-laaast* 2, p. 3.
- Maahtvjest Pa’ss Jevan’gelje (2022). Stockholm: Institute for Bible Translation.
- Mah’tveest Pas’-Jevangeli (1878). Helsinki: British and Foreign Bible Society.
- Németh, Zsigmond (1991). *96 gleiche Texte in uralischen Sprachen (Vaterunser)*. Szombathely.
- Normanskaja, Julia V. (2016). “Jevangelije ot Matfeja (1878) kak pamjatnik točnoj fikscii arhaičeskogo sostojanija kil’dinskogo saamskogo jazyka”. In: *Uralo-altajskije issledovanija* 3.22, pp. 34–45.
- Normanskaja, Julia V. and Viktorija B. Bakula (2022). “Jevangelije ot Matfeja na saamskom jazyke kak dokazat’el’stvo točnosti prasaamskoj r’ekonstrukcii J. Lehtiranta”. In: *Kirilličeskije pamjatniki na ural’skih i altajskih jazykah. 1. Grafikon’etičeskije osob’ennosti knig XIX v.* Ed. by Julia V. Normanskaja. Moskva, pp. 18–38.
- Odđa Testamenta (1998). Oslo: Norgga Biibbalsearvi.
- Rießler, Michael and Joshua Wilbur (2007). “Documenting the endangered Kola Saami languages”. In: *Språk og språkforhold i Sápmi*. Ed. by Tove Bull et al. 11. Berlin: Humboldt University of Berlin, pp. 39–82.
- Ščekoldin, Konstantin (1895). *Azbuka dlja loparej, živuščich v Kol’skom uezde Archangel’skoj gubernii*. Archangel’sk.
- Scheller, Elisabeth (2022). “Eevvtlessaa’nn”. In: *Maahtvjest Pa’ss Jevan’gelje*. Stockholm: Institute for Bible Translation, pp. 3–5.
- Sjögren, Andreas Johan (1828). *Anteckningar om församlingarne i Kemi-Lappmark*. Helsingfors: J. Simelii Enka.

Author Index

A Pirinen, Flammie, 123

Adamova, Maria, 115

Alnajjar, Khalid, 41

Borbála Bajzát, Tímea, 96

Bouton, Jules, 59

Chakraborty, Roshni, 12

Chanona Hernandez, Liliana, 49

Chelombitko, Iaroslav, 89

Davoodi, Laleh, 1

Dorkin, Aleksei, 104

Erina, Olga, 67

Ernštreits, Valts, 76

Gelbukh, Alexander, 49

Harkke, Ville, 1

Huovinen, Leo, 132

Hämäläinen, Mika, 36, 41, 132

Indig, Balázs, 96

Kabaeva, Nadezhda, 67

Kangur, Uku, 12

Kharlashkin., Lev, 132

Khokhlova, Maria, 115

Kolesnikova, Olga, 49

Komissarov, Aleksey, 89

Krasitskii, Mikhail, 49

Lüüsi, Lauri, 12

Macias, Melany, 132

Nazarenko, Alexander, 109

Ouzerrout, Samy, 29

Öörni, Anssi, 1

Partanen, Niko, 81

Purason, Taido, 104

Rießler, Michael, 138

Rueter, Jack, 41, 67

Sharma, Rajesh, 12

Sidorov, Grigori, 49

Sirts, Kairit, 104